



CAN UNCLASSIFIED



DRDC | RDDC
technologysciencetechnologie

Impact of Demographic Factors on Performance of Biometric Systems

Dr. John W. M. Campbell

Prepared by:
Bion Biometrics Inc.
236B Claridge Dr
Nepean ON K2J 5H1

PSPC Contract Number: W7714-135838-Task 30
Technical Authority: Brian Greene
Contractor's date of publication: March 2018

Defence Research and Development Canada

Contract Report

DRDC-RDDC-2018-C113

June 2018

CAN UNCLASSIFIED

IMPORTANT INFORMATIVE STATEMENTS

This document was reviewed for Controlled Goods by Defence Research and Development Canada (DRDC) using the Schedule to the *Defence Production Act*.

Disclaimer: This document is not published by the Editorial Office of Defence Research and Development Canada, an agency of the Department of National Defence of Canada but is to be catalogued in the Canadian Defence Information System (CANDIS), the national repository for Defence S&T documents. Her Majesty the Queen in Right of Canada (Department of National Defence) makes no representations or warranties, expressed or implied, of any kind whatsoever, and assumes no liability for the accuracy, reliability, completeness, currency or usefulness of any information, product, process or material included in this document. Nothing in this document should be interpreted as an endorsement for the specific use of any tool, technique or process examined in it. Any reliance on, or use of, any information, product, process or material included in this document is at the sole risk of the person so using it or relying on it. Canada does not assume any liability in respect of any damages or losses arising out of or in connection with the use of, or reliance on, any information, product, process or material included in this document.

Impact of Demographic Factors on Performance of Biometric Systems

Prepared By:

Dr. John W. M. Campbell

Bion Biometrics Inc.

March, 2018

EXECUTIVE SUMMARY

Biometric systems using facial, fingerprint and iris recognition have been widely deployed by Canada and other governments. These systems are used to issue passports and visas, permit travelers to cross the border and to allow employees access to physical and digital resources. Biometric performance, in terms of the risk of non-matches and of false matches, is critical to the successful and secure operation of these systems.

Recent research has uncovered significant performance differences in multiple biometric systems depending on the demographics (specifically age, sex and ethnicity) of the subjects whose biometric characteristics are being captured. In some cases, this simply leads to a bias that makes it more difficult for certain subjects to use the system. Existing studies show that this is the case for females with fingerprint recognition and facial recognition for older people with fingerprint recognition and iris recognition and for young children with all three modalities. In other cases, the performance difference leads to high rates of false matches and this can significantly undermine the security of the biometric system. So far this is known to be the case for certain ethnic groups and for young children when using facial recognition, but further studies are needed to investigate this phenomenon more thoroughly for all biometric modalities.

Several governments are actively researching this problem and Canada need to do the same. Unless systems, especially those using facial recognition, are tested properly to investigate the impact of demographic factors on system performance then they may be providing a false sense of security. In the worst case, one US study has shown that young children may be able to pass through a facial recognition based ABC system using a different child's travel document with a 40% success rate, which raises a huge problem for child trafficking and brings the entire security of the border into question. Proper testing and modifications to the ABC system can mitigate this problem, but as this research is very new, the only government agency currently known to be implementing such measures is New Zealand Customs Service.

CONTENTS

Executive summary	2
Introduction	5
Demographic factors.....	5
Measuring performance	7
Security Versus Convenience/Efficiency	8
Relevant Research – Academic Papers	9
Facial Recognition.....	9
Fingerprint Recognition	13
Iris Recognition	14
Summary of Findings From Academic Literature	15
Relevant Research – ISO Standards	16
ISO/IEC 29194:2015.....	16
ISO/IEC 30110:2015	17
ISO/IEC 20322 (draft).....	18
ISO/IEC 29196:2015	19
Summary of Findings From Standards.....	20
Relevant Research – Government Sponsored Studies	21
Facial Recognition – Passport Canada	21
Facial Recognition – US National Institute of Standards and Technology	23
Fingerprint Recognition	28
Iris Recognition	28
Summary of Findings From Government Sponsored Studies	30
Applications to automated border control Including CBSA Primary Inspection Kiosks (PIK).....	31

Applications to Canadian passport issuance System	32
Applications to Canadian Immigration Biometric Identification System (CIBIDS)	34
Final Summary of Research	35
Bibliography	37

INTRODUCTION

Biometrics are used in many areas of life, from unlocking mobile phones to facilitating border crossings to criminal investigations. The effectiveness of a biometric subsystem as part of a larger system is dependent on the performance of the biometric subsystem and the context in which it is designed to be used. In recent years, a lot of effort has been expended by both government and commercial entities in testing different biometric systems to measure their performance and fitness for purposes in different contexts. From the extensive but unpublished testing performed by Apple on their Touch ID fingerprint systems to the public testing of multiple facial recognition, fingerprint and iris recognition algorithms by the US National Institute of Standards and Technology, large amounts of time and money have been expended to quantitatively measure biometric system performance. The results have been significant as algorithm developers have used the information to improve their products. Facial recognition matching accuracy, for instance, improved by two orders of magnitude in NIST tests from 2002 to 2010 and by another 30% from 2010 to 2013 and by another 50% from 2013 to 2017.

As general biometric matching accuracy has reached levels where laboratory performance is acceptable for most purposes, research focus has moved to other areas. Protection against attacks such as spoofing or photo-morphing is now a major research area, as is managing operational environments so that operational performance is at least somewhat close to laboratory performance. As operational performance metrics have been studied quantitatively, however, researchers have come to realize that the traditional methods of measuring performance have limitations. One of these is that they are highly dependent on the specific group of individuals used in testing. If the test subjects are mostly young, well educated, Caucasian males (typical of test sets captured at US colleges and universities) then the results may not be applicable to elderly Asian females. In fact, it has become apparent that the test subjects need to match the demographic factors of the eventual users of the system in order to accurately predict the real system performance. Unfortunately, many biometric algorithms have been trained and tested for years with data sets that only represent a limited subset of demographic factors and so algorithms or systems that show excellent performance in most tests, may have hidden weaknesses for certain demographic groups.

DEMOGRAPHIC FACTORS

In order to perform scientific research, it is necessary to properly quantify the data that will be used for analysis. This can be difficult for demographic factors as there are many ways to divide

a population into subgroups. The most commonly used biometric modalities in government systems are face fingerprint and, to a lesser extent, iris. Focusing on these three modalities, an initial review of academic literature suggests that the demographic factors most likely to affect performance are age, sex, gender and ethnicity. Age is commonly understood, but different researchers seem to use different definitions of sex, gender and ethnicity. It is therefore useful to introduce the following definitions which will be used throughout this document.

Ethnicity - The state of belonging to a group with a common origin, set of customs or traditions

Gender - The state of being male or female as it relates to social, cultural or behavioural factors.

Note that gender is defined by society, culture and history, and varies from one culture to another and changes over time within each culture.

Sex - The state of being male or female as it relates to biological factors such as DNA, anatomy and physiology

Note that there are individuals who do not identify with the common definitions of male or female either in terms of gender or sex or both. These individuals may exhibit gender or sex characteristics that are different than those exhibited by the majority groups of male and female within their cultures.

Biometric matching accuracy, enrolment success and other aspects of performance can be affected both by sex and gender. The direct influence of gender on matching accuracy can be difficult to quantify, however, as the social and cultural practices associated with gender vary from culture to culture. For instance, males are more likely to perform manual labour in North America and Europe, but in some African countries, females are more likely to perform manual labour associated with agricultural production. Since manual labour can result in degradation of the quality of fingerprints and lower fingerprint matching accuracy, this is an example of a demographic factor that is related to gender but may be more significant for males in some cultures and for females in others. This aspect of performance is affected by gender correlated with ethnicity.

In most operational biometric systems there is no opportunity for individuals to identify their gender and the only information available is the sex field marked in an identity document. Typically, this is listed as Male, Female or Other/Unspecified. In the machine readable zone of a passport for instance, the choices for sex are "M", "F" or "X". This means that these simple categories are all that is available in large datasets and it is difficult to distinguish between the effects of sex and those of gender when correlating biometric performance with the available demographic information.

Biometric performance can vary substantially with a person's age. Effects such as skin elasticity, bone structure and comprehension of instructions can all affect the matching accuracy and usability of a biometric system. Separately, the process of ageing changes a subject's biometric characteristics such that biometric recognition accuracy generally decreases with the time elapsed between collection of two samples. Thus, biometric matching accuracy depends on both age and ageing. For example, enrolment of a face at age 5 and verification at age 15 will usually be much less accurate than enrolment of a face at age 45 and verification at age 55 because the aging effect is larger at the lower age.

Biometric performance also varies with 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, which is a reasonable proxy for ethnicity in some countries, but a rather poor one in others. Occasionally a survey is taken that allows individuals participating in a test data set to declare their ethnic identity and, in a few studies, images of the individuals have been examined and classified by a human into specific ethnic groups. None of these methods are perfect, but if there are significant biometric performance differences for a given ethnic group then a large study will still show this performance difference even if there are some misclassifications of ethnicity.

MEASURING PERFORMANCE

There are well defined performance metrics that are used to evaluate biometric systems. The most common of these are listed below, using definitions from the official ISO biometric vocabulary [1] available at <http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html>.

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

Note 1: Acceptance of the output of a biometric capture process for subsequent comparison will depend on policy. Failure to acquire includes failure to capture.

Note 2: Other possible causes of failure to acquire include poor biometric sample quality, algorithmic deficiencies and biometric characteristics outside the range of the system.

failure-to-acquire rate (FTAR) - proportion of a specified set of biometric acquisition processes that were failures to acquire

false match - comparison decision of match for a biometric probe and a biometric reference that are from different biometric capture subjects or different biometric characteristics from the same biometric capture subject

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

false non-match - comparison decision of “non-match” for a biometric probe and a biometric reference that are from the same biometric capture subject and of the same biometric characteristic

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

quality score - quantitative value of the fitness of a biometric sample to accomplish or fulfil the comparison decision

SECURITY VERSUS CONVENIENCE/EFFICIENCY

Most biometric systems require a trade-off between security and convenience or efficiency. Increasing the minimum acceptable quality score improves matching accuracy, but increases FTAR, which reduces convenience. Increasing the score threshold that defines a positive match reduces FMR, improving security, but also increases FNMR, reducing convenience if those who fail to match can't use the system or efficiency if they have to use a secondary process when the biometric fails to match. Typically, a biometric system uses fixed parameters for the minimum quality score, the maximum time to allow for a biometric characteristic to be captured, the comparison score threshold for a positive match, etc. that are based on the experience of the manufacturer in their internal testing or the testing of the system operator. These tests normally use an easily available set of test subjects. For manufacturers, it may be individuals in their office for live testing or one of several freely available data sets for offline testing. For the system operator, it may be local staff or a subset of operational users that were available during a pilot. These test groups are usually quite limited in their demographic composition and almost never selected specifically to cover an appropriate range of demographic factors. With the fixed set of system parameters, specific demographic subgroups may then negatively affect security or convenience or, in the worst case, both. This can be quantitatively measure by testing with a specific demographic group and measuring changes in FTAR, FMR or FNMR. If a full system test is not available, then quality score can also be used to investigate the impact of demographic factors, as a demographic group that exhibits lower quality scores will typically have higher FTAR or FNMR and, in some cases, higher FMR.

RELEVANT RESEARCH – ACADEMIC PAPERS

A survey of academic papers and other public documents with quantitative results on the impact of demographic factors reveals that there have been some interesting studies over the years that demonstrate the fact that demographic factors do affect biometric performance.

FACIAL RECOGNITION

Factors that influence algorithm performance in the Face Recognition Grand Challenge

Reference [2] looks at the matching accuracy of three matching algorithms on facial images from the US National Institute of Standards and Technology (NIST) Facial Recognition Grand Challenge. Each individual had multiple enrolment and verification images, resulting in a total of 134,760 match decisions. The probability of correct verification (1- FNMR) was calculated at 8 different values of FMR ranging from 1 in 10,000 to 1 in 10. The impact of different covariates on the probability of correct verification at all of these different values of FMR was combined together into a single likelihood ratio that kept all other factors equal while a single covariate was changed. This allowed numerous effects to be examined, including the impact of age, sex and ethnicity. It should be noted that in this reference, the authors incorrectly use the word “gender” to refer to sex.

All three 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 Algorithm B had a 2% higher probability of correct verification for males than for females, whereas Algorithm C had a 4% higher probability for males. Surprisingly, Algorithm A has a 3.5% lower probability of correct verification for males than for females, but since this probability of correct verification was in the single digits, the statistical reliability of data using algorithm A on this data set was somewhat questionable. The other two algorithms are therefore more reliable and confirm that 1:1 facial recognition is more difficult for female subjects than for male subjects.

The data set was also divided by ethnicity, but the only two ethnic groups with more than 13 subjects were Caucasian (253 subjects) and East Asian (91 subjects). For these two groups, Algorithm B showed 8% and Algorithm C showed 9% higher probability of correct verification for East Asian subject over Caucasian subjects. Of course, it is important to remember that in this case, the majority of subjects in this test were Caucasian, so it was not clear if the algorithms inherently worked better with the features of East Asian faces or if the East Asian faces were simply different from the majority of faces, reducing the probability of false matches.

Face recognition performance: Role of demographic information

Reference [3] looks at the matching accuracy of six algorithms on a large database of around one million mugshot images taken from the Pinellas County Sheriff's Office in Florida. 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, sex and ethnicity. These cohorts were selected so that all of the other demographic factors except the one being evaluated were approximately constant across the cohorts. Each cohort contained both a training set and a test set with one probe image and one gallery image for each subject, but the two sets were completely disjoint. The segment of the study that considered age had approximately 8000 subjects in each of the test and training sets for the 18-30 age range, 8000 in the 30-50 age range and approximately 2800 in the 50-70 age range. For the sex study there were approximately 8000 male and 800 female subjects in each of the test and training sets. For the ethnicity study, the FBI National Crime Information Centre code manual was followed so the main ethnic groups in the data set were identified as White, Black and Hispanic. There were approximately 8000 White, 8000 Black and 2800 Hispanic subjects in each of the test and training sets. Three of the six algorithms used in this study were commercial products and three were academic. One of the academic algorithms was trainable and so the disjoint training sets were used to investigate the impact of changing the training data for a specific algorithm.

The results are very clear. At a fixed false match rate of 0.1%, all six algorithms had a lower true accept rate ($1 - \text{FNMR}$) on the 18-30 age group than on either of the other two age groups. This difference ranged from 3% to 6% depending on the algorithm. There was less difference between the 30-50 age group and the 50-70 age group, with some algorithms reporting better matching accuracy on one and some on the other. This study suggests that older people are easier to recognize, but that the decreasing difficulty with age levels off somewhere in middle age. When the trainable algorithm was trained using only subjects from the training set for a particular age cohort, as opposed to training it on all age cohorts, then the true accept rate always increased for the cohort on which it was trained, but by a relatively small amount ranging from 1.3% to 1.8%. Interestingly, however, the true accept rate for all age cohorts increased when algorithm was trained only on the Age 18-30 cohort. It is not clear exactly why this result was observed, but it may mean that there is more individual distinctiveness among faces in the 18-30 cohort and therefore training the algorithm on this cohort improved its overall matching accuracy slightly.

For ethnicity, all six algorithms showed that the true accept rate at a FMR of 0.1% was lowest for the Black cohort, then 2% to 8% higher (depending on the algorithm) for the White cohort and then a further 1% to 7% higher for the Hispanic cohort. Training on a specific cohort was able to improve matching accuracy on that cohort by 1.8% for the Black cohort and 1.5% for the

White cohort, but reduced the true accept rate by 2.9% for the Hispanic cohort. This may be simply because the smaller number of subjects in the Hispanic cohort were insufficient to properly train the algorithm.

Finally, at FMR = 0.1%, the true accept rate for females was between 5% and 19% lower than for males, depending on the algorithm used. In this case, a version of the trainable algorithm trained only on females actually had a 0.6% lower matching accuracy on the female cohort and a version trained only on males only had a 0.2% higher matching accuracy on the male cohort. This suggests that there is an intrinsic difference between males and females that makes females more difficult for facial recognition algorithms to recognize. Whether this is related to sex or gender is not clear from this study.

Report on the FG 2015 Video Person Recognition Evaluation

Reference [4] looks at facial recognition of subjects in videos acquired as part of the Point-and-Shoot Face Recognition Challenge Problem (PaSC). These videos were acquired in Spring, 2011 on the campus of University of Notre Dame. They represent a 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 and because the subjects were not looking directly at the camera 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 1401 tripod mounted videos and 1401 handheld videos. Each matching algorithm was given two videos and required to return a match score indicating the likelihood that the subject in both videos was the same. This resulted in 3128 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 matching accuracy.

This study is different from the previous studies because it operates 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 that males were easier to recognize than females and that Asians were easier to recognize than Caucasians.

Demographic effects on estimates of automatic face recognition performance

Reference [5] looks at data from the NIST Facial Recognition Vendor Test (FRVT) 2006. In this study, the focus was on how the demographic distribution of the non-match pairs would impact the matching accuracy. 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 matching accuracy on each data set was measured by fusing together the match 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 matching scores for each image pair.

The images selected for this study were from the lower performing sets of images classified as “moderately difficult” and “difficult”. Set 1 contained 62% females and 38% males; 76% Caucasians, 13% East Asians and 11% other ethnicities; with 92% of subjects in the youngest age category of 18-29 years old. Set 2 contained 55% females and 45% males; 64% Caucasians, 21% Hispanics and 15% other ethnicities; and a fairly even spread of subject ages ranging from 18 to over 60 years old. Each image was then matched against all other images from the same data set to generate match scores and non-match scores. By selecting subsets of the non-match scores, it was possible to compare matching accuracy when there was no demographic matching (i.e. all non-matches were included), when the non-match images were matched by sex (i.e. each image was only matched against other images of the same sex), when the non-match images were matched by ethnicity or when the non-match images were matched by both sex and ethnicity.

As expected, the true accept rate at a specific false match rate became lower when the non-match pairs were more closely matched because the match threshold had to increase in order to maintain a specific FMR as the non-match pairs became more similar. This was true for both data sets. As an example, the true accept rate at FMR = 0.1% on the “moderately difficult” part of Set 1 was 79.25% for no demographic matching, 74.02% for sex matched pairs, 74.08% for ethnicity matched pairs and 68.51% for pairs with matching sex and ethnicity. The study also analyzed the results using different data subsets, 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 ethnicity and the matching faces of a different ethnicity.

The general conclusions were always the same. 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, for a fixed FMR, the match threshold must be increased and the chance of a false non-match is also increased. There are two practical lessons from this for operational systems. First, in order to correctly estimate the matching accuracy of a facial recognition system, it needs to 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. Second, 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 another individual with their approximate demographic characteristics.

It should be noted that all the academic papers listed were limited to studies of one-to-one verification outcomes. The demographic effects reported therefore do not automatically apply to identification systems using the same underlying recognition algorithms.

FINGERPRINT RECOGNITION

IDENT/IAFIS Image quality study

Reference [6] is a review of the Image Quality Study conducted by MitreTek Systems in 2000 to examine the use of the US Federal Bureau of Investigation's Integrated Automated Fingerprint Identification System (FBI 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 matching accuracy. Although their results are specific to the quality and matching algorithms that were in use by the IAFIS at that time, several of the findings should have general applicability.

First, the researchers found that there was poor correlation between the image quality score and the matching accuracy for an individual fingerprint image, except for those fingerprints with very low quality scores which were unlikely to be successfully matched. Second, 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. This led to the conclusion that "Clearly, performance and throughput will be engineering challenges for systems with large female populations." This shows that sex has a major impact on fingerprint recognition.

Impact of sex on fingerprint recognition systems

Reference [7] 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 [6].

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 matched against itself. For both the capacitive and the optical sensor, the matching accuracy for the female subset was better than for the male subset, in that the false non-match rate at every value of the false match rate below 0.1% was lower for the female subset. This is a surprising finding as the image quality score should be at least somewhat correlated to matching accuracy. In this case, however, the image quality score for female fingerprints was lower even though the matching accuracy for female fingerprints was better.

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

Reference [8] was a continuation of the work started in Reference [7]. 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 matching 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.

All three of these studies show a bias in fingerprint quality algorithms towards female fingerprints that may prevent female fingerprints from being properly processed even if they are of sufficient quality to be accurately matched.

IRIS RECOGNITION

It is very difficult to find academic papers that analyze the performance of top tier iris recognition algorithms. There are papers that review performance of iris recognition algorithms that were created by academics specifically for each study, but these are not relevant to commercial iris recognition. It may simply be that top tier iris recognition algorithms are so accurate that it requires very large databases to properly analyze performance and there are no suitable publicly available databases. This means that we don't currently have any relevant academic studies on how iris performance is affected by demographic factors. A few papers look at the impact of certain pathologies such as glaucoma on iris recognition and it is reasonable to expect a higher incidence of such pathologies among older people, but even these use very limited data or simulated data. There are, however, many papers describing how to use iris images to classify sex and ethnicity. Several of these papers report very high accuracy in determining sex and ethnicity from iris images, so there are clearly differences in the iris

images that are correlated with sex and ethnicity. Whether or not these affect the performance of commercial iris recognition algorithms is not yet clear.

Predicting Ethnicity and Gender from Iris Texture

Reference [9] is a sample paper that uses a database of five images of each of the left and right iris from a total of 60 Asian and 60 Caucasian subjects. In each ethnic group, 30 of the subjects were male and 30 were female. The authors divided the subjects into separate test and training sets so that no individual was contained in both, as they believed that there was correlation between the left and right iris of a single individual and so the data should be separated not just by iris but by subject. They trained multiple classifiers on the training data and were able to achieve 90.58% correct ethnicity prediction with their best classifier. Next, they trained the classifiers on data separated by sex into male and female groups. The best classifier only obtained 62% accuracy in predicting sex from the iris image. Other papers using different classifiers have performed much better for sex prediction, with prediction accuracy as high as 91% [10] being reported, but this paper looked at both sex and ethnicity together. In fact, the paper found that correct ethnicity classification was poorer for females than for males, but that correct gender classification was not affected by ethnicity.

SUMMARY OF FINDINGS FROM ACADEMIC LITERATURE

The academic literature reveals four important conclusions about the impact of demographic factors on the performance of biometric systems.

- Facial recognition seems to work better (in terms of lower FNMR for 1:1 verification) for older people versus younger people and for males versus females.
- Facial recognition seems to work better (in terms of lower FNMR for 1:1 verification) for Asians versus Caucasians and for Caucasians versus African Americans (denoted as Blacks by the FBI). It may also work better for Hispanics versus Caucasians.
- Retraining a facial recognition algorithm on data that includes large numbers of a specific demographic group usually helps improve the performance for that group, but this does not seem to work for females versus males.
- The chance of an imposter successfully impersonating a valid use of a facial recognition system increases if the imposter attempts to impersonate someone with similar demographic factors (age, sex and ethnicity) to their own. At a fixed match threshold, this can create a security risk.
- Fingerprints from females are usually classified as being of lower quality than fingerprints from males and this may result in females having more difficulty in using a fingerprint based system or even in being unable to use it because their fingerprints are

deemed to be of insufficient quality. Despite this, females' fingerprints don't appear to have lower performance than male fingerprints.

- Iris images show different characteristics for males than for females and for different ethnic groups, but it is unclear if this affects performance in any significant way.

RELEVANT RESEARCH – ISO STANDARDS

Since 2002, the International Organization for Standardization (ISO) has been developing biometric standards and technical reports through its ISO/IEC JTC-1 SC 37 Biometrics group. Some of these standards and technical reports include statements that are relevant to the impact of demographic factors on biometric performance. Unfortunately, very few standards provide reference research to back up the statements they make, as they are developed by consensus of experts from different countries and the ISO process does not require experts to provide the research that supports their opinions. The one exception is a new ISO Technical Report currently under development entitled “ISO/IEC TR 22116 Information technology - Identifying and mitigating the differential impact of demographic factors in biometric systems”. This technical report does provide reference papers and excerpts of results from specific government studies that are being shared by ISO member states to help understand the problem of bias in biometrics due to the impact of demographic factors. Most of the results discussed in this document have also been shared with ISO for inclusion in TR 22116, so there is no need for a separate discussion of TR 22116 in this section. The other relevant standards, with their less rigorous conclusions, are detailed below.

ISO/IEC 29194:2015

[ISO/IEC TR 29194:2015 Information Technology -- Biometrics -- Guide on designing accessible and inclusive biometric systems](#)

This technical report lists different ways in which individuals may find it difficult to use biometric systems and provides guidance to help make the systems more accessible for all individuals. Many of the factors considered in this document are more likely to be found in specific demographic groups and therefore systems need to make accommodation for these factors or those demographic groups will experience poor system performance. For instance, visual, auditory, motor and mental problems are all more likely to occur in the elderly. Similarly, young children also experience more difficulty with certain motor actions and may have difficulty reading signs or following directions. The list of the factors affecting individuals that may cause problems using a biometric system appears below.

- (Inability to) Perceive visual information

- People who are unable to perceive any visual information.
- People who have difficulty in perceiving visual information.
- (Inability to) Perceive auditory information
 - People who are unable to perceive any auditory information.
 - People who have difficulty in perceiving auditory information.
- (Inability to) Perform motor actions
 - People who are unable to walk unaided.
 - People who are unable to stand.
 - People who are unable to pitch, or yaw, or rotate head, or keep stationary.
 - People who are unable to raise and/or rotate arms/hands.
- (Inability to) Present physiological attribute
 - Unable to present the specified hand(s).
 - Unable to present specified finger(s) and/or palm(s).
 - Unable to present the specified eye(s) as attribute or as landmark.
 - People who are unable to present physical attribute within the specified field of the sensor.
 - Unable to present specified auditory input.
- (Inability to) Apply instructions due to mental impairment
 - People with cognitive or learning difficulties.
 - Where interaction and/or responses from system are counter intuition or familiarity.
- (Inability to) Follow guidance due to cultural discrepancies
 - People with language differences.

ISO/IEC 30110:2015

[ISO/IEC TR 30110:2015 Information technology -- Cross jurisdictional and societal aspects of implementation of biometric technologies -- Biometrics and children](#)

This technical report considers the impact of age, specifically of age below 18, on the implementation of biometric systems. It provides information on issues with multiple biometric modalities when used with children. Generally, biometric performance is degraded when young children are involved and systems need to either exclude or find alternative mechanisms to accommodate children below a certain age. The standard also examines the issues related to privacy and consent when collecting biometrics from children, but these are dependent on country specific legislation and are not directly related to performance issues. Specific findings and recommendations of this standard related to biometric performance are listed below.

Fingerprint patterns are fixed in the womb, but the size of the fingerprint increases as a child grows and this can cause distortions that make recognition difficult when the child has grown substantially in between enrolment and verification of the fingerprint. Young children have very small fingerprints and high skin plasticity which makes acquisition of a quality fingerprint very difficult. The standard concludes that children under four will frequently have problems using a fingerprint based system and those under eleven will have higher error rates than adults or teenagers.

Facial recognition can work with children over short periods, but accuracy diminishes significantly as the child's face and head change with growth. There are two periods when this makes facial recognition very difficult to use. The first is when the child is under the age of five and experiences rapid growth from a baby to a toddler. Facial recognition accuracy is very poor during this period. There is then a period of stability until puberty, but the significant changes caused by puberty mean that the false reject rate can be very high when images acquired before puberty are compared to images acquired after puberty. This means that automated matching of passport images at a border control point, for instance, is very difficult for younger teenagers if their passport is more than a couple of years old.

The iris has a unique pattern which is stable throughout life, but the size of the iris changes as a child grows and is not stable until the age of six to eight years. It is also very difficult for children under four to keep their eyes wide open and focused on an iris camera for the time required for iris acquisition, so most iris cameras have difficulty with children in this age group. The UIDAI programme in India has enrolled tens of millions of young children, but doesn't recommend the use of iris recognition for children under four.

ISO/IEC 20322 (DRAFT)

ISO/IEC TR 20322 Information technology - Cross jurisdictional and societal aspects of implementation of biometric technologies - Biometrics and elderly people

This technical report looks at the opposite end of the age spectrum and examines the issues that arise when trying to use biometric system with those over the age of 65. Poor vision and cognitive impairment are more common in the elderly and making it more difficult to use any biometric system. The standard suggests that biometric systems that will be used by populations with a large number of elderly people should include signage that is easy to read in larger font sizes and attempt to use additional markers such as tactile guides to assist in positioning fingerprints or flashing lights to attract attention.

Some pathologies that are more common in the elderly, such as arthritis or arthrosis can make it very difficult for elderly people to use fingerprint, finger vein or hand geometry based

systems due to the difficulty in correctly presenting the finger or hand. As individuals age the sweat glands are not as efficient at producing sweat and skin becomes drier, sags from the loss of collagen and elastin fibers, becomes thinner and loses fat. All these conditions decrease the firmness of the skin, causing wrinkles, and make it more difficult to obtain high quality fingerprints. This means that fingerprint systems are often unreliable for the very elderly and alternative modes of identification may be necessary. For instance, the US VISIT programme exempts travelers over the age of 79 from presenting their fingerprints when they cross the US border.

Iris recognition can also be problematic when used with elderly populations, as glaucoma or cataracts may reduce performance of iris recognition. Studies on this are divided and there does not seem to yet be a conclusion on whether this will cause a problem for operational systems. Any condition which significantly reduces vision quality, however, makes it more difficult to correctly present the iris to an iris recognition camera.

Facial recognition does not seem to be a problem for the elderly.

ISO/IEC 29196:2015

[ISO/IEC TR 29196:2015 Guidance for biometric enrolment](#)

One of the most important contributions to a successful biometric-based recognition system is a consistent enrolment service that generates the biometric data required for subsequent recognition of individuals. This standard provides a wealth of useful guidance on designing an enrolment system and is therefore applicable to all government programs that capture biometrics data for enrolment. Many of the recommendations are also useful at the verification stage to ensure that good quality biometric data can be captured.

Section 6.2.2.3 of this standard explains the metrics required to measure a successful enrolment. It points out the tension between the metrics related directly to enrolment, such as failure to enroll, number of retries during the enrolment process, etc. and metrics related to the subsequent verification process, such as false match rate and false non-match rate. Reducing the quality threshold for a successful enrolment will probably decrease the failure to enroll rate but will likely increase the false non-match rate. Therefore, the interests of the organization in charge of enrolment may be in conflict with the interests of the organization using the enrolment data. For instance, a passport agency typically wishes to ensure that very few submitted images are rejected, but this may result in an increase in errors when the passport images are used for matching at an automated border control system. Although this section does not specifically mention demographic factors, the same principle applies when making decisions to improve performance for specific demographic groups. If the fingerprint enrolment

quality threshold is lowered to prevent elderly people from experiencing failure to enroll errors, then the subsequent false non-match rate on verification for those same elderly people will likely be increased. One of the conclusions of this section of the standard is that metrics need to be collected and analyzed both from the enrolment and from any subsequent verification or identification operations so that the specific performance issues can be detected and mitigated. It even recommends periodic independent evaluations of each system. This is definitely required when the performance impact of demographic factors is being analyzed.

There are some specific references to demographic factors in this technical report that may affect error metrics for enrolment. For facial recognition, it is noted that some women may be uncomfortable removing their head garment or veil in the presence of a male and so it is advised that a female operator be available for these cases. It also mentions that very young and very old people both frequently have difficulty in being acquired by fingerprint systems. It also notes that facial recognitions systems sometimes have difficulty with certain ethnic groups or with people who wear glasses, which is more common among older people. Further studies are required, however, to quantify the significance of these issues.

In section 8.3.4 on tenprint systems, the standard notes that “The root cause analysis of problem areas is helped immeasurably if more detailed data are available relating distributions of NFIQ scores to specific age and ethnic groupings, analyzed by gender and NFIQ scores from previous enrolments (if available).”. This acknowledges the importance of age, gender and ethnicity as demographic factors that affect tenprint fingerprint systems such as those used by the RCMP.

SUMMARY OF FINDINGS FROM STANDARDS

Although the standards and technical reports published by ISO don’t provide details of the data used to determine their conclusions, they do make several conclusions relevant to the impact of demographic factors on biometric performance, as summarized below.

- Elderly people and young children will have more difficulty in using most biometric systems due to increased incidence of physical or mental impairment, including difficulty in understanding and following directions
- Fingerprint recognition has higher failure to acquire and failure to match errors for elderly people and is not recommended for those older than 79 years
- Fingerprint recognition may have problems achieving fingerprints of sufficient quality for young children, especially those under four years old

- Facial recognition performance is poor for children and teenagers, especially if there is an extended period between the enrolment and verification, as is typical with ePassports and other identity documents
- Children under four years old may have difficulty properly presenting their eyes to an iris recognition system

RELEVANT RESEARCH – GOVERNMENT SPONSORED STUDIES

As part of the current project on demographic factors causing bias in biometrics and the companion international research being conducted to support ISO TR 22116, several governments have begun to share research from analysis of large scale biometric databases or from operational systems. Contributions have already been received from IRCC and CBSA in Canada and from NIST in the US but performing such research and getting permission to share it can be time consuming and a lot more data is expected to be shared in 2018 and 2019. For now, this section highlights two studies for facial recognition and one for iris recognition.

FACIAL RECOGNITION – PASSPORT CANADA

Every time an individual applies for a Canadian passport, their facial image is searched against a large database of existing images. This is done to ensure that the individual has not previously applied for a passport using a different identity and also to ensure that they are not on any facial watchlists. If they have previously applied for a passport then a 1:1 match is performed against each of their previous passport images stored in the database to ensure that these do successfully match. Commercial 1:many facial recognition algorithms employ a variety of strategies to compare an image to a large database of other images and this is not the same as simply performing a 1:1 match of the subject or probe image against each of the millions of faces (the gallery images) in the database. The end result of the 1:many search is that a list of candidate faces from the gallery may be returned (if any candidates are above a minimum match threshold) and a human examiner then needs to review the candidate list and decide if one or more images from that list are actually matches to the probe facial image. The minimum match threshold can be set high enough so that no images are returned for human review, but this is not recommended as 1:many facial matching is still usually performed with a human in the loop. Another variable that is commonly manipulated is, N, the number of images returned in the candidate list. The metric then used to determine performance is the Cumulative Match Characteristic (CMC), which is the percentage of all searches where the probe image has a matching image in the gallery and that matching image is returned in the top N images that form the candidate list. As N increases, the likelihood of the correct matching image being

included in the candidate list increases, but so does the amount of time it takes the human examiner to process the candidate list, so the CMC as a function of N reflects a trade-off between accuracy and staffing requirements, which translates to a trade-off between accuracy and cost.

As part of its own evaluation of its face matching system, Passport Canada ran tests to compare the performance of the B4 and B7 versions of the commercial algorithm that they use for 1:many face matching. One part of this evaluation looked at the difference in performance between male and female probe images. As the table below shows, they isolated 110,072 probe images for males and another 114,401 probe images for females. They used the matching algorithm (either B4 or B7) to perform a search for each of these probe images against a gallery containing the matching probe image and an additional set of non-matching images. The non-matching images consisted of both males and females, and there were a total of 697,727 images in the gallery used for testing males and 757,793 images in the gallery used for testing females.

Sex	Probe	Probe Match	Gallery
Male	110,072	110,072	697,727
Female	114,401	114,401	757,793
Total	224,473	224,473	1,455,520

Table 1 - Passport Canada Databases for Testing for Sex Bias

The results of the 224,473 1:many matches are summarized in Figure 1 as a CMC Curve showing the percentage of searches where the probe match image was returned within the candidate list of length N (identified in the figure as Rank). It is important to note that both the B4 and B7 algorithms had higher CMC values for males than for females, but the B7 algorithm was significantly better in that the CMC value for females with B7 was higher than that of males with B4. The size of the performance gap between males and females was also reduced when moving from version B4 to version B7. This means that B7 showed better performance and less sex bias than B4, but that the sex bias still exists even with the newer algorithm. Although this difference is only 1.14% at Rank 1, this still represents an additional 1,304 images from the 114,401 female probes tested that were not found at Rank 1. It should be noted that Figure 1 was provided directly by Passport Canada and it uses the term “gender” to refer to the male / female difference when it is more like caused by sex, since sex is what is normally shown on a passport. This study shows convincing evidence of a sex bias against females in both versions of the facial recognition algorithm used by Passport Canada.

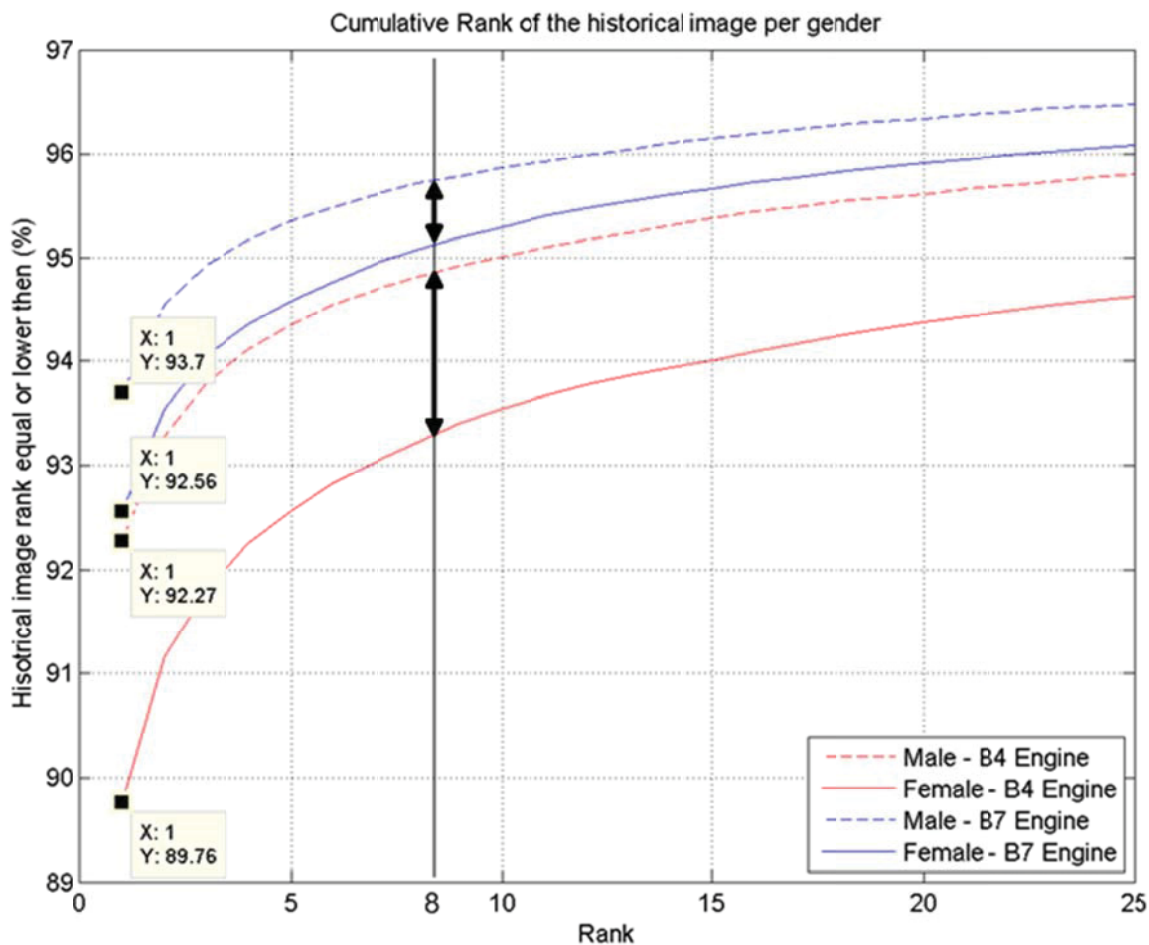


Figure 1 - CMC Curve for Passport Canada B4 and B7 Algorithms

FACIAL RECOGNITION – US NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

The US National Institute of Standards and Technology (NIST) has several different projects to evaluate biometric performance. One of these, the Ongoing Face Recognition Vendor Test (FRVT) [11] allows developers to submit algorithms at any time. Recently, NIST has realized the importance of demographic factors and has added them as a major consideration in the Ongoing FRVT. The submitted algorithms are tested using several databases, among them a set of nearly 250,000 visa photographs. Each photograph is accompanied by the subject's age, country-of-birth (which is used as a rough proxy for ethnicity), and sex. For each algorithm

tested, NIST reports multiple metrics related to biometric matching accuracy, including the following:

- How FMR and FNMR differ by sex. Figure 2 shows, for four algorithms, the FNMR and FMR measured for each sex at six different possible operating thresholds. In this experiment, impostors are of the same age, sex, and region of birth as the enrollee. The plots show that false match rates 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 enrollee. Figure 3 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 enrollee. If the imposters are selected to have the same sex and region of birth, then the FMR can be exceptionally high (up to 40% for children aged 0-4, which is 400 times the global imposter rate of 0.1%).
- How FMR depends on the country of the birth of the impostor and enrollee. Figure 4 shows the FMR expected when an impostor from one region of the world is compared to an enrollee from another. FMR varies globally, with high FMR in East Asia, South Asia, and sub-Saharan Africa. Notably FMR increases across regions, for example Caribbean to sub-Saharan Africa. NIST's report [11], which includes analogous figures for all algorithms tested shows that most algorithms exhibit wide variations in FMR geographically. At the math threshold which gives a global FMR of 0.001, an individual from South Asia who attempts to match against another South Asian individual of the same age and sex has an FMR of 0.08 or 80 times as high. Based on the data shown in Figure 3, this will be significantly higher for South Asians who are under 16 or over 64.

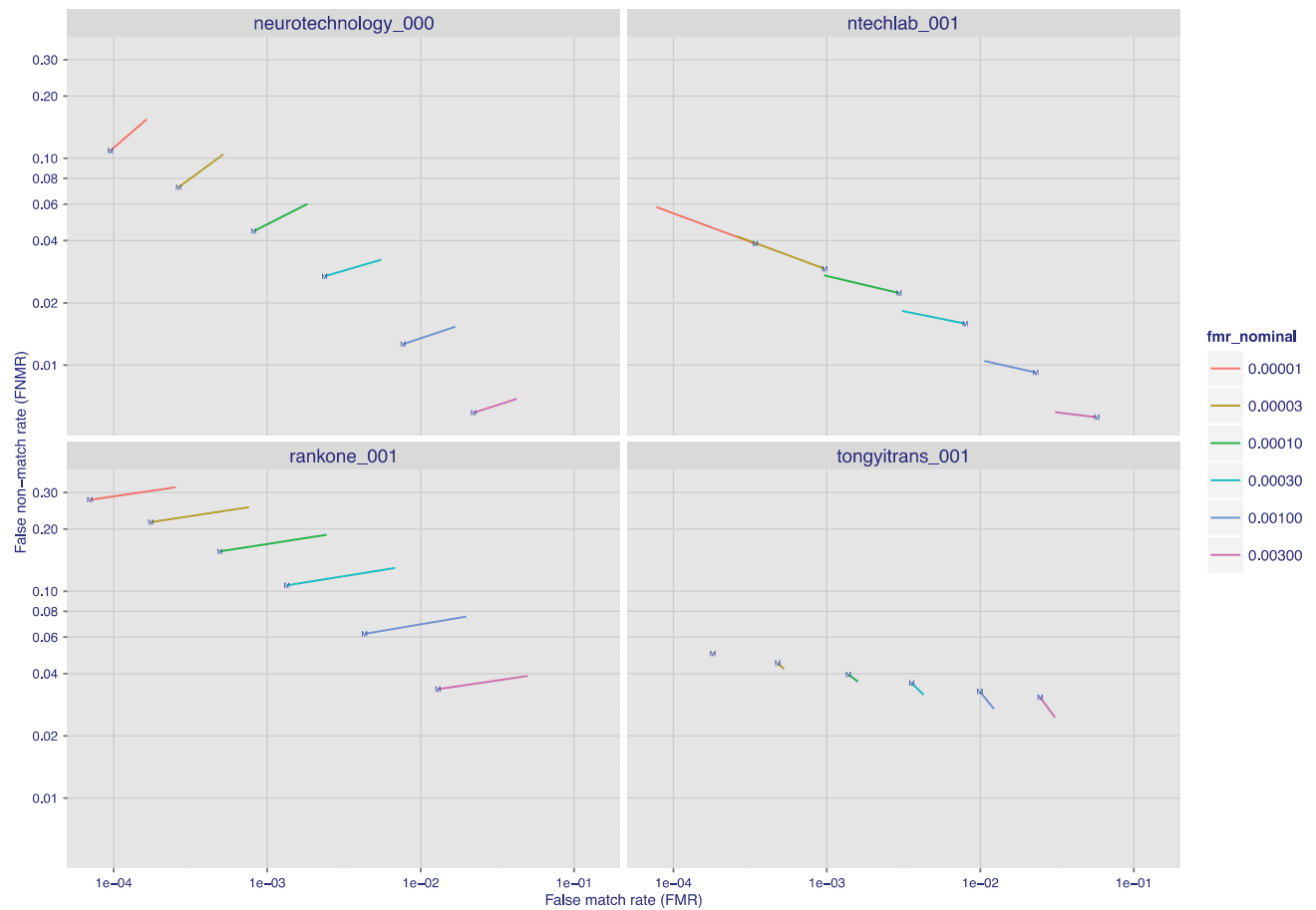


Figure 2 - FNMR and FMR at six different match thresholds. At each point a line is drawn between the operating points for males (denoted by M) and females (no marker). The six match thresholds are selected to give the nominal false match rates given in the legend, and are computed over all impostor pairs regardless of age, sex, 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-matched impostor pairs (i.e individuals of the same sex, from the same geographic region and age group)

Cross age FMR at threshold $T = 30.260$ for algorithm `neurotechnology_000`, giving $FMR(T) = 0.001$ globally.

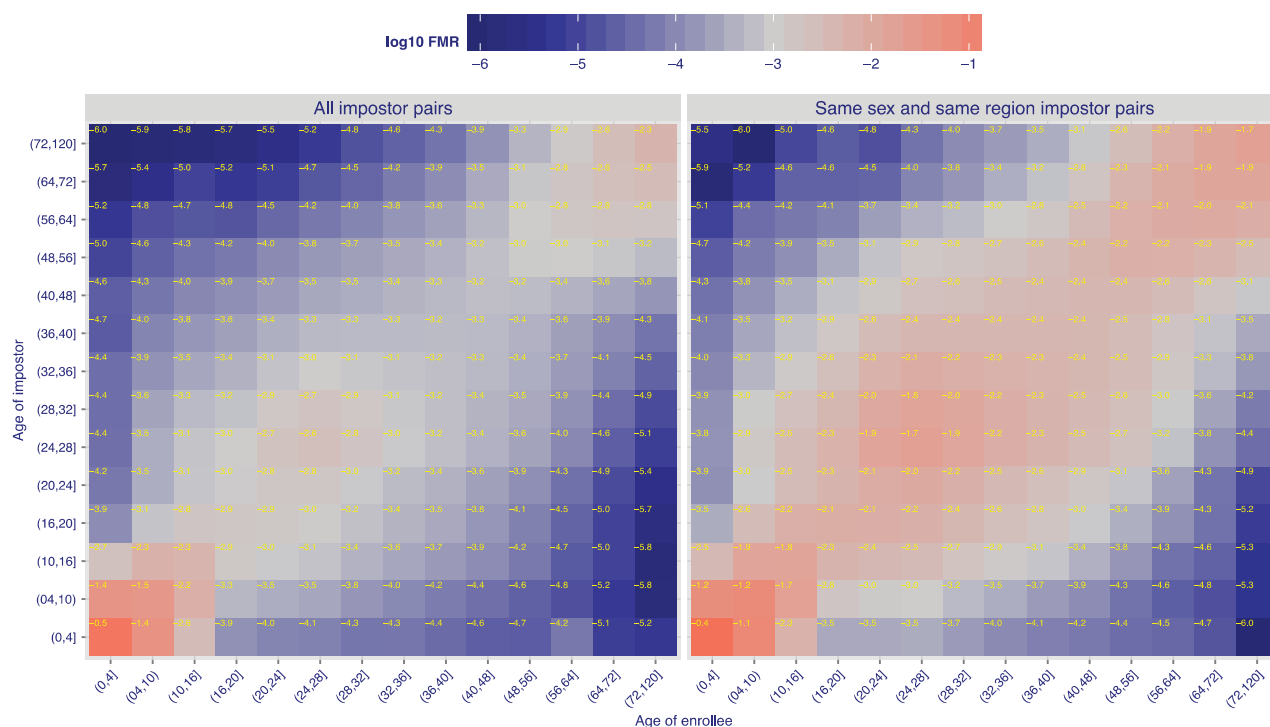


Figure 3 - For one algorithm operating on visa images, the heatmap shows false match observed over impostor comparisons of faces from different individuals who have the given age pair. False matches are counted against a recognition threshold fixed globally to give $FMR = 0.001$ over all impostor comparisons. The text in each box gives the same quantity as that coded by the color. Light colors represent a security vulnerability to, for example, an automated border control system using ePassports.

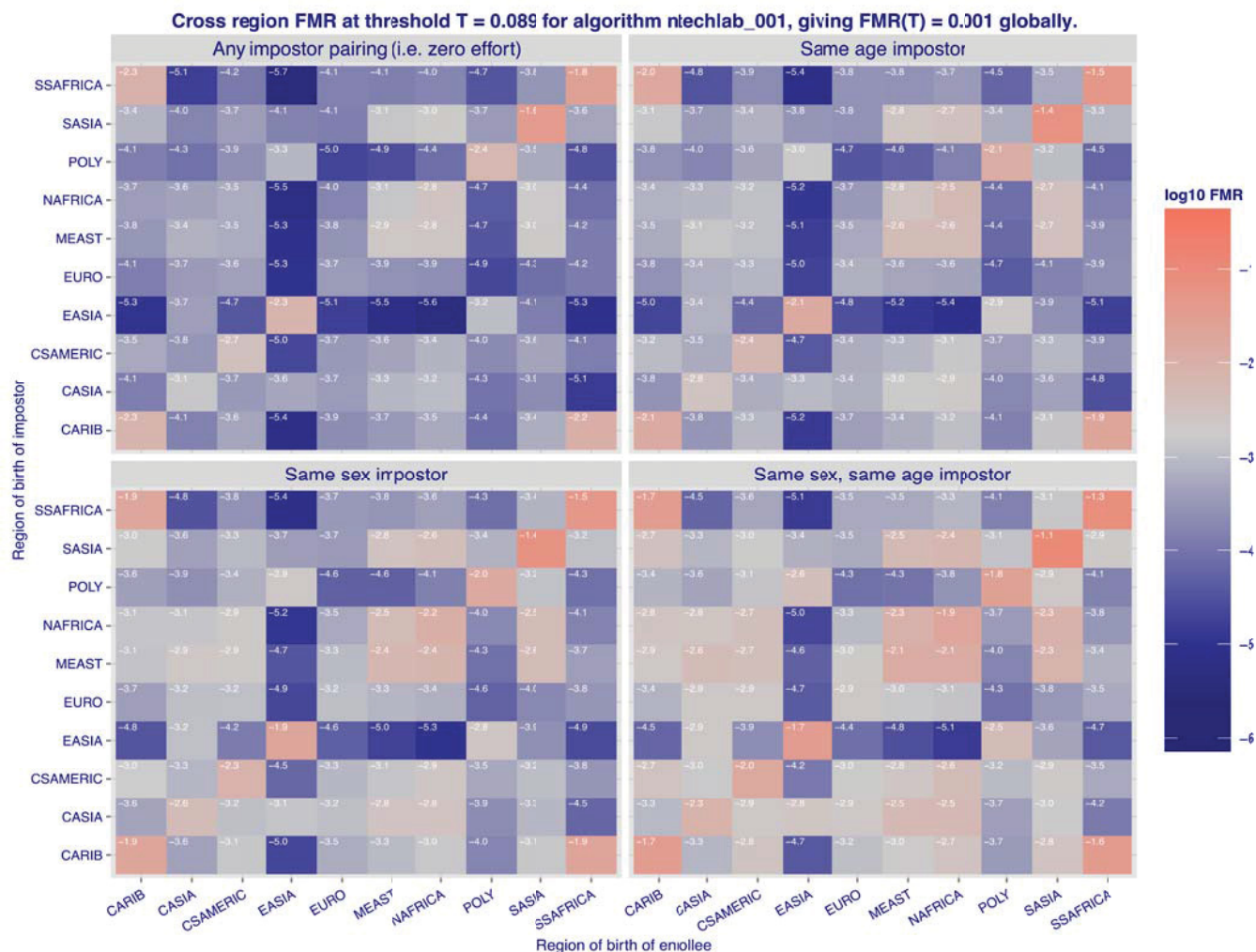


Figure 4 - For one algorithm operating on visa images, the heatmap shows false match rates observed over impostor comparisons of faces from different individuals who were born in the given region pair. False matches are counted against a recognition threshold fixed globally to give $FMR = 0.001$ over all impostor comparisons. If text appears in each box it gives the same quantity as that coded by the colour. Grey indicates FMR is at the intended 0.001 or 10^{-3} level. Light red colors represent a security vulnerability to, for example, an automated border control system using ePassports.

FINGERPRINT RECOGNITION

Unfortunately no large scale government studies on the impact of demographic factors on fingerprint recognition are currently available, even though the academic papers and the standards indicate that age, sex and ethnicity are all significant factors in both the quality scores and the matching performance of fingerprints.

IRIS RECOGNITION

In 2004, the Canada Border Services Agency (CBSA) began deployment of a system to authenticate pre-registered travelers in airports. This system, called, NEXUS-Air consisted of kiosks using iris recognition to authenticate the identity of travelers 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.

Enrolment and subsequent identification with the NEXUS-Air system work as follows. After being approved for participation in the program, each traveler attempts to enroll 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 traveler 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 match threshold, then the travelers' identity is confirmed as the identity associated with that iris template in the database.

CBSA developed the OPS-XING database from transaction logs of the NEXUS system for 702,626 distinct NEXUS participants who used the system to cross the border at airports, some of them more than 60 times, between 2007 and 2014. The enrolment data goes back to 2003. OPS-XING includes quality metrics, Hamming distance and other ancillary data, but the only demographic information provided is the age of the traveler. 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 traveler was able to use the system. This study showed some very interesting data related to the impact of age on iris recognition matching accuracy.

During enrolment, if only a single iris meets the quality threshold, then the traveler will be enrolled with only a single eye rather than two eyes. The number of travelers who enrolled in the system and the percentage of them who were only able to enroll a single eye are shown below in Figure 4. Travelers who could not enroll 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 65. This means that iris recognition may not be a recommended modality for populations with large percentages of young children or elderly people. In any event, any iris recognition system should provide extra help to aid the young and the elderly so that they can have a successful enrolment.

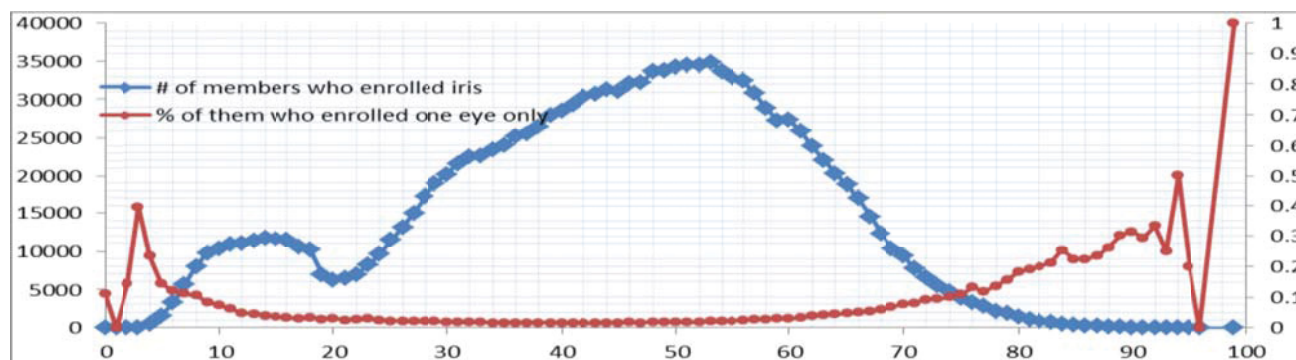


Figure 4 – Number of enrolees and percentage with one eye

When travelers who have successfully enrolled at least one iris use the kiosk, the Hamming distance between the iris template created at the kiosk and the matching iris template created during enrolment is calculated. Although the Hamming distance must be below the threshold for a match to occur and the data to be recorded, there is still variability in the average Hamming distance for travelers as a function of age. Since environmental factors within each airport also affect the Hamming distance, it is simplest to separate matching accuracy by airport for this calculation. Figure 5 shows a normalized version of the Hamming distance as a function of age for a single Canadian airport computed using a least-squares regression analysis. The gray zone represents the 95% confidence interval. It should be noted that the large gray zone for travelers over the age of 80 indicates that the least squares regression had insufficient data to be reliable and so this part of the curve should be ignored. The remainder of the curve shows that the average Hamming distance is lower for middle aged travelers than it is for young or elderly travelers. This agrees with the result from enrolment data that iris recognition is more difficult for those under 14 or over 60 years of age. It should be noted, however, that the middle-aged travelers also traveled the most and had the largest number of uses of the system, so part of this effect may be related to habituation.

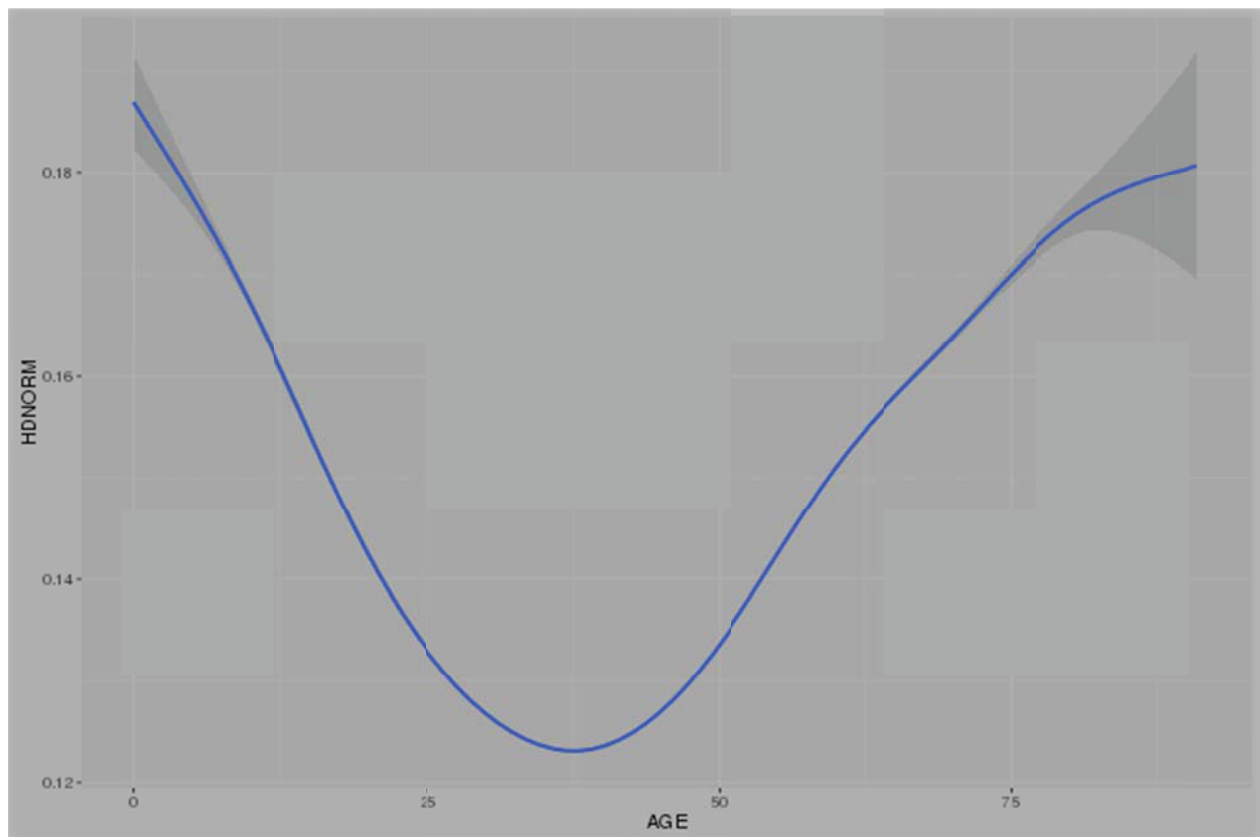


Figure 5 – Normalized Hamming Distance as a function of age

The final conclusion of this CBSA case study is that iris recognition is less reliable, but may still be usable, for younger and elderly individuals.

SUMMARY OF FINDINGS FROM GOVERNMENT SPONSORED STUDIES

There are no government studies about fingerprint recognition and only a single study about iris recognition which focuses only on the impact of age on iris recognition performance. There are however, two very important studies on facial recognition. At this stage, the most important conclusions relate to facial recognition and they highlight potential security issues with current passport and border control systems. A summary of the conclusions follows:

- Enrolment in an iris recognition system and (possibly) the successful use of that system is more difficult for those under 14 and those over 60 years of age and it becomes increasingly difficult as age increases above 60 or decreases below 14.
- Almost all facial recognition systems appear to exhibit worse performance for females than for males. The NIST study shows that this makes it more likely for females to be

unable to successfully use a facial recognition system using 1:1 matching, such as an ABC gate. The Passport Canada study shows that this mean it is more difficult to determine if a female already exists in a database and thus increases the likelihood of successful identity fraud by females.

- Facial recognition systems are much more likely to incorrectly declare two images to match if they are of the same age, sex and ethnicity (i.e. the False Match Rate (FMR) at a fixed match threshold will be much higher than expected).
- For young children under 4 this can cause FMR to be up to 400 times higher than would be estimated based on general performance tests
- For certain Asian ethnicities this can cause FMR to be up to 80 times higher that would be estimated based on general performance tests
- Systems that use facial recognition to determine if an individual matches a previously enrolled image (such as an access control system, a border control system or a 1:1 verification on passport renewal) will have a very high chance of allowing imposters to pass through the system unless they have conducted performance tests that investigate demographic factors including age, sex and ethnicity and then adjust the match threshold based on the results of such tests
- If a match threshold is adjusted so that an acceptable FMR is achieved for all demographic sub-groups, then FNMR is likely to become very high for certain demographic groups. Dynamic thresholds, where the match threshold is adjusted depending on the identity information of the subject using the system are likely to give better results.

APPLICATIONS TO AUTOMATED BORDER CONTROL INCLUDING CBSA PRIMARY INSPECTION KIOSKS (PIK)

Automated Border Control systems such as the CBSA Primary Inspection Kiosk use information which has already been captured during a passport or visa issuance process and are therefore unable to directly control the quality of the biometric enrolment. They need to focus on capturing good quality verification data using whatever modality is available from the enrolment. For the vast majority of travelers this will be a facial image to be matched against the enrolment image from an ePassport chip but for some travelers this may be a fingerprint to be matched against a fingerprint captured during visa enrolment. Operational efficiency is important so the FNMR should be as low as possible to facilitate travelers using the PIK, but security is critical in that no traveler should enter Canada using a different that is not their own.

The following recommendations are based on the research discussed in previous sections.

1. There should be an exemption from using fingerprints and an alternative mechanism provided for adults over 79 and children under 10. Facial recognition will have both high FNMR and FMR for young children, so children under 10 should also be exempted from facial recognition.
2. Data should be collected using the fingerprint quality and matching algorithm used in the PIKs at each airport to ensure that there is not a substantially higher fingerprint rejection rate for females than for males. If this is found to occur, then a variable quality threshold that is lower for females than for males should be implemented.
3. The lighting system used by the kiosks in different airports should be tested to ensure that it can obtain a facial image without saturation or underexposure for people of different skin tone ranging from very dark (Africans) to very light (Northern Europeans). This will help to reduce the FMNR for those with darker skin.
4. If the default facial match threshold suggested by the manufacturer is being used or if a threshold has been selected based on testing of all travelers, then it is likely that a significant security hole exists for travelers of certain ethnicities or from certain countries. The only way to ensure that the biometric system is secure for all travelers is to capture performance data from a large number of travelers at each airport and then analyze it to evaluate the effects of age, sex and ethnicity. Then an appropriate match threshold can be selected for each airport. Ideally, the match threshold would be dynamic and would change based on the demographic data contained in the passport.
5. If possible, the results from testing the PIK system testing, or a sanitized subset of those results, should be shared with the SC 37 committee as a Canadian contribution towards the developing standard on “Identifying and mitigating the differential impact of demographic factors in biometric systems”. This will allow the Canadian results to be combined with results from other countries so that the best way to mitigate any bias against particular demographic groups can be determined.

APPLICATIONS TO CANADIAN PASSPORT ISSUANCE SYSTEM

The Canadian passport issuance system uses facial recognition as the primary biometric. Facial images are scanned from a printed photograph submitted by the passport applicant. They are then processed to create an image that is stored in the passport database for 1:many matching to prevent duplicate identities and for 1:1 matching to verify the applicant when a passport is renewed. They are also compressed and stored in the Logical Data Structure of the contactless chip contained in the passport. In this system the quality of the enrolment image can be controlled by ensuring an accurate scanning and compression process that doesn't degrade the image quality and by rejecting poor quality submitted photos. The passport system controls the

facial recognition used internally but it has no control over the facial recognition performed at different borders as the Canadian passport holder attempts to use ABC systems around the world or even in Canada.

Since facial recognition is the primary biometric for the Canadian passport system, the following recommendations are applicable:

1. Facial recognition will have both high FNMR and FMR for young children, so children under 10 should be exempted from facial recognition.
2. It is likely that females will have higher error rates than males, both for 1:1 and 1:many facial recognition. This may represent an inconvenience for females when renewing passports (as they may undergo more scrutiny after their images fail to match previous passport images) and a security risk (as they are less likely to be found if they already exist in the database or if they are on a watchlist). The current facial recognition system should be tested to investigate the performance of females versus males and, if necessary, thresholds should be adjusted for females to try and mitigate this problem. Future systems should be required to demonstrate the performance on females versus males as part of future procurements.
3. Certain demographic groups are much more likely to match against other images from the same demographic group (higher FMR for individuals with similar demographics). This represents an issue for 1:1 matching on renewal where it may be possible to substitute a photo of a different person when renewing a passport so that this other person can successfully use the passport once it is issued. It also represents a potential inefficiency for the 1:many matching, since a high percentage of the candidate lists requiring human verification will likely come from specific demographic groups. The current system should be tested to evaluate this and a plan developed to adjust match thresholds and candidate list lengths based on the demographics of the individual.
4. Certain demographic groups are less likely to match against their own images (higher FNMR). This represents an inconvenience in 1:1 matching for passport renewal because they are more likely to fail to match their own previous images and to require additional scrutiny. The current facial recognition system should be evaluated to determine if this is a significant problem or not.
5. If possible, the results from the Canadian passport issuance system testing, or a sanitized subset of those results, should be shared with the SC 37 committee as a Canadian contribution towards the developing standard on “Identifying and mitigating the differential impact of demographic factors in biometric systems”. This will allow the Canadian results to be combined with results from other countries so that the best way to mitigate any bias against particular demographic groups can be determined.

APPLICATIONS TO CANADIAN IMMIGRATION BIOMETRIC IDENTIFICATION SYSTEM (CIBIDS)

The Canadian Immigration Biometric Identification System captures both fingerprints and facial images from applicants from certain countries seeking a visa to enter Canada. CIBIDS is primarily an enrolment system as the identification matching for deduplication and watch list checking is performed by RCMP using fingerprints and the verification against individual visa holders at the border is performed by CBSA using facial images or by CBSA and RCMP using fingerprints. The enrolments take place at multiple locations across the world so issues such as humidity and lighting will vary significantly from one enrolment site to the next. The system is used with individuals with varying age, sex, gender and ethnicity so this is an excellent example of a system that needs to manage the impact of demographic factors on biometric performance.

For a system involving enrolment of faces and fingers, the following recommendations are applicable:

1. Adults over 79 and children under 10 may have difficulty providing fingerprints of sufficient quality to be useful and if they can't be exempted from providing fingerprints then the quality threshold should be adjusted for these groups.
2. Although children under 10 are subject to high FNMR and FMR when using facial recognition, their images are still useful if only a limited time has elapsed between the enrolment and verification. For visa issuance, where the first entry to Canada will often be within a few months of the photo being captured, facial recognition may be usable. Fingerprint recognition should still be the preferred modality for children under 10, provided that fingerprints of sufficient quality can be captured.
3. In order to ensure that there is no bias against females, CIC should work with RCMP to ensure that there is not a substantially higher fingerprint rejection rate for females than for males. If this is found to occur, then a variable quality threshold that is lower for females than for males should be implemented.
4. Fingerprint capture devices should be ergonomically designed to accommodate both left and right handed individuals and those with limited motor skills, arthritis or arthrosis, which are more common in the elderly. This means that there should be an attendant to help position the hand and fingers of those who can't easily do it themselves.

5. For facial image capture, the lighting should be tested to ensure that it can obtain a facial image without saturation or underexposure for people of different skin tone ranging from very dark (Africans) to very light (Northern Europeans). Different illumination levels may be required for different enrolment sites.

FINAL SUMMARY OF RESEARCH

Research from academic papers, international standards and multiple government studies has shown that both fingerprint recognition and facial recognition exhibit performance differences as a function of age, sex and ethnicity. Iris recognition exhibits some performance differences with age, but other demographic factors have not been tested.

For iris recognition, more studies to examine the impact of sex and ethnicity would be helpful. The only recommendation from the current research is that that extra effort should be expended to achieve a successful iris capture for children and for the elderly.

For fingerprint recognition, the performance differences primarily make it more difficult for certain groups to properly enroll fingerprints or to have them successfully verified after enrolment. Therefore, fingerprint recommendation is not recommended for adults over 79 and for children under 10, unless no other options are available. The impact on large scale 1:many fingerprint searches, such as those used for deduplication in identity management systems or for forensic searches used in criminal investigations have not been properly investigated. More studies are definitely required in this area.

For facial recognition, females tend to have slightly higher error rates (both FMR and FNMR) than males and this makes it more difficult and less secure to use facial recognition on females. Facial recognition is also extremely unreliable (for both FMR and FNMR) for young children and is definitely not recommended for children under 10. Most importantly, however, the FMR increases substantially for a facial recognition system if the demographics of the two images being compared are similar. This represents a huge security risk for system using facial recognition if match thresholds are set based on manufacturer recommendations or on general performance testing using the entire population of users. Any facial recognition system needs to collect test results from its operational population and then analyze them as a function of demographic factors so that match thresholds can be set appropriately. If manufacturers can be convinced to train algorithms on appropriately demographically balanced data sets, then this situation may improve over time, but testing based on demographic factors should remain the default behavior until this has been clearly demonstrated to no longer be necessary.

More studies on fingerprint recognition and iris recognition and on 1:many facial recognition will help improve our understanding of these issues and Canada could benefit from the recent international interest in this subject by sharing its own studies and reviewing those of other nations through the appropriate forums at ISO, ICAO and through the Five Eyes alliance between Australia, Canada, New Zealand, UK and the US. All of the other Five Eyes countries are no actively looking at this problem and at least one, New Zealand, is currently updating its operational systems to minimize the security risks associated with demographic bias in operational performance of biometric systems.

BIBLIOGRAPHY

- [1] ISO/IEC 2382-37:2017 "Information technology -- Vocabulary -- Part 37: Biometrics"
- [2] "Factors that influence algorithm performance in the Face Recognition Grand Challenge", J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips and Bruce A. Draper, *Computer Vision and Image Understanding* 113, 2009
- [3] "Face Recognition Performance: Role of Demographic Information", Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge and Anil K. Jain, *IEEE Transactions on Information Forensics and Security*, Volume: 7, Issue: 6, 2012
- [4] "Report on the FG 2015 Video Person Recognition Evaluation", P J. Phillips, J. R. Beveridge, Hao Zhang, Bruce A. Draper, Patrick J. Flynn, Zhenhua Feng, Patrik Huber, Josef Kittler, Zhiwu Huang, Shaoxin Li, Yan Li, Meina Kan, Ruiping Wang, Shiguang Shan, Xilin Chen, Haoxiang Li, Hua Gang, Vitomir Struc, Janez Krizaj, Changxing Ding and Dacheng Tao, *Proceedings of the Eleventh IEEE Conference on Automatic Face and Gesture Recognition*, 2015
- [5] "Demographic effects on estimates of automatic face recognition performance", Alice J. O'Toole, P. Jonathon Phillips, Xiaobo An and Joseph Dunlop, *Image and Vision Computing* 30, 2012
- [6] "Implications of the IDENT/IAFIS Image Quality Study for Visa Fingerprint Processing", R. Austin Hicklin and Christopher L. Reedy, Technical Report, Mitretek Systems, October 31, 2002
- [7] "Impact of Gender on Fingerprint Recognition Systems", Michael D. Frick, Shimon K. Modi, Stephen J. Elliott and Eric P. Kukula, *Proceedings of the Fifth International Conference on Information Technology and Applications*, 2008
- [8] "The Impact of Gender on Image Quality, Henry Classification and Performance on a Fingerprint Recognition System", Kevin O'Connor and Stephen J. Elliott, *Proceedings of the Seventh International Conference on Information Technology and Applications*, 2011
- [9] "Predicting Ethnicity and Gender from Iris Texture", Stephen Lagree and Kevin W. Bowyer, *Proceedings of the 2011 IEEE International Conference on Technologies for Homeland Security (HST)*, 2011
- [10] "Gender Classification from Iris Images Using Fusion of Uniform Local Binary Patterns", Juan E. Tapia¹, Claudio A. Perez, and Kevin W. Bowyer, *Lecture Notes in Computer Science*, Volume 8926, Springer, 2015

[11] “Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification”, Patrick Grother, Mei Ngan, and Kayee Hanaoka, *NIST Interagency Report*, 2017 (See <https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing>)

DOCUMENT CONTROL DATA		
*Security markings for the title, authors, abstract and keywords must be entered when the document is sensitive		
1. ORIGINATOR (Name and address of the organization preparing the document. A DRDC Centre sponsoring a contractor's report, or tasking agency, is entered in Section 8.) Bion Biometrics Inc. 236B Claridge Dr Nepean ON K2J 5H1		2a. SECURITY MARKING (Overall security marking of the document including special supplemental markings if applicable.) CAN UNCLASSIFIED
		2b. CONTROLLED GOODS NON-CONTROLLED GOODS DMC A
3. TITLE (The document title and sub-title as indicated on the title page.) Impact of Demographic Factors on Performance of Biometric Systems		
4. AUTHORS (Last name, followed by initials – ranks, titles, etc., not to be used) Campbell, J. W. M.		
5. DATE OF PUBLICATION (Month and year of publication of document.) March 2018	6a. NO. OF PAGES (Total pages, including Annexes, excluding DCD, covering and verso pages.) 38	6b. NO. OF REFS (Total references cited.) 11
7. DOCUMENT CATEGORY (e.g., Scientific Report, Contract Report, Scientific Letter.) Contract Report		
8. SPONSORING CENTRE (The name and address of the department project office or laboratory sponsoring the research and development.) DRDC - Centre for Operational Research and Analysis Defence Research and Development Canada Carling Campus 60 Moodie Drive, building 7 Kanata ON K2H 8E9 Canada		
9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)	9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)	
10a. DRDC PUBLICATION NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.) DRDC-RDDC-2018-C113	10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)	
11a. FUTURE DISTRIBUTION WITHIN CANADA (Approval for further dissemination of the document. Security classification must also be considered.) Public release		
11b. FUTURE DISTRIBUTION OUTSIDE CANADA (Approval for further dissemination of the document. Security classification must also be considered.)		

12. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Use semi-colon as a delimiter.)

Biometrics; Demographic Bias

13. ABSTRACT/RÉSUMÉ (When available in the document, the French version of the abstract must be included here.)

Biometric systems using facial, fingerprint and iris recognition have been widely deployed by Canada and other governments. These systems are used to issue passports and visas, permit travelers to cross the border and to allow employees access to physical and digital resources. Biometric performance, in terms of the risk of non-matches and of false matches, is critical to the successful and secure operation of these systems.

Recent research has uncovered significant performance differences in multiple biometric systems depending on the demographics (specifically age, sex and ethnicity) of the subjects whose biometric characteristics are being captured. In some cases, this simply leads to a bias that makes it more difficult for certain subjects to use the system. Existing studies show that this is the case for females with fingerprint recognition and facial recognition for older people with fingerprint recognition and iris recognition and for young children with all three modalities. In other cases, the performance difference leads to high rates of false matches and this can significantly undermine the security of the biometric system. So far this is known to be the case for certain ethnic groups and for young children when using facial recognition, but further studies are needed to investigate this phenomenon more thoroughly for all biometric modalities. Several governments are actively researching this problem and Canada need to do the same. Unless systems, especially those using facial recognition, are tested properly to investigate the impact of demographic factors on system performance then they may be providing a false sense of security. In the worst case, one US study has shown that young children may be able to pass through a facial recognition based ABC system using a different child's travel document with a 40% success rate, which raises a huge problem for child trafficking and brings the entire security of the border into question. Proper testing and modifications to the ABC system can mitigate this problem, but as this research is very new, the only government agency currently known to be implementing such measures is New Zealand Customs Service.

Les systèmes biométriques utilisant la reconnaissance faciale, les empreintes digitales et l'iris ont été largement déployés par le Canada et d'autres gouvernements. Ces systèmes sont utilisés pour délivrer des passeports et des visas, permettre aux voyageurs de traverser les frontières et permettre aux employés d'accéder à des ressources physiques et numériques. Le rendement biométrique, en termes de risque de non-appariement et de fausses correspondances, est essentiel à la sûreté de ces systèmes et à leur bon fonctionnement. Des recherches récentes ont révélé des différences significatives de rendement entre plusieurs systèmes biométriques en fonction de la démographie (en particulier l'âge, le sexe et l'origine ethnique) des sujets dont les caractéristiques biométriques sont captées. Dans certains cas, cela conduit simplement à un biais qui rend plus difficile l'utilisation du système par certains sujets. Les études existantes montrent que c'est le cas avec la reconnaissance des empreintes digitales pour les femmes et la reconnaissance faciale, des empreintes digitales et de l'iris pour les personnes âgées et avec les trois modalités pour les jeunes enfants. Dans d'autres cas, la différence de rendement entraîne des taux élevés de fausses correspondances, ce qui peut compromettre considérablement la sécurité du système biométrique. Jusqu'à présent, cela est reconnu pour certains groupes ethniques et pour les jeunes enfants lors de l'utilisation de la reconnaissance faciale, mais d'autres études sont nécessaires pour étudier ce phénomène de manière plus approfondie pour toutes les modalités biométriques.

Plusieurs gouvernements étudient activement ce problème et le Canada doit faire de même. À moins que les systèmes, en particulier ceux qui utilisent la reconnaissance faciale, soient testés correctement pour étudier l'impact des facteurs démographiques sur le rendement du système, ils peuvent fournir un faux sentiment de sécurité. Dans le pire des cas, une étude américaine a montré que les jeunes enfants peuvent passer par un système ABC de reconnaissance faciale en utilisant un document de voyage d'un autre enfant avec un taux de réussite de 40 %, ce qui soulève un énorme problème pour la traite des enfants et remet en question tout l'aspect sécuritaire des frontières. Des essais appropriés et des modifications au système ABC peuvent atténuer ce problème, mais comme cette recherche est très récente, la seule agence gouvernementale dont on sait qu'elle met en œuvre de telles mesures est le Service des douanes de Nouvelle-Zélande.