FINAL REPORT

IRIS RECOGNITION STUDY 2006 (IRIS06)

AKA

STANDARDS-BASED PERFORMANCE AND USER COOPERATION STUDIES OF COMMERCIAL IRIS RECOGNITION PRODUCTS

1 September 2007
Version 1.0

Exceptional Service in Society's Interest

4405 E Baseline Road, Suite 118
Phoenix, AZ 85042
Phone: (480) 889-6400
Fax: (480) 889-6401

Abstract

Authenti-Corp performed a standards-based evaluation of online, offline, and off-axis performance for three commercially-available iris recognition products. A 300-person live test subject population with demographics representative of the US population was employed for online testing. Multiple sets of ISO/IEC 19794-6 compliant iris images were collected from each test subject with each product spanning time intervals from fifteen minutes to about six weeks. The iris images collected online were evaluated offline in verification mode using template generation and matching algorithms similar to those used by the commercial products (provided by Professor John Daugman, University of Cambridge). To explore user-cooperation factors, off-axis gaze experiments were performed with the larger test population, and off-axis pose experiments were performed with a small six-person test population using a specialized test apparatus. In all, about 29,000 iris images were collected, and millions of comparisons were performed. Evaluations were conducted using the ANSI INCITS 358-2002, BioAPI 1.1 biometric application programming interface and in accordance with the ISO/IEC 19795 standard for biometric performance testing. Online and offline performance metrics, such as true and false match rates, generalized true and false accept rates, and enrollment and recognition transaction times, along with the associated confidence intervals, are reported for the three products evaluated.

Results indicate that iris recognition performance improves when multiple attempts are allowed. Three-attempt Failure to Enroll rates as low as 0.35% were measured, and True Match Rates as high as 99.7% were measured. Three-attempt Generalized True Accept Rates (which include the influence of Failures to Enroll, Failures to Acquire, and Failures to Match) as high as 97.8% were measured. Three-attempt mean Recognition Transaction Times as short as 7.9 seconds were also measured. In addition, time separation between enrollment and recognition attempts (up to six weeks) did not have a measurable influence on iris recognition performance, and left and right eyes exhibited statistically similar performance.

The products evaluated demonstrated tradeoffs between speed and accuracy. Higher accuracy requires longer transaction times, and faster transaction times result in lower accuracy. Results also indicate that eyeglasses can degrade iris recognition performance and that interoperability matching performance is sometimes better than native matching performance.

Off-axis experiments indicate that the evaluated products perform well with yaw (rotate head as if saying “No”) and roll (ear-to-shoulder) angles of ±20° or more and perform better when test subjects face upwards relative to the camera rather than downwards.

While several important areas for further study are identified, the results of this evaluation indicate that the current crop of commercial iris recognition products can recognize cooperative and uncooperative individuals rapidly, reliably, and interchangeably in a variety of criminal justice and border control applications.

Executive Summary

Authenti-Corp performed a standards-based study to evaluate the online, offline, and off-axis performance of three commercially-available iris recognition products (Products A, B, and C). A 300-person live test subject population with demographics representative of the US population was employed for online testing. Multiple sets of standards-compliant iris images (ISO/IEC 19794-6) were collected from each test subject with each product spanning time intervals from fifteen minutes to about six weeks. The iris images collected online were evaluated offline in verification mode using template generation and matching algorithms similar to those used by the commercial products. To explore user-cooperation factors, off-axis gaze experiments were performed with the larger test population, and off-axis pose experiments were performed with a small six-person test population using a specialized test apparatus. In all, about 29,000 iris images were collected, and millions of comparisons were performed. Evaluations were conducted using a standardized biometric application programming interface (ANSI INCITS 358-2002, BioAPI 1.1) and in accordance with ISO standards for biometric performance testing (ISO/IEC 19795). Online and offline performance metrics, such as true and false match rates, generalized true and false accept rates, and enrollment and recognition transaction times, along with the associated confidence intervals, are reported for the three products tested.

Significant results of the study are summarized below:

- Cumulative Failure to Enroll (FTE) rates, Failure to Acquire (FTA) rates, False Non-Match Rates (FNMR), and Generalized False Rejection Rates (GFRR) generally decrease with increasing numbers of attempts. Similarly, True Match Rates (TMR) and Generalize True Accept Rates (GTAR) generally increase with increasing numbers of attempts. This performance enhancement is possibly due to improved effectiveness of the camera-human interface with increasing human practice and the removal of eyeglasses. In general, enrollment and recognition performance improve when multiple attempts are allowed.
- All products exhibit roughly the same three-attempt FTE rate for the left-or-right eye feature set. In this case, an FTE is declared when three attempts are allowed for each eye and neither eye successfully enrolls, which is a realistic operational enrollment policy. For the three products evaluated, this FTE rate varied from 0.35% to 3.39% with mean enrollment

Three-Attempt Enrollment Metrics for Left-or-Right Eye Feature Set

	Product A	Product B	Product C
FTE (%)	0.35	0.68	3.39
Mean Enrollment Transaction Time (sec)	40.4	32.2	70.1
See Section 6 for confidence intervals			

transaction times varying from 32.3 to 70.1 sec. Although the FTE rates varied by an order of magnitude, Figure 6-1 shows that the confidence intervals for each value overlap indicating that the differences are not statistically significant.

- Left and right eyes generally exhibit roughly the same FTE, FTA, FNMR/TMR, GFRR/GTAR, and mean Recognition Transaction Times (RTT) for each of the products evaluated. Overall, right and left eyes exhibit statistically similar iris recognition performance.
- Time separation between enrollment and recognition attempts (up to six weeks) does not have a measurable influence on FTA, FNMR, and GFRR and mean RTT.
- Product C appears to be influenced by ambient lighting conditions.
- FNMR/TMR is similar for all products and all iris-feature sets (during real-time online operation and in offline comparisons using the algorithms provided by Professor John Daugman, University of Cambridge). Results indicate the following general trends for three-attempt recognition metrics:
  - $FTA_A < FTA_B \sim FTA_C$ (online),
  - $GFRR_A < GFRR_B \sim GFRR_C$ (online),
  - $GTAR_A > GTAR_B \sim GTAR_C$ (offline), and
  - mean $RTT_B < RTT_C < RTT_A$ (online).

  These trends are illustrated in the table below for the left-or-right-eye feature set.

Overall Three-Attempt Recognition Metrics for Left-or-Right Eye Feature Set

	Product A	Product B	Product C
FTA (online) (%)	1.5	6.9	6.9
FNMR=1-TMR (online) (%)	0.0	1.8	0.4
TMR (offline@HD=0.32) (%)	99.7	97.3	99.4
GFRR=1-GTAR (online) (%)	1.9	9.3	10.7
GTAR (offline@HD=0.32) (%)	97.8	89.9	89.0
Mean RTT (online) (sec)	21.4	7.9	11.2
See Section 6 for confidence intervals			

- The products tested demonstrate tradeoffs between speed and accuracy. Higher accuracy requires longer transaction times, and faster transaction times result in lower accuracy. For example, Product A exhibits the best recognition performance but also has the longest

average transaction time. The "best" product depends on the specific needs of a particular operational scenario.

- Unless the exact instantiation of an iris recognition product's algorithms are used for offline testing, offline performance results do not necessarily indicate the online (real-world) performance of the product.
- In some cases, enrollment and recognition images from different cameras provide better matching performance than enrollment and recognition images from the same camera. That is, interoperability matching performance is better than native matching performance. Specifically, the best matching performance is obtained using enrollment images from Product C and recognition samples from Product A.
- Eyeglasses can degrade iris recognition performance. For Products B and C, matching performance is degraded for iris images acquired from test subjects wearing eyeglasses. Product A matching performance is similar both with and without glasses.
- Product A collects the highest percentage of high quality iris images compared to Products B and C but also exhibits the longest mean transaction time.
- Product B demonstrates a significantly larger collection volume than Products A and C and exhibits the shortest mean three-attempt recognition transaction time (7.9 seconds). As such, Product B is most appropriate for use with uncooperative users.
- The three iris recognition products evaluated perform better when test subjects gaze upward (with neutral pose) or face upward (with neutral gaze) relative to the camera rather than downward.
- The evaluated products generally performed well with yaw (rotate head as if saying "No") and roll (ear-to-shoulder) angles of ±20° or more when the test subjects were located at manufacturer-designated distances from the camera.

Several areas deserving further study using the collected iris-image database are identified, including the influence of iris image quality parameters on recognition performance, intra-individual correlation factors and the Doddington's Zoo phenomenon, the practicality of various methods to generate impostor match score distributions, and the evaluation of identification performance (as opposed to verification performance), which may be one of the greatest strengths of iris recognition technology.

While iris recognition technology continues to mature and improve, the results of the study indicate that the current crop of commercial iris recognition products can recognize cooperative and uncooperative individuals rapidly, reliably, and interchangeably in a variety of criminal justice and border control applications. The "best" product, of course, depends of the specific needs of a particular operational scenario.

Table of Contents

1. Introduction

Biometric products are used for automated recognition of individuals based on their behavioral and biological characteristics. Iris recognition biometric products recognize individuals based on their iris images, more specifically the distinctive patterns in the irises created by various structures, such as crypts, furrows, frills, ridges, ligaments, freckles, coronas, and collarettes (Figure 1-1). Other common biometric products use fingerprint features, facial images, hand geometry, characteristics of handwritten signatures, and voice recordings to recognize individuals. The US-VISIT program currently uses fingerprint recognition and digital facial photographs at ports of entry.[2] Fingerprint recognition, known for its low error rates, typically requires an individual to place their finger on a sensor to be recognized. The error rates for facial recognition technologies are typically higher than for fingerprint technologies, but facial recognition is often preferred because its operation is non-contact. Iris recognition combines the advantages of fingerprinting (low error rates) and facial recognition (non-contact operation) and as such may prove valuable for many criminal justice and border control applications.[3]

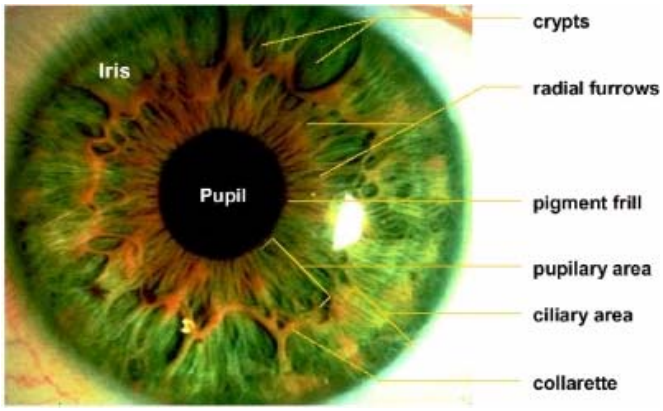


Figure 1-1. Structures of the Iris[1]

1.1 Iris Recognition Overview

Most commercial iris recognition products identify individuals using high-resolution images of iris patterns captured in the near-infrared (NIR) portion of the optical spectrum. Images of the iris are photographed with NIR-sensitive cameras located several inches to several feet in front of the eye. In the NIR, the visible "color" of the iris is not observed, and a monochromatic grayscale representation of the iris is used (Figure 1-2).

[1] Image from http://www.rdecom.army.mil/rdemagazine/200305/index.htm (accessed 1 September 2007).
[2] http://www.dhs.gov/xtrvlsec/programs/content_multi_image_0006.shtm (accessed 1 September 2007), http://www.dhs.gov/xtrvlsec/programs/editorial_0525.shtm (accessed 1 September 2007).
[3] Funding and oversight for this study were provided by the US Department of Justice, National Institute of Justice (NIJ) and the US Department of Homeland Security, Transportation Security Administration (TSA) under Award No. 2005-IJ-CX-K066.

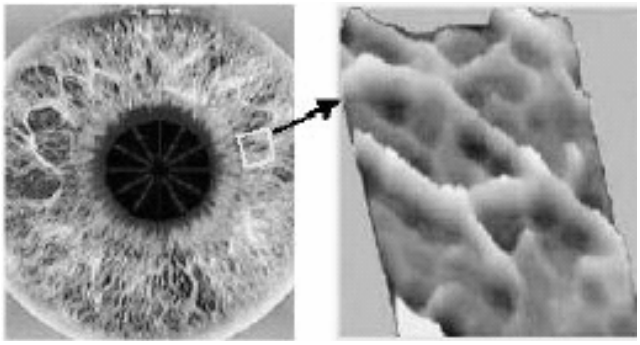


Figure 1-2. Grayscale Image and Texture of Iris[4]

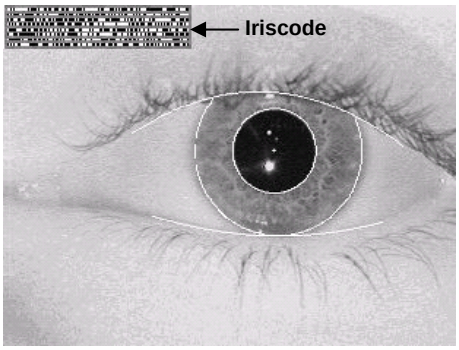


Figure 1-3. Segmented Iris[5]

Two basic biometric processes are performed using the grayscale iris images: ***enrollment*** and ***recognition***. During the enrollment process, a product's camera captures a high-resolution, NIR photograph of one or both of an individual's eyes. Using the iris portion of this photograph, the product creates a template that represents that individual's iris(es). The resulting enrollment template is then saved in a reference database. During the recognition process, which in real life occurs at some later time, often days, months, or even years later, the product takes another photograph of the individual's eye(s), creates a new template, and compares the new template with the enrollment template stored in the database. If the similarity, or match score, between the new template and the stored template is better than a specified discrimination level, typically known as the threshold, the individual is ***recognized***. If the similarity between the new template and the stored template is poorer than the discrimination level, the individual is ***not recognized***.

The basic operation of an iris recognition system is summarized in Figure 1-4. Light from an NIR illumination source, such as NIR light emitting diodes (LEDs) or a flashlamp, is reflected off the individual's iris, and an NIR image of the iris is collected with a camera. The iris image is then transferred to a computer. For most iris recognition systems, a segmentation algorithm then locates the iris and pupil and detects the eyelids. This segmentation process is illustrated by the white outlines around the iris in Figure 1-3. For iris recognition systems based on the algorithms developed by Professor John Daugman, University of Cambridge, the iris region is then remapped using a rectangular to polar transform (normalization), and converted into an "Iriscode" template. The Iriscode in Figure 1-3 is essentially a digital representation of the iris

[4] Figure from "Contribution a la verification biometrique de personnes par reconnaissance de l'iris," Christel -Loïc Tisse, Doctoral thesis, p. 24, 28 October 2003. Available at http://www.lirmm.fr/xml/en/0165-10.html (accessed 1 September 2007).

[5] Figure from "How Iris Recognition Works," John Daugman, PhD, OBE, http://www.CL.cam.ac.uk/users/jgd1000/csvt.pdf (accessed 1 September 2007). Used with permission.

texture in Figure 1-2 created by the iris features illustrated in Figure 1-1. More specifically, the Iriscode is an encoding of the random texture of the iris in terms of wavelet phase information spanning several scales of analysis.[6]

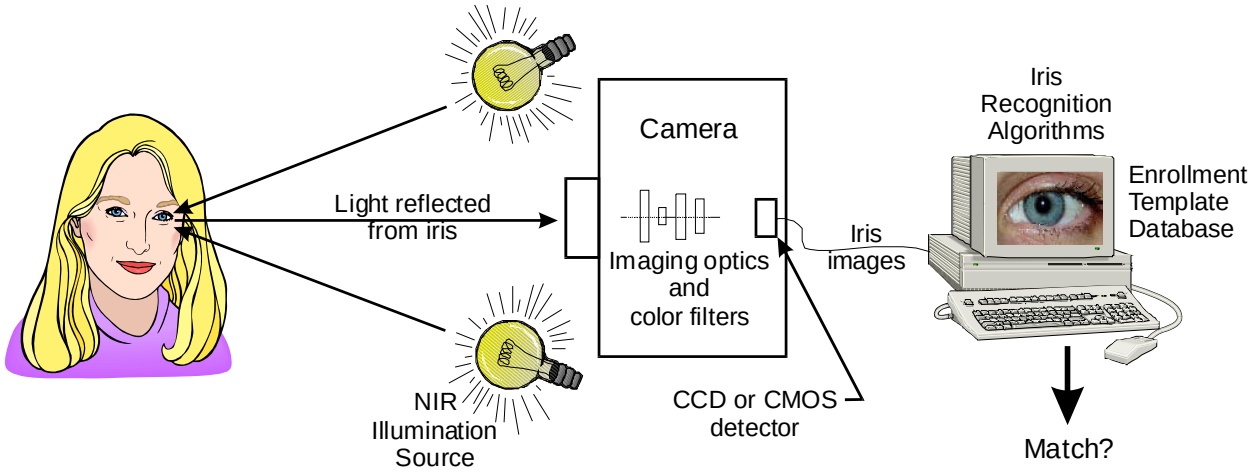


Figure 1-4. Iris Recognition Schematic

The recognition[7] process can be further subdivided into two categories: ***verification*** and ***identification***. For verification, the template created from the presented iris(es) is compared to the previously enrolled template for that individual. For identification, the template created from the presented iris(es) is compared to some or all of the previously enrolled templates in the reference database. An individual is recognized if the template created from the presented iris "matches" an enrolled template as determined by the iris recognition matching algorithm. For Daugman-based algorithms, a Hamming Distance (HD) score is computed when two Iriscode templates are compared. The HD score is lower when the templates are more similar and higher when the templates are less similar. As such, lower scores indicate better matching performance for Daugman-based algorithms. (For technical details on Iriscode generation and matching in Daugman-based iris recognition algorithms, see "How Iris Recognition Works," by Professor John Daugman[6] and Appendix 11.1.)

Regardless of the matching algorithm being used, iris recognition systems can encompass several different operational scenarios and challenges. For example, some iris recognition systems use one-eye cameras, and other systems use two-eye cameras. For one-eye cameras, the

[6] "How Iris Recognition Works," John Daugman, PhD, OBE, http://www.CL.cam.ac.uk/users/jgd1000/csvt.pdf (accessed 1 September 2007).
[7] We use the term "recognition" in this report to signify either a biometric verification or a biometric identification operation.

left and right eyes are presented to the camera separately (the camera optics collect only one iris image per user presentation). For two-eye systems, the left and right eyes are presented to the camera simultaneously (the camera optics collect both left and right iris images during one user presentation). If both left and right eyes are enrolled and both eyes are recognized, transactions with one-eye systems will likely require more time than transactions with two-eye systems.

In addition, different levels of user participation (predominantly active or predominantly passive) are required to interact with different systems. Most iris recognition systems require the user to look directly into the center of the camera (on-axis presentation) from within a zone known as the collection volume. The size and location of this volume is different for each camera, depending on the design specifications for that camera. In many cases, the user is responsible for placing themselves within this collection volume, which may be small or large. Often the camera will provide visual or auditory cues to help the user find the appropriate location. In some cases, the user must purposefully align their eye in the camera; mirrors are often employed to help the user provide an on-axis presentation (active user effort). In other cases, the user need only look in a specified direction once they are located within the collection volume (nominal user effort). For some systems, the user simply looks straight ahead and a trained operator aligns the camera (passive user effort) to obtain an on-axis presentation.

Further, the current generation of commercial iris recognition products is designed for operational scenarios where the eyes are placed in an optimal position relative to the product's camera to obtain ideal, on-axis eye alignment. Such on-axis alignment is possible with cooperative users. In some cases, however, it may be desirable to utilize iris recognition with uncooperative users, such as during the arrest of intoxicated or violent individuals. In this challenging operational scenario, non-ideal, off-axis alignment may be common. However, the performance of iris recognition technology in off-axis presentation conditions is not currently well characterized.

An additional operational challenge is the interoperability of iris images collected from different iris recognition products. While a level of interoperability between products has been demonstrated in a proprietary environment,[8] interoperability using ISO-standard iris interchange formats has not been demonstrated.

[8] http://www.biometricscatalog.org/itirt/itirt-FinalReport.pdf (accessed 1 September 2007).

1.2 IRIS06 Overview

To address the operational challenges described above, the Iris RecognItion Study 2006, or ***IRIS06***, studied native and interoperability performance of three commercially-available iris recognition products using ideal (on-axis) and non-ideal (off-axis) iris presentations. IRIS06 focused on the following primary questions:

- What are the realistic error rates and transaction times for various commercial iris recognition products?
- Can iris recognition products recognize individuals using ISO-standard iris images created by different iris recognition products? That is, are ISO-standard iris images interchangeable (interoperable) between products?
- What is the influence of off-axis user presentation parameters, such as linear position (X, Y, and Z) in the collection volume, head pose angle (pitch, roll, and yaw), and eye-gaze angle (up-and-down and side-to-side) on the ability of iris recognition products to acquire and recognize iris images?

Nomenclature:

- ***Native Performance***: recognition images from one product are compared to enrollment images from the same product
- ***Interoperability performance***: recognition images from one product are compared to enrollment images from a different product

The ultimate goal of IRIS06 is to predict the performance of iris recognition technology in various criminal justice, law enforcement, and border control biometric implementations, such as physical and logical access control, surveillance, and identification applications, for both cooperative and uncooperative individuals.

To this end, an open invitation to iris recognition product providers to participate in the IRIS06 study was issued in December 2005. Information about the test was posted on the IRIS06 website (www.authenti-corp.com/iris06), and the test protocol, product requirements, and test harness software were released to interested product providers. Ultimately, three commercial products were included in the IRIS06 study. To address concerns of ongoing litigation in the iris recognition domain, it was agreed that the products would not be identified. [9]

The IRIS06 study entailed two different concepts of operation: 1) a ***scenario evaluation*** using a relatively large human test population, and 2) a ***controlled off-axis experiment*** using a

[9] While the iris recognition product names are not provided in this report, the product providers may choose to publicly identify their product at their discretion.

specially-developed physical apparatus and small volunteer test population. Both studies were performed in a standard indoor office environment using a BioAPI[10] test interface.

For the scenario evaluation, on-axis iris images were collected from 264 live human test subjects over two visits separated by about six weeks. During the second visit, some off-axis images were also collected from each test subject. Online, on-axis performance was determined for the three commercial products, known as Product A, Product B, and Product C.[9] These performance results most closely emulate the performance expected for real-world deployments. All of the iris images acquired online were stored in ISO/IEC 19794-6[11] format for subsequent offline analysis, which was performed using iris recognition algorithms provided by Professor John Daugman, University of Cambridge.[12] Native (intra-product) and interoperability (inter-product) on-axis performance was determined during offline analysis. In addition, native off-axis performance was analyzed offline.

For the controlled off-axis experiment, off-axis iris images were collected from six volunteer test subjects (members of Authenti-Corp's data collection team) using a physical apparatus specifically designed to control and measure the relative linear position and the relative pitch, roll, and yaw angles between the test subject's iris and the iris recognition camera.

The scenario evaluation portion of the IRIS06 study was performed in conformance with the ISO/IEC 19795 standard for Biometric Performance Testing and Reporting to the extent possible given the constraints of the study. Specifically, care was taken in the report to not provide information that would allow the products tested to be identified – one of the constraints of the study.

1.3 Prior Related Work

Several similar types of scenario evaluations have been performed in the past. In 2001, the UK Communications-Electronics Security Group (CESG) Biometric Working Group published the "Biometric Product Testing Final Report"[13]. This study, led by Dr. Tony Mansfield of the UK National Physical Laboratory, evaluated the performance of seven biometric systems,

[10] ANSI INCITS 358-2002, Information technology - Biometrics Application Programming Interface (BioAPI) Specification (Version 1.1).
[11] ISO/IEC 19794-6:2005, Information technology - Biometric data interchange formats - Part 6: Iris image data
[12] www.cl.cam.ac.uk/~jgd1000/ (accessed 1 September 2007).
[13] http://www.cesg.gov.uk/site/ast/biometrics/media/BiometricTestReportpt1.pdf (accessed 1 September 2007).

including face, fingerprint, hand geometry, iris, vein, and voice recognition products. In 2002, the Department of Defense Counterdrug Technology Development Program Office sponsored the "Face Recognition at a Chokepoint" study, which was led by Duane Blackburn and Mike Bone.[14] In 2003, the German Bundesamt für Sicherheit in der Informationstechnik (BSI) performed a "Comparative Study of Facial Recognition Systems" referred to as BioFace I and II. In 2004, Authenti-Corp conducted a performance and interoperability scenario evaluation of fingerprint systems in a seafaring environment using a seafarer test population for the International Labour Organization (ILO).[15] The International Biometric Group (IBG), led by Michael Thieme, has performed several rounds of Comparative Biometric Testing.[16] In 2005, IBG performed Independent Testing of Iris Recognition Technology (ITIRT) using Iridian's proprietary templates, interfaces, and offline analysis tools.[8] Also in 2005, Authenti-Corp studied the performance and interoperability of several fingerprint sensors for a government client.[17]

Each of these studies provided valuable insight into the performance of biometric systems in fielded applications. However, none of these studies focused on the online and offline performance of iris recognition systems using standards-compliant interfaces and formats or on the performance of off-axis iris images. The IRIS06 effort utilized ANSI INCITS 358-2002 BioAPI 1.1 function calls and ISO/IEC 19794-6 data formats to study online and offline performance and interoperability of on-axis and off-axis images of live, human test subjects.

Protecting the rights and welfare of the IRIS06 test subjects is of the utmost importance. The measures implemented to ensure that test subjects' private, identifiable information remains confidential at all times are presented in the following section. The data collection processes and procedures for the scenario evaluation and off-axis experiment are presented in Sections 3 and 4, respectively. Data analysis techniques are described in Section 5, and the on-axis and off-axis results are presented in Section 6. Our conclusions are summarized in Section 7, and future research directions are described in Section 8. Implications of the IRIS06 results for knowledge and practice are addressed in Section 9.

[14] http://www.biometricscatalog.org/ApprovedDocuments/evaluation/ChokePoint-1.pdf (accessed 1 September 2007).
[15] http://www.ilo.org/public/english/dialogue/sector/sectors/mariti/sid.pdf (accessed 1 September 2007).
[16] http://www.biometricgroup.com (accessed 1 September 2007).
[17] Report not publicly releasable. Summary of results presented at Biometrics Conference (London, 2006), V. Valencia, "Biometric Sciences - The Good, the Bad, and the Ugly".

2. Test Subject Rights and Welfare

Substantial measures have been taken to protect the rights and welfare of the IRIS06 human test subjects. These measures, described below, include obtaining certification from an Institutional Review Board and implementing provisions to protect test subject privacy.

2.1 Institutional Review Board Certification

The US National Institute of Justice (NIJ), the programmatic sponsor for the IRIS06 study, requires that all NIJ-sponsored studies that involve human test subjects be conducted in compliance with the Code of Federal Regulations 28 CFR 22[18] (Confidentiality of identifiable research and statistical information) and with the US Code 42 USC 3789g[19] (Confidentiality of information). NIJ further requires that all studies utilizing live human test subjects be reviewed and certified by an Institutional Review Board (IRB) before biometric data collection can begin. (An IRB is a group of individuals who perform an independent review of research.)

To this end, we compiled, developed, and submitted over 20 required documents to the Western Institutional Review Board (WIRB).[20] The documentation included the NIJ grant solicitation, proposal, and award paperwork; a WIRB-mandated supplemental test protocol; test subject recruiting materials including a telephone screener script, a frequently asked questions handout, and a confirmation letter; two informed consent forms, one for the scenario evaluation and one for the controlled off-axis experiment; a test subject exit questionnaire; principal investigator credentials; the WIRB application; Authenti-Corp's NIJ privacy certificate, federal-wide assurance documentation, human research subject protection training certificates for all key personnel; and other miscellaneous documents. The WIRB-mandated test protocol was a substantial document and included sections covering the purpose of the study and background criteria for test subject selection; detailed descriptions of the methods and procedures used, including research design, data analysis and statistics, and data storage and confidentiality methods; a risk-benefit analysis; and test subject identification, recruitment, consent, and procedure details.

[18] http://www.access.gpo.gov/nara/cfr/waisidx_04/28cfr22_04.html (accessed 1 September 2007).
[19] http://www4.law.cornell.edu/uscode/html/uscode42/usc_sec_42_00003789---g000-.html (accessed 1 September 2007).
[20] http://www.wirb.com/ (accessed 1 September 2007).

Authenti-Corp submitted the requisite documents to WIRB on 8 February 2006. WIRB made substantial changes to the test subject recruiting materials and informed consent forms, including alterations that reflected a medical research study as opposed to a biometrics evaluation. In addition, many of the changes 1) were technically incorrect, 2) were contrary to NIJ requirements for test subject privacy assurance, and 3) adversely influenced the professional appearance of the documents. Much of the confusion can be attributed to the fact that the WIRB members were not familiar with biometrics, with the personally-identifiable nature of biometric data, or with 28 CFR 22. We corrected the WIRB-altered documents and resubmitted them, along with an email from Cheryl Crawford Watson, NIJ's Human Subjects Protection Compliance Officer, stating that the WIRB alterations violated NIJ-required confidentiality regulations, including 42 USC 3789g and 28 CFR Part 22.22.

WIRB notified Authenti-Corp that the corrected documents were approved on 11 April 2006, clearing the way to recruit test subjects and start collecting biometric data. The end-to-end IRB certification process, including document preparation and two approval iterations, lasted about four months.

2.2 Privacy Protection Provisions

In addition to IRB certification, significant measures were taken to protect the personally-identifiable information of the IRIS06 test subjects. First, all Authenti-Corp employees underwent extensive human research subject protection training prior to becoming authorized members of the Authenti-Corp Biometrics Test Team. Test Team members were advised of and agreed in writing to comply with the regulations in 28 CFR Part 22 and with the procedures outlined in the WIRB protocol document to protect the privacy and confidentiality of personally-identifiable information.

To further protect test subject confidentiality, iris images are identified using a unique identifier number (UIN) instead of the name of the test subject. A UIN was assigned during each test subject's first visit and affixed to the test subject's informed consent form. Iris images and demographic data stored in the ***test database*** are associated only with this number. The only link between data in the test database and test subject names are the ***informed consent forms*** and an ***Excel worksheet*** used for scheduling appointments. The Excel worksheet is stored in a

password-protected zip file and processed on a dedicated standalone computer that is never connected to any network, including the Internet. The test database resides on a standalone computer system that is separate from the computer used to process the Excel worksheet. The test database is never connected to any network outside of the test lab, including the Internet. The ***informed consent forms***, ***Excel worksheet***, and ***test database*** are physically protected at all times, either controlled by authorized members of Authenti-Corp's Biometrics Test Team or secured in a locked room that is accessible only to authorized members of the Authenti-Corp Biometrics Test Team.

All test subject personally-identifiable data, which includes the written informed consent form, the Excel worksheet used for scheduling, and the database information, will be securely destroyed on or before 20 years following completion of the IRIS06 study. The 20-year data storage period allows for future testing of iris recognition products to measure product performance enhancements and to study the influence of extended periods of time on the performance of iris recognition products. Test subjects may be contacted for additional data collection visits throughout the 20-year data storage period if they indicated permission to do so on their informed consent form.

All IRIS06 test subject personally-identifiable data is highly protected at all times. The data collected during this study is used only for statistical purposes. The statistical results of this study are published in this document and presented at meetings and conferences, however test subject names are never associated with their iris images, and test subject identities will never be disclosed. Iris images collected during the IRIS06 study will be provided to the NIJ for the purpose of data archiving upon project completion. As the iris images are proprietary, they can be shared only with US government agencies that 1) agree to abide by the confidentiality and data protection measures found in 28 CFR Part 22 and 2) process a privacy certificate with the NIJ.

The detailed data storage and confidentiality mechanisms utilized for the IRIS06 scenario evaluation and controlled off-axis experiment are described in the proprietary WIRB protocol document, which can be provided to US Government employees upon request.

3. Scenario Evaluation Data Collection

The purpose of the scenario data collection was two-fold: 1) to measure online (real-time operation) performance, and 2) to collect images for subsequent offline analysis. The scenario evaluation data were collected from 295 live human test subjects during a first visit. Of those test subjects, 264 returned for a second visit where additional data were collected. The logistics of recruiting the test subjects and collecting the data is presented in the following sections. The data collection test protocol, and the test harness software and hardware used to collect the data, are also described below. Details for the controlled off-axis data collection are covered in Section 4.

3.1 Test Subject Logistics

The demographic profile of the IRIS06 test subjects is presented in the following sections, along with our methodology for recruiting test subjects and the test subject visit statistics.

3.1.1. Recruitment

The human test subjects for the scenario evaluation were recruited from the Phoenix, Arizona metropolitan area. A local recruiting agency was hired to recruit 300 persons using the IRB-approved telephone screener and deliver them via scheduled appointments to Authenti-Corp's Biometrics Test Center.

Authenti-Corp provided the recruiting company with a tool for scheduling the first of two 30-minute appointments with each test subject and for documenting each test subject's demographic data. The scheduling and demographic information was securely transmitted to Authenti-Corp on a daily basis so that the test subject database could be populated prior to the scheduled appointments for that day. The recruiting agency scheduled the first visit for each test subject, sent reminder cards, and made reminder phone calls the day before the scheduled appointment. Authenti-Corp personnel scheduled each test subject's second visit (during their first visit), sent second-visit reminder cards, and made reminder phone calls the day before the second scheduled appointment.

After scheduling the first appointment over the phone, the recruiting agency mailed the IRB-approved Frequently Asked Questions (FAQs) handout, confirmation letter, and informed consent form to each scheduled test subject.[21]

- The Informed Consent Form describes the privacy protection measures in place to protect the test subjects' personally-identifiable information, provides confidentiality certifications, describes the usage of the personal information and the amount of time it will be retained, indicates that participation is completely voluntary, and provides spaces to enter the test subject's demographic information and informed consent signature. The informed consent form also has directions to Authenti-Corp's Biometrics Test Center and a location for the UIN, which is assigned to each test subject after consent has been provided.
- The FAQs provide an overview of biometric technology, address the purpose of the study, and attempt to belay any fears that test subjects may have about participating in the study.
- The Confirmation Letter provides a short description of what to expect during the test subject's visits to Authenti-Corp and confirms the test subject's scheduled appointment time.

Test subjects were paid a stipend upon completion of each visit. The goal was to have 300 test subjects participate in the first visit and 250 test subjects complete both visits, taking into account the fact that some test subjects would drop out and not participate in the second visit. As an incentive to entice the test subjects to return for the important second visit, the stipend for the second visit was slightly higher than the stipend for the first visit.

3.1.2. Demographics and visit statistics

The recruiting agency was responsible for delivering recruits to Authenti-Corp according to the target demographic profile presented in Table 3-1. The target demographics were based on the 2000 US Census. The demographics of the test subject population that completed both visits are also presented in Table 3-1. As shown, about 74% of the scheduled test subjects showed up for and completed the first visit (a 26% no-show rate). The return rate for the second visit was 89% (11% no-show rate). Ultimately, 264 test subjects completed both visits, which achieved the goal of 250 return test subjects.

[21] These documents can be provided to US Government employees upon request.

Table 3-1. Test Subject Demographic Profile and Statistics								
	Visit 1				Visit 2			
	Target		Actual		Target		Actual	
# of Test Subjects	300		295		250		264	
# of Appointments	401		295		295		264	
No-Show Rate	26%				11%			
Collection Dates	5/1/06-6/10/06				6/12/06-7/15/06			
Male	147	49%	139	47%	123	49%	121	46%
Female	153	51%	156	53%	128	51%	143	54%
18-24	30	10%	29	10%	25	10%	22	8%
25-34	60	20%	54	18%	50	20%	45	17%
35-44	66	22%	69	23%	55	22%	64	24%
45-54	57	19%	58	20%	48	19%	52	20%
55-64	36	12%	38	13%	30	12%	36	14%
65-74	27	9%	24	8%	23	9%	24	9%
75+	24	8%	23	8%	20	8%	21	8%
White	225	75%	220	75%	188	75%	199	75%
Black/African-American	36	12%	37	13%	30	12%	33	13%
Asian	12	4%	10	3%	10	4%	7	3%
Native American	3	1%	4	1%	3	1%	3	1%
Others	24	8%	24	8%	20	8%	22	8%
Hispanic	39	13%	35	12%	33	13%	33	12.5%
Non-Hispanic	261	87%	260	88%	218	87%	231	87.5%

Figure 3-1 illustrates the number of days between the two visits. The average separation between the first and second visits was 38 days. For about 28% of the test subjects, the separation was 42 days, or 6 weeks. The demographic statistics for the test subjects that completed both visits are illustrated graphically in Figure 3-2.

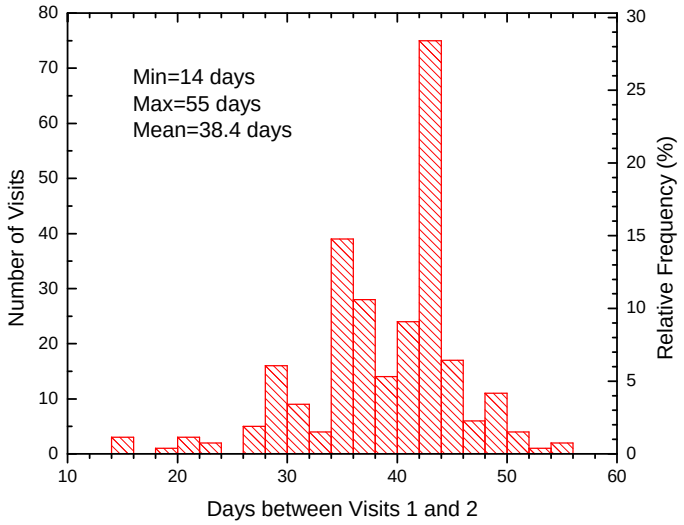


Figure 3-1. Number of Days between Visits

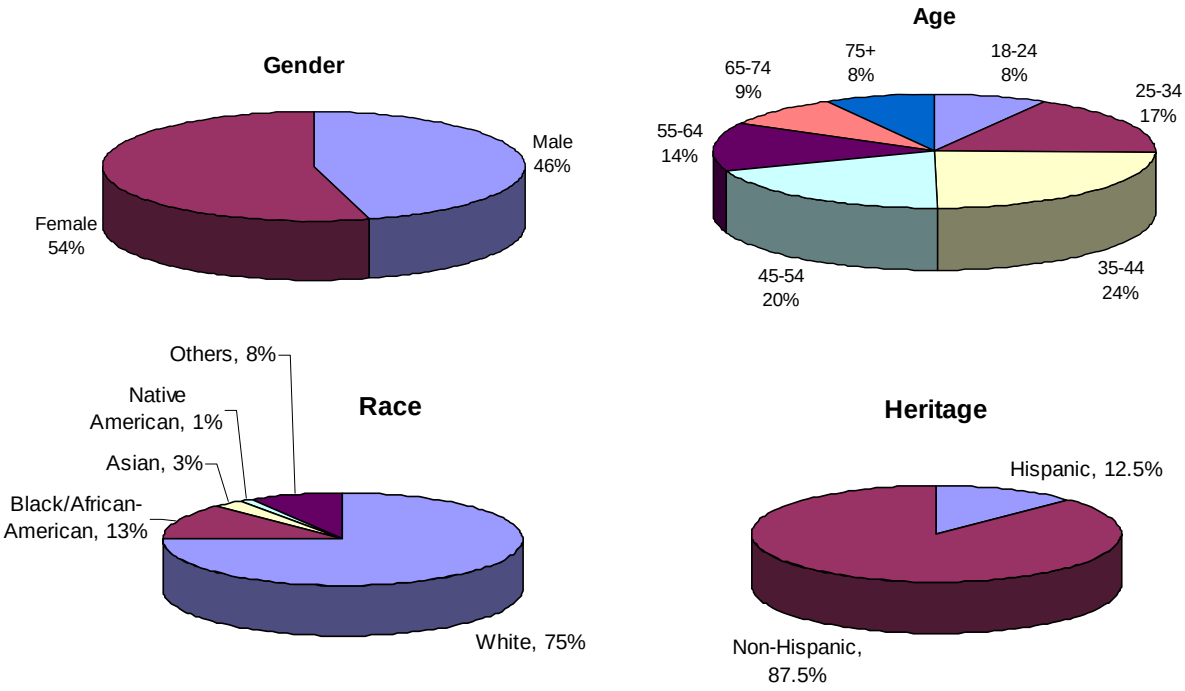


Figure 3-2. Second Visit Test Subject Demographics (from Table 1)

3.2 Data Collection Logistics

All scenario evaluation data were collected in a single room that emulated an indoor office environment. The test subjects, who were generally non-habituated, cooperative users, were guided through the data collection process by members of the Data Collection Test Team. Test team members included a Test Administrator, who operated the test harness computer and

software; a Test Observer, who demonstrated usage of each camera prior to enrollment, observed the data collection process, and noted any relevant information, conditions, and events;[22] and a Test Experimenter, who planned and managed the evaluation and audited data collection and results to ensure consistency and integrity. One test administrator and one test observer were in the room with the test subject at all times. The test experimenter audited the data collection process on an intermittent basis. All test team members underwent extensive training prior to collecting data to ensure uniform interaction between test team members and test subjects and to minimize test team influence on data collection.

The administrator and observer strictly followed a prescriptive computer-driven test protocol (see Section 3.4) and recited prepared scripts. Each test subject visit lasted nominally 30 minutes. All enrollment transactions were performed with the test subject sitting in a chair. Recognition transactions were performed standing for wall-mounted recognition units and sitting for all other units. For seated transactions, the test subject was asked to adjust the height and location of their chair such that their eyes were located at a specific height and distance from the desktop. The desktop cameras were located such that the specified seated-eye-height and chair location coincided with the optimal eye placement per manufacturer's guidelines. Visual alignment cues were provided to the administrator and observer so that they could verbally assist each test subject into optimal alignment with each camera.

Details about camera mounting configuration and test team alignment cues will not be described in this report as this information would allow the products to be identified. However, all products were mounted and handled in the data collection room according to manufacturer recommendations. The background for each desktop and wall-mounted product was a near-white, flat (as opposed to glossy) painted, solid wall. Product providers were invited to inspect the installation and usage of their products prior to data collection to ensure proper operation.

Overhead lighting intensity and positioning in the data collection room was consistent with that of an indoor office environment. Observations of the test team during the first visit indicated that the performance of one of the products was negatively influenced by the overhead fluorescent lighting. Specifically, the failure to acquire rate was very high for that product (~12%) and live images of the irises indicated intense, sporadic glare-like reflections and

[22] A sample Observer Log sheet can be provided to US government employees upon request.

scattered light effects. We did not observe these effects during our initial dry run testing. We hypothesize that for some test subjects, the angular relationship between the overhead lights, the test subject's irises, and the camera lens caused spurious reflections of the fluorescent lights into the camera lens. To minimize the spurious reflections for this product, we removed two fluorescent bulbs above that camera for the second visit. No lighting issues were observed for the other two cameras. Light intensity was measured at various points in the room using a Sekonic Handy Lumi Model-246 light meter for both visits. The measured illuminance at the optimal eye locations for each camera varied from 240 to 400 lux for the first visit and 110 to 270 lux for the second visit. The decrease in illuminance between Visit 1 and Visit 2 was 25% for Camera A, 32% for Camera B, and 32% for Camera C

Histograms indicating the temperature and relative humidity for each of the test subject visits are presented in Figure 3-3. The average temperature over all visits was 73.4º F and the

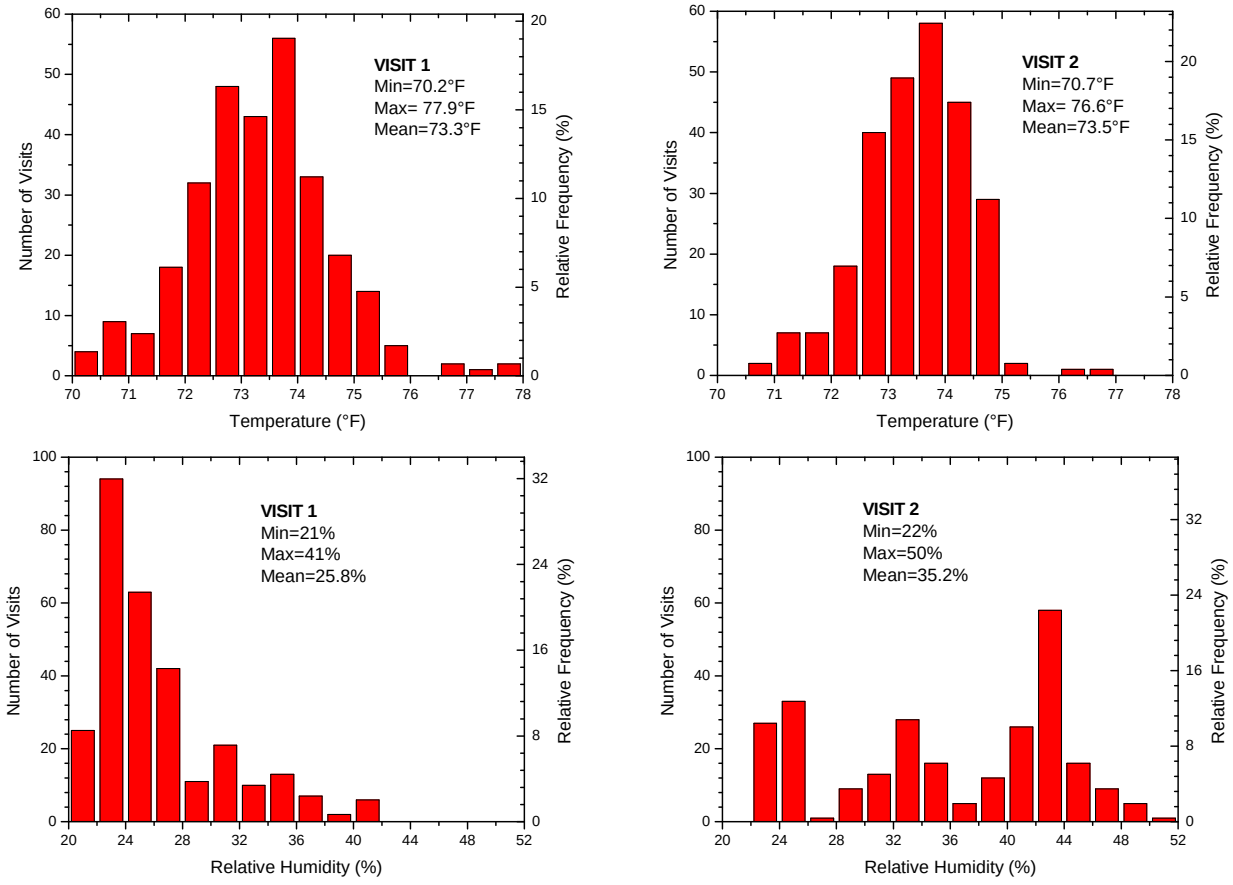


Figure 3-3. Test Room Temperature and Humidity

average relative humidity was 29.2%. A portion of the second visit occurred during Phoenix's monsoon season (July) resulting in higher relative humidity and a larger spread of values compared to the first visit.

3.3 Test Protocol

As noted above, the purpose of the scenario data collection was to measure online performance and to collect images for subsequent offline analysis. In online mode (real-time operation), the cameras tested do not provide meaningful match scores via the BioAPI function calls but do provide a match decision based on the internally configured threshold. With only match decisions, performance as a function of threshold score cannot be analyzed. As such, the images associated with each online enrollment and recognition attempt were saved for subsequent offline analysis, where match scores can be computed and a more thorough performance analysis conducted. Data analysis methods are discussed in detail in Section 5.

3.3.1. Data collection overview

The hierarchy of the various enrollment and recognition attempts is outlined in Figure 3-4. A summary of the test protocol follows. The detailed test protocol is described in Section 3.3.2.

Data were collected during two separate visits scheduled nominally six weeks apart. The six-week separation was selected for this study because it 1) was suitable for the logistics of data collection, 2) made optimum use of available human resources, and 3) reflected best practices per discussions in the international biometric performance testing standards community.

During the first visit, test subjects were enrolled in all three products. (One product arrived eight days after data collection began. For this product, the missed enrollments were conducted during the second visit.) The products were presented to the test subjects in random order to minimize any habituation effects. Three attempts were allowed to successfully enroll at least one eye in each product; 90 seconds were allowed for each attempt. Each product tried to enroll both left and right eyes during each attempt. Enrollment attempts ceased after at least one eye was successfully enrolled. Following enrollment, three verification attempts were conducted on each camera. Again, each camera was presented to the test subjects in random order. Twenty seconds were allowed for each attempt. Regardless of success or failure, three verification attempts were performed for each camera to maximize the images available for offline data

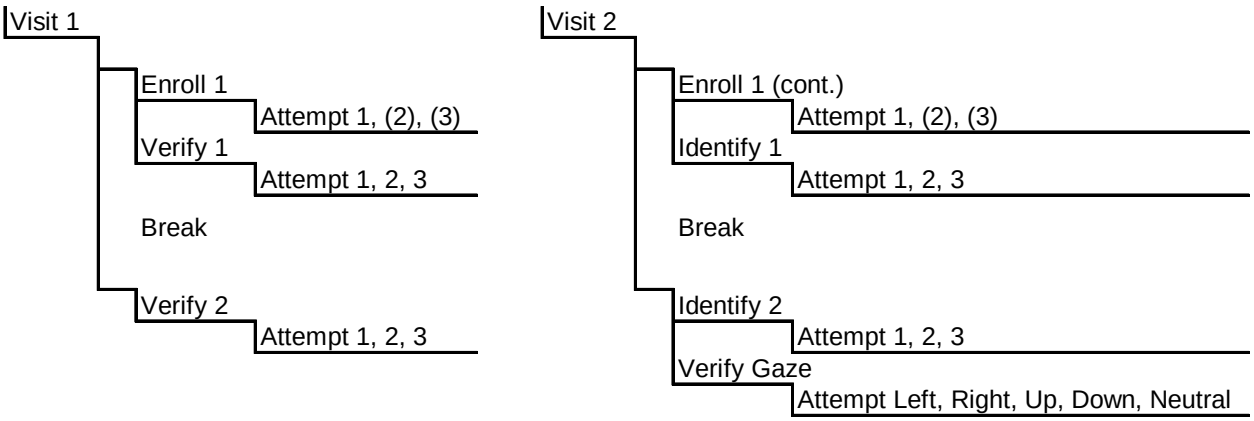


Figure 3-4. Data Collection Protocol Hierarchy

analysis. The test subjects were then invited to leave the data collection room for about five minutes to "disengage from the sensors". Following the short break, three additional verification attempts were conducted. Verification attempts were conducted even if an individual failed to enroll to collect additional data for offline analysis.

During the second visit, enrollments were conducted as needed for the late-arriving product, and then three identification attempts were conducted. As before, the cameras were presented in random order, and 20 seconds were allowed for each identification attempt. Again, the test subjects were invited to leave the data collection room for a short break and then three additional identification attempts were conducted. Following the identification attempts, the test subjects were guided through a series of off-axis gaze verification attempts where the head remained in an optimal position facing the camera and the eyes were directed to off-axis targets. Identification and verification attempts were conducted even if an individual failed to enroll to collect additional data for offline analysis. The test subjects were also asked to complete an optional, anonymous exit questionnaire during the second visit. The results of the exit questionnaire are presented in Appendix 11.2.

Online impostor recognition transactions were not conducted, primarily because the transactions could not be completed within the 30-minute time frame allotted for test subject visits. We had to choose between off-axis gaze attempts or impostor attempts to stay within the 30-minute time constraint. Since impostor recognition attempts could be performed during offline analysis and since the sponsor of the effort specifically requested off-axis results, we chose to perform the off-axis gaze attempts.

3.3.2. Detailed data collection protocol

The detailed data collection test protocol for each visit is described below. The protocols were administered by Authenti-Corp's Biometric Evaluation Test Harness (BETH) software and associated computer system. The test administrator operated BETH's graphical user interface (GUI), which directed the actions of the test administrator and test observer. In concert, BETH communicated with each product's biometric service provider via BioAPI function calls. Since the product-specific administrator and operator scripts essentially identify each product, they will not be provided in this report.

Throughout each visit, the observer recorded notable events and pertinent information on the observer log sheet and checked appropriate boxes (on paper) as each enrollment and recognition procedure was completed. The administrator also checked appropriate boxes (electronically) via the BETH data collection GUI throughout the visit as each script was recited and required action taken. The detailed test protocol for each visit is described below.

Visit 1

Step 1. Test subject validation and informed consent

Before entering the test room, the observer checked that the test subject signed the informed consent form and checked the test subject's government-issued identification (ID) to ensure the individual was the scheduled participant. Drivers' licenses and state ID cards were acceptable forms of identification. As required by WIRB protocol, the observer ensured that the test subject comprehended the informed consent form by asking the following questions:

i. Do you understand that your participation in this study is voluntary?
ii. Do you understand that we will be taking photographs of your irises?
iii. Do you understand that your name will never be associated with the photos of your irises?

If the test subject answered "yes" to the first three questions, the observer proceeded to answer any further questions that the test subject had and assumed that the test subject sufficiently comprehended the elements contained in the informed consent form.

If the test subject answered "no" to any of the first three questions, the observer verbally explained the elements and then asked the question(s) again. If the test subject still answered "no," the observer assumed that the test subject did not have the capacity to provide informed consent and informed the test subject that Authenti-Corp was not able to include them in the

study. If the test subject then answered "yes," the observer proceeded to answer any further questions that the test subject had and assumed that the test subject sufficiently comprehended the elements contained in the informed consent form.

After ensuring informed consent, the observer escorted the test subject into the test room, measured the test subject's standing eye-height, signed the informed consent form, provided a copy of the executed informed consent form to the test subject, affixed the pre-assigned UIN labels to the test subject's original informed consent form and to the observer's log sheet, and securely stored the UIN-labeled informed consent form. The UIN is used to ensure that the test subject's personally identifiable information, such as name and address, are not associated with the iris images collected during the study.

The administrator greeted the test subject, asked if they had any questions about the study or their participation in this study, answered any questions, recited a prepared introductory script, and asked the test subject to remove colored or patterned contact lenses if present.

The observer provided the UIN barcode label from the observer log sheet to the administrator, and the administrator scanned or manually entered the UIN to begin the BETH Visit 1 data collection software script.

BETH displayed the test subject's pre-populated demographic profile. The administrator reviewed this information with the test subject, made corrections to the existing information as directed by the test subject; entered the standing eye-height as measured by the observer; requested the test subject's eye color, height, and occupation; and asked if the test subject had any of the eye conditions shown in Figure 3-5. During this time, the observer recorded a variety of information on the observer log sheet, including temperature and humidity. The administrator then instructed the test subject to adjust the height of their chair using the visual alignment guides to achieve an optimal position for the tabletop cameras.

Step 2. Enrollment Transaction

Prior to beginning the enrollment process, the administrator recited a general introductory test script, asked the test subject to remove eye glasses and directed the test subject to open their eyes wide and look straight into each camera. The administrator then directed the test subject to enroll both eyes in each camera consecutively. The camera and eye presentation order (for one-

eye cameras) was prescribed by BETH in random order and communicated by the administrator. Each product's enrollment GUI was used during enrollment.

Before the test subject enrolled on each camera, the administrator recited product-specific test procedures, and the observer demonstrated correct positioning and usage of the camera. Both the administrator and the observer verbally guided the test subject into rough alignment with the camera and checked that the correct eye was being presented prior to initiating the product's enrollment GUI. The administrator was required to interact with the product enrollment GUI during the enrollment process. In addition, one or more of the products provided visual or auditory feedback to help the test subject correctly align with the camera after the enrollment GUI was initiated.

For each camera, up to three attempts were allowed to enroll at least one eye successfully. Enrollment attempts ceased after at least one eye was successfully enrolled. A 90-second timeout limit was imposed on each enrollment attempt. The 90-second time period started when the product's enrollment GUI was initiated. Within this time period, multiple presentations could be attempted and validation of the enrollment with a verification attempt could be processed as governed by the manufacturers' software and internal decision policies. Each product attempted to enroll both left and right eyes on each attempt. If the 90-second limit

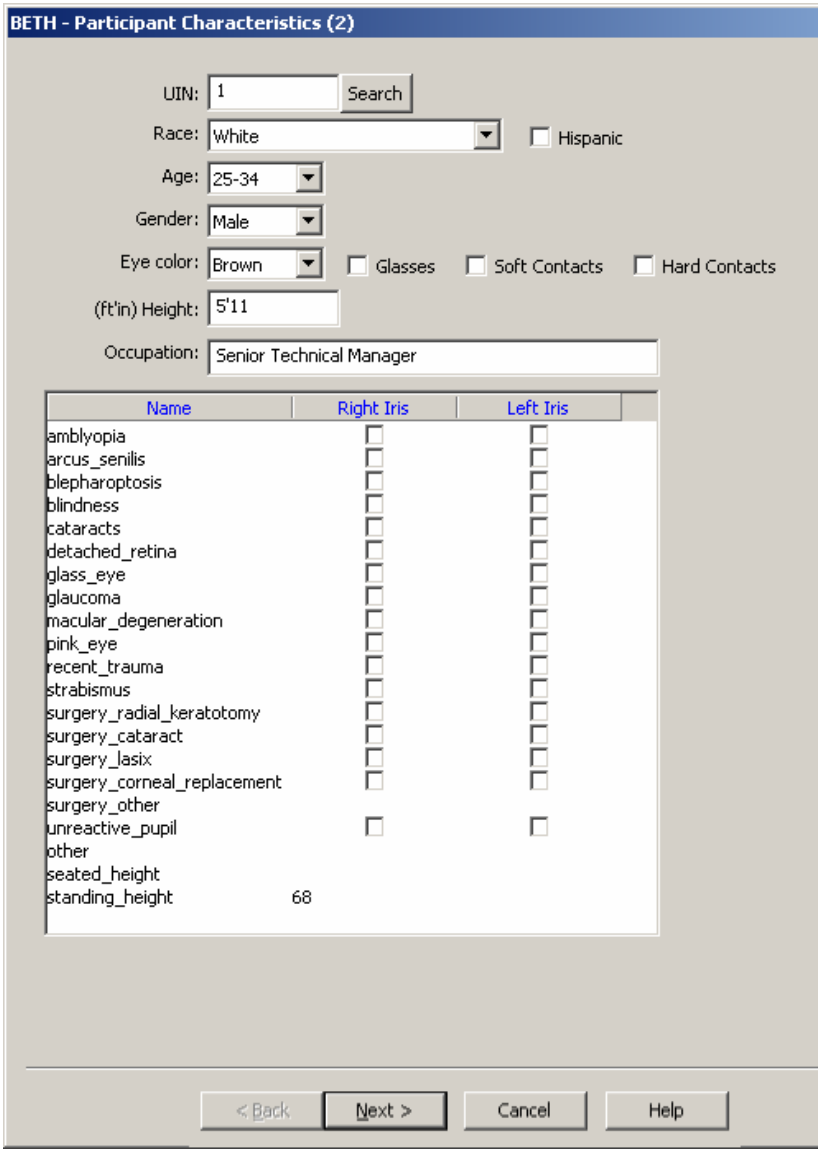


Figure 3-5. BETH Test Subject Information Screen

was reached before one eye was successful enrolled, the enrollment attempt was manually terminated and the administrator initiated a new enrollment attempt via BETH (up to three times per product). If at least one eye was enrolled successfully, the enrollment transaction was concluded.

The administrator provided the following verbal guidance to the test subjects for successive enrollment attempts:

- Attempt 1 – "Open your eyelids as wide as possible"
- Attempt 2 – "Tilt your chin up/down" (if doing so would increase the percentage of the iris presented to the camera)
- Attempt 3 – "Hold your eyelids open with your fingers"

Step 3. Verification Transaction 1

Prior to beginning the verification process, the administrator recited a general introductory test script. The administrator then directed the test subject to verify both eyes in each camera consecutively. The camera order was prescribed by BETH in random order and communicated by the administrator. For one-eye cameras, BETH randomly prescribed which eye was presented first and then alternated between right and left eyes on successive attempts. The eye to be presented was communicated by the administrator. Both the administrator and the observer verbally guided the test subject into rough alignment with the camera and checked that the correct eye was being presented prior to initiating a verification attempt. One or more of the products provided visual or auditory feedback to help the test subject correctly align with the camera during verification.

For one-eye cameras, three verification attempts were conducted for the right eye, and three verification attempts were conducted for the left eye. For two-eye cameras, three verification attempts were conducted with both eyes presented simultaneously. A 20-second timeout limit was imposed on each verification attempt, however the products often decided to terminate an attempt sooner than 20 seconds. Verification attempts against a “dummy” enrollment template were conducted if a test subject failed to enroll so that additional images for offline analysis could be collected.

If a test subject wore glasses, the administrator directed them to leave their glasses on for the first two attempts and remove their glasses for the third attempt. The administrator instructed the test subjects to look away from the camera (disengage) between successive verification

attempts and provided the verbal guidance below for successive verification attempts. If the prior attempt was successful, no additional guidance was given.

- Attempt 1 – no guidance
- Attempt 2 – "Open your eyelids as wide as possible" (if previous attempt was unsuccessful)
- Attempt 3 – "Tilt your chin up/down" (if previous attempt was unsuccessful and if doing so would increase the percentage of the iris presented to the camera)

Step 4. Break

After the first verification transaction was completed, the observer escorted the test subject to the waiting room, scheduled the second visit appointment, and provided the test subject with an appointment reminder card. The observer then escorted the test subject back into the test room. The short break disengaged the test subject from the cameras in an effort to emulate time separation between Verification Transaction 1 and Verification Transaction 2.

Step 5. Verification Transaction 2

Following the break, the observer provided the UIN barcode label affixed to the observer's log sheet to the administrator, and the administrator scanned or manually entered the UIN to initiate the BETH Visit 1 "after-the-break" data collection software script.

The Verification Transaction 1 protocol outline in Step 3 above was then repeated, with the exception that the administrator recited a slightly different introductory script.

Step 6. Payment and departure

Following completion of the second verification transaction, the administrator paid the test subject and obtained a signed receipt. The administrator also recited a "Thank you" script that emphasized the importance of the second visit and reminded the test subject that the stipend for the second visit was larger than today's stipend. The observer then escorted the test subject out of the building.

Visit 2

Step 7. Test subject validation

The observer checked the test subject’s government-issued identification (ID) to ensure the individual was the scheduled participant, measured the test subject's standing eye-height, and escorted the test subject into the test room.

The administrator greeted the test subject, recited an introductory script, and asked the test subject to remove colored or patterned contact lenses if present.

The observer provided the UIN barcode label affixed to the observer's log sheet to the administrator. (Each observer log sheet contains log information for Visit 1 and the UIN barcode label for a given test subject on one side, and log information for Visit 2 on the other side. The sheets were retrieved from secure storage prior to each scheduled appointment.) The administrator scanned or manually entered the UIN to begin the BETH Visit 2 data collection software script.

BETH displayed the test subject's demographic profile. The administrator asked the test subject if any information changed since the last visit, made corrections to the existing information as directed by the test subject, and entered the Visit 2 standing eye-height as measured by the observer. During this time, the observer recorded a variety of information on the observer log sheet, including temperature and humidity.

The administrator then instructed the test subject to adjust the height of their chair using the visual alignment guides to achieve an optimal position for the tabletop cameras.

Step 8. Enrollment Transaction (continued)

For those test subjects that were not enrolled in the late-arriving product, the Visit 1 enrollment transaction protocol was performed as outlined in Step 2 above.

Step 9. Identification Transaction 1

The Verification Transaction 1 protocol outlined in Step 3 above was then repeated with the following exceptions: 1) the administrator recited a slightly different introductory script, 2) identification was performed instead of verification, and 3) for one-eye cameras, BETH prescribed which eye was presented for each attempt in random order (as opposed to prescribing the first eye and then alternating between right and left eyes on successive attempts as was done in Visit 1). The eye to be presented was communicated by the administrator. Recall that for verification, the template created from the test subject's presented iris(es) is compared only to the previously enrolled template for that test subject (or to a dummy enrollment template if the test subject failed to enroll so that additional images for offline analysis could be collected). For identification, the template created from the test subject's presented iris(es) is compared to some

or all of the previously enrolled templates in the database (the Visit 1 enrollment templates in this case) in an effort to identify the test subject.

Step 10. Break

After the first identification transaction was completed, the observer escorted the test subject to the waiting room and invited the test subject to fill out an optional, anonymous exit questionnaire. The results of the exit questionnaire are presented in Appendix 11.2. The observer then escorted the test subject back into the test room. As with Visit 1, the short break disengaged the test subject from the cameras in an effort to emulate time separation between Identification Transactions 1 and 2.

Step 11. Identification Transaction 2

Following the break, the Verification Transaction 2 protocol outlined in Step 5 above was repeated with the following exceptions: 1) the administrator recited a slightly different introductory script, and 2) identification was performed instead of verification, and 3) for one-eye cameras, BETH prescribed which eye was presented for each attempt in random order (as opposed to prescribing the first eye and then alternating between right and left eyes on successive attempts as was done in Visit 1).

Step 12. Off-axis gaze verification attempts

Following completion of Identification Transaction 2 on each camera, several off-axis gaze verification attempts were conducted on that camera. Markers were placed around each camera to guide the test subject's gaze. Five gazes were studied: neutral (normal camera use), up (marked by pink square target marker), down (marked by black diamond target marker), left (marked by pink circle target marker) and right (marked by black triangle target marker). The markers were placed to achieve nominal gaze angles of ~20º up and down and ~28º left and right assuming the eye(s) were located at the ideal location in the center of each camera's collection volume. A detailed discussion of the off-axis gaze analysis is provided in Section 6.2.1.

Prior to beginning the off-axis gaze verification process, the administrator recited the following script: “We will be testing how different gazes work by having you look at the targets you see around the camera. For these gaze tests, you will be asked to look forward as before and hold your head still. Then, without turning your head, keep your face pointing forward and look at the target with your eyes. If you find any of these gazes to be uncomfortable for you, please

tell us and we will accommodate you." If looking at a target was uncomfortable for a test subject, the administrator directed them to look in the given direction as far as was comfortable. The administrator directed the test subject to remove their glasses for the off-axis gaze verification attempts.

One verification attempt was performed in the neutral position first, then BETH randomly prescribed up, down, left, and right gaze attempts. The gaze to be attempted was communicated by the administrator. For one-eye cameras, this process was performed for both left and right eyes. BETH randomly prescribed which eye was processed first. For two-eye systems, this process was performed once with both eyes presented simultaneously.

Before the test subject performed off-axis gaze verification attempts on each camera, the administrator recited product-specific test procedures. Prior to initiating a verification attempt the administrator and the observer 1) verbally guided the test subject into rough alignment with the camera, 2) checked that the correct eye was being presented, and 3) checked that the test subject was gazing at the requested target. The administrator directed the test subject to ignore instructions from the camera to look at the center of the camera. As with the previous verification and identification attempts, a 20-second timeout limit was imposed on each off-axis gaze verification attempt.

Once the off-axis gaze verifications were complete on a given camera, Identification Transaction 2 was performed on the next camera, as prescribed by BETH, until Identification Transaction 2 and the off-axis gaze verifications had been performed on each camera. The second identification transaction and the off-axis gaze verifications were performed on the same camera sequentially to keep the test subject visit time within the allotted 30-minute time frame.

Step 13. Payment and departure

Following completion of the second identification transaction, the final actions of the scenario evaluation test protocol were conducted: the administrator paid the test subject for the second visit and obtained a signed receipt, the administrator recited a "Thank you for participating" script, and the observer escorted the test subject out of the building.

3.4 Biometric Evaluation Test Harness (BETH)

As noted above, the scenario evaluation protocol was administered by Authenti-Corp's Biometric Evaluation Test Harness (BETH) software and associated computer systems. BETH is

an integrated solution for designing and executing biometric product evaluations. We describe below BETH's general capabilities and specific IRIS06 configurations.

BETH's evaluation management architecture provides a configurable and flexible platform that supports, regulates, and facilitates the biometric evaluation process. The experimenter customizes and configures BETH for a specific evaluation. The test administrator operates BETH's Administrator GUI, which directs the actions taken by the test team during data collection. BETH communicates with each product to perform the required actions and stores results in a relational database. The experimenter uses BETH to review data, query evaluation status, and analyze results.

During data collection, an administrator personal computer (PC) hosts the relational database and administrator software, and a second review monitor provides feedback to the observer and test subject, as illustrated in Figure 3-6. In general, any number of product PCs,

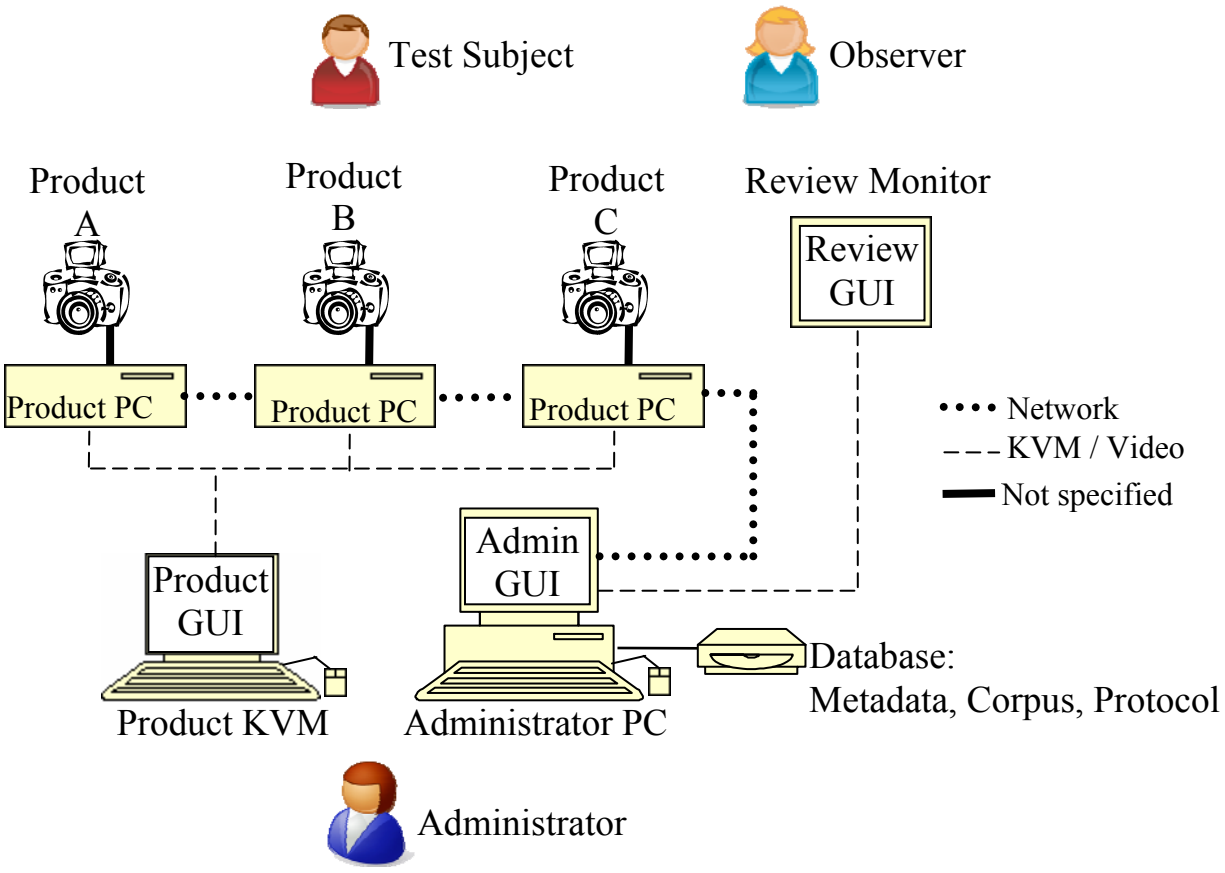


Figure 3-6. BETH Data Collection Architecture

which host the products to be evaluated, are accessible via a standard keyboard/video/mouse (KVM) switch. For IRIS06, three product PCs were required, and BioAPI 1.1 enabled communication between the products and the product PCs. BioAPI 1.1 is the ANSI standard for biometric application programming interfaces.[10]

3.4.1. Implementation

BETH provides a core architecture and implementation scheme that can be used for many evaluations. The IRIS06 products were integrated into the BETH environment, and BETH and its database were customized to support the ISO/IEC 19794-6 standard interchange format for iris images and to support the IRIS06 data collection test protocol.

Database

The BETH database consists of the "corpus," the "metadata," and relevant information about the data collection "protocol". The ***corpus*** contains the collected biometric images and templates, which are stored as BioAPI biometric information records (BIRs) for IRIS06. The ***metadata*** contains information about those images and templates, including the originating test subject UIN, product, and procedure, as well as any associated errors, recognition decisions, and timestamps. The metadata also includes test subject information, such as demographic profile (gender, age, race, and ethnicity), conditions (such as standing eye-height with shoes, cataracts, blindness, and various eye conditions for IRIS06), and completed data collection procedures (such as Enroll, Verify1, Verify2, Identify1, Identify2 for IRIS06). The ***protocol*** governs execution of data collection using a set of products, procedures, and rules (such as three 90-second attempts to enroll for IRIS06). The BETH database was configured for the IRIS06 corpus, metadata, and protocol requirements. BETH's administrator software used this configuration to drive the protocol during data collection.

Table 3-2 lists the IRIS06-specific metadata included in the database and indicates what information was provided by the test subjects (self reported), and what information was confirmed, assigned, or measured for each image by the experimenter during data collection or data review (ground truth). Data review is described in Section 5.2.1. BETH's analysis software uses the information in the database to provide filtering capabilities during offline data analysis as described in Section 5.2.5. The information that can be filtered using the current version of the

analysis software GUI is indicated in Table 3-2. Information filters that are not currently incorporated in the analysis software can be added as required.

Table 3-2. Test Subject and Image Information

		Self Reported	Confirmed during review	Assigned during review	Measured during data collection	Filterable
Demographics						
	Gender	✓				✓
	Race	✓				✓
	Ethnicity (Hispanic/Latino)	✓				✓
	Age	✓				✓
	Eye color	✓		✓		✓
	Height	✓				
	Standing eye height				✓	
	Occupation	✓				
Image Parameters (Hints)						
	Bad picture			✓		✓
	Bad features			✓		✓
	Obstruction			✓		✓
	Bad placement			✓		✓
	Bad environment			✓		✓
Features / Conditions						
	Left eye / Right eye		✓			✓
	Glasses	✓	✓			✓
	Contacts	✓				✓
	Hard Contacts		✓			✓
	Eye conditions: amblyopia (lazy eye), acrus senilis, blepharoptosis (droopy eyelid), blindness, cataracts, detached retina, glass eye, glaucoma, macular degeneration, pink eye, recent trauma, strabismus, surgery (radial keratotomy, cataract, lasix, corneal replacement, other), unreactive pupil, other	✓				

Data format integration

For IRIS06, BIRs are stored in the format used by BioAPI 1.1, which consists of a CBEFF header and a data block containing ISO/IEC 19794-6:2005 Biometric data interchange formats – Part 6: Iris image data. These standards specify how to store iris images with relevant metadata in an open and standard way. IRIS06 constrained the use of the ISO 19794-6:2005 standard to 640x480 uncompressed 8-bit gray-scale rectilinear images, as this is the most common format output by iris recognition products and accepted by 3rd-party template

generators. A single file in this format may contain multiple iris-feature sets (left iris, right iris, or left and right irises) and any number of images of each feature. For enrollment, all products produced and exported native audit BIRs with both left and right iris images. For recognition attempts, two-eye products produced and exported native audit BIRs with both irises, while one-eye products produced and exported native audit BIRs with only one iris.

BETH follows a strict file naming convention so that all BIRs are stored uniquely and can be easily identified. The naming convention, shown in Figure 3-7, consists of the test subject's 4-digit UIN; the datasource-product, feature, gaze, transaction, image (not used for IRIS06), attempt, and template generator identifiers.

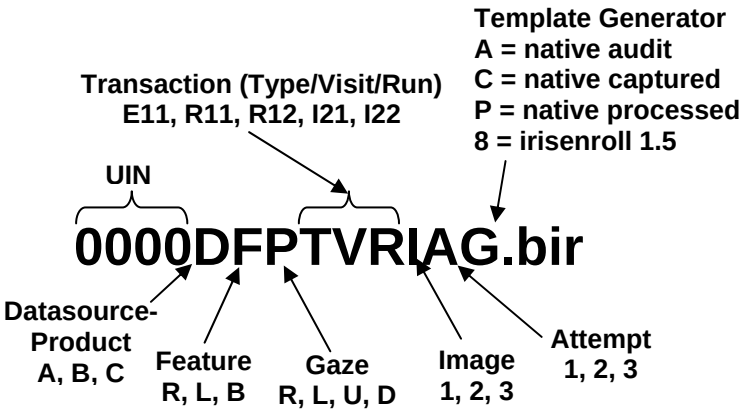


Figure 3-7. BETH BIR File Naming Convention

For IRIS06, the native audit BIRs contain the iris image data in the ISO/IEC 19794-6 data interchange format. BETH stored the native captured and native processed BIRs when available, but they were not used for IRIS06. When native audit BIRs containing iris images were processed by the Daugman "irisenroll" tool, the output was saved with the same filename except with the template generator identifier set to 8 instead of A.

Product integration

A custom BETH module (ACSourceBioAPI1_1.exe) was developed to integrate the products into BETH using the BioAPI 1.1 reference implementation.[23] The BETH module initialized a product's Biometric Service Provider (BSP) and communicated with that BSP to perform actions during data collection, such as enrollment, verification, and identification. A detailed product specification was developed and delivered to product vendors to ensure compatibility with IRIS06. Unfortunately, none of the product vendors delivered a fully compliant BSP, however adjustments were made to accommodate all products. When necessary, Authenti-Corp developed BSPs using the products' SDK and BioFoundry BioAPI BSP-builder

[23] http://www.bioapi.org/Downloads/BioAPI%201.1.doc (accessed 1 September 2007).

tools.[24] A set of identical PCs acquired and configured in exactly the same manner hosted the products and their BSPs.

The BioAPI 1.1 functions utilized included BioAPI_Capture, BioAPI_Enroll, BioAPI_CreateTemplate, BioAPI_DbStoreBIR, BioAPI_Verify, and BioAPI_Identify. BioAPI commands were executed on the PC hosting the product and were timed from the start of the command until it returned a result. The BioAPI enrollment process acquired and registered templates to be compared against later using recognition (verification and identification) commands. Recognition commands returned an error, such as failure to acquire, or a decision as to whether the probe template matched the enrolled template for verification or which enrolled template(s) it matched for identification. BioAPI functions that return audit BIRs provided the required image data in the ISO/IEC 19794-6 data interchange format as described above. Returned error messages, such as a failure to acquire an image, were logged in the BETH database.

One product was delivered late and without a fully compliant BSP. The SDK provided for this product was sufficiently complicated that only a portion of the required functionality for IRIS06 could be integrated in a timely manner. As such, during data collection, this product captured iris images using the BioAPI commands "Capture for purpose Enroll," "Capture for purpose Verify," and "Capture for purpose Identify" but did not perform enrollment or recognition decisions in real time. The captured images were used later for enrollment and recognition in the product's proprietary environment. As such, this product's proprietary environment established its online results instead of the BioAPI environment. Authenti-Corp does not believe that this constitutes a significant difference as the data were collected using the BioAPI environment and the match results would be the same whether calling a BioAPI function or a proprietary API function to get those results. Further, all comparisons for this product were performed in identification mode, instead of in verification mode. (The product did not support verification.) Transaction times for this product included only BioAPI capture times and did not include subsequent matching time. However, since matching time was found to be a small fraction of a second, Authenti-Corp believes that transaction times (typically many seconds) were not significantly impacted.

[24] The BioFoundry Division of OSS Nokalva (www.biofoundry.com) graciously provided BioAPI support for the IRIS06 effort free of charge.

3.4.2. Data collection tools

During data collection, the test team manages test subject visits using scheduling tools and data collection GUIs as described below.

Test subject management

Once test subjects agree to participate in an evaluation, they are entered into an Excel spreadsheet that is used to generate their UIN. The link between their personally-identifiable information and their UIN is kept securely in the Excel spreadsheet and on the printed informed consent form as documented in Section 2.2. All paper documents that include the UIN do so in plain text and as a barcode for automated data entry.

Prior to data collection each day, the test team imports a sanitized version of the spreadsheet file, containing only the UINs, demographics, and scheduled appointment dates and times into BETH. Once a test subject is registered in the BETH database, the test subject is able to participate in data collection procedures using their UIN. The test subject UINs and associated metadata are also available to the statistics and management tools used by the experimenter.

Data collection GUIs

Each product to be evaluated is integrated with BETH using a custom software module that communicates with the product in its own 'language'. For IRIS06, this module uses the BioAPI 1.1 reference implementation[25] to communicate with the product's BSP, to execute commands, and to report results. This software runs on each PC that is hosting a product to be evaluated and is registered with the BETH database so that BETH can communicate with each PC to execute commands using the products.

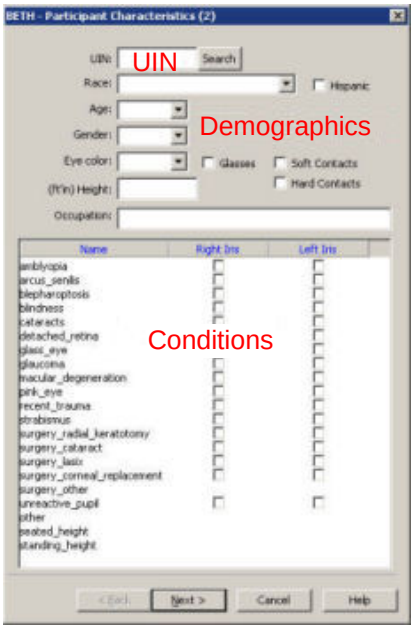


Figure 3-8. BETH Test Subject Characteristics

BETH's administrator software drives the required protocol during each test subject visit and stores results in the database. This software is easy to use and, to minimize the potential for human error during data collection, it requires very little manual data entry. First, the test subject's UIN is

[25] http://www.bioapi.org/Downloads/BioAPI%201.1.doc (accessed 1 September 2007).

entered and demographic and condition information is updated as shown in Figure 3-8. Next, the software guides progress through the data collection protocol by presenting the administrator with the proper task order, required instructions, and prompts at the correct times as shown in Figure 3-9 In some cases, interaction with a product's own GUI is required (as is the case during IRIS06 enrollment), so the administrator switches to the PC hosting the product during this procedure using a standard KVM switch. On the administrator PC, a second monitor shows prompts and results to the test subject and observer as illustrated in Figure 3-10.

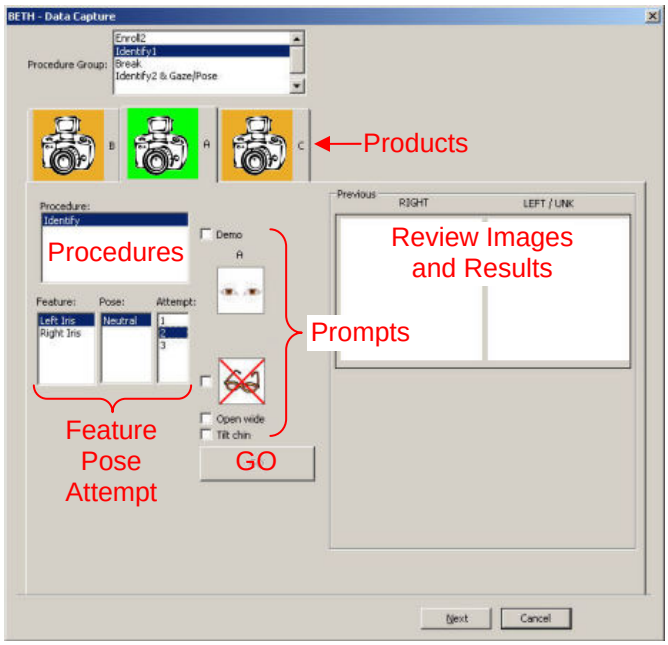


Figure 3-9. BETH Administrator GUI

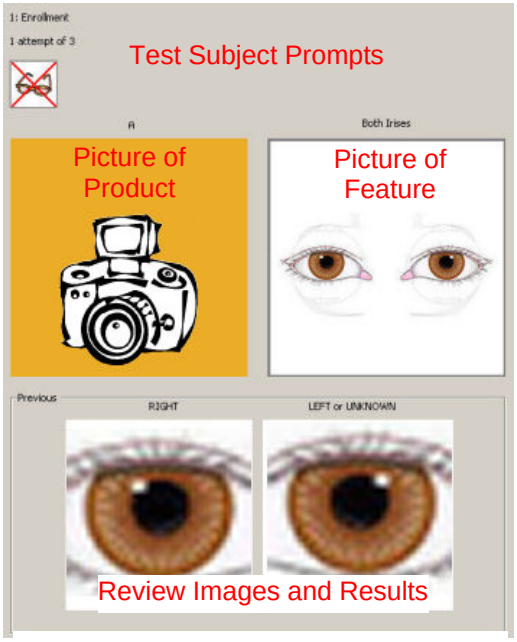


Figure 3-10. BETH Test Subject Review GUI

3.4.3. Analysis tools

BETH provides tools to review data, to compile and report online results, and to perform offline data analysis as described below.

Data review tools

The experimenter uses BETH tools to view summaries of test subject information and statistics (ACExperimenter.exe) and to review biometric samples for errors (ACReview.exe). If problems are discovered during the review process (such as an incorrect eye presentation), the affected data elements are flagged for exclusion in the database.

Biometric image review is performed in a flexible environment where the reviewer can access each and every image in isolation or in groups of interest (Figure 3-11). The software allows the reviewer to compare any two images and view the resulting match score, to manually enter a note attached to the image, and to mark the images with several common 'hints' (such as eyes closed, out-of-focus, or glasses-on). These hints can be used to filter the dataset during data analysis.

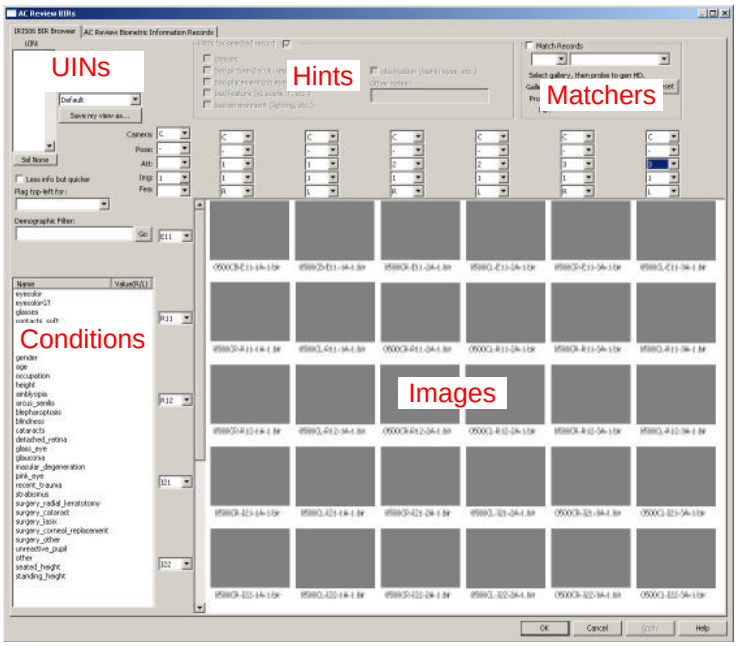


Figure 3-11. BETH ACReview Software

Online results tool

The experimenter exports raw online results from the database into a well-prepared spreadsheet to tabulate the overall online results of the evaluation. The spreadsheet includes calculations for various performance metrics and confidence intervals as described in Sections 5.1-5.3 below.

Offline template generation and matching

For offline evaluation, BETH provides a tool (ACTempgen.exe) to run template generators on unprocessed biometric images to produce templates that matching tools can use. The experimenter runs matching software on these generated templates to produce match score matrices (ACAnalyze.exe). Match score matrices are tables containing match scores between enrollment and recognition templates that are used to examine and plot performance metrics. BETH provides tools for flexible creation and labeling of templates and matrices. Each template generator or matcher is assigned a unique ID and is integrated with BETH using a custom software module that can communicate with the tool in its own 'language'. Each tool runs as a server and is registered with BETH so that its services can be used. BETH provides a list of available tools to the experimenter and provides the option to execute them on the collected biometric data. For IRIS06, Professor John Daugman's "irisenroll (release 1.5)" template generator and "matcher1" template matcher were integrated and used for offline analysis.

Offline data analysis

The experimenter uses BETH's offline analysis tool (ACAnalyze.exe) to 1) produce match score matrices; 2) filter and process the matrices based on the desired demographics, conditions, and image parameters listed in Table 3-2, quality scores (not available for IRIS06), feature sets, and transaction types; and 3) output or graph the results (histograms, Detection Error Tradeoff and Receiver Operating Characteristic curves, and confidence intervals as discussed in Sections 5.2.5 and 5.3.

The ACAnalyze software, illustrated in Figure 3-12, is configured to understand the products and test procedures for a given evaluation and has access to matching, graphing, and other analysis tools. Correct functioning of the ACAnalyze tool was validated by inputting the same source data into the "Program for Rate Estimation and Statistical Summaries" or PRESS tool and checking that the results from both tools were identical.[26]

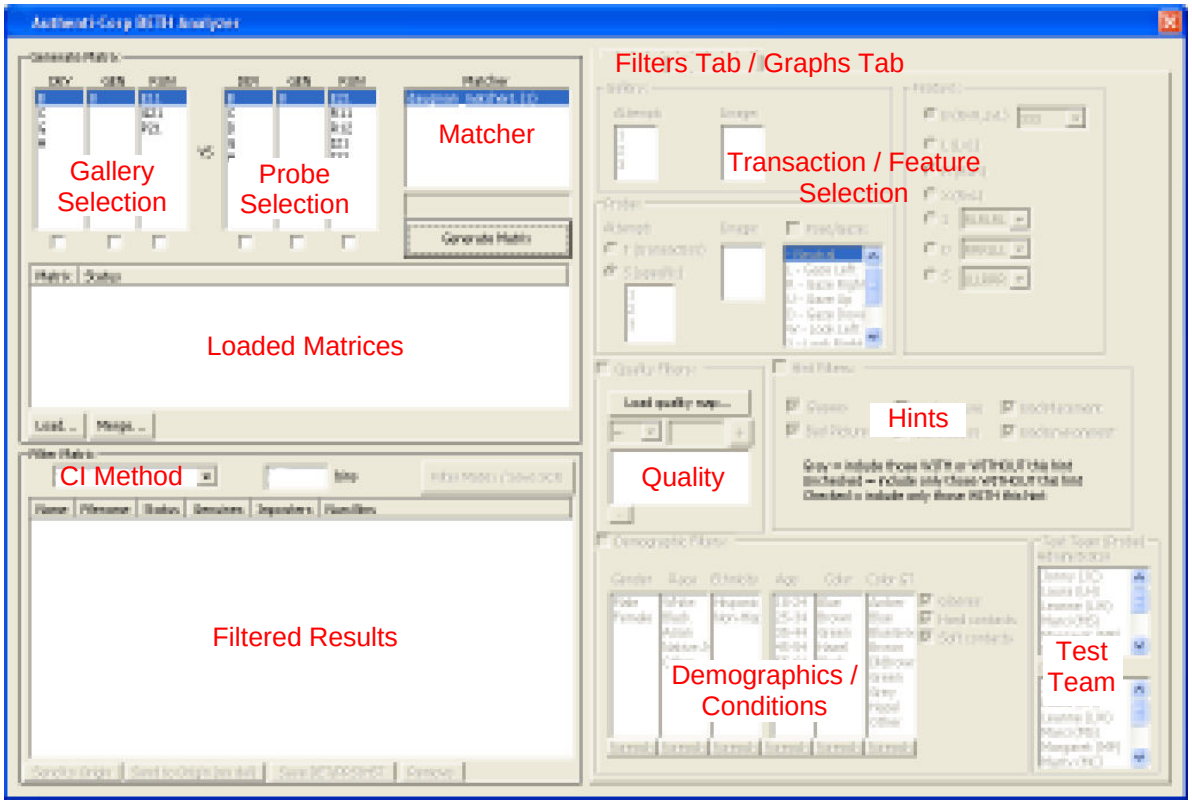


Figure 3-12. BETH ACAnalyze Software

[26] The development of PRESS, a very useful tool for analyzing data collected on biometric authentication devices, was funded by the Center for Identification Technology Research (CITeR) at West Virginia University (http://www.citer.wvu.edu/) and St. Lawrence University. The tools is available at http://it.stlawu.edu/~msch/biometrics/downloads.htm (accessed 1 September 2007).

4. Controlled Off-Axis Data Collection

In iris recognition systems, enrollment and recognition are typically performed under ideal placement conditions where the eyes are placed in an optimal on-axis position relative to the product's camera as shown in Figure 4-1. The Controlled Off-Axis experiment investigates the performance of iris recognition technology under non-ideal off-axis placement conditions with procedures designed to emulate the behavior of an uncooperative user. Off-axis recognition attempts were performed against ideal on-axis enrollment templates by a small test population using a physical apparatus specifically designed to control and measure the linear (X, Y, and Z) and rotational (Yaw, Pitch, and Roll) offset between the test subject's iris and the camera.

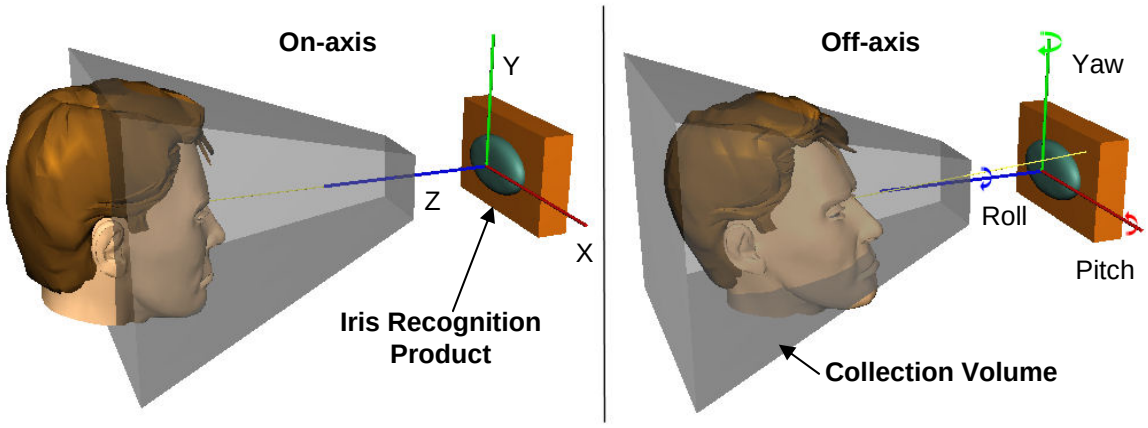


Figure 4-1: On-axis versus Off-axis Presentation

For this experiment, the origin is placed at the optical center of the camera with the Z axis pointing along the line of sight. Linear translation on the Y axis is up and down, X is side-to-side, and Z is forward and backward. Yaw can be thought of as shaking your head side-to-side as when saying "no", pitch as when saying "yes" (nodding head up and down), and roll as when saying "maybe" (ear to shoulder).

4.1 Data Collection Logistics

As with the scenario evaluation, all controlled off-axis data were collected in a room that emulated an indoor office environment. A single Test Administrator guided the six volunteer test subjects (members of Authenti-Corp's Data Collection Test Team) through the data collection process. The administrator underwent extensive training prior to collecting data to ensure uniform interaction with the test subjects and to minimize administrator influence on data

collection. Being members of the scenario data collection team, all test subjects had previous training and familiarity with the iris recognition products.

The administrator used a physical apparatus and a computer-driven test harness software application to run procedures for each iris recognition product with each test subject. An on-axis ideal enrollment and three off-axis recognition procedures, using different translational and rotational offset styles, were used to evaluate the off-axis performance of the products.

4.2 Physical Apparatus

A physical apparatus was designed and built to execute the protocol procedures (described in Section 4.4 below) in a repeatable fashion. The apparatus, shown in Figure 4-2, controls and measures the offset between the camera and the test subject for off-axis recognition attempts. Notice that instead of moving the test subject around the camera, the apparatus moves the camera around a stationary test subject.

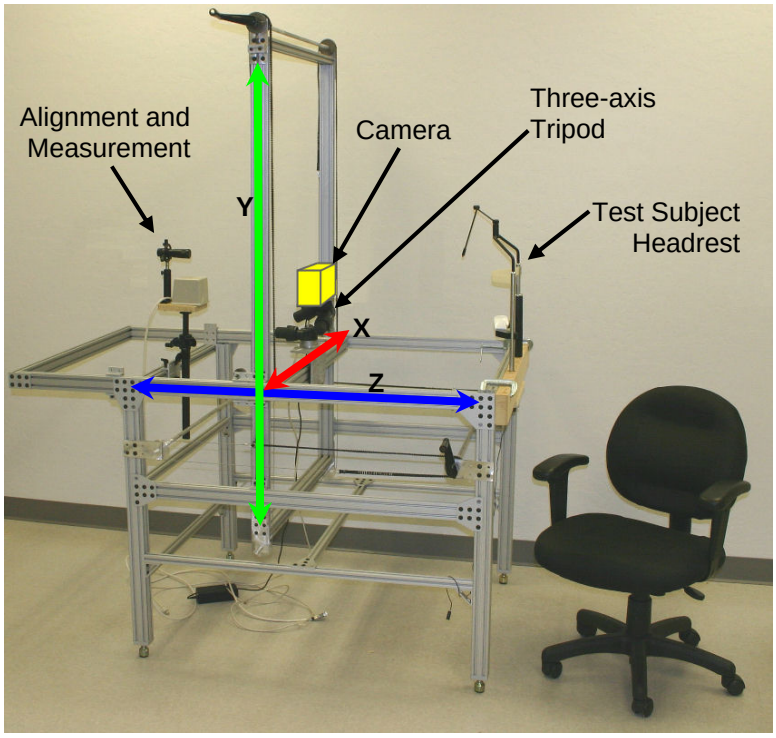


Figure 4-2: Physical Apparatus

The apparatus provides translation control along the X, Y, and Z axes labeled in Figure 4-2, as well as three-axis rotation control using a three-axis tripod. This allows six degree-of-freedom movement of the camera in relation to the test subject. A magnetic tracking system measures the offset between the camera and the test subject.

Each camera attaches to the three-axis tripod on the apparatus, using a custom built mount. The tripod, with mounted camera, is then able to move within the apparatus along the three axes of translation using belts, cranks, and slides. The magnetic sensor is rigidly fixed to each camera with reproducible, constant offset between the magnetic sensor and the camera's

optical center. This ***Sensor Offset*** is pre-configured in the test harness software for each camera and is shown by the blue line in Figure 4-3.

The test subject sits in the chair and leans their head against the ophthalmologist headrest. The height of the chair is raised and lowered so that the test subject can lean comfortably against the headrest. The head is in no way constrained and the test subject can easily back away from the headrest at any time. An eye gaze light is used to maintain a direct, forward looking gaze when it will not obstruct the camera's view of the test subject's iris.

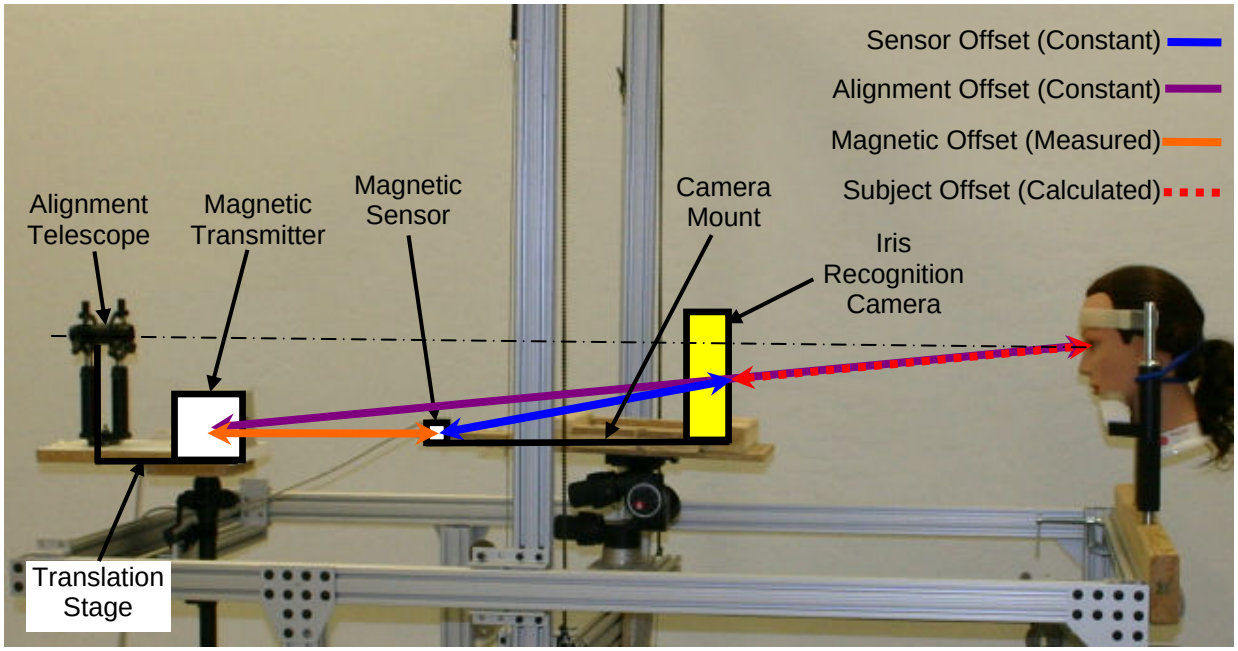


Figure 4-3: Apparatus Measurement System

Once the test subject is in position, the alignment telescope translation stage is used to set a pre-determined ***Alignment Offset*** (shown by the purple line in Figure 4-3) between the magnetic transmitter and the point of interest on the test subject's face. The point of interest is the midpoint between the test subject's eyes for two-eye cameras or the center of the iris for one-eye cameras.

As with the Sensor Offset, the constant Alignment Offset distance is pre-configured in the test harness software for each camera. To set the Alignment Offset, the chin height of the head rest is adjusted until the test subject's eyes are horizontally aligned with the alignment telescope's crosshairs. Then the point of interest is aligned with the vertical crosshairs of the

alignment telescope by moving the magnetic transmitter and telescope unit with the translation stage.

Once aligned and initialized, the magnetic tracking system continuously reports the ***Magnetic Offset*** between the magnetic transmitter and the magnetic sensor to the test harness software about 30 times per second (~30Hz). This dynamic Magnetic Offset and the pre-configured constant Sensor and Alignment Offsets provide the test harness software the information needed to calculate the ***Subject Offset*** between the camera's optical center and the test subject point of interest. The Subject Offset is recorded in the test harness software database with each recognition attempt providing a three-dimensional (3D) map of all recognition attempts.

The test harness software also provides a real-time 3D view of the collected data and the current offset as reported by the magnetic tracking system, as shown in Figure 4-4. Physical movement of the sensor in the real world affects the magnetic offset in the software, which then moves the sensor and camera indicators in the 3D view providing immediate feedback of measurement values. Top and side views as well as 3D navigation capabilities are provided for easy alignment and visualization of the data being collected.

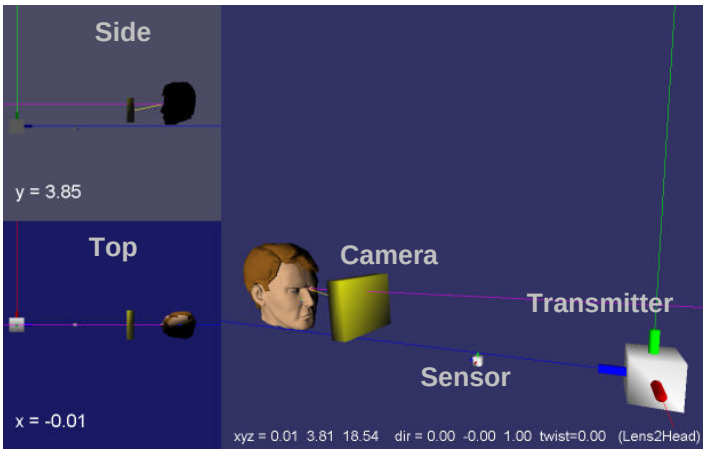


Figure 4-4: Test Harness 3D View

4.3 Test Subject Logistics

As mentioned above, the six test subject volunteers were Authenti-Corp employees. Each test subject was scheduled for testing when the iris recognition products were not in use for the scenario evaluation. Each testing session for an individual test subject lasted nominally between one and four hours and the average total time spent with a subject was nine hours. The test administrator created a testing schedule based on test subject availability in which individuals had nominally one to two testing sessions per week until all procedures were tested with each product.

The six volunteers consisted of: three females and three males; three 18-24 year olds, two 25-34 year olds, and one 35-44 year old; five Whites and one Asian, and no Hispanic/Latinos. If a test subject wore glasses they were always removed prior to testing, however contact lenses were allowed and were worn by one of the test subjects.

4.4 Test Protocol

As with the scenario evaluation, Authenti-Corp's BETH software administered the procedures for the off-axis experiment. The test administrator operated BETH's GUI to accomplish each of the procedures. In concert, BETH communicated with each product's BSP via BioAPI function calls and stored results in the BETH database. The test administrator physically positioned the cameras and verbally instructed the test subject for each recognition attempt. Since some product-specific testing details identify each product, they will not be provided in this report.

Enrollment and some recognition attempts were performed at the ***ideal*** test subject location, which was often used as a starting point or a reference point for measurements. The ideal test subject location was defined by calculating the center of the camera's advertised collection volume. The center of the volume always rests directly in front of the camera's optical center so that X and Y are zero. Z represents the optimal standoff distance between the camera's optical center and the test subject point of interest and varies per product. There is no rotational offset for the ideal location so yaw, pitch, and roll are all zero.

Each test subject first performed an on-axis enrollment in the ideal location. After successful enrollment, the test subject attempted recognition at a variety of linear and rotational offsets, or ***poses***, according to three procedure styles: ***Neutral***, ***Sweep***, and ***Translate***. Neutral procedures represent a user looking directly at the camera from different positions around the camera. Translate procedures represent a user looking directly forward but not positioned directly in front of the camera. Sweep procedures represent users located directly in front of the camera but looking away from the camera. These three procedures were performed as appropriate for each product. If a particular procedure did not produce revealing data, the results of that procedure are not presented. As such, if a procedure is not reported, it can be assumed that the product simply did not work off-axis in that case.

For all poses, the test subject held a direct gaze in which the eyes looked straight forward, orthogonal to the face as shown in Figure 4-5. Gazing at angles other than orthogonal to the face is explored in the guided gaze off-axis experiment (presented in Section 6.2.1).

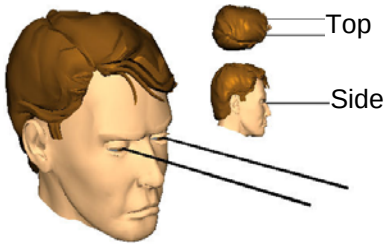


Figure 4-5: Direct Eye Gaze

During controlled-pose off-axis data collection, the administrator explored the performance of the product as the test subject was located in and around the product's collection volume. In areas of interest, such as near the edge of the collection volume or where the product began to show signs of failure to acquire or match, increments of movement were decreased to better define the edges of performance. Larger increments of movement were used in areas where a product was known to perform quite well so that the areas of interest could be most effectively explored.

All testing procedures involved performing off-axis iris verification against an on-axis enrollment template. Recognition attempts were limited by a 20-second timeout. If the timeout expired then the attempt was terminated and constituted a failure to acquire. Each of the procedures was performed once for two-eye cameras and once on each iris for one-eye cameras. The following steps describe the test protocol in detail.

Step 1. Physical apparatus setup and alignment

Before a session of data taking began, many steps were performed to prepare and align the camera, the test subject headrest, and other apparatus components. The test administrator attached the camera to be tested to the physical apparatus using the camera's custom tripod mount. The magnetic sensor was attached to the mount at the predetermined location for each camera and aligned with respect to the camera's optical center to match the pre-configured Sensor Offset. Finally, the administrator checked that the telescope and magnetic transmitter were aligned to the apparatus frame.

A mannequin with artificial eyes was used to verify the alignment. The artificial iris was placed in the telescope crosshairs by moving the translation stage holding the magnetic transmitter and telescope. The software then calculated the ideal position for the camera, and the camera was moved to that position. To verify that the physical apparatus ideal location and the software-calculated ideal location are the same, an image of the iris was captured. An iris

appropriately centered in the resulting image indicated that the software and physical apparatus were calibrated. This process was repeated before the start of each testing session and after changing products during testing.

Step 2. Test subject processing and alignment

New test subjects were given the informed consent form, and their demographic information was collected using the process described in Section 3.3.2, Visit 1, Step 1. After ensuring informed consent, test subjects were assigned a UIN to uniquely identify them in the test harness software during the evaluation.

The test administrator demonstrated how to properly position the head in the headrest to ensure consistent results. Test subjects were instructed that they could remove their head from the headrest or take breaks to rest their eyes at any time during testing. Before testing, the administrator asked the test subjects to remove eye glasses and colored or patterned contacts.

To align the apparatus measurement system, the test administrator and subject performed the alignment steps as described above in Section 4.2. In addition, the focus of the crosshairs on the point of interest was checked periodically throughout testing and any time the test subject withdrew their head from the headrest. This process was repeated whenever the point of interest changed, such as when switching eyes.

Step 3. Enrollment procedure

The first procedure performed with a product and test subject was on-axis enrollment under the ideal manufacturer-specified conditions. The test administrator gave verbal cues to open the eyelids wide and to look at the camera to ensure the highest quality enrollment template. Up to three 90-second attempts were allowed to enroll successfully.

To reduce realignment time, both irises were not generally enrolled consecutively with single eye products. One iris was enrolled then tested with the first recognition procedure before the other iris was enrolled. Subsequent recognition procedures, after both irises had already been successfully enrolled, were generally performed on alternating irises to give each eye a resting period.

Step 4. Neutral verification procedures

Neutral procedures represent a user looking directly at the camera from different positions around the camera as shown in Figure 4-6. These procedures are accomplished by positioning the camera directly in the line of sight of the test subject (X=0, Y=0) and then adjusting the distance and rotation of the camera (Z, Pitch, and Roll) using the apparatus. Because the camera was positioned directly in front of the user for these procedures they were asked to look directly at the camera and open their eyes wide for each recognition attempt.

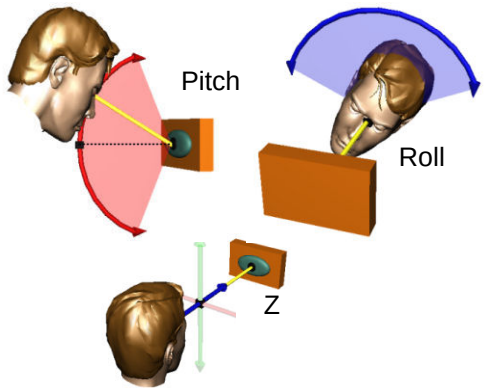


Figure 4-6: Neutral Procedures

Step 4a. Neutral Z

Neutral Z emulates the user located directly in front of and looking directly at the camera from various distances. For this procedure, the administrator placed the camera in the ideal location at the center of the camera's collection volume and moved it forward and backward in increments. Recognition attempts were performed at each increment.

Step 4b. Neutral Pitch

The Neutral Pitch procedure emulates a user positioned above or below the camera at various distances while looking directly at the product. This procedure was performed only for products with large collection volumes because products with small collection volumes did not produce reportable results.

To perform this procedure the camera was rotated in front of the stationary test subject about the X-axis (Pitch) at various distances using the camera's optical center as the center of rotation. As previously described, the camera was always located along the test subject's line of sight (X=0, Y=0). First, the camera was pitched (rotated around the X axis) by an angle with respect to the test subject's gaze axis. Then the camera was translated in increments along the Z-axis to perform recognition attempts at various distances from the product at that angle.

Step 4c. Neutral Roll

The Neutral Roll procedure emulates the user located directly in front of and looking at the camera but tilting the head to the side, "to touch an ear to a shoulder". This procedure was

performed for all cameras at one or three discrete distances from the camera, depending on the capabilities of the camera.

To perform this procedure the camera was rotated in front of the stationary user about the Z-axis (Roll) using the camera's optical center as the center of rotation at various distances. As previously described, the camera was always located along the test subject's line of sight (X=0, Y=0). First, the camera was moved along the Z axis to a specific distance from the user. Then, the camera was rolled (rotated about the Z axis) in increments where recognition attempts were performed.

Step 5. Sweep procedures

Sweep procedures represent users located directly in front of the camera but turning the face away from the camera as shown in Figure 4-7. These procedures were performed for all cameras at one or three discrete distances from the camera, depending on the capabilities of the camera. These procedures were accomplished by rotating the camera around the stationary test subject using the test subject point of interest (center of the eye or midpoint between the eyes) as the center of that rotation while maintaining a fixed distance. Because the camera was not positioned in front of the test subject they were asked to look directly forward, not at the camera, and open their eyes wide for each recognition attempt. When it would not interfere with the camera's line of sight, the gaze indicator light on the headrest was used to help the test subject keep a forward gaze.

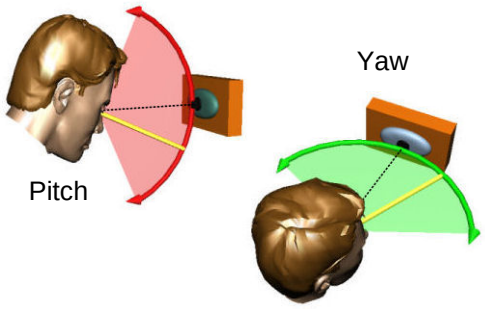


Figure 4-7: Sweep Procedures

Step 5a. Sweep Pitch

The Sweep Pitch procedure emulates a user positioned directly in front of the camera but looking up or down away from the camera. At each of the one to three distances, the camera was moved and rotated around the test subject to maintain the fixed distance and to increment the pitch angle between the camera and test subject. Recognition attempts were performed as the pitch angle was incremented.

Step 5b. Sweep Yaw

Sweep Yaw emulates a user positioned directly in front of the camera but looking to the right or left of the camera. At each of the one to three distances, the camera was moved and

rotated around the test subject to maintain the fixed distance and to increment the yaw angle between the camera and test subject. Recognition attempts were performed as the yaw angle was incremented.

Step 6. Translate procedures

Translate procedures represent a user looking directly forward but not positioned directly in front of the product as shown in Figure 4-8. These procedures were performed for all products at one or three discrete distances from the product, depending on the capabilities of the product. These procedures were accomplished by translating the product in the X and Y axis at the one or three Z distances. No rotation was involved in these procedures. Because the product was not positioned in front of the test subject they were asked to look directly forward, not at the product, and open their eyes wide for each recognition attempt. When it would not interfere with the camera's line of sight, the gaze indicator light on the headrest was used to help the test subject keep a forward gaze.

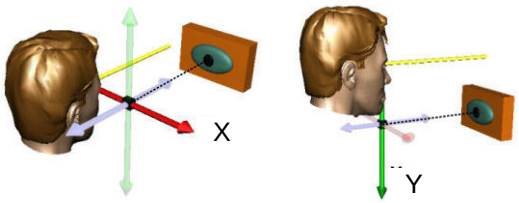


Figure 4-8: Translate Procedures

Step 6a. Translate X

Translate X emulates a user looking directly forward and positioned to the side of the product. At each of the one to three Z distances, the product was translated along the X axis in increments. Recognition attempts were performed at each of the increments.

Step 6b. Translate Y

The Translate Y procedure emulates a user looking directly forward and positioned below or above the product. At each of the one to three Z distances, the product was translated along the Y axis in increments. Recognition attempts were performed at each of the increments.

4.5 BETH for Off-Axis Experiment

The biometric evaluation test harness (BETH) used for the controlled off-axis experiment is comprised of the same tools and systems used in the scenario evaluation as described in Section 3.4 with a few exceptions. This experiment used a different set of tools for analysis of the online results and did not include offline analysis. In addition, this experiment used the interactive 3D view, described in Section 4.2, for realtime feedback of apparatus measurements.

Data collection tools

For each test subject, the test administrator selected the product, procedure, and iris-feature set, and performed the required recognition attempts for that procedure. BETH also recorded the measured 3D offset between the test subject and the product (Subject Offset) for each recognition attempt.

Online results analysis tools

The results from this evaluation were summarized and analyzed using a spreadsheet and graphing tools. The experimenter exported the raw results from the database into a well-prepared spreadsheet and organized the data according to test subject and procedure. The experimenter reviewed the results for each procedure for each test subject and used formulas and summary tools to generate the results graphed in Section 6.2.2 below.

5. Data Analysis Techniques

In this section we describe the basic biometric performance metrics and how we determine the statistical significance of the results by estimating confidence intervals. The data analysis techniques used for the scenario evaluation and for the off-axis controlled pose experiment are also presented in detail.

5.1 Biometric Performance Metrics

Before delving into the specific data analysis techniques used for the scenario evaluation and off-axis pose experiment, it is important to have a general understanding of biometric performance metrics and how these metrics were applied in the IRIS06 study. There are two main types of performance metrics: error rates and transaction times. The values we obtain for these metric types depend profoundly on how we define the biometric transactions, which are characterized by the level-of-effort and the biometric features that are taken into consideration.

For IRIS06, three types of biometric feature sets were considered: 1) right eye only, 2) left eye only, and 3) left or right eye. Online data were collected for the right-eye, left-eye, and left-or-right-eye feature sets. Additionally, the combined performance of the right and left eyes was investigated during offline data analysis.[27] Various types of single-attempt and multiple-attempt levels-of-effort were also explored. The application of these principles to the IRIS06 error rate and transaction time performance metrics are explored in the following sections.

Three types of iris-feature sets:

- Left eye only
- Right eye only
- Left or right eye

5.1.1. Error rates

Basic error rates, generalized error rates, and levels-of-effort for the various IRIS06 iris-feature sets and decision polices are defined and described below. Basic error rates highlight specific errors that can occur in a biometric system, such as failure to enroll, failure to acquire, or failure to match a biometric sample. In contrast, generalized error rates incorporate errors across the whole system so that real-world performance can be better understood. Generalized error rates allow us to estimate the percentage of the population that will not be able to utilize the biometric system.

[27] In this offline combined-performance case, the left and right eyes from one individual are treated as if they originated from two different individuals.

Basic error rates

- Failure to Enroll rate (FTE) – proportion of the test population that failed to enroll the selected feature set. For the right-or-left-eye feature set, a failure to enroll means that both the left and right eye failed to enroll for a given test subject. An enrollment transaction consisted of three 90-second attempts to enroll at least one eye.
- Failure to Acquire rate (FTA) – proportion of the enrolled test population for which the system failed to capture a feature set during a recognition (verification or identification) attempt. FTA includes only those feature sets that were previously successfully enrolled. (FTE rates are incorporated in the generalized error rate metrics described below.) For the right-or-left-eye feature set, a failure to acquire means that either the left or right eye enrolled and that both the left and right eye failed to acquire for a given test subject.
- False Non-Match Rate (FNMR) – proportion of the enrolled test population for which the system acquired but failed to recognize (verify or identify) a feature set during a recognition attempt. FNMR includes only those feature sets that were previously successfully enrolled. FTE and FTA rates are not incorporated in FNMR but are incorporated in the generalized error rate metrics described below. For the right-or-left-eye feature set, a false non-match means that either the left or right eye enrolled, that one or both of the enrolled eyes acquired, and that both the left and right eye failed to match for a given test subject.
- True Match Rate (TMR=1-FNMR) – proportion of the enrolled test population for which the system acquired and succeeded to recognize (verify or identify) a feature set during a recognition attempt. TMR is a "success rate" as opposed to an "error rate" and is commonly plotted in Receiver Operating Characteristic (ROC) curves as discussed in Section 5.2.5.
- False Match Rate (FMR) – proportion of the users not enrolled in the system (impostor test population) for which the system acquired a feature set during a recognition attempt and falsely declared that feature set to match an enrolled feature set. FTA rates are not incorporated in FMR but are incorporated in the generalized error rate metrics described below. For the right-or-left-eye feature set, a false match means that either the left or right eye acquired and that either the left or right eye falsely matched for a given test subject.

Levels of effort

In the IRIS06 data analysis, FNMR and FMR are reported for several types of single-attempt and multiple-attempt levels-of-effort.

The first type is a ***simple single attempt***, such as the first, second, or third attempt in any one of the four recognition transactions (Verify 1, Verify 2, Identify 1 and Identify 2). There are 12 simple-single-attempts (3 attempts x 4 transactions) in the IRIS06 study for each of the iris-feature sets. For simple single attempts, only one genuine match score for each iris-feature set (that is, for each test subject) is included in the analysis.

Also of interest are combinations of the simple single attempts. For example, we may wish to know the combined FNMR and FMR for all of the Verify 1 attempts or for all 12 simple-single-attempts or perhaps for all first attempts of each transaction. Here each attempt is treated as a separate member of the test population in the FNMR and FMR calculations. While multiple attempts from each test subject are included in this approach, the resulting FNMR and FMR values represent the average single-attempt performance over all considered attempts for all test subjects. This type of ***combined-single-attempt*** analysis is common in the biometrics community. It has the advantage of including multiple match scores per iris-feature set (per test subject), which effectively "smoothes" the resulting curves by increasing the signal-to-noise ratio. In addition, the statistical significance of the result may be improved if the match scores for a given test subject are independent, that is, if the intra-test-subject comparison-sample correlation is low. (We have not found this to be the case for iris recognition.) Statistical significance and comparison sample correlations are explored in detail in Section 5.3.

Another level-of-effort approach that is particularly operationally relevant is ***cumulative multiple attempts***. For example, the cumulative FNMR and FMR after one attempt, after two attempts, and after three attempts for each of the four recognition transactions can be computed. FNMR for the second attempt would include the proportion of the enrolled test population that failed to recognize during both the first and second attempts, and FNMR for the third attempt would include the proportion of the enrolled test population that failed to recognize during all three attempts. This approach emulates performance in many real-world applications where a user is given three tries to be successfully recognized. For cumulative multiple attempts, as with simple single attempts, only one match score for each iris-feature set (that is, for each test subject) is included in the analysis.

The final level-of-effort approach is ***combined cumulative multiple attempts***. For example, combining the cumulative multiple attempt results for each of the four recognition transactions would provide the overall cumulative multiple attempt performance for the selected

iris-feature set and product, which effectively represents the average real-world performance over multiple transactions. For combined cumulative multiple attempts, as with combined single attempts, multiple match scores for each iris-feature set (that is, for each test subject) are included in the analysis.

During offline analysis, we include a special case where both the left and right-eye features sets are included individually in the analysis. In this case, the size of the test population is double the number of test subjects. Each eye is treated as a separate test subject.

Generalized error rates

FNMR and FMR do not incorporate FTE and FTA results. In systems that readily accept poorer quality images, FTE and FTA may be quite low but the resulting FNMR and FMR may be quite high. Alternatively, systems that accept only high quality images may have high FTE and FTA but correspondingly lower FNMR and FMR. Clearly, there are tradeoffs between FTE/FTA and FNMR/FMR. As such, generalized error rate equations that take both components into account allow us to compare systems more readily. The generalized error rate equations for verification and identification transactions are provided below. When the cumulative-third-attempt FNMR and FMR values are used to compute the generalized error rates, results represent the three-strikes-and-you're-out recognition transaction performance.

Verification

- Generalized False Reject Rate (GFRR) – proportion of verification transactions with truthful claims of identity that are incorrectly denied

  $$GFRR = FTE + (1 - FTE) \times FTA + (1 - FTE) \times (1 - FTA) \times FNMR$$

- Generalized True Accept Rate (GTAR=1-GFRR) – proportion of verification transactions with truthful claims of identity that are correctly confirmed
- Generalized False Accept Rate (GFAR) – proportion of verification transactions with wrongful claims of identity that are incorrectly confirmed

  $$GFAR = (1 - FTA) \times FMR$$[28]

[28] In this equation, FTA refers to the proportion of the impostor test population for which the system failed to capture a feature set during a recognition attempt. Our equation for GFAR differs from that published in IS0/IEC 19795 (GFAR = (1 – FTE) x (1 – FTA) x FMR) in that we do not include FTE. To emulate real-life operation of a biometric system, we choose to compare impostor feature sets only with features sets of users that are enrolled in the system.

Identification

- Generalized False-Negative Identification-Error Rate (GFNIR) – proportion of identification transactions in which the user's correct identifier is not among those returned

  $$GFNIR = FTE + (1 – FTE) \times FTA + (1 – FTE) \times (1 - FTA) \times FNMR$$

- Generalized False-Positive Identification-Error Rate (GFPIR) – proportion of identification transactions by users not enrolled in the system where an identifier is returned

  $$GFPIR = (1 – FTA) \times (1 - (1 - FMR)^N),$$
  where N is the number of iris-feature sets in the database

These generalized error rate equations allow us to estimate the percentage of the population that will not be able to successfully utilize the biometric product, and thus how many individuals will need to utilize a secondary or backup system. Note that the equations for GFRR and GFNIR are identical. FMR in the GFAR equation is replaced by $(1 - (1 - FMR)^N)$ in the GFPIR equation. This adjustment presumably takes into account the fact that for verification a non-enrolled user is compared to one randomly selected enrolled user while for identification a non-enrolled user could be compared to all enrolled users.

In reality, we compute performance metrics from the raw data, as described in Section 5.2.3, as opposed to applying the equations above directly.

5.1.2. Transaction times

Transaction times are another important biometric performance metric. We anticipate that faster transaction times might lead to higher error rates; and slower transaction times, to lower error rates. As such, we can expect a tradeoff between transaction times and error rates. For IRIS06, transaction times were calculated as the sum of (up to three) attempt durations required to successfully complete the transaction. Recall that attempt durations were measured as the time difference between executing the BioAPI function call, such as BioAPI_Enroll, and receiving the BioAPI function call response. If the transaction was unsuccessful, the transaction time is the sum of the three attempt durations. Recall that the products did not always honor the timeouts specified by BETH during data collection; sometimes products terminated attempts in less than 20 seconds.

For enrollment, the transaction time is computed only for the right-or-left-eye feature set in accordance with the IRIS06 test protocol. The enrollment transaction time is defined as the

sum of the attempt durations required to successfully enroll at least one eye or the sum of the three attempt durations if enrollment was unsuccessful. For two-eye cameras, both eyes were presented simultaneously. For one-eye cameras, left and right eyes were presented consecutively as directed by BETH. As such, one-eye camera enrollment transaction times could be twice as long as those for two-eye cameras.

For recognition, transaction times were computed for the right-or-left-eye feature set, and for the right-eye and left-eye feature sets where possible. For the single-eye feature sets, the recognition transaction time is defined as the sum of the attempt durations required to successfully match the eye or, if unsuccessful, the sum of the three attempt durations. For the right-or-left-eye feature set, the recognition transaction time is defined as the sum of the attempt times required to successfully match at least one eye or, if unsuccessful, the sum of the six attempt durations.

For two-eye cameras, where both eyes are presented simultaneously, the right-or-left-eye recognition transaction time is clearly defined. However for one-eye cameras, where the left and right eyes are presented consecutively, a decision policy defining the presentation order of the eyes is required to clearly define the transaction time. For example, a user could present the one eye, and then present the other eye, and then continue to alternate between the two eyes up to three times until one of the eyes is recognized. Alternatively a user could present one eye up to three times in a row and if not recognized present their other eye up to three times in a row. Recall that our test protocol prescribed that the left and right eyes were presented randomly to minimize habituation effects. This was intentional so that the relative performance of right and left eyes could be examined independently. We therefore constructed the IRIS06 recognition transaction times by adding BioAPI attempt durations in the order prescribed by the eye order decision policy until the user was recognized. We felt that operationally, users were more likely to try three times with one eye, and then try with the other eye if the first eye was unsuccessful (as opposed to shifting back and forth between right and left eyes). As such, we constructed recognition transaction times based on right/right/right/left/left/left eye-order attempts and on left/left/left/right/right/right eye-order attempts. The transaction times we present in the results are the average of these two cases.

5.2 Scenario Evaluation

As noted in Section 3, the purpose of the scenario evaluation data collection was to measure online (real-time operation) performance metrics and to collect images for a more thorough offline analysis. Online and offline data analysis techniques are described below. The data review process, which must be performed prior to data analysis, is also discussed.

5.2.1. Data review

Data review is performed with the data review tools described in Section 3.4.3 above to 1) identify human errors and 2) flag image properties of interest, such as eyes-closed, out-of-focus, or glasses-on.

While the greatest of care is taken during data collection to minimize the potential for human errors, human errors do occur. Examples include presenting the right eye to the camera when the left eye was expected, or incorrectly entering a test subject's UIN. The data review tools can display each biometric sample (each iris image in this case) in isolation or in groups of interest along with associated test subject information and statistics. For example, the left and right eyes from the same test subject taken on all cameras can be displayed side by side to look for left-eye/right-eye presentation errors. The tools also flag extremely poor genuine match scores and extremely good impostor match scores for further investigation. When errors are detected, the erroneous data elements are flagged for exclusion during data analysis.

When reviewing images, performance-relevant image properties, such as glasses-on, are flagged and explanatory notes can be attached to each image. The image-property flags, which we call "hints," can be used to filter the data set during subsequent offline data analysis. For example, we can investigate the influence of glasses by generating performance curves for test subjects wearing glasses and for test subjects not wearing glasses.

Data review is typically conducted by the experimenter on an ongoing basis throughout the data collection process so that problems can be identified and corrected in a timely fashion. Every single image is examined by the Experimenter. The data review process is required to ensure that performance-relative image properties are flagged, that ground truth is known, that the collected data is valid, and that the measured performance of the products under test is not adversely influenced by human errors made by members of the test team.

The image properties flagged during data review for the IRIS06 images include: glasses (eyeglasses present in image), hard contacts (hard contacts present in image), bad environment (dark images), bad feature (pupil and iris indistinguishable), obstructions (e.g., eyelashes in pupil area, eyeglass frames covering part of iris, hair in eye, hat brim obscuring eye), bad picture (e.g., out of focus, extremely large pupil), bad placement (e.g., image is a non-eye feature, such as hair, nostril, hat, glasses, bridge of nose, etc., eye closed, droopy eyelids, eyes squinting or blinking, image off center or out of frame). Unless specifically noted, these images ***are*** included in data analysis since these are the actual images that the cameras collected and used for enrollment and recognition attempts.

The IRIS06 data review revealed 81 human errors that occurred at the attempt level in 9,498 total attempts, which corresponds to a human error rate of 0.85%. Two types of human errors occurred. The first was incorrect eye presentation (for example, the left eye presented when the right eye expected). Though the test administrator, the test observer, and the test subject all try hard to ensure correct eye presentation, errors still occur. The second type of human error was incorrect data entry of UIN. While a barcode reader was provided to scan the UIN, the test team members realized that manual entry of the UIN was slightly faster than automated entry with barcode reader. As such, they sometimes chose to enter the UIN manually instead of using the barcode reader, and errors occurred. In future evaluations, manual UIN entry will be disabled.

5.2.2. Data exclusions

Iris images that were either not collected or collected improperly were excluded from the error rate, confidence intervals, and transaction time calculations. The exclusion cases that arose during IRIS06 and how they were handled are described below:

- When a product was not available to a test subject (for example, the product that arrived late was not available to the initial test subjects during the first visit), that test subject's UIN was excluded from that product's online and offline analysis for the affected procedure.
- Test subjects that did not return for the second visit or whose UIN was entered incorrectly during the second visit were excluded from online and offline analysis of second visit procedures for all products.

- Test subjects who presented an incorrect feature, such as the right eye when the left eye was expected, were excluded from online and offline analysis for that product and procedure.
- If a software license problem occurred during data collection, the affected test subject UIN was excluded from online and offline analysis for the affected product and procedure.
- If a software malfunction occurred during data collection, the affected test subject UIN was excluded from online and offline analysis for the affected product and procedure.

Exclusion tables were constructed to ignore excluded UINs for the affected product-procedure combinations. For transaction-based analyses, where multiple attempts are combined to produce a single result, if a UIN-product-procedure exclusion occurred anywhere in that transaction, the UIN was excluded from that entire transaction. As such, the total number of UINs considered for each product-procedure combination vary depending on the number of exclusions for that combination.

5.2.3. Numerical methods

For both online and offline analyses, we calculate basic error rates (FTE, FTA, FNMR, and FMR) using raw data. That is, the total number of errors (failures to enroll, failures to acquire, failures to match, or false matches) are simply tabulated and divided by the total number of tries (to enroll, to acquire, or to match) at each threshold of interest. The raw data count approach is valid for all level-of-effort approaches (simple-single-attempt, cumulative-multiple-attempt, combined-single-attempt, and combined-cumulative-multiple-attempts approaches) as described in Section 5.1.1. The difference lies in how the tries and the errors for each try are defined for each level-of-effort approach.

Similarly, we calculate generalized error rates (GFRR and GFAR) using raw data counts as opposed to the equations presented in Section 5.1.1. The results are the same with both approaches, however using the raw data directly improves numerical precision and is computationally simpler. For example, assume 100 test subjects try to enroll and 10 fail (FTE=10/100=10.0%), then 90 test subjects attempt recognition and 10 fail to acquire (FTA=10/90=11.$\overline{1}$%), then 80 test subjects attempt to match and 10 fail (FNMR=10/80=12.5%). Then GFRR is the total number of errors (10+10+10=30) divided by the total number of tries (100), GFRR=30/100=30%. Using the equation in Section 5.1.1:

$$GFRR = FTE + (1 - FTE) \times FTA + (1 - FTE) \times (1 - FTA) \times FNMR$$

$$= 0.100 + (1\text{-}0.100) \times 0.\overline{1} + (1\text{-}0.100) \times (1\text{-}0.\overline{1}) \times 0.125 = 29.\overline{9}\ \%.$$

Similarly for GFAR, assume 100 impostor test subjects attempt to be recognized by the biometric system, 10 fail to acquire (FTA=10/100=10%) and 2 erroneously match (FMR=2/90=2.$\overline{2}$ %). Then GFAR is the total number of errors (2) divided by the total number of tries (100), GFAR=2/100=2%. Using the equation in Section 5.1.1:

$$GFAR = (1 - FTA) \times FMR = (1\text{-}0.1) \times 0.0\overline{2} = 1.\overline{9}\ \%.$$

The small numerical errors obtained when using the formulas can be minimized by using double precision arithmetic; however the direct count method is substantially faster and easier to perform via computer than are the equations.

5.2.4. Online analysis

For each product, the online (real-time) native performance metrics, namely error rates and transaction times, were compiled from the raw data returned from the BioAPI commands.

The error rates available online are FTE, FTA, FNMR, GFRR as described in Section 5.1.1. Each of these error rates is tabulated for the right-eye-only, left-eye-only, and right-or-left-eye feature sets. The FNMR and GFRR metrics reflect each product's internal operating point (threshold) as set by the manufacturer, such as a Hamming distance of 0.32. (The influence of different thresholds on performance is investigated in the offline analysis.) We compute the error rates for each attempt (on a cumulative basis) and for each transaction. We compute the average transaction error rates for each visit and overall. For one of the cameras, FNMR and GFRR could be computed only for the left-or-right-eye feature set due to the internal functioning of the product's BSP.

The online enrollment and recognition transaction times (average times and histograms) are also computed for each transaction as defined in Sections 5.1.2. The average recognition times per visit and overall are also computed.

5.2.5. Offline analysis

While the online analysis provides false match results for only one threshold setting, offline analysis allows us to study false match and false accept performance as a function of

threshold. In addition, offline analysis allows us to study the interoperability of images collected with different camera systems and the performance under various conditions, such as age, gender, race, glasses, and eye gaze direction. The images collected per the IRIS06 test protocol are used for offline analysis. Templates are created for each image, and the templates are selectively compared (matched) per the desired filtering parameters listed in Table 3-2 above. Professor John Daugman's "irisenroll (release 1.5)" template generator and "matcher1" template matcher is used for offline analysis.

Offline recognition performance results are typically reported in the form of Detection Error Tradeoff (DET) or Receiver Operating Characteristic (ROC) curves, which are described in detail below. To obtain these curves, we must first generate a similarity matrix with the biometric enrollment and recognition samples of interest. In the simulated similarity matrix shown in Table 5-1, match scores are generated for all of the enrollment samples from a Product N compared to all of the recognition samples from a Product M[29] for each test subject in the database (assuming one feature set per test subject) using a Matching Algorithm A.[30] In general, biometric samples could be biometric templates, models, images, recordings, etc. For IRIS06, the primary biometric samples are ISO-compliant iris images.

For the example illustrated in Table 5-1, Comparison Algorithm A produces a higher score when the two compared biometric samples are more similar and a lower score when the samples are less similar. (Note that the opposite is the case for the Daugman iris matching algorithm.) The diagonal elements in the simulated similarity matrix (denoted by green numbers in Table 5-1) indicate the genuine scores. For example, Person 3's enrollment sample compared to Person 3's recognition sample results in a score of 72. For the example Matching Algorithm A, genuine match scores should be high. The off-axis elements in the example similarity matrix (denoted by red numbers in Table 5-1) are cross-comparison impostor scores, for example,

[29] If the enrollment and recognition samples are from the same product (M=N), results indicate native performance. If the enrollment and recognition samples are from different products (M≠N), results indicate interoperability performance.

[30] Depending on the technology, templates or models may be generated from the samples, which are then compared to obtain the match scores presented in the similarity matrix.

Person 6's enrollment sample compared to recognition samples from all other persons. [31] These impostor match scores should be low for Algorithm A.

Table 5-1. Simulated Similarity Matrix showing Match Scores

Matching Algorithm A		Product N *Enrollment* Samples							
		Person 1	**Person 2**	**Person 3**	**Person 4**	**Person 5**	**Person 6**	**Person *n-1***	**Person *n***
Product M *Recognition* Samples	**Person 1**	**65**	**30**	**14**	**26**	**24**	**31**	**20**	**5**
	Person 2	**21**	**85**	**25**	**15**	**22**	**6**	**40**	**30**
	Person 3	**16**	**11**	**72**	**22**	**45**	**26**	**14**	**27**
	Person 4	**31**	**19**	**35**	**45**	**28**	**7**	**21**	**26**
	Person 5	**42**	**2**	**17**	**24**	**66**	**27**	**22**	**22**
	Person 6	**5**	**28**	**46**	**8**	**16**	**78**	**21**	**19**
	Person *n-1*	**18**	**33**	**22**	**27**	**38**	**42**	**58**	**31**
	Person *n*	**24**	**10**	**9**	**17**	**22**	**31**	**18**	**70**

The next step is to use the scores in the similarity matrix to generate histograms of the genuine and impostor match score distributions as illustrated in Figure 5-1. The green bars in Figure 5-1 indicate the number of genuine comparisons that achieved each score, or the genuine score distribution. Similarly, the red bars indicated the impostor score distribution. Ideally, the genuine and impostor score distributions would be distinct, with no overlap. A threshold match score could then be set in the biometric system such that all impostors are rejected and all genuine attempts are granted. In reality, there is almost always some overlap between the genuine and impostor distributions. The histogram helps us understand the performance that will

[31] Ideally, impostor scores would be generated by comparing each sample with a non-self sample only once. For example, Person 3 is compared only to Person 4 and Person 5 is compared only to Person 6, etc. This is possible in large-scale offline evaluations where many samples are available. However, for smaller scenario tests we allow full cross comparisons and address the correlation of impostor comparison scores when determining confidence intervals. This is standard practice as documented in the international standard for biometric performance testing, ISO/IEC 19795-1.

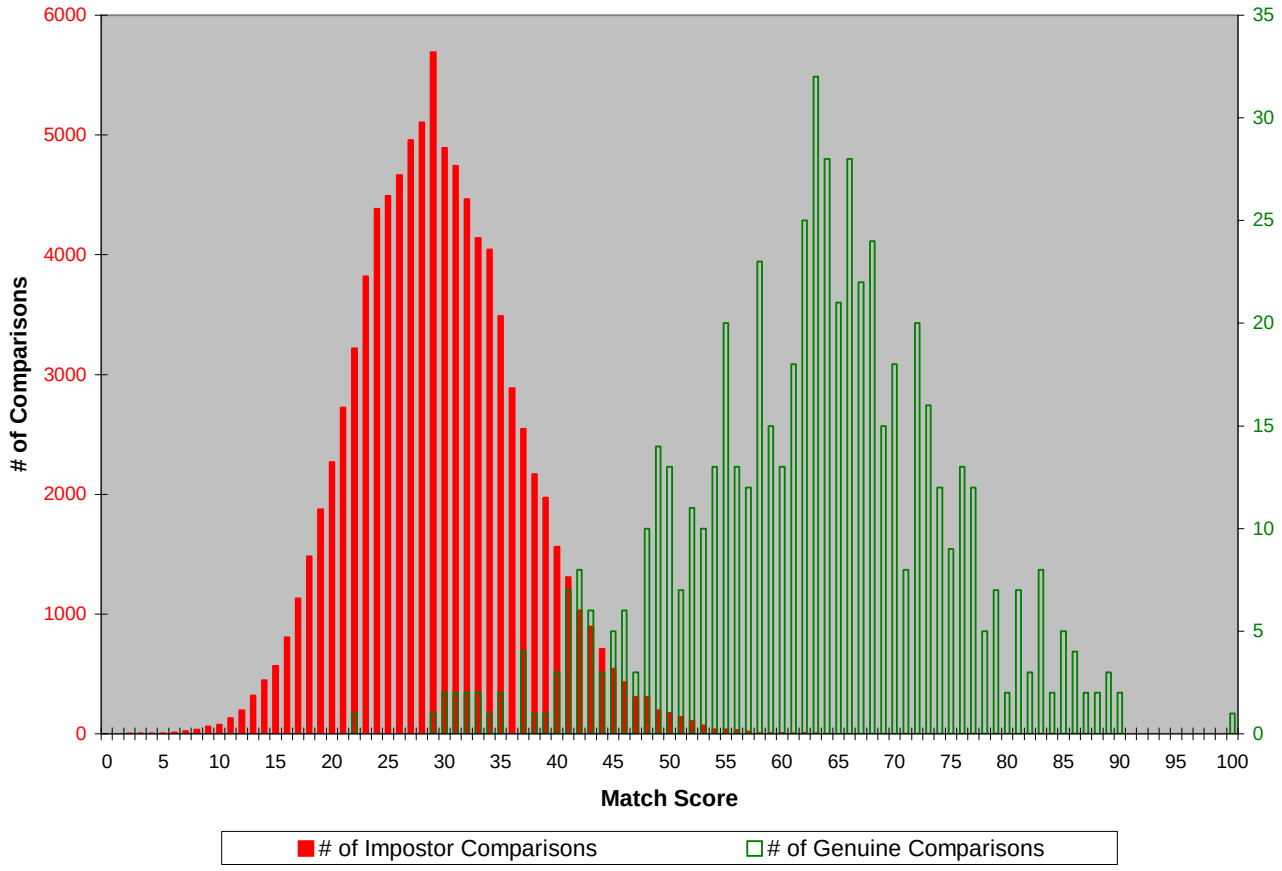


Figure 5-1. Example Histogram of Genuine and Impostor Comparison Score Distributions

be achieved depending on where we set the threshold. For example, if we place the threshold at match score = 55 in Figure 5-1, most of the impostors will be rejected, and we will achieve a low false match rate (FMR). However, a good portion of the genuine scores will also be rejected resulting in a high false non-match rate (FNMR). Conversely, if we set the threshold at match score = 25, most of the genuine scores will be accepted yielding a low FNMR. However, many impostor scores will be accepted resulting in a high FMR. These tradeoffs between security (low FMR) and convenience (low FNMR) must be considered for each individual implementation of biometric technology.

To quantify FNMR and FMR performance as a function of threshold score, DET or ROC curves are generated. Each point on a DET or ROC curve represents a different threshold match score (operating point) on the genuine and impostor distribution histogram in Figure 5-1. Recall that for our example, higher match scores indicate that the biometric samples are more similar. Then for a given threshold score, the number of genuine scores that occur below the selected

threshold score divided by the total number of genuine scores is the FNMR. This represents the recognition attempts that match but will be rejected at that threshold score. Similarly, the number of impostor scores that occur above that same selected threshold score divided by the total number of impostor scores is the FMR. This represents the recognition attempts that should be rejected but will be accepted as matches at that threshold score. These two values become one point on the DET curve. The process is then repeated for all other threshold scores of interest. Simulated DET and ROC curves are presented in Figure 5-2.

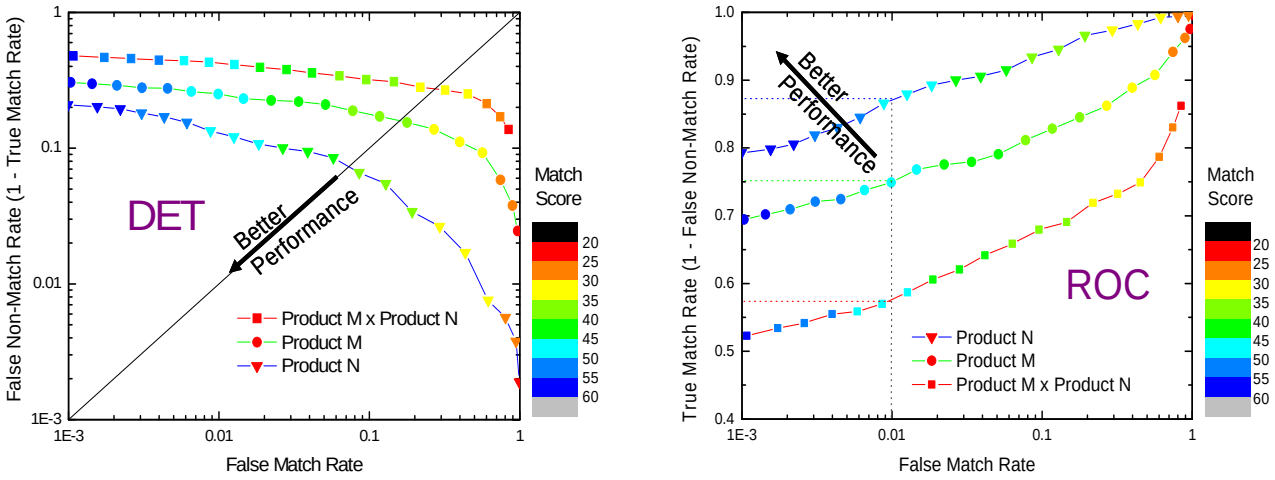


Figure 5-2. Simulated DET and ROC Curves

In reality, it is not necessary to construct a genuine and impostor distribution histogram. The FNMR and FMR at various thresholds can be computed directly using the data from the similarity matrix. However, the histogram provides a valuable visualization tool to understand the meaning of the DET and ROC curves.

For DET curves, false match error rate is plotted on the horizontal axis using a $\log_{10}$ scale and false non-match error rate is plotted on the vertical axis using a $\log_{10}$ scale. Since we want the error rates to be lower, curves that are closer to the origin of the graph (lower left corner) indicate better performance (lower error rates or higher accuracy). The intersection of the performance curves and the diagonal line in the DET plot indicates the equal error rates where FMR=FNMR.

For ROC curves, false match error rate is plotted on the horizontal axis using a $\log_{10}$ scale and the true match rate (TMR=1-FNMR) is plotted on the vertical axis using a linear scale. In this case, curves that are closer to the upper left hand corner of the graph indicate better

performance (higher accuracy). The vertical axis can be interpreted as the probability that a genuine match will be detected, or the probability of detection.

DET and ROC curves allow decision makers to select the most appropriate operating point (threshold setting) based on the requirements of their particular system. Some prefer to see biometric performance results presented in ROC curves while others prefer DET curves. In this report, we plot ROC curves.

The simulated performance curves in Figure 5-2 indicate that the native performance of Product N is better than that of Product M, and that interoperable performance (Product M enrollment images compared to Product N recognition images, denoted Product M x Product N) is poorer than the native performance of both Products M and N.

The points on the curves are color coded to represent the threshold score that will result in the performance at each point, as indicated in the Match Score legend. If we choose an operating point of FMR=0.01 (1%), which is typical in many operational scenarios, then the probability of detection is roughly 87.5%, 75.0%, and 57.5% for Product N, Product M, and Product M x Product N, respectively, for this simulated example. Further, the threshold match score required to achieve this performance is between 45 and 50 (from the color legend for turquoise symbols) for all three cases. A more precise threshold score can be obtained from the data used to construct the graph.

For IRIS06, basic (FTE and FTA not included) and generalized (FTE and FTA included) ROC curves are plotted for a variety of levels-of-effort and with a variety of filters of interest, such as “with and without glasses” and up, down and sideways gaze directions.

Results of the online and offline scenario analyses are presented in Sections 6.1.1 and 6.1.2, respectively. These results indicate the estimated performance of the iris recognition product under normal operating conditions. Results of the offline scenario analysis performed with off-axis gazes, presented in Section 6.2.1, are indicative of performance that might be obtained with uncooperative users.

5.3 Uncertainty Estimates

It is important to estimate uncertainty in the measured biometric performance so that the level of confidence in the results can be specified. In particular, we wish to understand the degree to which the results obtained with our test population are representative of a larger population. We estimated uncertainties for the binary matching data (match/no-match) and for transaction times (number of seconds) in the form of 95% confidence intervals.

While transaction times are not normally distributed, we assume that the central limit theorem is valid in our case and that normal distribution-based descriptive statistics can be used. The central limit theorem asserts that when sample size becomes large (e.g., n>100), then the sample mean will follow the normal distribution even if the respective variable is not normally distributed in the population. We used the descriptive statistics tools in Origin 7.5 SR6[32] to calculate 95% confidence intervals for transaction times.

We use two different methods to estimate confidence intervals for matching decisions (match/no-match).[33] When the match decisions (comparison samples) are uncorrelated (one comparison sample per person[34]), we use the adjusted Wald method to estimate confidence intervals. [35] When the match decisions (comparison samples) are correlated (multiple comparison samples per person[34]), we use the Logit Beta-binomial to estimate confidence intervals.[36]

5.3.1. Uncorrelated comparison samples

Per the advice of Dr. Michael Schuckers at St. Lawrence University,[37] when no information on the correlation between comparison samples is available, that is, when only one single match decision is available for each iris-feature set (as is the case for our simple-single-attempt and cumulative-multiple-attempt analyses), we used the adjusted Wald method for determining confidence intervals as recommended by Agresti and Coull.[35] This method is based

[32] http://www.OriginLab.com.
[33] Note that since confidence intervals are computed only at specific thresholds, if a match score is provided (as opposed to a binary match decision), the specified threshold is applied to that score to generate a match decision.
[34] "Person" and "iris-feature set" are used interchangeably when appropriate in an effort to make the discussion easier to understand.
[35] A Agresti and BA Coull, "Approximate is better than 'Exact' for interval estimation of binomial proportions," The American Statistician **52**, 119 (1998).
[36] See Dr. Michael Schuckers' manuscript entitled "Estimation and sample size calculations for matching performance of biometric authentication," (http://it.stlawu.edu/~msch/biometrics/papers/Schuckers-PatternRecognition.pdf, accessed 1 September 2007) for a thorough discussion of estimating confidence intervals for biometric data.

on approximating the sampling distribution of the binomial proportion with a normal distribution, as justified by the central limit theorem. In this method, 2 successes and 2 failures are added to the dataset. For example, FNMR = p = F / N, where p is the proportion of failures to match, F is the number of failures to match, and N is the number of attempts to match (or number of comparison samples).[38] Using the adjusted Wald method, the lower and upper values for the 95% confidence interval for p or FNMR[39] are defined as:

$$\text{lower limit} = p' - 1.96 \times \sqrt{\frac{p'(1-p')}{N+4}}, \quad \text{upper limit} = p' + 1.96 \times \sqrt{\frac{p'(1-p')}{N+4}},$$

$$\text{where } p' = \frac{\#\text{ of failures} + 2}{\#\text{ of feature sets} + 4} = \frac{F+2}{N+4}$$

and 1.96 is the z-score for 95% confidence intervals.[40]

Note that the confidence interval is not centered around p but around p', which is closer to 0.5 than p. Agresti and Coull have shown that the adjusted Wald method performs well for all values of N and p and provides intervals that are closer to 95% confidence than the so-called "exact" method of estimating confidence intervals. Computer simulations by several investigators have demonstrated that "exact" confidence intervals are wider than they need to be and thus typically provide greater than 95% confidence.[41] While p in the example above represents FNMR, these equations are equally valid for FMR, FTE, and FTA when one binary comparison decision is available for each iris-feature set.

5.3.2. Correlated comparison samples

Again per the advice of Dr. Schuckers, when information on the correlation between comparison samples is available we used the Logit Beta-binomial approach for determining confidence intervals. This is relevant when more than one match decision (multiple comparison samples) is available for each iris-feature set, as is the case for our combined-single-attempt and combined cumulative-multiple-attempt analyses.

---

[37] http://it.stlawu.edu/~msch/biometrics/ (accessed 1 September 2007).

[38] For the right-eye-only, left-eye-only, and right-or-left-eye feature sets, N is also the number of test subjects since we are using one sample per test subject. N is also the number of iris-feature sets since there is one iris-feature set per test subject. For the offline, combined performance of the right and left eyes, N is twice the number of test subjects. In this case, left and right eyes from the same person are assumed to be uncorrelated and are treated as if they originate from different test subjects.

[39] This approach is also valid for p=FMR, where the number of genuine failures (F = # of failures) is replaced by the number of impostor successes.

[40] See, for example, http://web.uccs.edu/lbecker/SPSS/confintervals.htm (accessed 1 September 2007).

The Logit Beta-binomial confidence interval is derived from the moments of the Beta-binomial. The Beta-binomial distribution[42] is a generalization of the binomial distribution that allows for 1) correlation (lack of independence) between multiple comparison samples (multiple recognition attempts) for a given iris-feature set, and 2) variations in p (FNMR or FMR) for different iris-feature sets (e.g., for different test subjects). For both the Logit Beta-binomial and the Beta-binomial, p is determined as follows. Estimated parameters are denoted with a "hat".

Let n be the number of feature sets (typically the number of test subjects) and m_i the number of attempts for the i^{th} feature set. If X_{ij} is the error status for the i^{th} feature set for the j^{th} attempt (error=1, non-error=0), and X_i is the total number of errors for the i^{th} feature set, then

$$\hat{p}_i = \frac{\sum_{j=1}^{m} X_{ij}}{m_i} = \frac{X_i}{m_i}$$

is the mean number of errors for the i^{th} feature set and the mean error rate for all feature sets is

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} m_i}.$$

The denominator of this equation is the sample size, which can also be represented as $n\bar{m}$, where $\bar{m}$ is the mean number of attempts for all feature sets

$$\bar{m} = \frac{\sum_{i=1}^{n} m_i}{n}.$$

For the Beta-binomial, the confidence interval for p is defined as[43]

$$\hat{p} \pm 1.96 \times \sqrt{\frac{\hat{p}(1-\hat{p})(1+(m_0-1)\hat{\rho})}{\bar{m}n}} \quad \text{where}$$

$$\hat{\rho} = \frac{BMS - WMS}{BMS + (m_0 - 1)WMS},$$

[41] http://www.graphpad.com/articles/CIofProportion.htm (accessed 1 September 2007).
[42] Schuckers, M E, "Using the beta-binomial distribution to assess performance of a biometric identification device," International Journal of Image and Graphics **3**, 523 (2003).
[43] Michael E. Schuckers, *et. al*., "A comparison of statistical methods for evaluating matching performance of a biometric identification device: a preliminary report" Proceedings of SPIE **5404**, 144 (2004).

$$m_0 = \overline{m} - \frac{\sum_{i=1}^{n}(m_i - \overline{m})^2}{\overline{m}n(n-1)},$$

$$BMS = \frac{\sum_{i=1}^{n} m_i (p_i - \hat{p})^2}{n-1} = \text{"between mean squares," and}$$

$$WMS = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(X_{ij} - p_i\right)^2}{n(\overline{m}-1)} = \frac{\sum_{i=1}^{n} m_i p_i (1-p_i)}{n(\overline{m}-1)} = \text{"within mean squares".}$$

An important feature to note is that the confidence interval is inversely proportional to sample size. The larger the sample size, the smaller the confidence interval. However the confidence interval also depends on ρ, which represents the degree of correlation or dependence between multiple feature-set comparison samples. Schuckers refers to ρ as the intra-*individual* correlation. We define ρ as the intra-*feature-set* correlation. If ρ is large, little additional independent information is provided with multiple comparison samples, and the confidence interval is larger. If ρ is small, multiple attempts provide additional independent information, and the confidence interval is smaller. Another way to look at this[44] is to consider the "effective" sample size as

$$\frac{n\overline{m}}{1+(\overline{m}-1)\rho},$$

which is equivalent to the number of independent observations in the data. In the limit as $\rho \to 0$, each attempt provides completely new information, the effective sample size is $n\overline{m}$, and the resulting confidence intervals are smaller. In the limit as $\rho \to 1$, each attempt provides no new information, the effective sample size is n, and the resulting confidence intervals are larger.

For the Logit Beta-binomial,[43] $\hat{p}$ and ρ are computed as for the Beta-binomial, but the confidence intervals are computed differently. In this approach, the log-odds of p or Logit(p) is computed:

[44] Travis J. Atkinson and Michael E. Schuckers, "Approximate Confidence Intervals for Estimation of Matching Error Rates of Biometric Identification Devices," *Biometric Authentication* (Springer, Berlin / Heidelberg, 2004) pp. 184-194. Also available at http://it.stlawu.edu/~msch/biometrics/papers.htm (accessed 1 September 2007).

$$\text{Logit}(p) = \log_e\left(\frac{p}{1-p}\right).$$

The transformed Logit lower confidence interval (L') and upper confidence interval (U') values are then defined as:

$$\text{L}' = \text{Logit}(\hat{p}) - 1.96 \times \sqrt{\frac{1+(m_0-1)\hat{\rho}}{\hat{p}(1-\hat{p})\overline{m}n}},\ \text{U}' = \text{Logit}(p) + 1.96 \times \sqrt{\frac{1+(m_0-1)\hat{\rho}}{\hat{p}(1-\hat{p})\overline{m}n}}.$$

The inverse Logit function is then used to transform L' and U' into the Logit Beta-binomial lower (LCI) and upper (UCI) confidence interval values:

$$\text{UCI} = \text{Logit}^{-1}(\text{U}') = \frac{e^{\text{U}'}}{1+e^{\text{U}'}} \text{ and } \text{LCI} = \text{Logit}^{-1}(\text{L}') = \frac{e^{\text{L}'}}{1+e^{\text{L}'}}.$$

The Logit Beta-binomial approach gives an asymmetric confidence interval because the Logit transformation is nonlinear.[45]

Studies have shown that the Logit Beta-binomial approach performs well when $np > 5$ or when $np > 0.5$ and $\rho < 0.1$ and $\overline{m} > 5$.[43] When these assumptions are valid during our analyses, we compute the Logit Beta-binomial 95%-confidence intervals for error rates. For cases where zero errors occur during correlated-comparison-sample analyses, the Logit Beta-binomial cannot be used, and we apply the Rule

Note that the Beta-binomial explicitly takes the animals in "Doddington's Zoo"[46] into account. Namely, individuals that are

- exceptionally unsuccessful at being accepted (chronic high false rejecters or "goats"),
- exceptionally vulnerable to impersonation (good false matchees or "lambs"), or
- exceptionally successful at impersonation (good false matchers or "wolves")

are treated as crucial elements of the data and are appropriately weighted when estimating ρ and p. Conversely, the binomial method, which treats the animals in the zoo with the same weight as all other individuals, does not account for the Doddington's Zoo phenomenon.

[45] Michael E. Schuckers and C. J. Knickerbocker, "Documentation for Program for Rate Estimation and Statistical Summaries (PRESS)" (2004), http://it.stlawu.edu/~msch/biometrics/downloads.htm (accessed 1 September 2007).
[46] G. R. Doddington, *et. al*., "Sheep, Goats, Lambs and Wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation" *Proc. of 5th Int'l. Conf. on Spoken Language Processing* (Australian Speech Science and Technology Association, Incorporated, Melbourne, 1998) pp. 608-611.

of 3,[47] where the upper 95% confidence interval value is approximately $3/n$. Note that since the IRIS06 comparison samples are highly correlated,[48] we use n in the denominator as opposed to $n\bar{m}$. For cases where confidence intervals cannot be computed with statistical significance, they are not shown.

5.3.3. Confidence interval interpretation

There is some disagreement regarding how to interpret 95% confidence intervals (CIs). Some believe that a 95% CI means that we can be 95% certain that the interval contains the true population parameter. While this is intuitive, it is not strictly correct. The precise and agreed upon definition per the statistics literature: If one generates many 95% CIs from many data sets, then we expect the CIs to include the true population parameter in 95% of the cases and not to include the true population parameter in the other 5%. Said in a slightly different way, if the same population is sampled on numerous occasions, and CI estimates are calculated on each occasion; the resulting CIs would bracket the true population parameter in approximately 95% of the cases.

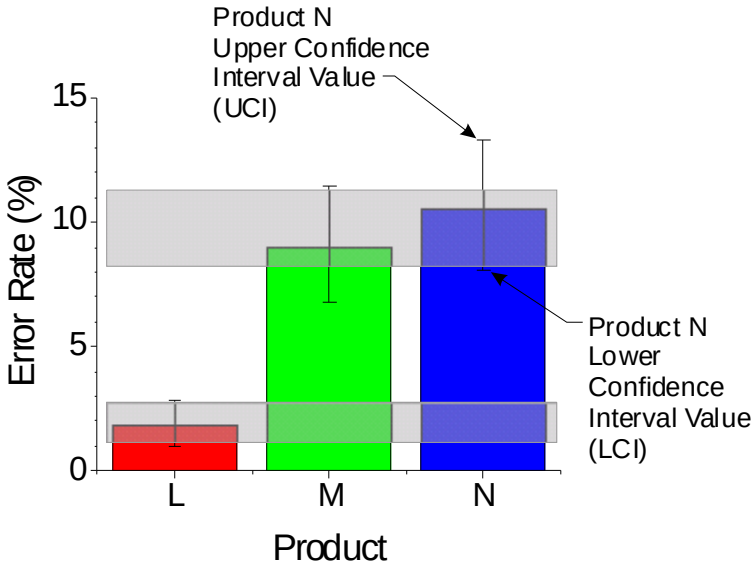


Figure 5-3. Confidence Interval Example

There is also some controversy regarding how to interpret the statistical significance of results from different parameters when the CIs for the parameters overlap and do not overlap. An example illustrating y-axis (Error Rate) CIs is presented in Figure 5-3. The CIs are indicated by the vertical lines with horizontal caps centered in each colored product column. For Products M and N, there is overlap between the CIs (as indicated by the upper gray shaded area). The CI for Product L, however, does not overlap with the CIs for Products M and N (as indicated by the lower gray shaded area).

[47] T.A. Louis, "Confidence intervals for a binomial parameter after observing no successes" The American Statistician **35**, 154 (1981).
[48] Detailed analysis of correlations between IRIS06 iris images is a subject for future study.

For the IRIS06 analysis, we will assume that if the CIs for different parameters overlap, the difference between the parameters is not statistically significant. Similarly, if the CIs for different parameters do not overlap, the difference between the parameters is statistically significant. For the example in Figure 5-3, we would conclude that the error rates for Products M and N are statistically similar and that the error rate for Product L is statistically lower than the rates for Products M and N. Of course, the reader is free to interpret the IRIS06 CIs as they see fit.

The adjusted Wald method and Logit Beta-binomial approaches described above for determining error rate CIs apply primarily to data collected during the scenario evaluation where sample size is reasonably large. For the controlled off-axis pose experiments, sample size is quite small, and we were unable to estimate uncertainties with any level of statistical confidence. However, the results of the controlled off-axis pose experiments do indicate performance trends that can be expected during off-axis use of different types of iris recognition camera systems.

5.4 Controlled Off-Axis Experiment

Performance with uncooperative users is further explored in the controlled pose off-axis experiment, where test subjects are placed in non-ideal locations for each camera. Performance is measured for each product using the Neutral, Sweep, and Translate procedures described in Section 4.4. Error rates from the Neutral, Sweep, and Translate procedures are plotted for each procedure using the reported offsets (X, Y, Z, Yaw, Pitch, Roll). The online (real-time) true match rate (TMR) is tabulated from the raw data returned from the BioAPI commands. To organize and plot TMR, results are grouped in offset bins such as 10-15 cm or 10-15 degrees. The procedures are designed to focus on one specific motion (for example, X translation at Z=10 cm, or Roll at Z=20 cm) so the offset bins are defined along that motion of interest for each procedure. For one-eye products performance is measured for each eye independently and for two-eye products the result for the combination of both eyes is used.

To tabulate the data, a simple spreadsheet is used to determine when a test subject was successfully recognized for a particular bin. A sample spreadsheet is shown in Table 5-2. Multiple attempts from the same test subject may be made within a single bin. In any given bin, a single successful recognition results in a match (1) for that bin, while a lack of any successful recognition attempts in a bin results in a non-match (0) for that bin. The percentage of successfully recognized test subjects is plotted for each bin.

Table 5-2. Sample Controlled Off-Axis Results Spreadsheet

Bin (distance or angle)	UIN				# of Matches
	101	102	103	104	
-10 to <-8	0	0	0	0	0
-8 to <-6	0	0	1	0	1
-6 to <-4	1	1	1	0	3
-4 to <-2	1	1	1	1	4
-2 to <0	1	1	1	1	4
0 to <2	1	1	1	1	4
2 to <4	1	1	1	1	4
4 to <6	1	1	0	0	2
6 to <8	0	1	0	0	1
8 to 10	0	0	0	0	0

6. Results

We present the IRIS06 on-axis and off-axis results below. The on-axis results include the online and offline analyses for the scenario evaluation. The off-axis results include the off-axis gaze experiment performed during the scenario evaluation (with about 250 test subjects) and the off-axis pose experiments performed with the specialized apparatus (with 6 test subjects).

6.1 On-Axis

The online (real-time) and offline biometric performance results for the scenario evaluation are presented in the following sections. The iris images collected during the online testing were saved for subsequent offline analysis.

6.1.1. Online results

The online error rates include FTE, FTA, FNMR, GFRR measured for each product during the data collection process. We also measured the enrollment and recognition times for each enrollment and recognition transaction. These results along with 95% confidence intervals are presented below.

Failure to enroll

The cumulative failure to enroll (FTE) rates for all three cameras and all three iris-feature sets are presented in Table 6-1, along with the average enrollment transaction time required to enroll at least one eye. Figure 6-1 presents these results graphically along with the 95% confidence intervals (CIs) for the 3rd-attempt FTE.[49]

Table 6-1. Cumulative Failure to Enroll (FTE)

	Product A			Product B			Product C		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
# of Persons	288			295			295		
Attempt 1	1.04%	0.69%	0.35%	7.80%	5.42%	4.07%	21.69%	21.02%	15.59%
Attempt 2	1.04%	0.69%	0.35%	5.76%	3.05%	1.36%	13.56%	13.22%	6.44%
Attempt 3	**1.04%**	**0.69%**	**0.35%**	**5.76%**	**2.37%**	**0.68%**	**11.86%**	**10.51%**	**3.39%**
Average Enrollment Transaction Time (sec)			40.4			32.2			70.1

[49] The upper and lower confidence interval values (UCI and LCI) are presented in tabular form in Appendix 11.3 for all IRIS06 online performance metrics.

We observe that the cumulative FTE rates decrease with increasing numbers of enrollment attempts for Products B and C. For Product A, FTE remains constant over all attempts. For Product A, if the test subject failed to enroll during the first attempt, they failed for all subsequent attempts.

We note that the FTE rates for Product C appear higher than the FTE rates for Products A and B. Figure 6-1 illustrates that $FTE_A < FTE_B < FTE_C$ for all iris-feature sets and all attempts. As shown in Figure 6-1b, the 3rd-attempt confidence intervals (CIs) for each product do not overlap significantly for the left-eye feature set, indicating that these results may be statistically significant. For the right-eye feature set, the 3rd-attempt CIs for Products A and B overlap significantly with each other but do not overlap with the CIs for Product C, indicating that the right-eye FTEs for Products A and B are statistically similar and lower than the FTE for Product C. For the left-or-right eye feature set, the CIs for Products A and B overlap substantially and both overlap slightly with the CI for Product C. This indicates that the difference in FTE between the products may not be statistically significant. That is, all products exhibit roughly the same FTE rates when successful enrollment is defined as "either the left or right eye enrolls successfully after three attempts with each eye".

Figure 6-1b also shows that the right-eye CIs for Product A significantly overlap the left-eye CIs for Product A, indicating that the left and right eyes exhibit roughly the same FTE rates for Product A. The CIs for the left and right eyes overlap somewhat for Product B and substantially for Product C, again indicating that the left and right eyes exhibit roughly the same FTE rates for Products B and C. We conclude that for each product, the left and right eyes exhibit roughly the same FTE rates after three attempts.

It is interesting to note that the one test subject that failed to enroll in Product A, the two test subjects that failed to enroll in Product B, and the ten test subjects that failed to enroll in Product C, were all different test subjects. No single test subject failed to enroll in more than one product.

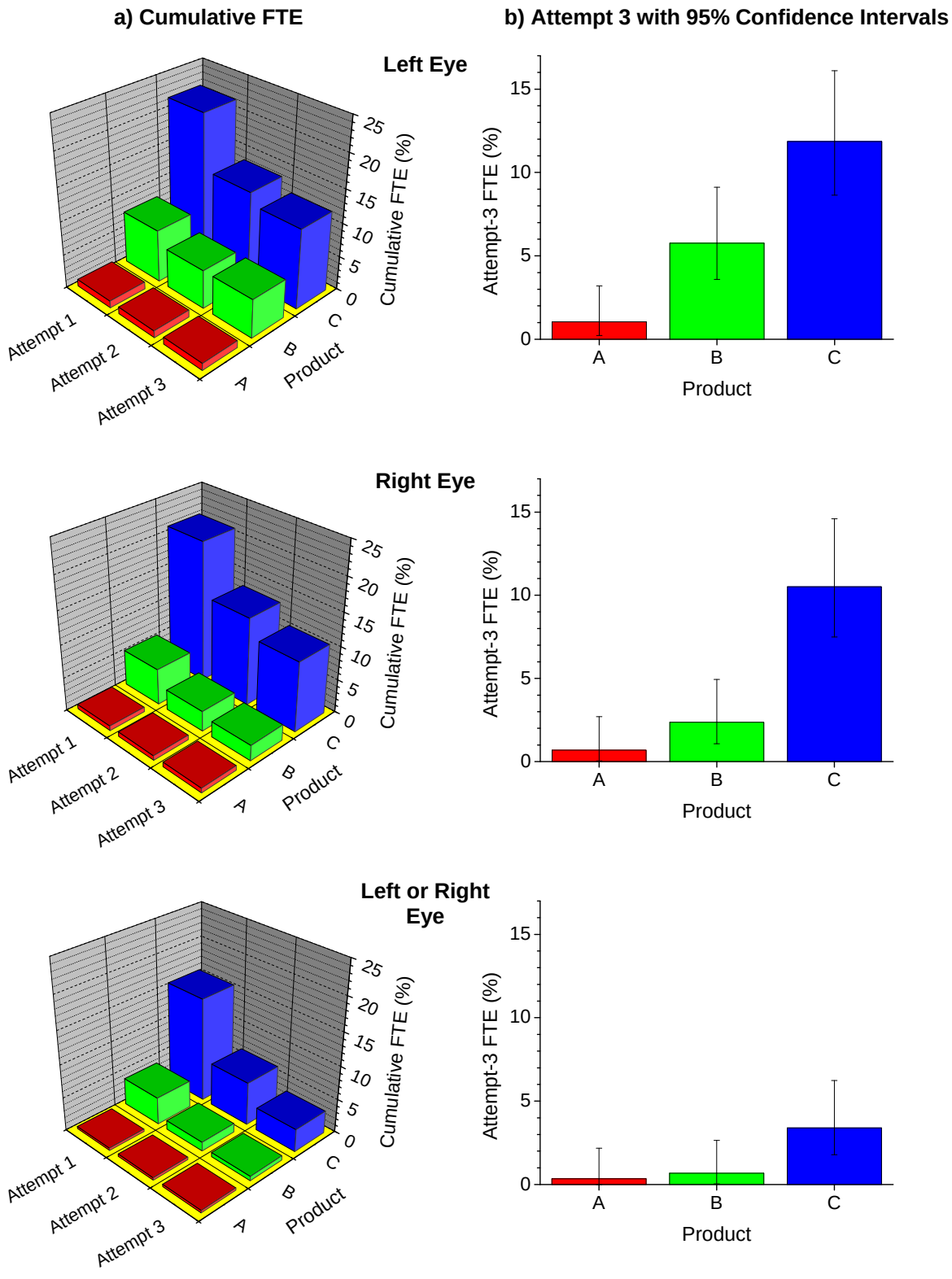


Figure 6-1. FTE Results

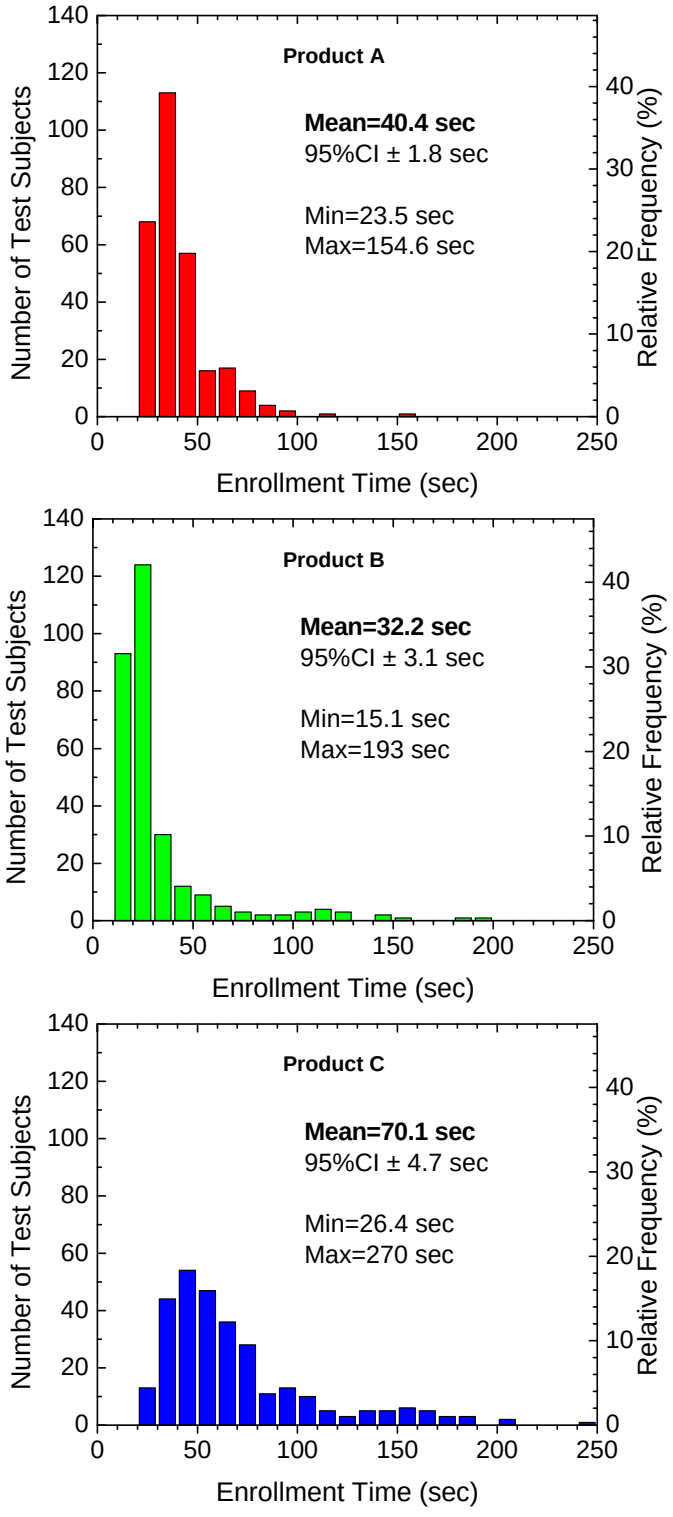


Figure 6-2. Enrollment Transaction Time Histograms

Enrollment transaction times

Recall that the cumulative FTE for the 3rd attempt for the Left-or-Right feature set is the enrollment transaction FTE per the IRIS06 test protocol. Recall also that the enrollment transaction time is defined as the sum of the attempt durations required to successfully enroll at least one eye or the sum of the three attempt durations if enrollment was unsuccessful. The enrollment transaction time histograms are presented in Figure 6-2; the associated mean enrollment transaction times (ETT) and 95% CIs are presented graphically in Figure 6-3. Figure 6-3 shows that the $ETT_B < ETT_A < ETT_C$. The CIs in Figure 6-3 do not overlap indicating that the difference is statistically significant. That is, Product B exhibited the shortest mean enrollment time, followed by Product A and then by Product C.

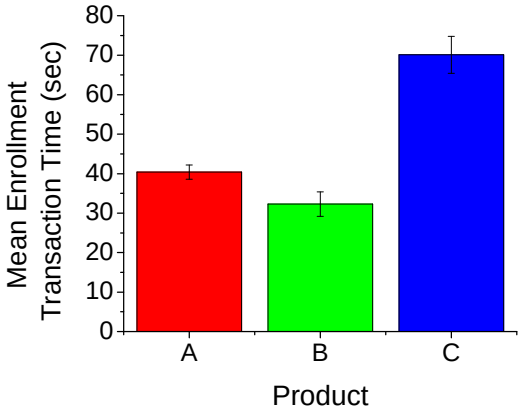


Figure 6-3. Mean Enrollment Transaction Times with 95% Confidence Intervals

Failure to acquire

The cumulative failure to acquire rates (FTA) for all three products and for all three iris-feature sets are presented in Table 6-2 for each of the transactions – two verify transactions during Visit 1 and two identify transactions during Visit 2.

Table 6-2. Cumulative Failure to Acquire (FTA)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	# of Persons	278	279	280	277	287	292	255	260	280
	Attempt 1	22.30%	17.56%	9.64%	17.69%	19.16%	15.07%	14.12%	13.85%	12.86%
	Attempt 2	11.15%	9.68%	4.64%	10.47%	9.76%	7.88%	11.76%	11.54%	12.50%
	Attempt 3	**2.52%**	**2.87%**	**0.71%**	**7.22%**	**6.62%**	**5.82%**	**11.76%**	**11.54%**	**12.50%**
	Verify 2									
	# of Persons	280	281	282	255	259	280	278	288	293
	Attempt 1	18.21%	23.49%	11.70%	23.02%	24.65%	17.75%	13.73%	13.51%	13.93%
	Attempt 2	8.93%	10.68%	5.32%	12.59%	12.50%	10.24%	12.55%	12.36%	12.86%
	Attempt 3	**2.86%**	**2.49%**	**1.77%**	**8.27%**	**10.76%**	**8.19%**	**11.76%**	**12.36%**	**12.86%**
Visit 2	**Identify 1**									
	# of Persons	167	167	168	202	208	213	222	225	243
	Attempt 1	26.35%	24.55%	14.29%	20.79%	23.08%	17.37%	5.86%	7.11%	2.88%
	Attempt 2	12.57%	13.77%	5.95%	9.90%	11.06%	7.04%	1.80%	1.33%	2.06%
	Attempt 3	**4.79%**	**4.79%**	**1.79%**	**5.94%**	**7.69%**	**6.10%**	**0.45%**	**0.44%**	**0.41%**
	Identify 2									
	# of Persons	167	167	168	202	208	213	220	222	241
	Attempt 1	26.95%	27.54%	16.67%	22.28%	21.15%	16.43%	3.64%	5.86%	1.66%
	Attempt 2	15.57%	16.17%	8.93%	13.37%	13.46%	10.33%	0.00%	1.80%	0.00%
	Attempt 3	**3.59%**	**6.59%**	**1.79%**	**8.42%**	**9.13%**	**7.51%**	**0.00%**	**0.45%**	**0.00%**

We observe that the cumulative FTA rates decrease with increasing numbers of attempts for all iris-feature sets for Products A and B. Presumably for these two products, the more a test subject practices, the more familiar they become with the camera, and the better the ability of the camera and test subject to acquire acceptable iris images. In addition, eyeglasses, if worn, were removed for the 3rd attempt, which may help decrease the 3rd attempt FTA. For Product C, FTA rates decreased or remained roughly the same upon subsequent attempts. Generally, little or no improvement was observed between the 2nd and 3rd attempts. We offer no specific explanation for this phenomenon.

The 3rd-attempt FTA results with 95% CIs are presented graphically in Figure 6-4. For each recognition transaction, the 3rd-attempt FTA CIs for the left and right eyes overlap

significantly for each individual product. For example, from Figure 6-4 we find for Visit 1-Verify 1, Product A, ***Left*** Eye, that the lower CI value L≈1.1% and the upper CI value U≈5.2%. Similarly, for Visit 1-Verify 1, Product A, ***Right*** Eye, L≈1.4% and U≈5.7%. Since these CIs overlap, we conclude that the left and right eyes exhibit roughly the same FTA rates for the Verify 1 transaction for Product A. We can see by visual inspection of Figure 6-4 that the same is true for each product and each transaction. As such, we conclude that left and right eyes exhibit roughly the same FTA rates for each of the three products evaluated. Recall that the same was true for FTE.

Turning our attention to the relative FTA rates between transactions for each individual product, we observe that the CIs for each individual feature set overlap for all transactions for Products A and B. Continuing with the previous example, recall that for Visit 1-Verify 1, Product A, Left Eye, L≈1.1% and U≈5.2%. Table 6-3 shows the CIs for Product A, Left Eye, for all four transactions. We see (both in Table 6-3 and graphically in Figure 6-4) that the CIs overlap substantially. We can see by visual inspection of Figure 6-4 that the same is true for all feature sets for Products A and B. As such, we conclude that for Products A and B, the FTA rates are roughly equivalent over all transactions for each individual feature set. This indicates that time separation between transactions does not have a measurable influence on FTA rates for Products A and B. Recall that the time separation between Visits 1 and 2 was roughly six weeks and the time separation between transactions within a visit was about 15 minutes. Recalling that the FTA for the right and left eyes is roughly the same, we can further conclude that for Product A and for Product B, the FTA is roughly the same for both right and left eyes over time.

Table 6-3. FTA 95% Confidence Intervals for Product A, Left Eye

Transaction	L (%)	U (%)
Verify 1	1.1	5.2
Verify 2	1.4	5.7
Identify 1	2.3	9.4
Identify 2	1.5	7.8

However for Product C we observe a different behavior, namely a significant difference between the FTA rates for Visits 1 and 2. The Product C FTA rates for Visit 1 are significantly larger than the Product C FTA rates for Visit 2 for all feature sets. In addition, the Product C FTA rates for Visit 1 are larger than the Visit 1 FTA rates Products A and B for corresponding iris-feature sets, while the Product C FTA rates for Visit 2 are smaller than the corresponding Visit 2 FTA rates for Products A and B.

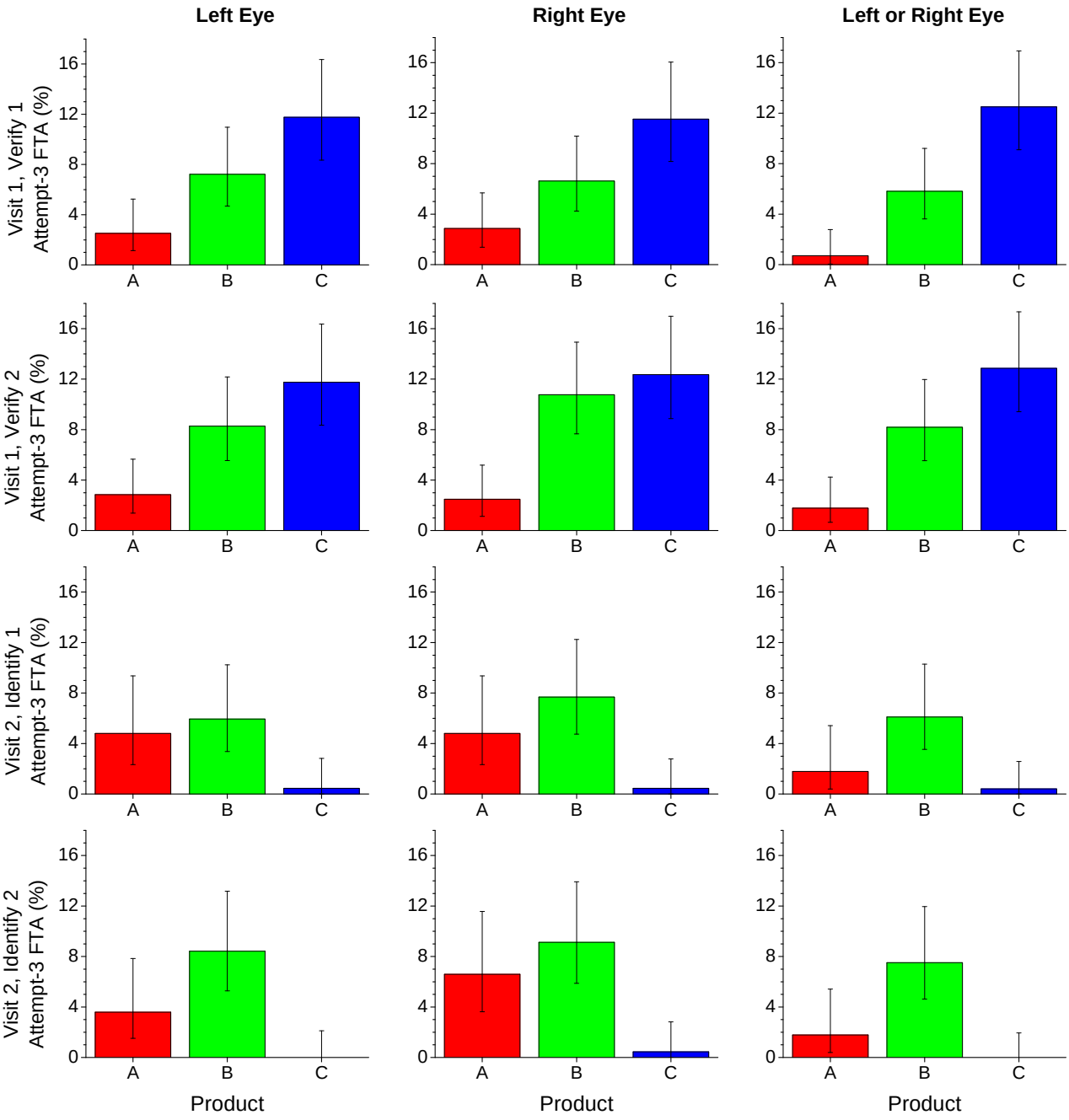


Figure 6-4. Cumulative 3rd-Attempt FTA with 95% Confidence Intervals by Transaction

Recall from Section 3.2 that the lighting was changed between Visits 1 and 2 specifically to address this issue. The results in Figure 6-4 indicate that the lighting change, removing two fluorescent lights above Product C, did improve FTA for Product C between Visits 1 and 2. Since the FTA rates are roughly equivalent over all transactions for each individual feature set for Products A and B, we conclude that the lighting change did not adversely influence the FTA

of Products A and B. We conclude that the FTA rate for Product C is influenced by ambient fluorescent lighting.

Before turning our attention to the relative FTA rates between products, we note that there is no statistically-significant difference in the 3rd-attempt FTA rates between the Visit 1 Verify 1 and Verify 2 transactions for a given product and a given feature set, and similarly for the Visit 2 Identify 1 and Identify 2 transactions. As such we combine these results to obtain average 3rd-attempt FTA rates for Visits 1 and 2 (Table 6.4). These results with 95% CIs are presented graphically in Figure 6-5.

Table 6-4. Cumulative 3rd-Attempt FTA by Visit

Product A			Product B			Product C		
L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)								
2.69%	2.68%	1.25%	7.75%	8.70%	7.01%	11.76%	11.95%	12.68%
Visit 2 (Identify 1 and Identify 2)								
4.19%	5.69%	1.79%	7.18%	8.41%	6.81%	0.23%	0.45%	0.21%

In Table 6-4, we observe that the FTA for Product A is higher for Visit 2 compared to Visit 1 for the left and right eyes, Product B FTA is slightly lower for Visit 2 compared to Visit 1 for all feature sets, and Product C FTA is substantially smaller for Visit 2 compared to Visit 1 for all feature sets. Inspecting Figure 6-5, we observe that the change in FTA between visits for Product A is not statistically significant (the error bars overlap for the Left Eye and Right Eye feature sets), in agreement with the analysis by transaction.

Note that when confidence intervals are not shown, the assumptions required to obtain valid CI estimates using the Logit Beta-binomial approach were not satisfied. This typically happens when the error rates are quite low. This is the case for the Left or Right Eye feature set for Product A, and for the Left Eye and Right Eye feature sets for Product C, Visit 2. While we cannot calculate meaningful Visit 2 FTA CIs for Product C due to the low error rates, we hypothesize that the differences between the Visit 1 and Visit 2 FTA are statistically significant based on the CI transactional data presented in Figure 6-4.

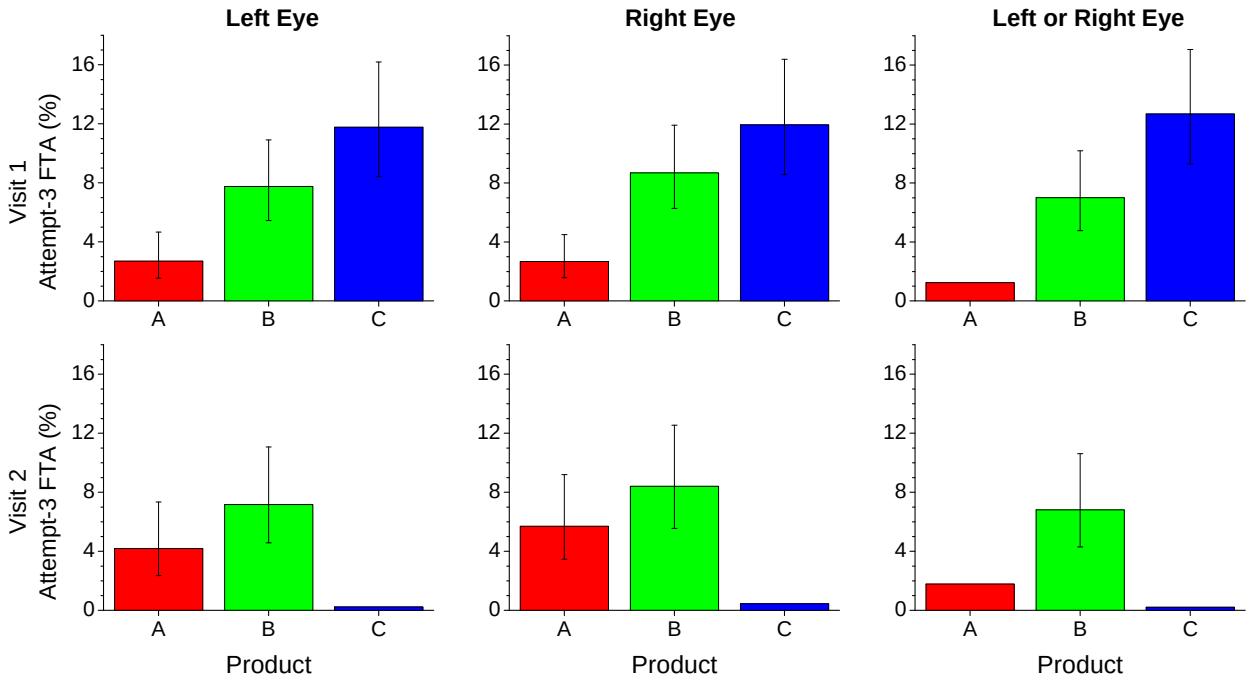


Figure 6-5. Cumulative 3rd-Attempt FTA with 95% Confidence Intervals by Visit

To investigate the relative FTA rates between products, we further distill the 3rd-attempt FTA dataset by combining the results for both visits as shown in Table 6-5 and Figure 6-6. These results represent the average performance over time for each camera. In addition, for Product C these results represent the average FTA in different lighting conditions.

Table 6-5 Overall Cumulative 3rd-Attempt FTA

Product A			Product B			Product C		
L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)								
3.25%	3.80%	1.45%	7.51%	8.58%	6.92%	6.41%	6.63%	6.90%

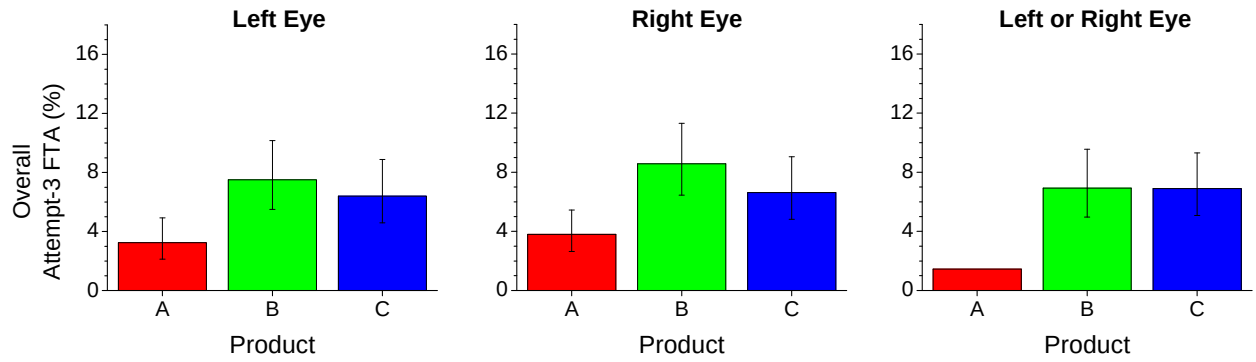


Figure 6-6. Overall Cumulative 3rd-Attempt FTA with 95% Confidence Intervals

As before, we observe no statistically-significant differences between left and right eye FTA for each individual product. Further, the overall cumulative FTA for Products B and C is roughly the same for each feature set (the CIs overlap). In addition, the average cumulative FTA for Product A is significantly lower than that for Product B and marginally lower than that for Product C for the left-eye and right-eye feature sets. For the left-or-right-eye feature set, the overall cumulative FTA for Product A may be less than that for Products B and C. In general, we note that $FTA_A \lesssim FTA_B \sim FTA_C$ for all feature sets.

False non-match rate

We now perform a similar analysis for cumulative false non-match rate (FNMR). The cumulative FNMR for all three products and all three iris-feature sets are presented in Table 6-6 for each of the transactions. Product B does not provide individual match results for the left-eye-only and right-eye-only feature sets.

Table 6-6. Cumulative False Non-Match Rate (FNMR)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	Attempt 1	3.24%	4.35%	0.40%			10.89%	10.96%	16.52%	6.97%
	Attempt 2	0.40%	0.00%	0.00%			2.23%	4.44%	9.13%	3.27%
	Attempt 3	**0.37%**	**0.00%**	**0.00%**			**0.36%**	**0.00%**	**0.87%**	**0.00%**
	Verify 2									
	Attempt 1	3.06%	1.40%	0.00%			13.69%	13.18%	16.96%	7.47%
	Attempt 2	0.39%	0.00%	0.00%			5.32%	4.93%	4.85%	2.87%
	Attempt 3	**0.37%**	**0.00%**	**0.00%**			**2.23%**	**0.00%**	**0.88%**	**0.41%**
Visit 2	**Identify 1**									
	Attempt 1	4.88%	3.17%	0.69%			18.18%	18.18%	24.88%	9.75%
	Attempt 2	0.68%	0.00%	0.00%			6.57%	8.26%	11.26%	4.20%
	Attempt 3	**0.00%**	**0.00%**	**0.00%**			**3.00%**	**2.26%**	**3.13%**	**1.24%**
	Identify 2									
	Attempt 1	2.46%	3.31%	0.00%			21.35%	18.40%	22.49%	11.39%
	Attempt 2	0.71%	0.00%	0.00%			9.42%	8.18%	12.39%	5.39%
	Attempt 3	**0.62%**	**0.00%**	**0.00%**			**2.03%**	**3.64%**	**1.81%**	**0.00%**

We observe, as with FTE and FTA, that cumulative FNMR generally decreases with increasing recognition attempts. Again, this is presumably because the effectiveness of the human-camera interface improves with practice, or habituation. The decrease in 3rd-attempt FNMR may also be affected by the removal of eyeglasses. In many cases the FNMR goes to zero after two or three attempts. We note that the 3rd-attempt FNMR values for Product C are near

zero for Visit 1 and generally higher for Visit 2. However, inspection of the 3rd-attempt FNMR 95% CIs presented in Figure 6-7 indicates that the difference is not statistically significant. We also observe in Figure 6-7 that, with one exception (Visit 2 – Identify 2, Left eye, Product C), the 3rd-attempt FNMR CIs for all feature sets, for all transactions, and for all products are relatively large and all overlap each other to some degree. For Visit 2 – Identify 2, Left eye, Product C, the lower CI value is slightly higher than the upper CI value for a few of the Visit 1, Product A CIs;

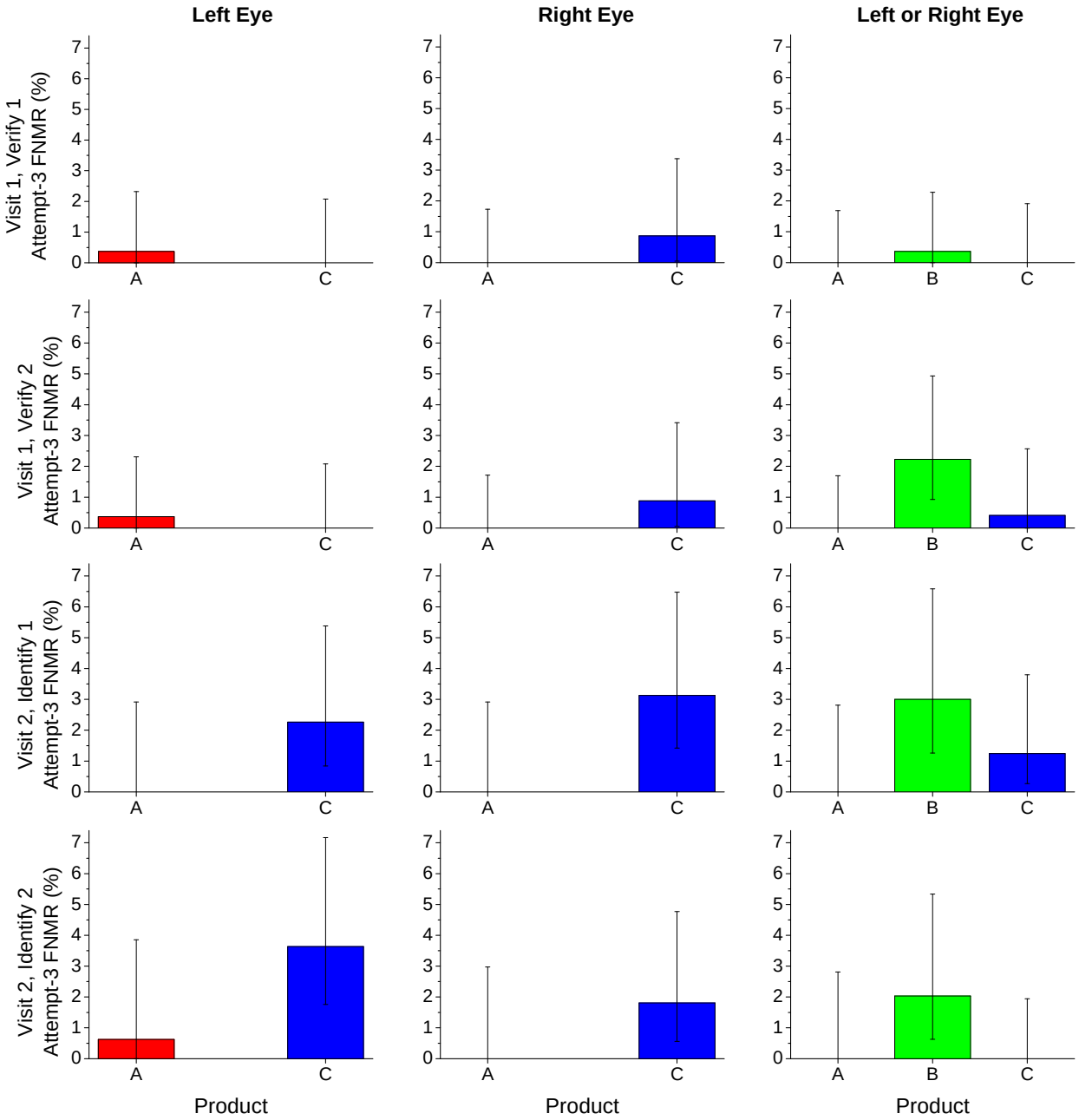


Figure 6-7. Cumulative 3rd-Attempt FNMR with 95% Confidence Intervals by Transaction

thus the CIs do not overlap. This may be a consequence of Product C's sensitivity to ambient lighting conditions. Noting this one exception, we observe that 3rd-attempt FNMR is roughly the same for all products, for all iris-feature sets, and for transactions separated by as much as six weeks.

For completeness, we present cumulative 3rd-attempt FNMR results with 95% CIs by visit and overall in Tables 6-7 and 6-8 and Figures 6-8 and 6-9. Since the error rates are low, we are unable to estimate CIs in many cases. As such, these results are challenging to analyze. More meaningful conclusions can be drawn from the GFRR analysis below, which incorporates FTE, FTA, and FNMR rates and allows us to compare the overall error rates between products. A more thorough FNMR analysis is presented in the offline analysis (Section 6.1.2).

Table 6-7. Cumulative 3rd-Attempt FNMR By Visit

Product A			Product B			Product C		
L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)								
0.37%	0.00%	0.00%			1.29%	0.00%	0.88%	0.20%
Visit 2 (Identify 1 and Identify 2)								
0.31%	0.00%	0.00%			2.52%	2.95%	2.47%	0.62%

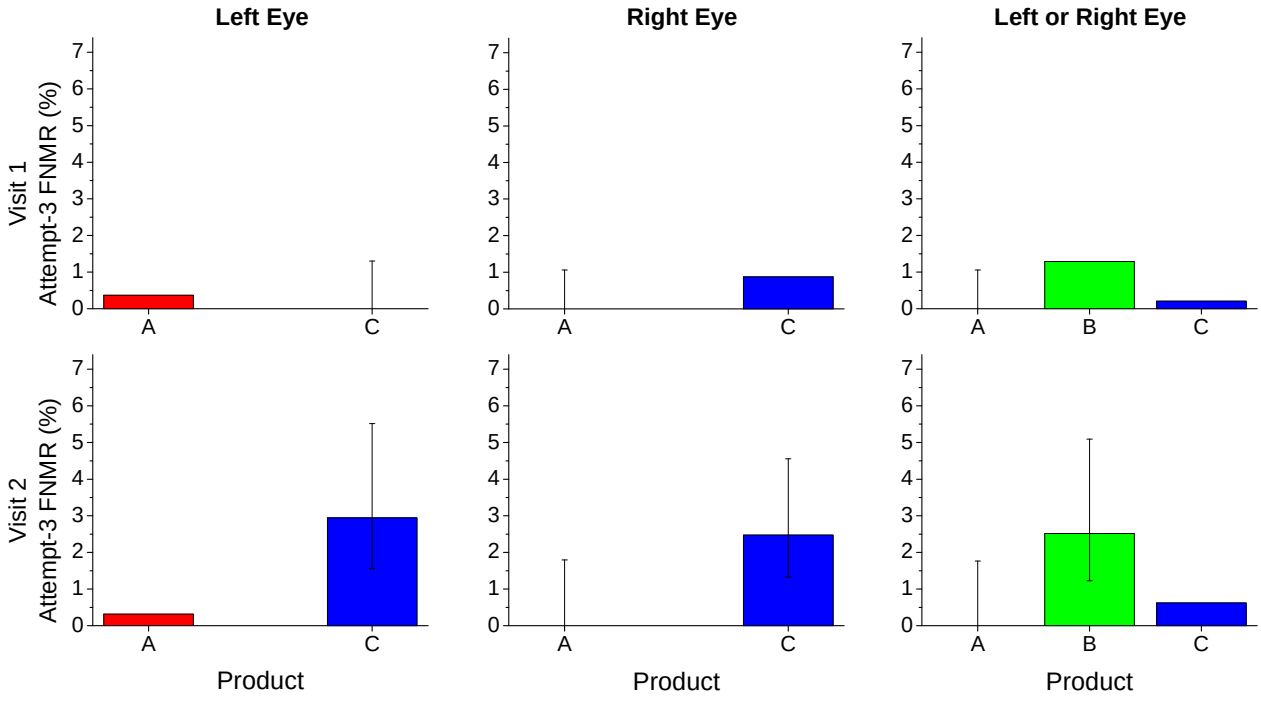


Figure 6-8. Cumulative 3rd-Attempt FNMR with 95% Confidence Intervals by Visit

Table 6-8. Overall Cumulative 3rd-Attempt FNMR								
Product A			Product B			Product C		
L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)								
0.35%	0.00%	0.00%			1.81%	1.46%	1.66%	0.41%

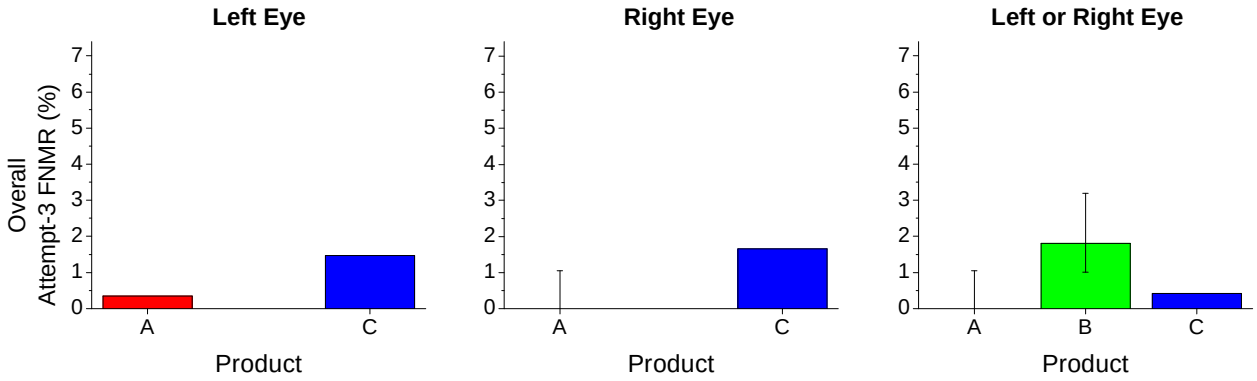


Figure 6-9. Overall Cumulative 3rd-Attempt FNMR with 95% Confidence Intervals

Generalized false reject rate

As explained in Section 5.1.1, the generalized false reject rate (GFRR) incorporates enrollment, biometric sample acquisition, and matching errors, allowing us to estimate the percentage of the population that will not be able to successfully utilize the biometric product, whether due to FTE, FTA, or FNMR. As such, we can readily compare and contrast the expected operational performance of different systems.

Recall that the equations for GFRR (for verification) and GFNIR (for identification) are identical. For simplicity, we use GFRR to represent GFNIR in this analysis. The cumulative GFRR for all three products and all three iris-feature sets are presented in Table 6-9 for each of the transactions. Also shown in Table 6-9 are the recognition transaction times, which are discussed in the following section.

Table 6-9 shows, as before, that the cumulative GFRR rates decrease with increasing numbers of attempts for all iris-feature sets, for all products, and for all transactions. Again this is likely caused by improved effectiveness of the human-camera interface with increasing human practice and the removal of eyeglasses on the 3rd attempt.

	Table 6-9. Cumulative Generalized False Reject Rate (GFRR)									
		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	Attempt 1	25.62%	21.71%	10.32%			24.83%	32.76%	35.74%	21.72%
	Attempt 2	12.46%	10.32%	4.98%			10.54%	25.86%	28.18%	18.28%
	Attempt 3	**3.91%**	**3.56%**	**1.07%**			**6.80%**	**22.41%**	**21.65%**	**15.52%**
	RTT (sec)	19.45	17.84	21.57			7.42	8.20	9.55	11.27
	Verify 2									
	Attempt 1	21.55%	25.09%	12.01%			29.49%	34.14%	35.86%	23.10%
	Attempt 2	10.25%	11.31%	5.65%			15.59%	26.90%	25.52%	18.28%
	Attempt 3	**4.24%**	**3.18%**	**2.12%**			**10.85%**	**22.41%**	**22.41%**	**16.21%**
	RTT (sec)	17.60	17.52	20.31			7.58	8.28	8.96	10.99
Visit 2	**Identify 1**									
	Attempt 1	31.18%	27.81%	15.38%			33.02%	33.46%	38.67%	15.81%
	Attempt 2	14.71%	14.79%	6.51%			13.95%	22.18%	23.05%	9.88%
	Attempt 3	**6.47%**	**5.92%**	**2.37%**			**9.77%**	**15.95%**	**15.23%**	**5.53%**
	RTT (sec)	18.64	17.57	21.77			8.78	9.49	11.43	12.06
	Identify 2									
	Attempt 1	30.00%	30.77%	17.16%			34.88%	32.16%	35.97%	16.33%
	Attempt 2	17.65%	17.16%	9.47%			19.53%	20.78%	24.51%	9.16%
	Attempt 3	**5.88%**	**7.69%**	**2.37%**			**10.23%**	**16.86%**	**14.23%**	**3.98%**
	RTT (sec)	17.45	17.78	21.95			7.96	8.15	10.15	10.30

The 3rd-attempt GFRR results with 95% CIs are presented graphically in Figure 6-10. For each transaction, the 3rd-attempt GFRR CIs for the left and right eyes overlap significantly for each Product A and Product C. (Product B does not provide individual eye match results.) As such, the left and right eyes exhibit roughly the same GFRR for Products A and C. In addition, the CIs for each individual feature set overlap significantly for all transactions for Products A and B (scanning vertically in Figure 6-10). As such, time separation between transactions does not have a measurable influence on GFRR for Products A and B.

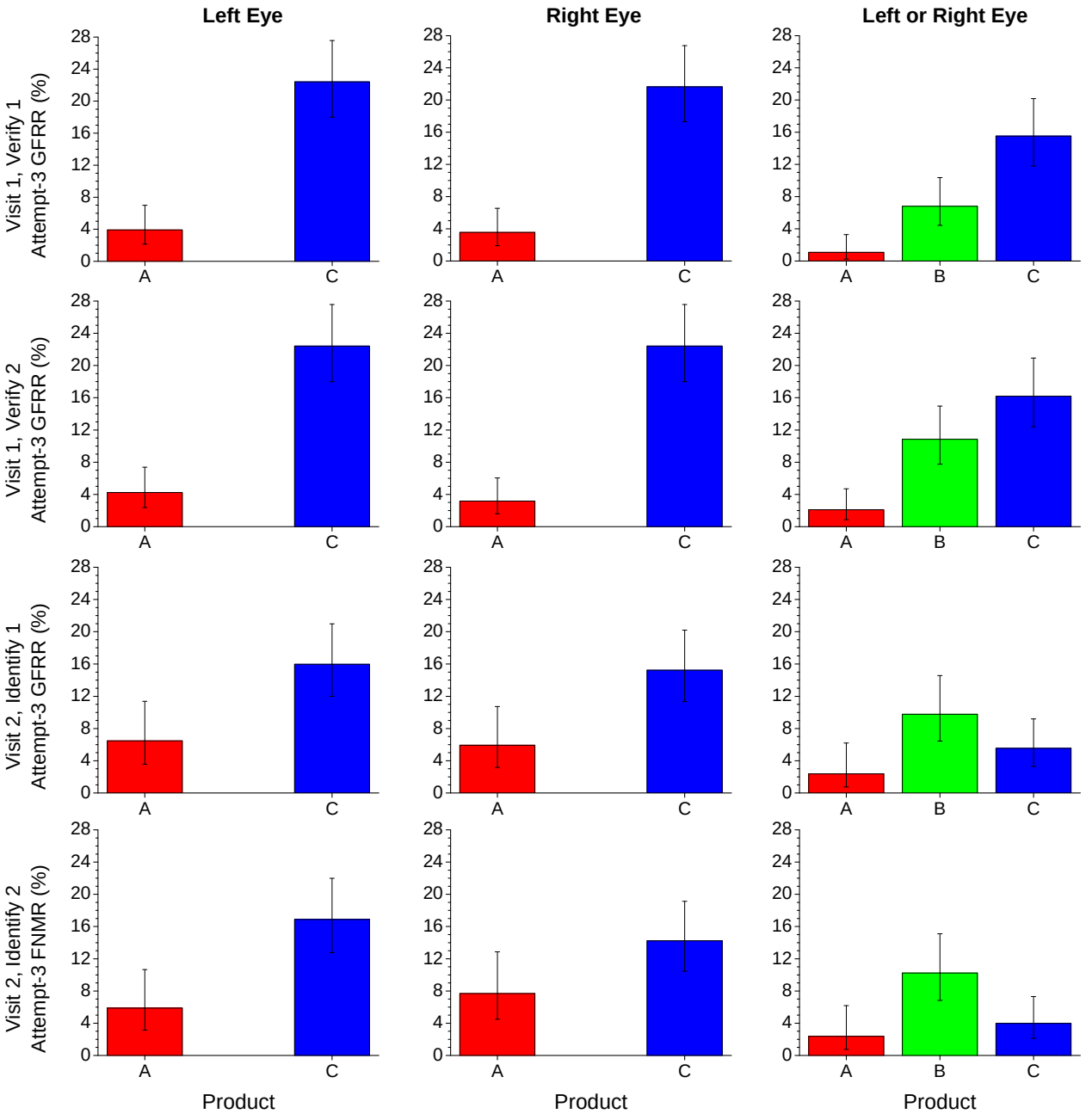


Figure 6-10. Cumulative 3rd-Attempt GFRR with 95% Confidence Intervals by Transaction

However, as seen in the GFRR by Visit data in Table 6-10 and Figure 6-11, the 3rd-attempt GFRR for Product C behaves differently. The GFRR is higher for Visit 1 than for Visit 2. For the left and right eyes, the CI's overlap somewhat, indicating that the difference is not significant, while for the left-or-right-eye feature set, the CIs do not overlap, indicating that the overall error rate for left-or-right-eye feature set was improved by changing the ambient fluorescent lighting. We note that Table 6-10 and Figure 6-11 indicate that the 3rd-attempt GFRR

slightly increased for Products A and B and slightly decreased for Product C for Visit 2 compared to Visit 1. As such, it appears that the lighting change may have negatively impacted (increased) the GFRR of Products A and B and improved (reduced) the GFRR for Product C. Inspection of the CIs in Figure 6-11 however indicates the only statistically significant change was the reduced GFRR for the left-or-right-eye feature set for Product C. The most striking component of GFRR that contributed to this change was FTA (per Figure 6-5). As such, we hypothesize that ambient fluorescent lighting can influence Product C's ability to capture usable iris images.

Table 6-10. Cumulative 3rd-Attempt GFRR by Visit

	Product A			Product B			Product C		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)									
	4.08%	3.37%	1.60%			8.83%	22.41%	22.03%	15.86%
RTT (sec)	18.52	17.68	20.94			7.50	8.24	9.25	11.13
Visit 2 (Identify 1 and Identify 2)									
	6.18%	6.80%	2.37%			10.00%	16.41%	14.73%	4.76%
RTT (sec)	18.04	17.67	21.86			8.37	8.82	10.79	11.18

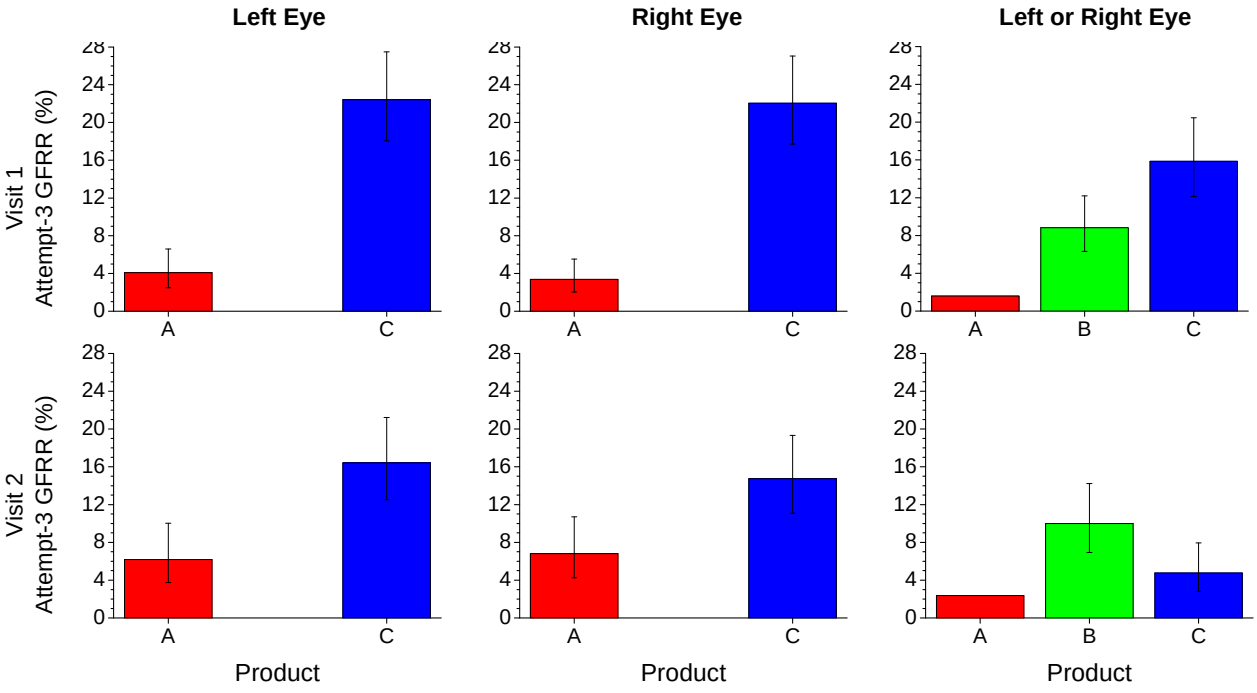


Figure 6-11. Cumulative 3rd-Attempt GFRR with 95% Confidence Intervals by Visit

Table 6-11. Overall Cumulative 3rd-Attempt GFRR									
	Product A			Product B			Product C		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)									
	4.87%	4.66%	1.88%			9.32%	19.60%	18.62%	10.70%
RTT (sec)	18.28	17.68	21.40			7.93	8.53	10.02	11.16

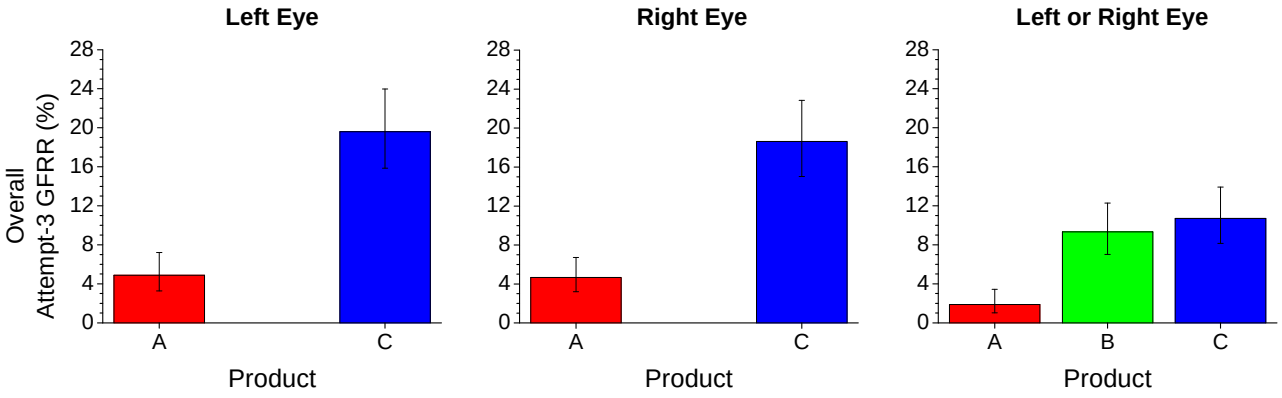


Figure 6-12. Overall Cumulative 3rd-Attempt GFRR with 95% Confidence Intervals

Finally, we review the overall 3rd-attempt GFRR data in Table 6-11 and Figure 6-12. In agreement with the analysis by transaction, left and right eyes exhibit roughly the same GFRR for Products A and for Product C.

In addition, the 3rd-attempt GFRR for product A was significantly lower than that for Product C for all iris-feature sets. For the left-or-right-eye feature set, the GFRR for Products B and C are roughly equivalent, and the GFRR for Product A is significantly lower than that for Products B and C. Recall that for Product C, this is an averaged GFRR from two different ambient lighting conditions. (Though Products A and B also experienced the different ambient lighting conditions, their performances was not significantly influenced.) In general, we conclude that $GFRR_A < GFRR_B \sim GFRR_C$.

We also note that the left-or-right-eye GFRR is statistically lower than the single-eye GFRR for Products A and C. Recall that the single-eye rates are based on three attempts with the selected eye and that the left-or-right-eye FTA is based on three attempts for each eye, which allows six attempts overall to acquire at least one iris image. For Products A and C, the additional attempts reduced the overall error rate somewhat. (Product B did not provide single-eye information.)

Recognition transaction times

The average recognition transaction times for each transaction are listed in Table 6-9 for each product and each iris-feature set. Figure 6-13 displays these results graphically with 95% confidence intervals. The CIs for each feature set for each individual product overlap to some extent for each of the transactions (scanning vertically in Figure 6-13). This indicates that the recognition transaction times did not change substantially over time. We conclude that the time duration between recognition attempts has no significant influence on recognition transaction times for the products tested.

The mean recognition times averaged over all transactions are shown in Table 6-11 and Figure 6-13. Reviewing Table 6-11, it appears that the right eye transaction times are slightly shorter than the left eye transaction times for Product A, while for Product C the left eye transaction times are slightly shorter than the right eye transaction times. Upon inspection of the CIs in Figure 6-13, we see that the left-eye and right-eye CIs for Product A and for Product C overlap, indicating that the difference between right-eye and left-eye mean recognition transaction times is probably not statistically significant.

However, the mean recognition transaction times for Product C are statistically shorter than those for Product A for all feature sets (CIs do not overlap). In addition, for the left-or-right-eye feature set, $RTT_B < RTT_C < RTT_A$. This is also clearly illustrated in the associated RTT histograms shown in Figure 6-14. We note that while Product A exhibited the lowest error rates, it also exhibited the longest mean recognition transaction times. As such, we observe a tradeoff between speed and accuracy.

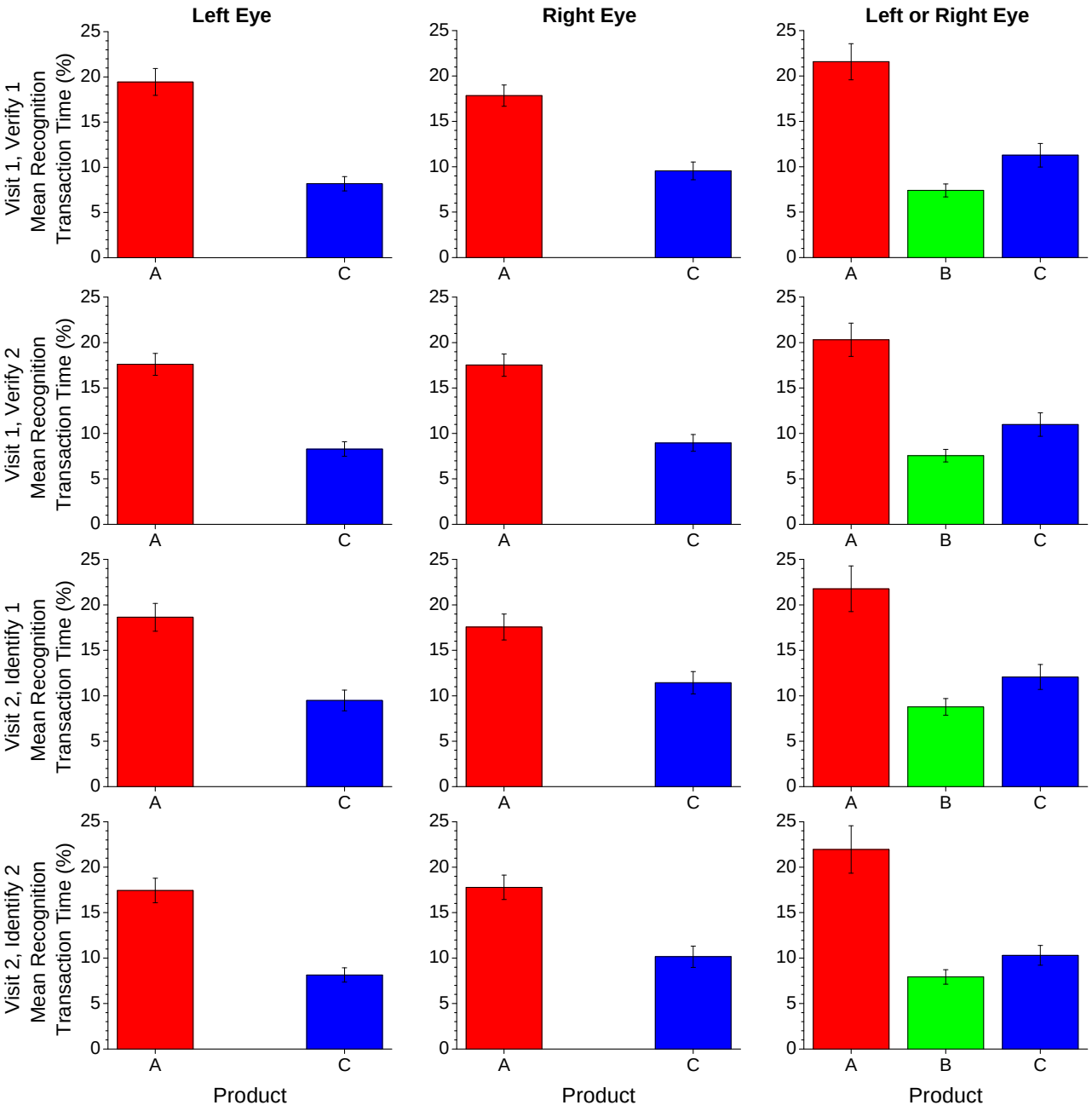


Figure 6-13. Mean Recognition Transactions Times with 95% Confidence Intervals by Transaction

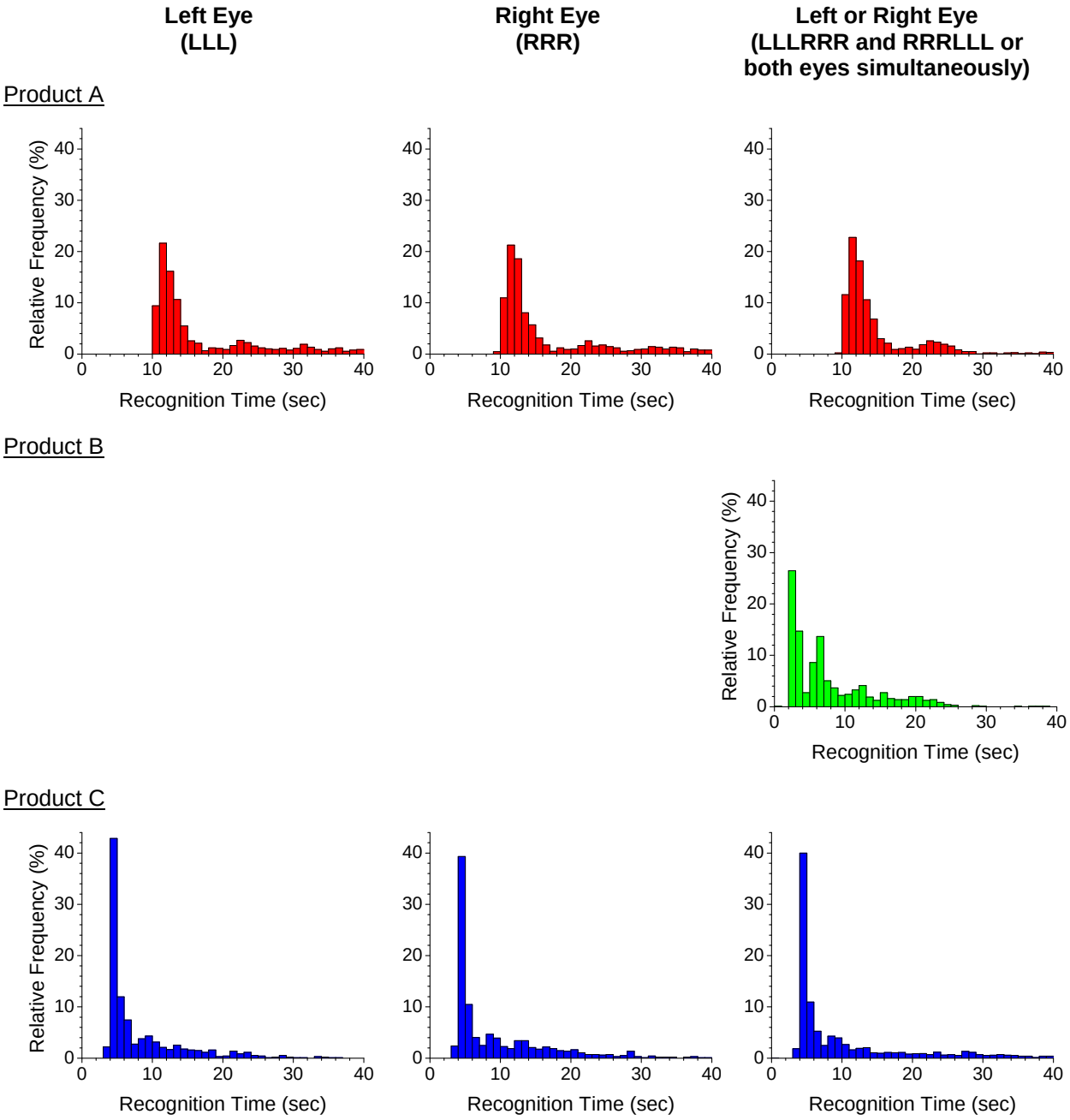


Figure 6-14. Overall Recognition Transaction Time Histograms

Summary of key online findings

- Cumulative FTE, FTA, FNMR, and GFRR rates generally decrease with increasing numbers of attempts, possibly due to improved effectiveness of the camera-human interface with increasing human practice and the removal of eyeglasses for the third attempt.
- All products exhibit roughly the same FTE rate when a successful enrollment means that at least one eye successfully enrolled after three attempts with each eye
- Left and right eyes generally exhibit roughly the same FTE, FTA, FNMR, GFRR, and RTT for each of the products evaluated
- Time separation between transactions does not have a measurable influence on FTA, FNMR, and GFRR for Products A and B
- Time separation between transactions does not have a measurable influence on mean RTT for all products tested
- FNMR is similar for all products and all iris-feature sets
- Product C is influenced by ambient fluorescent lighting
- In general, the 3rd-attempt $FTA_A \lesssim FTA_B \sim FTA_C$
- In general, the 3rd-attempt $GFRR_A < GFRR_B \sim GFRR_C$
- Mean enrollment transaction times: $ETT_B < ETT_A < ETT_C$
- Mean recognition transaction times for left-or-right eye feature set: $RTT_B<RTT_C<RTT_A$
- Mean recognition transaction times for single-eye feature sets: $RTT_C < RTT_A$
- The products tested demonstrate tradeoffs between speed and accuracy. The “best” product depends on the needs of a particular operational scenario.

6.1.2. Offline results

The iris images collected during the online scenario evaluation were utilized for offline analysis. Professor John Daugman's "irisenroll (release 1.5)" template generator and "matcher1" template matcher were used to generate iris templates (Iriscodes) and to perform matching operations, respectively. We present the offline error rates using ROC curves where TMR is plotted as a function of FMR at different decision thresholds. (Recall that TMR=1-FNMR.) We also present generalized ROC curves where GTAR is plotted as a function of GFAR at different decision thresholds. (Recall that GTAR=1-GFRR.) The offline generalized performance curves are generated using the FTE and FTA rates measured during the online evaluation. All offline results presented in this report are generated in verification mode, even if the images were collected in identification mode.[50]

Online versus offline performance

We begin the offline analysis by comparing the offline results with the online results to understand how well offline results emulate online performance. As an example, Figure 6-15 shows the Visit 1, Verify 1 basic offline ROC curves for cumulative attempts for Products A, B, and C along with the color-coded Hamming distance (threshold score) map. The color map indicates the Hamming distance that corresponds to each point on the curve. For example, for Product A, left eye, there is an orange-colored point and a red-colored point, both at TMR~0.98, at the two lowest FMR values. From the color map, orange indicates threshold scores between 0.32 and 0.34, and red indicates threshold scores between 0.30 and 0.32.

Figure 6-15 also shows the corresponding online results from Table 6-6. Since we did not perform online impostor attempts, we cannot estimate online FMR. As such, we plot the online cumulative TMR on the left vertical axis. (Recall that TMR=1-FNMR.) It is our understanding that most commercial instantiations of Professor Daugman's algorithm use a Hamming distance (HD) of 0.32 as the threshold. As such, we would expect the online TMR to agree with the offline TMR values at HD=0.32, which according to the color map corresponds roughly to the lowest FMR value on each curve in Figure 6-15. However, we observe that this is generally not the case. For example, for Product C, Left Eye, the online Attempt 1 TMR is 0.89 (89%) while

[50] We expect the error rate differences between verification and identification modes to be minimal for the IRIS06 evaluation conditions. The analysis of identification performance is a subject for future study.

the lowest offline value is about 0.93 (93%). For Attempt 2, $TMR_{online} \approx TMR_{offline} \approx 96\%$, and for Attempt 3, $TMR_{online} \approx 100\%$ while $TMR_{offline} \approx 98\%$. Similar discrepancies between online and offline results are observed for other products and eye-feature sets. Note that for Product A, two or three online points for each feature set lie on top of each other at or near 100% TMR while the corresponding offline values are lower in most cases. Also recall that Product B did not provide individual-eye online match results.

A similar comparison between online and offline cumulative-attempt TMR performance for Visit 2, Identify 2 is presented in Figure 6-16. We again observe significant discrepancies between the online match results (the points on the left vertical axis) and the offline match results (the points corresponding to low FMR on the ROC curves).

To further explore the performance differences, we list in Table 6-12 the offline FNMR values at HD=0.32 ($FNMR_{HD=0.32}$) for Visit 1, Verify 1 and the associated online FNMR values from Table 6-6. In some cases, the ROC point corresponding to HD=0.32 is not shown in the graph because the $FMR_{HD=0.32}$ is zero, and $\log_{10}(0) = -\infty$ cannot be plotted on our finite graphs. The $FNMR_{HD=0.32}$ values presented in Table 6-12 are taken from the source data for the ROC curves. Table 6-12 also lists the number of genuine match errors (false non-matches) and the number of genuine comparisons for each attempt in parenthesis for the Visit 1, Verify 1 transaction. In agreement with the graphical results in Figures 6-15 and 6-16, we observe that the number of match errors for corresponding online and offline cases is different. For example, for Product A, Left Eye, Attempt 1, 7 match errors occurred online while 9 match errors occurred offline. Similar differences are observed throughout Table 6-12. An analogous table for the Visit 2, Verify 2 transaction is presented in Appendix 11.4

It is interesting to note in Table 6-12 that the number of genuine comparisons increases with increasing attempts for each transaction. This is because the failure to acquire rate decreases with increasing attempts, and FTA is not included in FNMR calculations. Continuing with our example, for Product A, Left Eye, Attempt 1, there are 216 genuine comparisons. Referring to Table 6-2 (p. 81), 278 persons participated in the Visit 1, Verify 1, Product A, Left eye transaction, and the FTA for Attempt 1 was 22.3%. Then 278 persons x (1-0.223)=216 genuine comparisons were performed for Attempt 1. Similarly for Attempt 2, 278 x (1-0.1115)=247 genuine comparisons, which agrees with the value in Table 6-12.

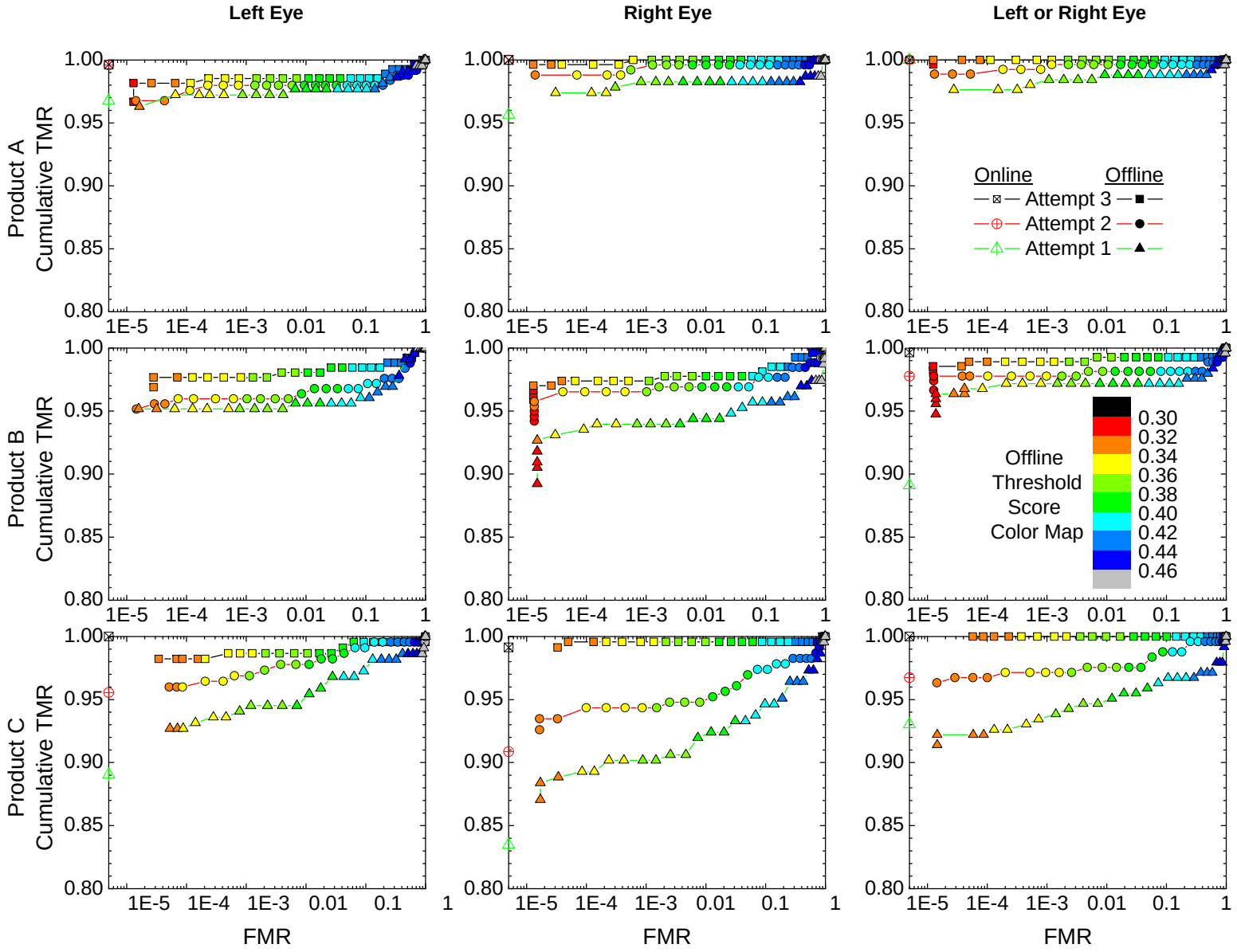


Figure 6-15. Visit 1, Verify 1 Basic Cumulative Performance Curves by Attempt

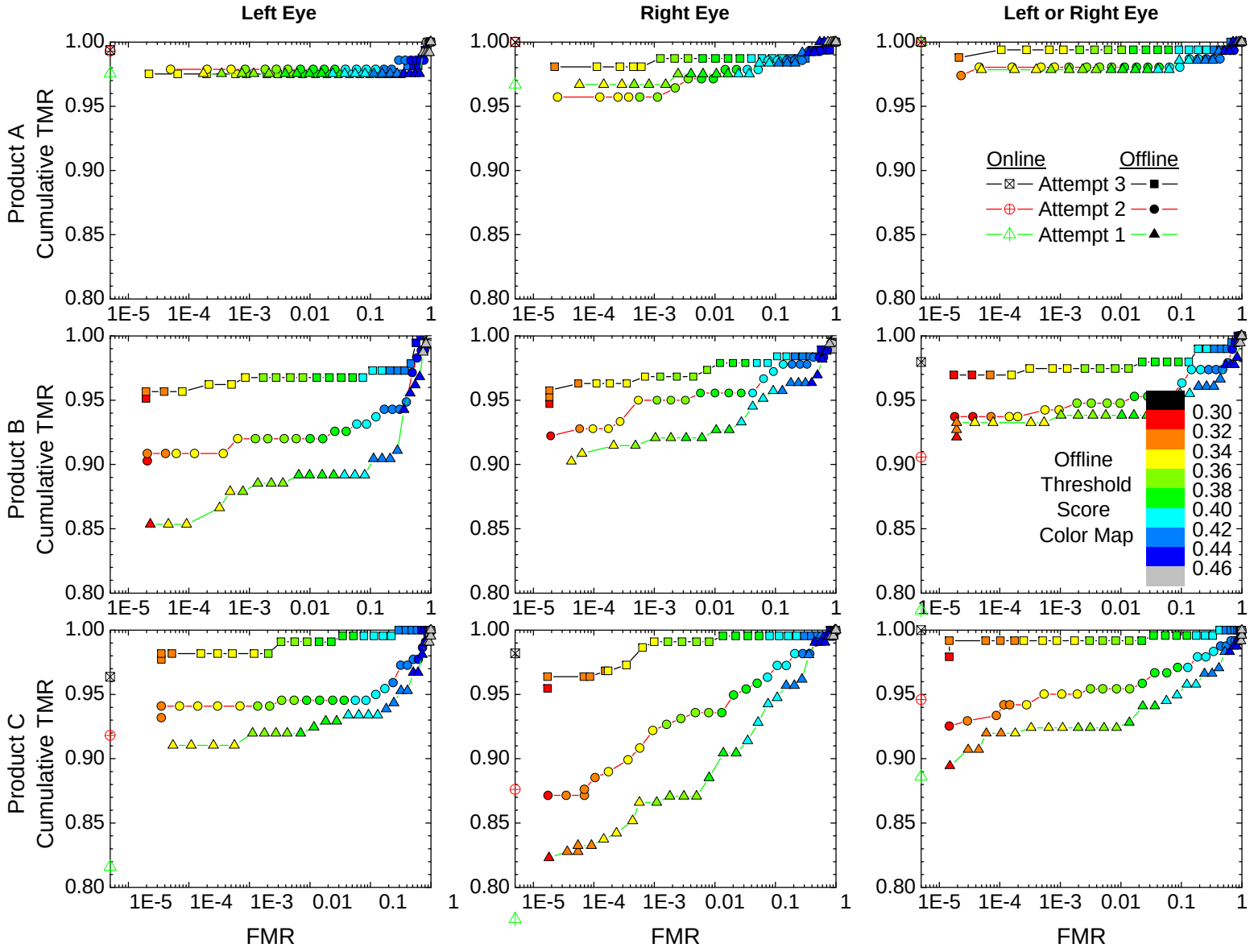


Figure 6-16. Visit 2, Identify 2 Basic Cumulative Performance Curves by Attempt

Table 6-12. Visit 1, Verify 1 Online and Offline FNMR						
	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product A						
Attempt 1	0.03241	0.04167	0.04348	0.03043	0.00395	0.03162
	(7/216)	(9/216)	(10/230)	(7/230)	(1/253)	(8/253)
Attempt 2	0.00405	0.03644	0	0.01587	0	0.01124
	(1/247)	(9/247)	(0/252)	(4/252)	(0/267)	(3/267)
Attempt 3	0.00369	0.01845	0	0.00738	0	0
	(1/271)	(5/271)	(0/271)	(2/271)	(0/278)	(0/278)
Product B						
Attempt 1					0.10887	0.03629
					(27/248)	(9/248)
Attempt 2					0.02230	0.02230
					(6/269)	(6/269)
Attempt 3					0.00364	0.01455
					(1/275)	(4/275)
Product C						
Attempt 1	0.10959	0.08219	0.16518	0.12946	0.06967	0.08607
	(24/219)	(18/219)	(37/224)	(29/224)	(17/244)	(21/244)
Attempt 2	0.04444	0.04889	0.09130	0.07391	0.03265	0.04082
	(10/225)	(11/225)	(21/230)	(17/230)	(8/245)	(10/245)
Attempt 3	0	0.01778	0.00870	0.00870	0	0
	(0/225)	(4/225)	(2/230)	(2/230)	(0/245)	(0/245)
(# of false non-matches / # of genuine comparisons)						

To further investigate the online and offline match differences, we re-present the data in Table 6-12 along with a list of the UINs for each test subject that failed to match for each attempt in Table 6-13. The yellow highlight UINs (xxxx) in Table 6-13 matched offline but did not match online. The green highlight UINs (xxxx) matched online but did not match offline. Clearly there are significant differences in the matching performance between the online and offline scenarios. For comparison, a similar analysis for the Visit 2, Identify 2 transaction is presented in Appendix 11.4.

Table 6-13. Visit 1, Verify 1 Online and Offline FNMR with UINs						
	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product A						
Attempt 1	0.03241 (7/216)	0.04167 (9/216)	0.04348 (10/230)	0.03043 (7/230)	0.00395 (1/253)	0.03162 (8/253)
UINs	1002 1003 1070 3069 5069 6080 7003	1047 1060 3063 3069 4027 4081 5040 5068 6022	1002 1021 2080 3030 3031 3043 3064 4024 5020 6007	0507 3030 3031 3043 3060 6050 6069	1002	1047 3030 3031 3043 4081 5068 6022 6069
Attempt 2	0.00405 (1/271)	0.03644 (9/247)	0 (0/252)	0.01587 (4/252)	0 (0/267)	0.01124 (3/267)
UINs	3069	1060 2044 3063 3069 4027 5040 5068 5080 5086		0507 3030 3043 6069		3043 5068 6069
Attempt 3	0.00369 (1/271)	0.01845 (5/271)	0 (0/271)	0.00738 (2/271)	0 (0/278)	0 (0/278)
UINs	3069	1060 3063 4027 5080 5086		0507 6069		
Product B						
Attempt 1					0.10887 (27/248)	0.03629 (9/248)
UINs					0518 0520 1002 1022 1047 1048 1049 1065 2027 2061 2081 3006 3011 3025 3026 3044 3063 3081 4006 4041 4047 4084 5024 5080 5083 6065 6080	1002 1022 1049 1065 2081 3044 4080 5040 6080
Attempt 2					0.02230 (6/269)	0.02230 (6/269)
UINs					1002 1065 2027 3006 4047 6065	1027 1049 1065 2029 2030 2082
Attempt 3					0.00364 (1/275)	0.01455 (4/275)
UINs					3006	1065 2029 2030 2082

	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product C						
Attempt 1	0.10959 (24/219)	0.08219 (18/219)	0.16518 (37/224)	0.12946 (29/224)	0.06967 (17/244)	0.08607 (21/244)
UINs	0516 0517 0521 1067 2006 2026 2028 2060 3006 3031 3040 3044 3049 3088 4024 4060 5024 5026 5049 5068 6001 6022 7009 7041	0517 0521 1003 2060 3006 3031 3040 3081 3088 4024 4027 4060 5024 5026 5049 5068 6001 6022	0521 0532 1000 1004 1006 1021 1050 1061 1067 1083 2006 2025 2030 2044 2068 3011 3025 3030 3040 3050 3090 4009 4024 4042 4060 4087 4088 5024 5045 5049 5067 5071 5080 6040 6065 6089 7004	0500 0521 1000 1004 1024 1050 1083 2006 2030 2042 2044 2068 3025 3040 3090 4024 4027 4042 4046 4060 4088 5024 5045 5049 5067 6040 6089 7004 7008	0517 0521 1004 1067 2006 3025 3030 3040 3044 4024 4060 4088 5024 5049 5071 6001 6040	0517 0521 1004 2042 3001 3025 3040 3081 3088 4024 4027 4042 4046 4060 4088 5024 5045 5049 5067 6022 6040
Attempt 2	0.04444 (10/225)	0.04889 (11/225)	0.09130 (21/230)	0.07391 (17/230)	0.03265 (8/245)	0.04082 (10/245)
UINs	3006 3031 3044 3088 4060 5024 5049 5068 6001 6022	0514 3006 3031 3088 4024 4060 5024 5049 5068 6001 6022	0532 1004 1050 1061 1083 2025 2030 2044 3011 3025 3030 3040 3090 4024 4042 4060 4087 4088 5067 6040 6065	1004 1050 1083 2030 2044 3025 3040 3088 3090 4010 4024 4042 4060 4088 5067 6040 7004	1004 3025 3030 3044 4060 4088 6001 6040	1004 3025 3088 4024 4042 4060 4088 5067 6001 6040
Attempt 3	0 (0/225)	0.01778 (4/225)	0.00870 (2/230)	0.00870 (2/230)	0 (0/245)	0 (0/245)
UINs		0514 3088 4024 5068	1061 2025	2044 4010		

(# of false non-matches / # of genuine comparisons),
yellow highlight UINs (xxxx) matched offline but did not match online, green highlight UINs (xxxx) matched online but did not match offline

The differences between the offline and online match results are interesting because the same exact matching pairs of images are used in both cases. Discussions with Professor Daugman indicate that software releases of his algorithm for different cameras contain internal parameters that are optimized for specific cameras.[51] Enhanced software releases were tailored for specific cameras to take into account properties of the specific camera system, such as resolution, iris size, illumination wavelength, and camera illumination geometry. Within the Iridian-proprietary PrivateID environment, header information in the data packet specifies from which camera a given image originates, and several algorithm parameters are adjusted accordingly. Since we were unable to share with Professor Daugman which cameras would be evaluated during the IRIS06 effort, he provided a "one size fits all" release for the IRIS06 offline testing. A thorough evaluation of the various non-matching image pairs and camera-specific parameters to determine the root causes of the different behavior between online and offline matching performance is a subject for future study.

We conclude that for current commercially-available iris recognition products based on Professor Daugman's algorithm, offline performance results computed outside of the Iridian proprietary environment do not necessarily correspond to real-world online performance. In general, we surmise that unless the exact instantiation of an iris recognition product's algorithm that is used for online operation is used during offline testing, offline results will not necessarily indicate real-world performance for that product.

It is worth noting that subsequent attempts, in some cases, do not result in fewer false matches (more true matches). For example, in Figure 6-16 for Product A, Right eye, the 2nd attempt TMR is less than the 1st attempt TMR. Intuitively, we would expect TMR to increase with increasing attempts. See Appendix 11.4 for an explanation of this phenomenon.

TMR by attempt

While offline results do not necessarily align with online results for specific cameras, offline results do represent performance that can be expected in non-proprietary open-architecture environments. As such, we next turn our attention to the statistical significance of the TMR cumulative-attempt results. As an example, Figure 6-17 re-presents the Visit 1, Verify 1 Basic Cumulative Performance Curves by Attempt from Figure 6-15 with adjusted-Wald 95%

[51] Professor John Daugman, University of Cambridge, personal communication, 27 March 2007.

confidence intervals.[52] (The Hamming distance color map has been removed to improve clarity.) For comparison, the corresponding online TMR values are also plotted with adjusted-Wald 95% confidence intervals. The online data points are slightly offset from each other near the left vertical axis to improve clarity. Recall that these points do not have a corresponding FMR value. It is interesting to note that the confidence intervals for corresponding online and offline TMR values overlap, indicating that online matching performance and the offline matching performance, while not exactly the same, are statistically similar.

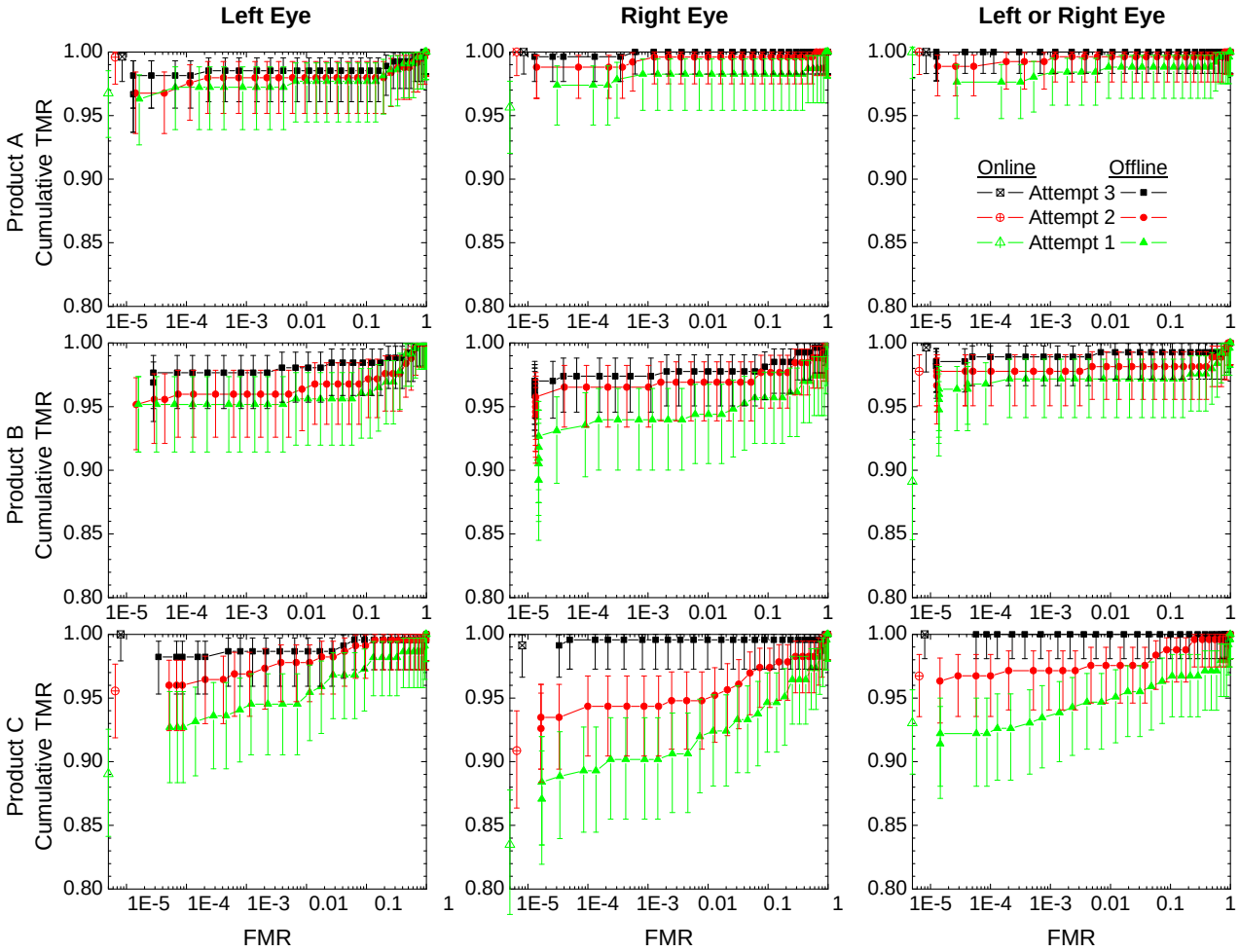


Figure 6-17. Visit 1, Verify 1 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

For Products A and B, TMR is statistically similar for all attempts for all iris-feature sets (confidence intervals overlap). There is a general trend to higher TMR values with increasing

[52] Note that the modified-Wald method is used here because there is only one comparison score for each test subject.

attempts for all products. (This performance improvement with attempts was also observed in the online FNMR analysis in Table 6-6.) The improvement between Attempt 1 and Attempt 3 is statistically significant for the right eye and for the left-or-right eye for Product C (no overlap between confidence intervals).

This trend to higher TMR with increasing attempts may indicate that the ability of the test subject-camera interface to obtain higher quality images improves with practice. By higher quality images, we mean images that have a better chance of matching the enrollment image. Recall that eyeglasses, if worn, were also removed for Attempt 3, which may contribute to improved Attempt-3 performance. (Performance without the influence of eyeglasses is presented in Figure 6-29 below.)

The trends observed for the Visit 1, Verify 1 transaction discussed above are also observed in the other transactions, which are presented in Appendix 11.5.1.[53] Although we observe a trend to higher TMR with increasing attempts, a clearer indication of performance with attempts is provided in the GTAR analysis below, which includes the influence of FTE and FTA as well as TMR and FMR.

TMR by transaction

To explore the performance of the images from each product as a function of time between enrollment and recognition and to compare the relative performance between products and iris-feature sets, we plot the cumulative 3rd-attempt basic ROC curves for each product, each transaction, and each iris-feature set in Figure 6-18. The data in Figure 6-18 indicates that in general the TMR performance of Products A and C are similar and that of Product B is slightly lower. However, there is substantial overlap of the confidence intervals between all products, for all transactions, and for all iris-feature sets, indicating that the data set does not support the statistical significance of this trend. The confidence intervals also overlap substantially between iris-feature sets (scanning horizontally in Figure 6-18) and between transactions (scanning vertically in Figure 6-18). We conclude that the matching performance of Professor Daugman's "irisenroll (release 1.5)" template generator and "matcher1" template matcher is fairly consistent with images from the three products tested, is about the same for left and right eyes, and changes little with time between enrollment and recognition.

[53] The offline basic and generalized ROC curves for all products, for all iris-feature sets, for all transactions, and for all attempts are presented in Appendix 11.5 for the convenience of the reader.

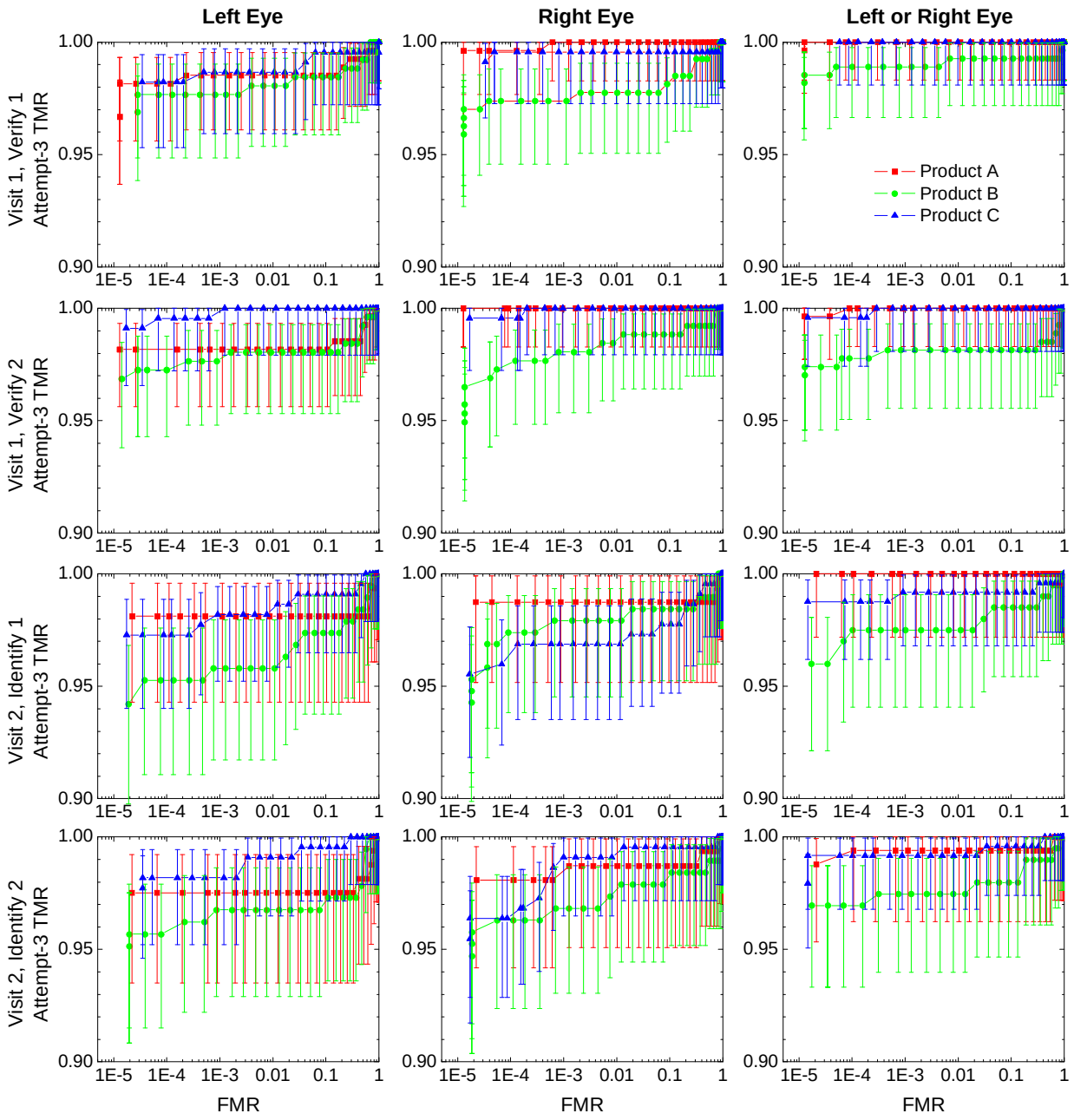


Figure 6-18. Basic Cumulative Performance Curves by Transaction with 95% Confidence Intervals

However, recall that these match results do not take into account eyes that failed to acquire or failed to enroll. The following GTAR analysis allows us to explore the overall performance of each product (as opposed to only the matching performance of the images from each camera) as a function of time and provides a clearer indication of the relative performance between products and between iris-feature sets.

GTAR by attempt

The generalized ROC curves include the influence of failure to enroll and failure to acquire along with false non-match and false match such that the overall performance of products can be compared. Failures to enroll and failures to acquire are essentially treated as failures to match (false non-matches) at all thresholds. The generalized ROCs allow us to estimate the percentage of the population that will not be able to successfully utilize the biometric product, and thus how many individuals will need to utilize a secondary or backup system.

To explore the overall performance of each product as a function of attempts, we present the generalized ROCs by cumulative attempt for Visit 1, Verify 1 in Figure 6-19. (Generalized ROCs for the other transactions are presented in Appendix 11.5.2.)

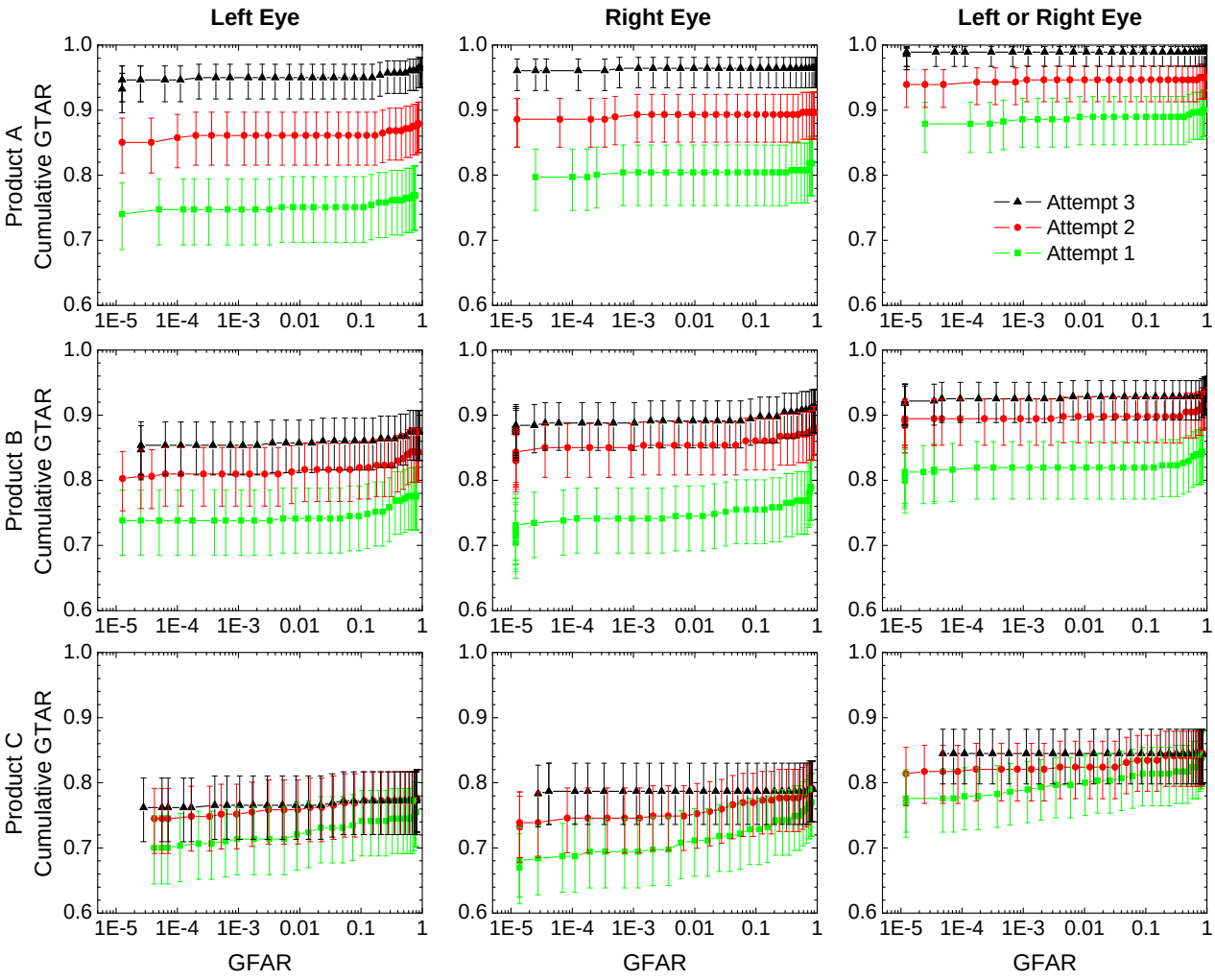


Figure 6-19. Visit 1, Verify 1 Generalized Performance Curves by Attempt

Figure 6-19 indicates a general trend to higher GTAR values with increasing attempts. This performance improvement with cumulative attempts was also observed in the online GFRR analysis in Table 6-9. For Product A, the improvement from Attempt 1 to Attempt 2 to Attempt 3 is statistically significant (confidence intervals do not overlap) for the left and right eyes. For the left-or-right eye feature set, GTAR for Attempt 3 is statistically higher than the GTAR for Attempt 1. For Product B, GTAR for Attempt 3 is statistically higher than the GTAR for Attempt 1 for all feature sets, while for Product C GTAR is statistically similar for all attempts for all iris-feature sets (confidence intervals overlap). Similar results are also observed for the other transactions.

As with TMR, the removal of eyeglasses for the 3rd attempt may contribute to the improved Attempt-3 performance. However, since FNMR changes little with increasing attempts, and FTE is a constant for each feature set, we conclude that the improvement in FTA with increasing attempts is predominantly responsible for higher GTARs with increasing attempts. In general, the overall recognition performance for Products A and B improves with increasing attempts while the overall recognition performance of Product C remains about the same. Recall from Table 6-2 that FTA improves substantially with attempts for Products A and B and improves only slightly with attempts for Product C. The ability of the test subject-camera interface to acquire useable images improves substantially with short-term practice for Products A and B and to a lesser extent for Product C.

Left-eye versus right-eye performance

We next turn our attention to the performance of right eyes versus the performance of left eyes. We observe in Figure 6-19 that the curves for left and right eyes are quite similar. To demonstrate this more clearly, we plot the 3rd-attempt generalized ROC curves for Visit 1, Verify 1, for left and right eyes for all three products along with the confidence intervals in Figure 6-20. The confidence intervals for the left and right eyes overlap substantially for all products. This was also observed in the online results for Products A and C. While Product B did not provide individual eye match results online, the right and left iris images are available for offline analysis, and we note that left and right eyes exhibit roughly the same performance for Product B as well as for Products A and C.

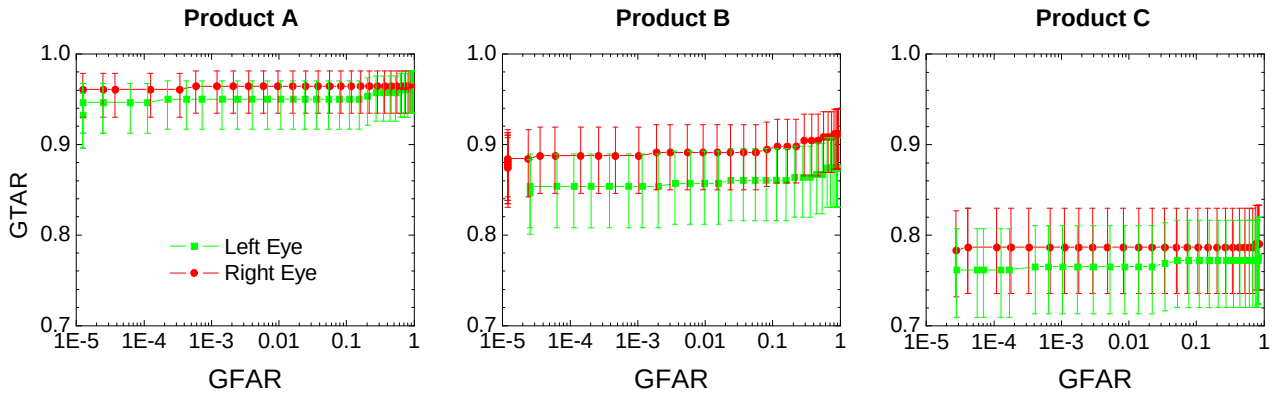


Figure 6-20. Visit 1, Verify 1, 3rd-Attempt Performance Curves for Left and Right Eyes

We conclude that right and left eyes exhibit statistically similar iris recognition performance. Recall that this was also observed in the online data. As such, for the remainder of the offline analysis, we will present combined "left and right eye" results. These curves are the average of the left eye and right eye curves and represent the performance of a single eye. This case is particularly relevant for single-eye iris recognition cameras where only one eye is presented in a single attempt. As before, the "left or right eye" curves represent performance when a successful recognition is defined as "either the left or the right eye matches". This case is particularly relevant for two-eye iris recognition cameras where both eyes are presented in a single attempt.

GTAR by transaction

Figures 6-21 and 6-22 present the generalized 3rd-attempt ROC curves for each transaction and each camera for both the single-eye and the left-or-right eye feature sets. Figure 6-21 includes the Hamming distance color map to indicate the threshold score for each point. The same data is presented in Figure 6-22 with 95% confidence intervals sans the color map for clarity. We observe in these figures that GTAR increases slightly with increasing GFAR but generally remains relatively flat over a large range of GFAR values. The relative stability of GTAR for many GFAR values is one of the attributes of Professor Daugman's iris recognition algorithms.

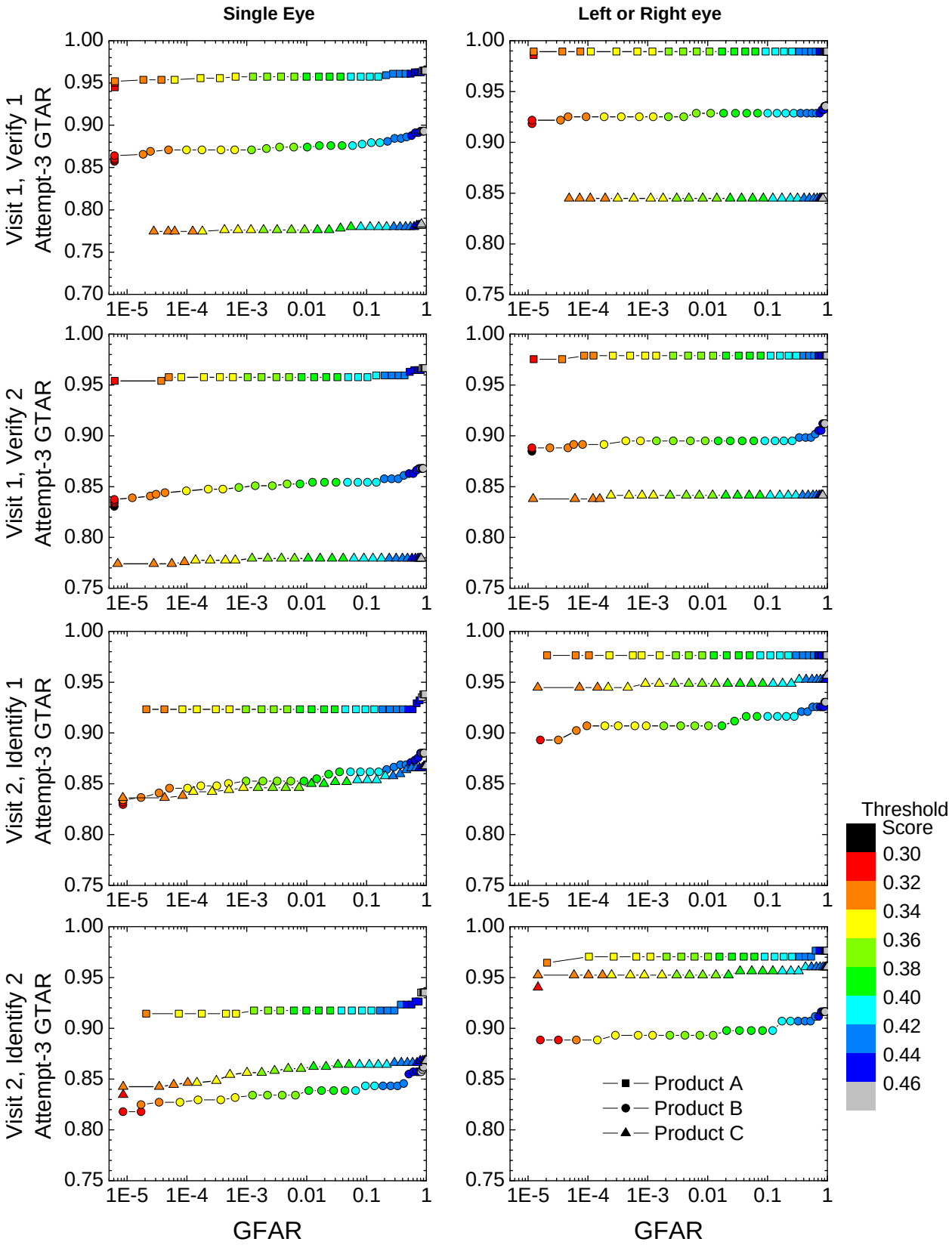


Figure 6-21. 3rd-Attempt Performance Curves by Transaction with Scores

As discussed previously, we presume that most commercial products that are based on Professor Daugman's iris recognition algorithm operate in the low-GFAR range of the operating curves (per the color map, the Hamming distance of 0.32 occurs at the interface between the red and orange points). For access control applications, increasing GFAR decreases security (more impostors can gain access), and increasing GTAR improves convenience (easier for authorized user to gain access). For identification applications, increasing GFAR increases the number of false identifications (decreasing convenience), and increasing GTAR increases the possibility that a person of interest will be identified (increasing security). The appropriate operating point, as determined by the selected threshold setting (Hamming distance), should be determined based on the requirements of each operational scenario. While, the iris recognition products evaluated during IRIS06 do not typically allow modification of the default threshold setting, there is little benefit to setting the threshold to a higher Hamming distance value as this would increase GFAR but have little effect on or slightly increase GTAR.

Figures 6-21 and 6-22 also reveal the relative performance between products and between iris-feature sets for each transaction. The left-or-right-eye feature sets typically have a slightly higher GTAR (accuracy) than the corresponding single-eye feature sets for each product and each transaction. Recall that the Attempt-3 GTAR for the single-eye feature set represents the recognition performance after three attempts with one eye (best of three tries). Similarly, the left-or-right-eye Attempt-3 GTAR represents the recognition performance after three attempts with two eyes (best of six tries – three tries with the left eye and three tries with the right eye). As such, we would expect the left-or-right-eye recognition performance to be better than the single eye performance. However, when we inspect the confidence intervals in Figure 6-22 (scanning horizontally for each product), we find that with two exceptions, the single-eye and left-or-right eye confidence intervals overlap. For the two Visit 2, Product C transactions, the single-eye and left-or-right eye confidence intervals do not overlap, indicating that the 3rd -attempt performance improvement using two eyes compared to one eye is statistically significant. For all other product transactions, the recognition performance for 3 tries with two eyes is not significantly better than for 3 tries with one eye.

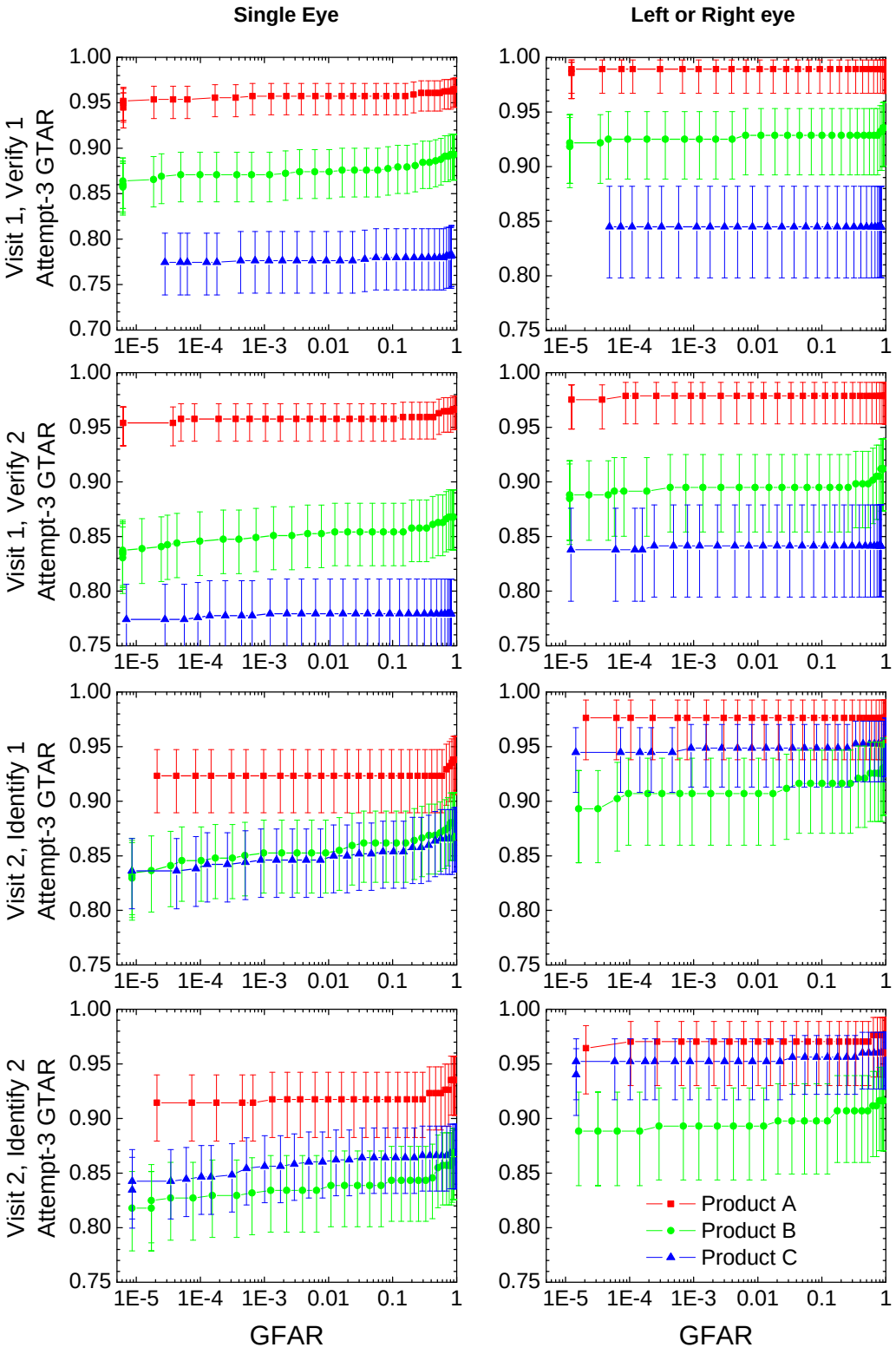


Figure 6-22. 3rd-Attempt Performance Curves by Transaction with 95% Confidence Intervals

Turning our attention to the relative performance between products, we observe that Product A generally performs better (higher GTAR) than Product B for all transactions with little or no overlap between the respective confidence intervals in Figure 6-22. We conclude that $GTAR_A > GTAR_B$ for both single-eye and left-or-right-eye feature sets. The GTAR for Product C is not consistent and is discussed further below.

GTAR by visit

We also observe that the performance for each product and iris-feature set is statistically similar within a given visit (confidence intervals overlap scanning vertically in Figure 6-22 within Visit 1 and within Visit 2). This indicates that recognition attempts separated by about 15 minutes yield statistically similar performance. As such, we combine the intra-visit recognition transactions and plot in Figure 6-23 the generalized ROCs by visit. Note that since there are now two comparison samples available for each test subject, one for each of the recognition transactions per visit, the Logit Beta-binomial method is used to compute confidence intervals. Recall also that when the assumptions required to obtain valid confidence interval estimates are not satisfied, namely $np>5$, confidence intervals are not shown.

Figure 6-23 shows that the performance of Product C changes substantially between Visits 1 and 2, probably because of the lighting change between visits. Figure 6-24 presents the data in Figure 6-23 in a slightly different format to more easily compare the performance changes between visits. For Products A and B, the performance for Visit 1 is slightly better than that for Visit 2 for both single-eye and left-or-right-eye feature sets. This trend could be caused by the 6-week time separation between Visits 1 and 2, or by the lighting change between Visits 1 and 2. However, since the confidence intervals overlap, the data set does not indicate that the trend is statistically significant. For Product C, we see the opposite effect; the performance for Visit 2 is substantially better than that for Visit 1 for both feature sets. The difference is statistically significant except at very low GFAR values for the single-eye feature set. We conclude that the lighting change between Visits 1 and 2 had the desired effect; it improved the performance of Product C without substantially degrading the performance of Products A and B.

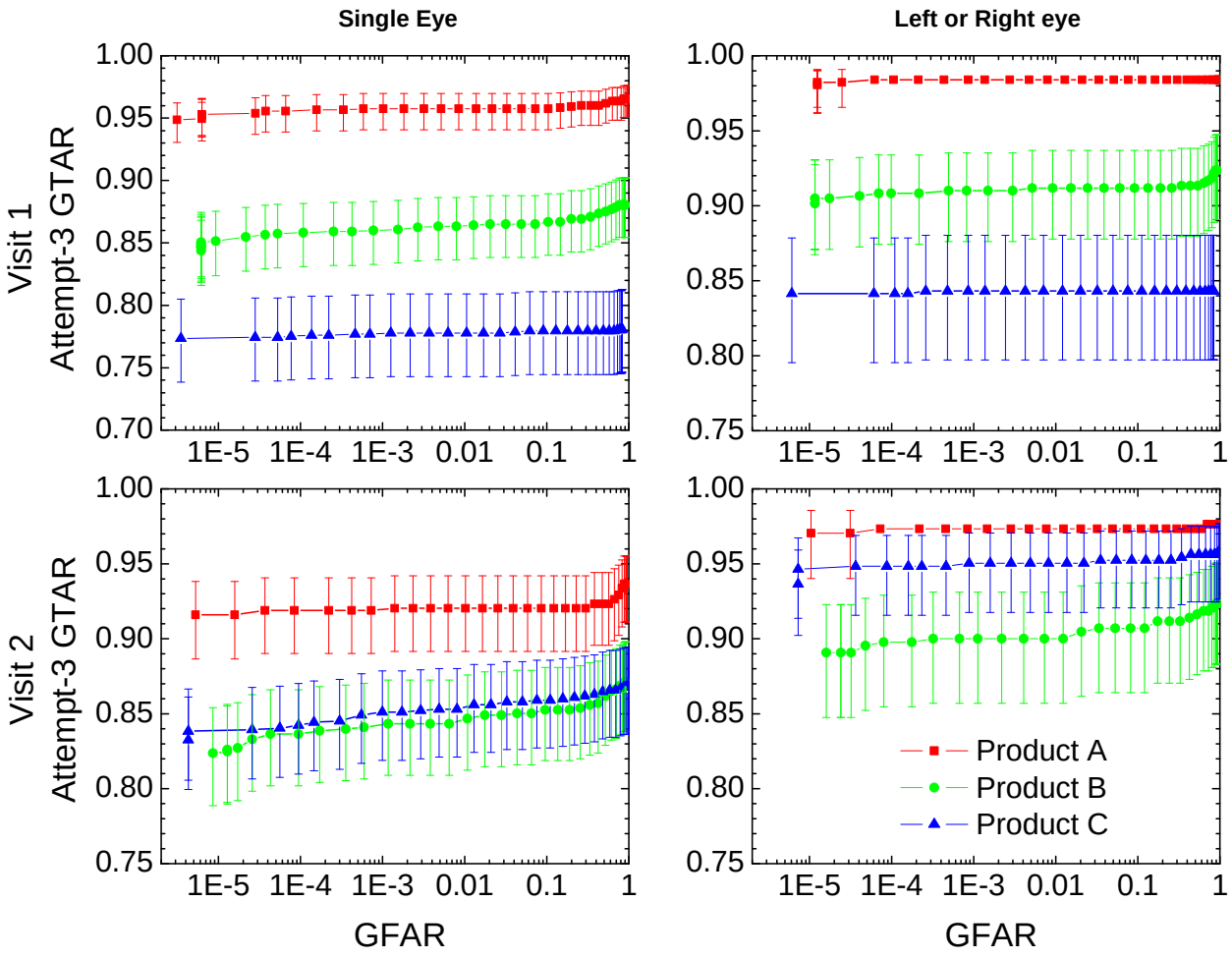


Figure 6-23. 3rd-Attempt Performance Curves by Visit with 95% Confidence Intervals

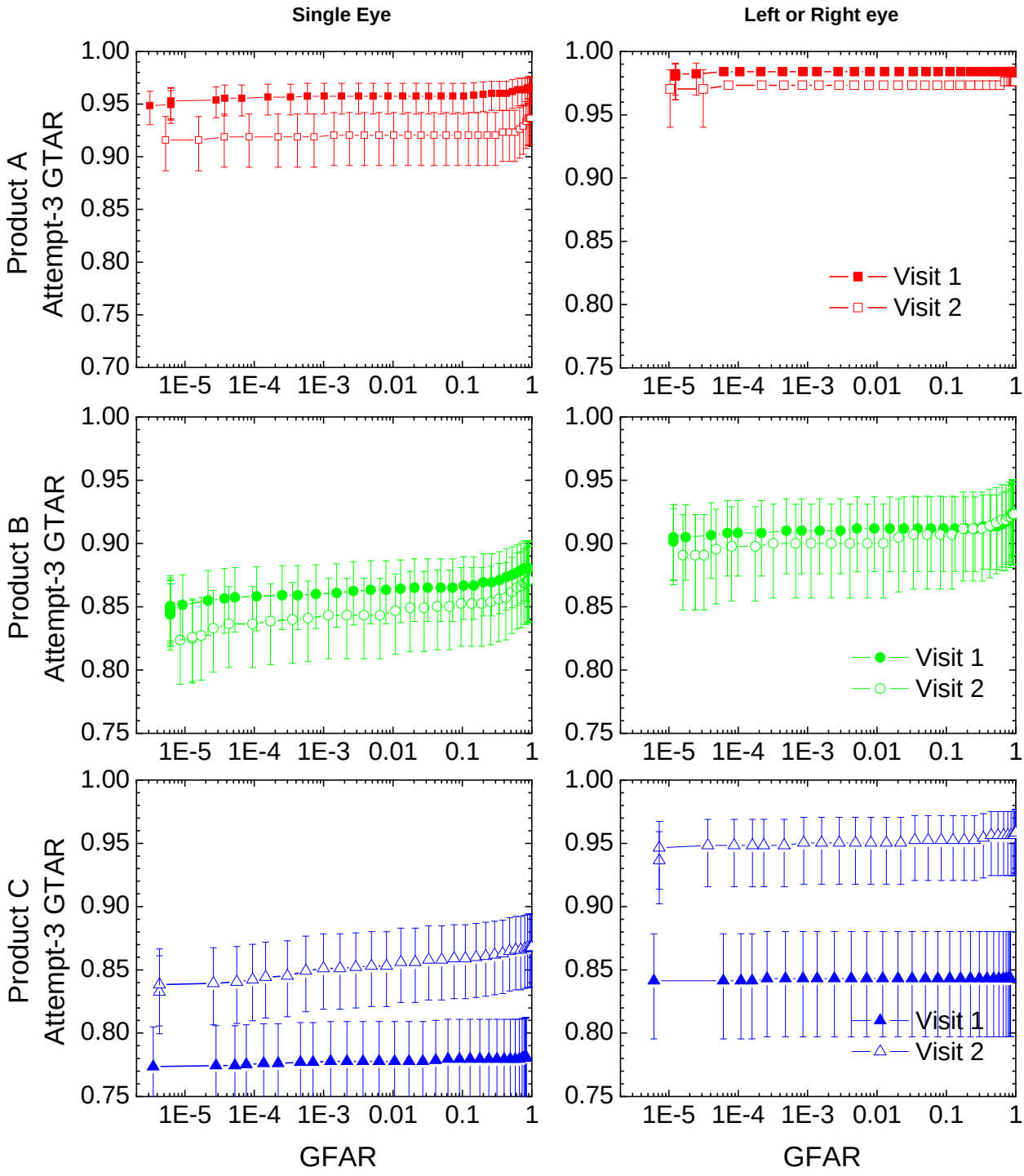


Figure 6-24. 3rd-Attempt Performance Curves by Visit by Product with 95% Confidence Intervals

GTAR overall

We next compare the overall performance (the combination of the 3rd-attempt results for all transactions) for the three products. The results presented in Figure 6-25 suggest that averaged over all transactions, Product A exhibits the best generalized performance, and Products B and C exhibit similar generalized performance (3rd-attempt $GTAR_A > GTAR_B \sim GTAR_C$). Recall that this conclusion was also reached from the online analysis. Overall TMR results are also presented in Figure 6-25 for comparison, and the GTAR and TMR performance results at a Hamming distance of 0.32 are summarized in Table 6-14.

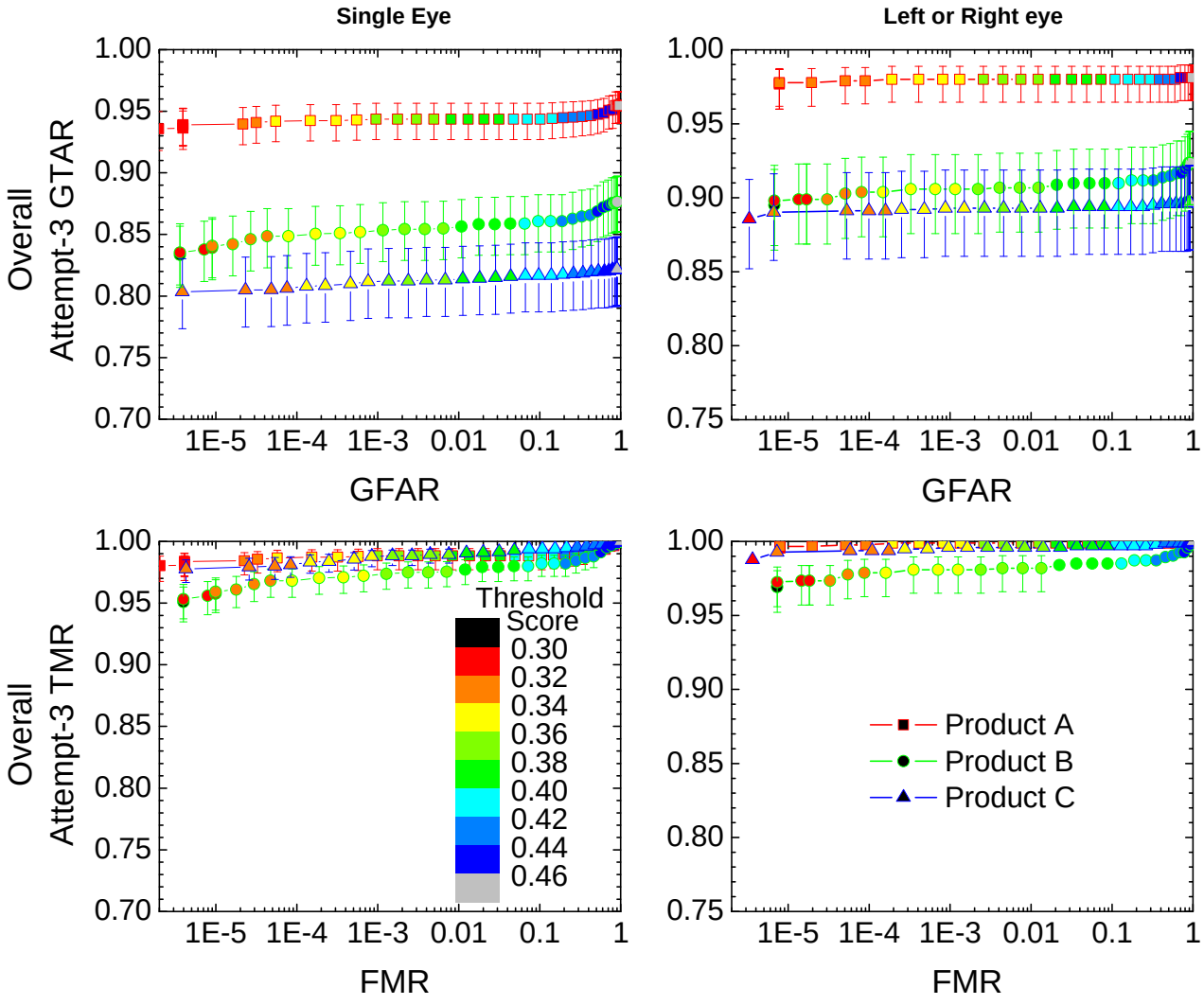


Figure 6-25. Overall 3rd-Attempt Performance Curves with 95% Confidence Intervals

Table 6-14. Overall Offline Attempt-3 GTAR and TMR at HD=0.32						
	GTAR (%)		**TMR (%)**		**Average Recognition Transaction Time (sec)**	
	Single Eye	Left or Right Eye	Single Eye	Left or Right Eye	Single Eye	Left or Right Eye
Product A	93.9	97.8	98.4	99.7	18.0	21.4
Product B	84.1	89.9	95.9	97.3		7.9
Product C	80.3	89.0	97.8	99.3	9.3	11.2
See Figures 6-12 and 6-25 for confidence intervals						

From Table 6-14 we conclude that when using a "three tries with one eye" decision policy, ~6% of the user population will not be able to use Product A, ~16% will not be able to use Product B, and ~20% will not be able to use Product C. However, when failures to enroll and failures to acquire (difficult cases) are not considered, only ~1.6% of the user population will not match when using Product A, ~4.1% will not match when using Product B, and ~2.2% will not match when using Product C. The performance improves substantially when both left and right eyes are used, such as with two-eye systems: ~0.3% of the population will not match using Product A, ~2.7% will not match using Product B, and ~0.7% will not match using Product C.

Table 6-14 also shows the associated average recognition times (from Table 6-11). As with the online analysis, we observe an inverse relationship between matching performance and recognition transaction times. While Product B has the lowest TMR values (lowest accuracy) at HD=0.32, it also has the shortest transaction times. While Product A has the highest TMR (highest accuracy), it has the longest transaction times.

Multiple attempts versus multiple features

We next turn our attention to the accuracy of performing several recognition attempts with the same feature compared to the accuracy of performing a single recognition attempt with multiple features. For example, is it more accurate to present the right eye twice or to present both the left and right eye once? Figure 6-26 compares these two scenarios for both GTAR and TMR by combining the cumulative 2nd-attempt data for left and right eye feature sets for each of the four transactions and by combining the 1st attempt data for the left-or-right eye feature set for each of the four transactions. We observe substantial overlap of the confidence intervals in all cases with the exception of

Product C GTAR. In this case, the CIs overlap, but not to a large extent. (Recall that if the assumptions required for CIs are not met CIs are not shown, such as for Product A TMR.) Figure 6-26 indicates that we can expect similar performance presenting one eye twice or presenting two eyes once. We assume that for a single-eye camera presenting two eyes twice would take longer, however we have not computed the associated time durations.

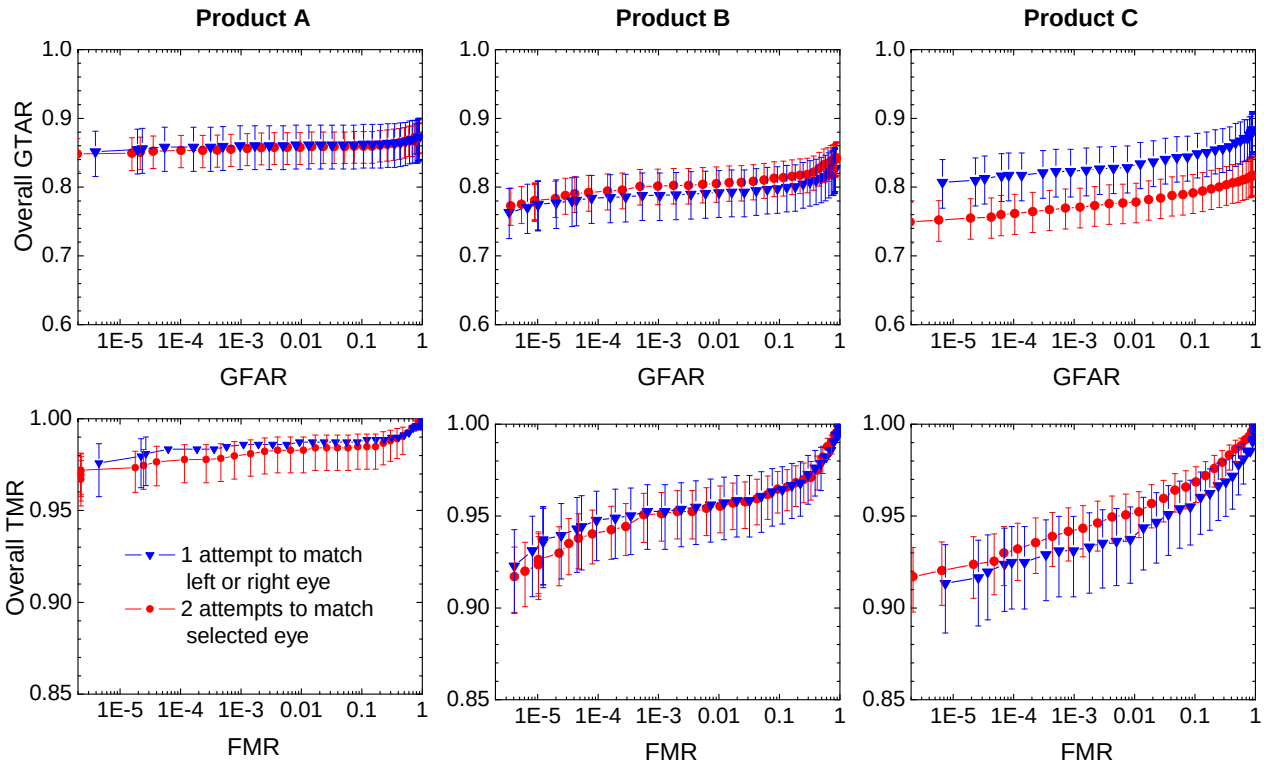


Figure 6-26. Two Attempts versus Two Features (combining all four transactions)

To expand upon this analysis, we plot in Figure 6-27 additional multiple-attempt and multiple-feature transactions. The general trend, from lowest accuracy to highest accuracy, is presented in Table 6-15. The overlap of the confidence intervals, and thus the statistical significance of the difference, for each product varies as illustrated in Figure 6-27. Note that the CIs are narrower than in previous analyses since comparison scores from all four transactions are combined, and more comparison scores yield smaller CIs. We note that for Product A, the TMR curves significantly overlap compared to the other products, while the GTAR curves show substantial separation. This may indicate that the GTAR improvement with additional attempts and features for Product A is more attributable to an improvement in FTA than to an improvement in matching performance.

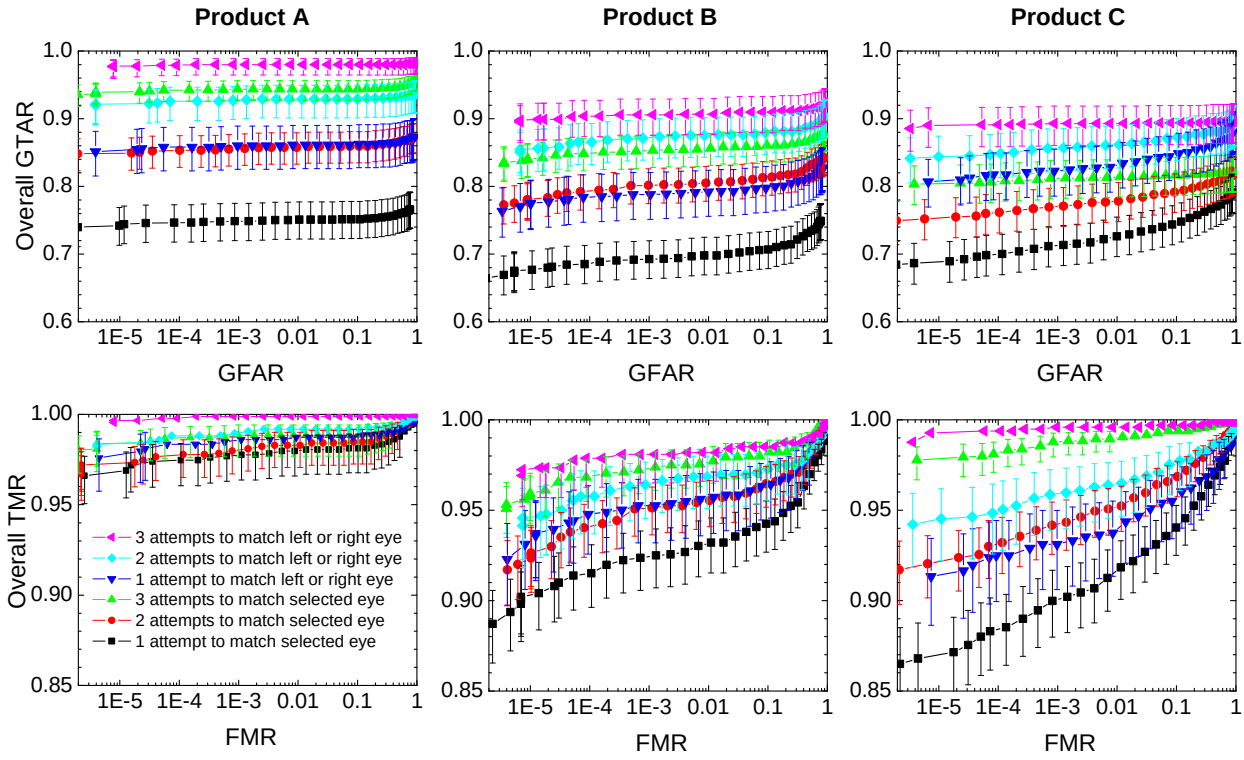


Figure 6-27. Multiple Attempts versus Multiple Features (combining all four transactions)

Table 6-15. Accuracy Ranking General Trends	
1	3 attempts to match left or right eye
2	2 attempts to match left or right eye ≈ 3 attempts to match single eye
3	1 attempt to match left or right eye ≈ 2 attempts to match single eye
4	1 attempt to match single eye

The above analyses were performed by combining 1st, 2nd, and 3rd cumulative attempt data from the four transactions (Visit 1–Verify 1, Visit 1–Verify 2, Visit 2–Identify 1, and Visit 2–Identify 2) for both one-eye and two-eye feature sets. For example, a 3rd-attempt analysis for the two-eye feature set utilizes the best match score from six comparisons: right eye enrollment with 1st right eye attempt, right eye enrollment with 2nd right eye attempt, right eye enrollment with 3rd right eye attempt, left eye enrollment with 1st left eye attempt, left eye enrollment with 2nd left eye attempt, and left eye enrollment with 3rd left eye attempt. The best scores for each test subject are used to generate the ROC curve for the selected transaction or combination of transactions (Visit 1, Visit 2, or Overall). This represents a combined-cumulative-multiple-attempt level-of effort approach as described in Section 5.1.1.

Interoperability

We now perform a combined-single-attempt analysis treating the left and right eyes from each test subject as two separate individuals. For this analysis, the match scores between the enrolled eye and every single attempt for that eye are utilized to create the ROC curves. This ignores the improvement of the user-sensor interface with successive attempts and represents the average performance over all attempts from all transactions. The Logit Beta-binomial method is used to compute confidence intervals since there are multiple comparison scores for each individual eye.

Further, all results presented so far have demonstrated native performance for each product. That is, match scores are generated from enrollment and recognition samples from the same product. We explore in Figure 6-28 interoperability performance, where enrollment and recognition samples from different products are compared. In the legend in Figure 6-28, "Product M x Product N" indicates that Product M enrollment samples and Product N recognition samples are compared. Native curves are indicated by yellow-filled symbols.

The first row of curves in Figure 6-28 show performance when enrollment samples from one product are compared to recognition samples from all three products. For example, we

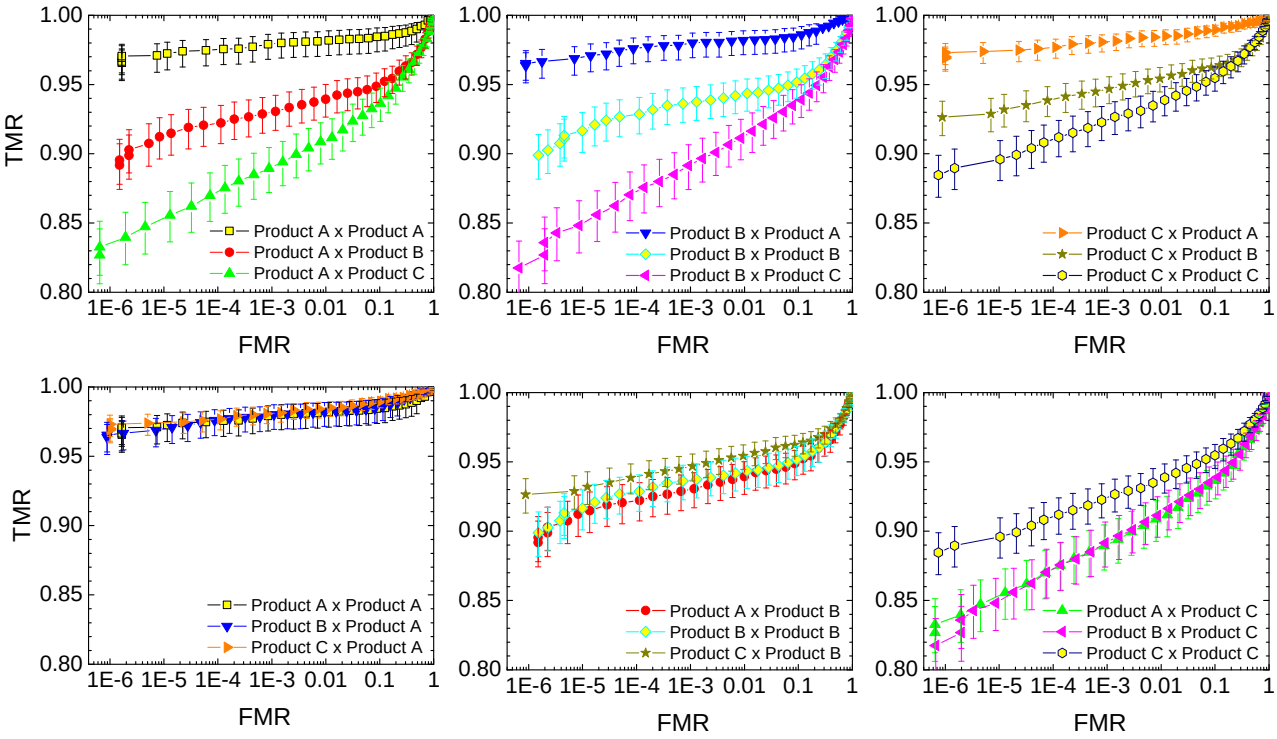


Figure 6-28. Single-Attempt Native and Interoperability Performance (combining all attempts)

observe that Product A enrollment samples have higher accuracy when they are compared with Product A recognition samples. The performance with Product B recognition samples is somewhat degraded, and the performance with Product C samples is further degraded. That is, native performance is better than interoperable performance. However, for Product B enrollment samples, we observe that the best performance is obtained with Product A recognition samples, followed by Product B and then Product C recognition samples. In other words, Product B enrollment samples achieve higher accuracy using recognition samples from a different product – interoperability performance is better than native performance. For Product C enrollment samples, we observe that the highest accuracy is obtained with Product A recognition samples, followed by Product B; and Product C recognition samples actually have the lowest accuracy. Again, interoperability performance is better than native performance.

The second row of curves in Figure 6-28 show performance when recognitions samples from one product are compared to enrollment samples from all three products. (This is the same data presented in the first row but organized differently.) We note that very good performance is obtained whenever Product A recognition samples are used, and slightly degraded performance is obtained when Product B recognition samples are used. Product C recognition samples exhibit the worst performance except when using native Product C enrollment images.

Table 6-16 summarizes the native and interoperability performance at FMR=10^{-5}. We see that for recognition, the highest mean TMR across all products is obtained when using Product A recognition samples (TMR_{mean}=97.2%) and the lowest mean TMR is obtained using Product C recognition samples (TMR_{mean}=86.6%). For enrollment, the highest mean TMR across all products is obtained using Product C enrollment images (TMR_{mean}=93.4%). In summary, the best performance is obtained when using Product C enrollment samples and Product A recognition samples. Recall that the same template generator and the same matching algorithm were used for all images. Further research into the nature of the differences between the images from the different cameras is needed to understand the observed improvement in interoperability performance compared to native performance.

Table 6-16. Single-Attempt Interoperability TMR (%) at FMR=10^{-5} (0.001%)					
		Recognition Samples			
		Product A	Product B	Product C	Mean
Enrollment Samples	Product A	97.2	91.3	85.3	91.3
	Product B	97.0	91.7	84.9	91.2
	Product C	97.4	93.1	89.6	93.4
	Mean	97.2	92.0	86.6	91.9
See Figure 6-28 for confidence intervals.					

Performance excluding eyeglasses

All of the analyses presented above include the influence of eyeglasses. While enrollment was performed without eyeglasses, all test subjects that wore glasses performed the first two recognition attempts for each transaction while wearing their eyeglasses and removed their eyeglasses for the third attempt. Figure 6-29 presents the single-attempt ROC curves generated with all images (with and without glasses) and the single-attempt ROC curves generated only with images without glasses (glasses excluded). We observe little difference in performance with and without glasses for Product A. However for Products B and C, we observe statistically-significant improvement in performance when glasses are not worn. We conclude that iris images acquired with and without glasses match equally well for Product A. However, for Products B and C, matching performance is degraded for iris images acquired with glasses.[54]

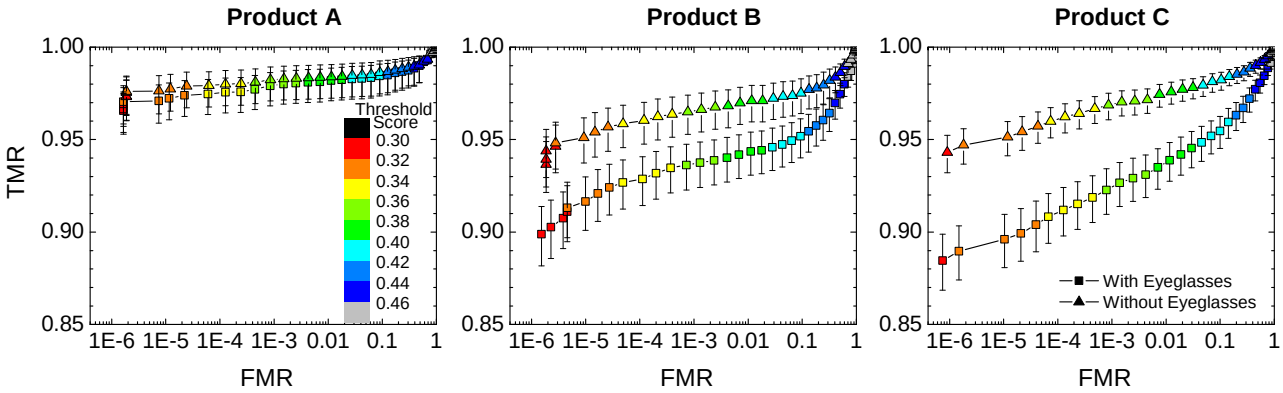


Figure 6-29. Single-Attempt Performance with and without Eyeglasses (combining all attempts)

[54] Note that FTE and FTA are not included in this attempt-level analysis. The transactional influence of glasses is a subject for future analysis.

Performance excluding flagged images

Finally, we investigate the performance of the three iris recognition products when all images with performance-relevant image-property flags are excluded from the analysis. Flagged image properties include glasses, hard contacts, bad environment, bad feature, obstructions, bad picture and bad placement as described in Section 5.2.1. Figure 6-30 presents the single-attempt ROC curves for 1) all images (labeled "With Eyeglasses"), 2) all images without eyeglasses, and 3) all non-flagged images (labeled "Without Flagged Images"). The first row in Figure 6-30 shows performance data sorted by product as a function of image property filter, and the second row shows the same data sorted by image property filter as a function of product.

We see that the performance for Product A is statistically similar for all image property filters, indicating that all images for this product match well. For Products B and C we see an improvement in performance when flagged images are excluded from the analysis, indicating that some images from these cameras are of poor quality. We also observe that the performance for Product B and C is similar with all images and without glasses, and the performance for

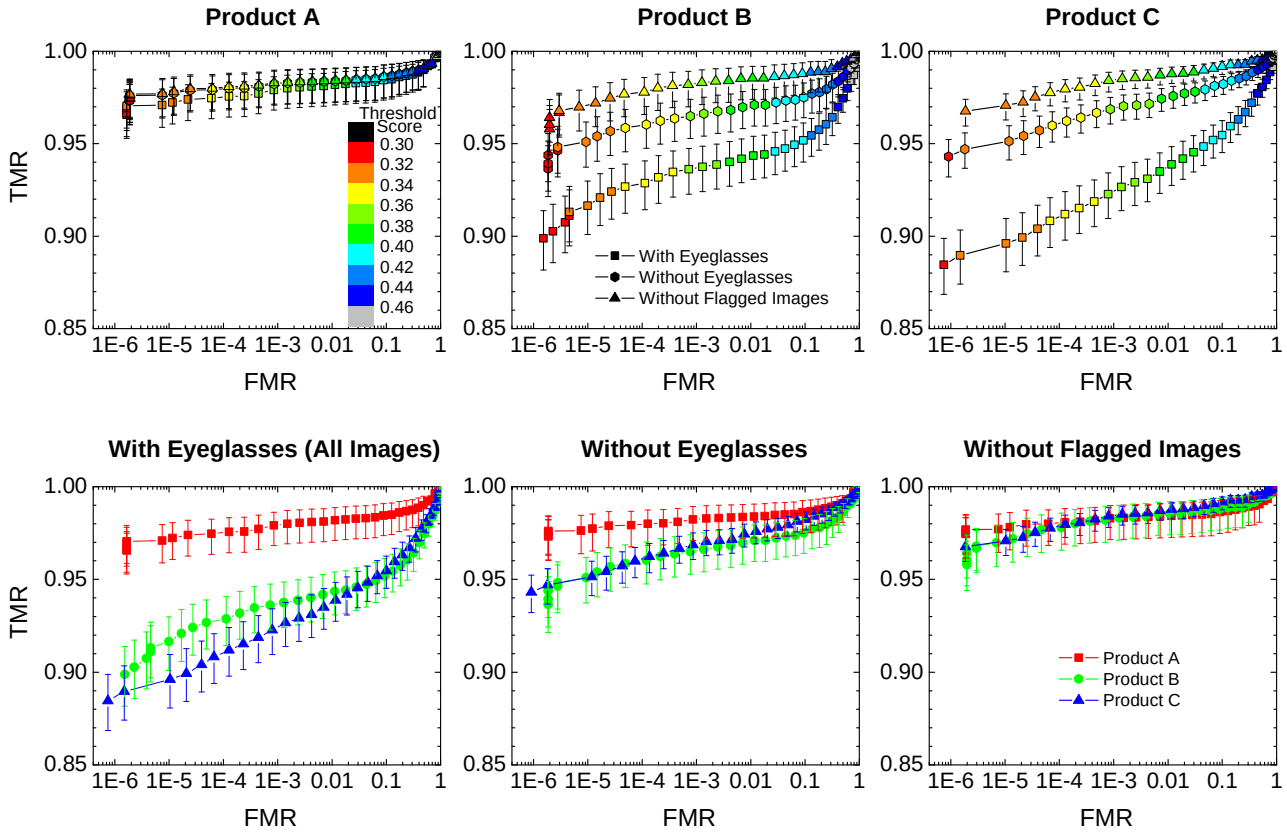


Figure 6-30. Single-Attempt Performance with and without Flagged Images (combining all attempts)

Product A is superior in both cases. However, when all flagged images are excluded, we note that Products A, B, and C exhibit similar performance. That is, iris images from all three products perform equally well (with Professor Daugman's algorithms) when poor quality images are excluded.[54]

For completeness, we present the associated genuine and impostor match score distributions in Figure 6-31. As expected, for Products B and C we observe a change in the overlap between the genuine and impostor distributions with glasses, without glasses, and without flagged images. More specifically, the overlap of the genuine scores (green bars) with the impostor scores (red bars) decreases when eyeglasses are excluded, and decreases further when flagged images are excluded for Products B and C. For Product A, we observe no

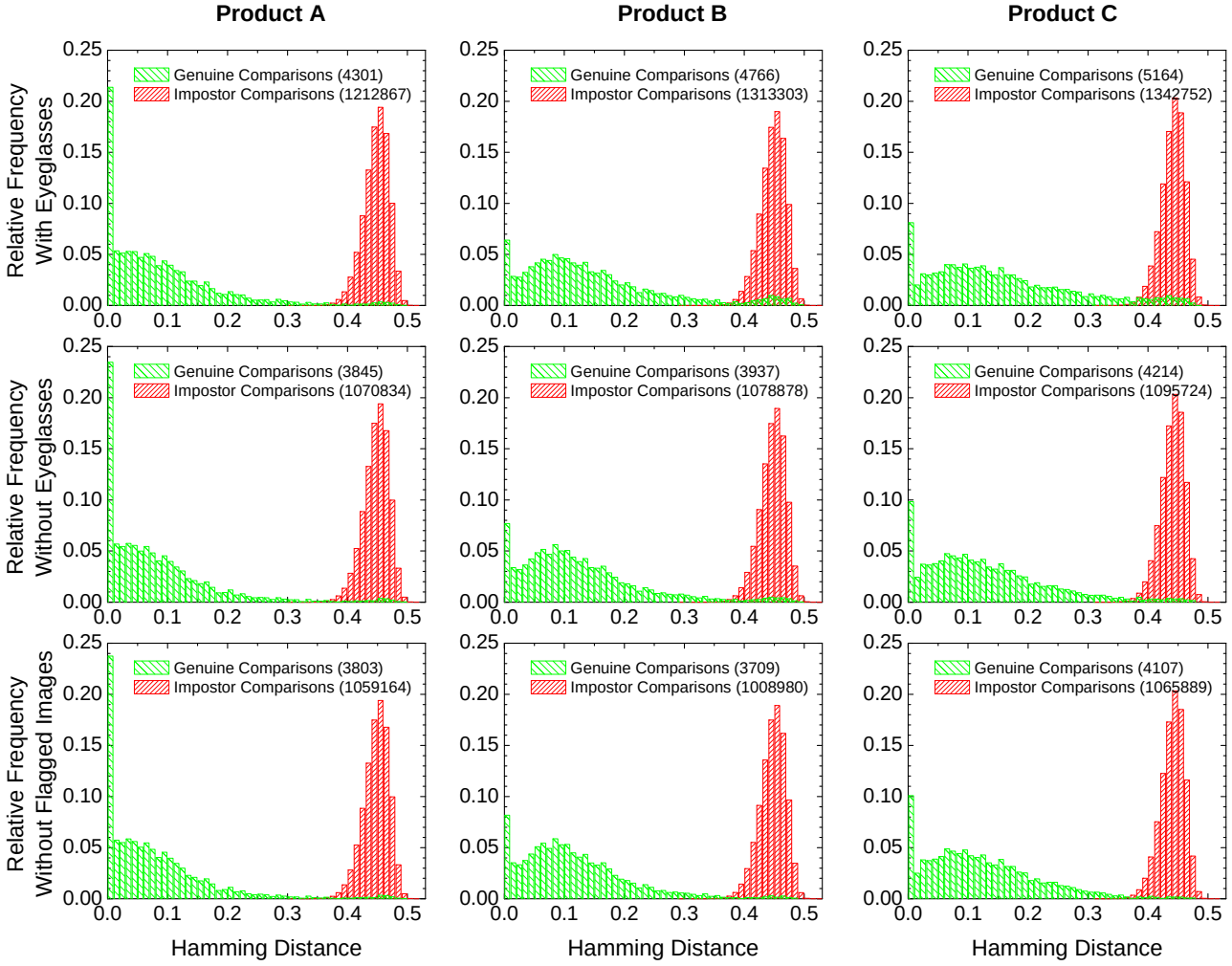


Figure 6-31. Single-Attempt Histograms with and without Flagged Images (combining all attempts)

significant changes in the score distributions as a function of image property filters. We also observe that the genuine score distribution for Product A is shifted to lower Hamming distance scores relative to Products B and C.

Summary of key offline findings

- Unless the exact instantiation of an iris recognition product's algorithm that is used for online operation is used for offline testing, offline performance results do not necessarily indicate online performance.
- The matching performance of Professor Daugman's "irisenroll (release 1.5)" template generator and "matcher1" matching algorithms is fairly consistent with images from the three products tested, is about the same for left and right eyes, and changes little with time (up to six weeks) between enrollment and recognition.
- Cumulative true match rates tend to increase with increasing attempts.
- Cumulative generalized true accept rates, which take failure to enroll and failure to acquire into account, tend to increase with increasing attempts, predominantly due to a decrease in failure to acquire rates with increasing attempts.
- Right and left eyes exhibit statistically similar iris recognition performance.
- The relative cumulative 3rd-attempt generalized true accept rate (GTAR) between products is $GTAR_A > GTAR_B \sim GTAR_C$.
- For the left-or-right eye feature set, Product A exhibited the highest cumulative 3rd-attempt generalized (GTAR=97.8%) and basic (TMR=99.7%) performance at the operational threshold Hamming distance of 0.32.
- Matching performance has an inverse relationship with transaction times. Higher accuracy requires longer transaction times, and faster transaction times result in lower accuracy.
- Interoperability matching performance is better than native matching performance. Specifically, the best matching performance is obtained using enrollment images from Product C and recognition samples from Product A.
- Matching performance is degraded for Products B and C when iris images are acquired when test subjects are wearing eyeglasses. Product A matching performance is similar both with and without glasses.
- Product A collected the highest percentage of high quality iris images but also exhibited the longest transaction times.
- When all images of questionable quality are excluded, the images from Products A, B, and C exhibit similar matching performance with Professor Daugman's algorithms.

6.2 Off-Axis

6.2.1. Guided gaze experiment

For the guided gaze experiment, test subjects were asked to gaze nominally 20º up, 20º down, 28º to the left, and 28º to the right relative to the center of each camera as described in Step 12 of the scenario evaluation test protocol (p. 32). One verification attempt was performed in each off-axis gaze direction, and one verification attempt was performed in the neutral position looking directly at the center of each camera. Glasses, if worn, were removed for all attempts. Approximately 250 test subjects participated in the off-axis gaze experiment, which was performed during Visit 2.

To analyze the off-axis data, templates were generated from the off-axis and neutral gaze images using Professor Daugman's irisenroll (release 1.5) algorithm and compared to the "ideal" Visit 1 enrollment templates using Professor Daugman's matcher1 algorithm. The images were thoroughly reviewed and flagged prior to analysis. Specifically, all images where the eyes were closed or nearly closed and where glasses were accidentally worn were flagged and excluded from the analysis. Since we could not force the test subjects to look at the targets mounted on or around the cameras (we could only request that they do so), we carefully inspected each and every image to ensure that the image was indeed off axis in the appropriate direction. This often entailed determining where the reflection was positioned relative to the pupil or where the iris was positioned relative to the eye socket. All images where the eye was not looking in the appropriate direction were flagged and excluded from the analysis. In general we note that the off-axis angles are greater than zero and less than or equal to the 20° up and down and less than or equal to 28° left and right. Further, all of the off-axis irises are within the frame of the image though most are not dead-center in the image.

The off-axis gaze results are summarized in Figure 6-32. For all products, matching performance was significantly degraded when gazing downwards and when gazing away from the nose (left eye gazing left and right eye gazing right). For Products A and B, the matching performance for neutral gaze and upward gaze are about the same, and performance for gaze towards the nose (left eye gazing right and right eye gazing left) was slightly lower. For Product C, performance was about the same when gazing upward and towards the nose but significantly lower than neutral performance. We are not at liberty to discuss camera geometries

here, which likely influence off-axis gaze performance. However in general matching performance is degraded when looking down and away from the camera center, perhaps because of excessive occlusion of the iris by the eyelids.

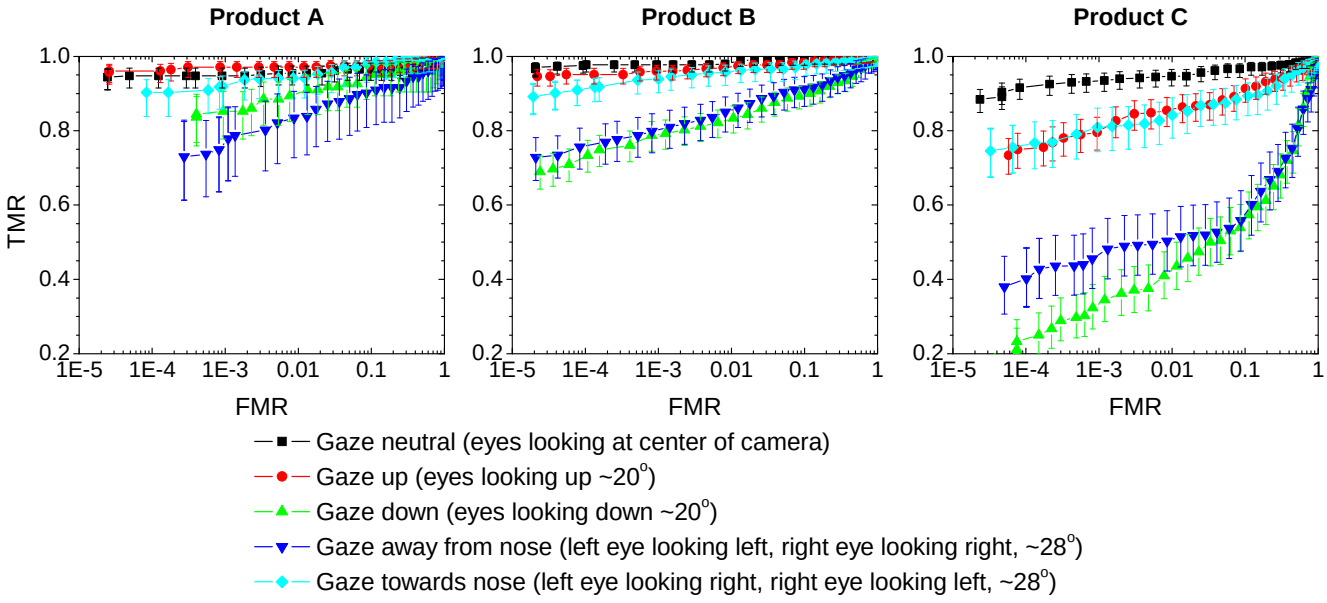


Figure 6-32. Single-Attempt Off-Axis Gaze Performance

6.2.2. Controlled pose experiment

For the controlled off-axis pose experiment, the TMR results of the Neutral, Sweep, and Translate procedures (as described in Sections 4.4 and 5.4) are presented using figures that best illustrate the procedure as appropriate for each camera. All of the figures use the TMR color key presented in Figure 6-33 to indicate the level of performance (perfect, good, moderate, poor, and bad) at various poses. Depending on how many test subjects participated in each procedure the actual percentages for Good (lime) and Poor (orange) vary. Recall that if a procedure is not reported, it can be assumed that the product simply did not work off-axis in that case.

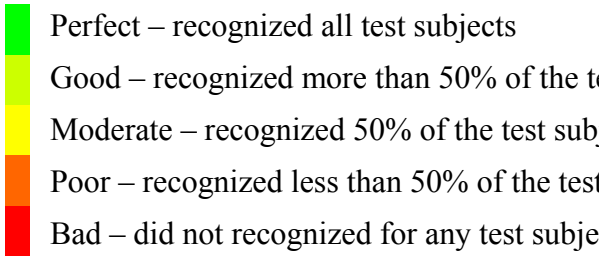


Figure 6-33. Controlled Pose TMR Color Key

Product A

Results of the Neutral Z procedure for Product A are illustrated in Figure 6-34. Recall that the Neutral Z procedure emulates a user located directly in front of and looking directly at the camera from various distances. Both left and right eyes performed Perfect or Good for Z distances of about 9 to 13 cm.

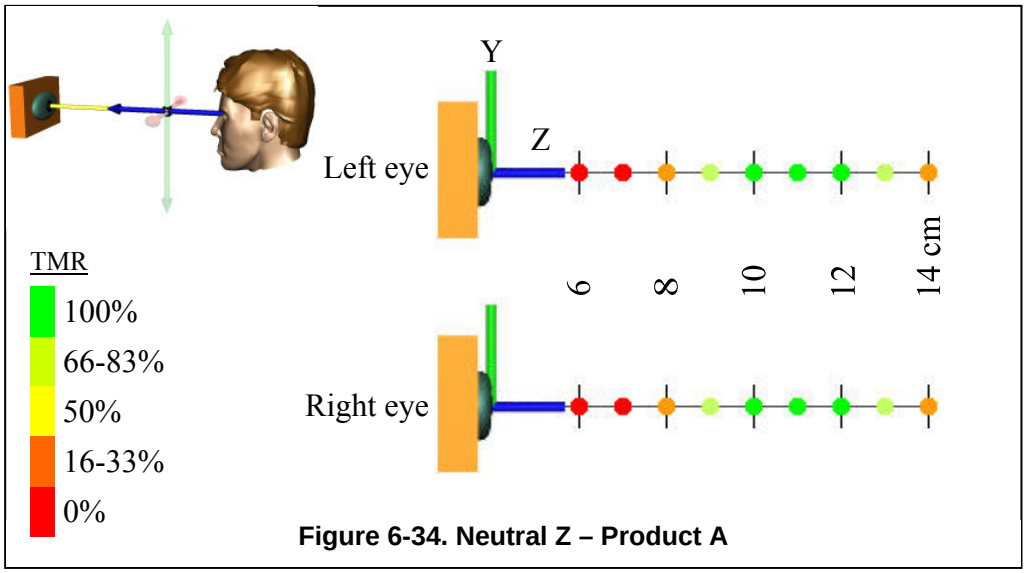


Figure 6-34. Neutral Z – Product A

Neutral Roll results for Product A are illustrated in Figure 6-35. Recall that Neutral Roll emulates the user located directly in front of and looking at the camera but tilting the head to the side "ear to a shoulder". Both eyes performed Perfect or Good with ±35° of head tilt at the ideal distance from the camera Z=10 cm.

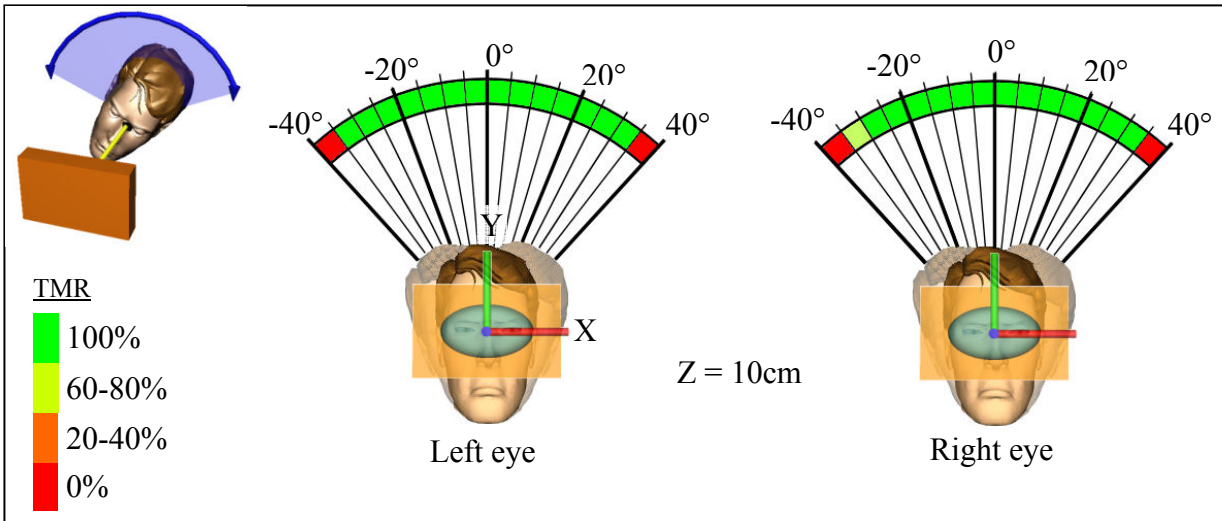


Figure 6-35. Neutral Roll – Product A

Sweep Pitch and Sweep Yaw for Product A are presented in Figures 6-36 and 6-37, respectively. Recall that Sweep Pitch emulates a user positioned directly in front of the camera but looking up or down away from the camera, and Sweep Yaw emulates a user positioned directly in front of the camera but looking to the right or left of the camera.

For Sweep Pitch, both eyes performed Perfect or Good from a downward facing pose of about 15° to an upward facing pose of about 30° at the ideal Z distance of 10 cm. For Sweep Yaw, the left eye performed Perfect or Good from a rightward facing pose of about 30° to a leftward facing pose of about 35°. The right eye performed Perfect or Good from a rightward facing pose of about 15° to a leftward facing pose of about 30°. We note that the left and right eye yaw results are not symmetrical. This may be an artifact of the small test population size.

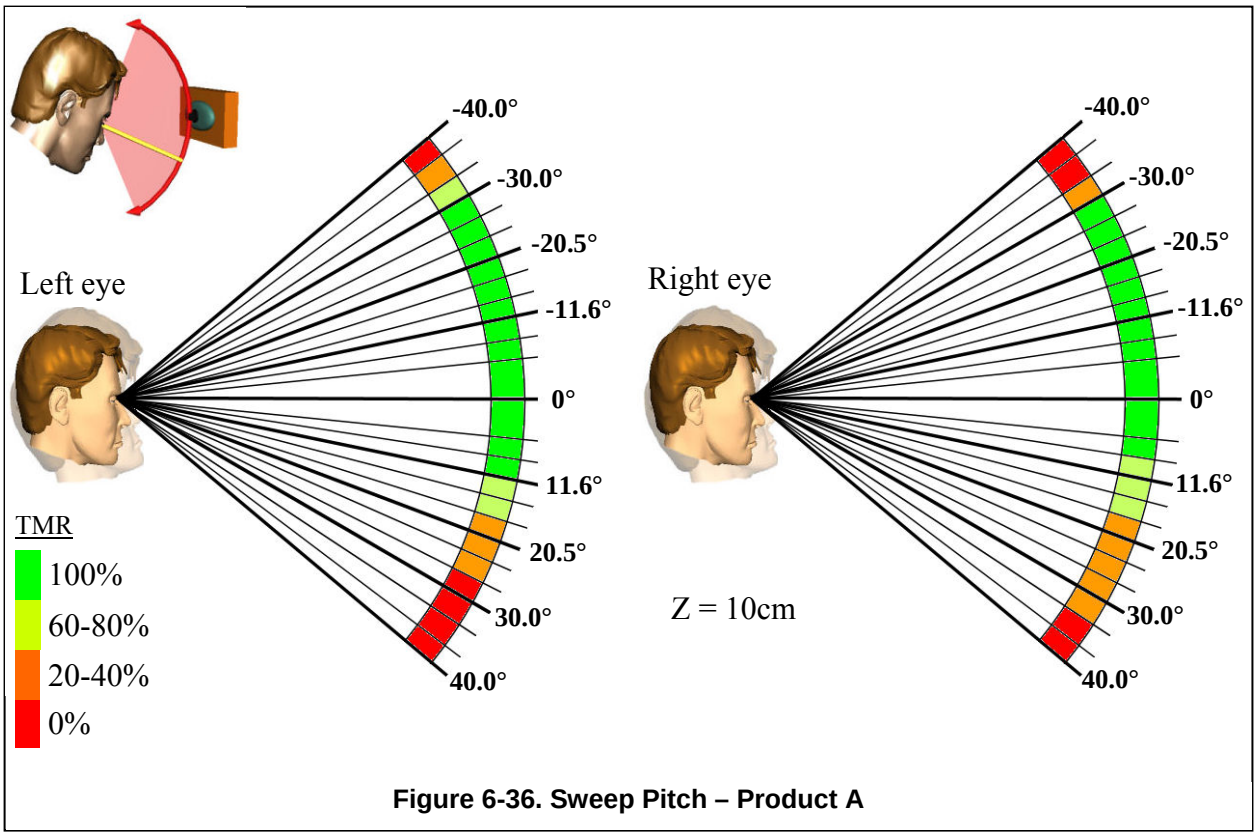


Figure 6-36. Sweep Pitch – Product A

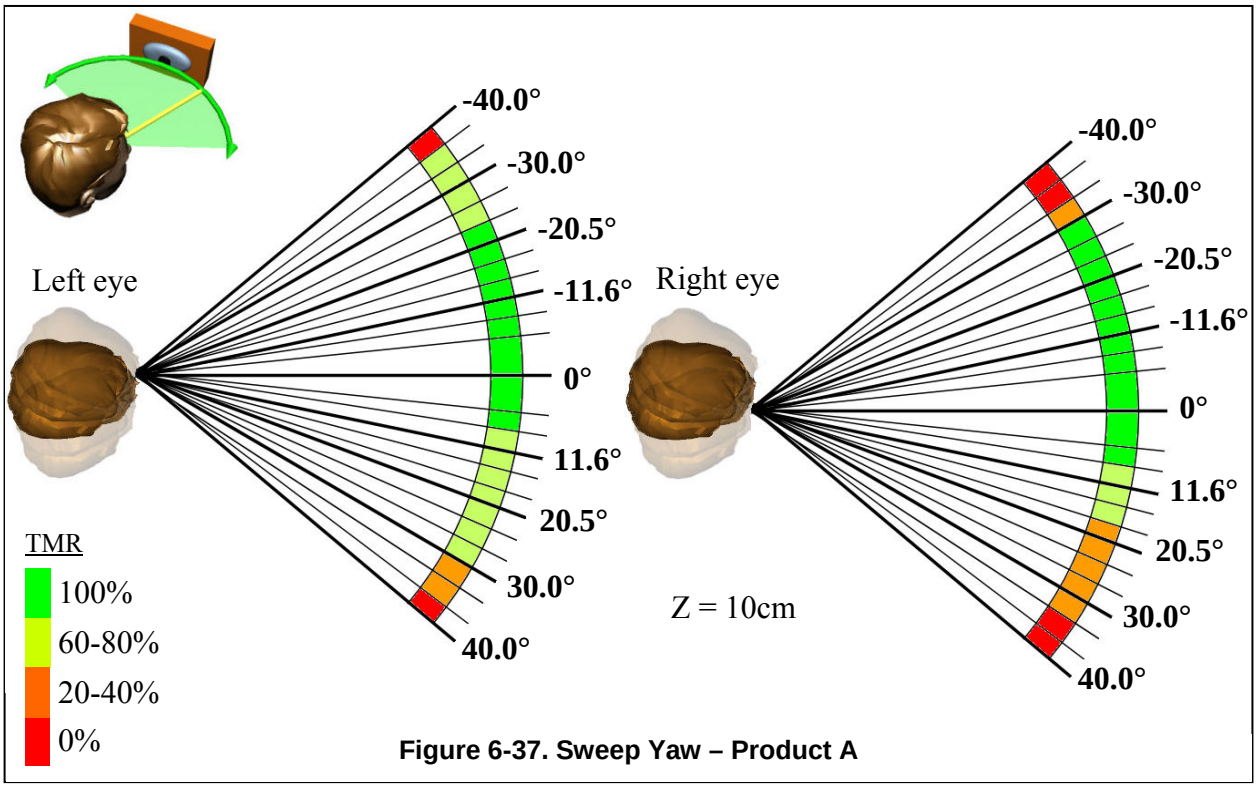


Figure 6-37. Sweep Yaw – Product A

Translate X and Y results are presented in Figure 6-38 at Z=10 cm. Recall that Translate procedures emulate a user positioned to the left, right, above, or below the camera and looking directly forward instead of toward the camera. Both eyes performed Perfect or Good at a distance of 1 cm away from the centerline of the camera system in any direction but did not perform well outside of that distance.

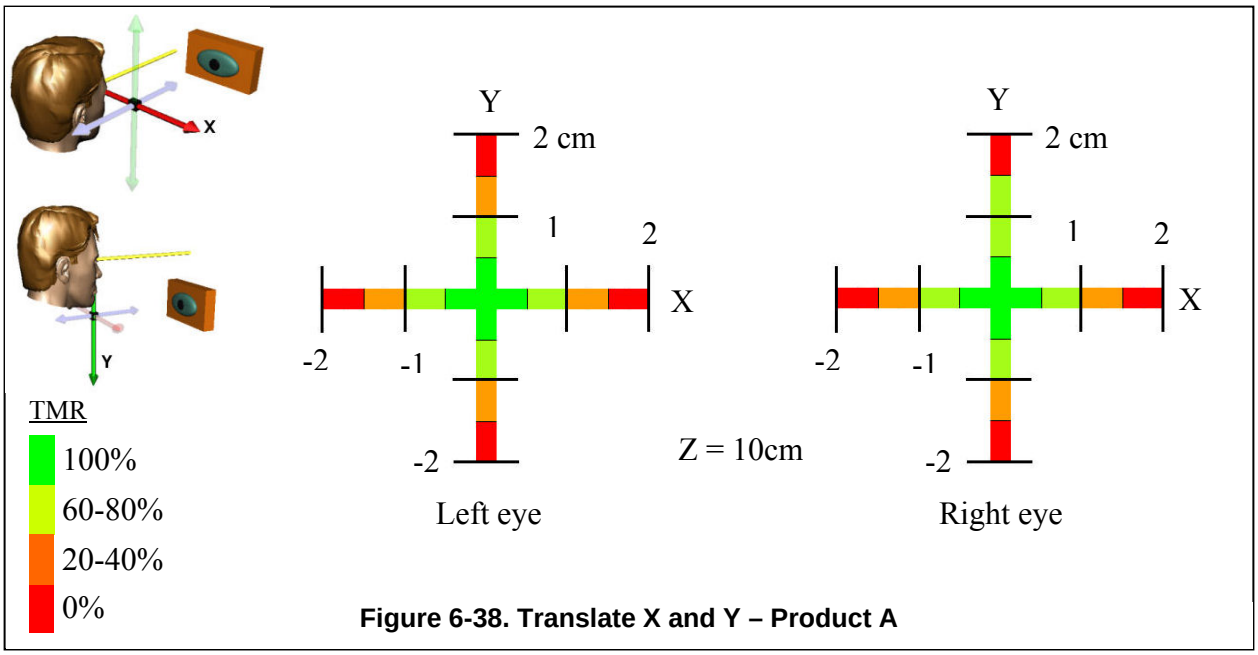


Figure 6-38. Translate X and Y – Product A

Product B

Results of the Neutral Pitch procedure for Product B are illustrated in Figure 6-39. Recall that Neutral Pitch emulates a user positioned above or below the camera at various distances while looking directly at the camera. Product B performed Perfect at 0° from Z distances of 20 to 45 cm and Good out to 50 cm. As the pose angle increased such that the test subjects were above the camera looking down, the range at which the product worked Perfect or Good shifted to closer distances. When the test subjects were below the camera looking up at it, the product performed Perfect or Good at slightly larger distances than when above the camera. The performance envelope is illustrated in Figure 6-39.

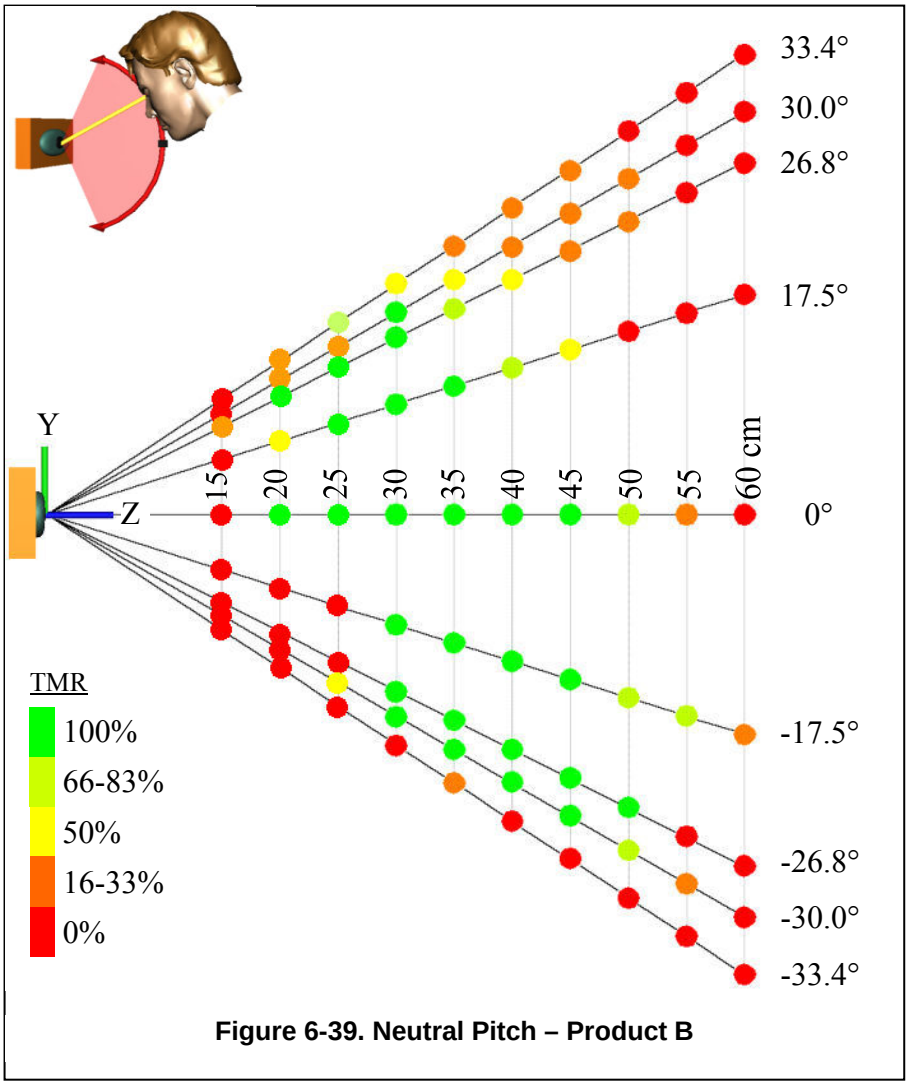


Figure 6-39. Neutral Pitch – Product B

For Neutral Roll (Figure 6-40), Product B performed Perfect or Good for head tilt angles up to about 20° at Z distances of 30 cm and 45 cm. At Z=55 cm, all head tilt angles performed poorly.

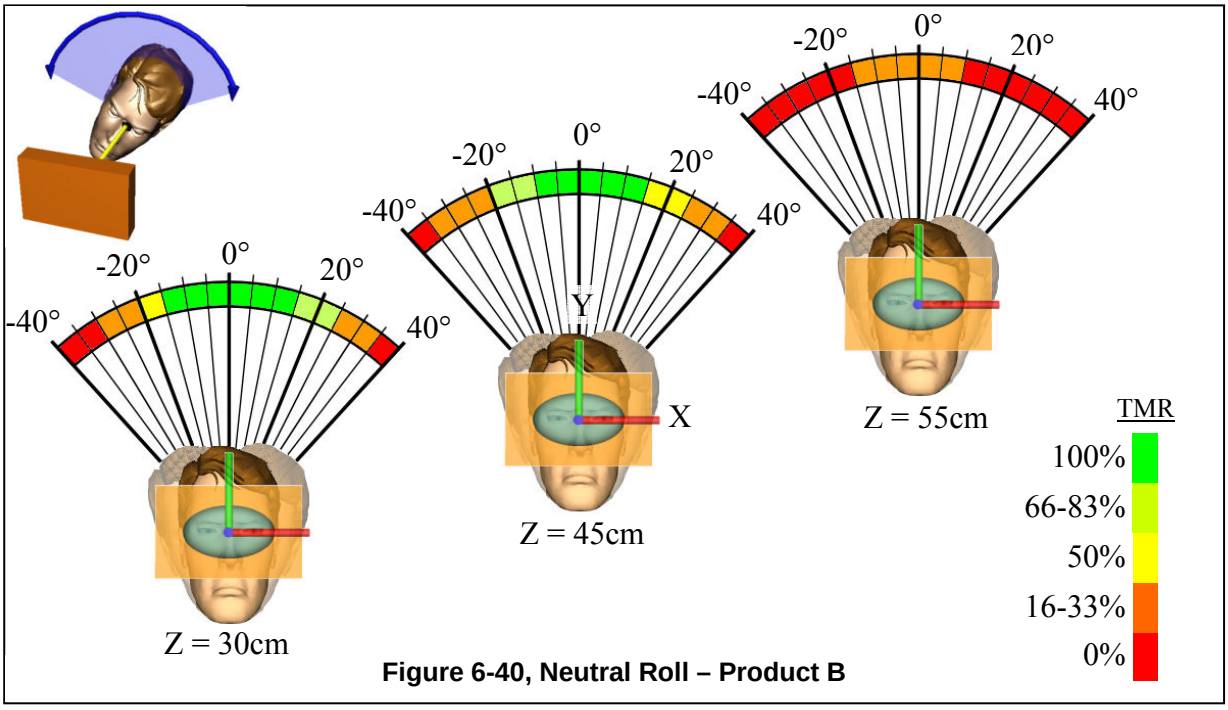


Figure 6-40, Neutral Roll – Product B

Figure 6-41 illustrates Product B's performance for Sweep Pitch. At Z=30 cm, the product performed Perfect or Good from a downward facing pose of about 20° to an upward facing pose of about 30°. At the ideal distance Z=45 cm, the product performed Perfect or Good from a downward facing pose of about 9° to an upward facing pose of about 37°. At Z=55 cm, Product B performed Perfect or Good from a downward facing pose of about 6° to an upward facing pose of about 30°. Product B performed significantly better when looking up at the camera, especially at greater distances away from the camera.

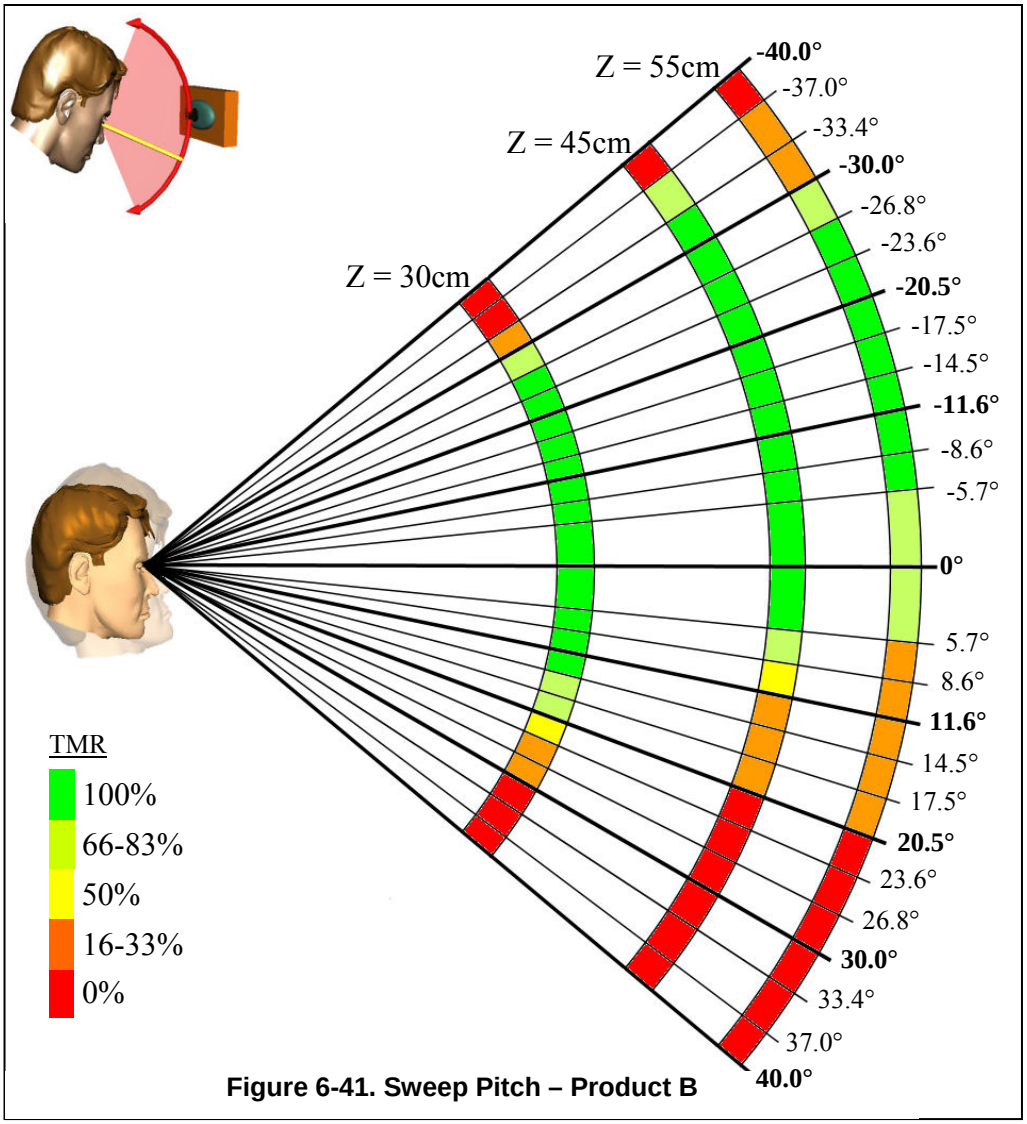


Figure 6-41. Sweep Pitch – Product B

For Sweep Yaw, Figure 6-42 shows that Product B performed Perfect or Good for head yaw angles of about ±30° at Z=30 cm and about ±20° at Z=45 cm. No head yaw angles performed well at Z=55 cm.

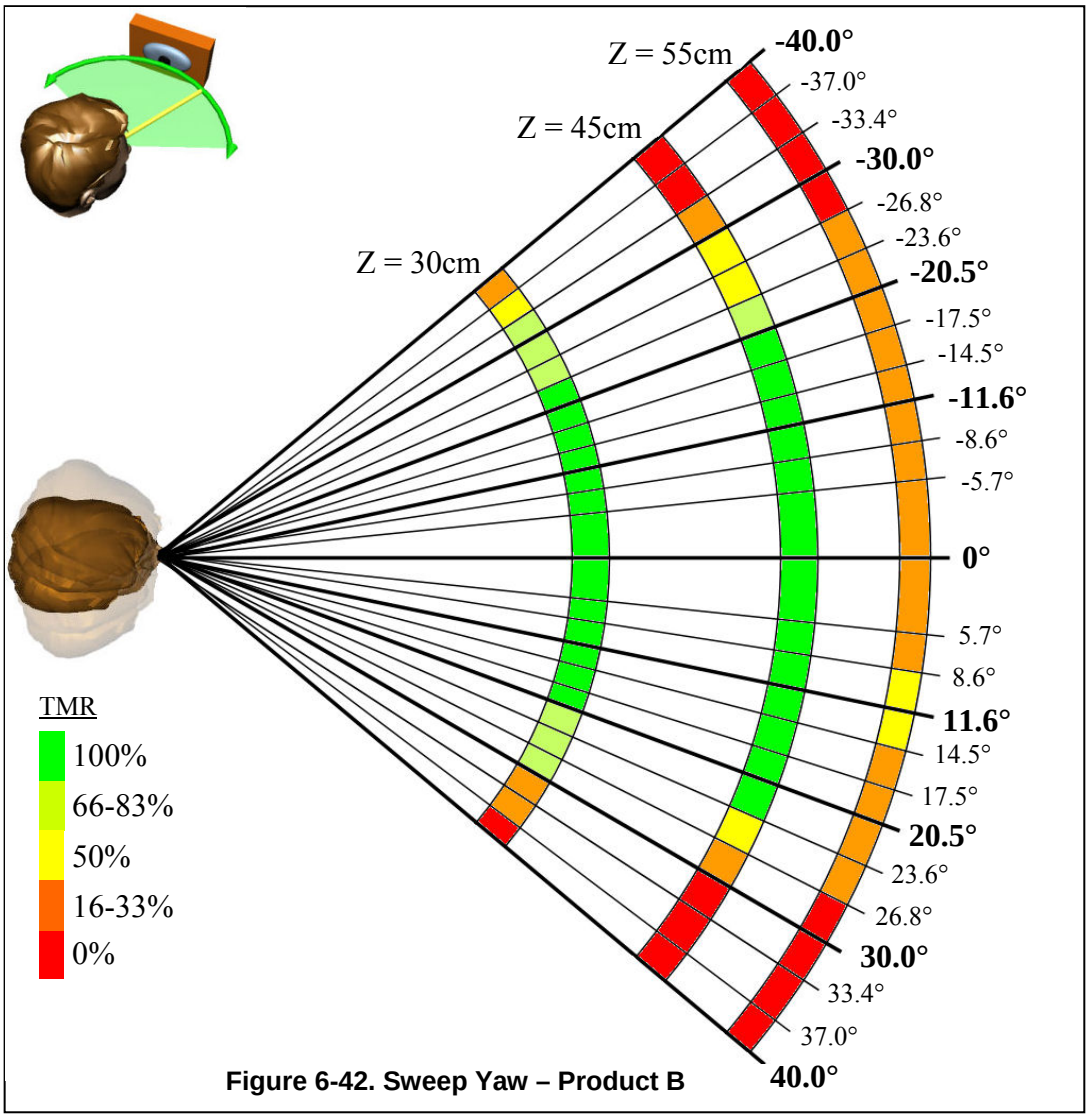


Figure 6-42. Sweep Yaw – Product B

Product C

Figure 6-43 illustrates Neutral Z performance for Product C. Both eyes performed Perfect or Good for Z=12 cm to Z=26 cm. The performance for the left and right eyes are slightly offset, which may be a result of the small test population size for this experiment.

For Neutral Roll, Figure 6-44 indicates that both eyes performed Perfect or Good with ±35° of head tilt at the ideal distance Z=20 cm.

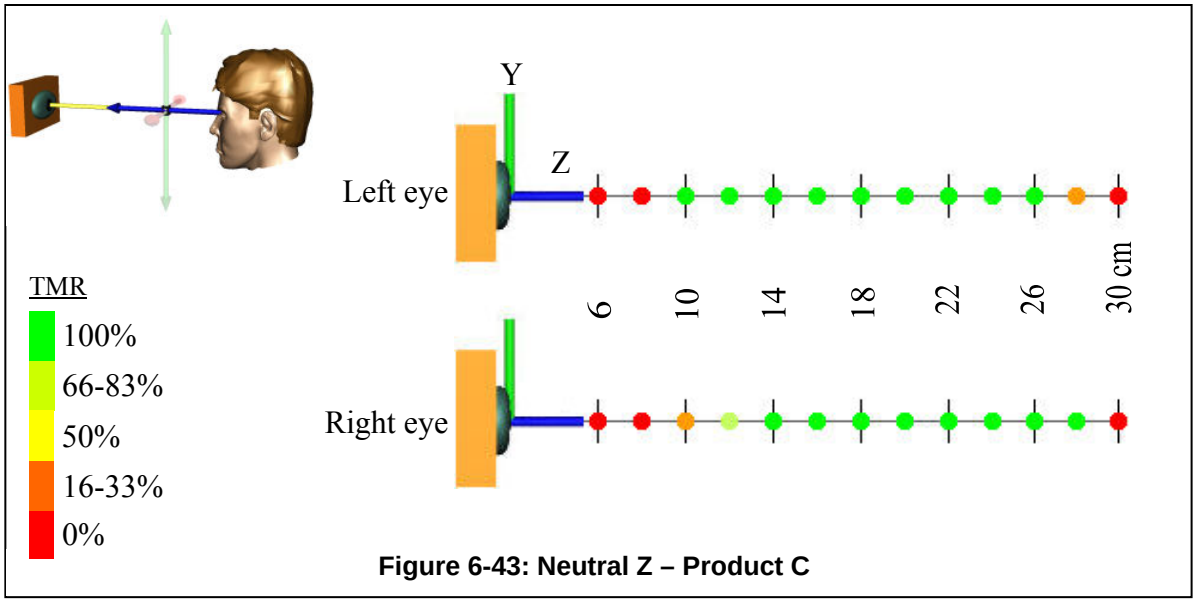


Figure 6-43: Neutral Z – Product C

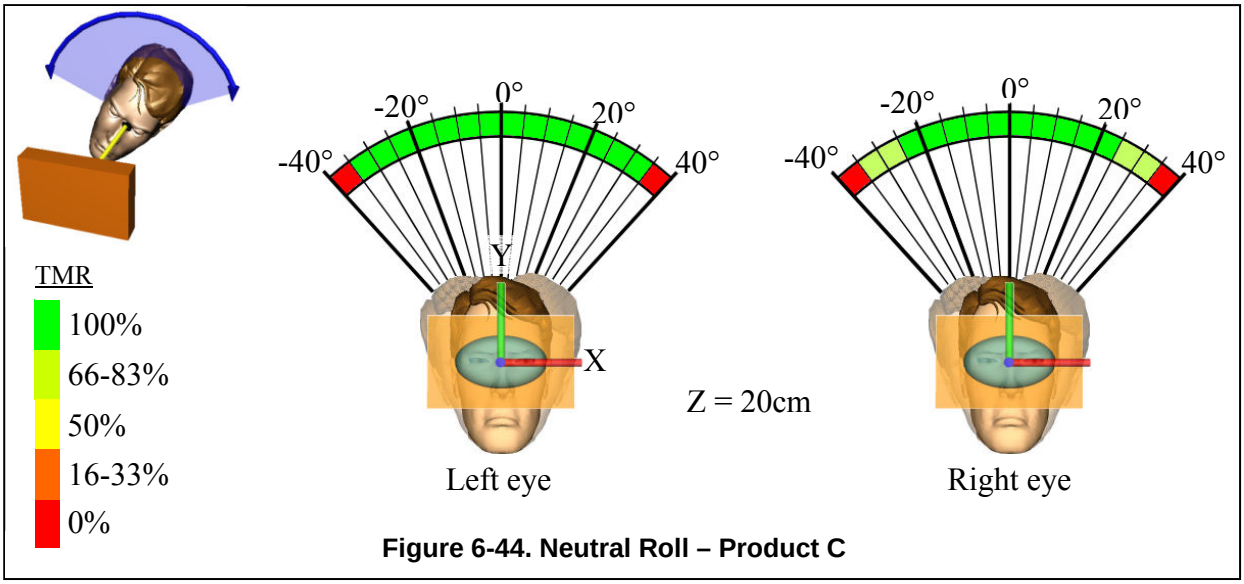


Figure 6-44. Neutral Roll – Product C

Sweep Pitch and Sweep Yaw for Product C are presented in Figures 6-45 and 6-46, respectively. For Sweep Pitch, both eyes performed Perfect or Good from a downward facing

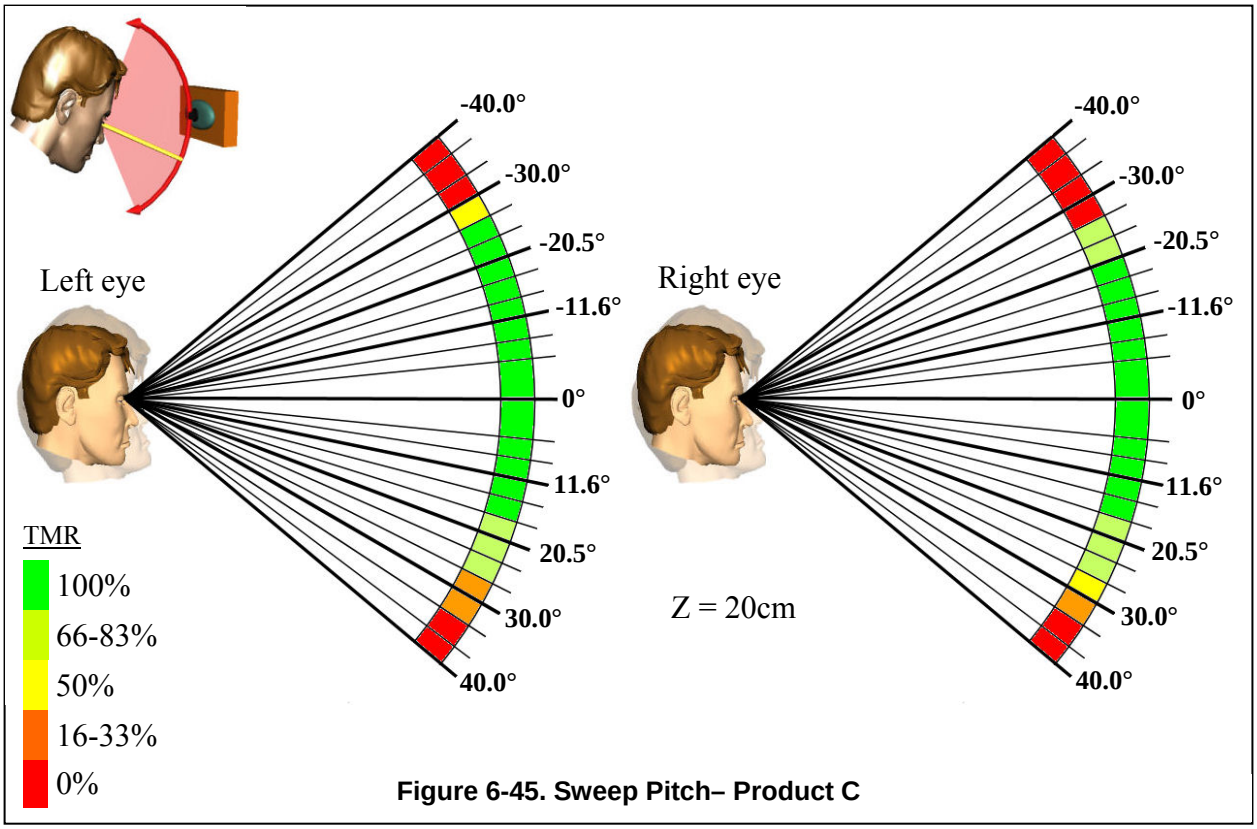


Figure 6-45. Sweep Pitch– Product C

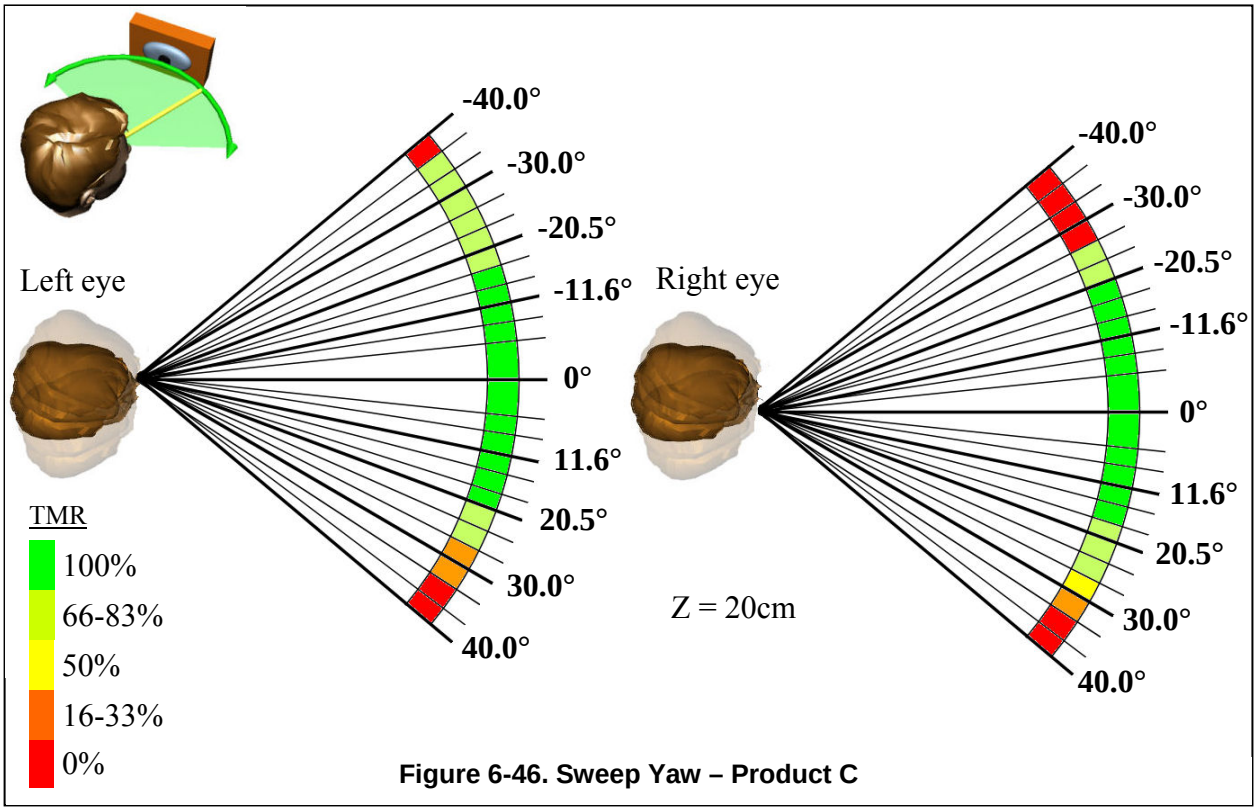


Figure 6-46. Sweep Yaw – Product C

pose of about 25° to an upward facing pose of about 25° at Z=20 cm producing the most symmetric performance of the three products for this procedure.

For Sweep Yaw at Z=20 cm (Figure 6-46), the left eye performed Perfect or Good from a rightward facing pose of about 25° to a leftward facing pose of about 35°; and the right eye performed Perfect or Good from a rightward facing pose of about 25° to a leftward facing pose of about 25°.

Translate X and Y results for Product C are presented in Figure 6-47 for Z=20 cm. Both eyes performed Perfect or Good up to X=±1 cm and Y=+1.5 cm. The left eye also performed Perfect or Good up to Y= -0.5, and the right eye performed Perfect or Good up to Y= -1.0 cm.

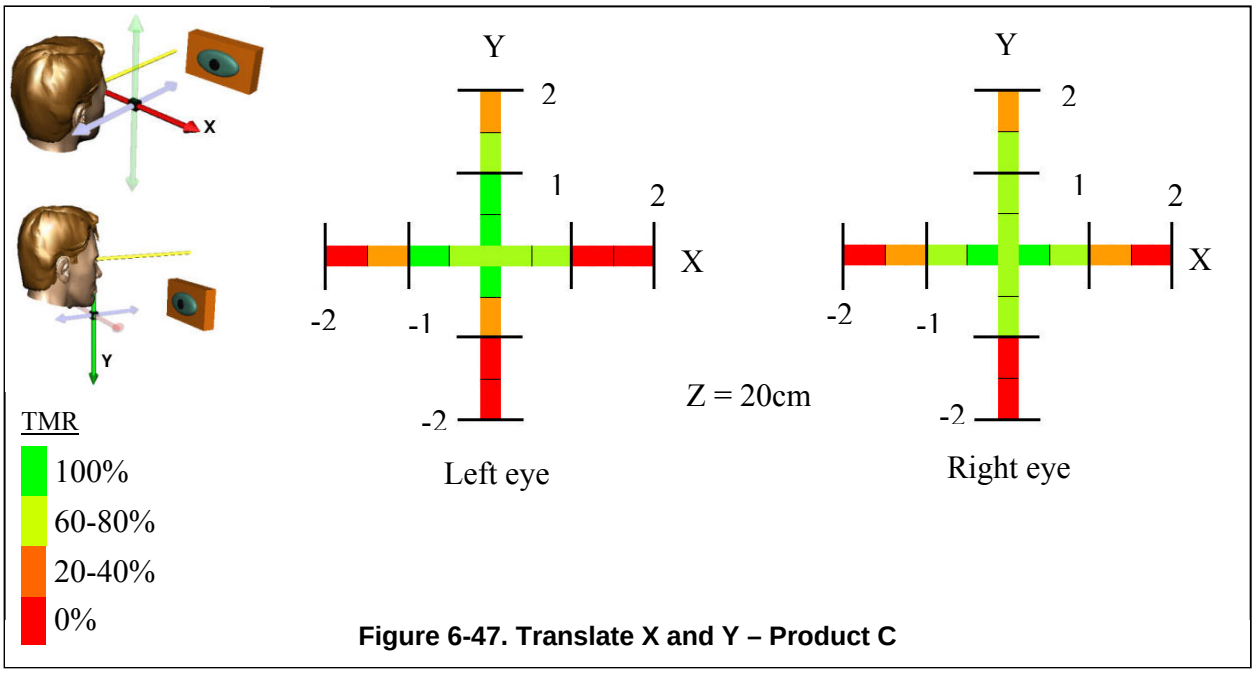


Figure 6-47. Translate X and Y – Product C

Summary of key off-axis pose findings

- The evaluated products generally performed well with yaw (rotate head as if saying "No") and roll (ear-to-shoulder) angles of ±20° or more when the test subject was located at manufacturer-designated distances from the camera.
- In general, the products performed better when the test subject looks upwards relative to the camera rather than downwards. This agrees with the results of the off-axis gaze experiments.
- Product B demonstrates a significantly larger collection volume than Products A and C. As such, Product B is most appropriate for use with uncooperative users.

7. Conclusions

The IRIS06 effort studied the real-time (online) performance of three commercially-available iris recognition products with about 300 live test subjects. Multiple sets of iris images were collected from each test subject with each product spanning time intervals from fifteen minutes to about six weeks. The images collected online were further evaluated offline using template generation and matching algorithms similar to those used by the commercial products. Off-axis gaze and pose experiments were also performed.

We draw the following conclusions from the online, offline, and off-axis results:

- Cumulative FTE, FTA, FNMR, and GFRR rates generally decrease (TMR and GTAR rates generally increase) with increasing numbers of attempts, possibly due to improved effectiveness of the camera-human interface with increasing human practice and the removal of eyeglasses. In general, enrollment and recognition performance improve when multiple attempts are allowed.
- Within the statistical confidence limitations of this evaluation, all products exhibit roughly the same three-attempt FTE rate for the left-or-right eye feature set. In this case, an FTE is declared when three attempts are allowed for each eye and neither eye successfully enrolls, which is a realistic operational enrollment policy. For the three products evaluated, this FTE rate varied from 0.35% to 3.39% with mean enrollment transaction times varying from 32.3 to 70.1 sec as indicated in Table 7-1.

Table 7-1. Three-Attempt Enrollment Metrics for Left-or-Right Eye Feature Set

	Product A	Product B	Product C
FTE (%)	0.35	0.68	3.39
Mean Enrollment Transaction Time (sec)	40.4	32.2	70.1
See Section 6 for confidence intervals			

- Left and right eyes generally exhibit roughly the same FTE, FTA, FNMR/TMR, GFRR/GTAR, and mean recognition transaction times (RTT) for each of the products evaluated. Overall, right and left eyes exhibit statistically similar iris recognition performance.
- Time separation between enrollment and recognition (up to six weeks) does not have a measurable influence on FTA, FNMR, and GFRR and mean RTT.
- Product C appears to be influenced by ambient lighting conditions.

- FNMR/TMR is similar for all products and all iris-feature sets (during real-time online operation and in offline comparisons using Professor Daugman's algorithms). Results indicate the following general trends for three-attempt recognition metrics:
  - $FTA_A < FTA_B \sim FTA_C$ (online),
  - $GFRR_A < GFRR_B \sim GFRR_C$ (online),
  - $GTAR_A > GTAR_B \sim GTAR_C$ (offline), and
  - mean $RTT_B < RTT_C < RTT_A$ (online).

  These trends are illustrated in Table 7-2 for the left-or-right-eye feature set.

Table 7-2. Overall Three-Attempt Recognition Metrics for Left-or-Right Eye Feature Set

	Product A	Product B	Product C
FTA (online) (%)	1.5	6.9	6.9
FNMR=1-TMR (online) (%)	0.0	1.8	0.4
TMR (offline@HD=0.32) (%)	99.7	97.3	99.4
GFRR=1-GTAR (online) (%)	1.9	9.3	10.7
GTAR (offline@HD=0.32) (%)	97.8	89.9	89.0
Mean RTT (online) (sec)	21.4	7.9	11.2
See Section 6 for confidence intervals			

- The products tested demonstrate tradeoffs between speed and accuracy. Higher accuracy requires longer transaction times, and faster transaction times result in lower accuracy. For example, Product A exhibits the best recognition performance but also has the longest average transaction time. The "best" product depends on the specific needs of a particular operational scenario.
- Unless the exact instantiation of an iris recognition product's algorithms are used for offline testing, offline performance results do not necessarily indicate the online (real-world) performance of the product.
- In some cases, enrollment and recognition images from different cameras provide better matching performance than enrollment and recognition images from the same camera. That is, interoperability matching performance is better than native matching performance. Specifically, the best matching performance is obtained using enrollment images from Product C and recognition samples from Product A.

- Eyeglasses can degrade iris recognition performance. For Products B and C, matching performance is degraded for iris images acquired from test subjects wearing eyeglasses. Product A matching performance is similar both with and without glasses.
- Product A collects the highest percentage of high quality iris images compared to Products B and C but also exhibits the longest mean transaction time.
- Product B demonstrates a significantly larger collection volume than Products A and C and exhibits the shortest mean three-attempt recognition transaction time (7.9 seconds). As such, Product B is most appropriate for use with uncooperative users.
- The three iris recognition products evaluated perform better when test subjects gaze upward (with neutral pose) or face upward (with neutral gaze) relative to the camera rather than downward.
- The evaluated products generally performed well with yaw (rotate head as if saying "No") and roll (ear-to-shoulder) angles of ±20° or more when the test subjects were located at manufacturer-designated distances from the camera.

8. Future Efforts

The IRIS06 results indicate several areas deserving of further study. The IRIS06 data set is very rich and worthy of additional analyses. For example, it would be highly useful to study factors contributing to FTE, FTA, and FNMR, such as test subject demographics and conditions, and image and camera properties. Understanding these factors will enable technology providers and implementers to improve operational performance of iris recognition. In addition, a thorough evaluation of the genuine iris image pairs that failed to match will allow us to determine the root causes of the different behavior between online and offline matching performance. Further research into the nature of the differences between the images from the different cameras is also needed to understand the observed improvement in interoperability performance compared to native performance. Such studies of recognition performance as a function of 1) user population demographics and characteristics, such as age, race, gender, eye color, eye conditions, and general health; 2) iris image quality metrics, such as pixel resolution, focus (blur), iris radius, pupil radius, pupil-iris ratio, iris-sclera contrast, iris intensity, texture energy, and visible iris; and 3) camera system properties, such as optical resolution, illumination wavelength, camera illumination geometry, and optical aberrations; will enable optimal usage of iris recognition technology with various user populations and in various operational environments.

Anecdotal review of the IRIS06 data suggests that recognition performance may be influenced by character traits of test team members. For example, fewer failures occurred when a young assertive male was serving as test administrator, while more failures occurred when a soft-spoken, nonassertive female was serving as test administrator. A comprehensive study of performance as a function of personnel behavior could help biometric system implementers select the most effective biometric system operators for various field applications.

Another interesting topic for future study is intra-individual correlation, or correlation between iris image comparisons for individuals. If intra-individual correlation is low, sufficient information can be gleaned by collecting multiple iris images from a limited test population to estimate operational recognition performance. However, if intra-individual correlation is high, it is necessary to collected iris images from a larger test population to accurately predict operational performance. This information will be invaluable for when planning future iris recognition evaluations.

Along these lines, an analysis of frequent false matchers (wolves), frequent false matchees (lambs), and frequent false non-matchers (goats) for the IRIS06 data set would also significantly bolster the knowledge base of the biometrics community. For example, if most false non-matches are due to only a few individuals, these individuals could be treated differently in fielded applications to minimize false non-match operational disruptions. In IRIS06, one iris image from a test subject strongly matched a non-self iris image. However, other images from the same individual's iris did not strongly match the same non-self iris image (or other iris images from the same non-self eye). This and other anecdotal evidence suggests that the Doddington's Zoo phenomenon may be image-specific as opposed to individual specific. This highlights the importance of studying image quality metrics such that potential problem images can be identified at the source.

It would also be of value to study the influence of various approaches to generating impostor score distributions. The shape of the impostor distribution, and thus the resulting ROC and DET performance curves, can change slightly depending on how impostor scores are computed. For example, for this evaluation impostor score distributions were generated by cross-comparing all non-self combinations of the enrolled test subject population, comparing right eyes only to right eyes, and comparing left eyes only to left eyes. How might the impostor score distribution differ: If scores resulting from non-self left eyes compared to non-self right eyes were included? If match scores from right and left eyes of the same person were included? If only one match score per test subject eye was included? If only match scores from true impostors

(individuals not used for genuine comparisons) were included? In addition, while FMR and GFAR confidence intervals were computed during the IRIS06 data analysis, presentation and discussion of these results was considered beyond the scope of IRIS06. Understanding the influence of various impostor-score-generation methodologies on the impostor score distribution and on the resulting FMR and GFAR values and confidence intervals will allow biometric system evaluators to better predict the performance of real-world biometric systems.

The IRIS06 offline evaluation was performed in verification mode for ease of comparison with other biometric evaluations. However, iris recognition has a reputation for performing well in identification mode, which is a more difficult task as an individual must be correctly selected from the entire database of enrolled samples, as opposed to simply being matched to the individual's enrolled sample. It would be of significant value to calculate offline identification performance of the IRIS06 dataset. Effective identification performance may be one of the primary advantages of iris recognition technology relative to other biometric modalities.

In this vein, comparing operationally-relevant performance metrics for iris recognition with those for fingerprint and facial recognition and identifying commercial biometric products for each of the modalities that meet requirements for specific real-world applications would also be useful to the biometrics community at large.

9. Implications for Knowledge and Practice

The ultimate goal of the IRIS06 effort was to predict the performance of iris recognition products in various criminal justice, law enforcement, and border control biometric applications, such as physical and logical access control, surveillance, and identification systems. IRIS06 results indicate that iris recognition is a viable biometric modality. Multiple products are available that meet various operational needs. For example, Product B is ideal for high throughput applications and for usage with uncooperative users. Product A is ideal when higher accuracy is more critical than faster transaction times. While IRIS06 did not specifically evaluate identification performance, it is likely to be very similar to the reported verification performance thus indicating very low false match error rates. As such, iris recognition may be an optimal solution for large-scale identification applications.

The ISO-compliant iris images collected from the products evaluated exhibit reasonable interoperability performance. Native and interoperability matching performance can be further improved if image comparison algorithms know the conditions under which the iris images are collected. The iris recognition community needs to agree on an approach for obtaining optimal

interoperable performance between images from different cameras and with different recognition algorithms. For example, iris image information (such as the image-collection camera system) can be shared in the standardized iris data interchange format so that appropriate algorithm parameters can be adjusted for different types of iris images. Alternatively, tighter standards can be levied on iris images and iris image collection parameters such that all algorithms can be optimized *a priori* for standardized iris images.

While iris recognition technology will continue to evolve and improve, based on the results of IRIS06 Authenti-Corp advises policymakers, practitioners, and public officials that iris recognition technology can currently be used to recognize cooperative and uncooperative individuals rapidly, reliably, and interchangeably in criminal justice and border control applications thus improving public safety, security, and quality of life for all citizens. The "best" biometric product – be it based on iris, fingerprint, or face – depends on the specific needs of a particular operational scenario.

10. Acknowledgements

Funding and oversight for this study were provided by the US Department of Justice, National Institute of Justice (NIJ) and the US Department of Homeland Security, Transportation Security Administration (DHS/TSA) under Award No. 2005-IJ-CX-K066. The opinions, findings, conclusions, and recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice or the Department of Homeland Security. Authenti-Corp is grateful for the support and guidance provided by the IRIS06 program managers, Christopher Miles (NIJ) and Valerie M. Lively (DHS/TSA).

Authenti-Corp also gratefully acknowledges BioAPI support from the BioFoundry Division of OSS Nokalva,[24] the provision of offline template generation and matching algorithms from Professor John Daugman, University of Cambridge,[12] and confidence interval technical advice from Dr. Michael Schuckers.[37] We also express sincere gratitude to our colleagues who critically reviewed this report and provided suggestions for improvement.

The primary contributors to the IRIS06 effort were Dr. Valorie S. Valencia, Roger Cottam, and Leanne J. Walker from Authenti-Corp, and Stephen A. Borota from the College of Optical Sciences, University of Arizona.

11. Appendices

11.1 Iriscode Generation and Comparison Technical Details

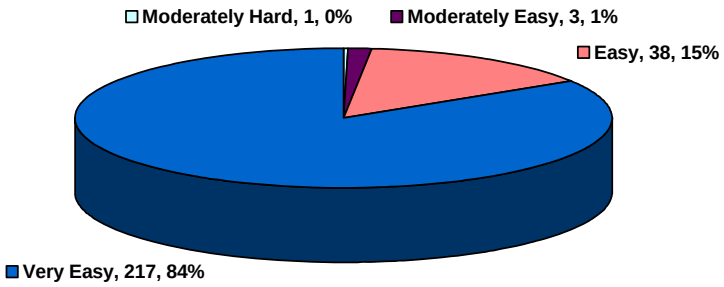
The following Iriscode discussion is from the paper "Iris on the Move: Acquisition of Images for Iris Recognition in Less Constrained Environments," J. R. Matey, *et. al.* Proceedings of the IEEE, **94**(11), 1937 (2006).

> "The iris code is generated by performing a dot product between complex Gabor wavelets and an $N X M$ grid of locations on the normalized image. The phase angles of resulting complex dot products are then quantized to 2 bits and assembled into the iris code as an $N X M$ array of 2-bit cells.
>
> "The comparison step computes a fractional Hamming distance between the bit array of one template and that of another and compares that distance to a predetermined threshold. The fractional Hamming distance (hereafter referred to as the Hamming distance, in accordance with popular usage) is the fraction of the bits that differ between the two templates.
>
> Using a statistical argument to extrapolate from a limited dataset, Daugman predicted that the probability of obtaining a Hamming distance less than 0.33 for templates that arise from different irises is about one part in 4 million and drops to less than one part in a billion at 0.30. At the time, the available datasets were not good enough to test the prediction. In a later paper, he presented extensive real world tests of the argument and found that the probability at 0.297 is about one part in 18 million.
>
> "The distance reported by the Iridian implementation of the Daugman algorithm is a modified (sometimes called normalized) fractional Hamming distance that is adjusted to provide a constant probability of a false match, based upon the fraction of valid bits in the compared templates.
>
> "In many commercial implementations of the Daugman algorithm, the templates are transformed using random XOR and permutation matrices to generate iris templates that are specific to the particular instantiation of the implementation in which the algorithm is used."

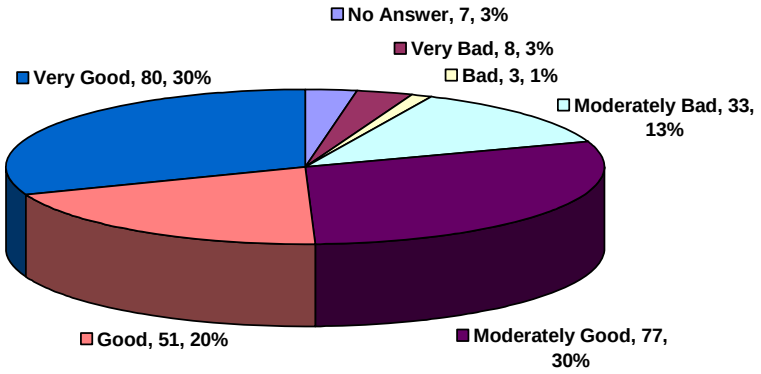
11.2 Test Subject Exit Questionnaire Results

Of the 264 test subjects who completed both visits, 259 filled out the optional exit questionnaire. Out of the 259, 8% had used biometric devices before. Test subject responses to the questions are summarized below.

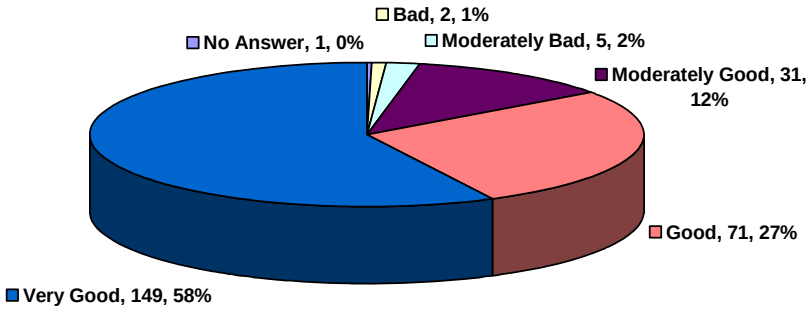
How hard/easy was the overall process?



What was your opinion of biometrics *before* today?

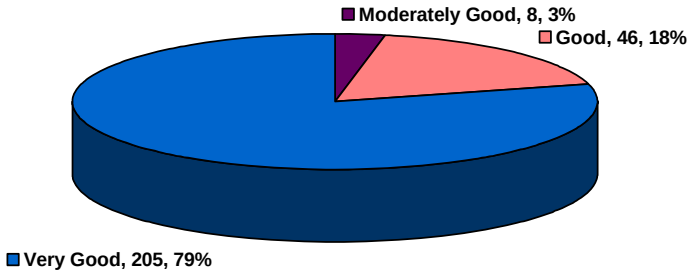


What is your opinion of biometrics now?

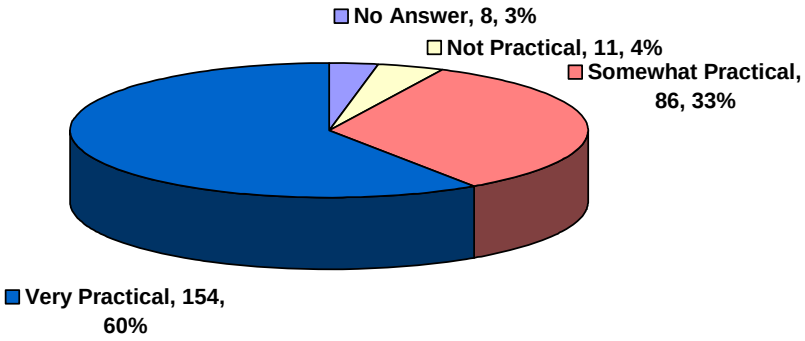


Participation in this study positively shifted test subjects' opinions of biometrics. Of the test subjects who responded to the questionnaire, 80% had a positive opinion of biometrics before the study, and 97% had a positive opinion of biometrics following the study.

How would you rate your overall experience today?



How practical would it be for you to use biometric security devices on a daily basis?



When asked what they liked about the study, test subjects most liked the ease of the study, followed by the friendliness of the test team (Authenti-Corp employees) and the short time the test took. The most common test subject concerns were data security and confidentiality and the possibility of eye damage; however 85% of the respondents had no concerns. When asked on the Informed Consent form if we could contact them at a later date to collect additional biometric data, 95% responded positively, 4.7% did not answer and 0.3% (1 person) responded negatively.

11.3 Online Confidence Interval Tables

Table 11-1. Cumulative Failure to Enroll (FTE)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
	# of Persons	288			295			295		
Visit 1	UCI	3.20%	2.70%	2.18%	11.50%	8.72%	7.08%	26.77%	26.05%	20.21%
	Attempt 1	**1.04%**	**0.69%**	**0.35%**	**7.80%**	**5.42%**	**4.07%**	**21.69%**	**21.02%**	**15.59%**
	LCI	0.22%	0.04%	0.00%	5.22%	3.32%	2.29%	17.37%	16.76%	11.89%
	UCI	3.20%	2.70%	2.18%	9.12%	5.81%	3.60%	17.99%	17.61%	9.92%
	Attempt 2	**1.04%**	**0.69%**	**0.35%**	**5.76%**	**3.05%**	**1.36%**	**13.56%**	**13.22%**	**6.44%**
	LCI	0.22%	0.04%	0.00%	3.59%	1.55%	0.42%	10.11%	9.81%	4.13%
	UCI	3.20%	2.70%	2.18%	9.12%	4.95%	2.64%	16.11%	14.59%	6.24%
	Attempt 3	**1.04%**	**0.69%**	**0.35%**	**5.76%**	**2.37%**	**0.68%**	**11.86%**	**10.51%**	**3.39%**
	LCI	0.22%	0.04%	0.00%	3.59%	1.07%	0.04%	8.64%	7.49%	1.79%
	UCI			42.22			35.36			74.74
	TT			**40.39**			**32.23**			**70.07**
	LCI			38.56			29.09			65.40

Table 11-2. Cumulative Failure to Acquire (FTA)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	# of Persons	278	279	280	277	287	292	255	260	280
	UCI	27.58%	22.50%	13.73%	22.66%	24.15%	19.67%	18.98%	18.63%	17.34%
	Attempt 1	**22.30%**	**17.56%**	**9.64%**	**17.69%**	**19.16%**	**15.07%**	**14.12%**	**13.85%**	**12.86%**
	LCI	17.81%	13.54%	6.69%	13.64%	15.03%	11.41%	10.36%	10.16%	9.42%
	UCI	15.45%	13.78%	7.88%	14.70%	13.80%	11.61%	16.36%	16.06%	16.94%
	Attempt 2	**11.15%**	**9.68%**	**4.64%**	**10.47%**	**9.76%**	**7.88%**	**11.76%**	**11.54%**	**12.50%**
	LCI	7.95%	6.71%	2.68%	7.37%	6.82%	5.28%	8.35%	8.18%	9.11%
	UCI	5.24%	5.68%	2.78%	10.97%	10.19%	9.21%	16.36%	16.06%	16.94%
	Attempt 3	**2.52%**	**2.87%**	**0.71%**	**7.22%**	**6.62%**	**5.82%**	**11.76%**	**11.54%**	**12.50%**
	LCI	1.14%	1.38%	0.04%	4.69%	4.24%	3.63%	8.35%	8.18%	9.11%
	Verify 2									
	# of Persons	280	281	282	255	259	280	278	288	293
	UCI	23.19%	28.81%	16.04%	28.35%	29.97%	22.57%	18.55%	18.27%	18.52%
	Attempt 1	**18.21%**	**23.49%**	**11.70%**	**23.02%**	**24.65%**	**17.75%**	**13.73%**	**13.51%**	**13.93%**
	LCI	14.13%	18.91%	8.44%	18.46%	20.03%	13.80%	10.02%	9.87%	10.35%
	UCI	12.92%	14.89%	8.68%	17.06%	16.87%	14.30%	17.24%	16.98%	17.34%
	Attempt 2	**8.93%**	**10.68%**	**5.32%**	**12.59%**	**12.50%**	**10.24%**	**12.55%**	**12.36%**	**12.86%**
	LCI	6.10%	7.56%	3.20%	9.18%	9.15%	7.25%	9.01%	8.87%	9.42%
	UCI	5.66%	5.19%	4.24%	12.18%	14.93%	11.97%	16.36%	16.98%	17.34%
	Attempt 3	**2.86%**	**2.49%**	**1.77%**	**8.27%**	**10.76%**	**8.19%**	**11.76%**	**12.36%**	**12.86%**
	LCI	1.38%	1.13%	0.66%	5.55%	7.67%	5.54%	8.35%	8.87%	9.42%
Visit 2	**Identify 1**									
	# of Persons	167	167	168	202	208	213	222	225	243
	UCI	33.55%	31.65%	20.47%	26.96%	29.30%	23.08%	9.88%	11.35%	5.98%
	Attempt 1	**26.35%**	**24.55%**	**14.29%**	**20.79%**	**23.08%**	**17.37%**	**5.86%**	**7.11%**	**2.88%**
	LCI	20.25%	18.64%	9.76%	15.76%	17.87%	12.86%	3.39%	4.37%	1.31%
	UCI	18.56%	19.92%	10.78%	14.90%	16.13%	11.41%	4.75%	4.08%	4.90%
	Attempt 2	**12.57%**	**13.77%**	**5.95%**	**9.90%**	**11.06%**	**7.04%**	**1.80%**	**1.33%**	**2.06%**
	LCI	8.34%	9.32%	3.17%	6.46%	7.45%	4.26%	0.56%	0.29%	0.76%
	UCI	9.36%	9.36%	5.42%	10.23%	12.24%	10.29%	2.82%	2.78%	2.58%
	Attempt 3	**4.79%**	**4.79%**	**1.79%**	**5.94%**	**7.69%**	**6.10%**	**0.45%**	**0.44%**	**0.41%**
	LCI	2.33%	2.33%	0.40%	3.36%	4.74%	3.54%	0.00%	0.00%	0.00%
	Identify 2									
	# of Persons	167	167	168	202	208	213	220	222	241
	UCI	34.18%	34.81%	23.11%	28.55%	27.25%	22.05%	7.17%	9.88%	4.38%
	Attempt 1	**26.95%**	**27.54%**	**16.67%**	**22.28%**	**21.15%**	**16.43%**	**3.64%**	**5.86%**	**1.66%**
	LCI	20.79%	21.34%	11.77%	17.09%	16.15%	12.05%	1.76%	3.39%	0.51%
	UCI	21.92%	22.58%	14.34%	18.83%	18.84%	15.23%	2.12%	4.75%	1.94%
	Attempt 2	**15.57%**	**16.17%**	**8.93%**	**13.37%**	**13.46%**	**10.33%**	**0.00%**	**1.80%**	**0.00%**
	LCI	10.83%	11.33%	5.42%	9.33%	9.46%	6.89%	0.00%	0.56%	0.00%
	UCI	7.84%	11.57%	5.42%	13.17%	13.93%	11.96%	2.12%	2.82%	1.94%
	Attempt 3	**3.59%**	**6.59%**	**1.79%**	**8.42%**	**9.13%**	**7.51%**	**0.00%**	**0.45%**	**0.00%**
	LCI	1.51%	3.63%	0.40%	5.27%	5.88%	4.63%	0.00%	0.00%	0.00%

Table 11-3. Cumulative FTA by Visit									
	Product A			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)									
UCI	4.66%	4.50%	2.57%	10.91%	11.93%	10.19%	16.19%	16.41%	17.06%
Attempt 3	**2.69%**	**2.68%**	**1.25%**	**7.75%**	**8.70%**	**7.01%**	**11.76%**	**11.95%**	**12.68%**
LCI	1.54%	1.58%	0.60%	5.45%	6.28%	4.77%	8.43%	8.57%	9.30%
Visit 2 (Identify 1 and Identify 2)									
UCI	7.35%	9.20%	4.37%	11.08%	12.55%	10.62%	1.54%	1.73%	1.41%
Attempt 3	**4.19%**	**5.69%**	**1.79%**	**7.18%**	**8.41%**	**6.81%**	**0.23%**	**0.45%**	**0.21%**
LCI	2.36%	3.47%	0.72%	4.58%	5.56%	4.30%	0.03%	0.11%	0.03%

Table 11-4 Overall Cumulative 3rd-Attempt FTA									
	Product A			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)									
UCI	4.93%	5.45%	2.61%	10.16%	11.32%	9.56%	8.88%	9.05%	9.30%
Attempt 3	**3.25%**	**3.80%**	**1.45%**	**7.51%**	**8.58%**	**6.92%**	**6.41%**	**6.63%**	**6.90%**
LCI	2.13%	2.64%	0.80%	5.50%	6.45%	4.98%	4.59%	4.82%	5.08%

Table 11-5. Cumulative False Non-Match Rate (FNMR)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	UCI	6.71%	7.95%	2.48%			15.45%	15.87%	21.99%	10.97%
	Attempt 1	**3.24%**	**4.35%**	**0.40%**			**10.89%**	**10.96%**	**16.52%**	**6.97%**
	LCI	1.47%	2.30%	0.00%			7.57%	7.45%	12.22%	4.35%
	UCI	2.54%	1.86%	1.76%			4.93%	8.13%	13.64%	6.45%
	Attempt 2	**0.40%**	**0.00%**	**0.00%**			**2.23%**	**4.44%**	**9.13%**	**3.27%**
	LCI	0.00%	0.00%	0.00%			0.93%	2.35%	6.01%	1.58%
	UCI	2.32%	1.73%	1.69%			2.29%	2.08%	3.37%	1.91%
	Attempt 3	**0.37%**	**0.00%**	**0.00%**			**0.36%**	**0.00%**	**0.87%**	**0.00%**
	LCI	0.00%	0.00%	0.00%			0.00%	0.00%	0.05%	0.00%
	Verify 2									
	UCI	6.34%	4.26%	1.88%			18.67%	18.36%	22.48%	11.59%
	Attempt 1	**3.06%**	**1.40%**	**0.00%**			**13.69%**	**13.18%**	**16.96%**	**7.47%**
	LCI	1.39%	0.30%	0.00%			9.90%	9.32%	12.61%	4.73%
	UCI	2.46%	1.87%	1.76%			8.84%	8.75%	8.60%	5.96%
	Attempt 2	**0.39%**	**0.00%**	**0.00%**			**5.32%**	**4.93%**	**4.85%**	**2.87%**
	LCI	0.00%	0.00%	0.00%			3.15%	2.70%	2.66%	1.30%
	UCI	2.31%	1.71%	1.69%			4.93%	2.08%	3.41%	2.57%
	Attempt 3	**0.37%**	**0.00%**	**0.00%**			**2.23%**	**0.00%**	**0.88%**	**0.41%**
	LCI	0.00%	0.00%	0.00%			0.93%	0.00%	0.05%	0.00%
Visit 2	**Identify 1**									
	UCI	10.52%	8.22%	4.30%			24.61%	24.02%	31.19%	14.28%
	Attempt 1	**4.88%**	**3.17%**	**0.69%**			**18.18%**	**18.18%**	**24.88%**	**9.75%**
	LCI	2.07%	1.01%	0.00%			13.17%	13.53%	19.51%	6.55%
	UCI	4.24%	3.21%	2.93%			11.04%	12.78%	16.18%	7.69%
	Attempt 2	**0.68%**	**0.00%**	**0.00%**			**6.57%**	**8.26%**	**11.26%**	**4.20%**
	LCI	0.00%	0.00%	0.00%			3.81%	5.24%	7.72%	2.22%
	UCI	2.92%	2.92%	2.81%			6.59%	5.38%	6.47%	3.80%
	Attempt 3	**0.00%**	**0.00%**	**0.00%**			**3.00%**	**2.26%**	**3.13%**	**1.24%**
	LCI	0.00%	0.00%	0.00%			1.26%	0.84%	1.42%	0.27%
	Identify 2									
	UCI	7.38%	8.55%	3.30%			27.99%	24.21%	28.66%	16.14%
	Attempt 1	**2.46%**	**3.31%**	**0.00%**			**21.35%**	**18.40%**	**22.49%**	**11.39%**
	LCI	0.56%	1.05%	0.00%			15.96%	13.75%	17.35%	7.93%
	UCI	4.39%	3.30%	3.03%			14.51%	12.66%	17.50%	9.12%
	Attempt 2	**0.71%**	**0.00%**	**0.00%**			**9.42%**	**8.18%**	**12.39%**	**5.39%**
	LCI	0.00%	0.00%	0.00%			6.00%	5.19%	8.63%	3.12%
	UCI	3.86%	2.97%	2.81%			5.34%	7.17%	4.77%	1.94%
	Attempt 3	**0.62%**	**0.00%**	**0.00%**			**2.03%**	**3.64%**	**1.81%**	**0.00%**
	LCI	0.00%	0.00%	0.00%			0.63%	1.76%	0.56%	0.00%

Table 11-6. Cumulative 3rd-Attempt FNMR By Visit									
	Product A			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)									
UCI	2.49%	1.06%	1.06%			2.91%	1.30%	2.28%	1.41%
Attempt 3	**0.37%**	**0.00%**	**0.00%**			**1.29%**	**0.00%**	**0.88%**	**0.20%**
LCI	0.05%	0.00%	0.00%			0.56%	0.00%	0.33%	0.03%
Visit 2 (Identify 1 and Identify 2)									
UCI	2.08%	1.80%	1.76%			5.09%	5.51%	4.56%	1.87%
Attempt 3	**0.31%**	**0.00%**	**0.00%**			**2.52%**	**2.95%**	**2.47%**	**0.62%**
LCI	0.05%	0.00%	0.00%			1.23%	1.56%	1.33%	0.20%

Table 11-7. Overall Cumulative 3rd-Attempt FNMR									
	Product A			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)									
UCI	1.45%	1.06%	1.05%			3.20%	2.76%	2.96%	1.24%
Attempt 3	**0.35%**	**0.00%**	**0.00%**			**1.81%**	**1.46%**	**1.66%**	**0.41%**
LCI	0.08%	0.00%	0.00%			1.01%	0.77%	0.93%	0.14%

Table 11-8. Cumulative Generalized False Reject Rate (GFRR)

		Product A			Product B			Product C		
		L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1	**Verify 1**									
	UCI	31.06%	26.92%	14.49%			30.10%	38.37%	41.41%	26.85%
	Attempt 1	**25.62%**	**21.71%**	**10.32%**			**24.83%**	**32.76%**	**35.74%**	**21.72%**
	LCI	20.87%	17.29%	7.26%			20.24%	27.62%	30.46%	17.37%
	UCI	16.88%	14.49%	8.29%			14.64%	31.22%	33.62%	23.17%
	Attempt 2	**12.46%**	**10.32%**	**4.98%**			**10.54%**	**25.86%**	**28.18%**	**18.28%**
	LCI	9.08%	7.26%	2.94%			7.51%	21.16%	23.32%	14.25%
	UCI	6.98%	6.54%	3.28%			10.35%	27.58%	26.76%	20.18%
	Attempt 3	**3.91%**	**3.56%**	**1.07%**			**6.80%**	**22.41%**	**21.65%**	**15.52%**
	LCI	2.14%	1.88%	0.23%			4.41%	17.99%	17.30%	11.80%
	UCI	20.93	19.01	23.57			8.11	8.99	10.52	12.57
	RTT (sec)	**19.45**	**17.84**	**21.57**			**7.42**	**8.20**	**9.55**	**11.27**
	LCI	17.96	16.67	19.58			6.68	7.41	8.58	9.97
	Verify 2									
	UCI	26.74%	30.47%	16.38%			34.95%	39.78%	41.54%	28.31%
	Attempt 1	**21.55%**	**25.09%**	**12.01%**			**29.49%**	**34.14%**	**35.86%**	**23.10%**
	LCI	17.16%	20.40%	8.71%			24.58%	28.93%	30.57%	18.62%
	UCI	14.39%	15.59%	9.08%			20.21%	32.30%	30.85%	23.17%
	Attempt 2	**10.25%**	**11.31%**	**5.65%**			**15.59%**	**26.90%**	**25.52%**	**18.28%**
	LCI	7.21%	8.11%	3.47%			11.89%	22.12%	20.85%	14.25%
	UCI	7.37%	6.05%	4.69%			14.97%	27.58%	27.58%	20.93%
	Attempt 3	**4.24%**	**3.18%**	**2.12%**			**10.85%**	**22.41%**	**22.41%**	**16.21%**
	LCI	2.39%	1.61%	0.88%			7.77%	17.99%	17.99%	12.41%
	UCI	18.80	18.75	22.15			8.25	9.08	9.88	12.28
	RTT (sec)	17.60	17.52	20.31			7.58	8.28	8.96	10.99
	LCI	16.39	16.30	18.47			6.86	7.48	8.04	9.70
Visit 2	**Identify 1**									
	UCI	38.52%	35.04%	21.67%			39.58%	39.45%	44.77%	20.86%
	Attempt 1	**31.18%**	**27.81%**	**15.38%**			**33.02%**	**33.46%**	**38.67%**	**15.81%**
	LCI	24.70%	21.61%	10.70%			27.09%	27.98%	32.92%	11.82%
	UCI	20.90%	21.01%	11.44%			19.29%	27.68%	28.61%	14.25%
	Attempt 2	**14.71%**	**14.79%**	**6.51%**			**13.95%**	**22.18%**	**23.05%**	**9.88%**
	LCI	10.14%	10.20%	3.59%			9.93%	17.53%	18.31%	6.76%
	UCI	11.38%	10.72%	6.19%			14.56%	20.98%	20.20%	9.18%
	Attempt 3	**6.47%**	**5.92%**	**2.37%**			**9.77%**	**15.95%**	**15.23%**	**5.53%**
	LCI	3.56%	3.15%	0.74%			6.44%	11.97%	11.34%	3.27%
	UCI	20.17	18.99	24.29			9.70	10.62	12.67	13.43
	RTT (sec)	18.64	17.57	21.77			8.78	9.49	11.43	12.06
	LCI	17.11	16.15	19.25			7.86	8.35	10.19	10.70
	Identify 2									
	UCI	37.30%	38.12%	23.63%			41.48%	38.13%	42.06%	21.46%
	Attempt 1	**30.00%**	**30.77%**	**17.16%**			**34.88%**	**32.16%**	**35.97%**	**16.33%**
	LCI	23.62%	24.31%	12.20%			28.84%	26.73%	30.31%	12.27%
	UCI	24.15%	23.63%	14.95%			25.40%	26.22%	30.19%	13.45%
	Attempt 2	**17.65%**	**17.16%**	**9.47%**			**19.53%**	**20.78%**	**24.51%**	**9.16%**
	LCI	12.63%	12.20%	5.85%			14.78%	16.25%	19.62%	6.15%
	UCI	10.66%	12.86%	6.19%			15.10%	21.99%	19.13%	7.31%
	Attempt 3	**5.88%**	**7.69%**	**2.37%**			**10.23%**	**16.86%**	**14.23%**	**3.98%**
	LCI	3.13%	4.48%	0.74%			6.82%	12.76%	10.45%	2.11%
	UCI	18.79	19.12	24.56			8.72	8.94	11.32	11.39
	RTT (sec)	17.45	17.78	21.95			7.96	8.15	10.15	10.30
	LCI	16.10	16.44	19.34			7.13	7.35	8.98	9.22

Table 11-9. Cumulative 3rd-Attempt GFRR by Visit									
	Product A			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Visit 1 (Verify 1 and Verify 2)									
UCI	6.60%	5.53%	3.23%			12.21%	27.48%	27.05%	20.46%
Attempt 3	**4.08%**	**3.37%**	**1.60%**			**8.83%**	**22.41%**	**22.03%**	**15.86%**
LCI	2.49%	2.04%	0.78%			6.32%	18.05%	17.72%	12.14%
UCI	19.87	18.88	22.86			8.18	9.03	10.20	12.43
RTT (sec)	**18.52**	**17.68**	**20.94**			**7.50**	**8.24**	**9.25**	**11.13**
LCI	17.17	16.48	19.02			6.77	7.44	8.31	9.83
Visit 2 (Identify 1 and Identify 2)									
UCI	10.02%	10.70%	5.36%			14.22%	21.23%	19.33%	7.94%
Attempt 3	**6.18%**	**6.80%**	**2.37%**			**10.00%**	**16.41%**	**14.73%**	**4.76%**
LCI	3.74%	4.26%	1.03%			6.93%	12.51%	11.08%	2.82%
UCI	19.48	19.06	24.42			9.21	9.78	11.99	12.41
RTT (sec)	**18.04**	**17.67**	**21.86**			**8.37**	**8.82**	**10.79**	**11.18**
LCI	16.60	16.29	19.30			7.50	7.85	9.59	9.96

Table 11-10. Overall Cumulative 3rd-Attempt GFRR									
	Product G			**Product B**			**Product C**		
	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye	L Eye	R Eye	L or R Eye
Overall (Verify 1, Verify 2, Identify 1, Identify 2)									
UCI	7.21%	6.71%	3.44%			12.29%	23.96%	22.84%	13.93%
Attempt 3	**4.87%**	**4.66%**	**1.88%**			**9.32%**	**19.60%**	**18.62%**	**10.70%**
LCI	3.26%	3.21%	1.03%			7.02%	15.86%	15.03%	8.15%
UCI	19.68	18.97	23.64			8.69	9.41	11.10	12.42
RTT (sec)	**18.28**	**17.68**	**21.40**			**7.93**	**8.53**	**10.02**	**11.16**
LCI	16.89	16.39	19.16			7.13	7.65	8.95	9.90

11.4 Offline versus Online False Non-Matches

Table 11-11. Visit 2, Identify 2 Online and Offline FNMR						
	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product A						
Attempt 1	0.0246 (3/122)	0.0410 (5/122)	0.0331 (4/121)	0.0413 (5/121)	0.0000 (0/140)	0.0429 (6/140)
Attempt 2	0.0071 (1/141)	0.0355 (5/141)	0.0000 (0/140)	0.0429 (6/140)	0.0000 (0/153)	0.0261 (4/153)
Attempt 3	0.0062 (1/161)	0.0373 (6/161)	0.0000 (0/156)	0.0192 (3/156)	0.0000 (0/165)	0.0121 (2/165)
Product B						
Attempt 1					0.2135 (38/178)	0.0787 (14/178)
Attempt 2					0.0942 (18/191)	0.0628 (12/191)
Attempt 3					0.0203 (4/197)	0.0305 (6/197)
Product C						
Attempt 1	0.1840 (39/212)	0.1226 (26/212)	0.2249 (47/209)	0.1770 (37/209)	0.1139 (27/237)	0.1055 (25/237)
Attempt 2	0.0818 (18/220)	0.0909 (20/220)	0.1239 (27/218)	0.1284 (28/218)	0.0539 (13/241)	0.0747 (18/241)
Attempt 3	0.0364 (8/220)	0.0273 (6/220)	0.0181 (4/221)	0.0362 (8/221)	0.0000 (0/241)	0.01245 (3/241)
(# of false non-matches / # of genuine comparisons)						

Table 11-12. Visit 2, Identify 2 Online and Offline FNMR with UINs						
	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product A						
Attempt 1	0.0246 (3/122)	0.0410 (5/122)	0.0331 (4/121)	0.0413 (5/121)	0.0000 (0/140)	0.0429 (6/140)
UINs	3044 4065 6005	3044 3063 4027 4081 5080	3000 5046 5067 6022	3000 3041 5030 5067 6021		3041 3044 4081 5030 5067 6021
Attempt 2	0.0071 (1/141)	0.0355 (5/141)	0.0000 (0/140)	0.0429 (6/140)	0.0000 (0/153)	0.0261 (4/153)
UINs	3044	3063 3069 4027 4081 5080		3000 3031 5007 5067 6021 6088		4081 5067 6021 3031
Attempt 3	0.0062 (1/161)	0.0373 (6/161)	0.0000 (0/156)	0.0192 (3/156)	0.0000 (0/165)	0.0121 (2/165)
UINs	3044	3063 3069 4027 4081 5080 6021		3000 6021 7004		4081 6021
Product B						
Attempt 1					0.2135 (38/178)	0.0787 (14/178)
UINs					0517 0530 0533 1024 1065 1080 2026 2030 2049 2061 2081 2082 2085 3005 3006 3025 3026 3040 3049 3088 4006 4042 4047 4049 4060 4081 4084 5020 5021 5025 5046 5060 5082 6001 6022 6065 6069 6071	0517 1065 2030 2082 3001 3025 3040 5021 5025 5040 5046 5082 6022 6071
Attempt 2					0.0942 (18/191)	0.0628 (12/191)
UINs					0517 1065 2030 2081 2082 3025 3088 4047 4049 4084 5021 5025 5046 6001 6022 6065 6069 6071	0517 2030 2042 2082 3040 4048 5025 5040 5046 6071 6080 7071
Attempt 3					0.0203 (4/197)	0.0305 (6/197)
UINs					1065 2082 5021 5046	0517 2042 2082 5040 5046 7071

	Left Eye		Right Eye		Left or Right Eye	
	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)	Online	Offline (at HD=0.32)
Product C						
Attempt 1	0.1840 (39/212)	0.1226 (26/212)	0.2249 (47/209)	0.1770 (37/209)	0.1139 (27/237)	0.1055 (25/237)
UINs	0514 0517 0520 0521 0533 1022 1024 1050 1061 2010 2030 2050 2068 2085 2086 2091 3006 3031 3043 3050 3082 3088 3090 4024 4042 4046 4047 4087 5020 5040 5045 5068 5082 5089 6001 6069 7004 7023 7041	0514 0520 0521 0533 1003 1047 2030 2050 2085 3031 3060 3082 4005 4024 4042 4047 4087 5020 5040 5045 5046 5068 5089 6001 7004 7041	0501 0506 0514 0520 0521 0524 0530 0533 1009 1021 1029 1049 1050 1088 2002 2029 2030 2049 2050 2061 2065 3005 3006 3031 3040 3049 3081 3082 3088 4010 4024 4060 4081 4084 4087 5000 5007 5020 5025 5040 5045 5047 5049 5067 5080 6080 7041	0514 0520 0521 0530 0533 1029 1049 1050 1084 1088 2029 2030 2041 2049 2050 2061 2065 3011 3025 3027 3031 3040 3081 3082 3088 4010 4024 4060 4081 4087 5000 5025 5045 5049 5067 6080 7049	0501 0514 0517 0520 0521 0533 1050 1088 2030 2050 2085 3005 3006 3031 3082 3088 4024 4047 4087 5020 5025 5040 5045 5067 5082 6001 7041	0514 0520 0521 0533 1029 1049 1088 2030 2050 2085 3011 3025 3031 3082 4005 4024 4042 4047 4087 5025 5045 5046 5067 5089 6001
Attempt 2	0.0818 (18/220)	0.0909 (20/220)	0.1239 (27/218)	0.1284 (28/218)	0.0539 (13/241)	0.0747 (18/241)
UINs	0514 0517 0520 0533 1022 1050 2030 2091 3031 3082 3088 4024 4042 4047 4087 5045 5089 7041	0514 0521 0533 1003 1027 1047 2030 2085 3031 3065 3082 4005 4024 4042 4047 4087 5045 5080 6022 7041	0501 0520 0530 0533 1029 1049 1050 2002 2061 3005 3006 3031 3040 3081 3082 4010 4024 4060 4081 4084 5007 5020 5025 5067 5080 6080 7041	0530 0533 1004 1029 1049 1050 2030 2041 2043 2049 2065 3011 3025 3027 3031 3040 3081 3082 4010 4024 4042 4060 4087 5049 5067 5089 6022 6080	0501 0517 0520 0533 1050 3005 3031 3082 4024 4047 5025 5067 7041	0533 1004 1027 1029 1049 2030 2085 3011 3025 3031 3082 4005 4024 4042 4047 4087 5067 6022
Attempt 3	0.0364 (8/220)	0.0273 (6/220)	0.0181 (4/221)	0.0362 (8/221)	0.0000 (0/241)	0.01245 (3/241)
UINs	0520 0533 1022 3031 3082 4042 4047 5089	0533 3065 3082 4005 4042 5080	1049 3006 5007 5067	1049 2043 3027 4010 4042 4087 5089 6049		1049 4005 4042
(# of false non-matches / # of genuine comparisons), yellow highlight UINs (xxxx) matched offline but did not match online, green highlight UINs (xxxx) matched online but did not match offline						

It is worth mentioning that subsequent attempts, in some cases, do not result in fewer false matches. For example, for Product A, Right eye, the offline 2nd attempt FNMR is greater than the 1st attempt FNMR in Table 11-11, and the 2nd attempt TMR=1-FNMR is less than the 1st attempt TMR in Figure 11-1. Intuitively, we would expect FNMR to decrease (and TMR to increase) with increasing attempts. However, we observe in Table 11-12 that while UINs 3041 and 5030 failed to match offline during the 1st attempt and successfully matched in the 2nd attempt, three additional UINs (3031, 5007, and 6088) failed to match offline during the 2nd attempt. These three UINs failed to acquire during the 1st attempt. Sixteen other UINs failed to acquire during the 1st attempt and successfully acquired and matched offline during the 2nd attempt. In summary, $FNMR_{Attempt\text{-}1}$=5/121=4.13% and $FNMR_{Attempt2}$=6/140=4.29%, and the inclusion of additional test subjects in subsequent attempts that failed to acquire during previous attempts explains why subsequent-attempt FNMR is sometimes lower than the FNMR for previous attempts.

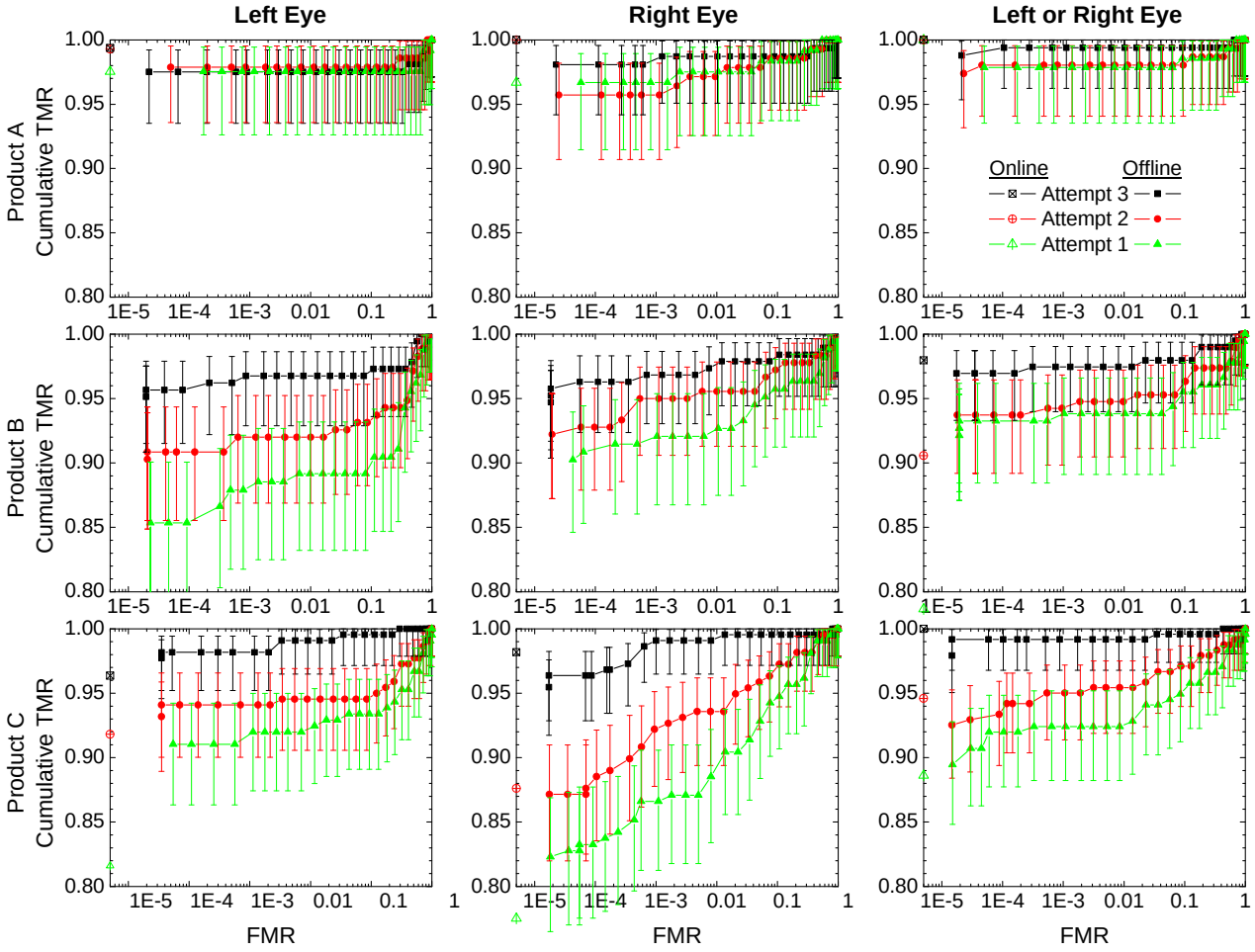


Figure 11-1. Visit 2, Identify 2 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

11.5 Offline ROC curves

11.5.1. Basic

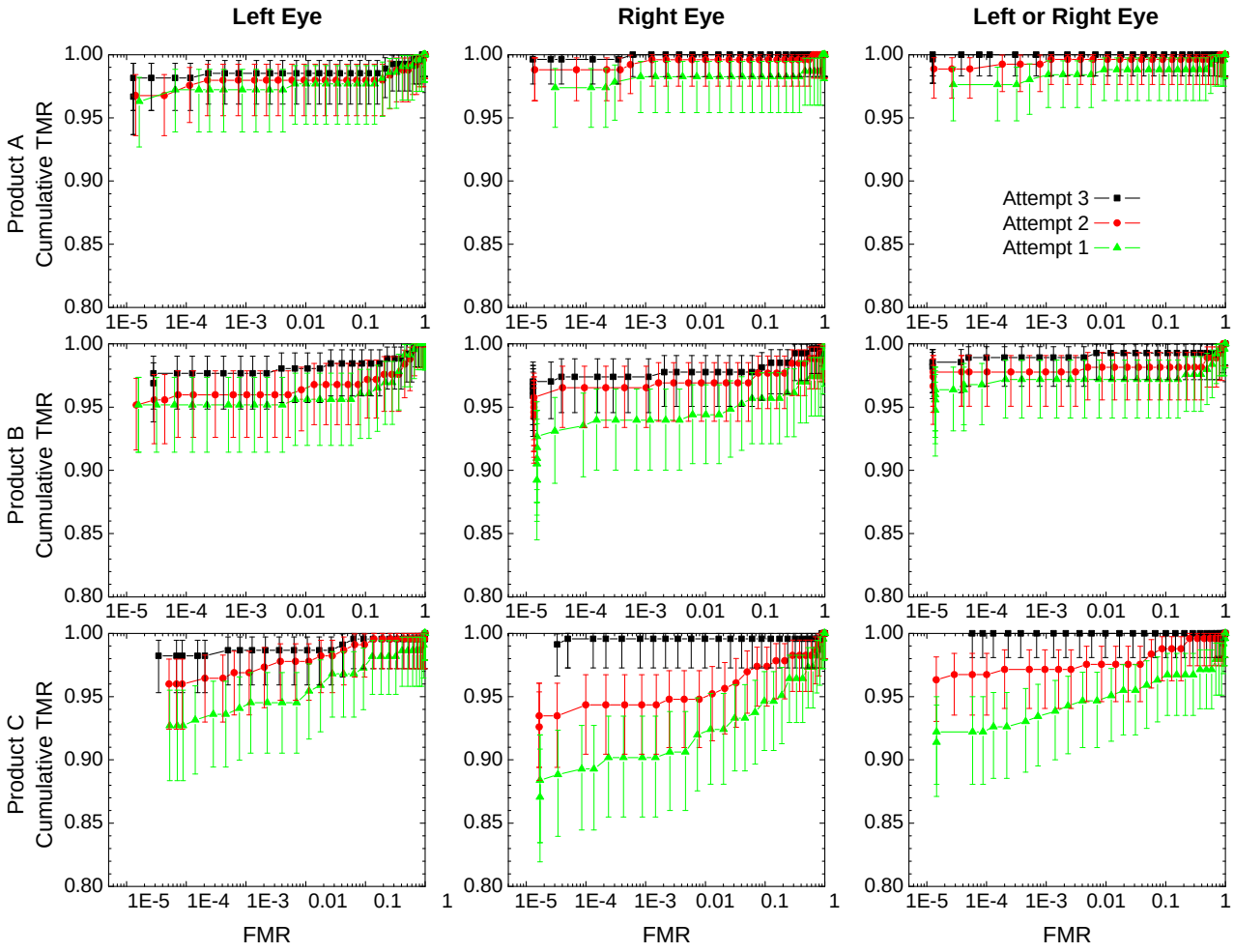


Figure 11-2. Visit 1, Verify 1 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

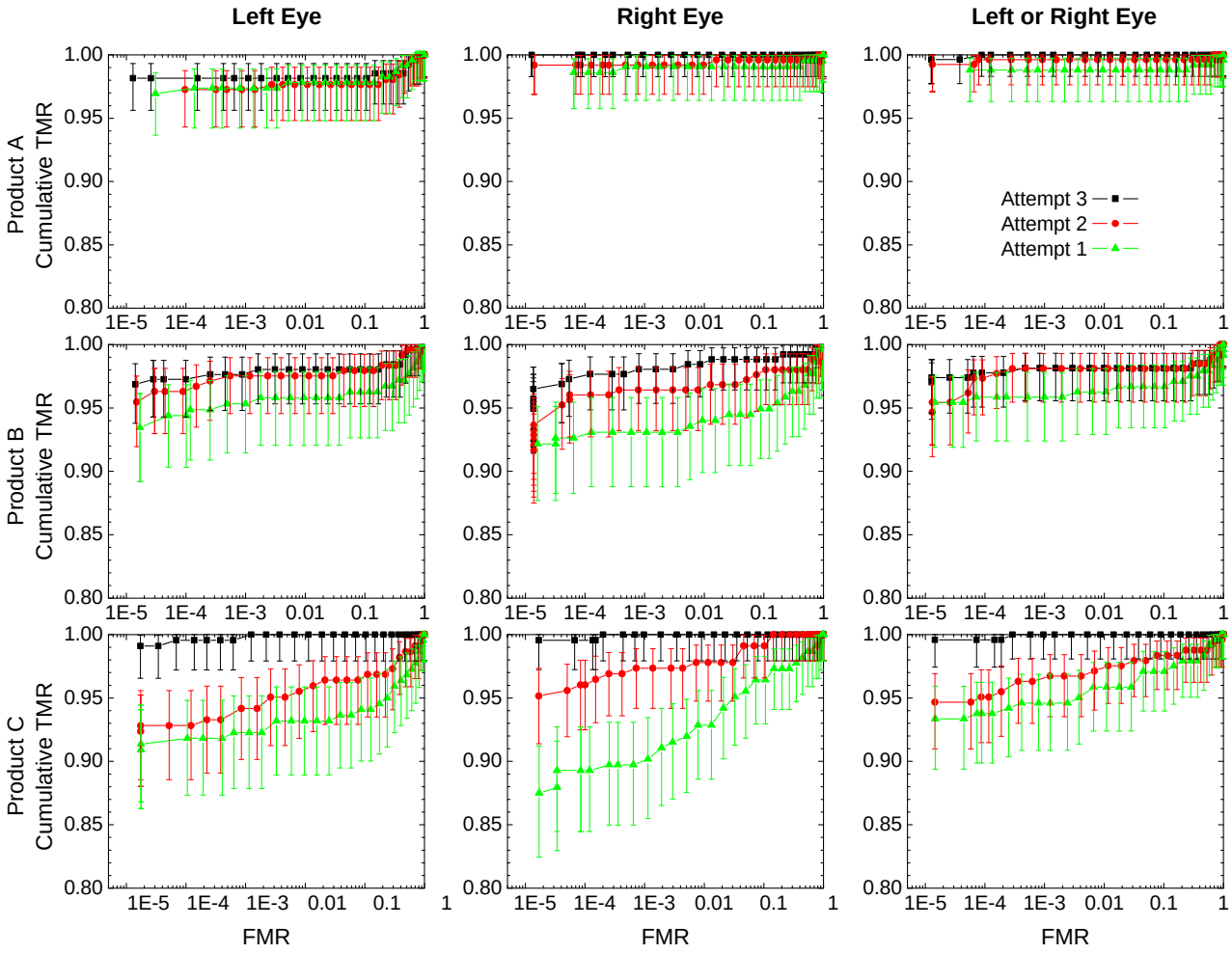


Figure 11-3. Visit 1, Verify 2 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

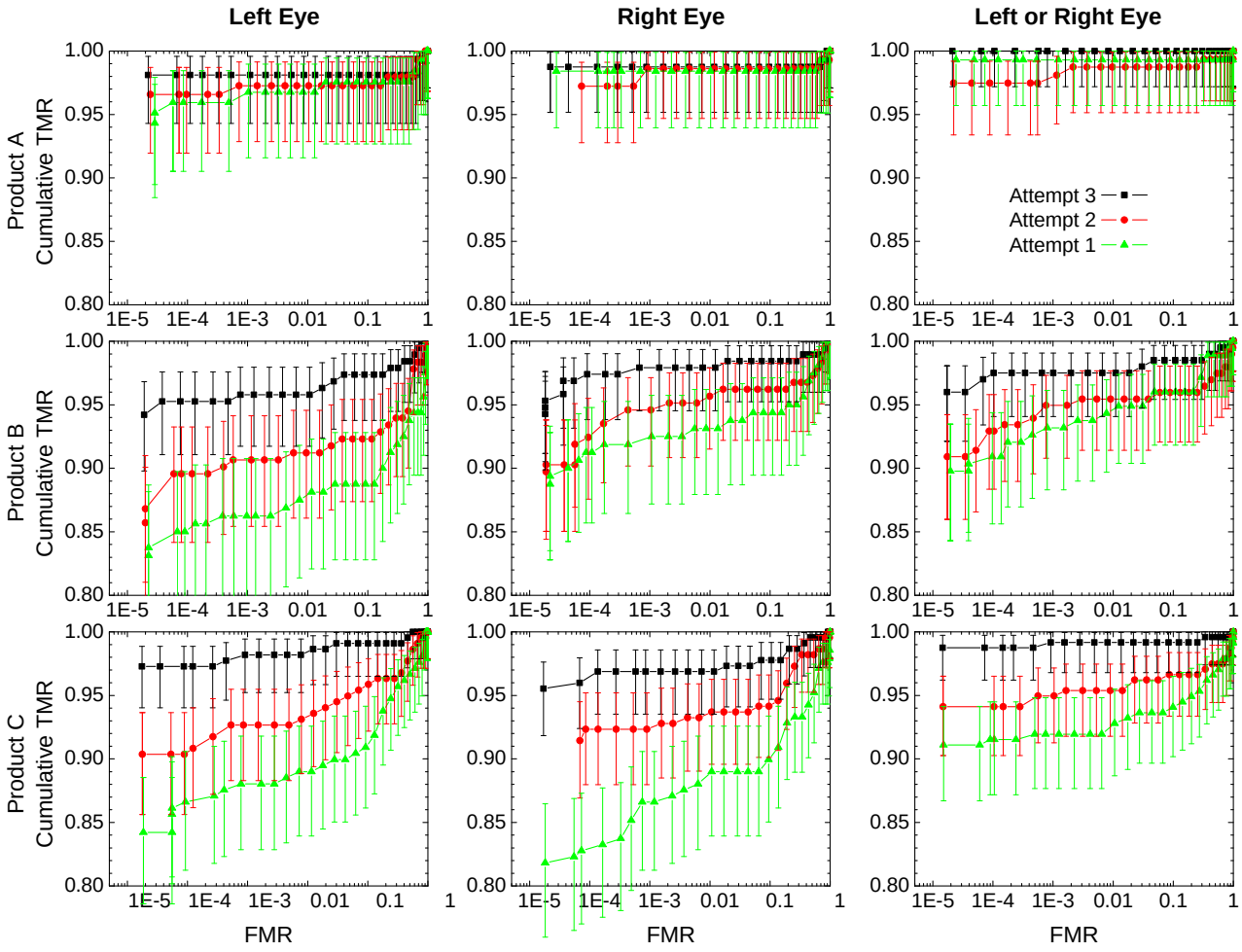


Figure 11-4. Visit 2, Identify 1 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

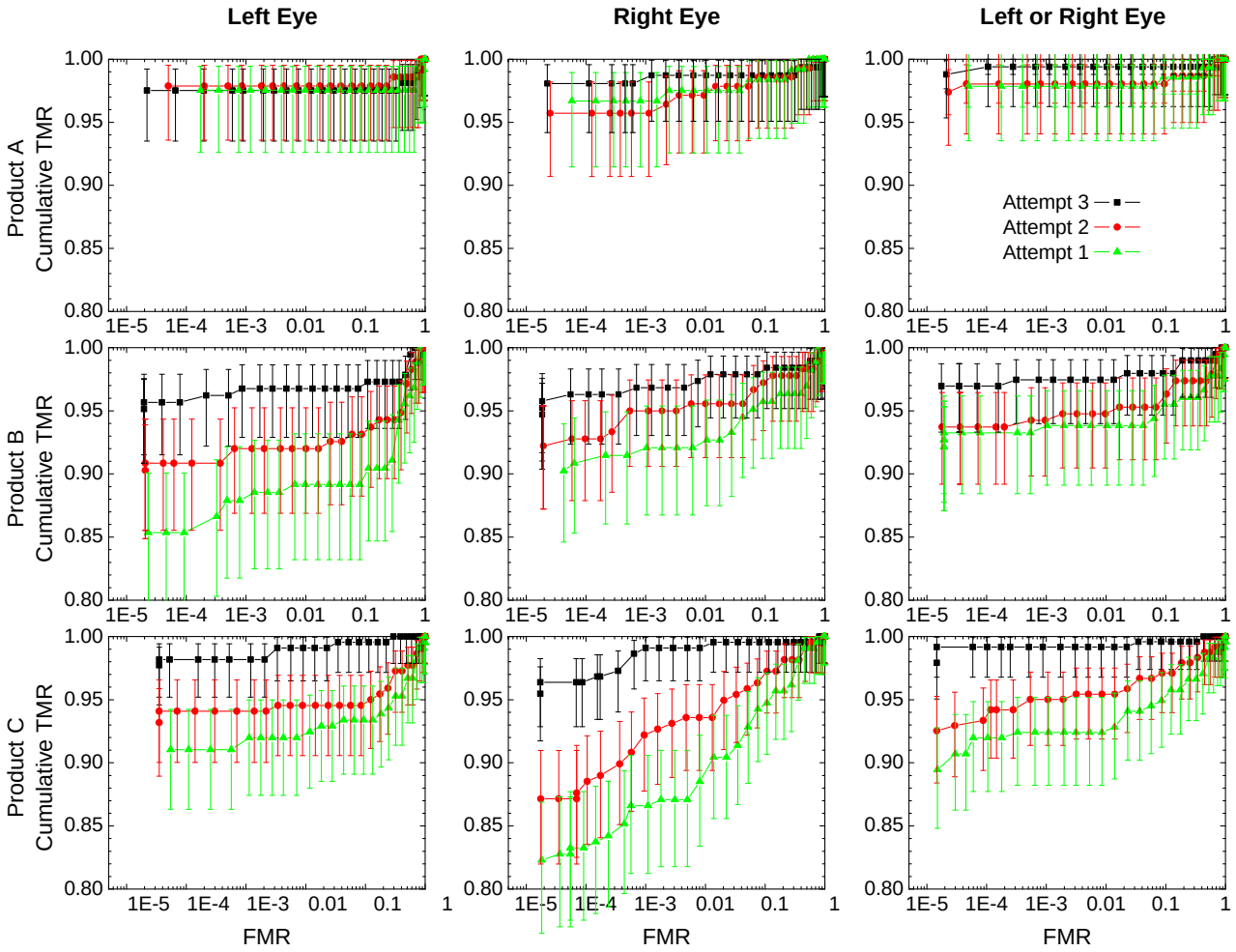


Figure 11-5. Visit 2, Identify 2 Basic Cumulative Performance Curves by Attempt with 95% Confidence Intervals

11.5.2. Generalized

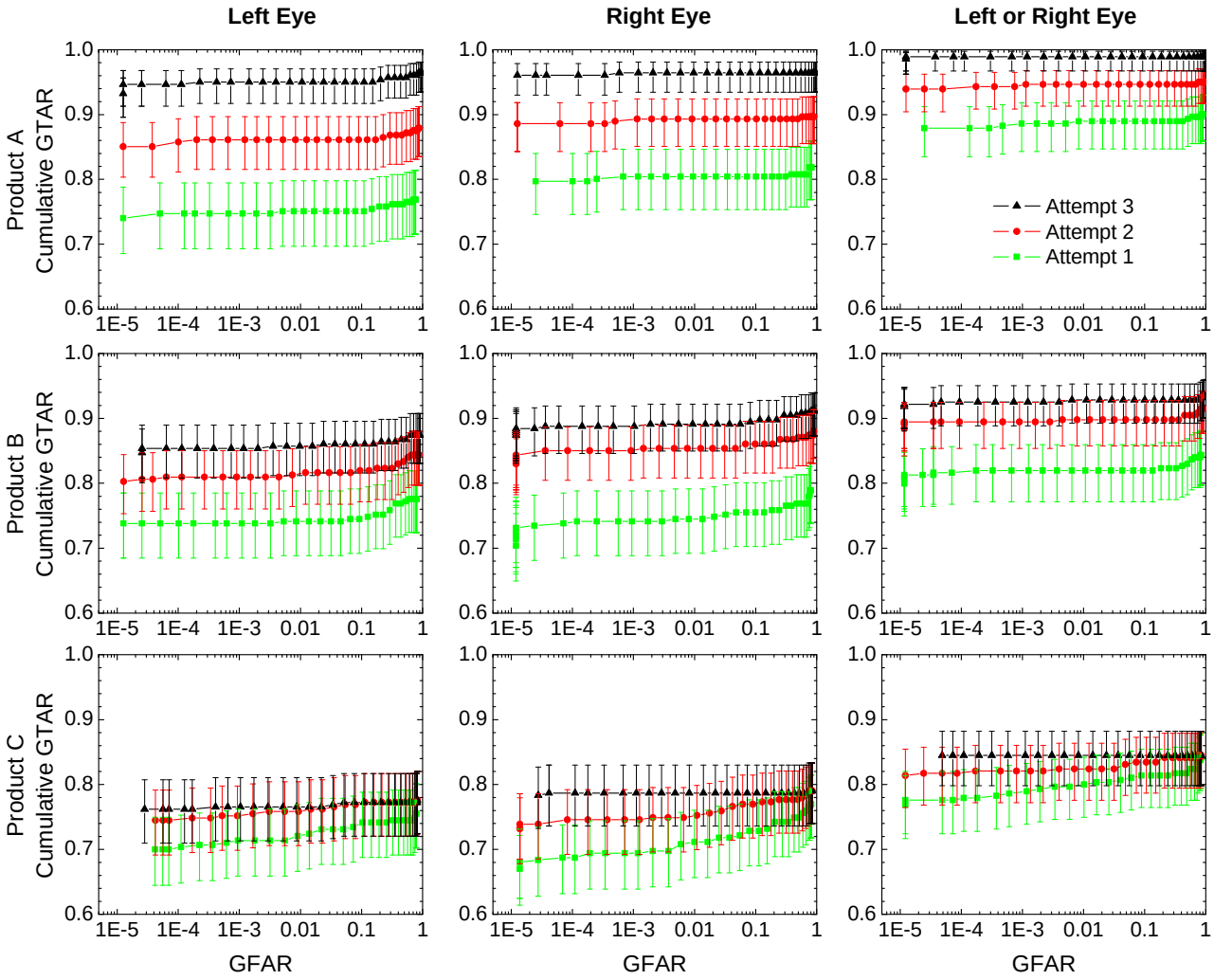


Figure 11-6. Visit 1, Verify 1 Generalized Cumulative Performance Curves by Attempt with 95% Confidence Intervals

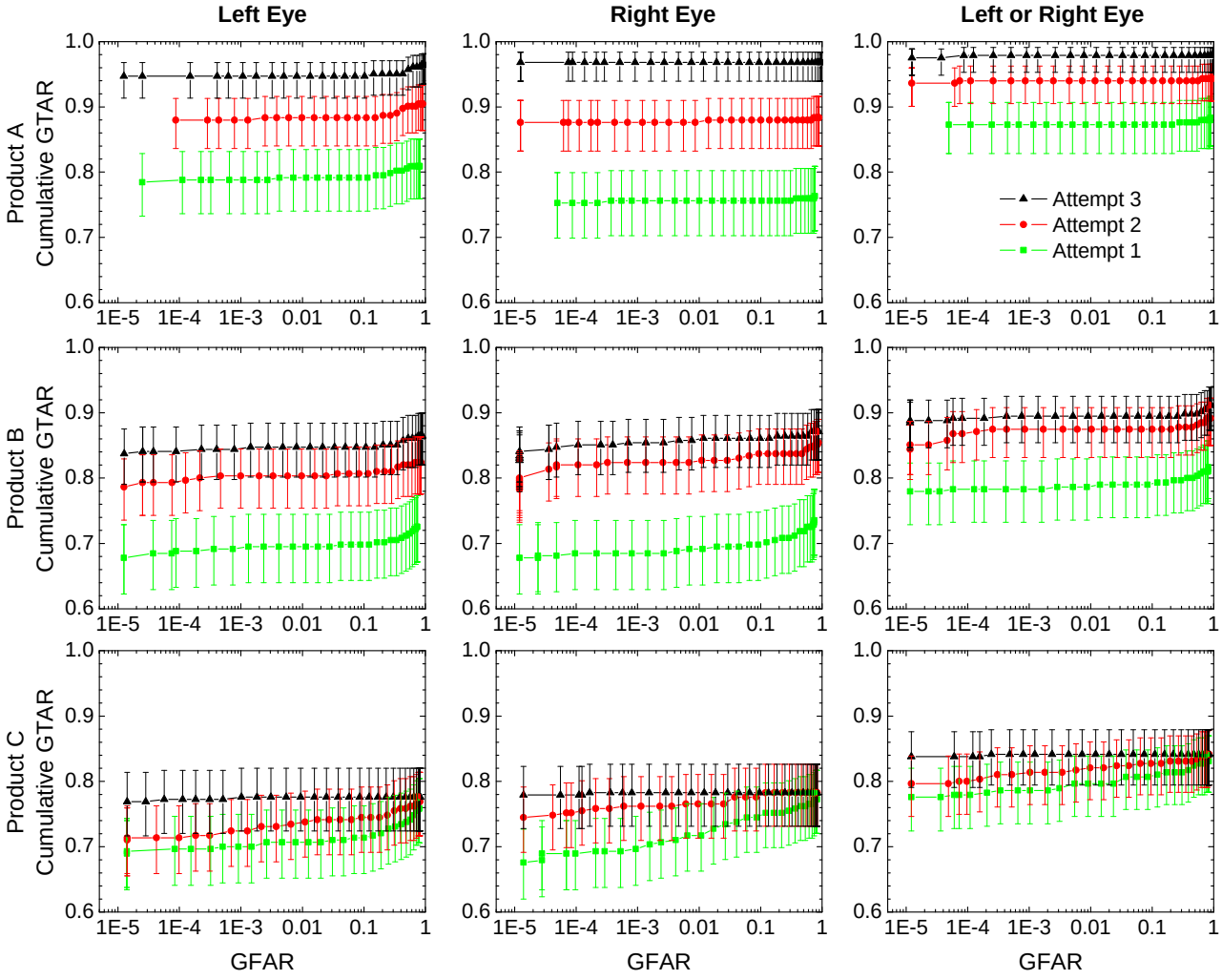


Figure 11-7. Visit 1, Verify 2 Generalized Cumulative Performance Curves by Attempt with 95% Confidence Intervals

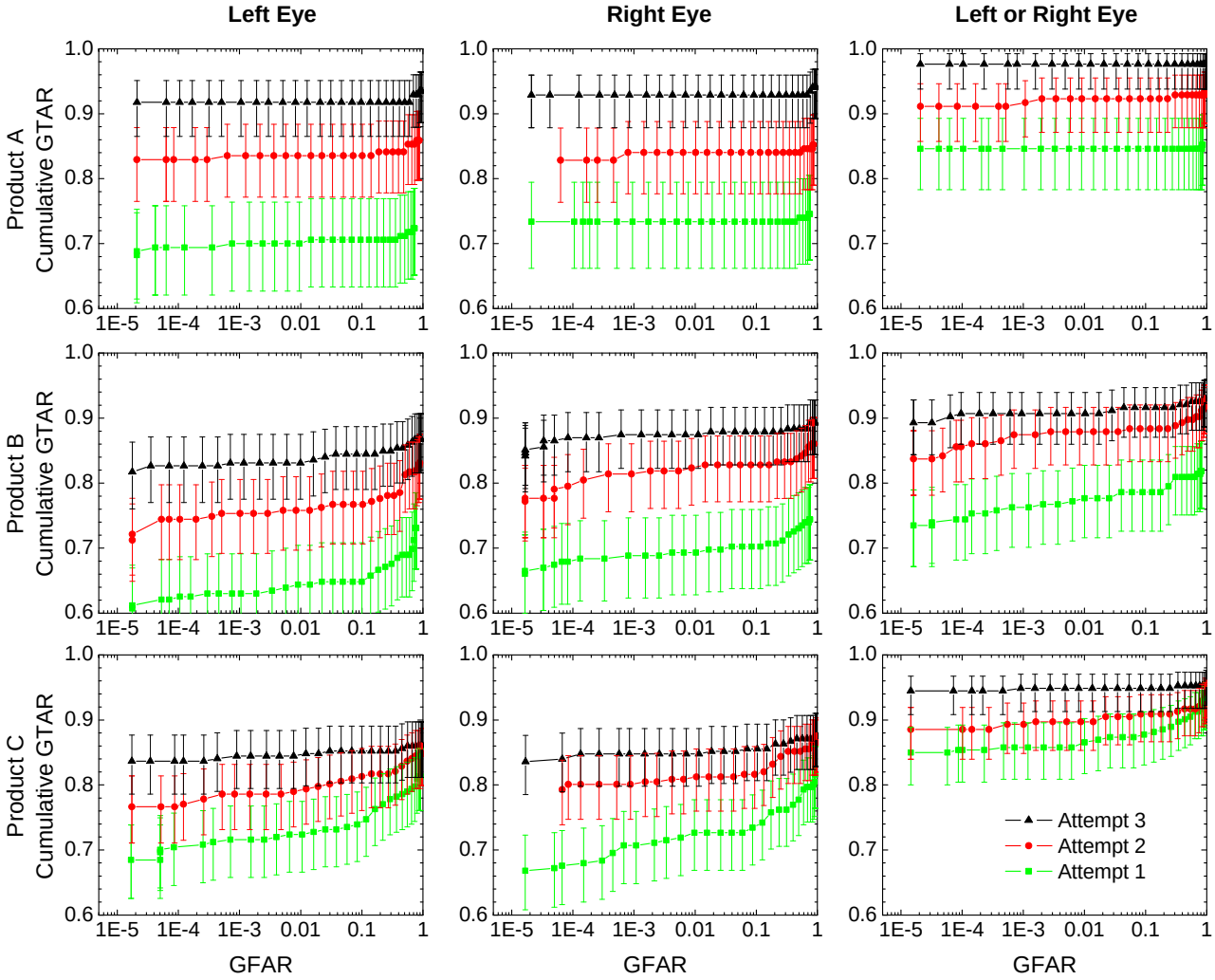


Figure 11-8. Visit 2, Identify 1 Generalized Cumulative Performance Curves by Attempt with 95% Confidence Intervals

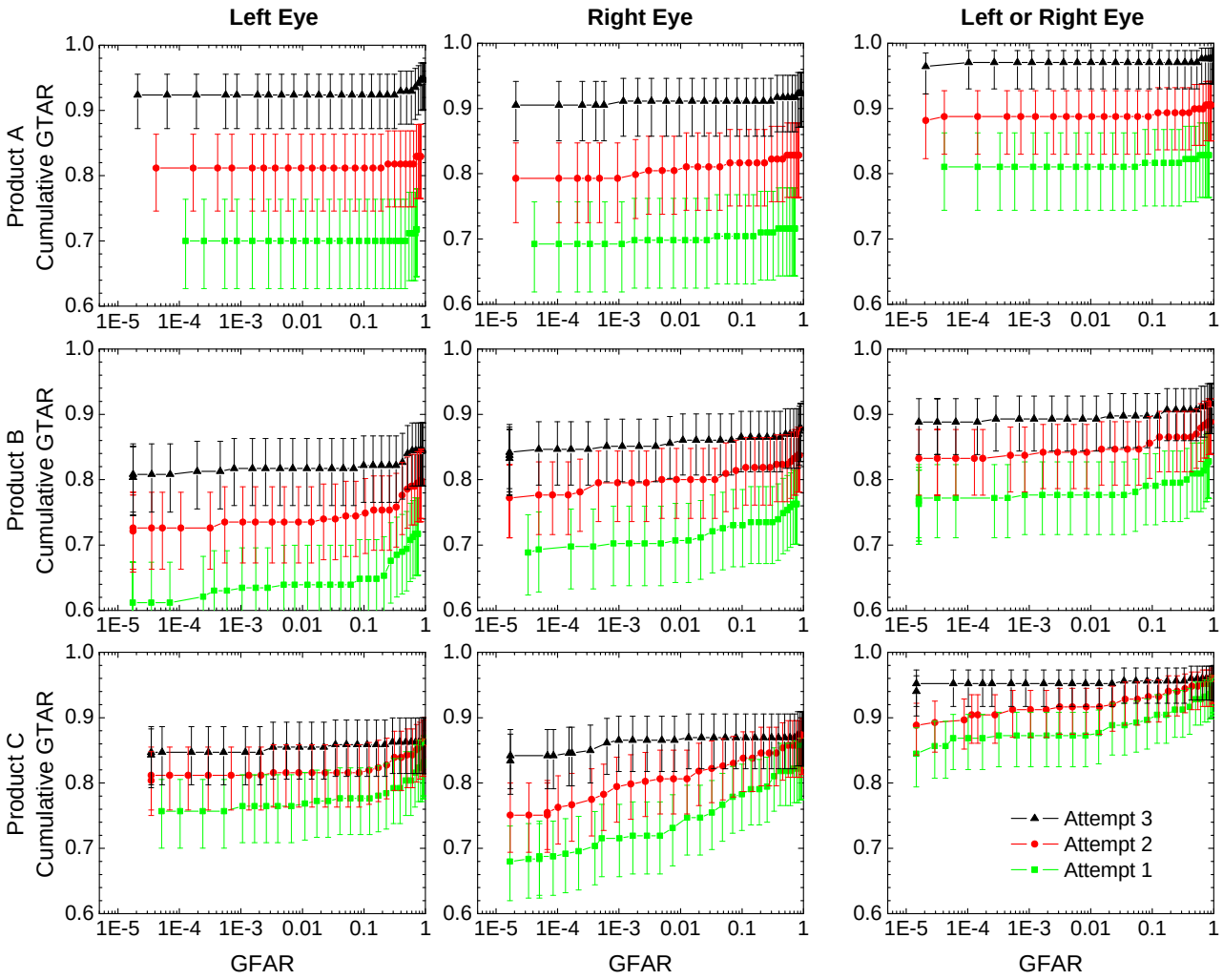


Figure 11-9. Visit 2, Identify 2 Generalized Cumulative Performance Curves by Attempt with 95% Confidence Intervals