NISTIR 8252

IREX IX Part Two
Multispectral Iris Recognition

George W. Quinn
Patrick Grother
James Matey

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8252

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

NISTIR 8252

IREX IX Part Two
Multispectral Iris Recognition

George W. Quinn
Patrick Grother
James Matey
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8252

June 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

National Institute of Standards and Technology Interagency or Internal Report 8252
Natl. Inst. Stand. Technol. Interag. Intern. Rep. 8252, 42 pages (June 2019)

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8252

Executive Summary

Introduction

Iris Exchange (IREX) IX is an evaluation of automated iris recognition algorithms. The first part of the evaluation was a performance test of both verification (one-to-one) and identification (one-to-many) recognition algorithms over operational test data. Those results are summarized in NIST IR 8207 [1]. The second part of the evaluation constitutes a multispectral evaluation of iris recognition. Those results are summarized in this report.

All currently deployed iris recognition systems operate on images of the iris illuminated in the near infrared band of the electromagnetic spectrum. The ISO/IEC 19794-6 [2] and 29794-6 [3] standards require the eye to be illuminated between "approximately 700 and 900 nanometers (nm)". Near infrared light is specified because melanin, the pigment that makes dark eyes dark [1], is nearly transparent in the near infrared. This evaluation assesses the accuracy of state-of-the-art iris matchers over a much wider range of the spectrum. Particular attention is placed on the visible (400 nm to 700 nm) and short wave infrared (700 nm to 1550 nm) bands.

Principle testing was performed over the Consolidated Multispectral Iris Dataset (CMID) which was provided to NIST by the Southern Methodist University. Collected in well controlled laboratory settings, the dataset is ideal for multispectral analysis. It contains about a quarter-million iris samples from over 200 subjects, by far the largest of its kind as of this writing. Thirteen research institutions submitted matching algorithms for testing.

Key Results

- **Near-Infrared (NIR) Matching:** Matching accuracy is very dependent on illumination wavelength, even within the standard 700 nm to 900 nm band. For some iris matchers, error rates vary by more than an order of magnitude depending on whether matching is performed at 700 nm or 910 nm. Nearly every matcher performs better at 910 nm than at 700 nm, 800 nm, or 970 nm. Matching accuracy was only measured at discrete wavelengths so an "optimal wavelength" was difficult to identify but is probably around 850 nm.
- **Visible Wavelength (VW) Matching:** The VW band spans from about 400 nm to 700 nm. Matching accuracy tends to be much better at the longer end of the VW band. At the shortest end (405 nm) matching is not viable. One matcher performs significantly better than the others at visible wavelengths. At 620 nm, this matcher is capable of correctly matching irises 95.5 % of the time while falsely matching only once every ten thousand attempts when using both eyes for matching.
- **Effect of Eye Color:** Lighter irises (*i.e.* blue, grey, green) match better than darker irises (*i.e.* brown, black) at visible wavelengths. However, at standard NIR wavelengths darker irises tend to match better than lighter irises. The reason for the former result is obvious: the melanin pigments in darker irises are obscuring the iris texture. As for the latter result, it is unclear if this is due to more rigorous algorithm tuning over darker irises (since brown is by far the most common eye color), or if there is something intrinsic in the features of dark irises that makes them easier to recognize. Despite lighter irises matching better than darker irises at visible wavelengths, overall accuracy for both eye hues is still better in the NIR.
- **Cross-Wavelength Matching:** Matching tends to be more accurate when both compared iris samples were acquired at the same (or similar) wavelengths. Accuracy is best for most matchers when both samples were acquired at 910 nm. One matcher that performs well on VW iris samples can compare VW samples to each other about as well as it can compare VW samples to NIR samples. False matches are more common when both compared samples were acquired at visible wavelengths.
- **Impact of Wavelength on the False Match Rate:** Generally, matching accuracy is assessed by quantifying two properties: 1) the ability to recognize that two iris samples represent the same eye, and 2) the ability of the matcher to distinguish when two iris samples represent different eyes. All previous research on multispectral iris recognition have focused on the first property. This is the first study to address the second property. As Daugman has often noted, a strength of iris recognition is its ability to distinguish samples that come from different sources. This is evidenced by the extremely low false match rates (FMRs) iris matchers are able to achieve. Although this is true for conventional (near-IR) matching, FMR tends to be much less predictable when comparing VW samples. FMR can vary by orders of magnitude depending on the wavelength at which the samples were acquired. Moreover, the variation in FMR is

[1] 70 to 90 % of the world's population have dark brown eyes.

inconsistent across matchers. Thus, calibrating a deployed iris recognition system to achieve a desired FMR could be difficult if the system operates over samples acquired at non-standard wavelengths.

- **Comparison to Earlier Research:** The multispectral results in the current investigation do not perfectly align with existing literature. This study found that accuracy is highest somewhere between 800 and 910 nm for most matchers. Past studies have found that the mean SQRT-normed Hamming Distance[2] is minimized at shorter wavelengths (800 nm [5] and 590 nm [6]). Additionally, the current study found that accuracy is highly sensitive to the wavelength at which the samples were acquired while the previous studies found that accuracy changes little across samples acquired between 590 nm and 970 nm. A possible explanation for the discrepancies is that previous studies used the mean SQRT-normalized Hamming Distance while the current study uses the false non-match rate (FNMR) at a fixed decision threshold or false match rate (FMR) to assess accuracy. The latter places greater emphasis on the behavior of the crucial right-tail of the mated distribution and is therefore the recommended statistic for assessing accuracy. The current study also uses a much larger dataset consisting of images collected in highly controlled environments.

- **Low-resolution Iris Matching:** There are forensic applications of iris recognition that are likely to require matching iris samples acquired at resolutions below the ISO/IEC 19794-6:2011 recommended minimum spatial sampling rate of 10 pixels / mm. This study found that the ability of matchers to recognize that two samples represent the same iris remains relatively stable until the radius of the iris is reduced to about 20 pixels (corresponding to $\approx$4 pixels / mm). However, the ability of matchers to distinguish that two samples represent *different* irises deteriorates much earlier, at radii of 64 pixels ($\approx$12.8 pixels / mm). This suggests that accuracy could be improved by acquiring iris samples at spatial sampling rates above the current ISO/IEC 19794-6:2011 recommendation. The ability of most matchers to distinguish between irises is poorest when the radius is between 10 and 16 pixels. No single matcher yields the best accuracy across all resolutions. NeuroTechnology's submission achieves the best accuracy at higher resolutions while Tafirt's and IrisID's matchers achieve the best accuracy at lower resolutions. The latter two matchers are capable of correctly matching the iris more than half the time when the radius of the iris is only 8 pixels (approximately 0.8 pixels/mm). When low-resolution iris samples are compared, it appears to be the lower resolution of the two that dictates matching accuracy.

- **Temporal Case Study:** A short experiment was conducted to demonstrate that two images of a person's iris can be successfully y matched despite being captured 48 years apart. Both compared iris images were acquired "in the wild" with conventional (non-iris) cameras and video equipment. The first sample is a fairly high resolution (red channel) image of an actor's eye from the film *2001:A Space Odyssey*. The second is from a photograph of the same actor captured at a celebrity event in 2015. The radius of this iris is 43 pixels, although tiling artifacts from JPEG compression were evident. All but one of the 13 matchers produce low measures of dissimilarity. Nine of the 13 matchers produce scores corresponding to an FMR less than 10^{-5}. This indicates an extremely high degree of confidence that the two samples represent the same iris. A further iris2pi matcher (not submitted to IREX IX) produced a Hamming Distance of 0.179. The probability of two samples of different eyes producing such a low Hamming Distance is less than one in one hundred billion.

- **Multiwavelength Fusion:** Fusion at the score level involves combining scores acquired at different illumination wavelengths into a single fused score. This fused score is then used to make a final match / nonmatch decision. Accuracy improved significantly when scores acquired at 910 nm were combined with scores acquired at 700 nm or 800 nm. Score fusion at visible wavelengths led to only minor improvements in accuracy (and even then, primarily only for lighter-eyed subjects).

[2] Daugman[4] recommended a correction factor to take into account the change in the width of the non-mated distribution as the fraction of the iris useful for iris recognition varies due to e.g. specularities and occlusion. This is an important correction for operational systems. It can confuse results in laboratory experiments. Some implementations of iris2pi have an option to turn this off; some do not. For algorithms used in IREX-IX, the use of such normalization is an unknown.

Acknowledgements

The authors would like to thank the sponsor of this activity, the Federal Bureau of Investigation. The authors would also like to thank David Ackerman of Princeton Identity for his critique and insight on early drafts of this report.

Disclaimer

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

Caution is advised when attempting to extrapolate numerical results from this evaluation to arbitrary applications. This evaluation measures performance over a particular test dataset collected under specific environmental conditions with specific hardware. It is difficult to predict how changing any of these parameters might affect performance.

Institutional Review Board

The National Institute of Standards and Technology Human Subjects Protection Office (HSPO) reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

Contents

Executive Summary 1

1 Introduction 6
- 1.1 Purpose . . . 6
- 1.2 The IREX Program . . . 6

2 Evaluation Procedures 8
- 2.1 Test Environment . . . 8
- 2.2 Test Dataset . . . 8
- 2.3 Matching Software . . . 10
- 2.4 Performance Metrics . . . 10
  - 2.4.1 Accuracy . . . 10
  - 2.4.2 Uncertainty Estimation . . . 11

3 Results 12
- 3.1 Intra-spectral Matching . . . 12
- 3.2 Cross-spectral Matching . . . 15
- 3.3 Effect of Eye Color . . . 17
- 3.4 Multispectral Fusion . . . 18
  - 3.4.1 Score Level . . . 18
  - 3.4.2 Sensor Level . . . 24
- 3.5 Forensic Iris . . . 27
  - 3.5.1 Visible Wavelength Matching . . . 27
  - 3.5.2 Matching Low Resolution Iris Samples . . . 28

4 References 34

Appendix A Uncertainty Estimation 37

List of Figures

1.1 The IREX program . . . 6

2.1 A blue iris acquired at different illumination wavelengths from the CMID. . . . 9
2.2 Depiction of Manually Identified Iris Boundary Points . . . 10
3.1 FNMR vs. Wavelength . . . 12

3.2 FNMR vs. Wavelength . . . 13
3.3 FMR vs. Wavelength . . . 14
3.4 FMR vs. Decision Threshold . . . 14
3.5 FNMR Cross-Wavelength Heatmap . . . 15
3.6 FMR Cross-Wavelength Heatmap . . . 16
3.7 Effect of Eye color on FNMR . . . 17
3.8 Diagram of Score Level Fusion . . . 18
3.9 Neyman Pearson Boundaries . . . 19
3.10 Score-level Multiwavelength Fusion in the NIR band . . . 20
3.11 DET Plot of Score-level Fusion in the NIR Band . . . 21
3.12 Score-level Multiwavelength Fusion in the VW band . . . 22
3.13 DET Plot of Score-level Fusion in the VW Band . . . 23
3.14 DET Plot of Score-level Fusion Partitioned by Eye Color . . . 23
3.15 Composite Image Example . . . 24
3.16 Sensor-level Fusion in the NIR band . . . 25
3.17 Sensor-level Fusion in the NIR band . . . 26
3.18 VW vs. NIR DET Comparison . . . 27
3.19 Down sampled iris image. . . . 28
3.20 Effect of resolution on FNMR. . . . 29
3.21 Effect of resolution on FMR. . . . 30
3.22 Heatmap for down sampled mated image pairs. . . . 31
3.23 Heatmap for downsampled non-mated image pairs. . . . 32
3.24 Heatmap, down sampled images, FNMR at fixed FMR. . . . 33

List of Tables

2.1 IREX IX Participation List . . . 10

A.1 Correlation Structure for One-to-one Comparisons . . . 37

1 Introduction

1.1 Purpose

This report constitutes a multispectral evaluation of iris recognition. The accuracy of automated iris matchers over a wide band of the electromagnetic spectrum was assessed. Particular focus was placed on the visible (400 nm to 700 nm) and near infrared (700 nm to 1550 nm) wavelength bands. Iris Exchange (IREX) IX Part 2 utilized a test dataset of over 220 000 images and several state-of-the-art iris matchers, making it the largest evaluation of multispectral iris to date. For this reason, more detailed analyses could be performed compared to previous studies (*e.g.* determining whether certain behaviors hold across different eye colors). Additionally, conclusions could be drawn to a much higher degree of confidence than in prior studies.

Testing was performed over the Consolidated Multispectral Iris Dataset (CMID) which was provided to the National Institute of Standards and Technology (NIST) by the Southern Methodist University. Collection occurred in highly controlled laboratory environments using volunteers. Although iris recognition is particularly well suited for large-scale *identification mode* (a.k.a. one-to-many mode) deployments, for academic evaluation *verification mode* (a.k.a. one-to-one mode) is preferred because it can more precisely test core algorithm capability. Thirteen research institutions submitted matching algorithms for testing. Only results for their *verification mode* submissions are reported.

The main focus areas of this evaluation are:

- **Optimal Wavelength for Matching:** Not all iris cameras illuminate the iris at the same illumination wavelength. The ISO/IEC 19794-6 standard requires the iris to be illuminated between "approximately 700 and 900 nanometers (nm)". This study aims to assess the degree to which accuracy is dependent on wavelength and to determine if one specific wavelength consistently yields better accuracy than the others.
- **Cross-wavelength Matching Accuracy:** Cross-wavelength matching refers to comparing iris samples acquired at different wavelengths. The fact that different iris cameras utilize different emission spectra poses an interoperability problem. This study aims to quantify the accuracy penalty that results from comparing iris samples acquired at different wavelengths. This includes comparing iris samples acquired at visible wavelengths to ones acquired at near infrared wavelengths.
- **Use as a Forensic Tool:** Forensic iris is a burgeoning field. Forensic science is the application of science to criminal and civil law. This study explores the potential for using iris as an investigational tool for law enforcement. This is expected to involve iris samples acquired "in the wild" from photographs, videos, or other sources not originally intended for use in iris recognition. The limits of iris recognition over such images is tested.
- **Multiwavelength Fusion:** The appearance of the iris differs across spectral bands, introducing the possibility of improving accuracy by combining the information garnered at each band. This study attempts to improve matching accuracy by applying score and sensor level fusion techniques to combine this information.

1.2 The IREX Program

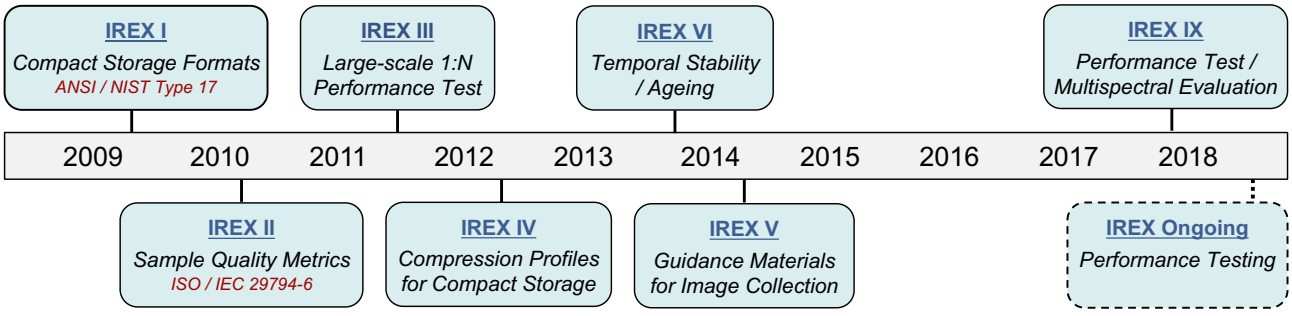


Figure 1.1: Timeline of the IREX program, including a *possible* future installment.

The IREX Program was initiated by NIST to support an expanded marketplace of iris-based applications. IREX provides quantitative support for iris recognition standardization, development, and deployment. To date, seven activities have been completed and one is tentatively planned. Each is summarized below.

- **IREX I** [7] was a large-scale, independently administered, evaluation of one-to-many iris recognition. It was conducted in cooperation with the iris recognition industry to develop and test standard formats for storing iris images. Standard formats are important for maintaining interoperability and preventing vendor lock-in. The evaluation was conducted in support of the ISO/IEC 19794-6 [8] and ANSI/NIST-ITL 1-2011 [9] standards.
- **IREX II** [10] supported industry by establishing a standard set of quality metrics for iris samples. Although iris recognition has the potential to be extremely accurate, it is highly dependent on the quality of the samples. The evaluation tested the efficacy of 14 automated quality assessment algorithms in support of the ISO/IEC 29794-6 standard [11].
- **IREX III** [12, 13] was a performance test of the latest iris recognition algorithms over operational data. Despite growing interest in iris-based technology, at the time there was a paucity of experimental data to support published theoretical considerations and accuracy claims. IREX III constituted the first public presentation of large-scale performance results using operational data.
- **IREX IV** [14, 15] built upon IREX III as a performance test of one-to-many iris recognition. In addition to providing participants from previous evaluations an opportunity to further develop and test their recognition algorithms, this evaluation explored the potential for using a cost equation model for optimizing algorithms for specific applications.
- **IREX V** [16] is an ongoing effort to provide best practice recommendations and guidelines for the proper collection and handling of iris images.
- **IREX VI** [17, 18] explored a possible aging effect for iris recognition. The intrinsic features of the iris may naturally change over time in a way that affects recognition accuracy. IREX VI found no evidence of a significant and widespread ageing effect up to nine years (data for intervals larger than nine years was unavailable). Later studies supported this conclusion [19, 20].
- **IREX VII** defines a framework for communication and interaction between components in an iris recognition system. By introducing layers of abstraction that isolate underlying vendor-specific implementation details, a system can become more flexible, extensible, and modifiable. NIST is currently using the framework internally, but no specifications or software have been publicly released as of this writing.
- **IREX VIII** is a placeholder for an as yet unimplemented conformance test of standard iris samples. The activity would constitute a laboratory evaluation of iris recognition algorithms capable of producing and consuming conformant iris samples according to ISO/IEC 19794-6:2011. There are currently no plans to move forward with this activity.
- **IREX IX** [1] is a performance test of the current state of the art over operational test data as well as an evaluation of multispectral iris recognition.
- **IREX Ongoing** is a possible successor to IREX IX. If conducted, it would be an ongoing, largely automated, evaluation of iris recognition algorithms similar to *MINEX III* [21] and *FRVT Ongoing* [22].

2 Evaluation Procedures

IREX IX is a *technology evaluation* in the sense that it is "an evaluation of multiple products providing the same capability" [23]. IREX IX's focus is on algorithm performance over other factors that might be relevant to the deployment and operation of a biometric system (e.g. societal and economic factors, policy drivers, legacy data). Performance is assessed using metrics that provide a general idea of the technology's capabilities. The relative importance of these metrics will depend on how the technology is applied.

Caution is advised when attempting to extrapolate numerical results from this evaluation to arbitrary applications. This evaluation measures performance over a particular test dataset collected under specific environmental conditions with specific hardware. It is difficult to predict how changing any of these parameters might affect performance.

2.1 Test Environment

The evaluation was conducted off-line at a NIST facility. Offline evaluations are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. However, they do not capture all aspects of an operational system. While this evaluation is designed to mimic operational reality as much as possible, it does not include a live image acquisition component or any interaction with real users.

Testing was performed on high-end PC-class blades running the Linux operating system (CentOS 7.2), which is typical of central server applications. Most of the blades had Dual Intel Xeon E5-2695 v3 3.3 GHz CPUs (56 total cores) with 192 GB of main memory. The test harness used concurrent processing to distribute workload across multiple blades.

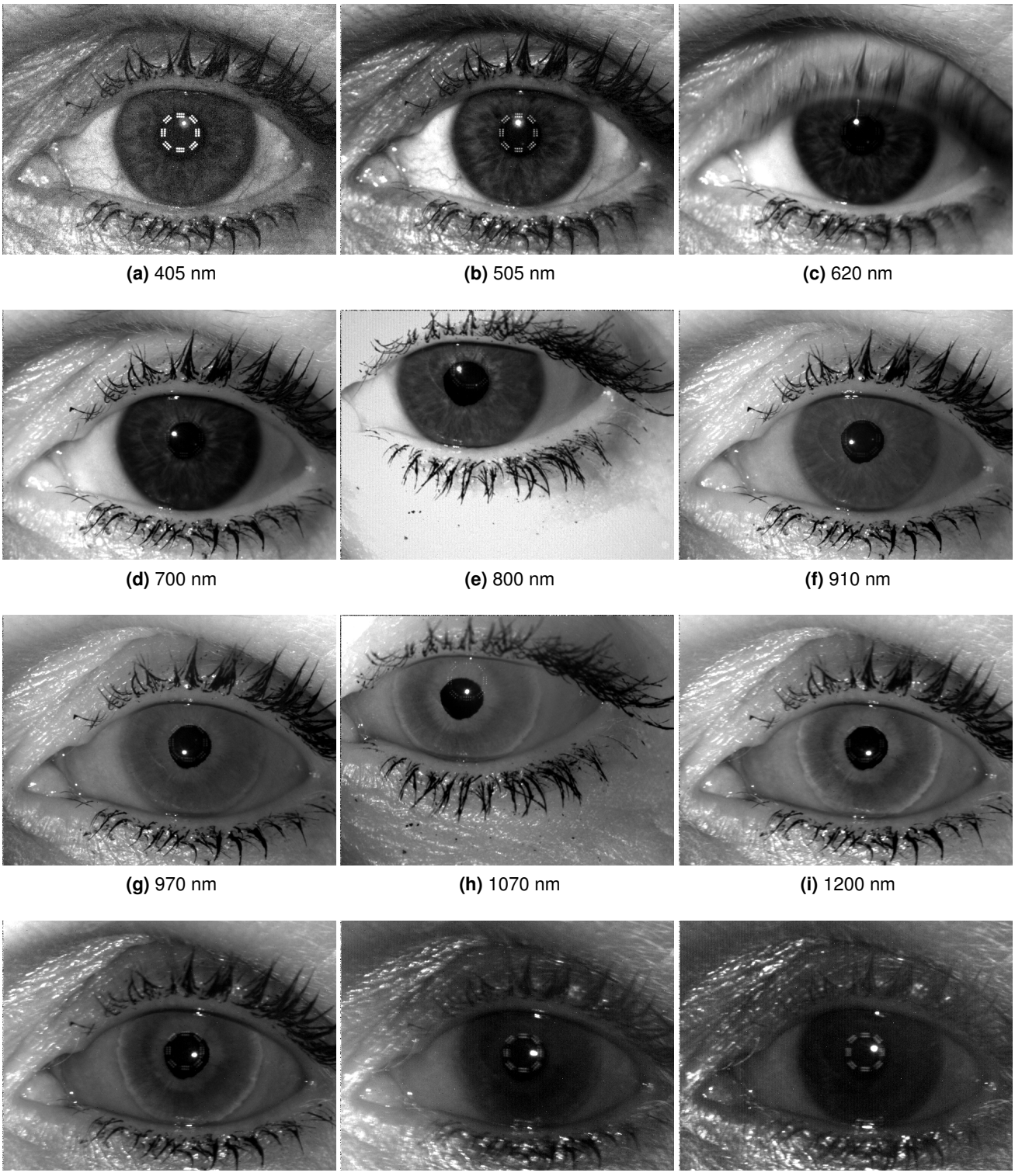
2.2 Test Dataset

The Consolidated Multi-Spectral Iris Dataset (CMID) was provided to NIST by the Southern Methodist University, which was sponsored by the US government to collect the images. Collection occurred in controlled laboratory environments using volunteers. For this reason, the quality of the samples is generally very high. The OPS-III dataset used in IREX IX: Part I was field-collected and contains poor quality samples (e.g. occluded and out-of-focus irides) with greater frequency. The following description of CMID is excerpted from [24]:

- **Nontraditional Spectrum –** Using a custom designed camera assembly, the CMID captures six images each of the right and left eye across a spectrum that ranges from 400 nm to 1600 nm. The LEDs used in this experiment have been certified as eye safe by multiple radiation safety experts as well as Institutional Review Boards at both Southern Methodist University (SMU) and the government sponsor. High-resolution visible light images of the ocular region are also taken using a professional photographic camera. Lastly, an image of the left and right iris is acquired using a commercial iris collection device.
- **Duration and Repetition –** The CMID collection is in its final (fourth) year with a goal of collecting each subject 16 times over that period.
- **Geographic Separation –** The CMID enrolled more than 400 subjects[1] across two geographically separated collection sites in order to increase the diversity of the collected subject pool. Roughly two-thirds of subjects are collected at the SMU research site.
- **Scale –** The CMID collects more than 160 iris images per session. The final CMID dataset is expected to contain more than 1 million laboratory quality iris images[2].

For samples acquired outside the normal 700 to 900 nm range, the pupil and limbus boundaries may be difficult to localize using standard techniques (e.g. Daugman's integrodifferentiable operator [25]). To address this problem, key points were manually identified for each iris image and passed to the matching software (see Figure 2.2). A final quality control step was performed to verify the overall accuracy of the boundary localizations. The full process is described in [24]. Figure 2.1 shows some example images from the dataset.

[1] The test set provided to NIST only contains samples from 220 subjects.
[2] The test set provided to NIST contains about 250 thousand images.



(a) 405 nm **(b)** 505 nm **(c)** 620 nm

(d) 700 nm **(e)** 800 nm **(f)** 910 nm

(g) 970 nm **(h)** 1070 nm **(i)** 1200 nm

(j) 1300 nm **(k)** 1450 nm **(l)** 1550 nm

Figure 2.1: A blue iris acquired at different illumination wavelengths from the CMID.

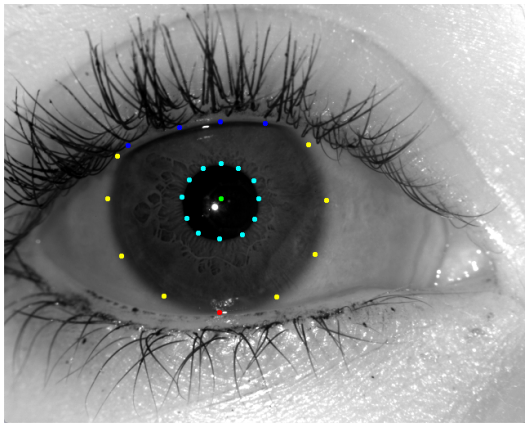


Figure 2.2: An example of an iris sample with manually identified boundary points. Eyelid, limbus, and pupil borders were all identified in addition to the iris center.

Participant	Phase 2	Phase 3
Aware Inc.	✓	✓
Decatur	✓	✓
DeltaID	✓	✓
Dermalog	✓	✓
FotoNation	✓	✓
IrisID	✓	✓
NEC	✓	✓
NeuroTechnology	✓	✓
Qualcomm	✓	✓
SOAR Advanced Technologies		✓
TAfIRT	✓	✓
Tiger IT	✓	✓
Unique Biometrics	✓	

Table 2.1: Participants of IREX IX along with the submission phases in which they participated. Phase 1 results are not published.

2.3 Matching Software

Thirteen commercial organizations and academic institutions submitted 46 iris recognition software libraries for evaluation. The participation window opened on October 7th, 2016 and closed on September 7th, 2017. Participation was open worldwide to anyone with the ability to implement an iris matching algorithm. There was no charge to participate.

Participants provided their submissions to NIST as static or dynamic libraries compiled on a recent Linux kernel. The libraries were then linked against NIST's test driver code to produce executables. A further validation step was performed to ensure that the algorithms produce identical output on both the participants' and NIST's test machines. The full process is described in the IREX IX API and CONOPS document [26].

Participants submitted their implementations in three rounds referred to as "phases". After the first two phases, participants were provided with rudimentary feedback on the performance of their submissions in the hope that it would assist with algorithm development for the next phase. Although only two phases were planned, a third phase was introduced and the first was designated a test phase. Table 2.1 lists the IREX IX participants along with the phases in which they participated. The deadline to submit to the second phase was January 21st, 2017 and the deadline for the third phase was September 1st, 2017. Each participant was required to submit at least one one-to-one implementation and one one-to-many implementation for each phase, although participants were allowed to submit up to two of each per phase. Some of the participants are new to the IREX program and some (Iris ID, NeuroTechnology, Delta ID, NEC, FotoNation) have participated in previous IREX evaluations.

2.4 Performance Metrics

Since this report has a more academic focus, accuracy is only assessed for one-to-one matching (a.k.a *verification mode*). Additionally, performance factors other than accuracy (*e.g.* computation time, template size, memory utilization) are not reported. Section 2.4.1 details the accuracy metrics while Section 2.4.2 provides a high-level introduction to estimates of uncertainty (*e.g.* confidence intervals) and how they should be interpreted in this report.

2.4.1 Accuracy

The degree of dissimilarity between two biometric templates is quantified by a dissimilarity score. In the case of John Daugman's IrisCode algorithm [27], the dissimilarity score is also known as a Hamming Distance. A dissimilarity score is referred to as *mated* if it is the result of comparing two templates representing the same iris (or pair of irides in the case of two-eye comparisons). It is known as a *nonmated* score if it is the result of comparing templates representing different irides. An identity claim is accepted if the dissimilarity score is below (or equal to) a preset decision threshold. Otherwise, the identity claim is rejected. As with any binary classification problem, two types of decision errors are possible. The first occurs when a nonmated comparison is misclassified as mated. This is known as a *false match*. The second type of decision error occurs when a mated comparison is misclassified as nonmated. This is known as a *false nonmatch*. The rates at which

these errors occur are the FMR and FNMR. Formally, let m_i $(i = 1...M)$ be a set of M mated dissimilarity scores and n_j $(j = 1...N)$ is a set of N nonmated dissimilarity scores. Then the accuracy statistics are computed as

$$\text{FNMR}(\tau) = \frac{1}{M}\sum_{i=1}^{M}[m_i > \tau], \qquad \text{and} \tag{2.1}$$

$$\text{FMR}(\tau) = \frac{1}{N}\sum_{j=1}^{N}[n_j \leq \tau]\,, \tag{2.2}$$

where τ is the decision threshold and $[...]$ is the Iverson Bracket [28] which denotes 1 if the boolean expression inside the bracket is true and 0 otherwise.

Since two types of decision errors are possible, no single numerical value can fully convey the accuracy of a biometric matching algorithm. Any fixed decision threshold yields a specific false match rate and false nonmatch rate. Adjusting the decision threshold reduces the rate of one type of error but at the expense of the other. This relationship is characterized by a DET (Detection Error Trade-off) curve [29], which plots the tradeoff between the two error rates. DET curves have become a standard in biometric testing, superseding the analogous ROC (Receiver Operating Characteristic) curve. Compared to ROC curves, the logarithmic axes of DET curves provide a superior view of the differences between matchers in the critical high performance region. A comprehensive introduction to the fundamentals of assessing the accuracy of binary classification systems can be found in [30].

2.4.1.1 Treatment of Feature Extraction Failures

Participants were instructed to provide submissions that always create comparable templates, even when no useful feature information could be extracted. These "blank templates" are expected to produce high measures of dissimilarity (effectively infinity) when compared to any other template. This was done for ease of testing but does not reflect operational reality. For example, a blank template would never be saved onto a smartcard and used for access control. If the template is being acquired in real-time from a cooperative user, the user could be prompted to provide a new sample or different accommodations could be made (e.g. using fingerprints instead). This inability to handle template creation errors in real time highlights a weakness of off-line testing.

2.4.2 Uncertainty Estimation

The Multispectral Dataset is a sampling of students and professors from the Southern Methodist University. This is not a perfect representation of the overall adult human population, but we nevertheless use the results in this report to draw conclusions about how automated iris recognition behaves over adults in general. Thus, we are effectively treating the set of all human adults as the *population*.

The confidence intervals presented in this report show how well the accuracy statistics calculated over our test data estimate the true population values. All of our confidence intervals are computed at the 90 % confidence level. This does not mean there is a 90 % probability that the true population value falls within the interval. Rather, it means that if the population is repeatedly sampled and an interval estimate is computed each time, the interval estimates would contain the true population value 90 % of the time.

The iris images in the Multispectral Dataset are paired in various ways to form comparison sets. These pairings introduce a correlation structure. For example, samples of a person's left and right eye captured during the same session are expected to be highly correlated in terms of sample quality. Wayman [31] found that failing to account for these dependencies can lead to overly optimistic estimates of confidence intervals. Thus, we took steps to factor the correlation structure into our estimates of uncertainty. The full procedure is detailed in Appendix A.

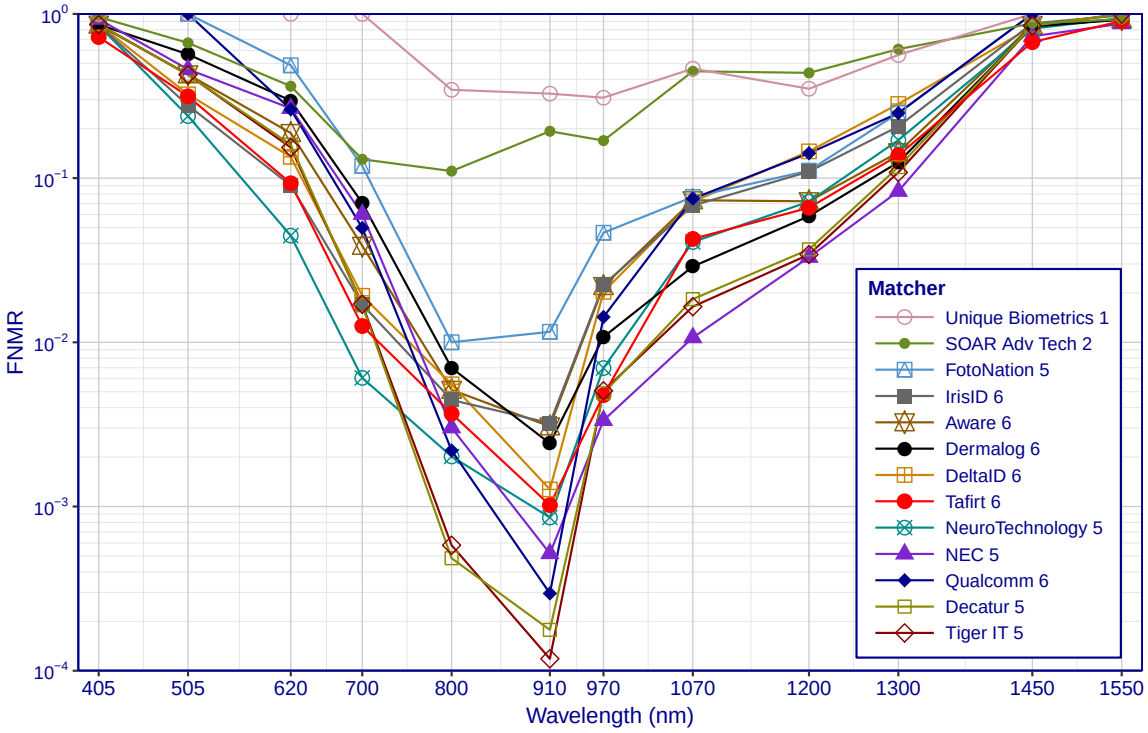


Figure 3.1: *FNMR as a function of the wavelength at which both of the compared samples were acquired. Results for two-eye comparisons are presented for 13 matchers. The decision threshold is fixed to elicit an FMR of 10^{-4} at 800 nm. Each point is generated from about half a million mated comparisons. FMR is computed from about 40 million nonmated comparisons.*

3 Results

3.1 Intra-spectral Matching

All currently deployed iris recognition systems operate on iris images illuminated in the near infrared (NIR) band of the electromagnetic spectrum. The ISO/IEC 19794-6 & 29794-6 standards require the eye to be illuminated between "approximately 700 and 900 nanometers (nm)" [2, 3]. NIR light is specified because melanin, the pigment that makes dark eyes dark[1], is nearly transparent in the NIR. This makes the stromal structure of dark brown irises easier to resolve. Operation at still longer wavelengths becomes problematic because fluids bathing the iris are strongly absorbing at wavelengths beyond 1000 nm and silicon based image sensors lose essentially all sensitivity. This section explores iris matching accuracy over wavelenegths ranging from a low of 405 nm (the nominal short edge of the visible spectrum) to a high of 1550 nm.

Figure 3.1 plots FNMR as a function of the wavelength at which the samples were acquired. All comparisons are between samples acquired at the same wavelength. When computing FNMR, the decision threshold was fixed to elicit an FMR of 10^{-4} when the samples were acquired at 800 nm. Only one matcher from each participant is shown because matchers submitted by the same participant tend to exhibit similar behavior. Eleven of the 13 IREX-IX matchers produce their lowest FNMR at 910 nm. FotoNation produces a slightly lower FNMR at 800 nm than at 910 nm but the difference is probably not statistically significant. The line segments connecting points are intended to make it easier for readers to track the results for specific matchers and should not be regarded as reliable interpolations. The true minimum FNMR for each of the 11 matchers could lie anywhere between 800 nm and 970 nm. With the exception of Unique Biometrics 1, none of the matchers produce their lowest FNMR at 700 nm or 970 nm. FNMR appears to be extremely dependent on wavelength, even within the standard 700-900 nm band. For example, FNMR is about 25 times greater at 700 nm compared to 800 nm for Qualcomm 6.

At the visible wavelengths 505 nm and 620 nm, NeuroTechnology 5 achieves the lowest FNMR among all matchers, although several other matchers perform better at longer wavelengths. At 620 nm, NeuroTechnology 5 produces an FNMR of 0.045

[1] 70 to 90 % of the world's population have dark brown eyes.

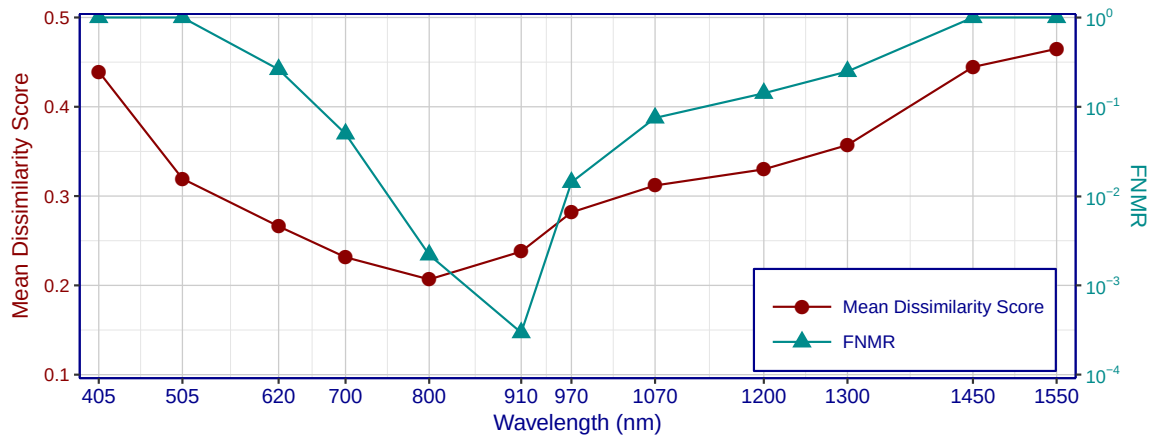


Figure 3.2: *FNMR and mean dissimilarity score as a function of the wavelength at which both of the compared samples were acquired. Two-eye matching results are presented for Qualcomm 6. When computing FNMR, the decision threshold was fixed to elicit an FMR of* 10^{-4} *at 800 nm. Each point is generated from about half a million mated comparisons.*

(corresponding to a "true match rate" of 0.955). Matching does not appear viable at 405 nm, near the blue edge of the visible spectrum. A visual inspection of the iris images revealed that little iris texture is visible at this wavelength (although blood vessels in the sclera are considerably more apparent than at any other tested wavelength). For all matchers, FNMR increases monotonically as the wavelength decreases from 800nm to 405 nm. Similarly, FNMR increases monotonically as the wavelength elongates from 800 nm to 1550 nm (Unique Biometrics 1 excepted).

The results presented here are not in perfect alignment with the existing literature. Ngo *et. al.* [5] found that Daugman's normalized Hamming distance[2] for mated comparisons is minimized when both of the compared iris images are acquired at 800 nm, suggesting this might be the best wavelength at which to acquire iris images. However, Figure 3.1 in this report seems to indicate FNMR is minimized at 910 nm. The dataset used by Ngo *et. al.* at the USNA was small, containing only 6 subjects, and no manual markup of the iris boundaries was provided. Ives *et. al.* [6], also at the USNA, expanded on Ngo and found that when boundary coordinates were not provided, the lowest normalized Hamming distance was achieved at 910 nm, with slightly increased distances at the neighboring wavelengths 810 nm and 970 nm. When boundary coordinates were provided, the mean distance score varied little between 500 and 900 nm, with the lowest mean score at 590 nm. This contrasts with the current report, which indicates accuracy is highly sensitive across this wavelength band.

The apparent discrepancy between the current results and previous research could be explained in part by the different accuracy metrics used. This report uses FNMR (at fixed FMR) to investigate the impact of wavelength on the ability of the matchers to recognize irises. Most previous studies used the mean SQRT-normed Hamming Distance (a common measure of dissimilarity) rather than FNMR. Figure 3.2 demonstrates how conclusions can differ depending on the metric used. Results are shown for Qualcomm 6 because it produces Hamming distance like scores[3]. The figure demonstrates that mean dissimilarity score is minimized at 800 nm while FNMR is minimized at 910 nm. Mean score is a more robust statistic than FNMR in the sense that it is less sensitive to outliers. However, in the context of iris recognition, outliers are of particular interest because they tend to dictate matcher accuracy. An exception would be if the outliers are caused by unwanted covariates. For example, capturing at 620 nm could be more distracting to the subject because the illuminators are visible to the naked eye. To mitigate such possibilities, SMU manually inspected the images in the CMID. Since the CMID illumination was provided by LEDs at the respective wavelengths and the LEDs are, by necessity, at different locations, we cannot rule out, at this time, the possibility that some of the observed variance is due to illumination position.

Figure 3.3 plots FMR as a function of the wavelength at which both of the compared samples were acquired. Results are less consistent compared to Figure 3.1. Most of the matchers produce higher FMRs at 505 nm and 620 nm compared to 800 nm. For NeuroTechnology 5, FMR increases from 0.0001 to ≈ 0.0006 when going from 800 nm to 620 nm. Within the range 505 nm to 1300 nm, FMR varies by about two orders of magnitude for several of the matchers. At the given decision threshold, the FMR is 500 times greater at 970 nm compared to 700 nm for DeltaID 6. With the exception of Tafirt 6, FMR never varies by more than about a factor of two between 970 nm and 1200 nm for any of the matchers.

[2] A correction factor applied to iris comparison scores first proposed by Daugman [4]. The score is adjusted based on the amount of overlapping iris texture between the compared iris images. The result is a more stable and predictable non-mated distribution. This can be useful for operational systems but can confound results in laboratory experiments.

[3] Algorithms submitted to IREX are typically black boxes; we are not given insight into the internal details of the algorithms.

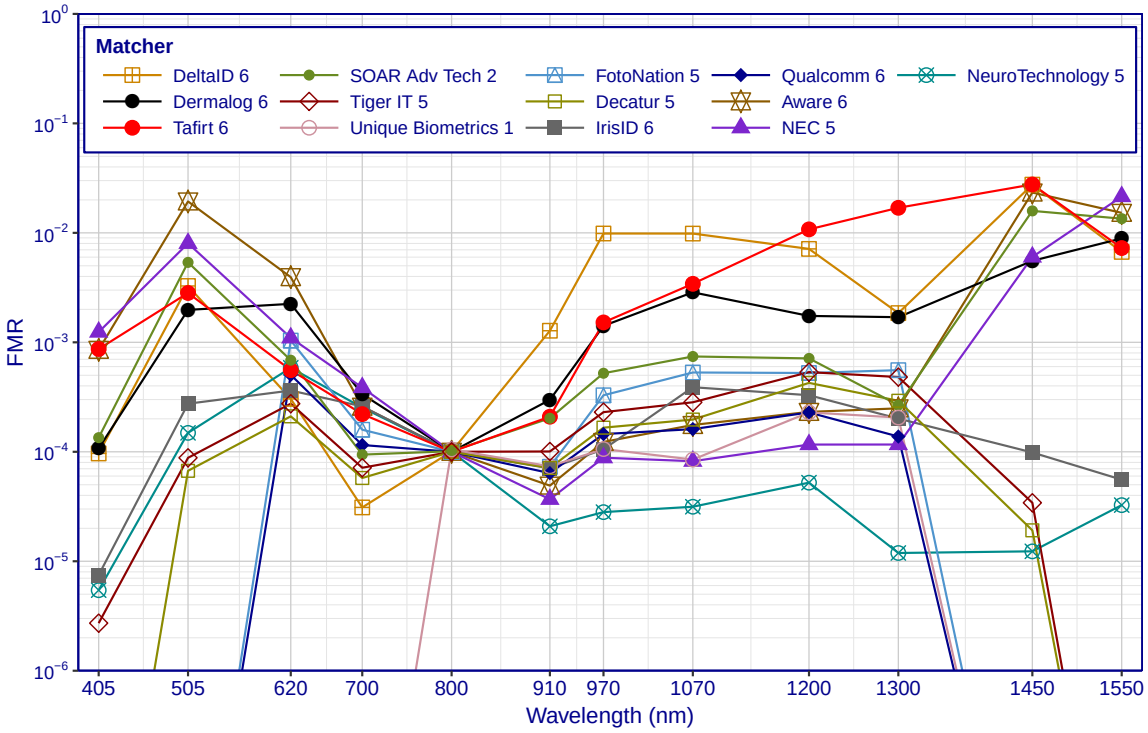


Figure 3.3: *FMR as a function of the wavelength at which both of the compared samples were acquired. Results for two-eye comparisons are presented for 13 matchers. The decision threshold is fixed to elicit an FMR of 10^{-4} at 800 nm. Each point is generated from about 40 million nonmated comparisons.*

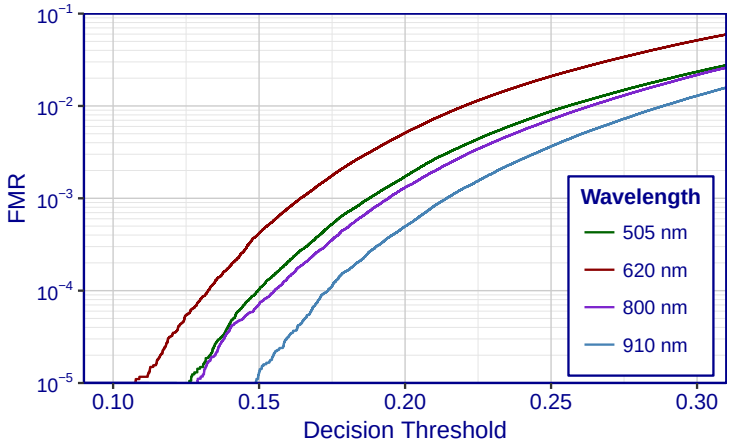


Figure 3.4: *FMR as a function of the decision threshold for different capture wavelengths. Results are presented for two-eye comparisons for NeuroTechnology 5.*

Figure 3.4 plots FMR as a function of dissimilarity score for NeuroTechnology 5. Such plots provide more information about the non-mated distributions but can be more difficult to interpret and analyze, especially when the distributions do not approximate smooth curves. The figure demonstrates that FMR varies by about a factor of 10 depending on the wavelength, with the lowest FMR at 910 nm and the highest at 620 nm, although the factor difference tends to increase as the decision threshold decreases.

3.2 Cross-spectral Matching

This section assesses accuracy when iris samples acquired at different wavelengths are compared to each other. Figure 3.5 is a heatmap where FNMR is indicated by color and the axes specify the illumination wavelengths at which the verification and enrollment samples were acquired. The main diagonal shows FNMR when both compared samples were acquired at the same illumination wavelength. Broadly speaking, FNMR tends to be lower when the samples were acquired at the same, or similar, wavelengths, and is lowest when both samples were acquired at 910 nm. The authors were curious about what might be causing the few misses that do occur at 910 nm. A manual inspection found that some quality-related problems are present in these images (despite attempts by SMU to filter such images out). The most prominent problems appear to be blur and limited visible iris texture due to eyelid occlusion (and often a combination of the two). The IREX III supplemental report found that eyelid occlusion compounded by blur is a common cause of misses. When both samples were acquired at 620 nm, the FNMR is 0.045. When one sample was acquired at 620 nm and the other at 800 nm, FNMR is slightly greater at 0.055. Even though 910 nm appears to be better for NIR matching, samples acquired at the shorter NIR wavelengths (700 and 800 nm) yield much lower FNMRs when compared to samples acquired at visible wavelengths.

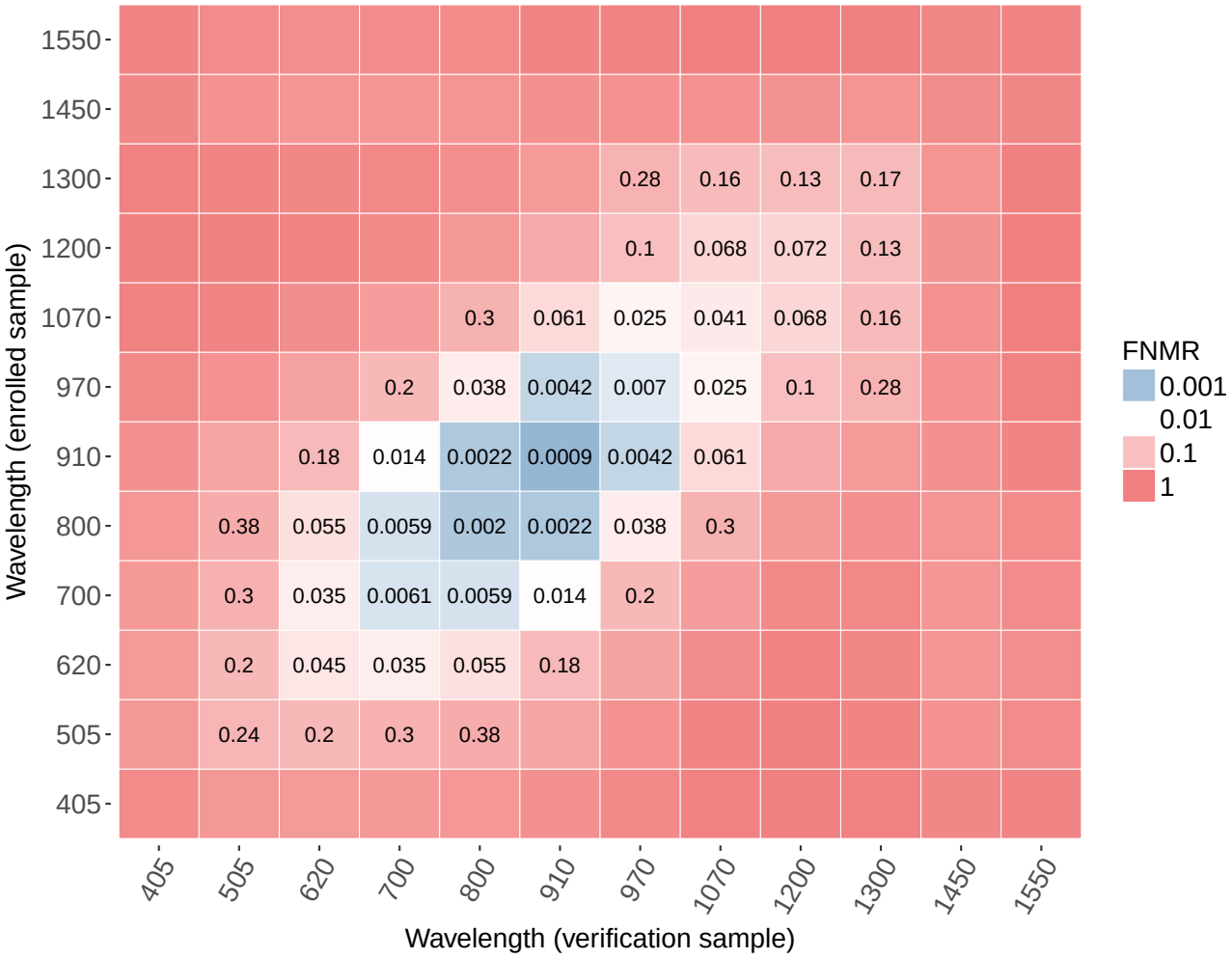


Figure 3.5: *Relationship between FNMR and the wavelength at which the compared samples were acquired. Capture wavelength for verification and enrollment samples is indicated by the axes. color indicates FNMR, with the precise number sometimes shown in the cell. Results are for two-eye comparisons for NeuroTechnology 5. Each FNMR value is generated from about 65 thousand mated comparisons. The decision threshold is set to yield an FMR of 10^{-4} when both of the compared samples were acquired at 800 nm.*

Figure 3.6 is a heat map where FMR is indicated by color and the axes specify the illumination wavelengths at which the verification and enrollment samples were acquired. It is similar to Figure 3.5 except the color indicates FMR rather than FNMR. FMR appears to be higher at shorter wavelengths (specifically between 505 nm and 800 nm). FMR is highest when both compared samples were acquired at 620 nm. FMR is almost always zero or very nearly zero when one of the compared samples was acquired at 405 nm, or at 1450 nm or above. When one of the compared samples was acquired between 910 nm and 1300 nm results are much less consistent. FMR is often unusually low when one of the samples was acquired at 910 nm, This is probably an artifact of NeuroTechnology's matcher as it does not occur with the other matchers. Figures 3.3 and 3.6 both demonstrate the lack of stability of the nonmated distribution across wavelengths.

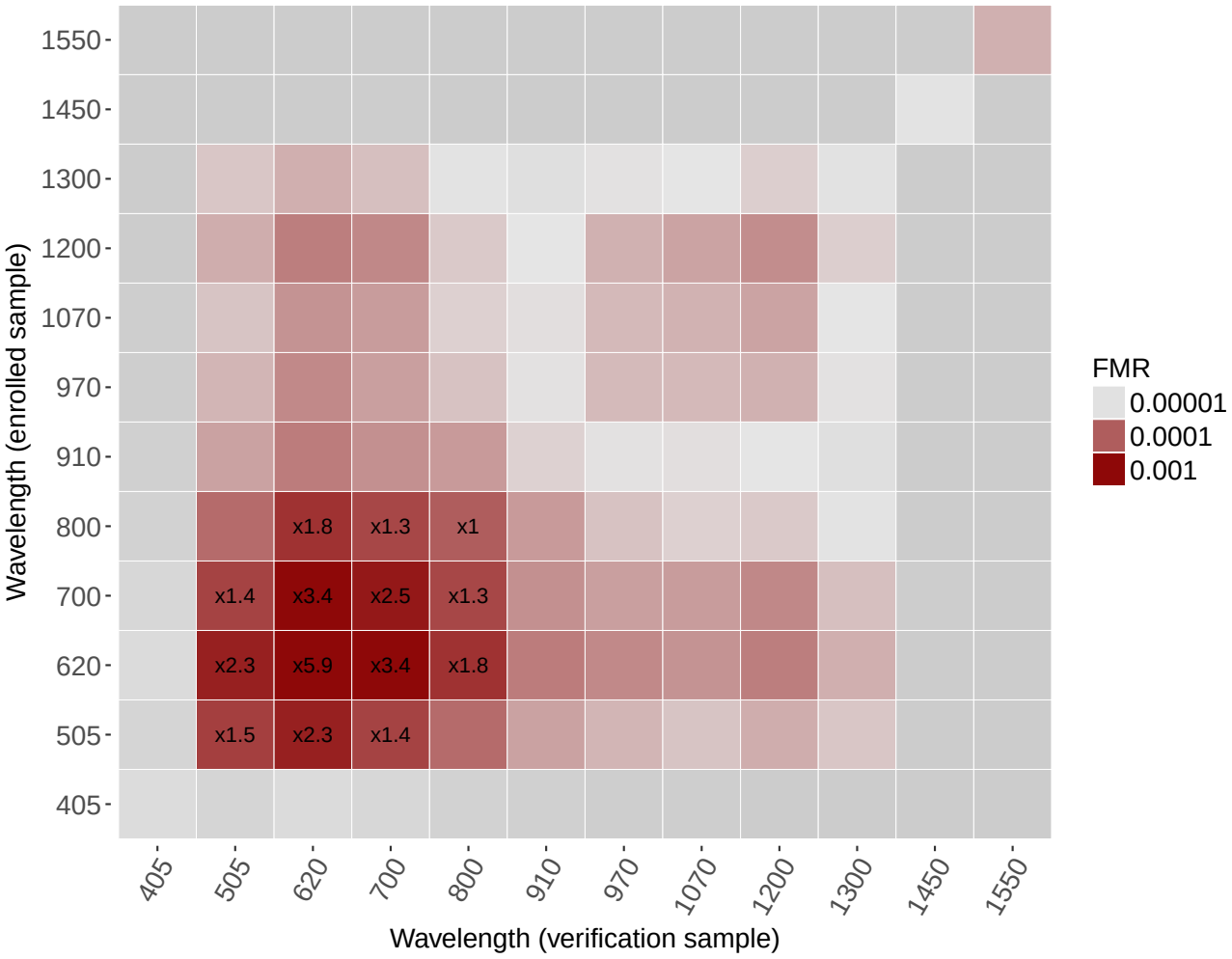
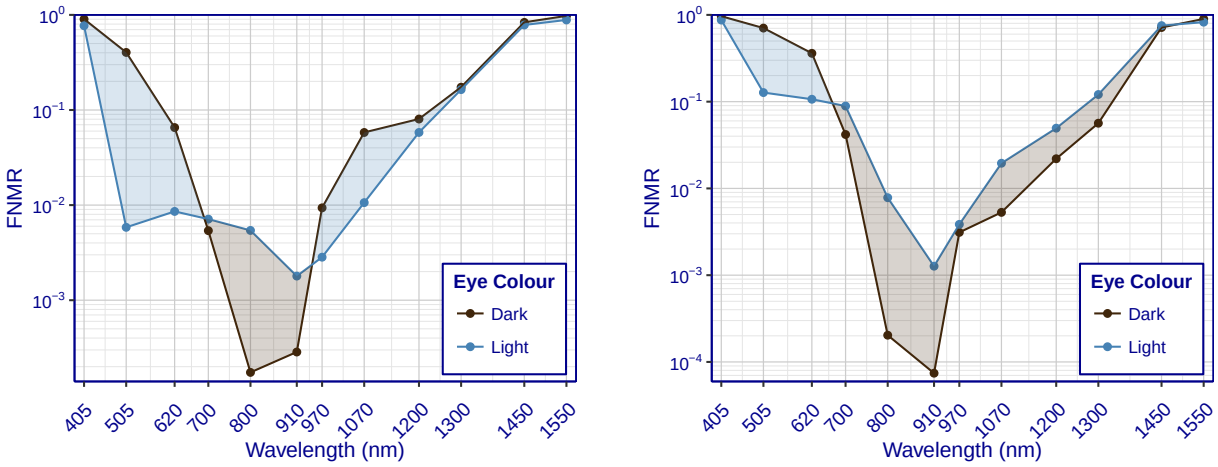


Figure 3.6: *Relationship between FMR and the wavelength at which the compared samples were acquired. Capture wavelength for verification and enrollment samples is indicated by the axes. color indicates FMR. Sometimes the factor increase in FMR compared to when both samples were acquired at 800 nm is shown in the cell. Results are for two-eye comparisons for NeuroTechnology 5. Each FMR value is generated from about a million and a half nonmated comparisons.*



(a) NeuroTechnology 5 **(b)** NEC 5

Figure 3.7: *FNMR as a function of capture wavelength when comparisons are broken out by eye hue (light or dark). Results are presented for two-eye comparisons for NeuroTechnology 5 and NEC 5. The decision threshold is fixed to elicit an FMR of* 10^{-4} *at 800 nm ignoring eye color. Each point is generated from about 25 thousand comparisons for light eyes and 40 thousand for dark eyes.*

3.3 Effect of Eye Color

Several studies have suggested that accuracy at visible wavelengths is likely to be impacted by eye color [32, 5, 33, 34] although they did not investigate the problem directly. As previously noted, melanin pigments in darker irises absorb light at visible wavelengths, making recognition more difficult. This is demonstrated for two matchers in Figure 3.7, which plot FNMR as a function of the wavelength at which both of the compared samples were acquired. Comparisons are further broken out by eye color. Most subjects in the dataset have brown eyes (297). Comparisons involving these subjects are classified as *Dark*. Comparisons involving subjects with green (78 subjects), grey (9 subjects), blue (148 subjects), and blue-green (28 subjects) are classified as *Light*. Comparisons involving hazel-eyed subjects are ignored. For both NeuroTechnology 5 and NEC 6, FNMR is lower for light-eyed comparisons than dark-eyed comparisons at 505 nm and 620 nm. This holds true for all of the matchers tested. At 700 nm, FNMR is comparable between the two eye hues. At this wavelength, absorption by the melanin is almost completely attenuated [35]. Interestingly, FNMR is typically lower for darker eyes than lighter eyes at 800 nm and 910 nm. This holds true for nearly all of the more accurate matchers over the CMID images (Qualcomm 6, Tiger IT 6, NeuroTechnology 5, DeltaID 6, and NEC 5). The reason is unclear. It could be the result of more rigorous training or tuning over individuals with darker eyes, or there could be something intrinsic in the features or behavior of individuals with darker eyes that makes them easier to recognize. Limbus, pupil, and eyelid boundary coordinates were provided for all iris images, so the difference cannot be due to the accuracy of boundary localization.

3.4 Multispectral Fusion

Multiwavelength fusion refers to combining multiple acquisitions of the iris at different wavelengths to improve performance. Fusion can occur at four possible levels depending on where along the matching process the data from multiple biometric sources are integrated. The four levels are sensor, feature, score, and decision level. The sensor and feature levels are referred to as pre-mapping (*i.e.* before matching) fusion while the score and decision levels are referred to as post-mapping (*i.e.* after matching) fusion [36]. Post-mapping fusion strategies tend to be fairly simple and straightforward to implement. The downside is that the data is combined further along in the matching process and is thus heavily reduced by the time it is fused. This can diminish the potential for improving performance since a lot of the original data is lost before fusion occurs.

3.4.1 Score Level

Score level fusion refers to combining comparison scores from different biometric sources into a single score. The fused score is then used to make the final match/nonmatch decision. Figure 3.8 depicts the process of combining two scores from different sources into a single score. Score level fusion occurs further along in the matching process than sensor and feature level fusion but earlier than decision level fusion. In the case of multispectral fusion, each score corresponds to a comparison performed at a different illumination wavelength. So *Verification Template 1* and *Enrollment Template 1* in the figure would be created from samples acquired at one illumination wavelength while *Verification Template 2* and *Enrollment Template 2* would be created from samples acquired at another illumination wavelength. It is common to use the same matcher to produce all of the comparison scores. A disadvantage of applying fusion at the score level is that multiple samples must be acquired and stored. Transferring the samples over a network could also take longer if the bandwidth is low. Although score level fusion increases computational complexity, many of the operations can be performed in parallel to mitigate the real-time increase in processing time.

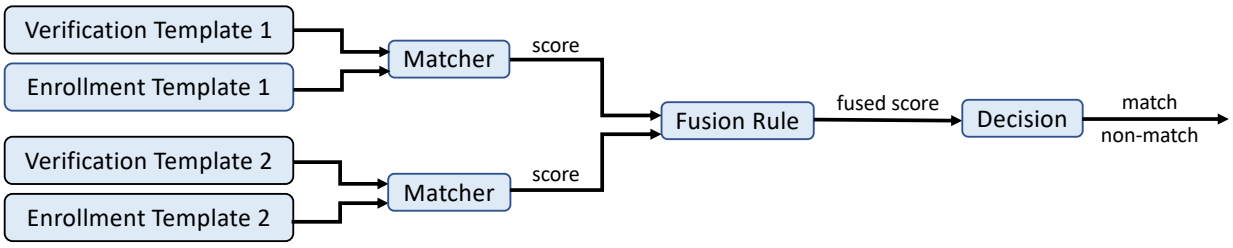


Figure 3.8: *Depiction of biometric fusion occurring at the score level. Two comparisons between verification and enrollment templates are performed, each producing a score. The scores are then combined using a fusion rule. A final decision is made using the fused score.*

Abdullah *et. al.* [37] applied simple sum-rule fusion to combine scores acquired from VW and NIR comparisons and found that it significantly improved accuracy over using NIR comparisons alone. Boyce [38] confirmed this result using a larger dataset. Gong *et. al.* [39] similarly found that accuracy could be improved by combining scores acquired at various illumination wavelengths. Most of the improvement in their analysis occurred when they combined scores acquired at 700 and 850 nm within the NIR band. Smaller improvements were achieved by incorporating scores acquired at shorter visible wavelengths. All of the aforementioned studies used sum-rule fusion which involves summing the scores acquired at the different wavelengths. Sum-rule fusion is typically regarded as occurring at the score level although a nearly identical result could be achieved at the feature level by concatenating (and then comparing) the IrisCodes created from iris images acquired at different illumination wavelengths.

The current analysis uses the CMID, which as of this writing is larger than any other dataset used to test multispectral iris fusion. During each capture session, the subject was positioned in front of the iris camera and several images of each iris were acquired at different wavelengths. Approximately 20 seconds elapsed between captures. Manually specified boundary coordinates for each image (see Section 2.2) were provided to the matchers during template creation. This prevents the accuracy of boundary localization from being a potential confounding factor in the analysis. Fusion is applied at the score level using the unweighted version of Neyman-Pearson Fusion (NPF) proposed by Hube [40]:

$$s_{\text{fused}} = s'_1 + \ldots + s'_k \tag{3.1}$$

where

$$s'_i = -\log \text{FMR}(s_i) \tag{3.2}$$

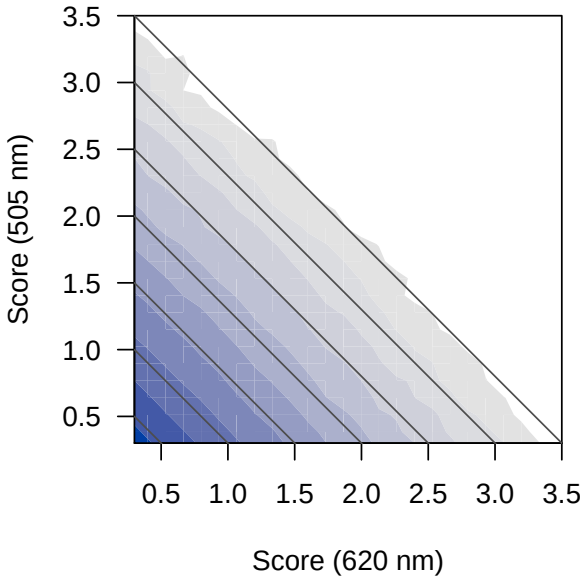


Figure 3.9: *Neyman Pearson boundaries for Aware 6. The axes correspond to LFMR scores acquired at two different illumination wavelengths (505 and 620 nm). The color indicates the density. The black lines show level curves for unweighted sum-rule fusion. The lines trace the joint density of the LFMR scores quite well.*

and s_i is the raw comparison score (often a Hamming Distance) at the wavelength indexed by i ($1 \leq i \leq k$). Equation 3.2 is typically referred to as the LFAR of the score but to remain consistent with ISO/IEC JTC 1/SC 37/WG 1 [41] terminology it will be referred to as the LFMR of the score. Note that Equation 3.2 converts the score from a measure of dissimilarity to a measure of similarity. The LFMR of the score is easier to interpret than the raw score. The probability of a nonmated comparison producing an LFMR score of 2 or higher is 10^{-2}. The probability of producing an LFMR score of 3 or higher is 10^{-3} and so on. The only downside to working with the LFMR score is that it requires precise knowledge of the nonmated distribution to perform the conversion.

There is a strong theoretical justification for using Neyman-Pearson Fusion. The name of the fusion rule references the Neyman-Pearson Lemma [42] which in the context of biometrics asserts that thresholding the likelihood ratio (*i.e.* the likelihood of the comparison being mated vs. nonmated) is the optimal method for making match/nonmatch decisions. When only a single score is available (i.e. no fusion is applied) the score is usually monotonic with the likelihood ratio. Hube's definition of Neyman-Pearson Fusion would be optimal but for the fact that it makes a few simplifying assumptions when combining the scores. Firstly, the LFMR score is typically computed using an empirical estimation of the FMR of the raw score. Secondly, the fusion rule assumes the scores are independent. Finally, the unweighted version of Neyman Pearson Fusion assumes FNMR increases linearly with the log of FMR (i.e. the ROC curve has a constant slope). To address the second problem, Hube suggests fusing the LFMR scores using a rule that closely aligns with the level curves of the joint density of the nonmated scores. Figure 3.9 shows such a density plot for Aware 6 given LFMR scores acquired at two different illumination wavelengths. The solid black lines show the level curves for unweighted sum-rule fusion which align quite well with the contour lines of the joint density plot.

Figure 3.10 shows the results of applying Neyman-Pearson Fusion using scores acquired at different wavelengths within the NIR band (700 nm to 910 nm). The bars show 90 % confidence intervals (computed using the method described in Appendix A). The figure demonstrates that combining scores acquired at 910 nm with scores acquired at either 700 nm or 800 nm leads to substantial improvements in accuracy for nearly every matcher. In the case of NEC 5, FNMR drops from 0.0132 to 0.0014, a factor of 9 improvement. Most of the matchers do not achieve significant improvements by incorporating a third wavelength. The exceptions, IrisID 6 and DeltaID 6, had high FNMRs compared to the other matchers before incorporating the third wavelength. Several matchers appear to bottom out at an FNMR around 0.0014 when all three wavelengths are used. A quick manual inspection of these images revealed that all, or nearly all, of the failed matches are due to ground truth errors (*i.e.* incorrect person identifiers being assigned to some of the iris images). Thus, the "matching errors" are actually mistakes in labeling the test data. Although no formal process for manually comparing iris images has been established, the iris images have superb sample quality and it was straightforward for the authors to determine that the textures did not match.

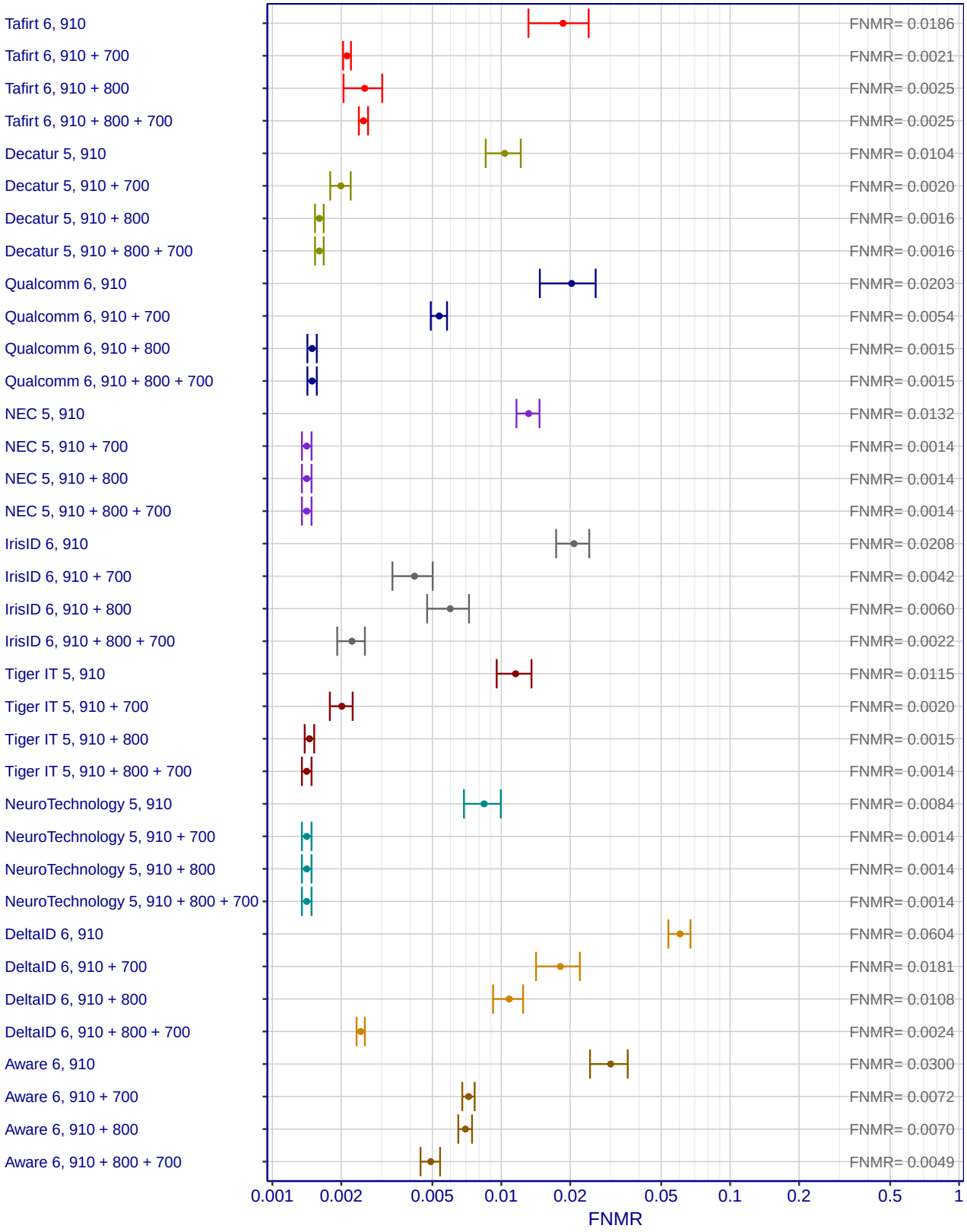


Figure 3.10: FNMR *(at* FMR=10^{-4}*) for selected matchers and wavelength combinations. Fusion of scores at each illumination wavelength was performed using Neyman-Pearson Fusion. Results are for single-eye comparisons. Each point was generated using about 40 thousand mated comparisons.*

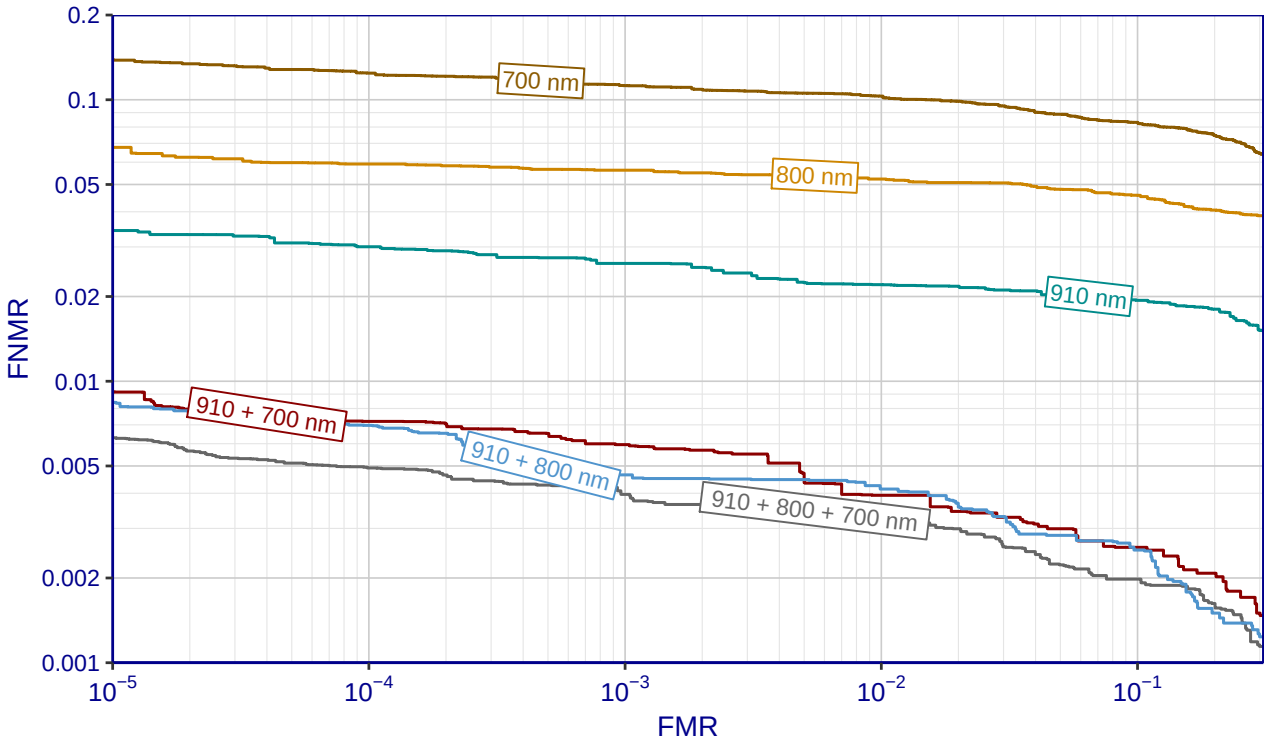


Figure 3.11: *Full DET plots for different combinations of illumination wavelengths for Aware 6. Fusion is performed at the score-level using unweighted Neyman-Pearson Fusion. Results show single-eye matching results. Each DET curve is generated from about 40 thousand mated scores and 13 million nonmated scores.*

Figure 3.11 shows full DET plots for Aware 6 when different combinations of wavelengths are fused together using unweighted Neyman-Pearson Fusion. As is the case with all of the matchers, any combination of wavelengths achieves better accuracy than any single wavelength by itself. No appreciable difference in accuracy appears to exist between $910+700$ nm and $910+800$ nm. The small apparent improvement in accuracy for $910+800+700$ nm may not be statistically significant. (The fact that the 90 % confidence bounds shown in the previous figure between $910+800+700$ nm and $910+800$ nm do not overlap does not necessarily mean that the difference is statistically significant). Neyman-Pearson Fusion does not appear to produce DETs that are radically different in overall shape or curvature compared to DETs produced by the original iris dissimilarity scores.

Figure 3.12 shows the results of using Neyman-Pearson Fusion to combine scores acquired at wavelengths in the VW band (405 - 620 nm). It is analogous to Figure 3.10 except it applies fusion within the VW rather than NIR band. Fusion does not appear to substantially improve accuracy for any of the matchers. It also does not appear to significantly *reduce* accuracy for any of the matchers. Matching at 405 nm produces extremely high error rates (FNMR $>$ 0.96 at $FMR = 10^{-4}$ for all matchers) and is therefore not expected to improve accuracy through fusion. Error rates at 505 nm are also very high (FNMR $>$ 0.96 at $FMR = 10^{-4}$) again limiting the potential benefit of fusion.

Fusion does appear to produce small improvements in accuracy for many of the matchers. For example, the FNMR for Tiger IT 5 drops from 0.245 to 0.195 when scores acquired at 620 are combined with scores acquired at 505 nm. Figure 3.14 demonstrates that this improvement is limited to comparisons involving lighter eyes (*i.e.* green, blue, grey). The figure shows full DETs for DeltaID 6 but the stated conclusion holds for all of the matchers where fusion produces a clear (but minor) improvement in accuracy. Fusing 620 nm with 505 nm produces an FNMR of 0.026 at $FMR = 10^{-4}$ for DeltaID 6 for lighter eyes, far below that of 505 nm (FNMR$= 0.070$) or 620 nm (FNMR$= 0.086$) alone. In contrast, fusing the scores for dark eyes (*i.e.* brown, black) produces DET curves that are nearly indistinguishable from the DET produced at 620 nm alone. The higher concentration of melanin in brown eyes is clearly obscuring the iris texture at the lower visible wavelengths, limiting the benefit of fusion.

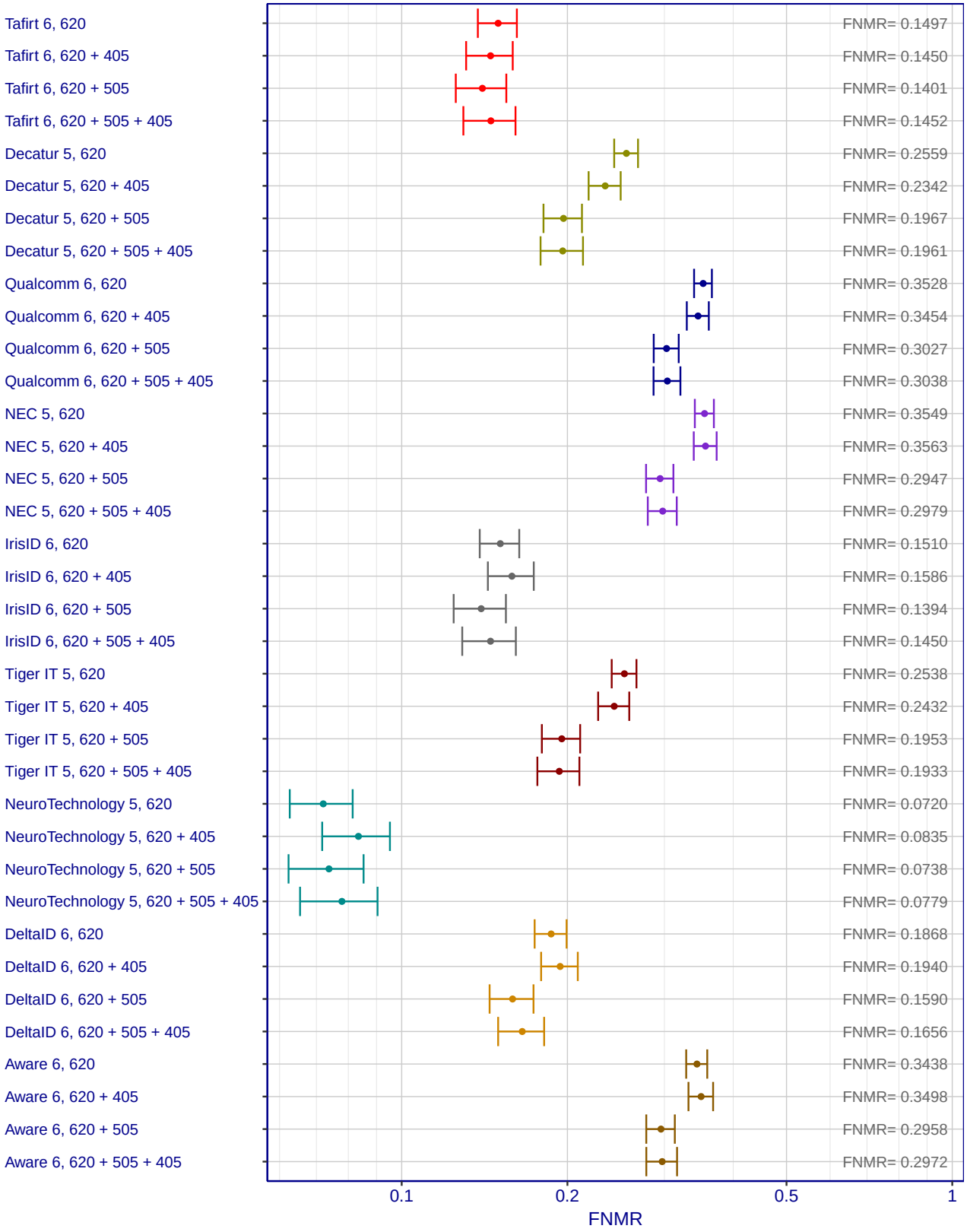


Figure 3.12: *FNMR (at FMR=10^{-4}) for selected matchers and wavelength combinations. Fusion of scores at each illumination wavelength was performed using Neyman-Pearson Fusion. Results are for single-eye comparisons. Each point was generated using about 80 thousand mated comparisons.*

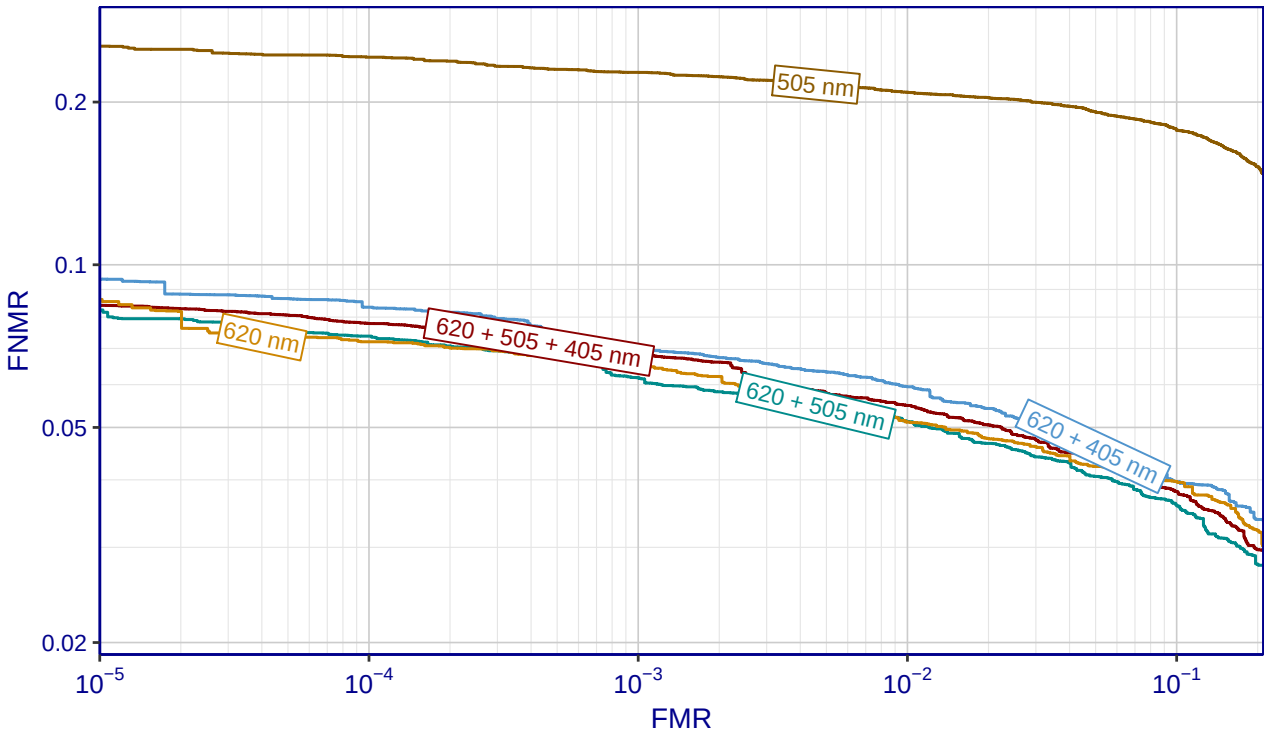


Figure 3.13: *Full DET plots for different combinations of illumination wavelengths for NeuroTechnology 5. Fusion is performed at the score-level using unweighted Neyman-Pearson Fusion. Results show single-eye matching results. Each DET curve is generated from about 80 thousand mated scores and 20 million nonmated scores.*

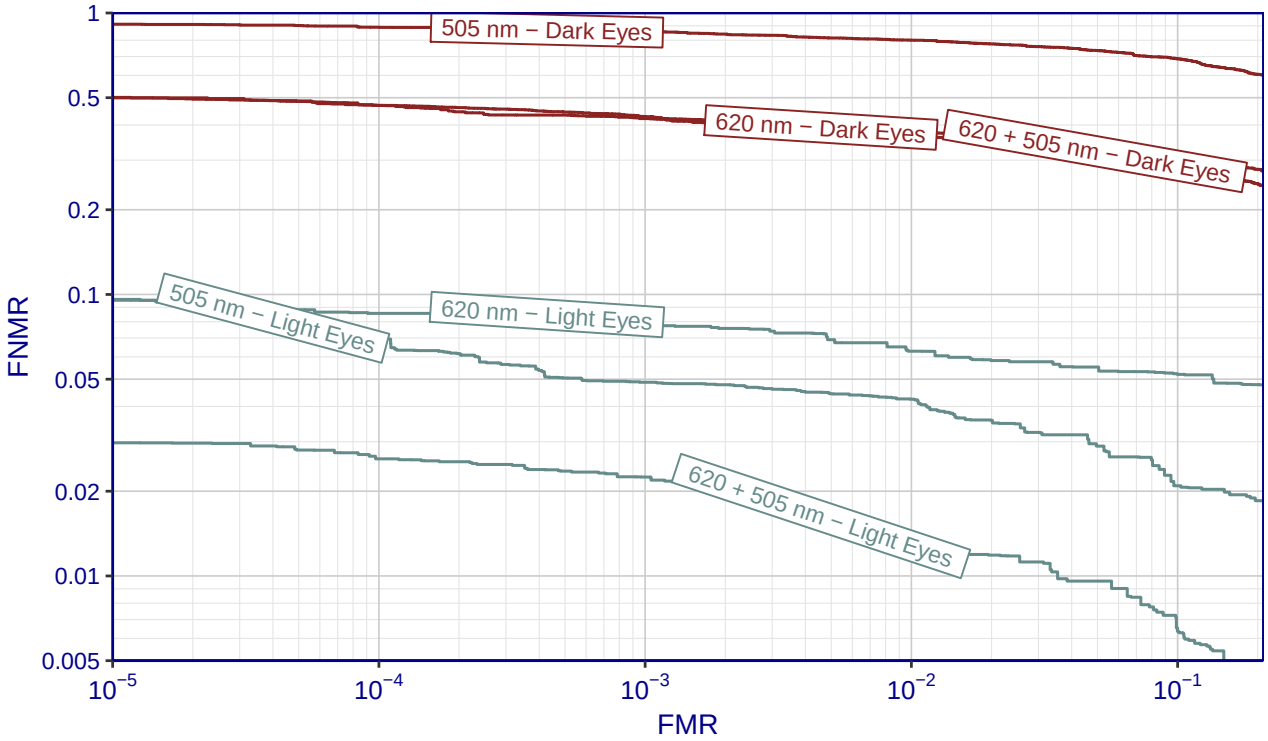


Figure 3.14: *Full DET plots for different combinations of illumination wavelengths for DeltaID 6. Fusion is performed at the score-level using unweighted Neyman-Pearson Fusion. Results show single-eye matching results.*

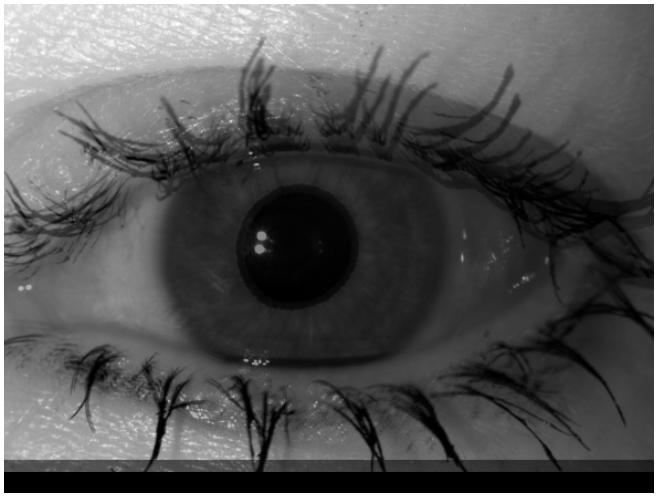


Figure 3.15: *Example of a composite image created by combining iris images acquired at 700, 800, and 910 nm. Each image was aligned and unweighted pixel averaging was used to create the composite image. Although the iris texture is clear, the eyelashes appear blurry because the eyelid was open by slightly different amounts across images.*

3.4.2 Sensor Level

Sensor level fusion refers to consolidating information acquired by multiple sensors (*i.e.* capture devices) into a single biometric sample. In the case of multiwavelength iris recognition this would entail acquiring the same image at different illumination wavelengths and combining the information into a single image. Features would then be extracted from the composite image to produce a template that is in turn is used for matching. A sample similar to a composite image can be produced by simply illuminating the iris at multiple wavelengths simultaneously during capture. In fact, many iris cameras do this to enhance interoperability.

Composite images were created by superimposing iris images acquired at 700, 800, and 910 nm. Accurately superimposing the images required precisely aligning them. The manually specified iris center was used for initial alignment. The OpenCV Library's [43] geometric transformation capabilities were used to achieve a precise pixel-level alignment. The library was used to align one iris with another by translating and rotating one of the images. Possible perspective distortions were ignored. Images were combined using equal-weight pixel averaging. The three combined images always came from the same capture session. Although the subject held their head in the same position throughout the capture session, acquisition of iris samples at different wavelengths was staggered, with each image being acquired a few seconds to a few minutes apart. The amount of dilation, the position of the eyelids, and the gaze angle could have changed during this time. Figure 3.15 shows an example of a composite image where the eyelid positions differed slightly across captures.

Figure 3.16 shows FNMR at fixed FMR for each matcher when comparing composite images to iris images acquired at various wavelengths. Comparisons involving composite images are never worse than the worst-case scenario: comparing samples acquired at opposite ends of the standard NIR band (700 nm to 910 nm). Nevertheless, composite images never perform as well as the best-case scenario: comparing samples that were both acquired at 910 nm. Figure 3.17 shows full DETs for IrisID 6. Although sensor-level fusion never catastrophically increases the error rates, neither does it lead to any substantial improvement in accuracy. For the case of IrisID 6, using a sample acquired at 910 nm seems to perform at least as well as the composite images in all cases and sometimes noticeably better (e.g. comparing samples both acquired at 910 nm always outperforms comparing composite images to images acquired at 910 nm). Section 3.4.1 also demonstrated that better results can be achieved using score-level fusion. The current investigation was unable to demonstrate that sensor level fusion offers any clear benefits.

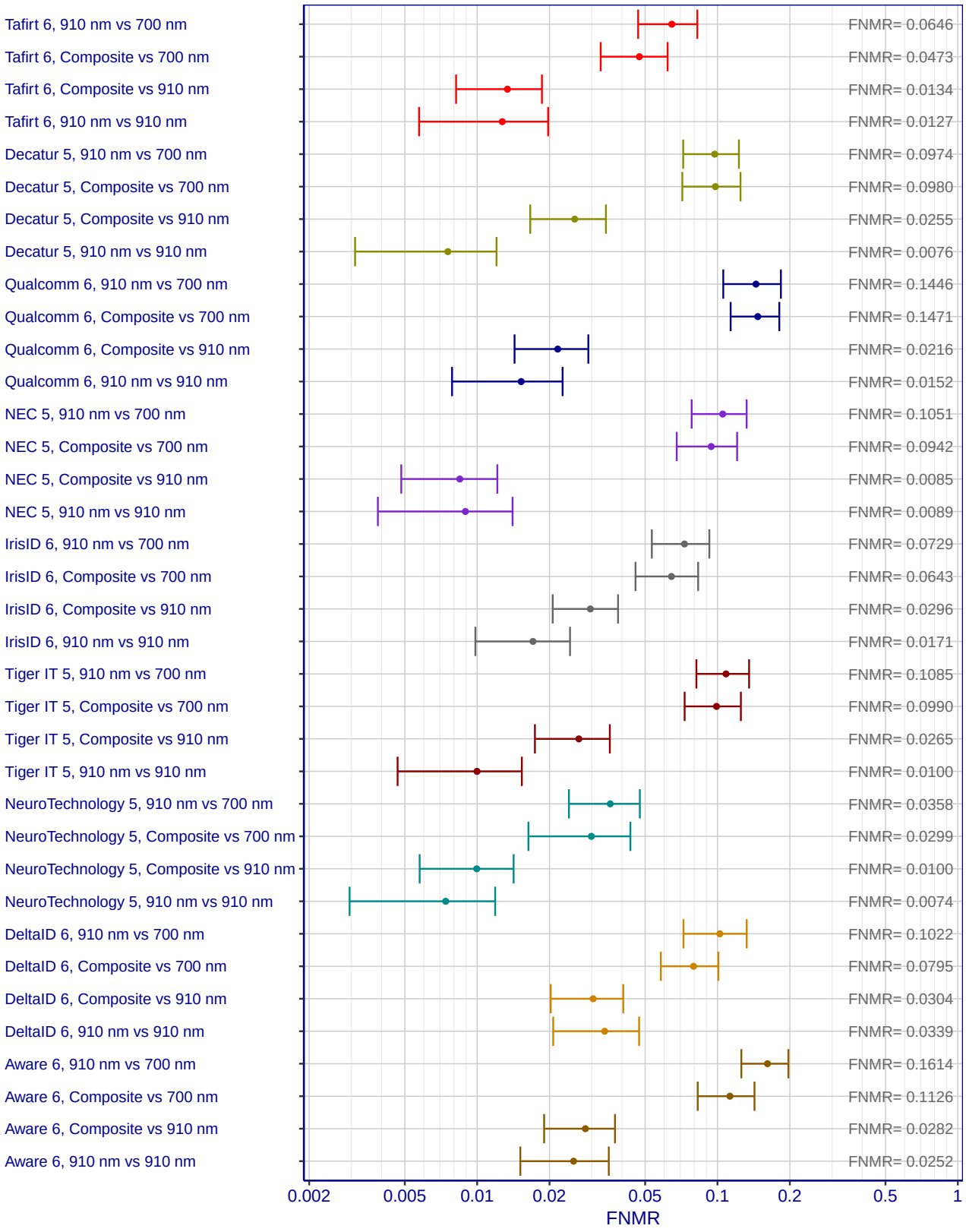


Figure 3.16: *FNMR (at FMR=10^{-4}) for selected matchers when composite images are used for matching. Results are for single-eye comparisons. Each point was generated using about 80 thousand mated comparisons.*

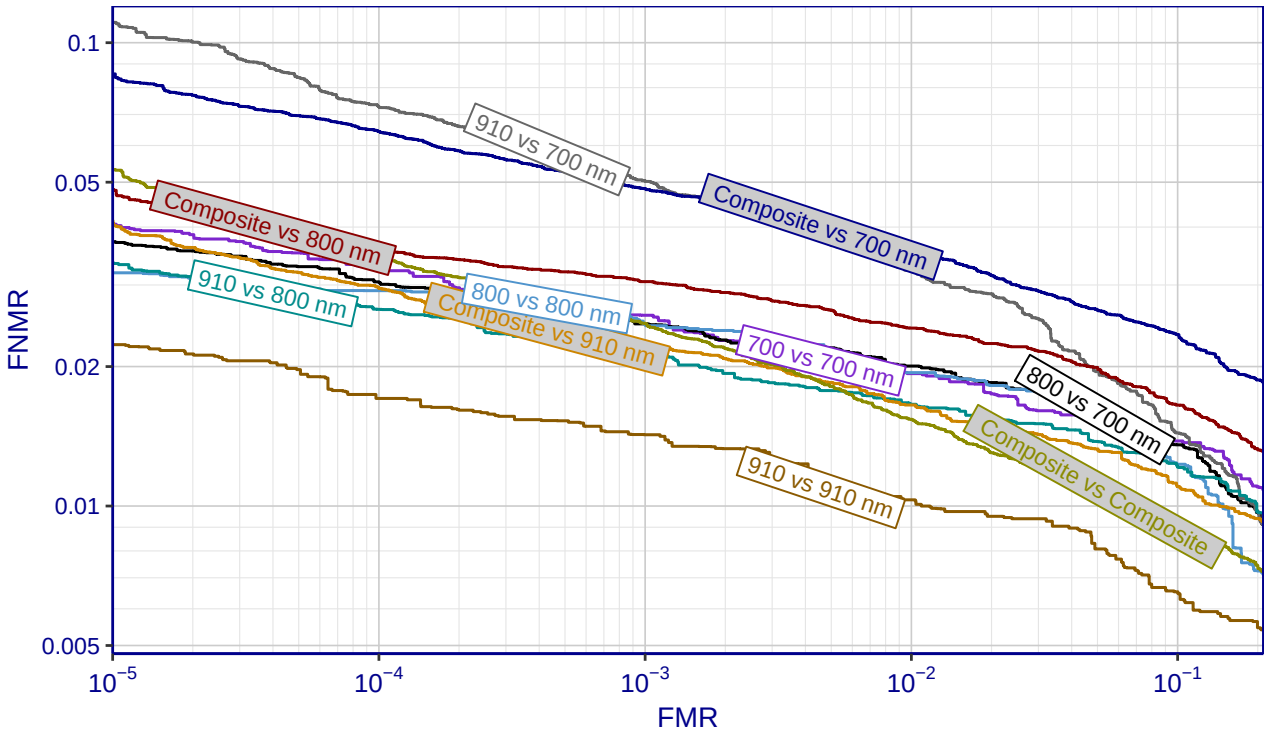


Figure 3.17: *Full DET plots for comparisons involving composite images for IrisID 6. Results show single-eye matching results. Each DET curve is generated from about 80 thousand mated scores and 20 million nonmated scores.*

3.5 Forensic Iris

Forensic iris is the application of iris recognition to civil and criminal law, including criminal investigation. Forensic iris is likely to involve comparing two iris images to determine if they represent the same person (inclusionary evidence) or to establish that they represent different people (exclusionary evidence). The comparisons may be performed by trained human examiners, automated matching algorithms, or some combination of the two.

Forensic iris is a nascent field. A decade ago leaders in the industry, including John Daugman [44], did not consider iris recognition viable for forensic applications. Although their observations were accurate at the time, advances in iris recognition technology, large-scale collection of iris data, and a growing corpus of research, have superseded those observations. There is now legitimate interest from government agencies in utilizing iris for forensic applications.

This section of the report explores the limits of automated iris recognition over "in the wild" iris images. "In the wild" refers to images of the iris acquired from sources not originally intended for iris recognition or images of the iris not acquired using traditional iris capture equipment. An example would be an image of a person's iris extracted from a high resolution photograph from the internet. Matey *et. al.* [45] recently demonstrated that iris recognition can operate reliably over images pulled from the internet. This report limits its investigation to automated matching and does not address other forensic topics such as manual iris recognition or interpretation of matching results in a court of law.

3.5.1 Visible Wavelength Matching

Nearly all forensic applications of iris recognition are expected to operate over iris samples acquired at visible wavelengths. Section 3.1 already demonstrated that operating within the VW band leads to a significant hit in accuracy. NeuroTechnology's matcher performed better than the other matchers over VW iris images. As Figure 3.18 demonstrates, FNMRs fall within the 0.04 to 0.06 range for two-eye matching, significantly worse than for NIR matching where FNMR is below 0.001 at any reasonable decision threshold.

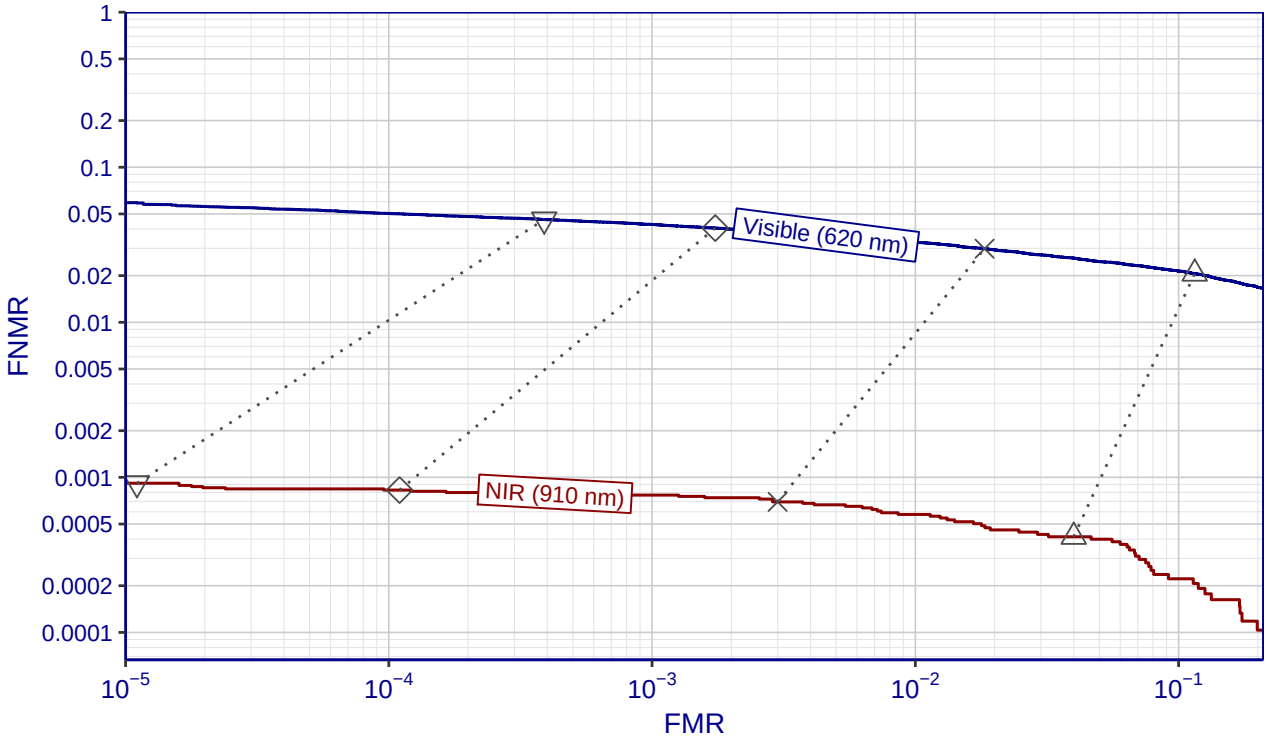
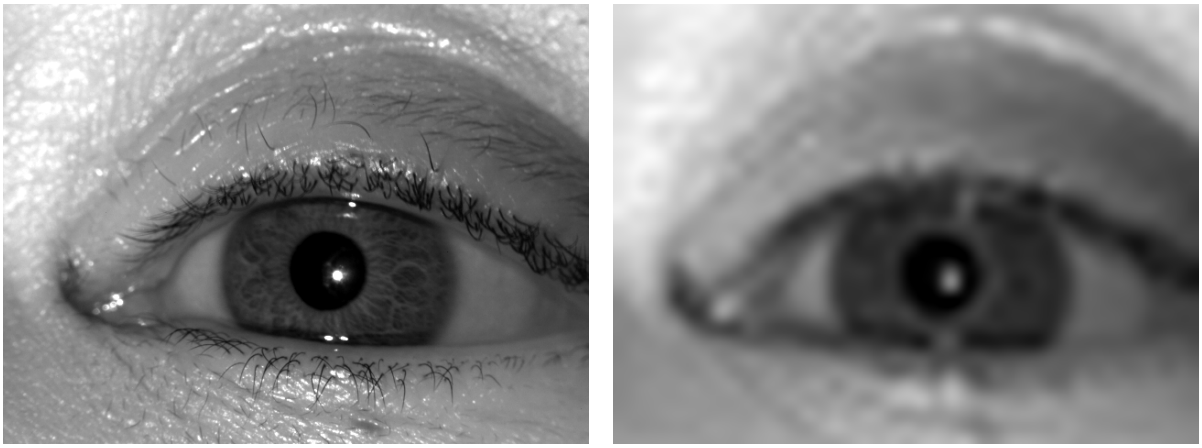


Figure 3.18: *DET plots when comparing VW iris images (blue) and NIR iris images (red) using NeuroTechnology 5. Results are for two-eye matching. Each grey line segment shows how a particular decision threshold produces different error rates depending on whether the matcher is comparing VW images or NIR images.*



(a) Original Iris Image **(b)** Low-resolution Iris Image

Figure 3.19: *Example of an image intentionally subjected to degradation to simulate a low-resolution capture. The first step involves decimating the image (i.e. reducing the sampling rate) so that the radius of the iris spans a specific number of pixels (12 in this example). Then, the image is upscaled back to its original pixel dimensions using bilinear interpolation. The interpolation used for upscaling leads to a blurry rather than pixelated appearance.*

3.5.2 Matching Low Resolution Iris Samples

Forensic iris is expected to sometimes involve images where the resolution of the iris is not well controlled. For optimal performance, ISO/IEC 19794-6:2011 [2] recommends a spatial sampling rate of no less than 10 pixels/mm and an MTF of no less than 0.6 at 2 cycles/mm; the earlier version of the standard [46] recommended 20 pixels/mm. Matey et al. [47, 48, 49] and Ackerman [50, 51] demonstrated that the Iris on the Move® applications could operate successfully at 10 pixels/mm. Typical commercial iris cameras such as the IrisAccess 4000 series [52] produce images with approximately 200 pixels across the nominally 10 mm wide iris in accord with the earlier recommendations of 20 pixels/mm. This section explores recognition at resolutions below the recommendations of the new standard. Consideration is further restricted to VW matching because it is expected to be the most common forensic use-case.

A sufficiently large dataset of low-resolution iris images was not available, so images from the CMID were decimated to simulate low-resolution captures. Decimation was performed with the *pamscale* command from the NetPGM package [53]. To minimize aliasing, the command filters out frequencies above half the new sampling rate before picking the new samples. The sampling rate is chosen so that the radius of the iris (i.e. the distance from the pupil center to the limbus boundary) in the decimated image spans a predetermined number of pixels. Figure 3.19 shows an example of an iris that was decimated by a factor of 15.7 to produce an image where the radius of the iris spans exactly 12 pixels ($\approx$1.2 pixels/mm). Upsampling was then applied to the decimated images to return them to their original pixel dimensions. Upsampling was performed using bilinear interpolation, which gives the images a blurry, rather than pixelated, appearance. The iris radius was determined using the manually marked boundary coordinates. These boundary coordinates were also provided to the matchers during template creation since boundary localization would otherwise be quite difficult in severely decimated images.

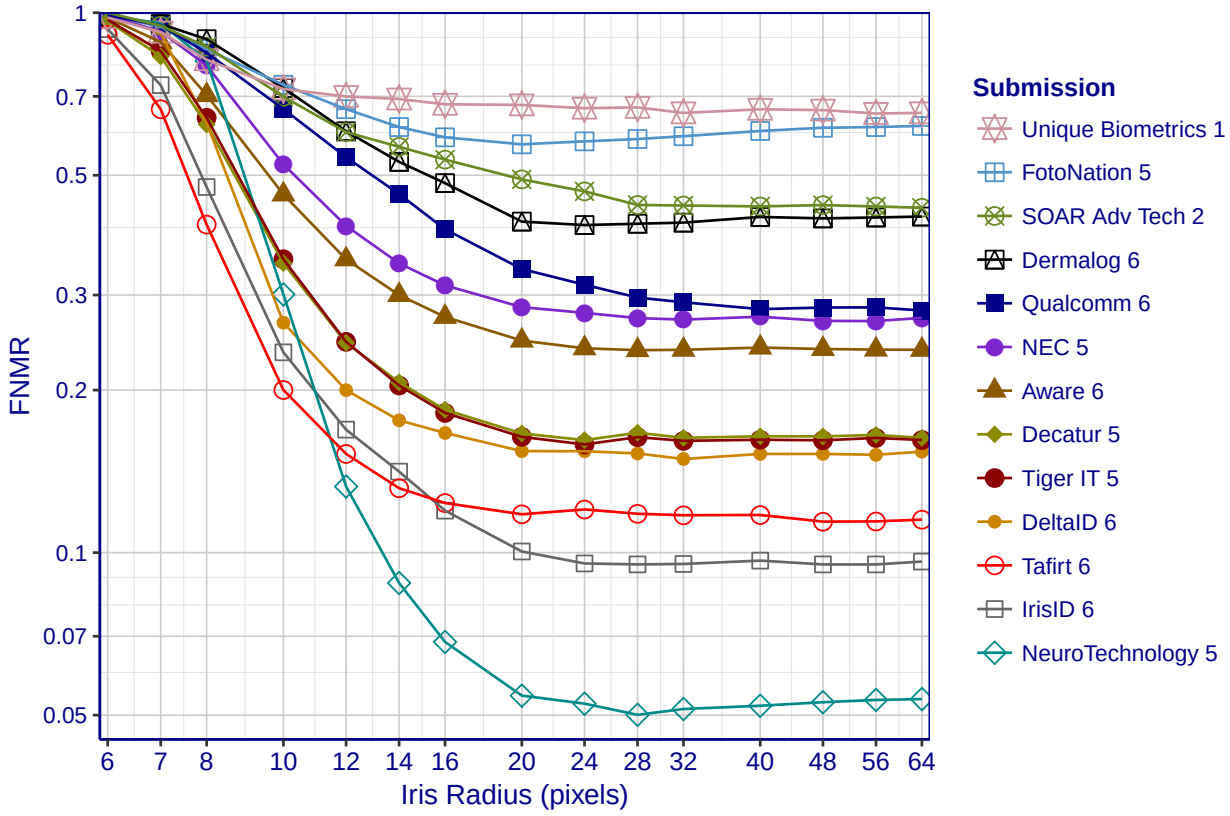


Figure 3.20: *FNMR as a function of the radius of the iris in pixels for various iris matchers. Only enrollment images were decimated (verification images were not decimated). The decision threshold is fixed to produce an FMR of* 10^{-4} *when no decimation is applied. Two-eye matching results are presented for samples acquired at 620 nm. Each point is produced from about 650 thousand mated comparisons. Note the sharp upturn in FNMR once the radius drops below 20 pixels and consider it in context with the internet image example discussed in the text.*

Figure 3.20 plots FNMR as a function of iris resolution for various matchers. Only enrollment images were decimated. Verification images were left at their original resolutions, which tended to be around 120 pixels (≈12 pixels/mm). The figure shows results for two-eye matching when all samples were acquired at 620 nm (which corresponds to orange-red in the VW band). FNMR appears to remain stable for most matchers until the radius is reduced to about 20 pixels (≈2 pixels/mm). Further reductions in resolution lead to exponential increases in FNMR. NeuroTechnology 5 produces the lowest FNMRs at higher resolutions (iris radii ≥ 12 pixels). At lower resolutions it is surpassed by both Tafirt 6 and IrisID 6. These two matchers are still capable of correctly matching the iris more than half the time when the radius of the iris is only 8 pixels (≈0.8 pixels/mm). This result is surprising and auspicious but further research is recommended before drawing any solid conclusions. The CMID dataset consists of extremely high-quality samples collected in well controlled environments. It is possible that FNMR is sensitive to small deviations from optimal sample quality that could lead to different results operationally. For example, variations in pupil dilation[4], which are minor in the CMID, could have a significant impact on accuracy. More importantly, FNMR only presents one side of matching accuracy. Namely, the ability to recognize that two samples represent the same iris. The other problem, being able to distinguish that two samples represent *different* irises, is addressed next.

[4]Dilation and constriction are terms that describe the pupil diameter. Dilation is an increase in pupil diameter, normally a response to low light levels; constriction is a decrease in pupil diameter, normally a response to high light levels. Both dilation and constriction can be caused by drugs.

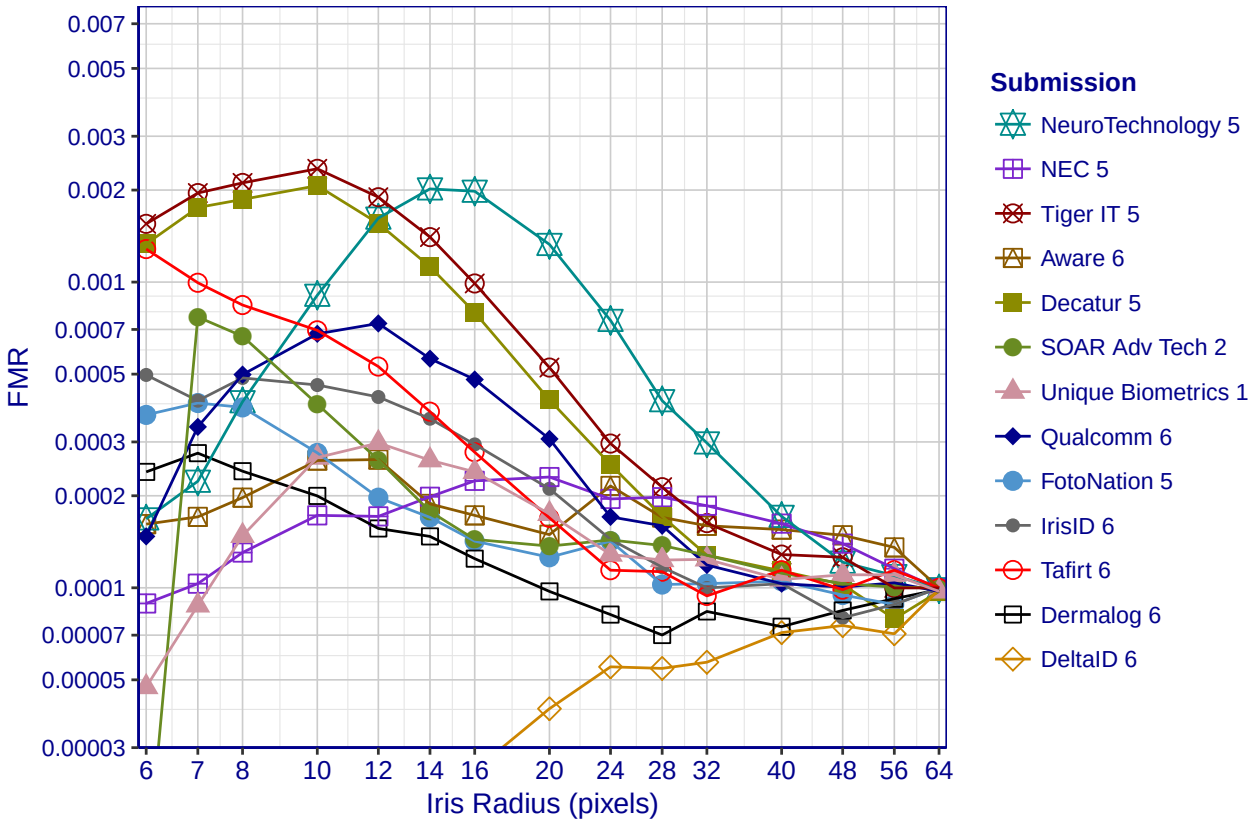


Figure 3.21: *FMR as a function of the radius of the iris in pixels for various iris matchers. Only enrollment images were decimated (verification images were not decimated). The decision threshold is fixed across all resolutions to produce an FMR of* 10^{-4} *when no decimation is applied. Two-eye matching results are presented for samples acquired at 620 nm. Each point is produced using about 1.5 million nonmated comparisons.*

Figure 3.21 plots FMR as a function of iris resolution along the same lines and setup as Figure 3.20. However, as the current figure demonstrates, behavior is much less consistent across matchers for FMR than for FNMR. FMR peaks between iris radii of 10 and 16 pixels for several of the matchers. Tiger IT 5, Decatur 5, and NeuroTechnology 5 experience the greatest fluctuations in FMR. In the case of NeuroTechnology 5, FMR peaks when the radius of the iris is about 14 pixels (≈1.4 pixels/mm). At this resolution, FMR is roughly 20 times higher than when the radius of the iris is 64 pixels (≈6.4 pixels/mm). For many of the matchers, FMR does not appear to level out as the iris radius increases, even when it reaches 64 pixels. This contrasts with FNMR, which appears to vary little for the matchers between radii of 24 pixels (≈2.4 pixels/mm) and 64 pixels (≈6.4 pixels/mm). In summary, it appears that less severe decimation tends to detrimentally impact the ability of matchers to distinguish between samples representing *different* irises, while more severe decimation detrimentally impacts the ability of matchers to recognize that two samples represent the *same* iris. Furthermore, some matchers might perform better if the iris samples were acquired at resolutions higher than the current ISO/IEC 19794-6 standard of 10 pixels/mm.

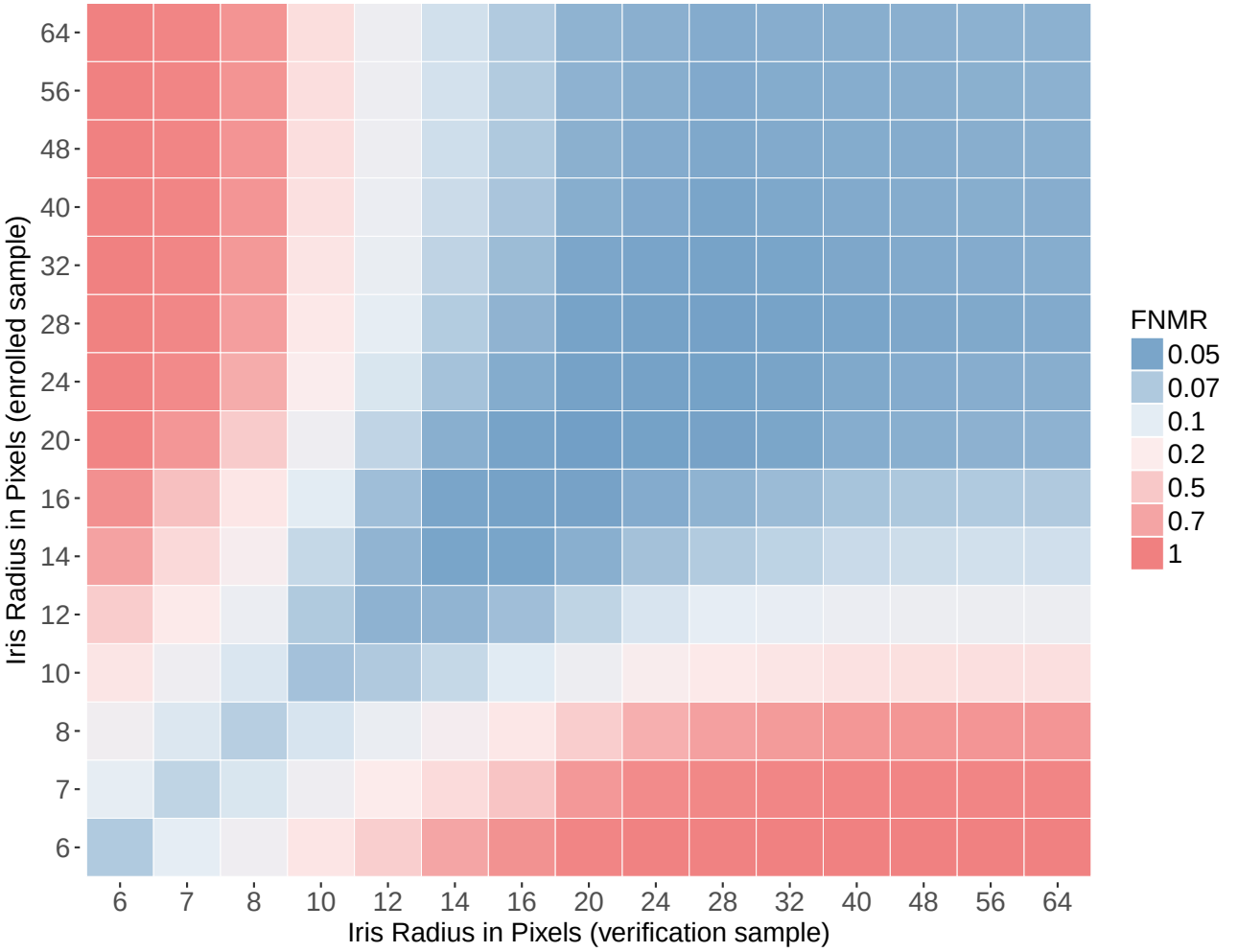


Figure 3.22: *FNMR as a function of the iris radius (in pixels) of both the verification and enrollment samples for NeuroTechnology 5. The decision threshold is fixed across all cells to produce an FMR of 10^{-4} when no decimation is applied. Two-eye matching results are presented for samples acquired at 620 nm. Each point is produced from about 650 thousand mated comparisons.*

A forensic application that is likely to arise is the need to compare iris samples where both samples were acquired at resolutions below the minimum ISO/IEC 19794-6 recommendations. Figure 3.22 shows a heatmap where color indicates FNMR and the axes specify the iris radius (in pixels) for the verification and enrollment samples. The decision threshold is universally fixed to elicit an FMR of 10^{-4} when no decimation is applied to the images. Generally, FNMR is lowest (*i.e.* best) when the iris resolution is high (radius ≥ 20 pixels) for both the verification and enrollment samples. At the lower resolutions, FNMR tends to be lowest when both the verification and enrollment samples are decimated to similar resolutions. This may seem auspicious, as it indicates that the matcher still has the ability to recognize that two iris samples represent the same source even when their resolutions are extremely low. While true, further analysis reveals that this benefit is offset by the fact that the matcher loses the ability to distinguish between samples representing different irises. This is demonstrated in the next figure.

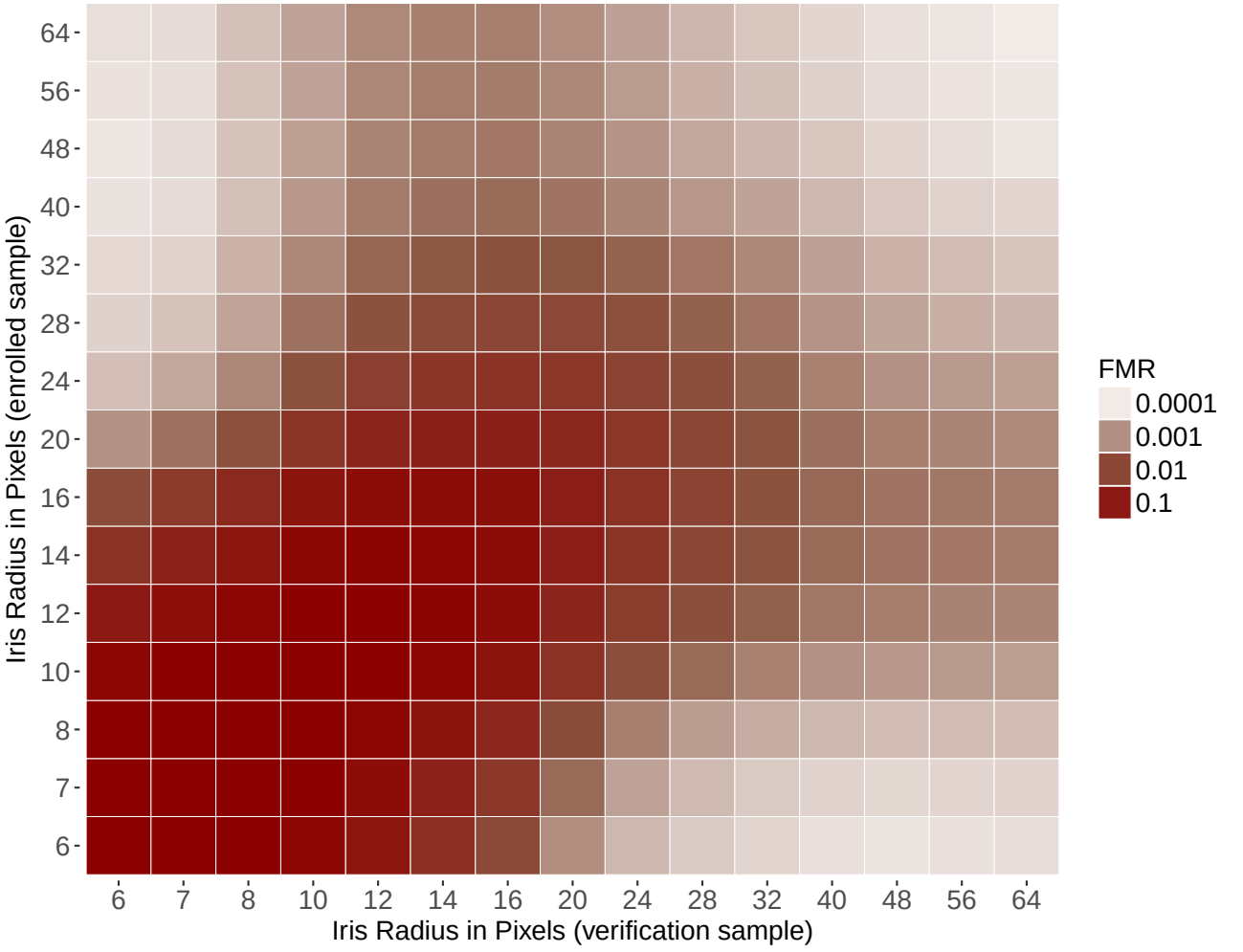


Figure 3.23: *FMR as a function of the iris radius (in pixels) of both the verification and enrollment samples for NeuroTechnology 5. The decision threshold is fixed across all cells to produce an FMR of 10^{-4} when no decimation is applied. Two-eye matching results are presented for samples at 620 nm. Each point is produced from about 1.5 million nonmated comparisons.*

Figure 3.23 shows a heatmap where color indicates FMR and the resolution and the axes specify the iris radius (in pixels) for verification and enrollment samples. The figure is identical to Figure 3.22 except color indicates FMR rather than FNMR. FMR varies considerably depending on the resolution of the samples. FMR balloons to around 0.1 when the resolution of both samples is low (*i.e.* the radius of the irises is $\leq$ 12 pixels, or $\approx$1.2 pixels/mm). The previous figure shows that matchers still have the ability to recognize that two samples represent the same iris even at extremely low resolutions. However, the current figure demonstrates that they lose the ability to recognize when two samples represent different irises. Figure 3.23 also reveals the need for threshold calibration based on the resolution of the compared samples to keep FMR low.

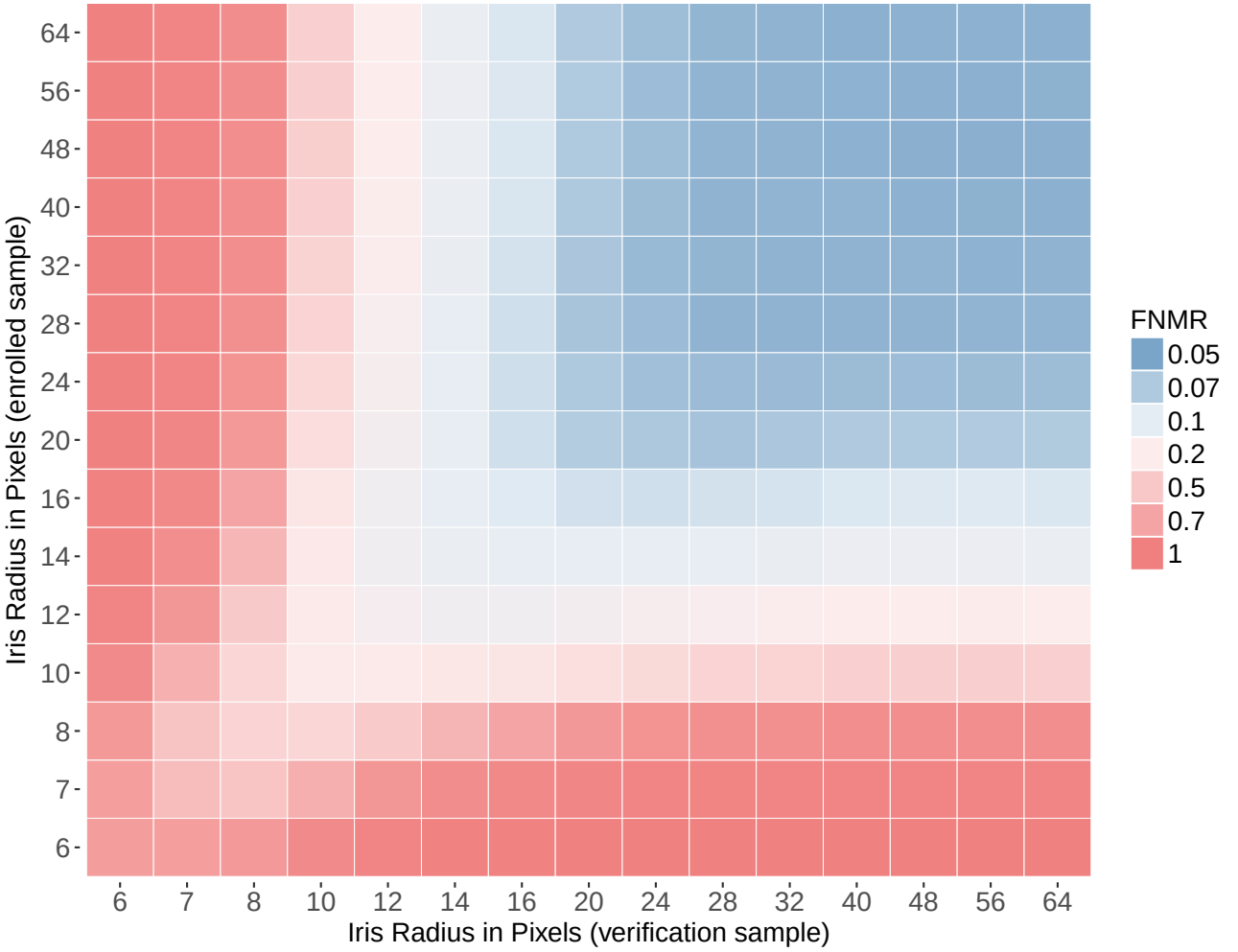


Figure 3.24: *Heatmap of FNMR at fixed FMR where the axes specify the iris radius (in pixels) of the verification and enrollment samples. Two-eye matching results at 620 nm are presented for NeuroTechnology 5. FNMRs are computed at FMR* $= 10^{-4}$*. Unlike Figure 3.22, where the same decision threshold is used to compute each FNMR, the decision threshold is adjusted in each cell to elicit an FMR of* 10^{-4}*. Each FNMR is calculated using about 650 thousand mated comparisons and each FMR is computed using 1.5 million nonmated comparisons.*

Biometric recognition accuracy for one-to-one matchers is characterized by the trade-off between FNMR and FMR as a decision threshold is adjusted [54]. For this reason, showing the effect of resolution on just FNMR (or FMR) at a fixed decision threshold is an incomplete representation of accuracy. Figure 3.24 attempts to address this limitation by plotting FNMR at fixed FMR for different iris resolutions. Unlike Figure 3.22, where the same decision threshold is used in each cell, the decision threshold is adjusted to ensure a fixed FMR of 10^{-4}. Accuracy drops precipitously when the radius of either the verification or enrollment samples dips below 20 pixels and continues to drop sharply as the resolution is further reduced. Generally speaking, it appears that FNMR (at FMR=10^{-4}) can be approximated reasonably well using the minimum of the resolution of the verification and enrollment samples. Thus, if the verification and enrollment samples have different iris resolutions, accuracy appears to be dependent upon the lower resolution of the two.

4 References

[1] G. W. Quinn, P. Grother, and J. Matey, "IREX IX Part One: Performance of Iris Recognition Algorithms." https://nvlpubs.nist.gov/nistpubs/ir/2018/NIST.IR.8207.pdf, 2018. 1, 7

[2] ISO/IEC, "ISO/IEC ISO-19794-6:2011, information technology — biometric data interchange formats — part 6: Iris image data," tech. rep., ISO/IEC, 2011. 1, 12, 28

[3] ISO 29794-6:2015, *Information technology – Biometric sample quality – Part 6: Iris image data*. ISO, Geneva, Switzerland, 2015. 1, 12

[4] J. Daugman, "Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1927–1935, 2006. 2, 13

[5] H. T. Ngo, R. W. Ives, J. R. Matey, J. Dormo, M. Rhoads, and D. Choi, "Design and implementation of a multispectral iris capture system," in *Proc. Conf Signals, Systems and Computers Record of the Forty-Third Asilomar Conf*, pp. 380–384, 2009. 2, 13, 17

[6] R. W. Ives, H. T. Ngo, S. D. Winchell, and J. R. Matey, "Preliminary evaluation of multispectral iris imagery," in *Proc. IET Conference of Image Processing (IPR 2012)*, pp. 1–5, July 2012. 2, 13

[7] P. Grother, E. Tabassi, G. W. Quinn, and W. Salamon, "Performance of Iris Recognition Algorithms on Standard Images." https://www.nist.gov/itl/iad/image-group/irex-i, 2009. 7

[8] *ISO/IEC 19794-6 - Biometric Data Interchange Formats - Iris Image Data*. 2011. 7

[9] *ANSI/NIST-ITL 1-2011 Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information*. 2011. 7

[10] E. Tabassi, P. Grother, and W. Salamon, "IREX - IQCE Performance of Iris Image Quality Assessment Algorithms." https://www.nist.gov/itl/iad/image-group/irex-ii-iqce, 2011. 7

[11] *ISO/IEC 29794-6 - Biometric Sample Quality Standard- Part 6: Iris Image*. 2012. 7

[12] P. Grother, G. Quinn, J. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson, "IREX III: Performance of Iris Identification Algorithms." https://www.nist.gov/itl/iad/image-group/irex-iii-homepage, 2011. 7

[13] G. Quinn and P. Grother, "IREX III Supplement I: Failure Analysis." https://www.nist.gov/itl/iad/image-group/irex-iii-homepage, 2011. 7

[14] G. W. Quinn, P. Grother, and M. Ngan, "IREX IV Part 1: Evaluation of Iris Identification Algorithms." https://www.nist.gov/publications/irex-iv-part-1-evaluation-iris-identification-algorithms, 2014. 7, 38

[15] G. W. Quinn, P. Grother, M. Ngan, and N. Rymer, "IREX IV: Part 2 Compression Profiles for Iris Image Compression." https://www.nist.gov/publications/irex-iv-part-2-compression-profiles-iris-image-compression, 2014. 7, 38

[16] G. W. Quinn, J. Matey, E. Tabassi, and P. Grother, "IREX V: Guidance for Iris Image Collection." https://www.nist.gov/itl/iad/image-group/irex-v-homepage, 2014. 7

[17] P. Grother, J. R. Matey, E. Tabassi, G. Quinn, and M. Chumakov, "IREX VI, temporal stability of iris recognition accuracy," tech. rep., NIST, July 2013. Interagency Report 7948. 7

[18] P. Grother, J. R. Matey, and G. W. Quinn, "Irex vi: Mixed-effects longitudinal models for iris ageing: Response to bowyer and ortiz," *IET Biometrics*, vol. 4, pp. 200–205, December 2015. 7

[19] K. Browning and N. Orlans, "Biometric Aging Effects of Aging on Iris Recognition," tech. rep., MITRE Corporation, 2014. 7

[20] H. Mehrotra, M. Vatsa, R. Singh, and B. Majhi, "Does Iris Change Over Time?," *PLoS One*, vol. 8, no. 11, 2013. 7

[21] "Minutiae Interoperability Exchange (MINEX) II." https://www.nist.gov/itl/iad/image-group/minutiae-interoperability-exchange-minex-iii. Accessed: 2018-07-09. 7

[22] P. Grother, M. Ngan, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) Part 1: Verification." https://www.nist.gov/programs-projects/face-recognition-vendor-test-frvt-ongoing, 2018. 7

[23] S. Brown, "Standardized Technology Evaluation Process (STEP) User's Guide and Methodology for Evaluation Teams." http://www2.mitre.org/work/sepo/toolkits/STEP. Accessed: 2018-07-27. 8

[24] D. Etter, J. Webb, and J. Howard, "Collecting Large Biometric Datasets: A Case Study in Applying Software Best Practices," *CrossTalk: The Immutable Laws of Software Development*, pp. 4–8, 2014. 8

[25] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, pp. 1148–1161, Nov. 1993. 8

[26] G. Quinn, P. Grother, and J. Matey, "Multi-Spectral Iris Evaluation Concept, Evaluation Plan, and API Specification Version 1.3." https://www.nist.gov/sites/default/files/documents/2017/02/23/irex9_conops.pdf, Sept. 2017. 10

[27] J. Daugman, "How iris recognition works," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, pp. 21–30, Jan 2004. 10

[28] K. E. Iverson, *A programming language*. New York, NY, USA: John Wiley & Sons, Inc., 1962. 11

[29] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, pp. 1895–1898, 1997. 11

[30] C. E. Metz, "Ce: Basic principles of roc analysis," in *Seminars in Nuclear Medicine*, pp. 8–283, 1978. 11

[31] J. L. Wayman, "Confidence Interval and Test Size Estimation for Biometric Data," in *IEEE Proc. AutoID*, pp. 177–184, 1999. 11, 39

[32] A. Ross, R. Pasula, and L. Hornak, "Exploring multispectral iris recognition beyond 900nm," in *Biometrics: Theory, Applications, and Systems, 2009. BTAS'09. IEEE 3rd International Conference on*, pp. 1–8, IEEE, 2009. 17

[33] C. Boyce, A. Ross, M. Monaco, L. Hornak, and X. Li, "Multispectral iris analysis: A preliminary study51," in *Computer Vision and Pattern Recognition Workshop, 2006. CVPRW'06. Conference on*, pp. 51–51, IEEE, 2006. 17

[34] C. K. Boyce, *Multispectral iris recognition analysis: Techniques and evaluation*. PhD thesis, West Virginia University, 2006. 17

[35] N. Kollias, "The spectroscopy of human melanin pigmentation," *Melanin: Its Role in Human Photoprotection*, vol. 38, pp. 31–38, 1995. 17

[36] U. Bubeck, "Multibiometric authentication-an overview of recent developments-term project cs 574 spring 2003 san diego state university," 18

[37] M. A. M. Abdullah, J. A. Chambers, W. L. Woo, and S. S. Dlay, "Iris biometrie: Is the near-infrared spectrum always the best?," in *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, pp. 816–819, Nov 2015. 18

[38] C. Boyce, *Multispectral Iris Recognition Analysis: Techniques and Evaluation*. West Virginia University Libraries, 2006. 18

[39] Y. Gong, D. Zhang, P. Shi, and J. Yan, "Optimal wavelength band clustering for multispectral iris recognition," *Appl. Opt.*, vol. 51, pp. 4275–4284, Jul 2012. 18

[40] P. J. Hube, "Neyman-Pearson biometric score fusion as an extension of the sum rule," in *Biometric Technology for Human Identification IV*, SPIE, 2007. 18

[41] ISO Working Group 1, "Standing Document 2 Harmonized Biometric Vocabulary," tech. rep., ISO/IEC JTC1 SC37 N1248, November 2005. 19

[42] and, "Ix. on the problem of the most efficient tests of statistical hypotheses," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 231, no. 694-706, pp. 289–337, 1933. 19

[43] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000. 24

[44] T. Thompson and S. Black, *Forensic human identification: An introduction*. CRC Press, 2006. 27

[45] J. R. Matey, G. W. Quinn, and P. Grother, "Iris Forensics: A Review," *Technical Note*, 2018. 27

[46] ISO/IEC, "ISO/IEC ISO-19794-6:2005, information technology — biometric data interchange formats — part 6: Iris image data," tech. rep., ISO/IEC, 2005. 28

[47] J. Matey, "Iris recognition," *Sarnoff Corporation, BCC*, 2005. 28

[48] J. Matey, D. Ackerman, J. Bergen, and M. Tinker, "Iris recognition in less constrained environments," *Advances in Biometrics*, pp. 107–131, 2008. 28

[49] J. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. LoIacono, S. Mangru, M. Tinker, T. Zappia, and W. Zhao, "Iris on the move: Acquisition of images for iris recognition in less constrained environments," *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1936–1947, 2006. 28

[50] D. Ackerman, "Spatial resolution as an iris quality metric," in *Biometrics Consortium Conference Tampa, Florida September*, vol. 28, p. 3, 2011. 28

[51] D. A. Ackerman, "Optics of iris imaging systems," in *Handbook of Iris Recognition*, pp. 367–393, Springer, 2013. 28

[52] "irisaccess 4000." 28

[53] "NetPBM: Extended portable bitmap toolkit." ftp://ftp.x.org/contrib/utilities. Accessed: 2018-09-10. 28

[54] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," tech. rep., DTIC Document, 1997. 33

[55] "Information Technology - Vocabulary, Part 37: Biometrics," standard, International Organization for Standardization, Geneva, CH, Feb. 2017. 37

[56] G. E. P. Box, J. S. Hunter, and W. G. Hunter, *Statistics for experimenters : design, innovation, and discovery.* Wiley-Interscience, 2005. 37

[57] G. Upton and I. Cook, *A Dictionary of Statistics*. Oxford University Press, 2 ed., 2008. 38

[58] L. D. Brown, T. T. Cai, and A. Dasgupta, "Interval Estimation for a Binomial Proportion," *Statistical Science*, vol. 16, pp. 101–133, 2001. 38, 39

[59] "Biometric Ideal Test: CASIA-IrisV4." http://biometrics.idealtest.org/dbDetailForUser.do?id=4. 38

[60] K. W. Bowyer and P. J. Flynn, "The ND-IRIS-0405 iris image dataset," *CoRR*, vol. abs/1606.04853, 2016. 38

[61] A. J. Mansfield and J. L. Wayman, "Best Practices in Testing and Reporting Performance of Biometric Devices," tech. rep., National Physical Laboratory. 39

[62] P. Bickel, "Response to SAG Problem 97-23." University of California, Berkeley, Department of Statistics. 39

[63] M. E. Schuckers, *Computational Methods in Biometric Authentication*. Information Science and Statistics„ London :: Springer London :, 2010. 39

[64] J. C. Wu, "Studies of Operational Measurement of ROC Curve on Large Fingerprint Data Sets Using Two-Sample Bootstrap." https://www.nist.gov/publications/studies-operational-measurement-roc-curve-large-fingerprint-data-sets-using-two-sample, 2007.
39

[65] G. Oehlert, *A first course in design and analysis of experiments*. W.H. Freeman, 2010. 39

A Uncertainty Estimation

This appendix describes how estimates of variability are computed in this report. Estimates of variability do not directly describe the population. Rather, they convey information about the primary statistics that are used to make inferences about the population. The primary statistics in this report are the core accuracy metrics defined in Section 2.4. Variability refers to how tightly these statistics represent the true population parameters.

The core accuracy metrics are computed over a sample of data subjects[1] selected from a larger population. In our case, we regard the larger population as the set of all adults in the United States. In truth, the sample was collected from volunteers at two geographically separated collection sites, which may not perfectly represent a random sample from the overall adult population. The iris images collected from the data subjects are paired in various ways to form comparison sets. These pairings introduce a correlation structure that must be incorporated into the estimates of variability.

The correlation structure for one-to-one comparisons is characterized by the positive correlations between comparisons. For example, two comparisons are expected to be positively correlated if they share a reference template in common. Table A.1 defines eight distinct types of dependency for single-eye mated comparison. The final column shows the strength of each type of dependency as the mean Pearson Correlation Coefficient (PCC) across all submissions for samples acquired at 800 nm. The correlations are measured with respect to the decisions made at an FMR of 10^{-4}. Although the correlation values are threshold dependent, they tend to change little for varying FMR due to the relative "flatness" of iris DET curves.

Correlation Type	Same Verification Session	Same Verification Samples	Same Enrollment Session	Same Enrollment Samples	Correlation at FMR $= 10^{-4}$
1	Yes	Yes	Yes	No	0.5 ± 0.3
2	Yes	Yes	No	No	0.2978 ± 0.2
3	Yes	No	Yes	Yes	0.5173 ± 0.3
4	Yes	No	Yes	No	0.3402 ± 0.3
5	Yes	No	No	No	0.1592 ± 0.2
6	No	No	Yes	Yes	0.3223 ± 0.2
7	No	No	Yes	No	0.15965 ± 0.2
8	No	No	No	No	0.148 ± 0.2

Table A.1: *A basic correlation structure for dual-eye mated comparisons. The rows describe different types of dependencies that can exist between comparisons. The values in the final column are mean correlation coefficients across all submissions (along with their standard deviations) when the images are acquired at 800 nm.*

The first type of dependency refers to comparisons that share their reference samples in common. The third dependency type refers to comparisons that share their verification samples in common. The second dependency type refers to comparisons where the verification samples are not identical but were acquired during the same capture session (recall that for the CMID, six images of each of the right and left eyes are acquired during a single session). The eighth type of dependency refers to comparisons that share neither a verification nor an enrollment session in common. Interestingly, the correlation is still positive. Furthermore, the correlation tends to be much stronger when the samples are acquired at wavelengths outside the normal 700 - 900 nm range. At a wavelength of 620 nm, the correlation coefficient is 0.5 ± 0.1. At a wavelength of 1070 nm, it is $X \pm X$. This value likely reflects the degree to which accuracy is dependent upon the person-specific features of the iris.

The mated comparison sets for CMID were constructed using all possible mated pairings where the reference sample was acquired first chronologically. Comparisons between samples acquired during the same capture session were also excluded. Given our strategy for set construction, Table A.1 fully captures the significant sources of dependency for dual-eye mated comparisons. For single-eye mated comparisons, we account for twelve different types of dependency. Identifying the appropriate correlation structure for a given comparison set can be a daunting task. Straightforward *factorial design* [56] may be the best approach in most cases. Certain approaches to constructing the comparison sets also lead to simplified correlation structures.

General equations are presented for estimating the variability of the core accuracy metrics given arbitrary correlation structures. Formally, let $\hat{p}(\tau)$ be the estimate of either FMR or FNMR as computed in Equations 2.1 and 2.2. Let $d_i(\tau)$ be the decision at threshold τ for the ith comparison ($i = 1, ..., N$). If the comparison is mated, then $d_i(\tau) = [m_i \leq \tau]$. From this point onward, the input, τ, will be omitted from all equations for clarity of presentation. Furthermore, let S_k be the set of comparison pairs fulfilling the criteria for dependency type k ($k = 1, ..., K$). The elements of S_k are pairs of indices where a given

[1] Terminology such as "data subject" is now used by NIST to conform with ISO/IEC 2382-3: Vocabulary, Part 37: Biometrics [55]

element (i, j) refers to comparisons i and j respectively. An unbiased estimate of the covariance for the kth dependency type is

$$\hat{\sigma}_k^2 = \frac{N}{N-1}\frac{1}{|S_k|}\sum_{(i,j)\in S_k}(d_i - \hat{p})(d_j - \hat{p}). \tag{A.1}$$

The first term is Bessel's Correction [57]. The rest of the equation is just the common estimate of covariance. The estimate of variance is

$$\hat{\sigma}^2 = \frac{\hat{p}(1-\hat{p})}{N} + \hat{c} \tag{A.2}$$

where

$$\hat{c} = \frac{1}{N^2}\sum_{k=1}^{K}|S_k|\hat{\sigma}_k^2. \tag{A.3}$$

Equation A.3 consolidates the contribution of all of the covariances to the overall estimate of the variance.

Confidence intervals can be constructed using estimates of both p and the σ^2. The simplest and most straightforward approach is to invoke the Central Limit Theorem (CLT) and define the interval as

$$\hat{p} \pm z\sqrt{\hat{\sigma}^2} \tag{A.4}$$

where z is the $(1-\alpha/2)$th quantile of the standard normal distribution and α is the desired significance level. (Typically $\alpha = 0.1$ or $\alpha = 0.05$ corresponding to confidence levels of 90% and 95% respectively.) However, Brown et. al. [58] identify several problems with this approach. First, they note that always defining $\hat{p}$ as the center of the interval can introduce a systematic negative bias to the coverage probability. Second, the actual distribution of $\hat{p}$ is significantly nonnormal when p is close to 0 or 1, even for large N. Finally, due to the fact that $\hat{p}$ is a discretized estimator, Equation A.4 severely underestimates the true coverage probability for certain "unlucky pairs" of p and N. For these reasons, we adopt their recommendation to use the Wilson Score method. However, the method must be modified to account for the correlation structure.

The Wilson Score interval is formed by inverting the normal approximation to the equal-tailed hypothesis test of $H_0 : p = p_0$. The hypothesis is accepted if $\hat{p}$ falls within the interval

$$\frac{|\hat{p} - p_0|}{\sqrt{\frac{1}{N}p_0(1-p_0) + \hat{c}}} \leq \pm z. \tag{A.5}$$

The denominator is the standard deviation of the test statistic. Unlike Equation A.4, it does not require a full estimate of the variance. The additional $\hat{c}$ term incorporates the contribution of the correlation structure to the estimate. Since knowing p_0 does not reveal the true value of c, the latter must be approximated using Equation A.3. The Wilson Interval is derived by regarding p_0 as the unknown parameter. Using the quadratic equation to solve Equation A.5 for p_0 yields the interval:

$$CI_W = \frac{\hat{p} + \frac{1}{2N}z^2 \pm z\sqrt{\frac{1}{N}\hat{p}(1-\hat{p}) + \frac{1}{4N^2}z^2 + (1+\frac{1}{N}z^2)\hat{c}}}{1+\frac{1}{N}z^2}. \tag{A.6}$$

The Wilson Score interval still loses accuracy when np or $n(p-1)$ is small (though not as severely as Equation A.4) . For this reason, we conservatively opt not to apply the Wilson Score Interval to cases where $np < 10^2$.

Previous NIST evaluations [14, 15] used the Wilson Score Method under the assumption that all comparisons are independent (effectively utilizing Equation A.6 under the assumption $\hat{c} = 0$). Failing to account for the dependencies probably led to overly optimistic estimates of variability. In the current evaluation, we found that the independence assumption leads to significant underestimates of variance, sometimes by a factor of 3 or more. Many academic iris datasets (*e.g.* CASIA [59], Notre Dame 0405 [60]) consist of iris samples collected from comparatively small numbers of subjects, typically a few hundred at most. Thus, the dependencies are expected to contribute considerably toward the variability of accuracy statistics computed over these public datasets.

Sometimes we report estimates of variability for FNMR at fixed FMR when in fact the decision threshold is fixed. Uncertainty with respect to what decision threshold corresponds to the targeted FMR results in increased uncertainty about the true value of FNMR. That said, our estimates of FMR are expected to be very tight given the large number of nonmated comparisons performed (often in excess of a billion). Additionally, even at very low FMRs, the lightly sloping nature of iris DET curves means that small discrepancies in FMR are not expected to significantly impact FNMR.

A.1 Discussion

Some previous literature exists on the topic of estimating uncertainty for biometric accuracy statistics. Mansfield *et. al.* [61] provided estimates of variability for FNMR and FMR given a particular sampling strategy. The sampling strategy assumes full cross comparisons and only accounts for a single source of dependency. Bickel [62] defined equations for estimating confidence bounds under similar restrictions, which were later tested by Wayman [31] over fingerprint data and found to be accurate. Schuckers [63] expanded upon their work by defining a general correlation structure for fingerprint recognition. Although his proposed method of estimating confidence intervals is common and asymptotically valid, it still suffers from the same weaknesses identified by Brown *et. al.* [58] in relation to Equation A.4. Bootstrapping [64] also fails to offer a viable alternative because it assumes independent and identically distributed (*iid*) comparison scores and thus ignores all sources of dependency. Altering the resampling strategy can sometimes compensate for one or two types of dependency [61], but typically no more than that.

The previous section provides a framework for computing unbiased estimates of variability. The downside of this approach compared to other methods is the added difficulty of having to identify the correlation structure. One approach is to conceptualize the problem as a *factorial experiment* [65] and construct the correlation structure as a design matrix.