IREX IV: Part 1

Evaluation of Iris Identification Algorithms

NIST Interagency Report 7949

George W. Quinn, Patrick Grother, and Mei Ngan

Information Access Division
National Institute of Standards and Technology

July 19, 2013

Acknowledgements

The authors would like to thank the sponsors of this activity. These are the Federal Bureau of Investigation, the U.S. VISIT office in the Department of Homeland Security (DHS), and the Science and Technology Directorate, also in DHS. Vitally, the authors would also like to thank the United States Department of Defense's Biometrics Identity Management Agency for their support and active collaboration.

Disclaimer

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

Executive Summary

Introduction

The Iris Exchange (IREX) IV evaluation aims to provide a scientific evaluation of the performance of automated iris identification algorithms. IREX IV evaluated the performance of 66 identification algorithm prototypes submitted by 12 research institutions. The greater number of participants in this evaluation compared to the previous IREX III evaluation reflects the growing market for iris-based technology. Performance was evaluated over a set of over 4 million field-collected iris images. IREX IV is the fourth activity in the IREX Program. The IREX Program was established by the National Institute of Standards and Technology (NIST) in 2009 to provide qualitative and quantitative support for iris recognition standardization, development, and deployment.

IREX IV's novel contribution to the IREX program is the use of a cost estimation model as a new performance metric. A cost equation is introduced to assess recognition performance for particular applications. The equation defines a preferred trade-off between security and convenience. A system that controls access to sensitive information should prioritize security, while a system that grants access to rides at a theme park is likely to place greater emphasis on the user's experience. The goal is to see if algorithm developers can optimize their algorithms for specific applications.

Biometric evaluations promote industrial competitiveness by providing a fair platform for comparison. As a technology test, IREX IV is also intended to identify areas for future research and development. It also offers algorithm developers, including participants from previous IREX evaluations, an opportunity to further develop and advance their technologies.

Industry Growth and Market Drivers

Since the expiration of the Flom-Safir patent in 2005 [1], the iris recognition industry has experienced rapid growth. This was evidenced by NIST's previous IREX III evaluation, which tested 95 iris identification algorithm prototypes provided by 11 research institutions, representing an order of magnitude increase in the number of providers over the past several years. The evaluation also affirmed the potential for iris recognition to accomplish large-scale identity management tasks.

Since then, several government agencies have deployed, or are in the process of procuring, iris recognition systems that operate on a national scale. The Unique Identity Authority of India (UIDAI) has collected iris images of hundreds of millions of Indian residents for the purpose of better managing the allocation of resources to its citizens [2, 3]. The United Arab Emirates (UAE) also employs iris recognition for its border-crossing control system, searching visitors against a watch list of several hundred thousand people expelled from the country for various violations. The Federal Bureau of Investigation (FBI) includes iris recognition technology on its technical roadmap. Since 2006, the Department of Defense (DOD) has been using handheld devices to collect iris images of people in various theaters of operation. The images are consolidated into a central repository known as the Automated Biometric Identification System (ABIS) and are currently used in several applications.

Key Results

The key results of the IREX IV evaluation are described below.

- **Core Accuracy and Speed:** Iris recognition is known to be one of the more accurate and computationally efficient biometric technologies. Some matchers are capable of searching a single iris image against an enrolled population of 1.6 million people in less than a second (using just one processing core) while achieving false negative identification (i.e. "miss") rates below 1.5 % at reasonably selective decision thresholds. This error rate drops by a factor of two when both eyes are used, but search time increases by a factor between 3 and 4.

  Identification failures for the most accurate matchers were almost always the result of poor sample quality, where the eye is closed, off-axis, highly rotated, etc. Many of these errors can be corrected through the use of more advanced cameras or improved image collection and data handling practices.

- **Cost Model Optimization:** Participants submitted two classes of recognition algorithms, one intended for high-security applications and one intended to prioritize convenience for the user. Rather than attempting to optimize matching behavior to either type of application, most participants strived for across-the-board improvements in accuracy. However, one participant submitted algorithms that adjust well to the application. Part of the reason only

one participant was successful at application-specific optimization might be that participants were not provided with sufficient development data that could lead to specific algorithm modifications.

- **Contingent Use of Second Eye for Rapid Identification:** When applied to iris recognition, contingent fusion uses the second eye for identification only if the first does not return a decisive match. IREX IV found that contingent fusion achieves accuracy comparable to that of two-eye matching while only using the second eye between 1 % and 2 % of the time. There are several advantages to requiring the use of only one eye for the majority of transactions. First, it is less computationally expensive. Second, it consumes less bandwidth when the samples are transferred over a network for a back-end search. Finally, it obviates the need for operators of single-eye cameras to capture an image of the second eye when the first proves sufficient.

- **Comparison to IREX III:** Although IREX III was conducted only 12 months prior to this evaluation, progress in automated recognition technology can occur fairly rapidly. Despite a moderately more difficult selection of iris images, the most accurate matcher in the current evaluation achieved slightly lower error rates than the most accurate one from IREX III. Participants who submitted algorithms that are not among the most accurate improved the most since IREX III. The disparity between best and worst matchers has reduced since IREX III.

  IREX III tested over iris images that had been compressed with JPEG at a quality of 75, which introduced compression artifacts that are visibly apparent in some images. IREX IV measured the impact of this compression on recognition accuracy and found that it increases miss rates by 2 to 11 % depending on the matching algorithm. The ISO/IEC 19794-6 and ANSI/NIST-ITL 1-2011 standards identify JP2 and PNG as superior compression methods for iris images.

- **Image Compression:** The ISO/IEC 19794-6 and ANSI/NIST-ITL 1-2011 standards define compact interoperable formats for storing iris images. Compact formats are useful for transmitting iris images across bandwidth-limited networks or for storage on space-limited devices such as smart-cards. It is also critical that such formats be interoperable to facilitate inter-agency exchange of iris data and to prevent vendor lock-in. The most compact format defined in the standards is the Kind 7 record. When converted to Kind 7 records, iris images can be compressed with JPEG 2000 to a kilobyte at the cost of a factor of two increase in the miss rate. Further investigation into JPEG 2000 image compression will be covered in a follow-up report.

Errata and Clarifications

- *July 19, 2013*: Changed color of box for algorithm U12B in Figure 27 to indicate that it is paired with F02P.
- *July 19, 2013*: Added curve for algorithm K01N to Figure 10 in the appendix.
- *July 19, 2013*: Corrected FNIR number for algorithm D02N in Figures 12 and 14. The previous number was erroneously high.
- *July 19, 2013*: In the summary tables in the appendix, it is possible for an algorithm's "best guess" decisions to achieve lower cost than the point of minimum cost along its DET curve. This can happen if the algorithm's decisions are better than decisions made solely by thresholding the dissimilarity score.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose

IREX IV builds upon IREX III [4] as a technology evaluation of iris identification using operational test data. It focuses on practical applications of iris recognition with an emphasis on large-scale deployments. Twelve commercial organizations and academic institutions submitted 66 iris recognition algorithm prototypes for evaluation.

Iris recognition is one of the more accurate and computationally efficient biometric technologies [5, 6]. This has led several government agencies to adopt iris-based methods of identifying people on a national scale. The largest is India's UID program [7] which contains iris images of hundreds of millions of Indian residents. The program was initiated to better manage the allocation of government resources and to provide improved services to citizens. The UAE also employs iris recognition as part of its border-crossing control system [8, 9]. At ports of entry, visitors are searched against a watch list of several hundred thousand people previously expelled from the country for various violations. Since 2006, the US Department of Defense has been using a variety of cameras (for example, Handheld Interagency Identity Detection Equipment (HIIDE) [10]) to collect iris images of subjects to support a variety of missions from tactical operations to detention management.

The main goals of this evaluation are to:

- *Provide an objective assessment of iris recognition technology:* The current evaluation aims to provide a fair scientific evaluation of the performance of automated iris recognition algorithms developed by research institutions. Rather than concentrating on a specific military or civilian application, the evaluation focuses on general algorithm performance and the common task of person identification to ensure its relevance to a wide range of applications.
- *Promote research and development:* Biometric evaluations promote industrial competitiveness by providing a fair platform for comparison. With an eye on the needs of sponsors and other current or potential users of the technology, the current evaluation seeks to identify areas for future research and development. IREX IV also offers algorithm developers, including participants from previous IREX evaluations, an opportunity to further develop and improve their recognition algorithms.
- *Evaluate both positive and negative identification performance:* Biometric systems take either a positive or negative approach to identification. Positive identification systems typically grant special privileges to enrolled users (e.g. access control) while negative identification systems often impose restrictions on them (e.g. watch lists). This report includes focused attention on both approaches to identification.
- *Introduce a cost model as a new performance metric:* The current evaluation introduces a cost equation for assessing recognition performance. Parameters in the model can be adjusted to reflect performance for a particular application. For example, systems that control access to highly-sensitive information should prioritize security, while systems that grant access to theme parks are likely to dedicate greater attention on the user's experience. Participants in IREX IV submitted recognition algorithms intended for use in both extreme scenarios. The goal is to see if algorithm developers can adjust algorithm behavior when the application is known in advance.
- *Investigate the problem of threshold calibration:* Biometric systems operate at an adjustable decision threshold that determines the strictness of the system's decisions. A poorly chosen decision threshold will lead to poor system performance. Identifying the best decision threshold for a particular application can be difficult if the probability distributions for comparison scores are not well known. This report investigates the problem of threshold selection by analyzing the stability and predictability of comparison score distributions.
- *Establish a standard JPEG 2000 compression profile:* Compact representation of the iris is important for applications that require samples to be stored on space-limited devices (such as smart cards) or transferred over limited-bandwidth networks. The second report stemming from this evaluation will establish a standard JPEG 2000 compression profile for the efficient compression of iris images. This report leads into that report with some initial results. This work lends support to the ISO/IEC 19794-6 [11] and ANSI/NIST-ITL 1-2011 [12] standards.

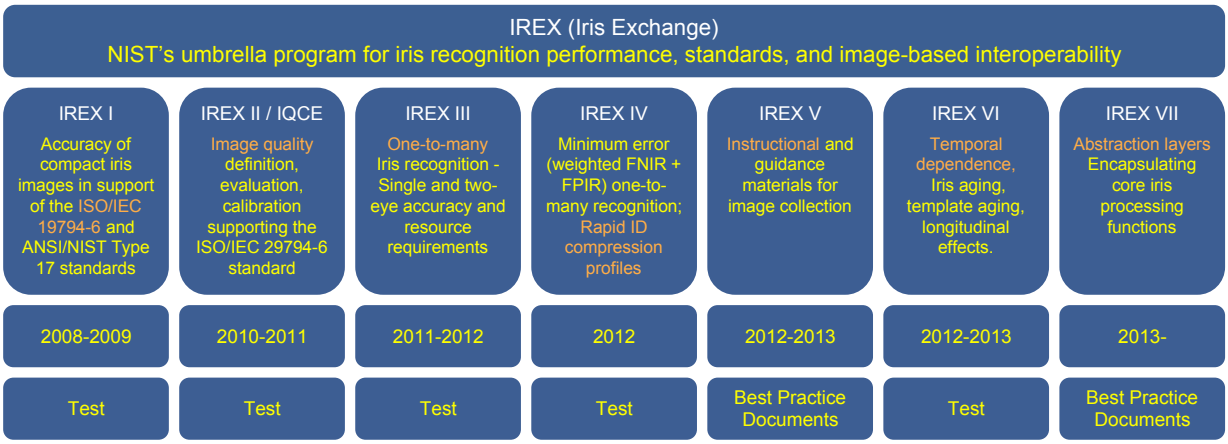


Figure 1: Timeline of the IREX program, including planned future installments.

1.2 The IREX Program

The IREX Program was initiated by NIST to support an expanded marketplace of iris-based applications. IREX provides quantitative support for iris recognition standardization, development, and deployment. To date, 4 activities have been completed and 3 more are tentatively planned (see Figure 1). Each is summarized below.

- **IREX I** [5] was a large-scale, independently administered, evaluation of one-to-many iris recognition. It was conducted in cooperation with the iris recognition industry to develop and test standard formats for storing iris images. Standard formats are important for maintaining interoperability and preventing vendor lock-in. The evaluation was conducted in support of the ISO/IEC 19794-6 and ANSI/NIST-ITL 1-2011 standards.
- **IREX II** [13] supported industry by establishing a standard set of quality metrics for iris samples. Although iris recognition has the potential to be extremely accurate, it is highly dependent on the quality of the samples. The evaluation tested the efficacy of 14 automated quality assessment algorithms in support of the ISO/IEC 29794-6 standard [14].
- **IREX III** was a performance test of the latest iris recognition algorithms over operational data. Despite growing interest in iris-based technology, at the time there was a paucity of experimental data to support published theoretical considerations and accuracy claims. IREX III constituted the first public presentation of large-scale performance results using operational data.
- **IREX IV** builds upon IREX III as a performance test of one-to-many iris recognition. In addition to providing participants from previous evaluations an opportunity to further develop and test their recognition algorithms, this evaluation explores the potential for using a cost equation model for optimizing algorithms for specific applications.
- **IREX V** will provide best practice recommendations and guidelines for the proper collection and handling of iris images.
- **IREX VI** will explore a possible aging effect for iris recognition. The intrinsic features of the iris may naturally change over time in a way that affects recognition accuracy. Factors such as subject habituation and aging of the camera may also introduce a time dependency.
- **IREX VII** intends to define a framework for communication and interaction between components in an iris recognition system. By introducing layers of abstraction that isolate underlying vendor-specific implementation details, a system can become more flexible, extensible, and modifiable.

The latest information on the IREX Program can be found on the IREX website [6].

1.3 Application Scenarios

Most real-world applications of biometrics perform identification (i.e. one-to-many matching) in an *open-set* mode, meaning individuals are searched against a database of previously enrolled persons, but without any guarantee that searched indi-

viduals are enrolled. *Closed-set* applications, which assume every searched individual is enrolled (and thus only concern themselves with identifying the searched person from among the enrolled population) are operationally uncommon for iris and are not tested in this evaluation.

Open-set biometric systems take either a *positive* or *negative* approach to identification. Positive identification systems verify the implicit claim that the user is enrolled in the database and typically grant special privileges or access to enrolled users. For example, the NEXUS Program [15] uses the iris to positively identify registered travelers for expedited security screening at ports of entry in the US and Canada. Negative identification systems verify the claim that the person is *not* enrolled and often impose restrictions on enrolled individuals. An example is the UAE's border-crossing control system, which uses the iris to prevent expelled individuals from re-entering the country.

A deployed biometric system should be optimized to make the right tradeoff between security and convenience. For example, the primary concern for a system that grants access to highly sensitive and restricted information should be preventing unauthorized access, while a biometric system such as the one used for ticketing at Walt Disney World [16] should prioritize user convenience. Each participant in this evaluation submitted two types of matching algorithms. Each type corresponds to a different general class of applications. Table 1 provides a description and comparison of each. The goal is to see if algorithm developers can optimize their matching algorithms to specific applications when the parameters and costs of errors are known in advance.

Parameter	**Class P (Positive Identification)**	**Class N (Negative Identification)**
Application Type	any open-set identification system (e.g. watchlist, database de-duplication, access control).	
Class Description	High cost associated with false positives	High cost associated with false negatives
Example Applications	Access control to high value information, resources, or facilities.	Watchlists for high-profile individuals. Investigational-mode searches.
Enrolled Database Size	From $O(10^2)$ to $O(10^6)$ subjects.	
Prior NIST References	IREX III Final Report [4] IREX III Supplement I: Failure Analysis [17] Multiple Biometric Evaluation (MBE) 2010 [18]	
Performance Criteria	Primarily accuracy and speed. Also memory usage, scalability, template-size, etc.	

Table 1: Description and comparison of Class P and Class N matching algorithms tested in this evaluation. Participants were required to submit at least one algorithm for each class.

2 Methodology

A technology evaluation [19] focuses on algorithm performance over other performance factors that might be relevant to the deployment and operation of a biometric system (e.g. societal and economic factors, policy drivers, legacy data). Performance is assessed with metrics that provide a general idea of the technology's capabilities. The relative importance of these metrics will depend on how the technology is applied. For example, biometric systems that perform identification in real-time require rapid response times, while offline tasks such as database de-duplication are usually performed under more relaxed speed constraints.

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

2.1 Test Environment

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

Testing was performed on high-end PC-class blades running the Linux operating system, which is typical of central server applications. Most of the blades had 6 quad-core CPUs running at 3.47 GHz with 192 GB of main memory. The test harness used concurrent processing to distribute workload across dozens of blades.

2.2 Matching Algorithms

Twelve commercial organizations and academic institutions submitted 66 iris recognition software libraries for evaluation. The participation window opened on May 16th, 2012 and closed on August 2nd, 2012. Participation was open worldwide to anyone with the ability to implement a large-scale one-to-many iris identification algorithm. There was no charge to participate.

Participants provided their submissions to NIST as static or dynamic libraries compiled on a recent Linux kernel. The libraries were then linked against NIST's test driver code to produce executables. A further validation step was performed to ensure that the algorithms produced the correct output on NIST's test machines. The full process is described in the IREX IV Application Programming Interface (API) and Concept of Operations (CONOPS) document [20].

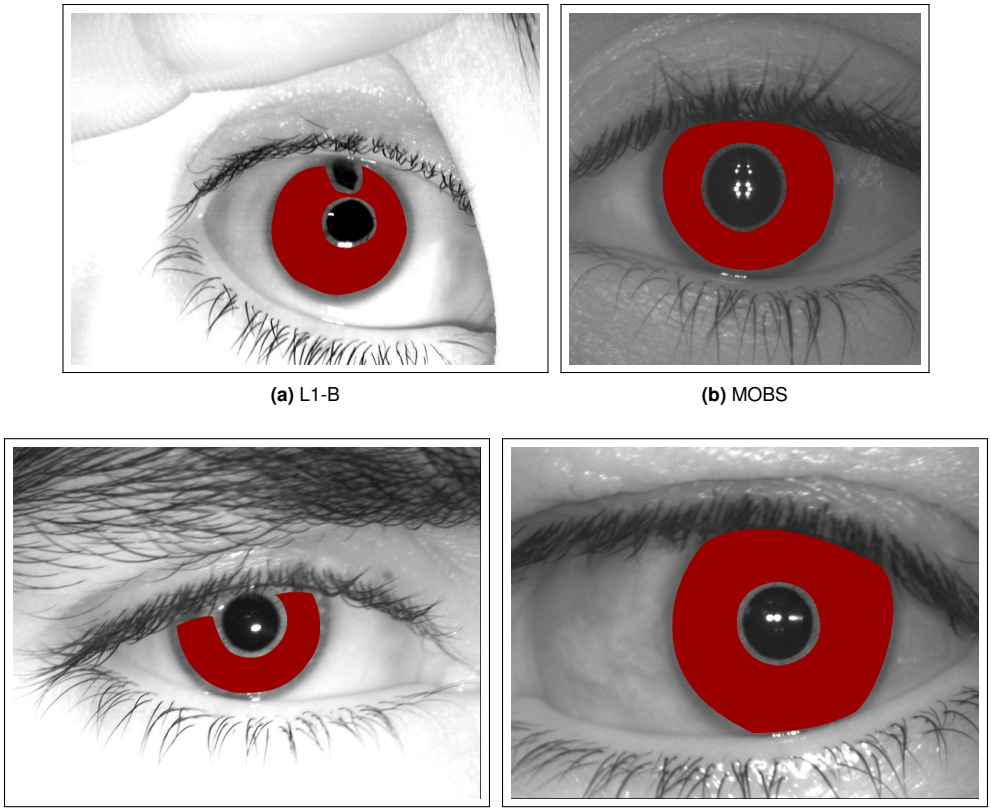
Table 2 lists the IREX IV participants along with the alpha-numeric codes assigned to each of their algorithms. Participants were allowed to submit up to 3 algorithms for each class. Four of the participants (University of Bath, iSciLab, Delta ID and Nihon Systems) are new to the IREX program while the other 8 have participated in previous IREX evaluations. For each participant, the algorithms are labeled numerically by chronological order of submission. The letter codes assigned to the participants are repeated at the bottom of each page for quick reference.

Participant	Letter Code	Class P Submissions	Class N Submissions
University of Bath	A	A00P, A01P, A02P	A00N, A01N, A02N
Neurotechnology	B	B00,P B01P, B02P	B00N, B01N, B02N
Smart Sensors	C	C00P, C01P, C02P	C00N, C01N, C02N
3M Cogent	D	D00P, D01P, D02P	D00N, D01N, D02N
IriTech	E	E00P, E01P, E02P	E00N, E01N, E02N
MorphoTrust	F	F00P, F01P, F02P	F00N, F01N, F02N
iSciLab	G	G00P, G01P, G02P	G00N, G01N, G02N
Delta ID	H	H00P, H01P	H00N, H01N
University of Cambridge	I	I00P, I01P, I02P	I00N, I01N, I02N
Iris ID	J	J00P, J01P	J00N, J01N
Morpho	K	K00P, K01P, K02P	K00N, K01N, K02N
Nihon Systems	L	L00P	L00N

Table 2: Participants of IREX IV along with their NIST-assigned letter codes and algorithm identifiers.

2.3 Image Dataset

The evaluation uses images from the Operational Set (OPS) II, which consists of approximately 7.5 million field-collected images from several commercial capture systems. The images occasionally suffer from poor sample quality (e.g. high



(a) L1-B **(b)** MOBS

(c) LSMS **(d)** COGT

Figure 2: Examples of iris images from the test dataset. The iris surface is masked with red in this report to prevent identification of the individuals. The subfigure captions refer to a four-letter software code that is an indicator of the type of camera used.

amounts of occlusion, specular reflections) that are typical of an operational system. Many were captured outside and contain heavily constricted pupils. Figure 2 shows some examples from the set. Most have a pixel resolution of 640x480, but some are 480x480. The pathological 330x330 images discussed in IREX III and its supplement are excluded from this evaluation. Some of the subjects' irides were captured by more than one camera model on different days. Further details on the images can be found in the IREX III Supplemental Report [17].

OPS-II is nearly identical to the OPS set used in IREX III with the exception that the images in the current set are free of lossy compression artifacts. The images in OPS were provided to NIST after they were compressed with JPEG [21] at a quality level of 75. Lossy compression selectively discards some of the information in an image in exchange for a more compact representation. This causes some of the pixel values to change.

2.4 Performance Metrics

2.4.1 Matching Accuracy

Open-set biometric systems are tasked with searching an individual against an enrollment database and returning zero or more candidates. A candidate is returned if the matcher determines that its dissimilarity to the searched image is below a pre-determined decision threshold. A false positive occurs when a search returns a candidate for an individual that *is not* enrolled in the database. A false negative occurs when a search *does not* return the correct candidate for an individual that *is* enrolled in the database. Raising the decision threshold increases the rate of false positives but decreases the rate of false negatives.

False positives are computed exclusively from *non-mated* searches (i.e. searches for which the searched individual is not enrolled in the database). This is more reflective of operation than if false positives had been computed from mated searches

with the correct candidates removed from the list. Similarly, false negatives are computed exclusively from mated searches.

Core matching accuracy is presented in the form of Detection Error Tradeoff (DET) plots [22], which show the trade-off between the False Positive Identification Rate (FPIR) and the False Negative Identification Rate (FNIR) as the decision threshold is adjusted. Formally, let $s_{i(j)}$ be the j-th lowest score for the i-th search against the enrollment database, and let $i = 1,...,K,K+1,...,M$ such that only the first K searches are mated. Additionally, let r_i be the lowest rank position of the correct mate for the i-th mated search. Then the statistics are

$$\text{FNIR}(\tau, r) = \quad 1 - \frac{1}{K} \sum_{i=1}^{K} \left[r_i \leq r \ \wedge \ s_{i,(r_i)} \leq \tau \right] \tag{1}$$

$$\text{FPIR}(\tau) = \quad \frac{1}{M-K+1} \sum_{i=K+1}^{M} \left[s_{i,(1)} \leq \tau \right], \tag{2}$$

where $[...]$ is the Iverson Bracket [23] which denotes 1 if the boolean expression inside the bracket evaluates to true and 0 otherwise, and $\wedge$ denotes logical-AND. Note that an additional rank constraint is imposed on mated searches. When evaluating algorithm accuracy, we did not look for the correct mate past the 20th most similar candidate (thus, $r = 20$). In practice, this has little effect on the computed FNIR values since at any reasonably selective decision threshold the correct mate will almost never satisfy the rank requirement without also satisfying the score requirement.

Equation 2 defines FPIR as the fraction of non-mated searches for which *at least one* candidate has a dissimilarity score at or below threshold. Selectivity-Reliability curves [24] compute *selectivity* as the average number of false positives returned for a non-mated search. This differs from our metric in that it takes into account the actual number of false positives returned for a particular search beyond just the first. *Selectivity* is a better metric for investigational mode applications where each candidate must be inspected by a human examiner (and thus workload scales with the number of returned candidates). That said, our definition of FPIR is grounded on the assumption that most operational uses of iris recognition result in similar outcomes regardless of whether the search returns one or several false positives. For example, one-to-many access control systems grant access to users as long as they match at least one enrolled individual.

Note that IREX III defined FPIR as the average number of false positives returned for a non-mated search divided by the length of the candidate list (i.e. the rank constraint). Since this differs from equation 2, a direct comparison of the results from the two evaluations requires computing FPIR according to a consistent definition. Section 3.7 is dedicated to comparing algorithm performance from the two evaluations.

Some DETs in this report include line segments between curves that connect points of equal threshold. The two curves might differ by enrolled population size or the number of iris samples used, and a connecting line segment shows how error rates for one curve compare to the other at the same decision threshold.

A matter of concern in any biometric evaluation is the integrity of ground truth information. Identity errors are known to exist in OPS II. To negate their impact on FPIR, we chose to horizontally flip search images prior to template generation when the searches were non-mated (see IREX III Section 6.3 for a detailed explanation and analysis). Unfortunately, this does not solve the problem where two or more persons are assigned the same identifier. Although this type of error can inflate estimates of FNIR, the IREX III Supplemental Report found it to be a rare occurrence (of 17 017 mated searches, only 28 failures were attributed to ground truth errors).

Due to the high frequency of erroneous (left/right) eye labelings in the OPS-II dataset, we chose to always enroll both eyes for an individual as separate entries and credit the algorithm with a match if either of the subject's eyes were matched. We suspect the mislabelings are due to ambiguity with respect to whether "left" is intended to represent the subject's left eye (correct) or the eye on the left from the perspective of the camera operator (incorrect).

2.4.2 Cost Model

A cost model is introduced that assigns fixed costs to each type of decision error (false positives and false negatives). Cost is computed as an average per user attempt as follows

$$\text{Cost}(\tau) = (1 - P_{\text{Mated}})\ \text{FPIR}(\tau)\ C_P + P_{\text{Mated}}\ \text{FNIR}(\tau)\ C_N\ , \tag{3}$$

where P_{Mated} is the known *a priori* probability that the user is mated, C_P is the known cost of a false positive, and C_N is the known cost of a false negative. Cost is in arbitrary units, which in practice could be units of time, workload, or money. The values of the cost parameters should be chosen to correspond to a specific application. Consider a biometric system that

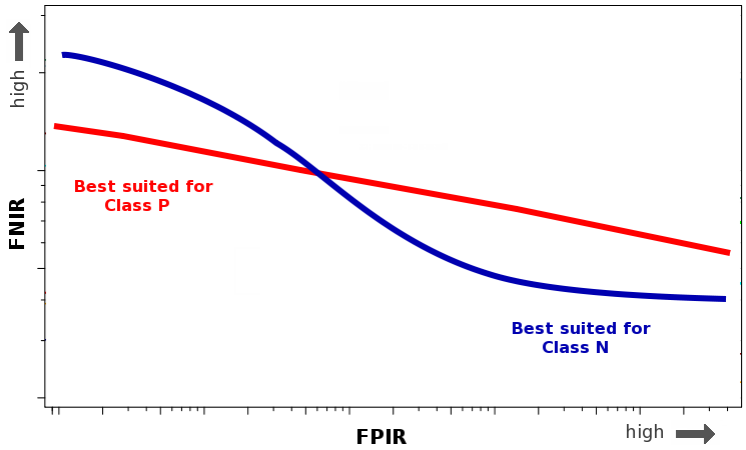


Figure 3: Notional DET plots demonstrating how the two algorithm classes place greater emphasis on different regions of the DET.

provides bank vault access to specific individuals. One might reasonably set the cost of a false positive to be the monetary value of whatever is in the vault, and the cost of a false negative to a value that reflects the amount of inconvenience incurred from having to open the vault by some other method. Setting P_{Mated} to 0.1 assumes that one out of every ten access attempts is by an allowed user.

Participants were required to submit two algorithm implementations at a time, each tasked with minimizing cost for a different set of cost parameters. The costs associated with each implementation class are shown in Table 3. Class P implementations penalize false positives heavily. Thus, suppression of false positives is paramount for this implementation class. Class N implementations penalize false negatives heavily so that suppression of false negatives is the priority.

Implementation Class	C_N	C_P	P_{mated}
Class P	1	10	0.01
Class N	200	1	0.1

Table 3: Cost parameters for each implementation class

The motivation for requiring participants to submit algorithms corresponding to two implementation classes is to see if developers can change the shape of their algorithms' DET curves by adjusting internal parameters. Figure 3 plots two notional DET curves for two possible algorithms. The two curves cross one another, making it impossible to state which is more accurate in any absolute sense. However, the algorithm corresponding to the blue curve achieves lower FNIR toward the right end of the figure, making it better suited for Class N (where comparatively greater penalty is assigned to false negatives). Conversely, the algorithm corresponding to the red curve achieves better performance toward the left end of the figure, where FPIR are rare, making it better suited for Class P.

Some DET curves in this report include a dot that marks the point of minimum cost, similar to plots in NIST Speaker Recognition Evaluations [25, 26]. Both points correspond to a decision threshold. Ideally, a biometric system will operate at the decision threshold that minimizes the cost. Selecting a poor decision threshold leads to greater cost than could otherwise be achieved.

2.4.3 Computation Time

Timing statistics are presented for primary operations (e.g. searching, template generation) as the actual time elapsed according to the C `time()` function, which has a resolution of one microsecond. Timing statistics were collected on an unloaded machine having the specifications described in Section 2.1. The alternative C function `clock()`, which measures the amount of processor time dedicated to the process, has insufficient resolution.

For ease of testing and fair comparison, algorithms were required to operate in single-threaded mode. Operationally, an algorithm can be designed to exploit multiple cores when available to expedite searching and enrollment.

2.4.4 Estimating Uncertainty

Some figures and tables convey information about the uncertainty associated with a statistic in the form of confidence intervals or estimates of standard deviation. These estimates are intended to capture random variation in the observed value if one assumes repeated *iid* sampling from the same population. They are *not* intended to reflect how the statistic might change over different test data or even different sampling strategies over the same data.

When our statistic of interest is the FNIR and a fixed operating threshold is assumed, each mated search result is a Bernoulli trial with unknown probability of success equal to the probability of making a correct identification. Let p be the probability of *failing* to make a correct identification so that FNIR$= p$. For a fixed number of n independent searches, the number of failed identifications is binomially distributed. The standard error of $\hat{p}$ is therefore

$$s_e = \sqrt{\frac{\hat{p}(1-\hat{p})}{(n-1)}}\,. \tag{4}$$

Boxplots that summarize $\hat{p}$'s distribution require estimates of its 25th, 50th, and 75th percentiles. The simplest formula for estimating a quantile for $\hat{p}$ involves approximating its distribution with a normal distribution and using equation 4 to fit the unknown standard deviation. However, Brown et. al [27] identify several problems with this approach stemming from the fact that (1) $\hat{p}$ is a discretized estimator, and (2) $\hat{p}$ tends to have a skewed distribution at values close to 0 or 1. Therefore, we adopt their recommendation to use the alternative Wilson Score method

$$\hat{F}(q) = \frac{\hat{p} + z_q^2 + z_q\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_q^2}{4n^2}}}{1 + \frac{1}{n}z_q^2}\,, \tag{5}$$

where z_q is the qth quantile of the standard normal distribution. The Wilson Score method still loses accuracy when np or $n(p-1)$ is small (say, less than 10), but the sample sizes used in this evaluation are always large enough to satisfy this condition.

We make several simplifying assumptions when applying the above methods of uncertainty estimation to the biometric identification problem. Most notably, separate searches against the same enrollment database are treated as independent samples, yet we know positive correlations are present due to the existence of Doddington's Zoo [28]. Ideally, each search would be performed against an entirely re-sampled set of enrollment samples, but we lack the data and doing so would be too computationally expensive.

We also report estimates of the variability of FNIR at a fixed FPIR when in fact the decision threshold is fixed. Uncertainty with respect to what decision threshold corresponds to the targeted FPIR results in increased uncertainty about the true value of FNIR. Chu et. al [29] propose a two-sample bootstrap approach to estimating the probability distribution of FNIR at a fixed FPIR. However, we chose to exploit the parametric nature of the problem. We also assume that, given the large number of non-mated searches we perform, our estimates of FPIR are so tightly bounded that they should have little, if any, impact on the estimates.

3 Results

3.1 Accuracy

3.1.1 Positive Identification

Positive identification systems verify the implicit claim that the user is enrolled in the system and often return the user's identity if found. Most positive identification systems grant special access or privileges to enrolled users. For example, the Pentagon Force Protection Agency plans to use iris and fingerprint scanners to control access to the Pentagon Building and Mark Center [30, 31]. Several state and local law enforcement agencies have also expressed interest in using iris recognition for identity management at their prison and jail facilities [32, 33, 34, 35]. Proposals include systems that would control access to sensitive areas, track the intake and release of inmates (such as those participating in work-release programs), and authenticate visitors, witnesses, and lawyers.

Recognition accuracy refers to how often the algorithm correctly identifies an individual or correctly rejects a claim of enrollment. False positives have the potential to grant special privileges or access to unauthorized users. False negatives deny those privileges to authorized users. In a positive identification system, a false negative occurs when either 1) the matcher fails to generate a matchable template from the user's biometric sample, or 2) the matcher fails to correctly identify the user after searching the template against the enrollment database. In our DET curves, failed attempts to enroll users also lead to false negatives since a user that is not enrolled can't possibly be identified during a search. However, operational treatment of failed enrollment attempts is likely to be different. If the enrollment is being attempted in real-time with a cooperative user, the user could be prompted to provide a new sample, or different accommodations could be made (e.g. using fingerprints instead of iris scans).

The costs associated with identification errors are highly scenario dependent and are often difficult to quantify. A false positive could result in free access to a theme park, or more seriously, unauthorized access to sensitive information. Many systems use multi-factor authentication (e.g. a password in addition to biometric authentication), so a false positive from the biometric matching step alone may not be enough to compromise the system.

3.1.1.1 Small Population

Applications

Some iris recognition systems may have only a few thousand users enrolled (e.g. those granting building or gym access). These systems may be cheaper to operate and more convenient for users if run in an *identification*, rather than *verification*, mode. The identification task separates itself from verification in that it does not require the user to provide a claim to identity. Thus, the user would not have to enter a pin number or carry around a smart card to use the system. The self-service iris kiosks used by UK IRIS [36] and NEXUS [15] operate in this way.

Iris recognition is particularly robust to relaxation of the identity claim requirement because of 1) its ability to rapidly perform searches against the entire enrolled population, and 2) the fact that false negatives are often the result of poor quality captures that would also fail in a verification system. This is reflected in the nearly-flat (i.e. low-slope) appearance of iris DET curves which has been noted in previous reports and evaluations [37, 38].

Results

Figure 4 presents DET accuracy for Class P (i.e. Positive ID) matching algorithms when searches are performed against an enrolled population size of 10 000. For simplicity, only the last (and usually most accurate) Class P submission from each participant is included in the figure. DET curves for all Class P submissions are presented in Appendix A. Figure 5 shows boxplots summarizing the uncertainty of FNIR at an FPIR of 10^{-3}. At an FPIR of 10^{-4}, the most accurate matcher (F02P) achieves an FNIR of 0.017, although the difference may not be statistically significant compared to the next most accurate algorithm (D02P) at the same FPIR.

The notable observations are:

- **Flatness:** The algorithms have lightly sloping DET curves such that the FNIR increases only slightly as the FPIR decreases (i.e. as one goes from right to left on the page). For most algorithms, false negatives that occur at discriminating (low) thresholds also occur at much more relaxed (higher) thresholds. For I02P, the FNIR rises from 0.030 to 0.043 (an increase of 44 %) as the FPIR decreases from 10^{-1} to 10^{-4}. For F02P, the FNIR rises from 0.010

to 0.017 (an increase of 71 %) over the same FPIR range.

- **Overlap/Crossing:** Although the DET curves occasionally intersect one another, they tend to have similar slopes (on a log scale). An exception is G02P, which has a slightly steeper slope such that it performs comparatively better at higher FPIR.
- **Improvements with Later Submissions:** Participants often achieved noticeable improvements in accuracy with later submissions (see also Figure 29). Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.

Operational Relevance

The lightly sloping DET curves in Figure 4 suggest that many of the false negatives would also fail for a comparable verification system. The IREX III Supplemental Report found that the majority of false rejections that occurred for the most accurate matchers in IREX III were due to poor presentation or poor fidelity rather than poor character. Character refers to the quality of the inherent features of the iris while presentation and fidelity refer to how accurately the iris features are represented in the sample [39]. In a live-capture environment, poor presentation can often be corrected through simple re-acquisition, but poor character is more problematic. The UIDAI is collecting iris samples along with fingerprints in part because residents' fingerprints can become heavily degraded from manual labor.

The accuracy numbers in an offline technology evaluation such as this may not be reflective of a system that includes a live-capture component. Human behavior is likely to be different, and failed identifications can often be corrected by re-acquisition of a new sample or by supplementing with additional information (e.g. multi-modal or multi-factor authentication).

When the DET curves for two algorithms cross one another, it becomes difficult to say that one is more accurate in any absolute sense. The best choice for a particular application will depend on the selected operating threshold in addition to other factors.

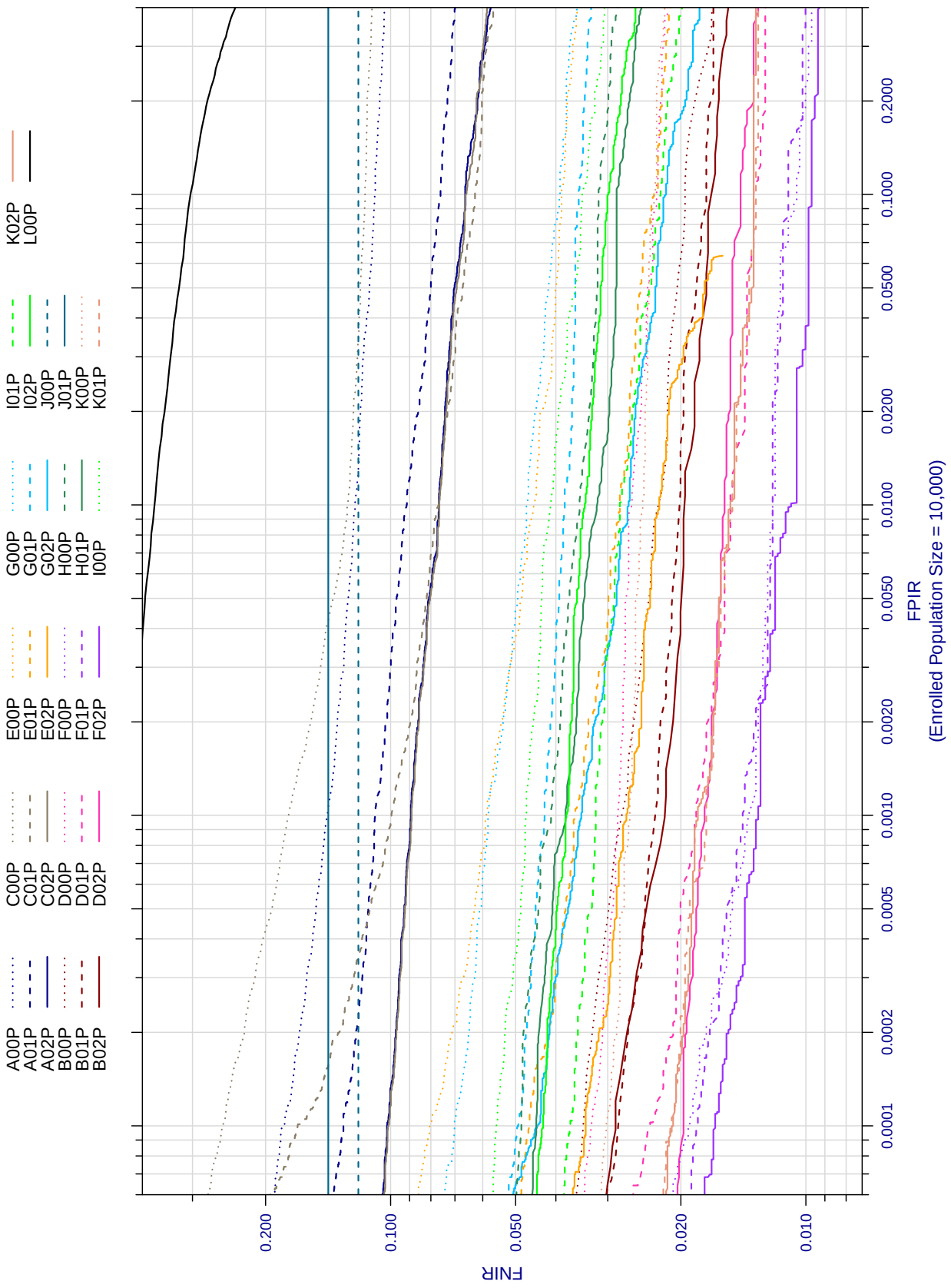


Figure 4: DET curves for single-eye searches against an enrolled population of 10 thousand. Only the final Class P submission from each participant is plotted. Plots were generated using 5 988 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

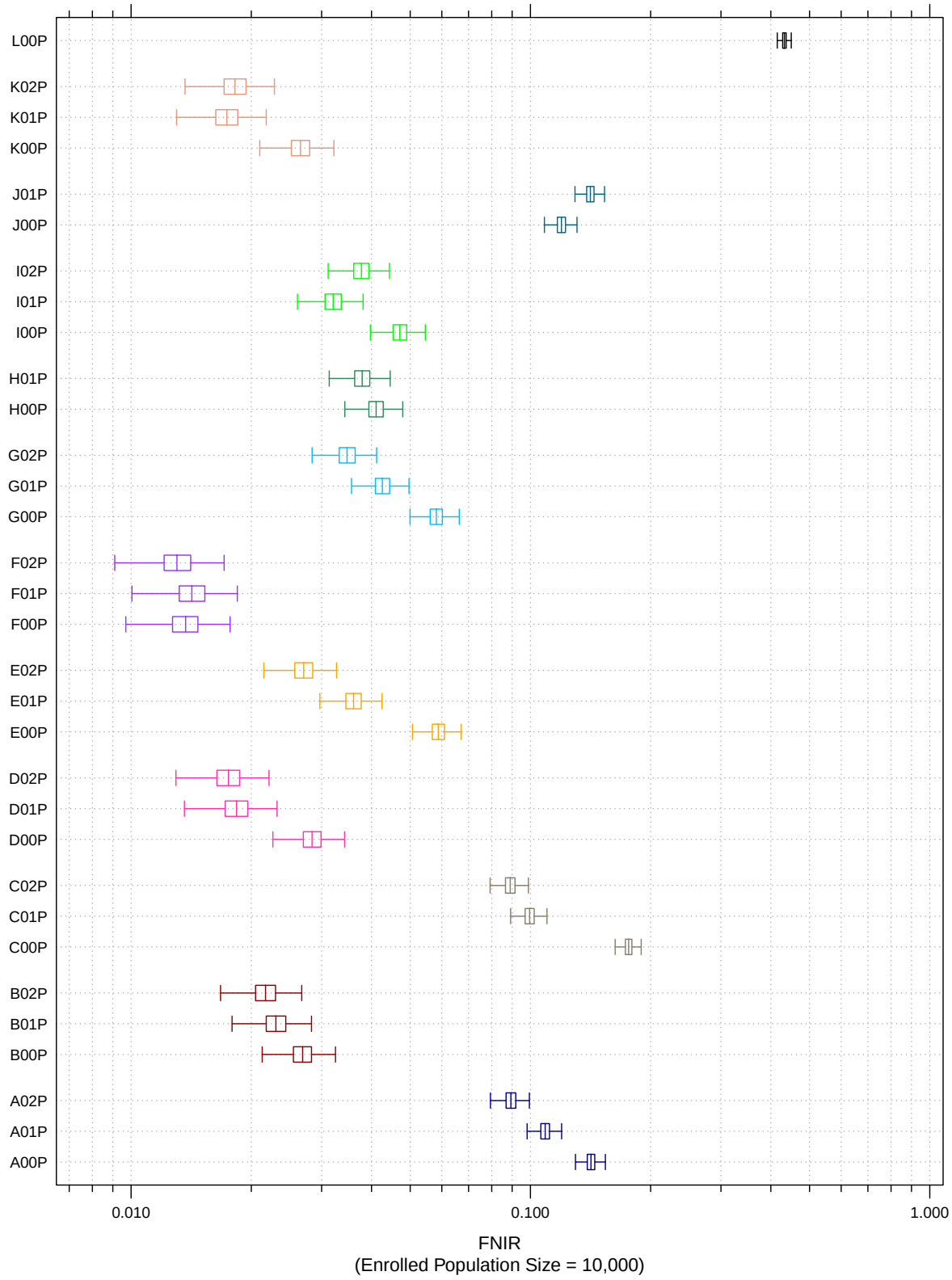


Figure 5: Boxplots summarizing the probability distribution of FNIR at an FPIR of 10^{-3} for each algorithm when single-eye searches are performed against an enrolled population of 10 thousand. Plots were generated using 5 988 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.1.1.2 Large Population

Applications

The IREX III Evaluation verified theoretical claims that iris recognition has the capability to distinguish an individual from among millions. The DOD has collected iris images of several million people in various theaters of operation. The system is used in military operations and base access control systems. India's UID program contains iris images for more than 320 million Indian citizens. The biometric is also used in border management applications including UAE expellee [9] and NEXUS system on the Canadian border [15].

Results

Figure 6 presents DET accuracy for Class P (i.e. Positive ID) matching algorithms when searches are performed against an enrolled population size of 1.6 million. Only the last (and usually most accurate) Class P submission from each participant is included in the figure. DET curves for all Class P submissions are included in Appendix A. Figure 7 shows boxplots summarizing the uncertainty of FNIR at an FPIR of 10^{-3}. At an FPIR of 10^{-4} the most accurate matcher (F02P) achieves an FNIR of 0.020.

Figure 8 shows how error rates change when the enrolled population is increased from 10 thousand to 1.6 million. Gray line segments connect points corresponding to the same operating threshold. For most algorithms, the increase in FNIR at a fixed FPIR is small (< 30 %) despite a 160-fold increase in the enrolled population size. Lines of equal threshold also indicate that at a fixed operating threshold, FPIR usually increases when the size of the enrolled population is increased.

The notable observations are:

- **Flatness:** The curves are lightly sloping, but steeper than for the smaller enrollment database (compare to Figure 6). This is likely due to the more difficult problem of identifying an individual within a greater population. Still, many false negatives that occur at more discriminating (low) thresholds also occurred at more relaxed (higher) thresholds. For B02P, the FNIR rises from 0.020 to 0.036 (an increase of 82 %) as the FPIR decreases from 10^{-1} to 10^{-4}. For F02P, the FNIR rises from 0.012 to 0.020 (an increase of 71 %) over the same FPIR range.
- **Overlap/Crossing:** Although the DET curves occasionally intersect, they tend to have similar slopes (on a log scale). However, G02P and E02P do not maintain a constant slope with the result that they perform comparatively better at higher FPIR.
- **Comparison to Small Enrolled Population:** Increasing the enrolled population size increases FPIR but has a negligible impact on FNIR for most algorithms. However, some algorithms (D02P, F02P, K02P) appear to perform score-normalization such that increasing the enrolled population size increases FNIR but has less of an impact on FPIR.
- **Improvements with Later Submissions:** Participants often achieved noticeable improvements in accuracy with later submissions (see also Figure 31). Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.

The most accurate algorithms are modestly better than those from the IREX III evaluation conducted two years ago (not shown). However, many participants have improved their algorithms such that the disparity in accuracy across the participants has reduced significantly. Section 3.7 compares results from this evaluation to IREX III.

Operational Relevance

The DET curves are lightly sloping compared to other biometric modalities, suggesting iris recognition is particularly advantageous at more discriminating (lower) operating thresholds where false positives are rare. Iris recognition accuracy also appears much less dependent on the enrolled population size than other biometric modalities. Thus, the enrolled population can be expanded without fear of ballooning error rates. However, a threshold adjustment may still be necessary.

The disparity in accuracy between the best and worst matchers has reduced since IREX III, making other performance factors (cost, speed, template size, etc.) more relevant when comparing matchers.

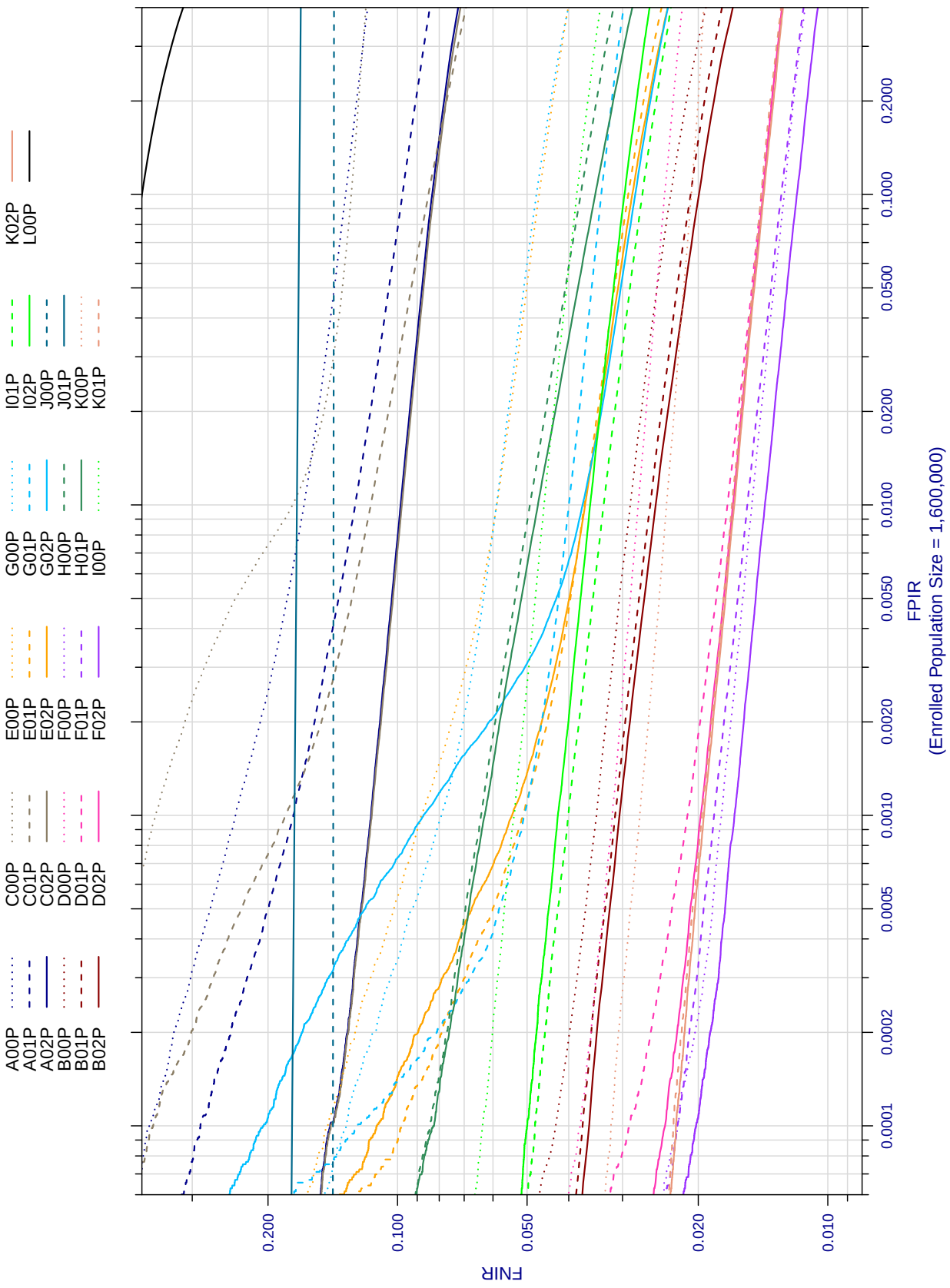


Figure 6: DET curves for single-eye searches against an enrolled population of 1.6 million. Only the final Class P submission from each participant is plotted. Plots were generated using 962 720 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

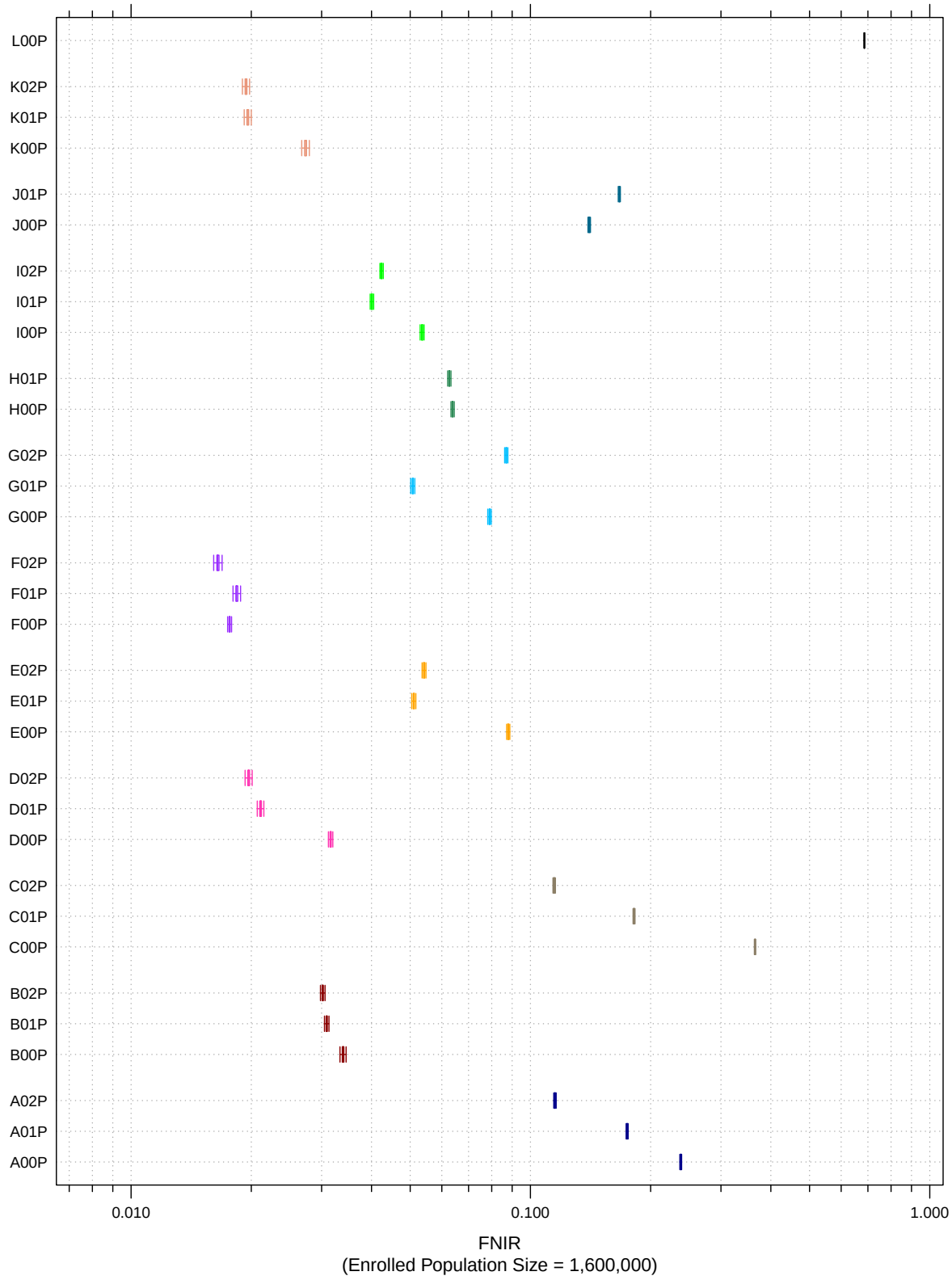


Figure 7: Boxplots summarizing the probability distribution of FNIR at an FPIR of 10^{-3} for each algorithm when single-eye searches are performed against an enrolled population of 10 thousand. Plots were generated using 962 720 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

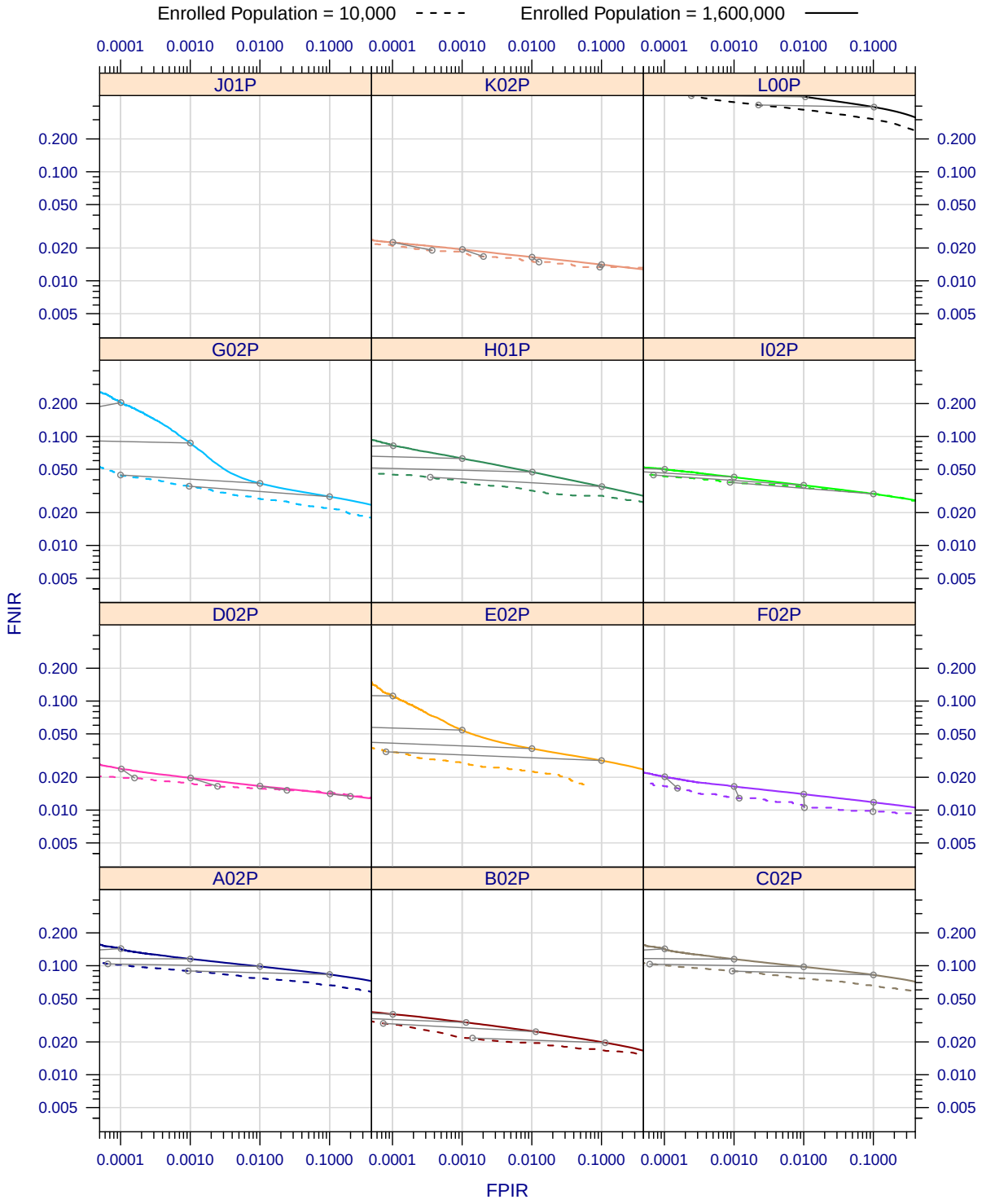


Figure 8: DET curves for single-eye searches against an enrolled population of 10 thousand (dotted line) and 1.6 million (solid line). Grey line segments connect points of equal threshold between curves.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.1.2 Negative Identification

A negative identification system searches a biometric sample against an enrolled population and may return one or more identities. Unlike positive identification systems, the implicit assumption is that the searched individual *is not* enrolled. Rather than granting special privileges to enrolled individuals, negative identification systems often impose restrictions on them. An example would be a system that uses biometric information to screen certain persons (e.g. terrorists, known criminals) from entering a country.

One of the largest iris systems operating in a negative identification mode is the UAE's border-crossing control system. The UAE's Ministry of Interior maintains a database of biometric data on approximately 440 000 individuals (as of 2004) expelled from the country for various violations, such as overstaying their work permits [40]. Most foreign visitors are checked against this database, and to date, more than 73 000 illicit attempts at re-entry have been caught. The Department of Homeland Security (DHS) could also use iris recognition to identify individuals attempting to enter the country illegally through Mexico. Illegal immigrants often have highly degraded fingers that may make iris scans a more suitable alternative to fingerprints.

Recognition accuracy refers to how often the algorithm correctly identifies an enrolled individual or correctly accepts that the user is not enrolled in the database. A false negative occurs when the algorithm fails to identify an enrolled individual. A false positive occurs when the searched person is falsely matched to at least one individual in the enrollment database. Additionally, we treat a Failure to Extract (FTX) iris feature information from either the search or mated enrollment sample(s) as a false positive. Most biometric evaluations treat template generation failures as false negatives, but we seek to model the situation where individuals who cannot be identified from their biometrics are subjected to scrutiny comparable to those who return "hits" against the enrollment database. From the perspective of game theory [41] it is certainly reasonable to assume that users who do not want to be identified might intentionally present a poor or improper sample (although we did not specifically look for this in our test set).

Many biometric systems do not treat feature extraction failures as false positives. Those with non-cooperative users (e.g. a camera performing face recognition at a choke-point at a railway station [42]) are often designed to operate transparently and cannot stop every user who does not present a high-quality image of their face to the camera. However, iris images are more difficult to acquire, and for this reason they usually require cooperation from users. If the system provides responses in real time, a user could be prompted to provide a new sample if the recognition software could not extract features from the first sample. However, some problems pertaining to poor sample quality cannot be corrected by acquiring a new sample (e.g. congenital defects such as aniridia or excessive pupil dilation due to opiate consumption), and some systems do not always attempt to extract features immediately after the image is acquired.

The costs associated with identification errors are scenario dependent and often difficult to quantify. A false negative can result in a failure to properly screen a suspected terrorist, criminal, or illegal immigrant at a port of entry. If it occurs during enrollment, it can lead to a duplicate entry for the person in the database, which is a common way of committing identity fraud. False positives inconvenience users when it requires them to undergo more rigorous screening processes. False positives also inconvenience system operators who must dedicate time and money to detecting and rectifying the errors.

3.1.2.1 Small Population

Applications

Some watch lists may only contain biometric information on a few hundred or thousand individuals. Even iris systems that normally operate in a positive identification mode (such as the ones described in Subsection 3.1.1.1) might still perform checks for duplicate entries during enrollment or periodic database consolidation operations.

Results and Notable Observations

Figure 9 presents DET accuracy for Class N (i.e. Negative ID) matching algorithms when searches are performed against an enrolled population size of 10 000. For simplicity, only the last (and usually most accurate) Class N submission from each participant is included in the figure. DET curves for all Class N submissions are presented in Appendix A. Some DET curves approach a vertical asymptote as the FPIR nears the algorithm's FTX rate (i.e. the fraction of searches for which the algorithm was unable to extract matchable feature information from either the search or corresponding enrollment sample). This occurs because an algorithm can never achieve an FPIR below its FTX rate when feature extraction failures are counted as false positives.

Participants were instructed to submit algorithms that minimize a cost equation that penalizes false negatives 50 times more severely than false positives. Since the cost equation is biased heavily in favor of reducing false negatives, the decision

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

threshold that achieves the lowest cost is always closer to the right side of the figure.

The notable behaviors are:

- **Flatness:** With the exception of the asymptotic behavior of some algorithms, the DET curves have similar slopes to those presented in Subsection 3.1.1.1 for a comparable positive recognition system. For F02N, the FNIR rises from 0.010 to 0.017 (an increase of 71 %) as the FPIR decreases from 10^{-1} to 10^{-4}. For H01N, the FNIR rises from 0.029 to 0.045 (an increase of 57%) over the same FPIR range.
- **Vertical asymptotes:** Algorithms H01N, I02N, and K02N report failed feature extraction attempts. Although this introduces a conspicuous vertical asymptote at each algorithm's FTX rate, it does not detrimentally affect the algorithm's point of minimum cost. Algorithms should not be penalized for properly reporting failed feature extraction attempts.
- **Minimum Cost:** The more accurate algorithms achieve their minimum cost at lower FNIR and usually lower FPIR as well. The most accurate algorithm (F02N) achieves the lowest cost among all algorithms at an FNIR of 0.011 and an FPIR of 0.010. Algorithm D02N achieves its minimum cost at a similar FPIR but at an FNIR of 0.016.
- **Improvements with Later Submissions:** Participants often achieved noticeable improvements in accuracy with later submissions (see Figure 34). Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.

Operational Relevance

Algorithms with curves that do not approach a vertical asymptote are not reporting their feature extraction failures. Failing to report these failures is undesirable in an operational system since it makes little sense to search or enroll a template that does not contain matchable feature information. However, DET curves presented in technology evaluations such as this one traditionally do not distinguish between matching errors and feature extraction failures, so algorithm developers may see little downside to searching or enrolling a template even if it is unlikely to match well.

Matching algorithms cannot achieve FPIR below their FTX rate, and the FNIR begins to trend upward sharply at FPIR less than twice the algorithm's FTX rate.

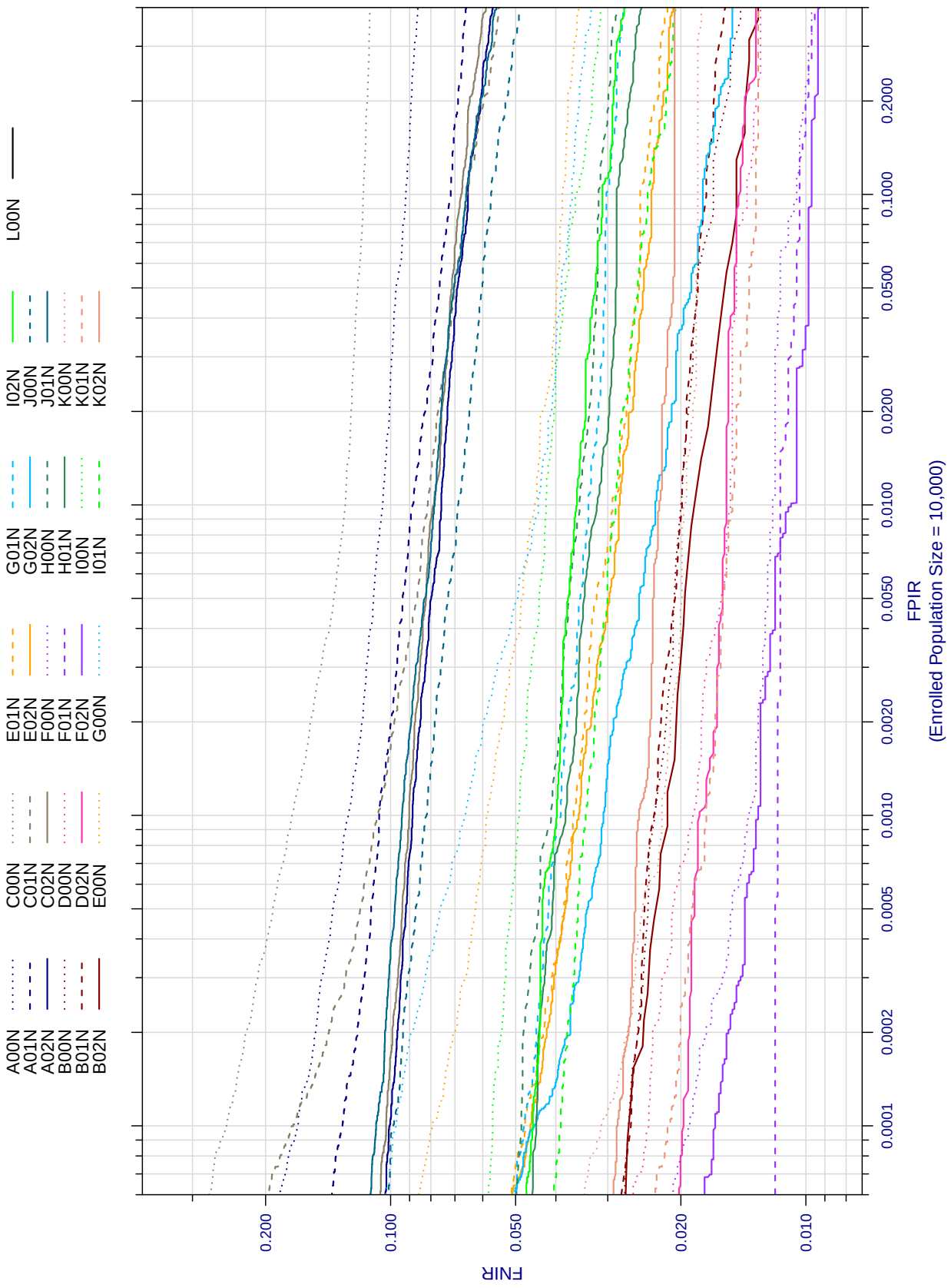


Figure 9: DET curves for single-eye searches against an enrolled population of 10 thousand. Only the final Class N submission from each participant is plotted. Plots were generated using 5 988 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.1.2.2 Large Population

Applications

The iris is currently being used for de-duplication of the Indian government's UID database. At present, the database contains biometric information on about 225 million residents, but that number is projected to grow to about 600 million by the end of 2014 [7]. The iris is also used by Biometric Identity Management Agency (BIMA) to screen law enforcement personnel at US facilities [43]. Iris recognition also has potential for improving census management, which in turn could facilitate better elections in places like Iraq and Afghanistan [44].

Results and Notable Observations

Figure 10 presents DET accuracy for Class N (i.e. Negative ID) matching algorithms when searches are performed against an enrolled population of 1.6 million. For simplicity, only the last (and usually most accurate) Class N submission from each participant is included in the figure. DET curves for all Class N submissions are presented in Appendix A.

Participants were instructed to submit algorithms that minimize a cost equation that penalizes false positives 50 times more than false negatives. Thus, the operating threshold that achieves the lowest cost according to this equation tends to be closer to the right side of the figure, where false positives are higher but false negatives lower.

The notable behaviors are as follows:

- **Flatness:** With the exception of the vertical asymptotes for some algorithms, the DET curves have similar slopes to those presented in subsection 3.1.1.2 for a comparable positive recognition system. For F02N, the FNIR rises from 0.012 to 0.020 (an increase of 73 %) as the FPIR decreases from 10^{-1} to 10^{-4}. A02N rises from 0.083 to 0.137 (an increase of 66 %) over the same FPIR range.
- **Vertical asymptotes:** Algorithms H01N, I02N, and K02N report failed feature extraction attempts. Although this introduces a conspicuous vertical asymptote at each algorithm's FTX rate, it does not detrimentally affect the algorithm's point of minimum cost. Algorithms should not be penalized for properly reporting failed feature extraction attempts.
- **Minimum Cost:** The more accurate algorithms achieve their minimum costs at lower FPIR. The most accurate algorithm, F02N, achieved a minimum cost at an FPIR of 0.018 while algorithm C02N achieved its minimum cost at an FPIR of 0.165.
- **Improvements with Later Submissions:** Participants often achieved noticeable improvements in accuracy with later submissions (see Figure 36). Curves translated downward in relation to previous submissions may indicate improvements in the feature extraction process. Changes in slope or shape may indicate alterations to the comparison strategy.

Operational Relevance

Iris recognition algorithms are robust to increases in the enrolled population size when operating in a negative identification mode. Cost typically increased by 30 % or less when the enrollment population was expanded from 10 thousand to 1.6 million. Additionally, the point of minimum cost was typically achieved at a lower threshold (and lower FPIR) when the enrollment population size was expanded. Iris recognition searches are also fast compared to other biometric modalities, making it particularly well suited for use with large enrolled populations.

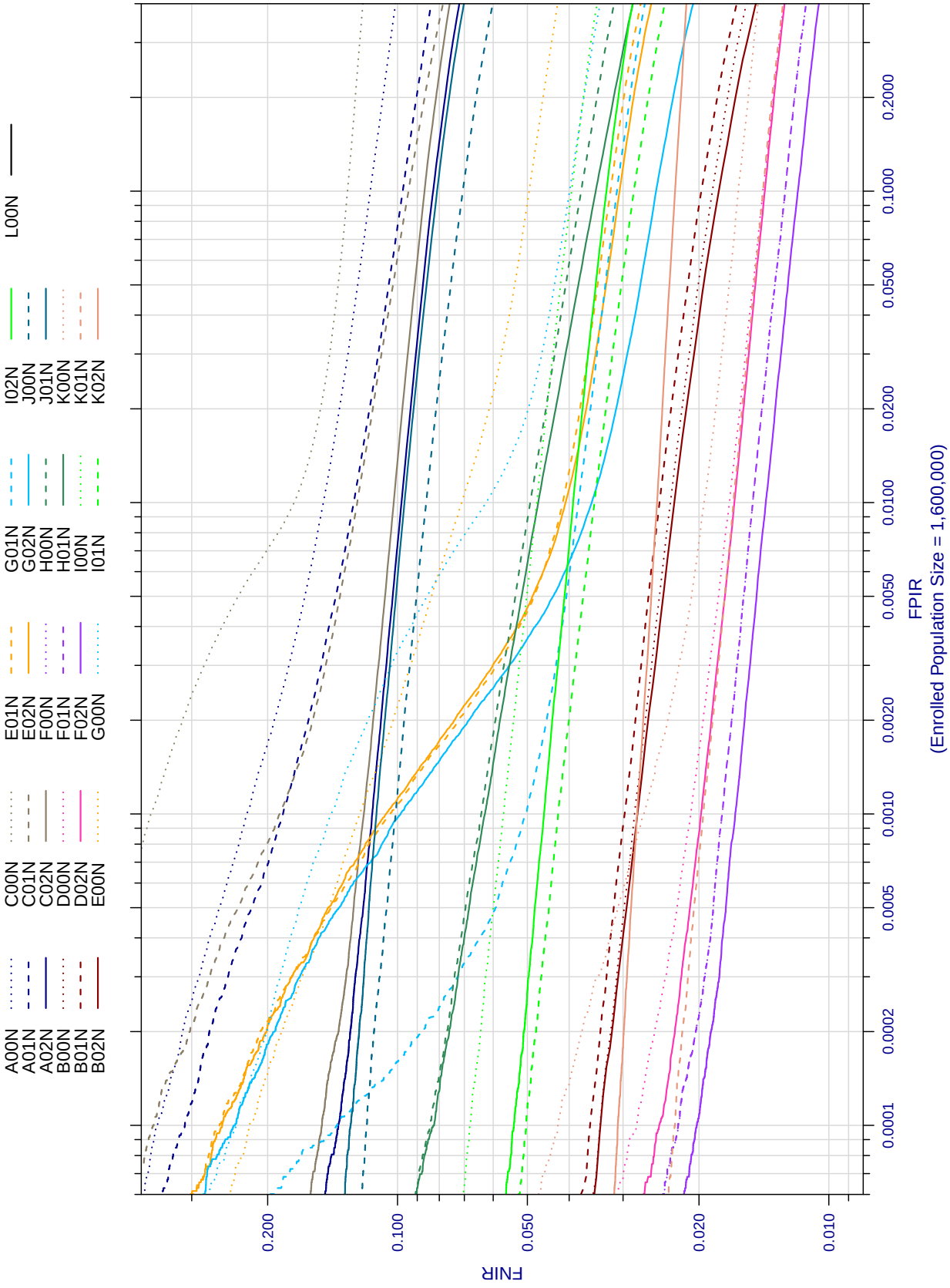


Figure 10: DET curves for single-eye searches against an enrolled population of 1.6 million. Only the final Class N submission from each participant is plotted. Plots were generated using 962 720 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.2 Speed

Speed is an important performance factor for iris recognition algorithms. The time it takes for an algorithm to generate a template and search it against a database can affect throughput rates at physical access control points, biometric scanning stations, service kiosks, etc. Iris recognition algorithms are capable of rapidly identifying users against large databases, which has lead several counties to adopt iris-based methods of authenticating travelers at airports and border crossings [36, 45]. Frequent travelers may enroll in these systems and enjoy expedited screening at security checkpoints. The United States and Canada jointly operate one such system as part of the NEXUS Program, which allows registered members to use self-service iris kiosks for expedited immigration clearance. Even when matching is performed offline, speed estimates factor into resource requirements for computing facilities.

In a typical live-capture system, a biometric sample is acquired and then transferred over a network to a central facility. Once at the facility, a template is generated from the sample and searched against the enrolled population. Finally, the results of the search are used to send a response back over the network. Turnaround time is affected by the time it takes to both generate a template and search it against the enrolled population. However, throughput rates may not be affected if the steps can be performed concurrently with other tasks. For example, a CBP officer could manually inspect a visitor's credentials while waiting for a response.

IREX III found that sometimes a speed-accuracy trade-off exists, where improved accuracy can be achieved through slower, but more involved, matching strategies.

3.2.1 Template Generation Time

Figures 11 and 12 present the distribution of template generation times for each algorithm. Template generation time refers to the amount of time that elapsed while a matchable template was generated from a raw iris image. It does not include any pre-processing steps performed by the testing harness such as loading the image from disk. The timing machine was a PC-class blade with a CPU running at 2.3 GHz. Further details on the testing environment are provided in Section 2.1.

Template generation time was not noticeably different for different enrolled population sizes, so figures report generation times when the algorithm expects an enrolled population of 10 thousand. For the same reason, template generation time is only reported for search templates and not enrollment templates.

The notable results and observations are:

- Template generation time varies considerably from one participant to another. Participant J's Class P algorithms usually generate a template in less than a twentieth of a second while algorithm C01P takes about a 0.8 seconds on average. Most algorithms generate templates in under half a second.
- Some algorithms appear to have a fixed template generation time (e.g. I02P, J02P) while others require different amounts of time depending on the image (e.g. C01P, A02P, G02P). The latter behavior is likely if the algorithm scans for the iris boundary over a range of parameters (e.g. radius, location) and stops once a boundary is found, but other explanations are possible.
- No clear speed-accuracy tradeoff exists. Participant F's algorithms are among the most accurate and are also among the fastest to generate templates. The most accurate algorithm for any given participant is also rarely the slowest one to generate templates.
- Class P and N algorithms tend to have similar template generation times. One exception is K00P and K00N, where the latter takes about 3 times longer to generate templates, although accuracy is comparable between the algorithms.

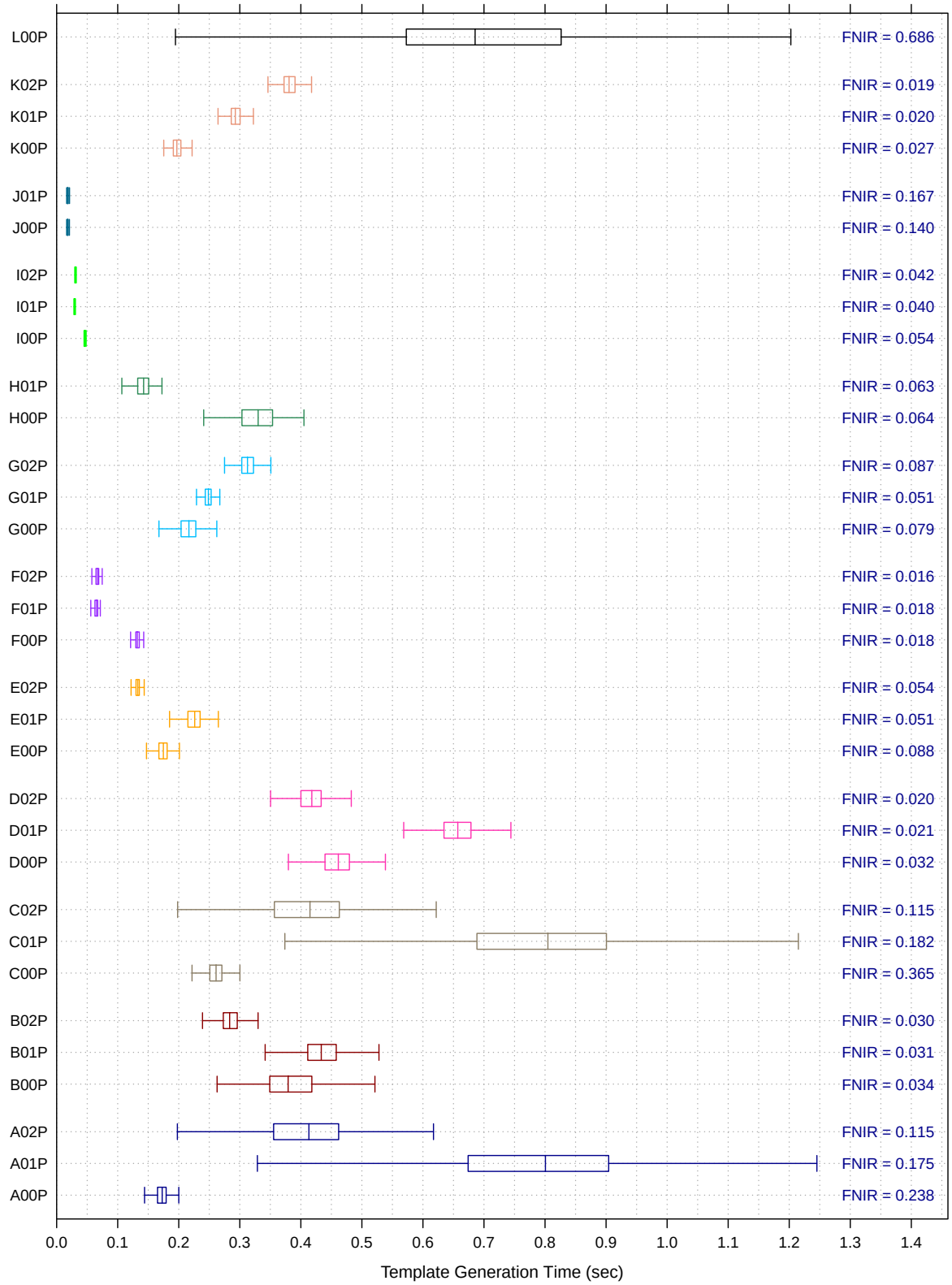


Figure 11: Boxplots summarizing the distribution of template generation times for Class P algorithms. For reference, FNIR at an FPIR of 10^{-3} against an enrolled population of 1.6 million is reported on the right. Boxplots were generated from 5 000 generated templates.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

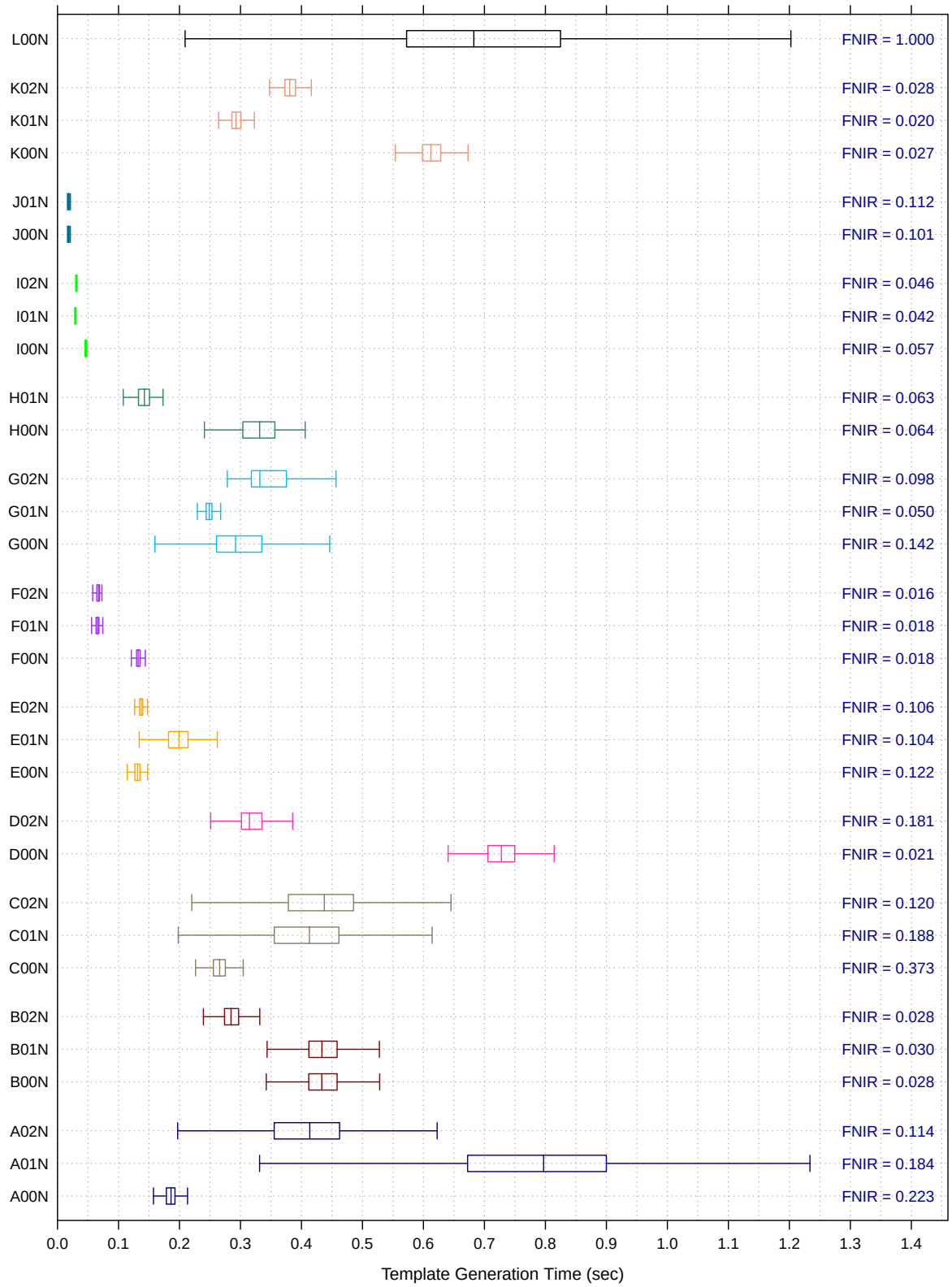


Figure 12: Boxplots summarizing the distribution of template generation times for Class N algorithms. For reference, FNIR at an FPIR of 10^{-3} against an enrolled population of 1.6 million (assuming positive identification) is reported on the right. Boxplots were generated using 5 000 generated templates.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.2.2 Search Time

Search time refers to the amount of time elapsed when a generated template is searched against an existing enrollment database. Timing statistics were collected on 5 000 mated searches using a single processing core. Machines that have multiple cores can perform concurrent processing to speed searches. IREX III found that using 16 cores simultaneously results in an 8 to 16 fold improvement in search time (for one matching algorithm that could operate in both single-threaded and multi-threaded mode). Figures 13 and 14 show search times for each algorithm when the enrolled population is 1.6 million. Figure 15 compares search times at different enrolled population sizes.

Results and Notable Observations

- For any given algorithm, search time almost always varies to some degree from one search to the next. Variation was greatest for algorithms submitted by participants G and J where some searches took more than 10 times longer than others. Search time may depend on the amount of feature information that could be extracted from the search image and used for comparison.
- Algorithms B01P, B02P, C01P, J01P, and K00P take less than a second to search against an enrolled population of 1.6 million for more than half their searches. Algorithms I02P and K02P take less than 2 seconds for more than half their searches.
- The most accurate algorithms, F02P and F02N, are also among the fastest. B02P is the fastest algorithm and is also among the more accurate matchers.
- Search time is not strongly correlated to feature extraction time when comparing algorithms. B01P, B02P and K00P are among the fastest algorithms to search, but their template generation times are not among the fastest.
- Search time increases linearly with the size of the enrollment database for all algorithms. A ten fold increase in the size of the enrollment database (from 160 thousand to 1.6 million) always leads to about a ten fold increase in search time.
- Search times were not appreciably different for non-mated searches (not shown). An algorithm that short-circuits the search operation after encountering a sample of sufficient similarity would be expected to complete mated searches more quickly, depending on the position of the mate in the enrollment database.

3.2.3 Operational Relevance

The relevant factor is turnaround time (i.e. the speed at which a search response can be returned after an iris sample is acquired). Short turnaround times are critical for maintaining high throughput rates at scanning stations or for rapid identification in the field. Some of the algorithms are capable of generating a template and matching it against an enrolled population of 1.6 million in under 4 seconds with just one processing core. If multiple cores were employed, this time could probably be reduced to under a second. A transaction in a centralized system involves several steps (presentation of the iris, image acquisition, network transfer, etc.) which together are likely to take longer than the matching step alone. Sometimes matching can be performed concurrently with other tasks (e.g. an operator can manually inspect credentials while waiting for the search to complete) such that matching incurs little if any additional time to the transaction.

No prominent speed-accuracy trade-off was present, and search time was always linearly proportional to the size of the enrolled population (i.e. a doubling of the enrolled population size doubles the search time).

The focus of this evaluation was not speed, although participants were informed that speed was a relevant performance factor before they submitted their algorithms. The fastest algorithm was capable of searching against an enrolled population of 1.6 million and returning a result in under half a second.

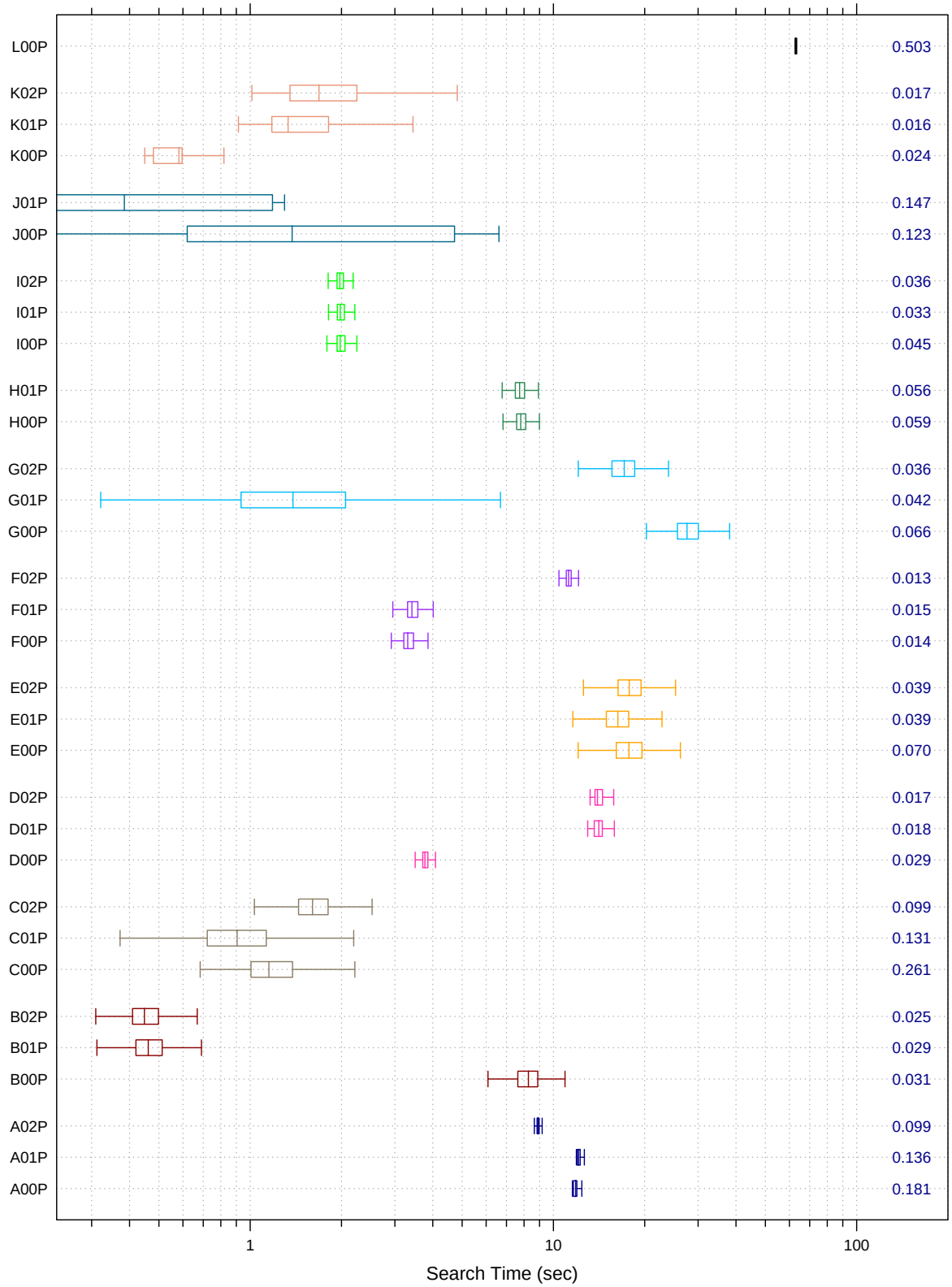


Figure 13: Distribution of search times for each Class P algorithm when 5 000 single-eye mated templates are searched against an enrolled population of 1.6 million. For reference, FNIR at an FPIR of 10^{-3} is presented on the right.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

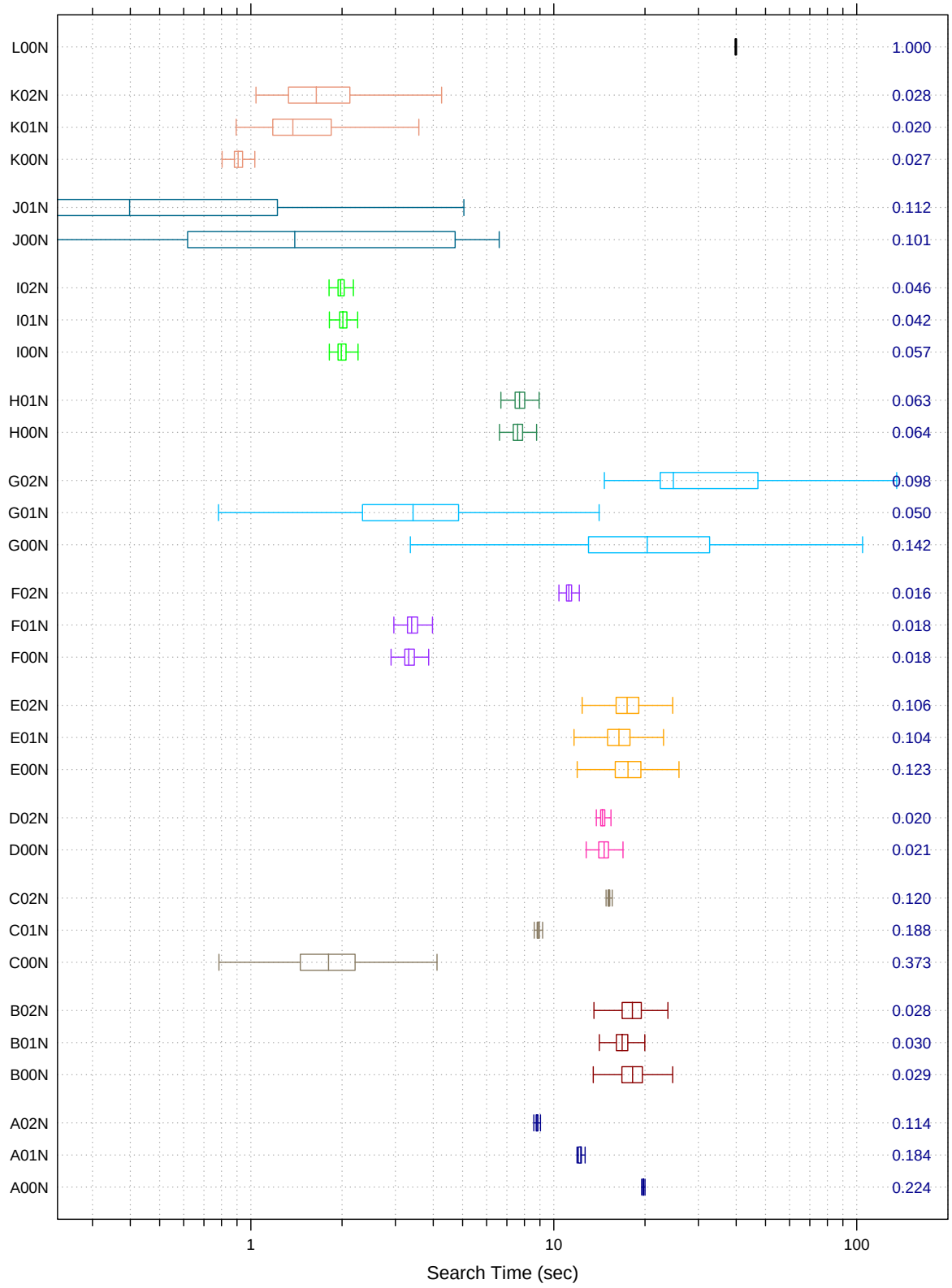


Figure 14: Distribution of search times for each Class N algorithm when 5 000 single-eye mated templates are searched against an enrolled population of 1.6 million. For reference, FNIR at an FPIR of 10^{-3} (assuming positive identification) is presented on the right.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

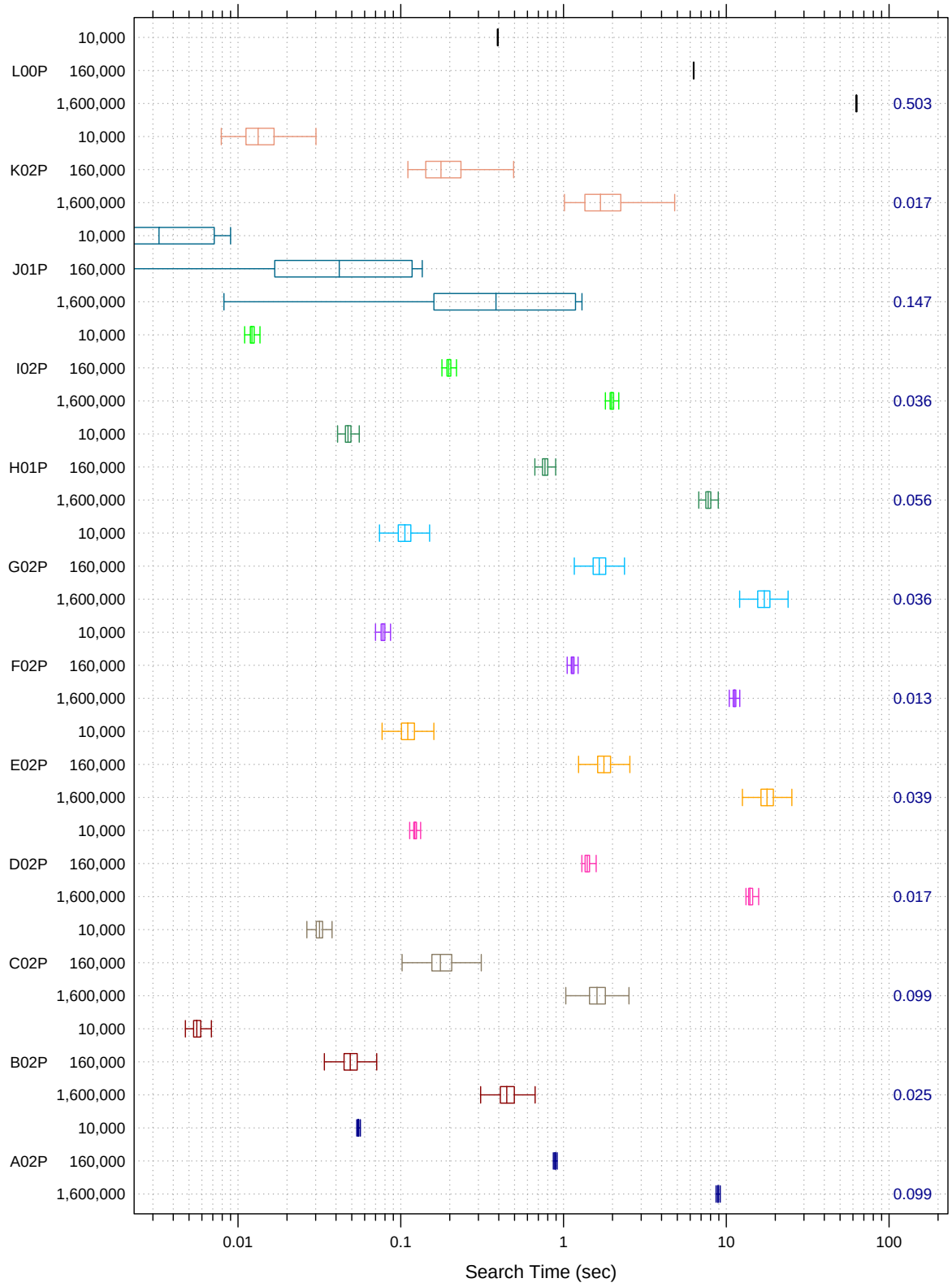


Figure 15: Search time for each Class P algorithm against different enrolled population sizes. For reference, FNIR at on FPIR of 10^{-3} is presented on the right when the enrolled population is 1.6 million.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

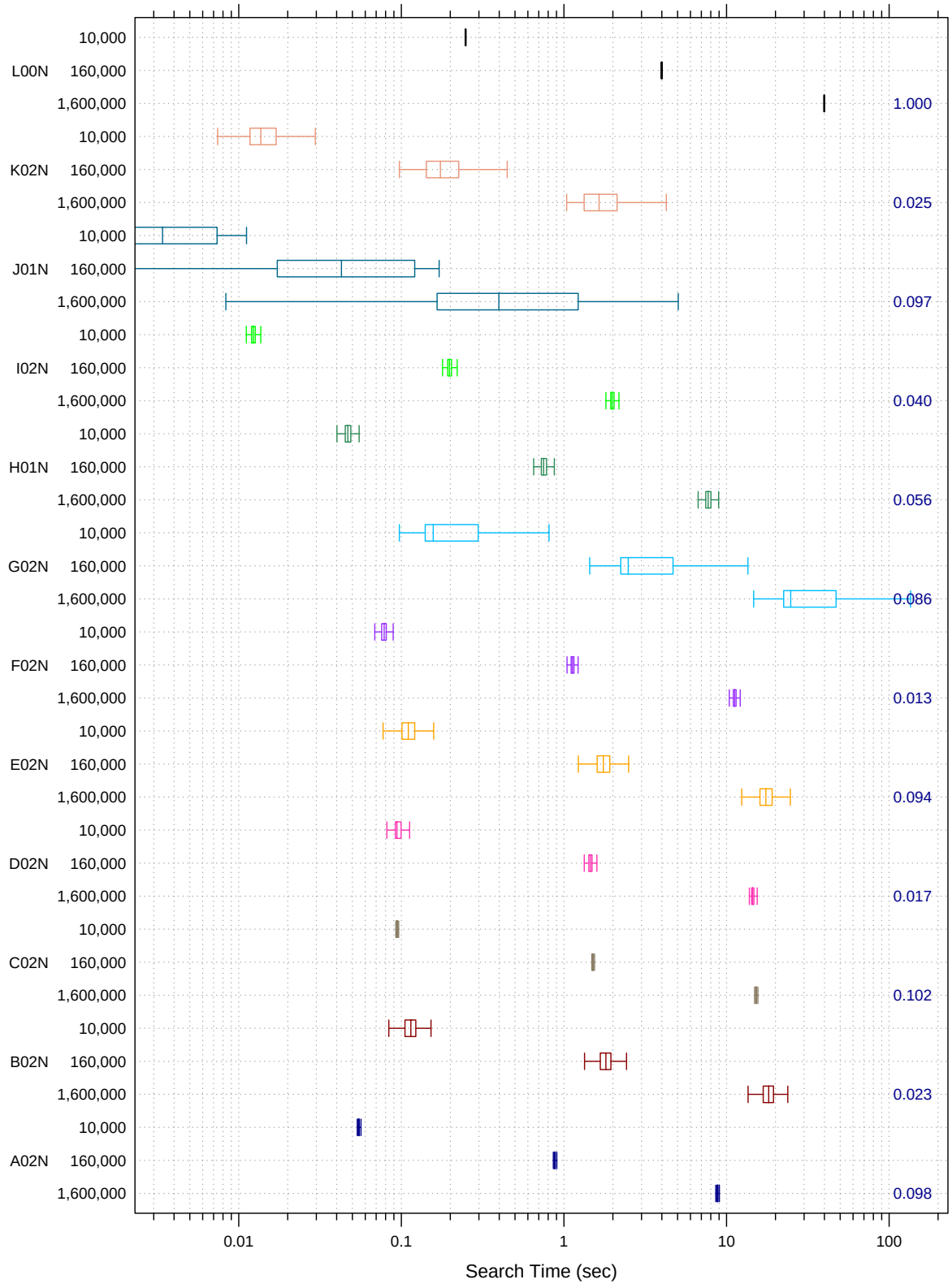


Figure 16: Search time for each Class N algorithm against different enrolled population sizes. For reference, FNIR at an FPIR of 10^{-3} assuming positive identification) is presented on the right when the enrolled population is 1.6 million.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.3 Template Size

Each recognition algorithm encodes feature information from an iris image into a proprietary representation that is later used for comparison. Even though proprietary templates often represent the useful features of the iris more compactly than raw images, storage and exchange of data between components should be performed using the standard image formats defined in ISO/IEC 19794-6:2011 and ANSI/NIST-ITL 1-2011 to maintain interoperability and prevent vendor lock-in.

Nevertheless, the sizes of proprietary templates can still dictate machine requirements. In addition to being stored on disk in centralized systems, enrollment templates are often permanently loaded into memory so comparisons can be performed more rapidly. Search templates need only be retained for the duration of the search and do not have to be permanently stored. For this reason, their sizes are less important.

Tables 4 and 5 show summary statistics for template sizes for each algorithm.

Results and Notable Observations

- Mean template sizes vary from 579 bytes (all enrollment templates from participants A and C) to 53 036 bytes (K02P). Most algorithms generate templates that have a fixed size. Some generate templates that are one of two sizes. The exception is K02P, where template sizes take on many values.
- Algorithms K00P, K02P, and those submitted by participants A and C produce asymmetrical search and enrollment templates such that their enrollment templates are much smaller. All other algorithms produce search and enrollment templates that have identical sizes.
- Eight providers' algorithms use larger templates than those provided by Cambridge University. This suggests a diversity of iris feature representations beyond the *IrisCode*.
- Templates generated using iris images of both eyes are always about twice the size of those generated from just one iris image. Most likely, the extracted feature information from both samples are simply being concatenated together.
- Broadly, templates for Class P and Class N algorithms are comparable in size.

	One-eye		Two-eye	
	Search Templates	Enrollment Templates	Search Templates	Enrollment Templates
A00P	7479 ± 0	579 ± 0	14830 ± 969	1148 ± 74
A01P	7479 ± 0	579 ± 0	14906 ± 621	1154 ± 50
A02P	14379 ± 0	579 ± 0	28672 ± 1108	1154 ± 46
B00P	2338 ± 0	2338 ± 0	4665 ± 49	4665 ± 45
B01P	4665 ± 36	4665 ± 41	9321 ± 52	9321 ± 47
B02P	4665 ± 36	4662 ± 95	9321 ± 52	9321 ± 46
C00P	14379 ± 0	579 ± 0	28449 ± 2086	1146 ± 81
C01P	14379 ± 0	579 ± 0	28758 ± 0	1158 ± 0
C02P	14379 ± 0	579 ± 0	28758 ± 0	1158 ± 0
D00P	2320 ± 5	2320 ± 6	4632 ± 9	4632 ± 10
D01P	9232 ± 0	9232 ± 0	18456 ± 0	18456 ± 0
D02P	9232 ± 0	9232 ± 0	18456 ± 0	18456 ± 0
E00P	4262 ± 0	4262 ± 0	8517 ± 115	8517 ± 112
E01P	4262 ± 0	4262 ± 0	8518 ± 100	8518 ± 92
E02P	4475 ± 917	4459 ± 883	8944 ± 1447	8913 ± 1386
F00P	1056 ± 0	1056 ± 0	2096 ± 0	2096 ± 0
F01P	1056 ± 0	1056 ± 0	2096 ± 0	2096 ± 0
F02P	2080 ± 0	2080 ± 0	4144 ± 0	4144 ± 0
G00P	5272 ± 0	5272 ± 0	10538 ± 137	10539 ± 127
G01P	5016 ± 0	5016 ± 0	10030 ± 26	10030 ± 18
G02P	5272 ± 0	5272 ± 0	10538 ± 145	10538 ± 140
H00P	2092 ± 0	2092 ± 0	4183 ± 38	4183 ± 37
H01P	2092 ± 0	2092 ± 0	4184 ± 10	4184 ± 13
I00P	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
I01P	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
I02P	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
J00P	1032 ± 0	1032 ± 0	2064 ± 7	2064 ± 8
J01P	1032 ± 0	1032 ± 0	2064 ± 7	2064 ± 8
K00P	5316 ± 0	607 ± 234	10632 ± 0	1214 ± 382
K01P	3836 ± 0	3836 ± 0	7672 ± 0	7672 ± 0
K02P	26504 ± 1664	20042 ± 1712	53008 ± 2913	40082 ± 3026
L00P	1025 ± 0	1025 ± 0	2050 ± 19	2050 ± 22

Table 4: Mean template size (in bytes) for Class P algorithms. Statistics were computed over 731,432 search and 160,000 enrollment templates.

	One-eye		Two-eye	
	Search Templates	Enrollment Templates	Search Templates	Enrollment Templates
A00N	14379 ± 0	579 ± 0	28512 ± 1864	1148 ± 74
A01N	7479 ± 0	579 ± 0	14958 ± 0	1155 ± 44
A02N	14379 ± 0	579 ± 0	28758 ± 0	1154 ± 49
B00N	4665 ± 36	4665 ± 41	9321 ± 52	9321 ± 47
B01N	1974 ± 15	1974 ± 17	3938 ± 22	3938 ± 20
B02N	4665 ± 36	4662 ± 95	9321 ± 52	9321 ± 46
C00N	14379 ± 0	579 ± 0	28758 ± 0	1158 ± 0
C01N	14379 ± 0	579 ± 0	28758 ± 0	1158 ± 0
C02N	22317 ± 1335	579 ± 0	44627 ± 2111	1158 ± 0
D00N	9232 ± 0	9232 ± 0	18456 ± 0	18456 ± 0
D02N	9232 ± 0	9232 ± 0	18456 ± 0	18456 ± 0
E00N	4262 ± 0	4262 ± 0	8513 ± 173	8514 ± 163
E01N	4262 ± 0	4262 ± 0	8520 ± 20	8520 ± 11
E02N	4476 ± 918	4459 ± 884	8947 ± 1447	8916 ± 1386
F00N	1056 ± 0	1056 ± 0	2096 ± 0	2096 ± 0
F01N	1056 ± 0	1056 ± 0	2096 ± 0	2096 ± 0
F02N	2080 ± 0	2080 ± 0	4144 ± 0	4144 ± 0
G00N	7258 ± 2206	7234 ± 2206	14513 ± 3696	14469 ± 3680
G01N	5016 ± 0	5016 ± 0	10030 ± 26	10030 ± 18
G02N	6746 ± 2170	6708 ± 2155	13489 ± 3440	13425 ± 3405
H00N	2092 ± 0	2092 ± 0	4183 ± 38	4183 ± 37
H01N	2092 ± 0	2092 ± 0	4184 ± 10	4184 ± 13
I00N	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
I01N	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
I02N	1028 ± 0	1028 ± 0	2056 ± 0	2056 ± 0
J00N	1032 ± 0	1032 ± 0	2064 ± 7	2064 ± 8
J01N	1032 ± 0	1032 ± 0	2064 ± 7	2064 ± 8
K00N	3836 ± 0	3836 ± 0	7672 ± 0	7672 ± 0
K01N	3836 ± 0	3836 ± 0	7672 ± 0	7672 ± 0
K02N	26519 ± 1650	20042 ± 1711	53036 ± 2887	40083 ± 3025
L00N	1025 ± 0	1025 ± 0	2050 ± 19	2050 ± 22

Table 5: Mean template size (in bytes) for Class N algorithms. Statistics were computed over 731,432 search and 160,000 enrollment templates.

3.4 Optimizing Matching for Specific Applications

Each participant submitted two classes of matching algorithms, each corresponding to a different general class of applications. Class P algorithms are intended for use in high-security applications, such as access control to highly sensitive information. Algorithms in this class should focus on optimizing accuracy at more strict (lower) decision thresholds. Class N algorithms are intended for applications that prioritize user convenience, such as access to theme parks. Algorithms in this class should focus on optimizing accuracy at more relaxed (higher) decision thresholds.

Figure 17 plots the lowest cost Class N and Class P algorithms for each participant when searches are performed against an enrolled population of 1.6 million. Cost is computed according to the cost model of equation 3. Error rates for both algorithm classes were computed for positive identification applications. (In Section 3.1.2 accuracy was computed for Class N algorithms assuming negative identification, where error rates are computed differently.)

3.4.1 Results and Notable Observations

Rather than optimizing accuracy to a specific application, many participants appear to have strived for across-the-board improvements with each submission. Some participants submitted the same algorithm for both classes. The only participant whose DET exhibited the predicted shape is G. At more strict (lower) operating thresholds, their lowest cost Class P algorithm is more accurate than their lowest cost Class N algorithm. At an FPIR of 10^{-4}, their lowest cost Class P algorithm achieves an FNIR that is 60 % lower than their Class N algorithm. At more relaxed (higher) thresholds, their lowest cost Class N algorithm is more accurate. At an FPIR of 10^{-2}, it achieves an FNIR that is 40 % lower than their Class P algorithm.

3.4.2 Operational Relevance

The purpose of requiring participants to submit matching algorithms for two classes is to see if algorithm developers can change the shape of their DET depending on whether the emphasis is the suppression of false positives or false negatives. Most participants were not able to achieve a desired shape change. Part of the reason for this may be that participants did not have access to meaningful development data. While algorithm developers were given limited performance data over the course of the evaluation, they were not provided with sufficient information that could lead to specific algorithm modifications. They may also have chosen to focus on across-the-board improvements in accuracy, which is understandable given the near-flatness of iris DET curves.

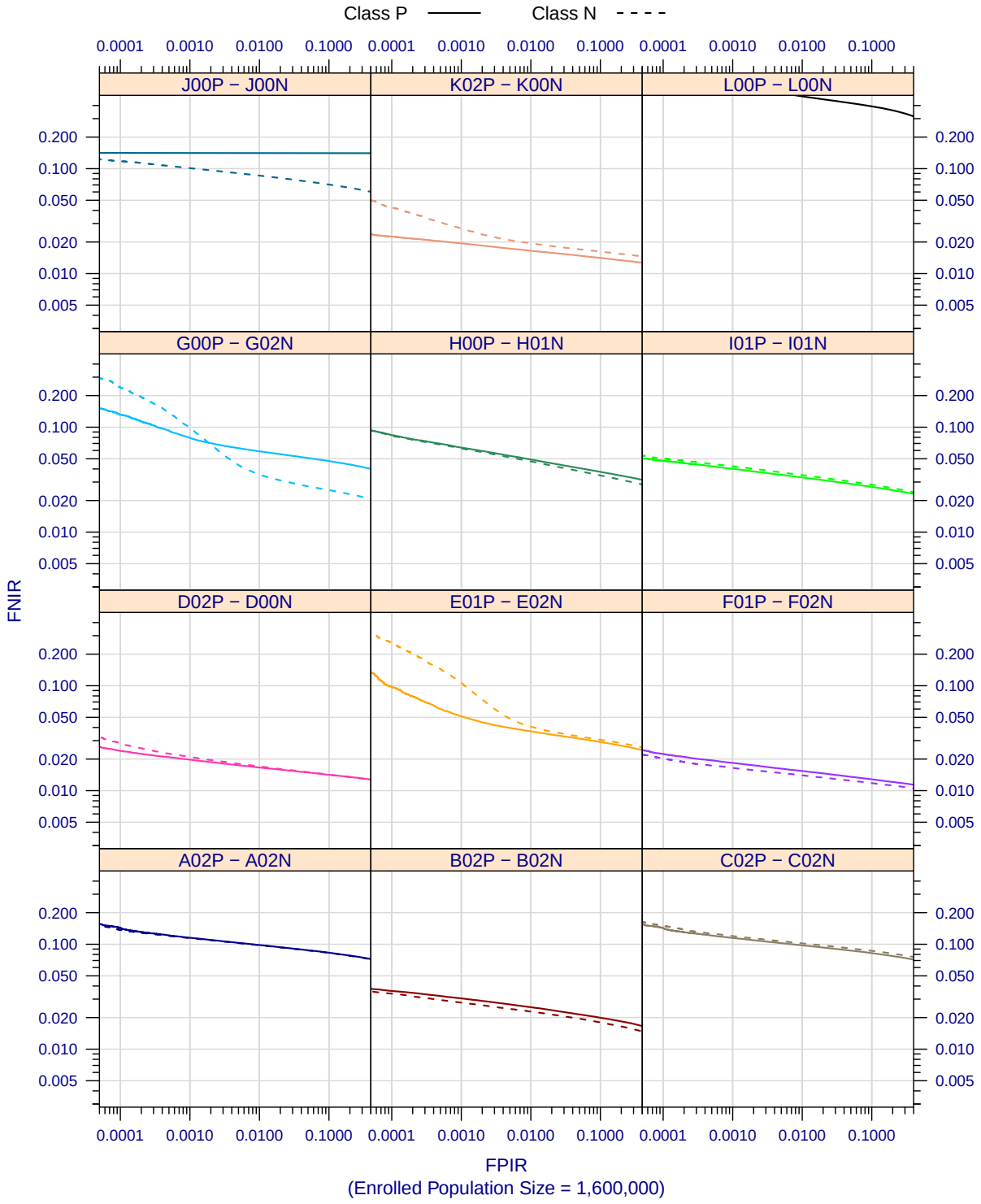


Figure 17: DET plots of the lowest cost Class P and Class N algorithms (according to the cost model of eq.3) from each participant. Plots were generated using 962 720 mated and 731 432 non-mated single-eye searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.5 One-eye vs. Two-eye Matching

Most iris recognition systems use both eyes for identification. When a dual-eye camera is used, images of both eyes are typically captured at nearly the same instant. As a result, their sample qualities are often highly correlated since people tend to blink, squint, or gaze off-axis equally with both eyes. Even when a single-eye camera is used to capture the images in succession, environmental conditions, behavioral habits (of both the camera operator and the subject), and other factors introduce strong correlations. These correlations are undesirable because they diminish the benefit of using both eyes for identification.

Iris cameras usually have a built-in quality filter that attempts to screen out poor quality images, but the filters occasionally fail, and some problems cannot be corrected through better presentation of the iris. Excessive pupil dilation caused by the consumption of certain drugs is a problem that afflicts both eyes and cannot be corrected at the time of capture. The IREX III Supplement also found that some subjects suffer from what appears to be a congenital defect characterized by a non-circular pupil that makes pupil localization more difficult. Since the condition is often present in both eyes, using both for matching may not solve the problem.

3.5.1 Accuracy

Figure 18 compares one-eye and two-eye DET accuracy for Class P algorithms. It should be noted that although images of both eyes were available for matching, the algorithms were not provided with information on which was the left, and which was the right. The IREX III report states that switching from one-eye to two-eye matching appears to result in a downward translation (on a log scale) of the DET curve. This implies the following simple relation

$$\rho = \frac{P_{FN}(FP_2)}{P_{FN}(FP_1)}\,, \tag{6}$$

where $P_{FN}(FP_1)$ and $P_{FN}(FP_2)$ are the probabilities of a false negative at a fixed FPIR for one-eye and two-eye matching respectively, and ρ is simply a constant indicating the degree of correlation between the left and right eyes. A value of $\rho = 1$ indicates a perfect correlation (i.e. using the second eye provides no accuracy benefit), while a value of 0 indicates a perfect negative correlation (i.e. using the second eye results in perfect accuracy). A value greater than 1 would indicate accuracy gets worse when the second eye is used. Figure 20 shows that the value of ρ remains fairly constant over a wide range of FPIR for most algorithms.

Figure 19 compares DET accuracy for one-eye and two-eye matching for Class N algorithms, where it appears most participants attempted alternative methods of fusion that generally yielded poorer results than their Class P algorithms.

Results and Notable Observations

- Switching to two eyes reduces FNIR by about a factor of two at a fixed FPIR for most Class P algorithms. For a few algorithms (E02P, G02P, L00P) the benefit to matching with both eyes diminishes at smaller FPIRs.
- At a fixed operating threshold, FPIR is greater for two-eye matching than single-eye matching for most algorithms. Some algorithms (D02P, F02P, K02P, B02N, D02N, F02N, K02N) appear to normalize their comparison scores such that FPIR changes little, or decreases, when both eyes are used.
- Participants often attempted different fusion strategies for their Class P and Class N algorithms. Participants B and D appear to use the same fusion strategy but with a different method of score normalization. FNIR decreases when both eyes are used for their Class N algorithms, but not for their Class P algorithms.
- I02N appears to use an incompatible range of comparison scores for one-eye and two-eye matching. Reasonable decision thresholds for one-eye matching would produce unacceptably high FPIR for two-eye matching.

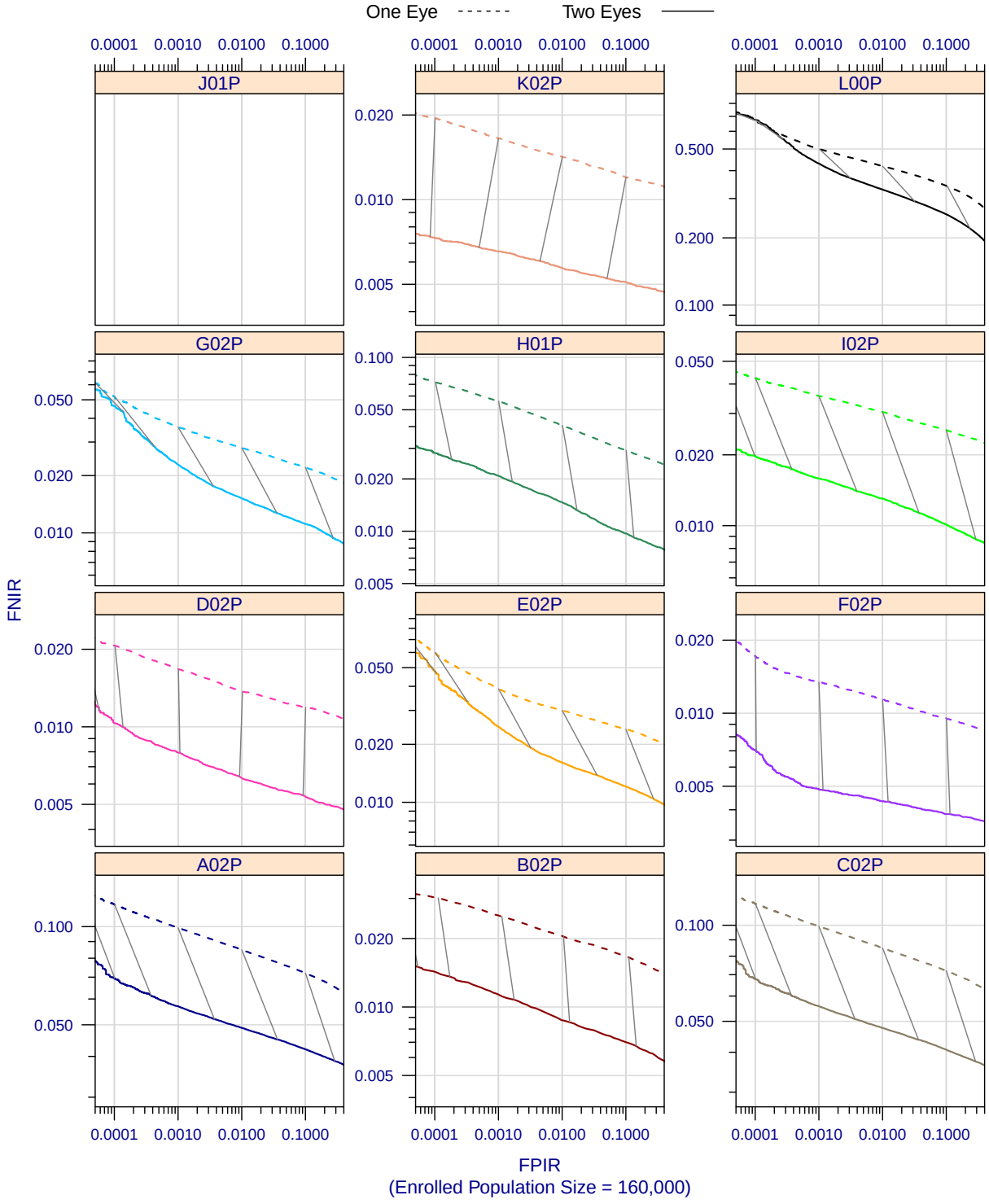


Figure 18: Comparison of one eye and two eye accuracy for Class P algorithms when searching against an enrolled population of 160 thousand. Plots were generated using 96 636 mated and 731 432 non-mated single-eye searches, and 96 076 mated and 365 716 non-mated two-eye searches. Gray line segments connect points of equal threshold, showing how error rates for single-eye searches compare to two-eye searches at the same decision threshold.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

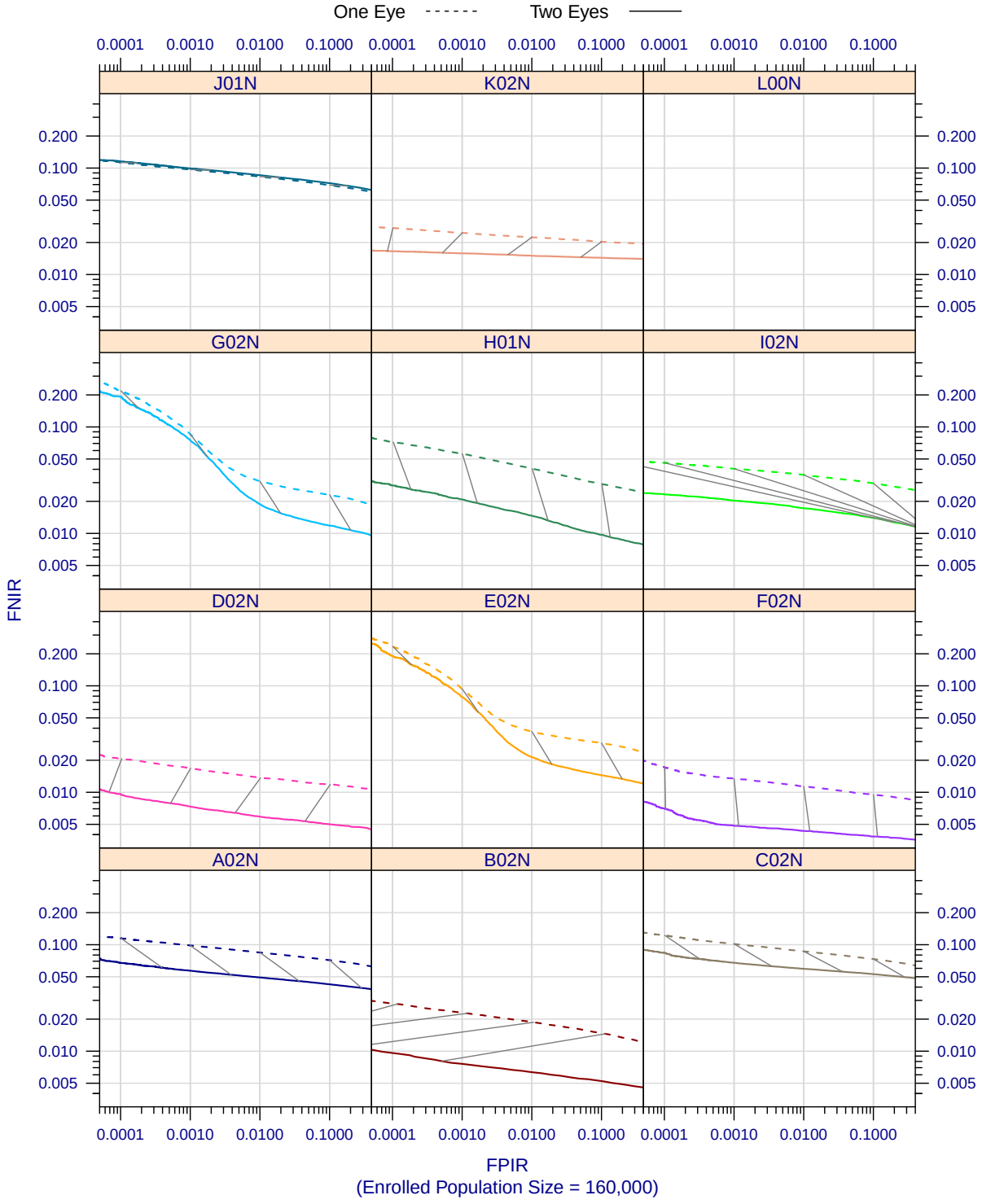


Figure 19: Comparison of one eye and two eye accuracy for Class N algorithms when searching against an enrolled population of 160 thousand. Plots were generated using 96 636 mated and 731 432 non-mated single-eye searches, and 96 076 mated and 365 716 non-mated two-eye searches. Error rates were computed assuming positive identification. Gray line segments connect points of equal threshold, showing how error rates for single-eye searches compare to two-eye searches at the same decision threshold.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

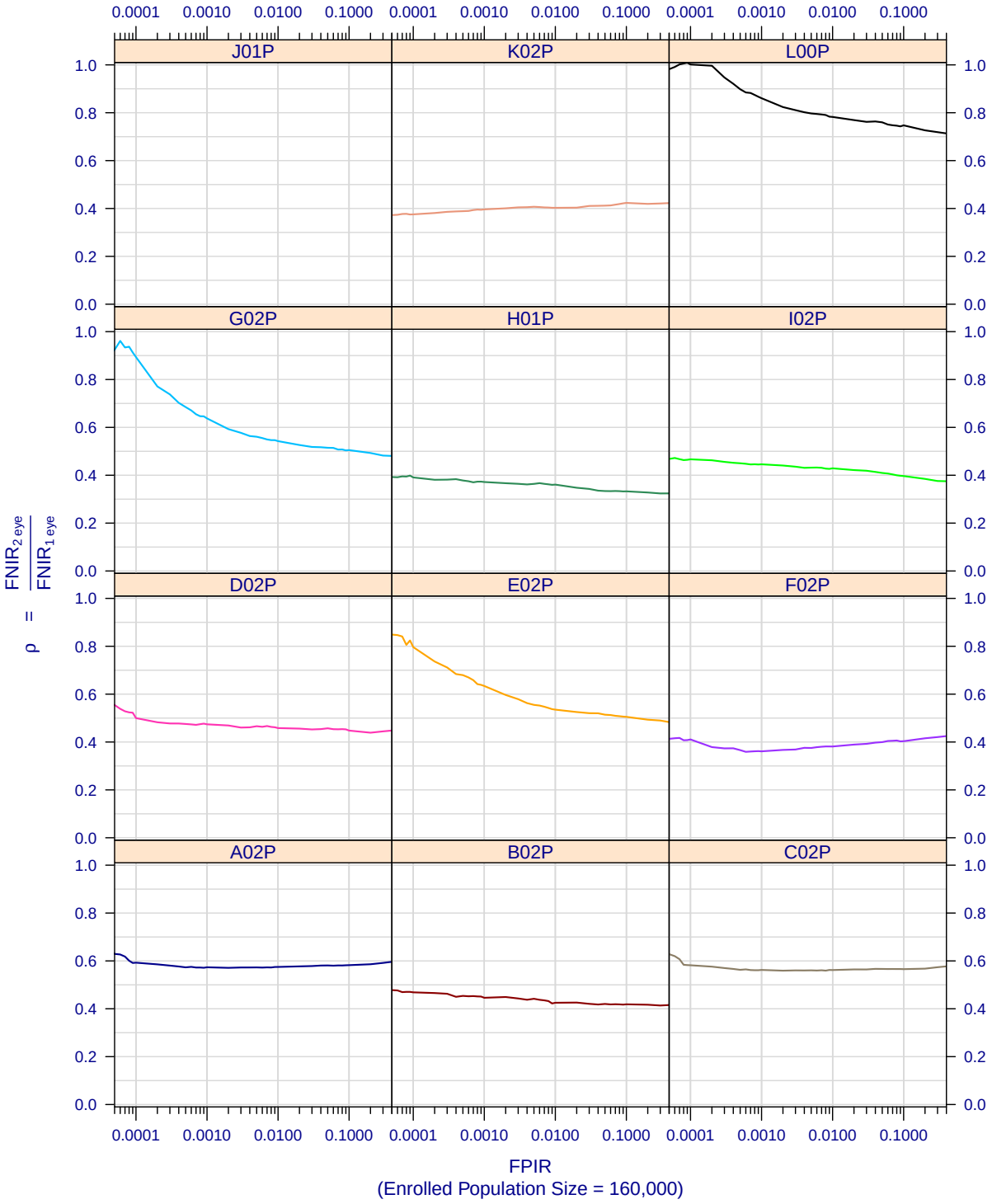


Figure 20: Plots showing the fractional change in FNIR as a function of FPIR when switching from single-eye to two-eye matching for Class P algorithms. For most algorithms, FNIR decreases by about a factor of two when the second eye is used. Plots were generated using 96 636 mated and 731 432 non-mated single-eye searches, and 96 076 mated and 365 716 non-mated two-eye searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.5.2 Speed

Matching with both eyes reduces error rates but is more computationally expensive. Figure 21 shows the difference in search time for one-eye and two-eye matching for Class P algorithms. Most algorithms take about 4 times longer to search when matching with both eyes. This would be expected if the algorithm internally compares each eye of the person being searched to each eye of the enrolled individuals. (The search time would be expected to only double if the algorithms were provided with information on which sample is of the left eye and which is of the right). Two exceptions are K02P, where the median search time only increases by about a factor of 3, and J01P, where the median search time only increases by about a factor of 2.

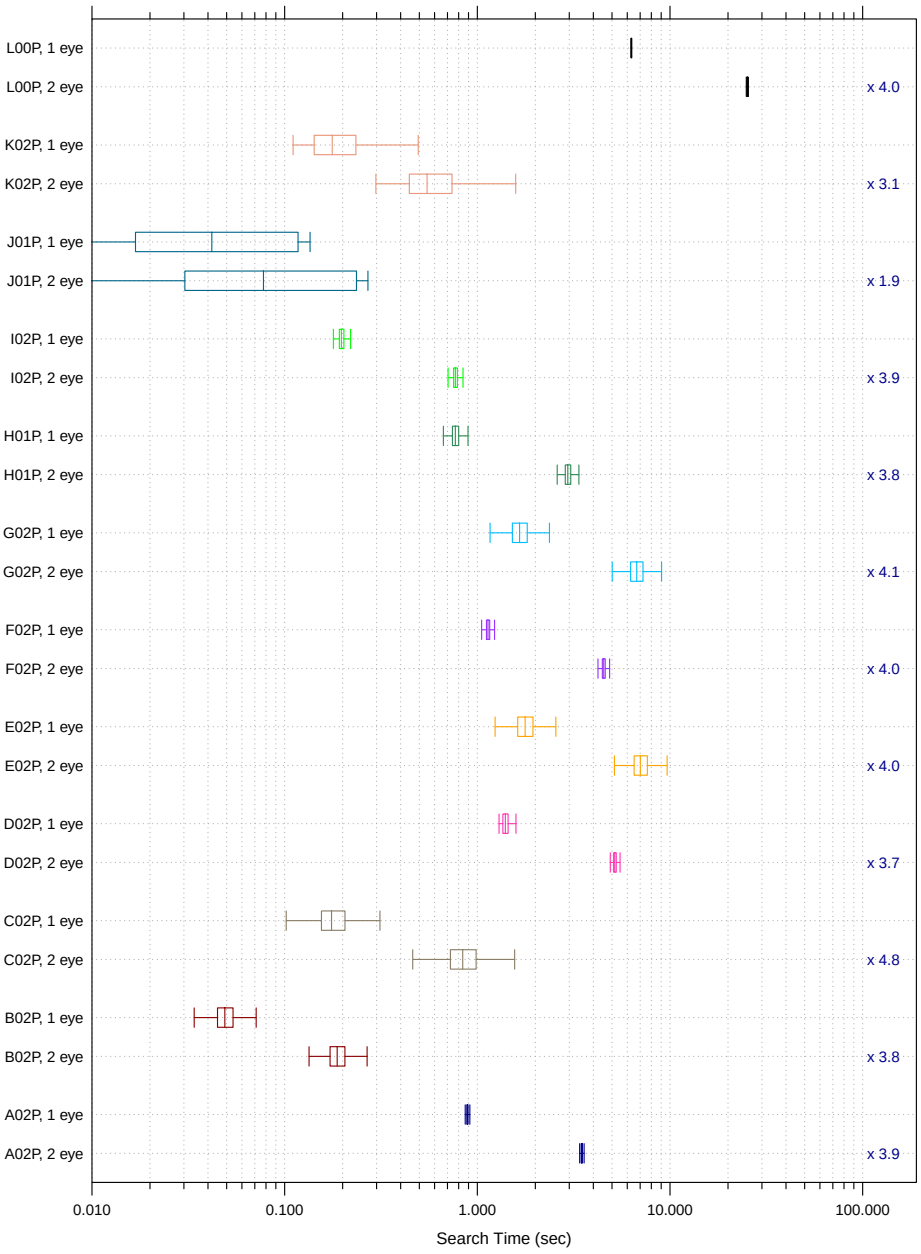


Figure 21: Comparison of search duration for one-eye and two-eye matching. Each boxplot shows the distribution of search times for 5 000 mated searches against an enrolled population of 160 thousand. The increase in median search time when switching from one-eye to two-eye matching for each algorithm is displayed on the right.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

3.5.3 Operational Relevance

When searching with two eyes, most iris recognition algorithms achieve a factor of 2 reduction in FNIR at a fixed FPIR, although the time to perform a search increases by a factor of 4. Sample quality is strongly correlated between a person's left and right iris images. This is true regardless of whether the images were captured concurrently with a dual-eye camera, or in succession with a single-eye camera. IREX III found that the dependency between images of a person's left and right irides is stronger than the dependency between fingerprint scans of a person's left and right index fingers.

Longer search durations can delay response times for systems that perform matching in real-time, which in turn can reduce throughput rates at physical access control points or biometric scanning stations. Nevertheless, iris recognition is very fast compared to other biometric modalities. The fastest algorithms only require an additional 3 to 4 seconds to search both eyes against an enrolled population of 1.6 million. By employing multiple CPU cores, this time could be reduced to a fraction of a second. If the samples must be transferred over a limited bandwidth network to a central facility, having to transfer both images will increase bandwidth requirements. Finally, using both eyes for identification can increase workload for camera operators using single-eye cameras. Whether it is more advantageous to perform identification with one or both eyes will depend on the application.

An iris recognition system could be designed to support both one-eye and two-eye matching. This would certainly be reasonable if images of both eyes were only available for a subset of the enrolled population. At a fixed decision threshold, most algorithms are more likely to return false negatives but less likely to return false positives for single-eye searches. However, some algorithms normalize their comparison scores such that (non-mated) single-eye searches are less likely, or about as likely, to return a false positive.

3.5.4 Contingent Fusion for Rapid Identification

In this section we apply *contingent fusion* [46] to the identification problem. Contingent fusion is a method of identification that applies fusion only if a search using the first presented sample does not result in a decisive match. Our aim is to achieve accuracy close to that of two-eye matching while minimizing the frequency at which both eyes must be used. There are several advantages to requiring only one eye for the majority of transactions. First, it is less computationally expensive than always searching with both eyes. Second, it consumes less bandwidth when the samples are transferred over a network for a back-end search. Finally, it obviates the need for operators of single-eye cameras to capture an image of the second eye if the first proves sufficient.

Contingent fusion is applied at the decision level [47] such that the second eye is used only if the first does not return a decisive match. A match is considered decisive if the top candidate has a dissimilarity score less than or equal to a critical threshold. Otherwise, the second eye is searched and the claim of enrollment is accepted only if the top candidate in either list is less than or equal to a second critical threshold. Formally, a particular identity is accepted only if

$$s_1 \leq \tau_1 \ \cup \ \min(s_1, s_2) \leq \tau_2 \tag{7}$$

where s_i is the person's score on the candidate list corresponding to the ith eye, $i = 1, 2$ ($s_i = \infty$ if the person is not on the list), and τ_1 and τ_2 are the first and second critical thresholds respectively. The first critical threshold determines the frequency at which the second eye is used while the second threshold is used to make final decisions. For simplicity, we set both critical thresholds to the same value, which also makes the minimum score for both eyes a sufficient statistic for equation 7.

Results

Figure 22 compares accuracy for single-eye, two-eye, and two-eye contingent matching. Two-eye matching delegates the responsibility of fusion to the matcher, which allows more flexibility with respect to how fusion is performed.

The notable observations are as follows:

- Contingent fusion always achieves accuracy comparable to that of two-eye matching. For some algorithms, contingent fusion actually appears to perform better, although the improvement is typically within the margin of error.
- For the two most accurate matchers (F02P, K02P), contingent fusion is almost as accurate as two-eye matching.
- For the two most accurate matchers (F02P, K02P), contingent fusion uses the second eye no more than 2 % of the time when the FPIR is 10^{-4} or greater. This is easy to determine from the figure since the rate at which the second eye is used is equal to the FNIR for single eye searches at the same threshold.

Operational Relevance

Contingent fusion is capable of achieving accuracy comparable to that of two-eye matching while only using the second eye between 1 % and 2 % of the time. Thus, for many applications contingent fusion may yield faster response times and consume fewer computational resources compared to always searching with both eyes. It can also provide a boost in accuracy over single-eye matching without appreciably affecting other performance criteria. However, for live-capture setups the same accuracy improvement could probably be achieved by simply re-acquiring an image of the same eye since most failures are due to poor presentation.

When captures are performed sequentially rather than concurrently, the accuracy improvement is likely to be even greater than what has been presented here. This is because concurrent captures of both eyes tend to produce more correlated samples in terms of their quality (e.g. if one eye is heavily occluded, the other is likely to be as well). The downside to acquiring an image of the second eye only after the first proves insufficient is that the camera operator must wait for the results of the first search before proceeding with the second capture. However, this is expected to occur so infrequently that, as long as response times are reasonable, it should not lead to substantial increases in operator workload or decreases in throughput rates compared to only capturing a single eye for all transactions.

We only explored one method of contingent fusion in this section. Others are possible, including strategies that employ multiple matchers or multiple captures of the same eye.

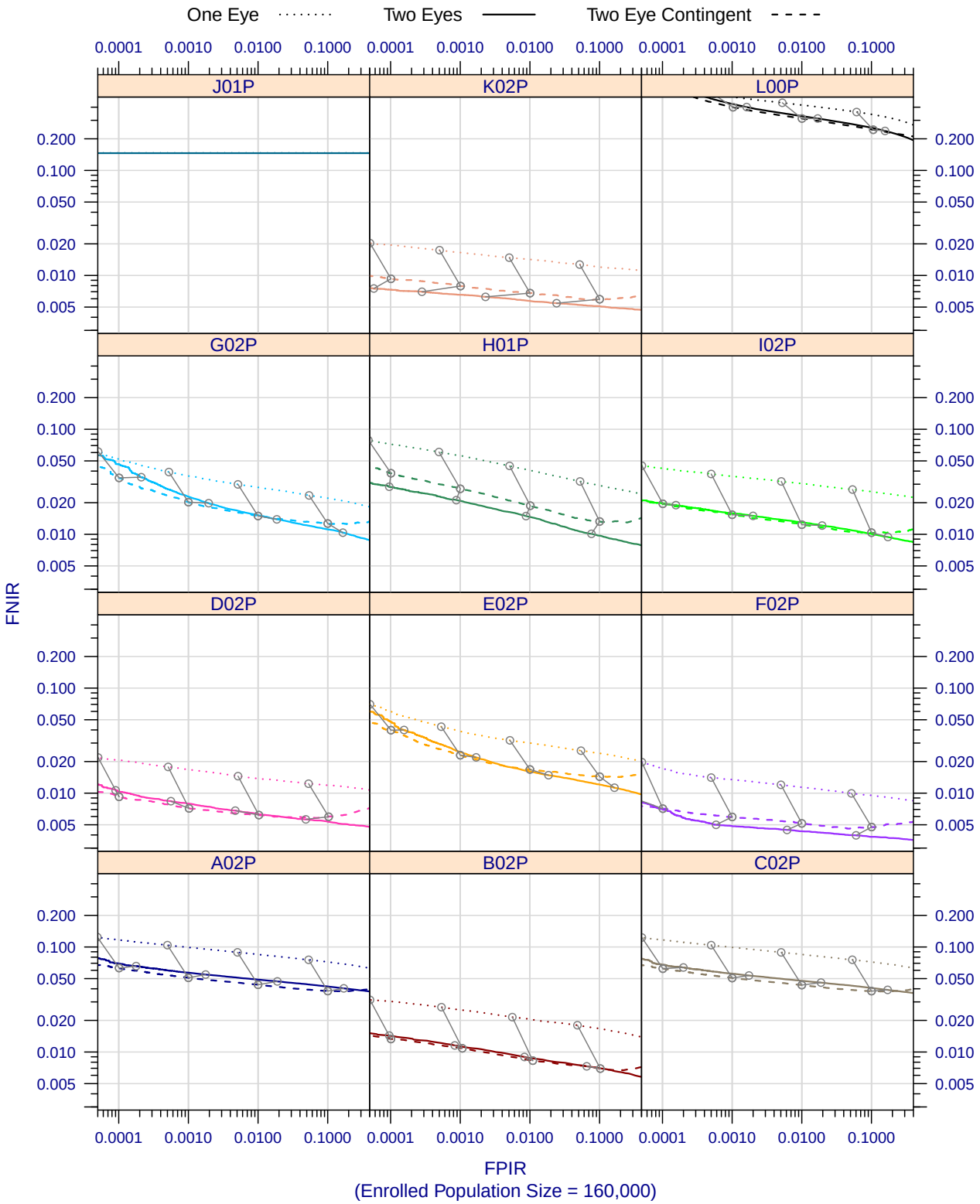


Figure 22: Comparison of DET accuracy for single-eye, two-eye, and two-eye contingent fusion matching strategies for several algorithms. Plots were generated using 96 636 mated and 731 432 non-mated single-eye searches, and 96 076 mated and 365 716 non-mated two-eye searches. Gray line segments connect points of equal threshold, showing how error rates for single-eye searches compare to two-eye searches at the same decision threshold.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

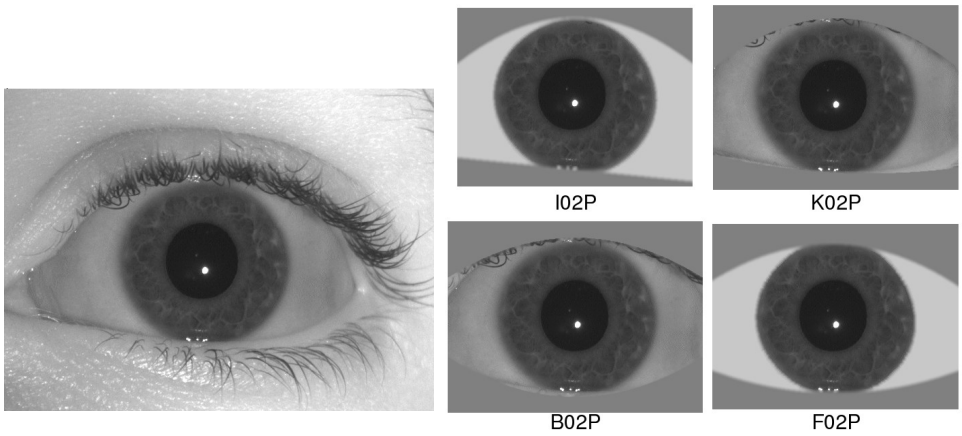


Figure 23: Examples of different algorithm attempts to create a Kind 7 record (right) from a raw iris image (left).

3.6 JPEG 2000 Compression of Iris Images

The ISO/IEC 19794-6 and ANSI/NIST-ITL 1-2011 standards define a set of specifications for the compact representation of iris images. Compact storage formats are useful for transmitting samples across bandwidth-limited networks or for storage on space-limited devices (such as ISO/IEC 7816 crypto-tokens [48]). The most efficient storage format defined in the standard is the Kind 7 record, which requires a cropping of the area around the iris followed by a masking of the eyelids and sclera with a solid color. Masking and cropping in this fashion ensures that a maximum of the encoding budget is dedicated to representing the iris features rather than the less useful region around the iris.

Four participants chose to submit algorithms that support the generation of Kind 7 records. Figure 23 shows examples of attempts by algorithms from each participant to generate a Kind 7 record from an iris image. Algorithms K02P and B02P both failed to mask the sclera and do not blur the borders between the eyelids and the rest of the eye. Algorithm B02P also tends to be overly conservative when masking the eyelids such that parts of the eyelashes are still visible. These behaviors make it difficult to compress converted iris records to small file sizes while retaining maximum fidelity of the iris features. In this example, algorithms F02P and I02P properly masked the eyelids and sclera and introduced a blur along the sclera-eyelid borders.

Figure 24 shows the effect of lossy JPEG 2000 compression on accuracy when the search images are converted to Kind 7 records and compressed to different file sizes. Searches were performed against an enrolled population of 10 thousand. At an FPIR of 10^{-3}, FNIR rises from 0.036 to 0.078 when search images are compressed to a kilobyte for algorithm I02P. FNIR rises from 0.013 to 0.047 for algorithm F02P. Error rates increase by a greater amount for algorithms K02P and B02P due to their failure to completely adhere to specifications.

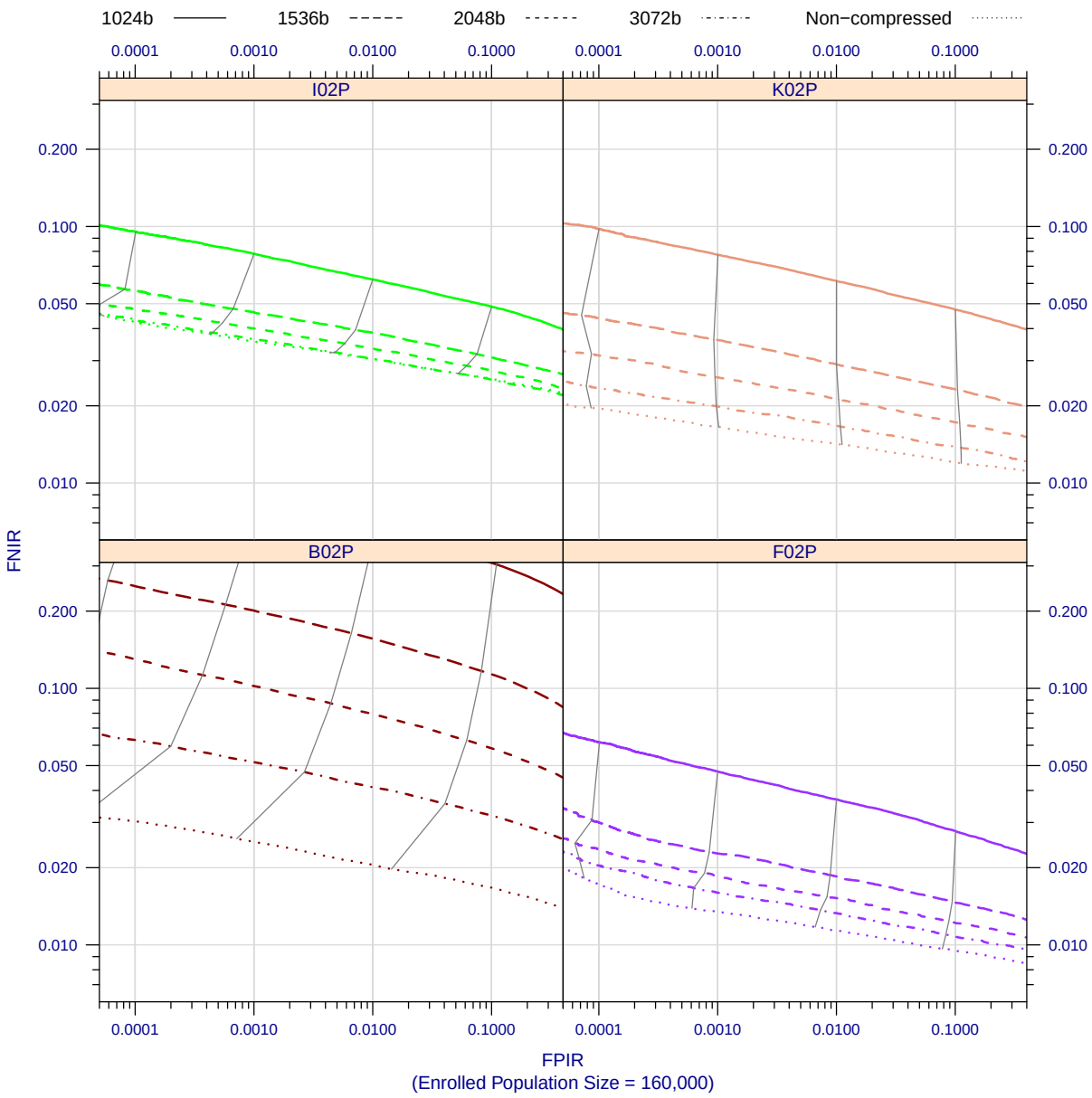


Figure 24: Effect of JPEG 2000 compression on single-eye DET accuracy when search images are converted to Kind 7 records and then compressed. The plots were generated from 96 636 mated and 731 432 non-mated searches against an enrolled population of 160 thousand.

3.7 Comparison to IREX III

3.7.1 Effect of JPEG Compression

IREX III tested algorithm performance using images that were provided to NIST after they had been compressed with JPEG at a quality of 75. This compression introduced blocking artifacts that are visible upon close inspection in some of the images. This section analyzes the effect that these compression artifacts have on recognition accuracy.

Figure 25 shows boxplots comparing the uncertainty of FNIR at an FPIR of 10^{-3} when search and enrollment images are compressed with JPEG (at a quality of 75) and when they are non-compressed. For simplicity, only the last (and usually most accurate) Class P submission from each participant is included in the figure. Figure 26 compares DET accuracy between compressed and non-compressed comparisons.

JPEG compression introduces a small but discernible effect on accuracy. At a fixed FPIR, FNIR is typically 2-11 % higher when the images are compressed. At an FPIR of 10^{-3}, the FNIR for the most accurate matcher (F02P) increases from 0.0134 to 0.0144, an increase of 7.0 %. For B02P, the FNIR increases from 0.0249 to 0.0277, an increase of 10.9 %. The lines of equal threshold in Figure 26 move up and to the left when the images are compressed, indicating that the changes make identification more difficult, but false matches also become slightly less frequent.

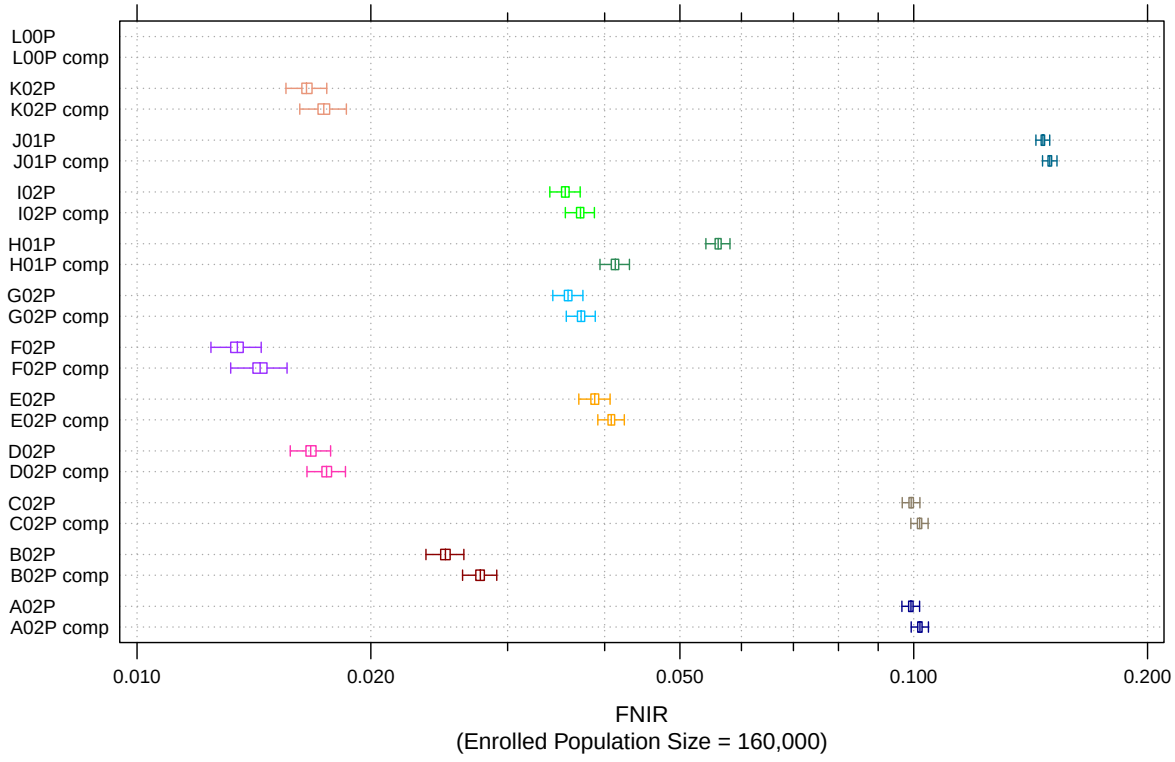


Figure 25: Boxplots summarizing the probability distribution of FNIR at an FPIR of 10^{-3} for single-eye searches when search and enrollment images are JPEG compressed at a quality of 75 (dotted lines), and when they are non-compressed (solid lines). Plots were generated using 96 636 mated and 731 432 non-mated searches.

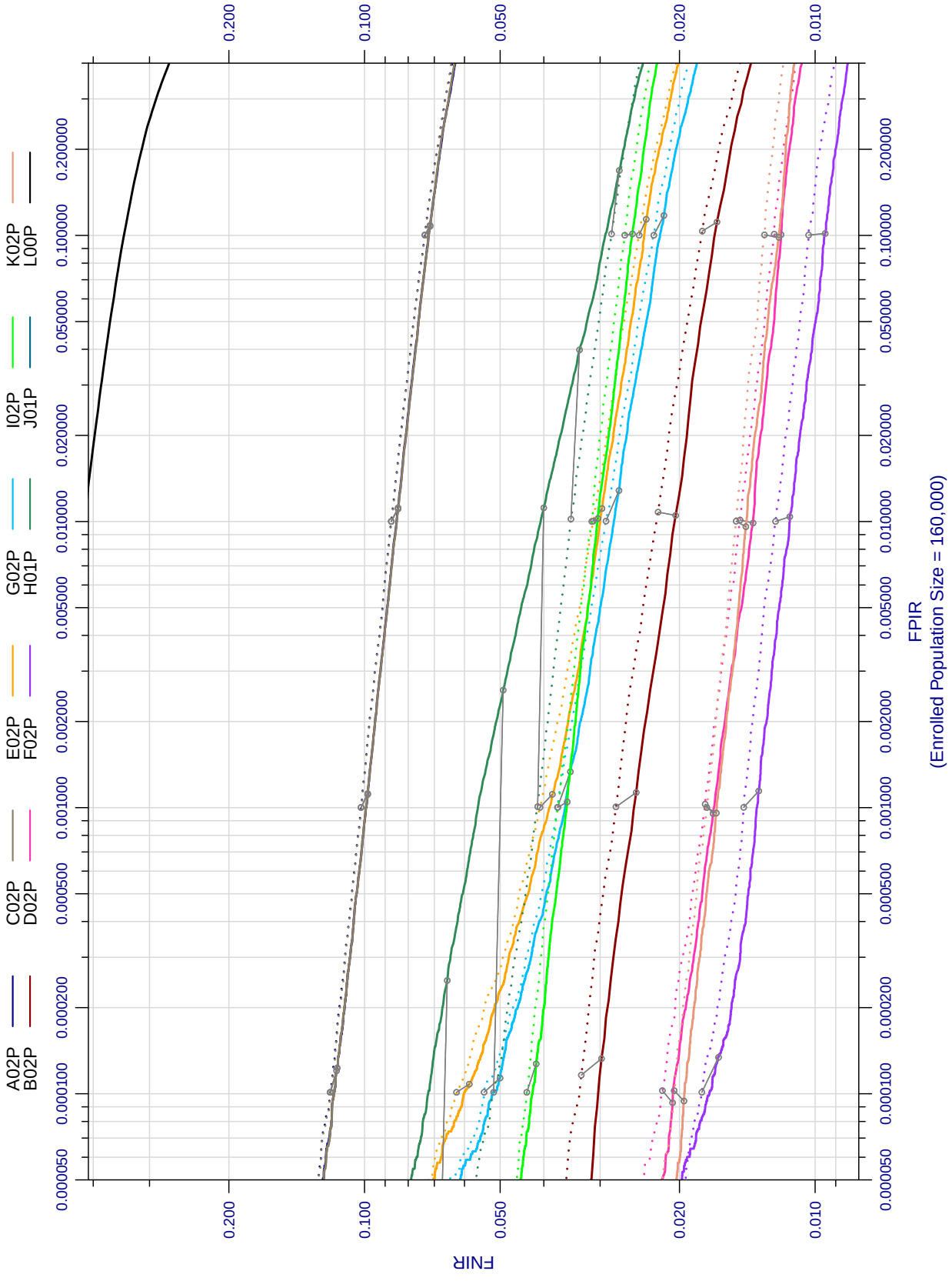


Figure 26: DET curves comparing accuracy when images are compressed (at a quality of 75) and when they are non-compressed. Plots were generated using 96 076 mated and 731 432 non-mated searches against an enrolled population of 160 thousand.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

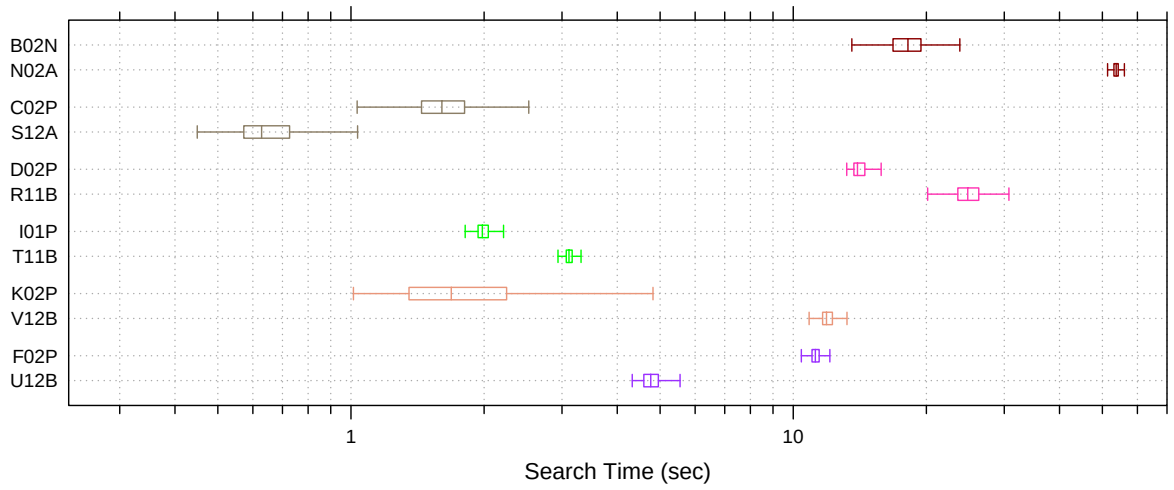


Figure 27: Comparison of search times between IREX III and IREX IV algorithms. Algorithms from the same participant have identical colors. Searches were performed against an enrollment population of 1.6 million. Boxplots for IREX IV algorithms were generated from 5 000 mated searches. Boxplots for IREX III algorithms were generated from 1 000 mated searches.

3.7.2 Performance Comparison

Although IREX III was conducted only 12 months prior to IREX IV, progress in automated recognition technology can occur fairly rapidly [49]. The concept of the iris as a biometric is fairly new, and most research activity did not begin until after the expiration of the Flom-Safir patent in 2005 [1]. Many companies have only recently begun developing their recognition algorithms and are likely to have made improvements since the conclusion of IREX III.

Since this evaluation defines FPIR differently than IREX III (see Section 2.4.1), it was necessary to recompute DET performance for the legacy IREX III algorithms. A direct comparison to the previous evaluation is still difficult because a different set of images were used. Although pulled from the same source, the images in this evaluation were never compressed with JPEG (see Section 3.7.1). The sampling strategy was also different and the current dataset contains many images that were not available at the time IREX III was conducted. These images were captured more recently and are more likely to have been acquired with newer cameras, although other properties of the images may also be different.

Results

Figure 28 compares DET accuracy for participants that submitted algorithms to both evaluations. Algorithms from the same participant are assigned identical colors. In IREX III, the most accurate algorithm achieves an FNIR of 0.018 at an FPIR of 10^{-3}. In this evaluation, F02P achieves an FNIR of 0.016 at the same FPIR. Figure 27 compares search times for the same set of algorithms.

The notable observations are:

- The most accurate algorithm in IREX IV produces slightly lower error rates than the most accurate algorithm from IREX III.
- Some participants achieved lower error rates in IREX III than in IREX IV. The most likely explanation is that IREX IV used a more difficult set of images for testing. Participants C and I improved the most since IREX III.
- Search times in IREX III and IREX IV are comparable. Sometimes a participant's most accurate algorithm from IREX IV is faster than their most accurate algorithm from IREX III. Sometimes the reverse is true. Participant R from IREX III submitted an algorithm (R12A) that matches nearly as accurately as R11B but has a median search time under 2 seconds.

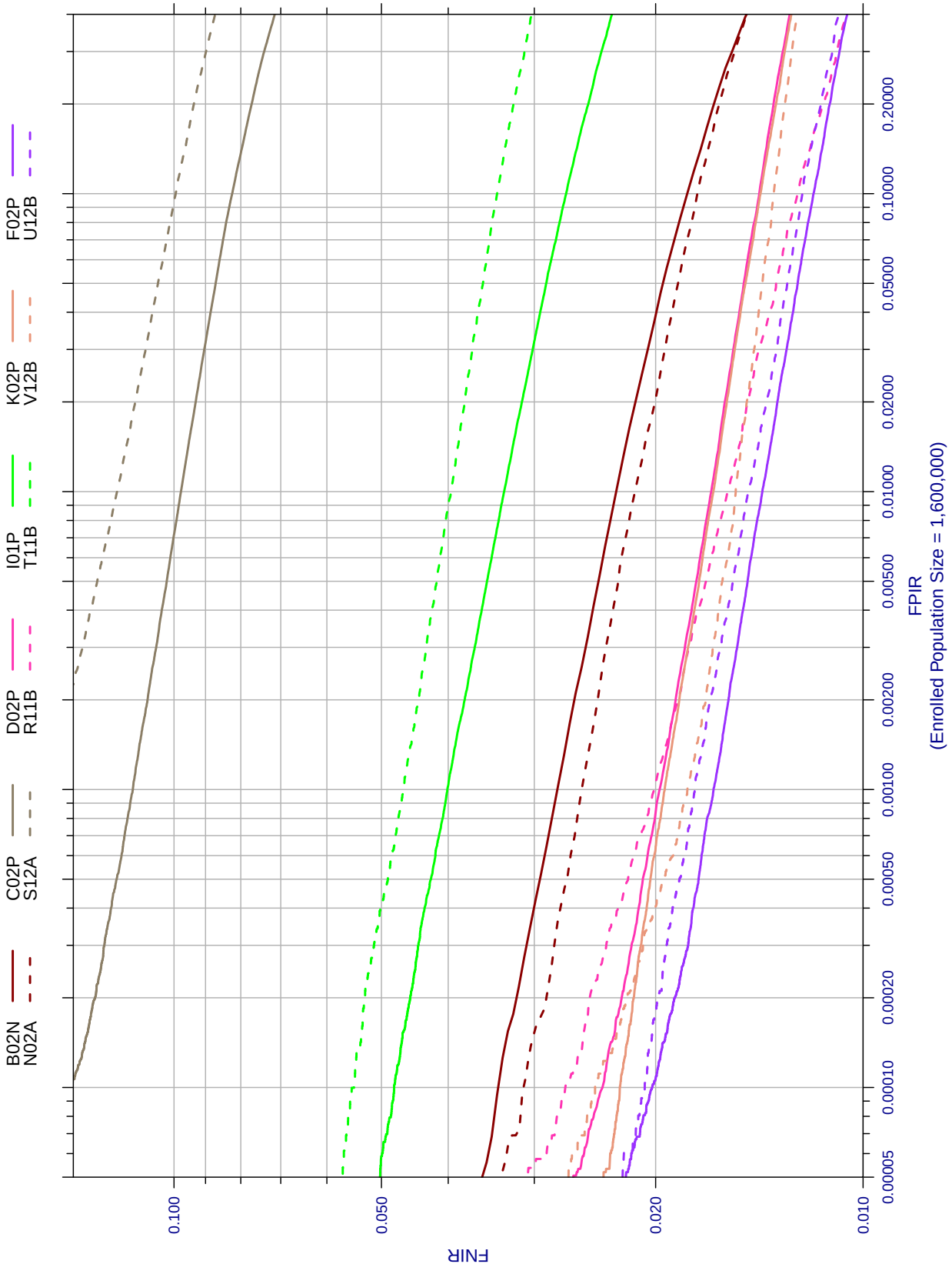


Figure 28: Comparison of single-eye DET accuracy between IREX IV (solid) and IREX III (dotted) algorithms when searches are performed against an enrolled population of 1.6 million. Algorithms from the same participant are given identical colors.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

4 References

[1] L. Flom and A. Safir, "Iris Recognitions System." http://www.google.com/patents?id=4I4tAAAAEBAJ, Feb. 1985. 2, 54

[2] F. Zelazny, "The Evolution of India's UID Program." http://www.cgdev.org/files/1426371_file_Zelazny_India_Case_Study_FINAL.pdf, 2012. 2

[3] "Social inclusion and aadhaar: Introduction and concept paper." http://uidai.gov.in/images/concept_paper_social_inclusion.pdf, 2012. 2

[4] P. Grother, G. Quinn, J. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson, "IREX III: Performance of Iris Identification Algorithms." http://www.nist.gov/itl/iad/ig/irexiii.cfm, 2011. 8, 10

[5] P. Grother, E. Tabassi, G. W. Quinn, and W. Salamon, "Performance of Iris Recognition Algorithms on Standard Images." http://www.nist.gov/itl/iad/ig/irex.cfm, 2009. 8, 9

[6] "The IREX Program." http://www.nist.gov/itl/iad/ig/irex.cfm. 8, 9

[7] "Prime Minister launches Aadhaar Enabled Service Delivery." Press Release, October 2012. http://uidai.gov.in/images/2nd_anniversary/uidai_press_release_for_oct_20.pdf. 8, 27

[8] A. N. Al-Raisi and A. M. Al-Khouri, "Iris recognition and the challenge of homeland and border control security in UAE," *Telematics and Informatics*, vol. 25, no. 2, pp. 117–132, 2008. 8

[9] "United Arab Emirates Iris System." http://www.adpolice.gov.ae/en/portal/eye.scan.aspx. 8, 20

[10] "HIIDE Gets an Update: New Software Adds Storage Capacity and More." http://www.biometrics.dod.mil/Newsletter/issues/2008/Oct/v4issue4_pm.html. 8

[11] *ISO/IEC 19794-6 - Biometric Data Interchange Formats - Iris Image Data*. 2009. 8

[12] *ANSI/NIST-ITL 1-2011 Data Format for the Interchange of Fingerprint, Facial & Other Biometric Information*. 2011. 8

[13] E. Tabassi, P. Grother, and W. Salamon, "IREX - IQCE Performance of Iris Image Quality Assessment Algorithms." http://www.nist.gov/itl/iad/ig/irexii.cfm, 2011. 9

[14] *ISO/IEC 29794-6 - Biometric Sample Quality Standard- Part 6: Iris Image*. 2012. 9

[15] "NEXUS Program Description." http://www.cbp.gov/xp/cgov/travel/trusted_traveler/nexus_prog/nexus.xml. 10, 16, 20

[16] G. Levin, "Real World. Most Demanding Biometric Applications," in *Biometric Consortium Conference*, 2007. 10

[17] G. Quinn and P. Grother, "IREX III Supplement I: Failure Analysis." http://www.nist.gov/itl/iad/ig/irexiii.cfm, 2011. 10, 12

[18] P. Grother, G. Quinn, and J. Phillips, "Report on the Evaluation of 2D Still-image Face Recognition Algorithms," tech. rep., NIST, 2010. 10

[19] P. J. Phillips, A. Martin, C. l. Wilson, and M. Przybocki, "An Introduction to Evaluating Biometric Systems," *Computer*, vol. 33, pp. 56–63, Feb. 2000. 10

[20] G. Quinn and P. Grother, "IREX IV: Evaluation of One-to-Many Iris Recognition ,Concept, Evaluation Plan, and API Specification." http://www.nist.gov/itl/iad/ig/irexiv.cfm, May 2012. 11

[21] G. K. Wallace, "The JPEG still picture compression standard," *Communications of the ACM*, pp. 30–44, 1991. 12

[22] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proc. Eurospeech*, pp. 1895–1898, 1997. 13

[23] K. E. Iverson, *A programming language*. New York, NY, USA: John Wiley & Sons, Inc., 1962. 13

[24] R. Bolle, J. Connell, S. Pankanti, N. Ratha, and A. Senior, *Guide to Biometrics*. Springer, 2004. 13

[25] "The NIST Year 2010 Speaker Recognition Evaluation Plan." http://www.itl.nist.gov/iad/mig//tests/sre/2010/NIST_SRE10_evalplan.r6.pdf. 14

[26] "The NIST Year 2008 Speaker Recognition Evaluation Plan." http://www.itl.nist.gov/iad/mig//tests/sre/2008/sre08_evalplan_release4.pdf. 14

[27] L. D. Brown, T. T. Cai, and A. Dasgupta, "Interval Estimation for a Binomial Proportion," *Statistical Science*, vol. 16, pp. 101–133, 2001. 15

[28] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds, "SHEEP, GOATS, LAMBS and WOLVES A Statistical Analysis of Speaker Performance in the NIST 1998 Speaker Recognition Evaluation," in *INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING*, 1998. 15

[29] J. C. Wu, "Studies of Operational Measurement of ROC Curve on Large Fingerprint Data Sets Using Two-Sample Bootstrap." http://www.nist.gov/customcf/get_pdf.cfm?pub_id=51242, 2007. 15

[30] "Interagency Advisory Board Meeting Agenda." http://fips201.com/resources/audio/iab_1111/iab_110111_nagel.pdf. 16

[31] "DoD to use iris scans, fingerprints for building security." http://www.federalnewsradio.com/396/2817314/DoD-to-use-iris-scans-fingerprints-for-building-security. 16

[32] "Cape May Country, New Jersey: Sheriff's Office Implements New Iris Recognition System." http://www.capemaycountygov.net/cit-e-access/photojournal/photopages.cfm?MID=1159&TID=5&TPID=1715. 16

[33] "Visitors for Inmates Entrance Procedures." http://www.hampshiresheriffs.com/index.php?option=com_content&task=view&id=23&Itemid=52. 16

[34] "Iris Recognition for Inmates." http://www.tarrantcounty.com/esheriff/cwp/view.asp?a=792&q=437580. 16

[35] "Biometrics in Prisons, Information for visitors." http://www.correctiveservices.wa.gov.au/_files/prisons/visiting-prisons/biometrics-brochure.pdf. 16

[36] "Iris recognition immigration system (IRIS)." http://www.ukba.homeoffice.gov.uk/customs-travel/Enteringtheuk/usingiris/. 16, 29

[37] J. Daugman, "Evolving Methods in Iris Recognition," in *Biometrics: Theory, Applications and Systems*, 2007. 16

[38] J. Daugman, "Iris Recognition: Algorithms, Performance, and Challenges," in *Biometrics Consortium Conference*, 2007. 16

[39] *ISO/IEC 29794-6 - Biometric Sample Quality - Framework*. 2009. 17

[40] J. Daugman and I. Malhas, "IRIS RECOGNITION BORDER-CROSSING SYSTEM IN THE UAE," *International Airport Review*, vol. 8, no. 2, pp. 49–53, 2004. 24

[41] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1991. 24

[42] "Face recognition as a search tool "Foto-Fahndung"," tech. rep., Bundeskasse Trier, 2007. 24

[43] J. Powell, "Enhancing Marine Corps Operations by Shortening the Intelligence and Targeting Cycle," 2010. 27

[44] "US army amasses biometric data in Afghanistan." http://www.guardian.co.uk/world/2010/oct/27/us-army-biometric-data-afghanistan. 27

[45] "CLEAR." http://clearme.com/. 29

[46] P. G. Elham Tabassi, George W. Quinn, "When to Fuse Two Biometrics, booktitle = IEEE Computer Society on Computer Vision and Pattern Recognition, Workshop on Multi-Biometrics," 2006. 48

[47] A. Hicklin, B. Ulery, and C. Watson, "A Brief Introduction to Biometric Fusion." http://www.noblis.org/MissionAreas/nsi/ThoughtLeadership/IdentityDiscovery_Management/Documents/IntroFusion_2006-06-16.pdf, 2006. 48

[48] *ISO/IEC 7816-15 - Cryptographic Information Application*. 2008. 50

[49] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 Large-Scale Results." http://www.frvt.org/FRVT2006/docs/FRVT2006andICE2006LargeScaleReport.pdf, last accessed October 2007. 54

A Additional Figures and Tables

Appendix A contains supplementary DET plots and additional summary statistics for all recognition algorithms.

A.1 Single-eye Positive Identification

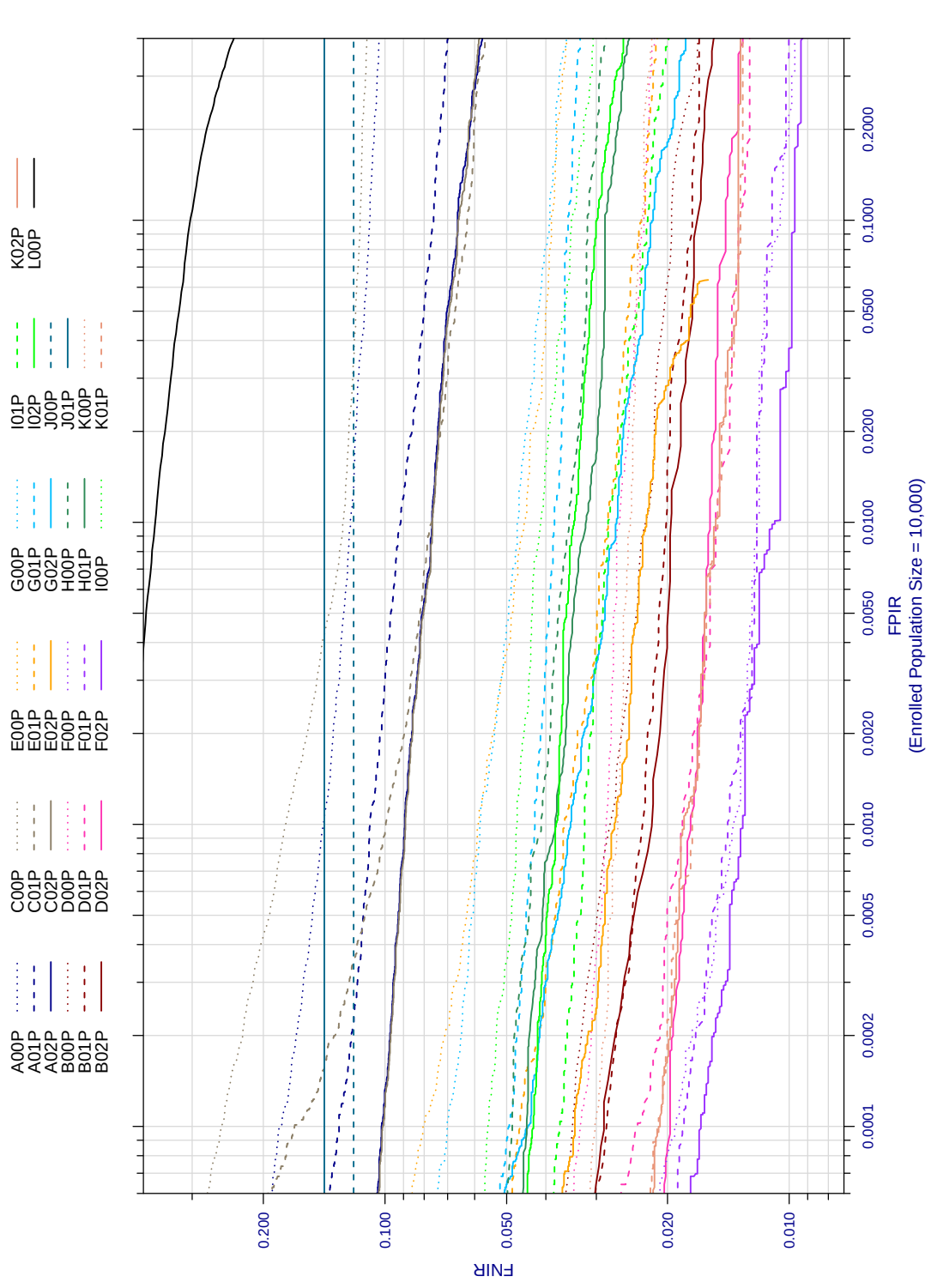


Figure 29: DET curves for single-eye searches for all Class P algorithms against an enrolled population of 10 thousand. Plots were generated using 5 988 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

	FNIR @ FPIR=0.001	FNIR @ FPIR=0.0001	Minimum Cost	Cost of Alg Decision
A00P	0.142 ± 0.005	0.182 ± 0.005	0.0024	0.023
A01P	0.109 ± 0.004	0.130 ± 0.004	0.0017	0.032
A02P	0.089 ± 0.004	0.102 ± 0.004	0.0012	0.012
B00P	0.027 ± 0.002	0.034 ± 0.002	0.00049	0.36
B01P	0.023 ± 0.002	0.028 ± 0.002	0.0004	0.67
B02P	0.022 ± 0.002	0.029 ± 0.002	0.0004	0.00041
C00P	0.176 ± 0.005	0.254 ± 0.006	0.0033	0.08
C01P	0.100 ± 0.004	0.167 ± 0.005	0.0025	0.066
C02P	0.089 ± 0.004	0.101 ± 0.004	0.0012	0.011
D00P	0.028 ± 0.002	0.033 ± 0.002	0.00043	0.40
D01P	0.018 ± 0.002	0.024 ± 0.002	0.00039	0.19
D02P	0.018 ± 0.002	0.020 ± 0.002	0.00029	9.90
E00P	0.059 ± 0.003	0.080 ± 0.004	0.0012	0.0017
E01P	0.036 ± 0.002	0.046 ± 0.003	0.00078	0.0017
E02P	0.027 ± 0.002	0.034 ± 0.002	0.00050	0.0016
F00P	0.014 ± 0.002	0.019 ± 0.002	0.00039	0.012
F01P	0.014 ± 0.002	0.018 ± 0.002	0.00027	0.00027
F02P	0.013 ± 0.001	0.017 ± 0.002	0.00035	0.00025
G00P	0.058 ± 0.003	0.070 ± 0.003	0.00099	0.0017
G01P	0.043 ± 0.003	0.049 ± 0.003	0.00084	0.0016
G02P	0.035 ± 0.002	0.044 ± 0.003	0.00083	0.0015
H00P	0.041 ± 0.003	0.048 ± 0.003	0.00064	0.00097
H01P	0.038 ± 0.002	0.045 ± 0.003	0.00057	0.0013
I00P	0.047 ± 0.003	0.056 ± 0.003	0.00071	0.010
I01P	0.032 ± 0.002	0.037 ± 0.002	0.00051	0.00052
I02P	0.038 ± 0.002	0.043 ± 0.003	0.00056	0.00057
J00P	0.120 ± 0.004	0.120 ± 0.004	0.0012	0.0012
J01P	0.141 ± 0.005	0.141 ± 0.005	0.0014	0.0014
K00P	0.027 ± 0.002	0.030 ± 0.002	0.00044	0.00046
K01P	0.017 ± 0.002	0.021 ± 0.002	0.00035	0.00036
K02P	0.018 ± 0.002	0.021 ± 0.002	0.00031	0.00033
L00P	0.433 ± 0.006	0.570 ± 0.006	0.0067	3.65

Table 6: Summary of algorithm performance.
(Number of eyes = 1, Enrolled Population Size = 10,000)

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

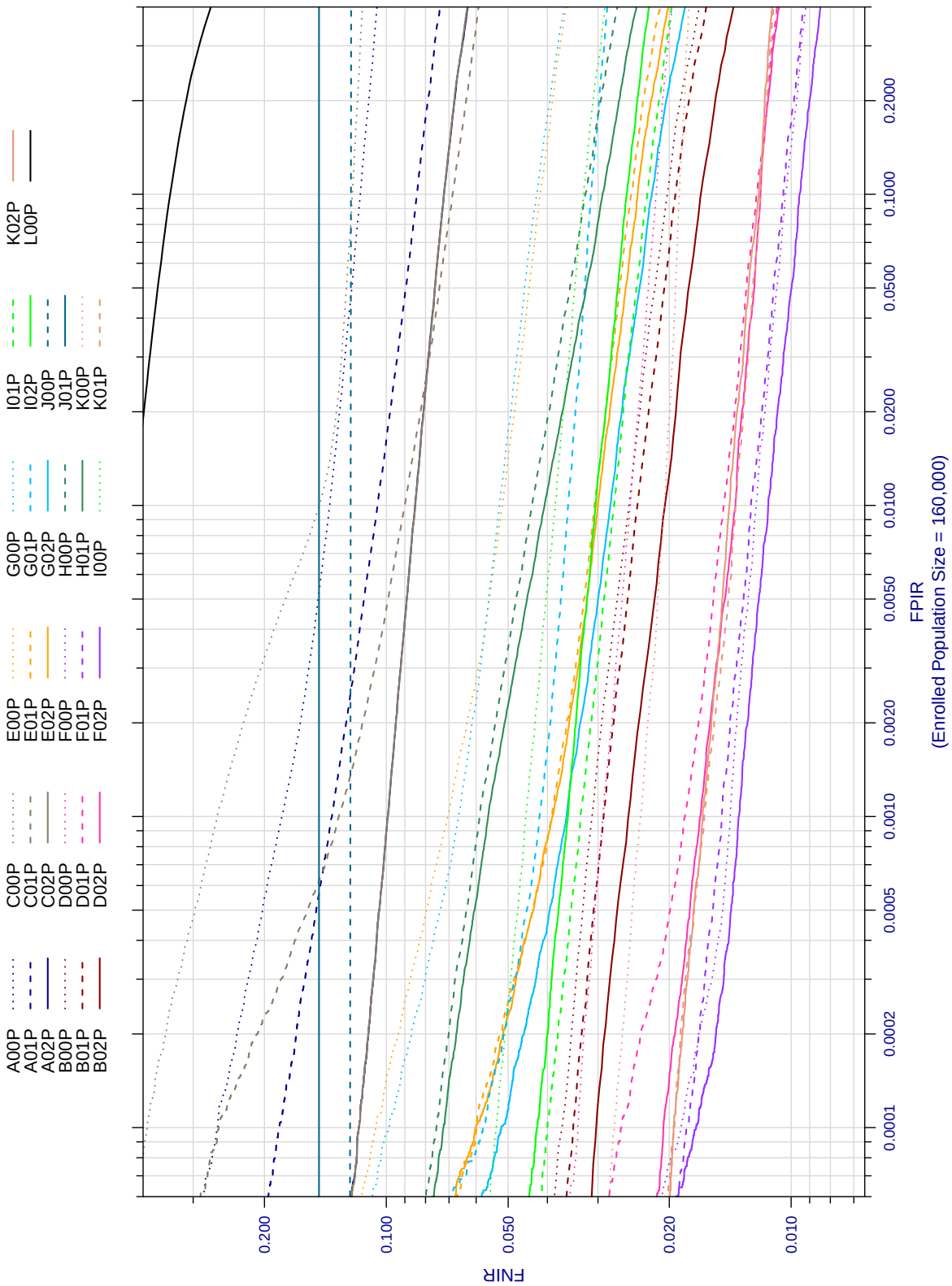


Figure 30: DET curves for single-eye searches for all Class P algorithms against an enrolled population of 160 thousand. Plots were generated using 96 636 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

	FNIR @ FPIR=0.001	FNIR @ FPIR=0.0001	Minimum Cost	Cost of Alg Decision
A00P	0.197 ± 0.001	0.568 ± 0.002	0.0054	0.36
A01P	0.136 ± 0.001	0.184 ± 0.001	0.0024	0.45
A02P	0.0994 ± 0.001	0.117 ± 0.001	0.0015	0.16
B00P	0.0296 ± 0.0005	0.619 ± 0.002	0.0035	2.80
B01P	0.0257 ± 0.0005	0.0312 ± 0.0006	0.00043	3.99
B02P	0.0250 ± 0.0005	0.0301 ± 0.0005	0.00041	0.00081
C00P	0.268 ± 0.001	0.423 ± 0.002	0.0051	0.44
C01P	0.131 ± 0.001	0.259 ± 0.001	0.0034	0.59
C02P	0.0992 ± 0.001	0.117 ± 0.001	0.0015	0.16
D00P	0.0275 ± 0.0005	0.0329 ± 0.0006	0.00048	0.084
D01P	0.0175 ± 0.0004	0.0252 ± 0.0005	0.00059	0.10
D02P	0.0167 ± 0.0004	0.0207 ± 0.0005	0.0003	9.90
E00P	0.0689 ± 0.0008	0.109 ± 0.001	0.0016	0.0018
E01P	0.0391 ± 0.0006	0.0599 ± 0.0008	0.0010	0.0016
E02P	0.0388 ± 0.0006	0.0601 ± 0.0008	0.0011	0.0015
F00P	0.0142 ± 0.0004	0.0187 ± 0.0004	0.00044	0.035
F01P	0.0148 ± 0.0004	0.0180 ± 0.0004	0.00027	0.00029
F02P	0.0135 ± 0.0004	0.0172 ± 0.0004	0.00036	0.00026
G00P	0.0640 ± 0.0008	0.0972 ± 0.001	0.0015	0.0018
G01P	0.0419 ± 0.0006	0.0577 ± 0.0008	0.0012	0.0017
G02P	0.0359 ± 0.0006	0.0519 ± 0.0007	0.001	0.0015
H00P	0.0588 ± 0.0008	0.0751 ± 0.0008	0.0011	0.0012
H01P	0.0560 ± 0.0007	0.0721 ± 0.0008	0.0011	0.0018
I00P	0.0452 ± 0.0007	0.0535 ± 0.0007	0.00072	0.15
I01P	0.0331 ± 0.0006	0.0400 ± 0.0006	0.00054	0.00054
I02P	0.0356 ± 0.0006	0.0423 ± 0.0006	0.00056	0.00059
J00P	0.122 ± 0.001	0.122 ± 0.001	0.0013	0.0013
J01P	0.147 ± 0.001	0.147 ± 0.001	0.0015	0.0015
K00P	0.0235 ± 0.0005	0.0274 ± 0.0005	0.00037	0.00038
K01P	0.0164 ± 0.0004	0.0194 ± 0.0004	0.00027	0.00030
K02P	0.0165 ± 0.0004	0.0195 ± 0.0004	0.00028	0.00028
L00P	0.500 ± 0.002	0.674 ± 0.002	0.0076	6.19

Table 7: Summary of algorithm performance.
(Number of eyes = 1, Enrolled Population Size = 160,000)

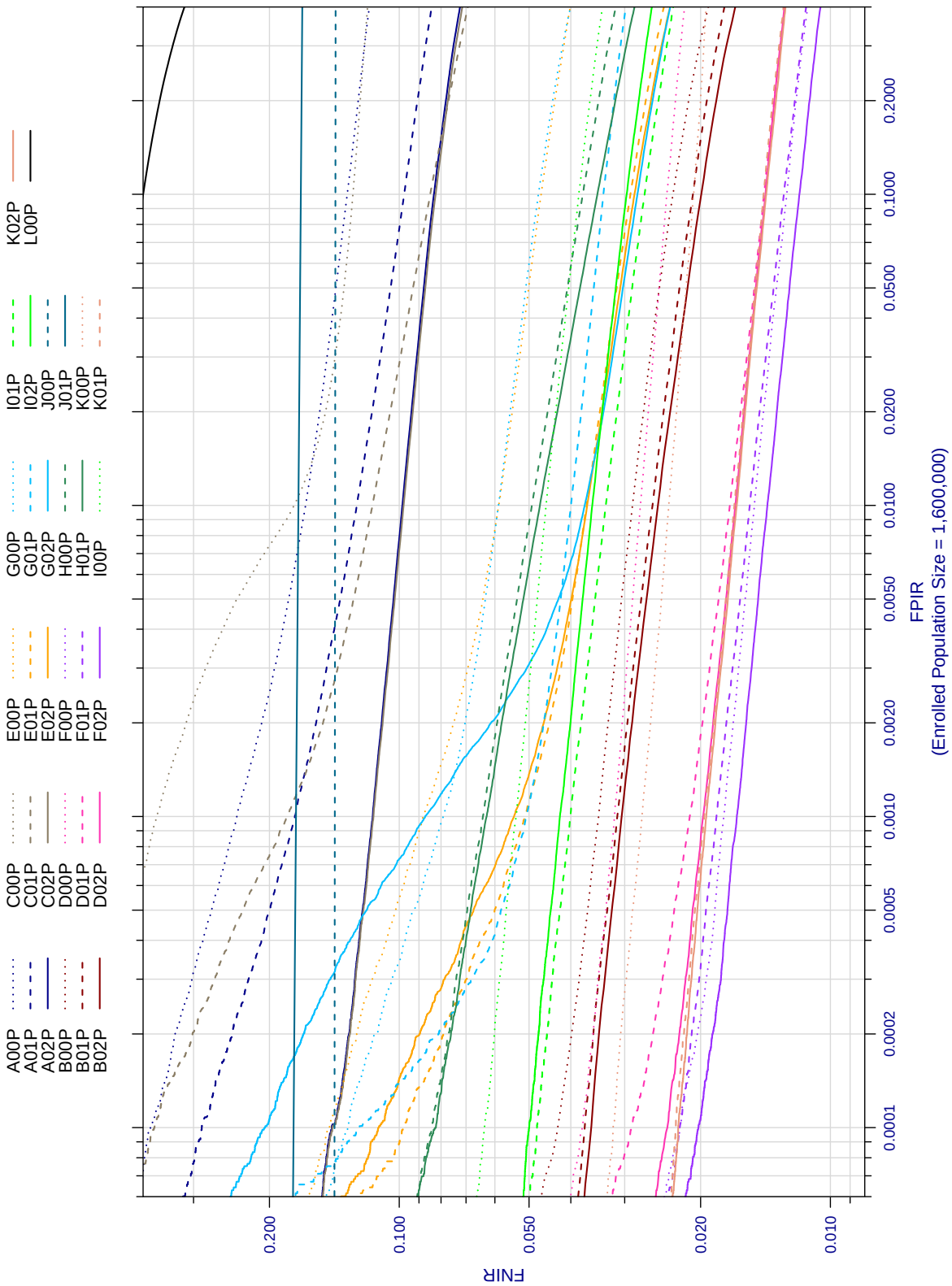


Figure 31: DET curves for single-eye searches for all Class P algorithms against an enrolled population of 1.6 million. Plots were generated using 962 720 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

	FNIR @ FPIR=0.001	FNIR @ FPIR=0.0001	Minimum Cost	Cost of Alg Decision
A00P	0.2379 ± 0.0004	0.3773 ± 0.0005	0.0045	2.01
A01P	0.1746 ± 0.0004	0.2871 ± 0.0005	0.0037	2.06
A02P	0.1152 ± 0.0003	0.1434 ± 0.0004	0.0020	1.12
B00P	0.0339 ± 0.0002	0.0425 ± 0.0002	0.00085	6.52
B01P	0.0309 ± 0.0002	0.0369 ± 0.0002	0.00052	7.71
B02P	0.0302 ± 0.0002	0.0358 ± 0.0002	0.0005	0.0063
C00P	0.3651 ± 0.0005	0.5318 ± 0.0005	0.0062	1.85
C01P	0.1818 ± 0.0004	0.3692 ± 0.0005	0.0046	2.90
C02P	0.1147 ± 0.0003	0.1428 ± 0.0004	0.0020	1.10
D00P	0.0316 ± 0.0002	0.0376 ± 0.0002	0.00075	0.056
D01P	0.0211 ± 0.0001	0.0288 ± 0.0002	0.0007	0.092
D02P	0.0197 ± 0.0001	0.0238 ± 0.0002	0.00049	9.90
E00P	0.0880 ± 0.0003	0.1451 ± 0.0004	0.0021	0.0020
E01P	0.0511 ± 0.0002	0.0972 ± 0.0003	0.0018	0.0017
E02P	0.0541 ± 0.0002	0.1114 ± 0.0003	0.0019	0.0018
F00P	0.0177 ± 0.0001	0.0223 ± 0.0002	0.00051	0.095
F01P	0.0184 ± 0.0001	0.0223 ± 0.0002	0.00036	0.00038
F02P	0.0165 ± 0.0001	0.0202 ± 0.0001	0.00041	0.00038
G00P	0.0790 ± 0.0003	0.1316 ± 0.0003	0.0019	0.0019
G01P	0.0508 ± 0.0002	0.1235 ± 0.0003	0.0022	0.0018
G02P	0.0870 ± 0.0003	0.2044 ± 0.0004	0.003	0.0025
H00P	0.0638 ± 0.0002	0.0839 ± 0.0003	0.0013	0.0016
H01P	0.0627 ± 0.0002	0.0821 ± 0.0003	0.0013	0.0024
I00P	0.0536 ± 0.0002	0.0634 ± 0.0002	0.00079	1.06
I01P	0.0401 ± 0.0002	0.0479 ± 0.0002	0.00062	0.00063
I02P	0.0424 ± 0.0002	0.0499 ± 0.0002	0.00063	0.00068
J00P	0.1404 ± 0.0004	0.1404 ± 0.0004	0.0015	0.0015
J01P	0.1670 ± 0.0004	0.1670 ± 0.0004	0.0018	0.0017
K00P	0.0273 ± 0.0002	0.0319 ± 0.0002	0.00042	0.01
K01P	0.0196 ± 0.0001	0.0228 ± 0.0002	0.00035	0.00041
K02P	0.0194 ± 0.0001	0.0225 ± 0.0002	0.00034	0.00036
L00P	0.6858 ± 0.0005	0.8667 ± 0.0003	0.0094	7.65

Table 8: Summary of algorithm performance.
(Number of eyes = 1, Enrolled Population Size = 1,600,000)

A.2 Two-eye Positive Identification

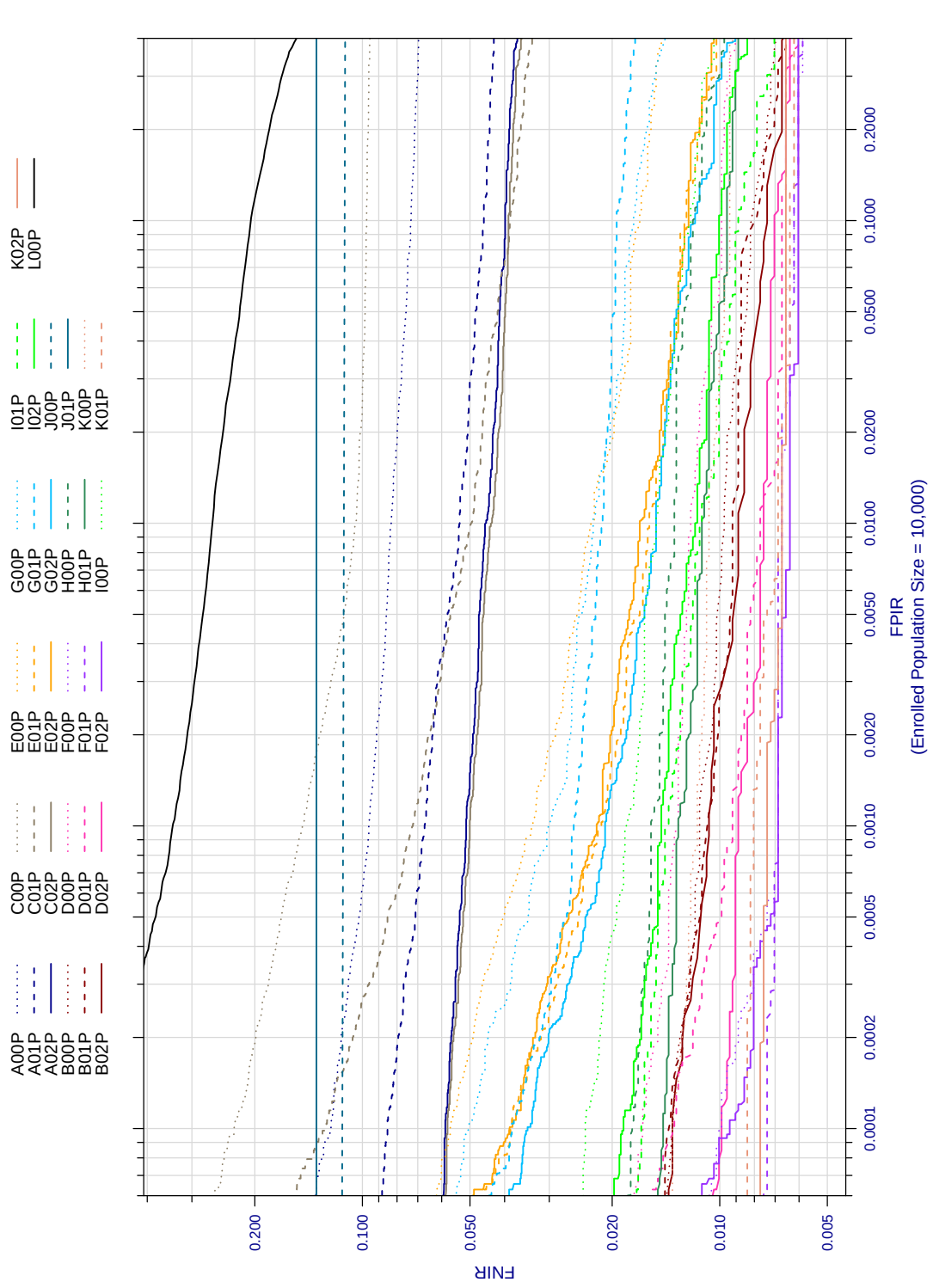


Figure 32: DET curves for two-eye searches for all Class P algorithms against an enrolled population of 10 thousand. Plots were generated using 5 972 mated and 365 716 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

	FNIR @ FPIR=0.001	FNIR @ FPIR=0.0001	Minimum Cost	Cost of Alg Decision
A00P	0.096 ± 0.004	0.122 ± 0.004	0.0018	0.023
A01P	0.067 ± 0.003	0.085 ± 0.004	0.0012	0.032
A02P	0.051 ± 0.003	0.059 ± 0.003	0.00075	0.012
B00P	0.011 ± 0.001	0.013 ± 0.001	0.00024	0.36
B01P	0.011 ± 0.001	0.014 ± 0.002	0.00021	0.67
B02P	0.011 ± 0.001	0.014 ± 0.001	0.00020	0.00041
C00P	0.147 ± 0.005	0.228 ± 0.005	0.0031	0.08
C01P	0.072 ± 0.003	0.127 ± 0.004	0.0021	0.066
C02P	0.050 ± 0.003	0.058 ± 0.003	0.00074	0.011
D00P	0.013 ± 0.001	0.016 ± 0.002	0.00025	0.40
D01P	0.0092 ± 0.001	0.013 ± 0.001	0.00028	0.19
D02P	0.0089 ± 0.001	0.0099 ± 0.001	0.00016	9.90
E00P	0.033 ± 0.002	0.057 ± 0.003	0.00087	0.0017
E01P	0.021 ± 0.002	0.039 ± 0.002	0.0008	0.0017
E02P	0.022 ± 0.002	0.039 ± 0.002	0.00078	0.0016
F00P	0.0069 ± 0.001	0.010 ± 0.001	0.00034	0.012
F01P	0.0069 ± 0.001	0.0074 ± 0.001	0.00012	0.00027
F02P	0.0069 ± 0.001	0.0094 ± 0.001	0.00033	0.00025
G00P	0.030 ± 0.002	0.051 ± 0.003	0.00082	0.0017
G01P	0.026 ± 0.002	0.038 ± 0.002	0.0008	0.0016
G02P	0.021 ± 0.002	0.035 ± 0.002	0.00074	0.0015
H00P	0.016 ± 0.002	0.018 ± 0.002	0.00027	0.00097
H01P	0.013 ± 0.001	0.014 ± 0.002	0.00026	0.0013
I00P	0.018 ± 0.002	0.024 ± 0.002	0.00032	0.010
I01P	0.014 ± 0.002	0.017 ± 0.002	0.00025	0.00052
I02P	0.015 ± 0.002	0.019 ± 0.002	0.00026	0.00057
J00P	0.112 ± 0.004	0.112 ± 0.004	0.0012	0.0012
J01P	0.134 ± 0.004	0.134 ± 0.004	0.0013	0.0014
K00P	0.012 ± 0.001	0.014 ± 0.001	0.00018	0.00046
K01P	0.008 ± 0.001	0.0084 ± 0.001	0.00012	0.00036
K02P	0.0074 ± 0.001	0.0077 ± 0.001	0.00012	0.00033
L00P	0.339 ± 0.006	0.534 ± 0.006	0.0062	3.65

Table 9: Summary of algorithm performance.
(Number of eyes = 2, Enrolled Population Size = 10,000)

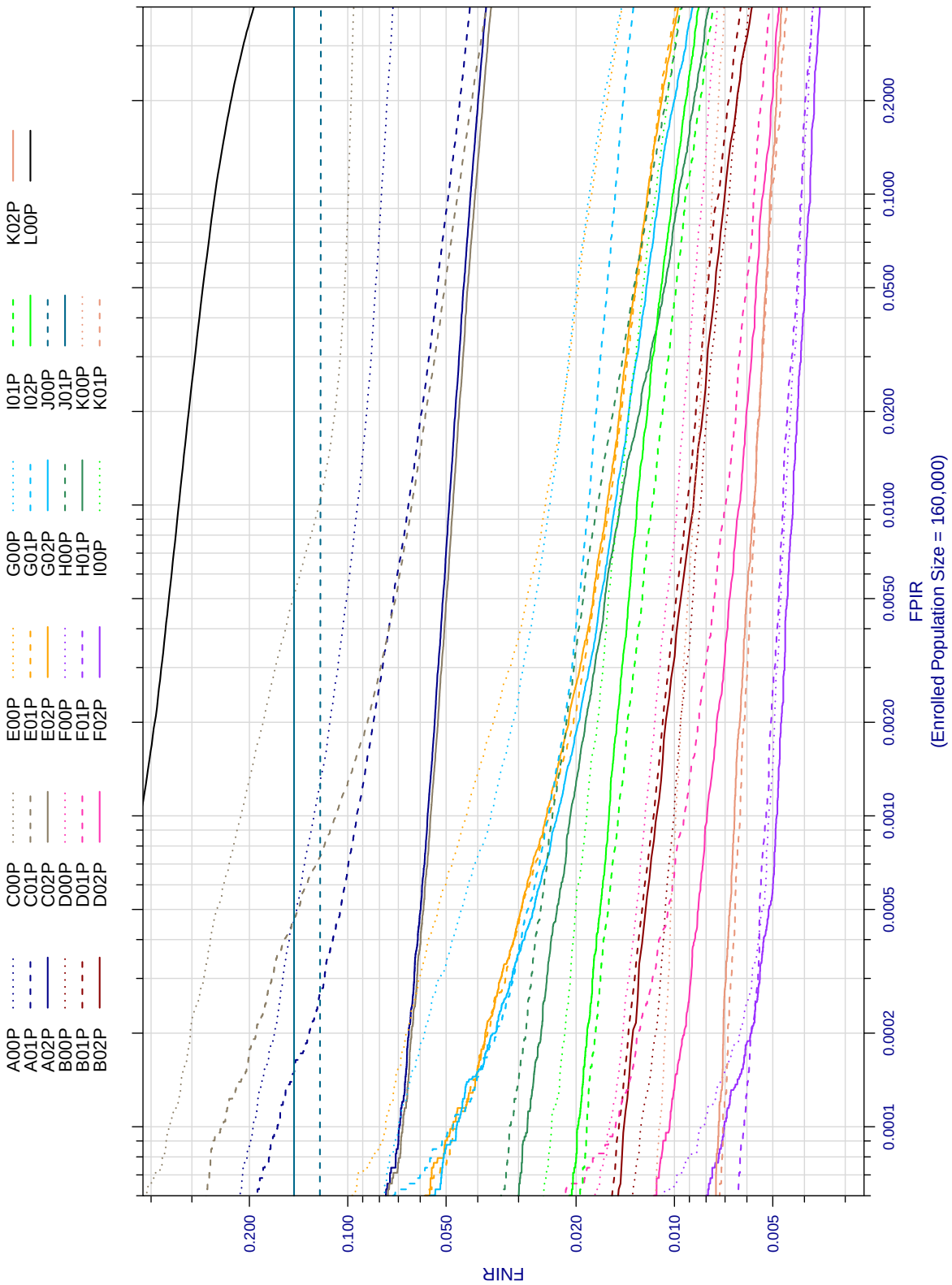


Figure 33: DET curves for two-eye searches for all Class P algorithms against an enrolled population of 160 thousand. Plots were generated using 96 076 mated and 365 716 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

	FNIR @ FPIR=0.001	FNIR @ FPIR=0.0001	Minimum Cost	Cost of Alg Decision
A00P	0.127 ± 0.001	0.195 ± 0.001	0.0027	0.36
A01P	0.0938 ± 0.001	0.164 ± 0.001	0.0024	0.45
A02P	0.0570 ± 0.0007	0.0695 ± 0.0008	0.0011	0.16
B00P	0.0100 ± 0.0003	0.0126 ± 0.0004	0.00022	2.80
B01P	0.0116 ± 0.0003	0.0147 ± 0.0004	0.00026	3.99
B02P	0.0111 ± 0.0003	0.0141 ± 0.0004	0.00026	0.00081
C00P	0.212 ± 0.001	0.351 ± 0.002	0.0044	0.44
C01P	0.109 ± 0.001	0.241 ± 0.001	0.0033	0.59
C02P	0.0558 ± 0.0007	0.0679 ± 0.0008	0.0011	0.16
D00P	0.0123 ± 0.0004	0.0158 ± 0.0004	0.00058	0.084
D01P	0.0093 ± 0.0003	0.0151 ± 0.0004	0.0007	0.10
D02P	0.0079 ± 0.0003	0.0103 ± 0.0003	0.00042	9.90
E00P	0.0433 ± 0.0007	0.0765 ± 0.0009	0.0014	0.0018
E01P	0.0242 ± 0.0005	0.0464 ± 0.0007	0.0010	0.0016
E02P	0.0246 ± 0.0005	0.0478 ± 0.0007	0.0010	0.0015
F00P	0.0051 ± 0.0002	0.0083 ± 0.0003	0.00039	0.035
F01P	0.0052 ± 0.0002	0.0061 ± 0.0003	0.00011	0.00029
F02P	0.0049 ± 0.0002	0.0070 ± 0.0003	0.00027	0.00026
G00P	0.0365 ± 0.0006	0.0698 ± 0.0008	0.0013	0.0018
G01P	0.0243 ± 0.0005	0.0476 ± 0.0007	0.0013	0.0017
G02P	0.0229 ± 0.0005	0.0464 ± 0.0007	0.00095	0.0015
H00P	0.0236 ± 0.0005	0.0319 ± 0.0006	0.00047	0.0012
H01P	0.0208 ± 0.0005	0.0282 ± 0.0005	0.00042	0.0018
I00P	0.0186 ± 0.0004	0.0238 ± 0.0005	0.00033	0.15
I01P	0.0148 ± 0.0004	0.0186 ± 0.0004	0.00028	0.00054
I02P	0.0159 ± 0.0004	0.0198 ± 0.0004	0.00032	0.00059
J00P	0.121 ± 0.001	0.121 ± 0.001	0.0013	0.0013
J01P	0.146 ± 0.001	0.146 ± 0.001	0.0015	0.0015
K00P	0.0099 ± 0.0003	0.0111 ± 0.0003	0.00017	0.00038
K01P	0.0063 ± 0.0003	0.0071 ± 0.0003	0.00011	0.00030
K02P	0.0065 ± 0.0003	0.0073 ± 0.0003	0.00012	0.00028
L00P	0.430 ± 0.002	0.675 ± 0.002	0.0077	6.19

Table 10: Summary of algorithm performance.
(Number of eyes = 2, Enrolled Population Size = 160,000)

A.3 Single-eye Negative Identification

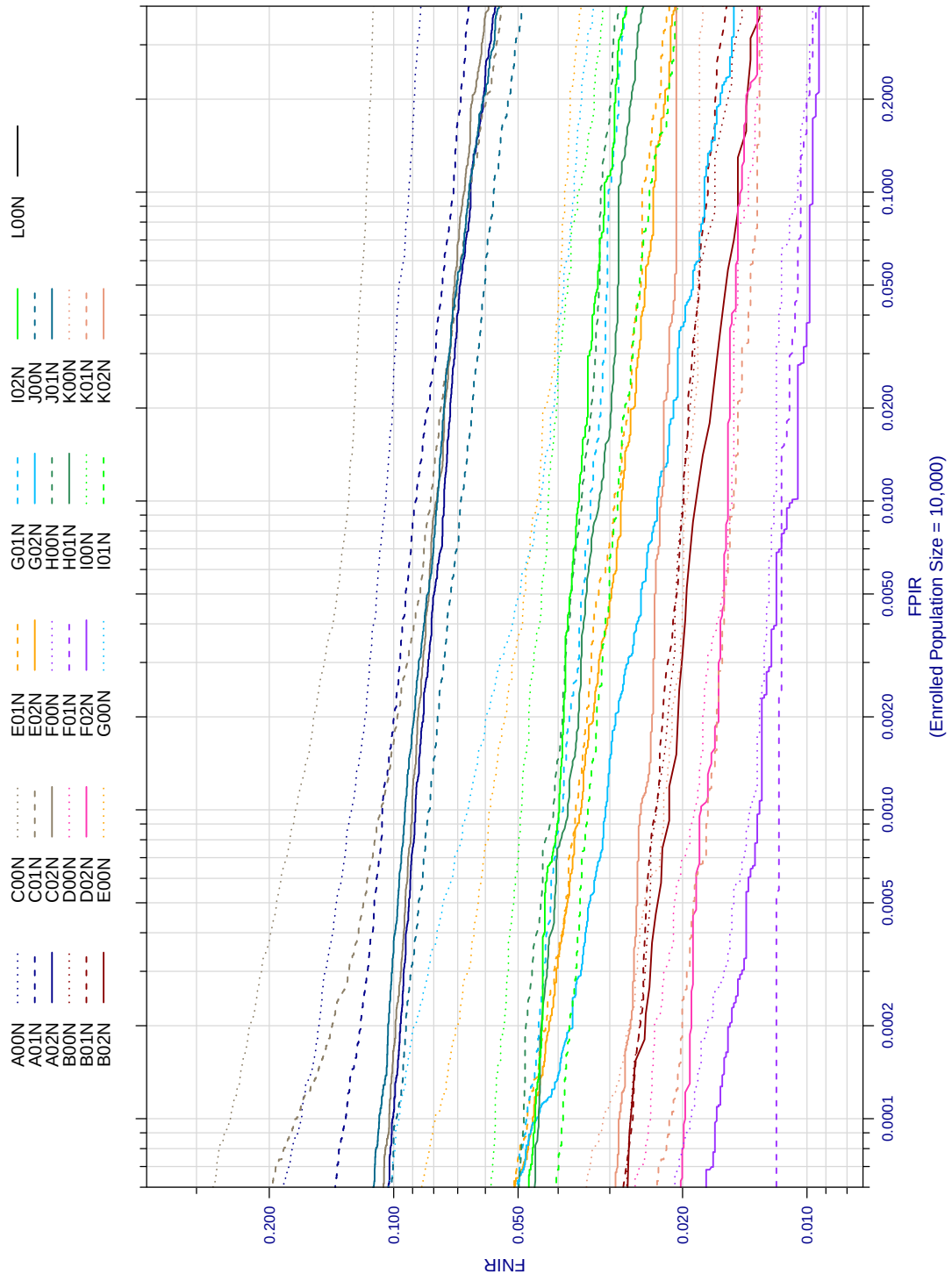


Figure 34: DET curves for single-eye searches for all Class N algorithms against an enrolled population of 10 thousand. Plots were generated using 5 988 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

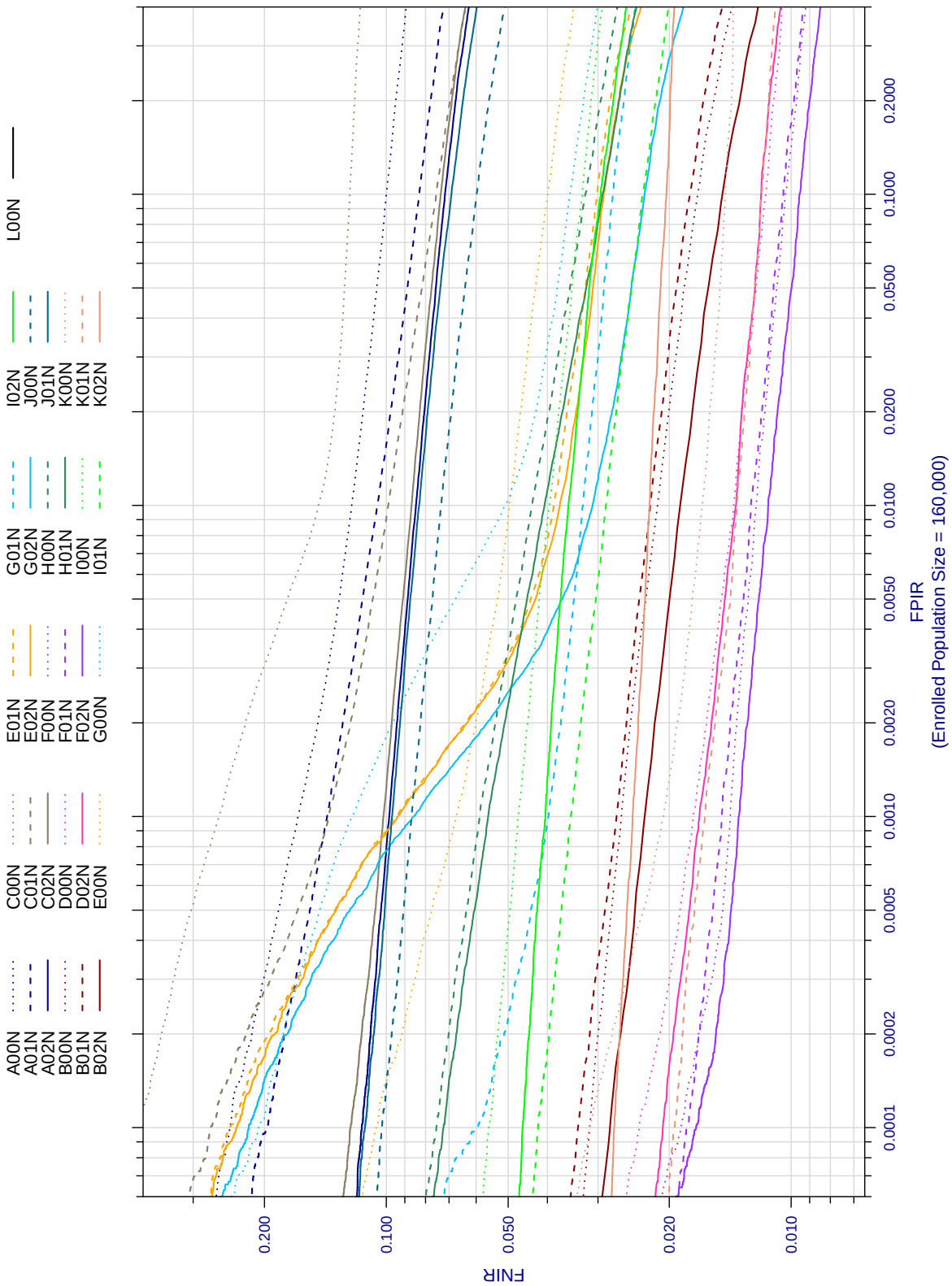


Figure 35: DET curves for single-eye searches for all Class N algorithms against an enrolled population of 160 thousand. Plots were generated using 96 636 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

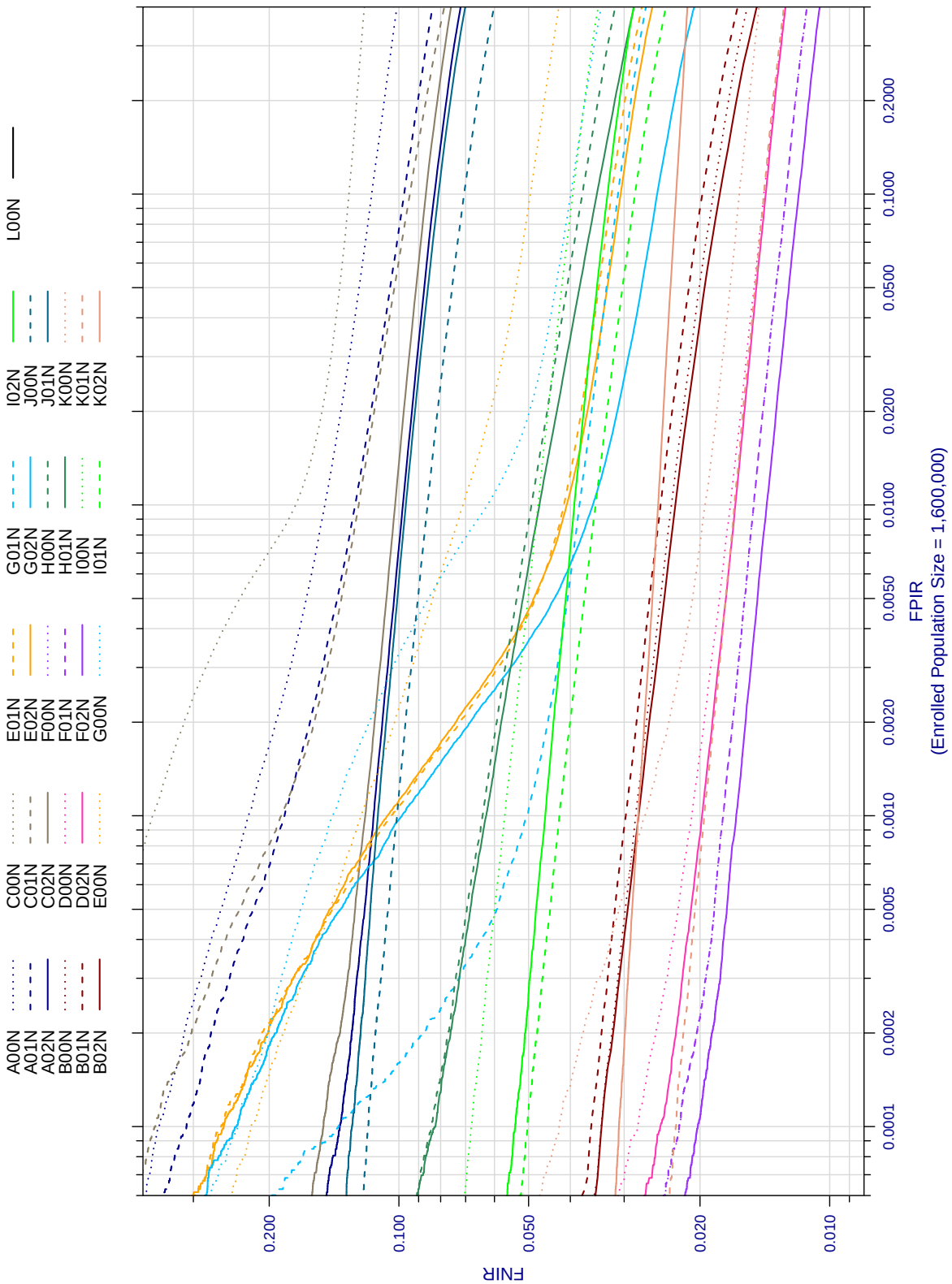


Figure 36: DET curves for single-eye searches for all Class N algorithms against an enrolled population of 1.6 million. Plots were generated using 962 720 mated and 731 432 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

A.4 Two-eye Negative Identification

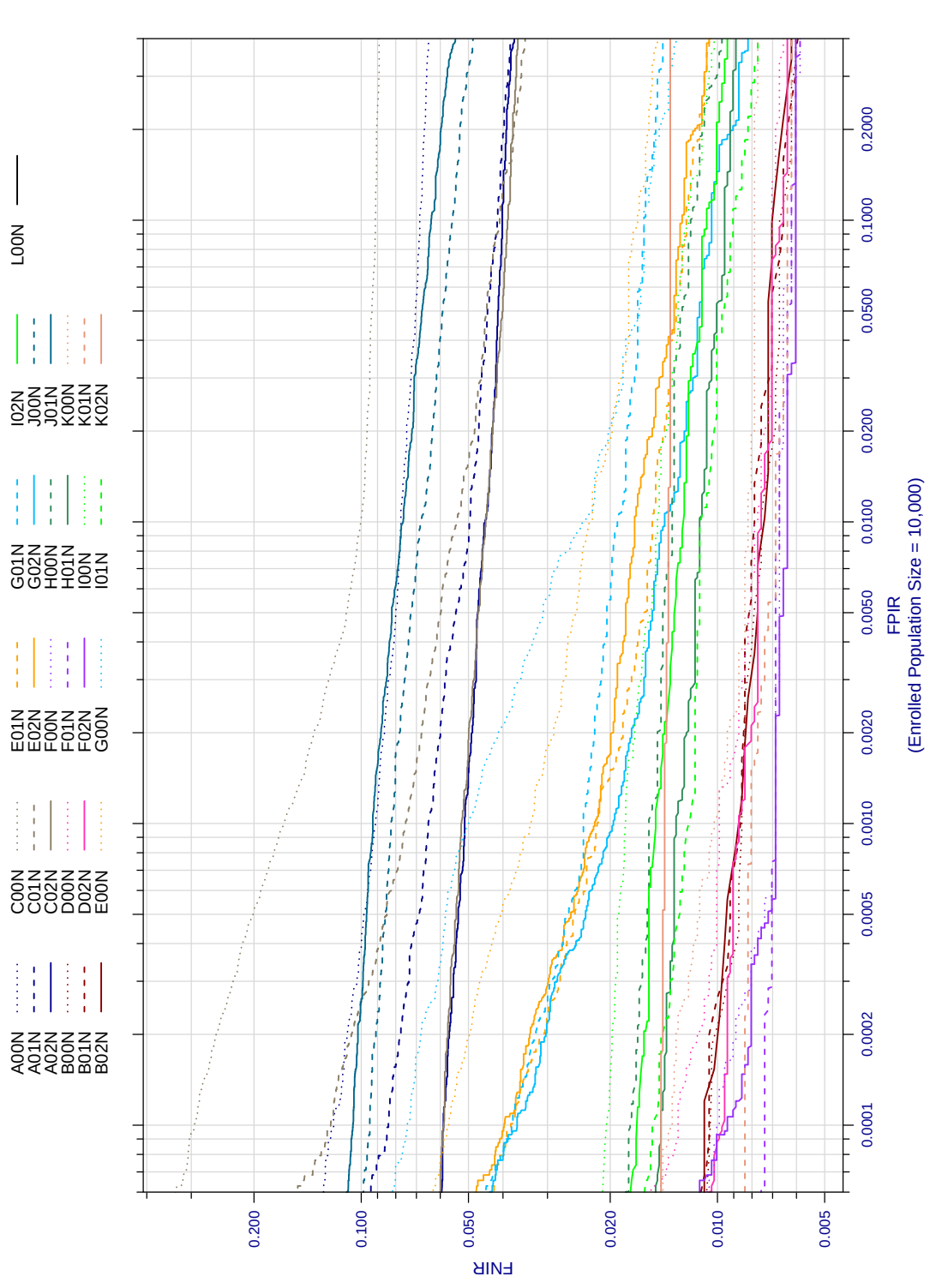


Figure 37: DET curves for two-eye searches for all Class N algorithms against an enrolled population of 10 thousand. Plots were generated using 5 972 mated and 365 716 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

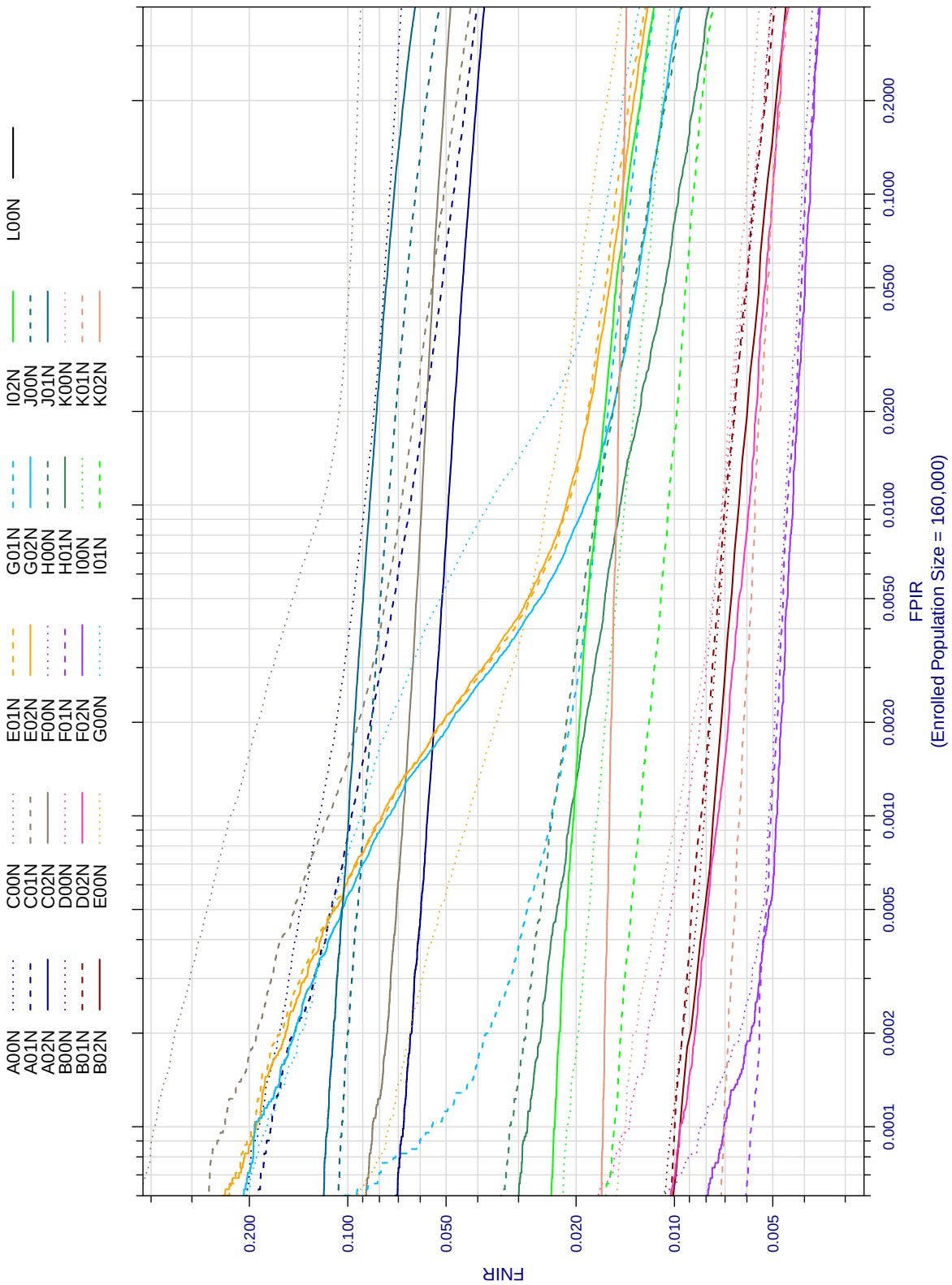


Figure 38: DET curves for two-eye searches for all Class N algorithms against an enrolled population of 160 thousand. Plots were generated using 96 076 mated and 365 716 non-mated searches.

A = U. of Bath	B = Neurotechnology	C = Smart Sensors	D = 3M Cogent	E = IriTech	F = MorphoTrust
G = IsciLab	H = DeltaID	I = U. of Cambridge	J = Iris ID	K = Morpho	L = Nihon Systems

B Glossary

ABIS Automated Biometric Identification System 2

ANSI/NIST-ITL 1-2011 US standard defining interchange formats for biometric information including iris data. 3

API Application Programming Interface 10

BIMA Biometric Identity Management Agency 19, 26

CONOPS Concept of Operations 10

DET Detection Error Trade-off. An accuracy curve for iris recognition algorithms that plots FNIR as a function of FPIR. 13, 15–17, 19, 22–27, 39–41, 48, 50–54, 57, 59, 61, 63, 65, 67–71

DHS Department of Homeland Security 23

DOD Department of Defense 2, 19

FBI Federal Bureau of Investigation 2

FNIR The False Negative Identification Rate, an accuracy statistic for iris recognition algorithms defined as the fraction of mated searches for which the mate was missed. 12–15, 18, 19, 21, 24, 26, 29, 30, 32–35, 39, 41, 44, 46, 47, 49, 51, 53

FPIR The False Positive Identification Rate, an accuracy statistic for iris recognition algorithms defined as the fraction of non-mated searches that return at least one false positive (i.e. a candidate having a dissimilarity score at or below the decision threshold. 12–16, 18, 19, 21, 23, 24, 26, 29, 30, 32–35, 39, 41, 44, 47, 49, 51, 53

FTX Failure to extract features, occurs when an iris recognition algorithm reports that it is unable to generate a matchable template from the iris image(s). 24, 26

HIIDE Handheld Interagency Identity Detection Equipment 7, 19

IREX Iris Exchange 2, 3, 7, 8, 10–12, 16, 19, 31, 41, 46, 53, 54

ISO/IEC 19794-6 International standard defining interchange formats for iris image data. 3

JP2 Joint Photographic Experts Group (JPEG) 2000 image compression standard and coding system. 3

NIST National Institute of Standards and Technology 2, 8, 10, 11, 13, 51

PNG Portable Network Graphics, an image file format. 3

UAE United Arab Emirates 2, 7, 9, 23

UID Unique Identity, a scheme initiated by the Indian government. 7, 19, 26

UIDAI Unique Identity Authority of India 2, 16