IREX I

Performance of Iris Recognition Algorithms on Standard Images

NIST Interagency Report 7629

P. Grother, E. Tabassi, G. W. Quinn, W. Salamon

Information Access Division

National Institute of Standards and Technology

AUTHENTICATED
U.S. GOVERNMENT
INFORMATION
GPO

October 30, 2009

Acknowledgments

The authors would like to thank the Department of Homeland Security's U. S. VISIT Office for supporting this work. In addition, the authors are grateful to the Department of Homeland Security's Science and Technology Directorate for supporting the ongoing standardization aspects of this activity. We extend our thanks to the U.S. Department of Defense and the University of Bath for providing iris imagery to the project.

Finally, thanks go IREX participants for their contributions toward the ISO/IEC 19794-6 standard and the development of the IREX test specification[1].

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.

[1] The formal CONOPS and API specification is available at http://iris.nist.gov/irex/irex_api.pdf.

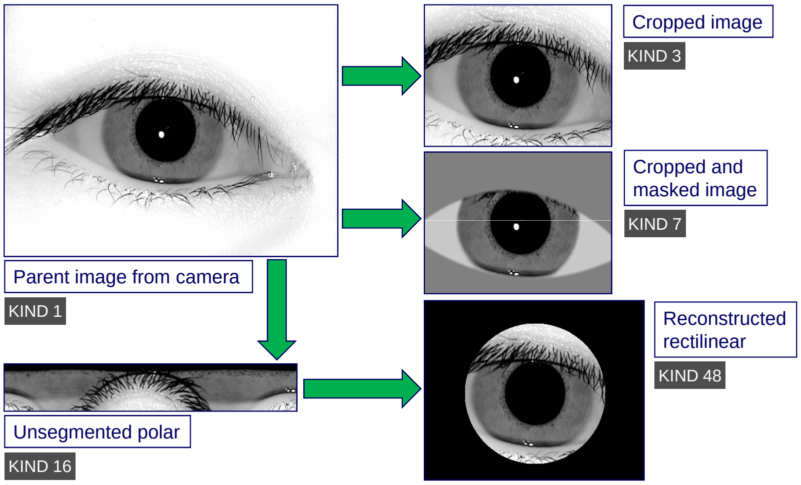
EXECUTIVE SUMMARY

The Iris Exchange (IREX) was initiated by NIST in late 2007 to support interoperable exchange of iris imagery in high performance biometric applications. The first activity in the program, the IREX I evaluation, was conducted in cooperation with the iris recognition industry to develop and test standard image formats, and to demonstrate that iris recognition algorithms can maintain their accuracy and interoperability with compact images. Standard formats are needed in federated applications in which iris data is exchanged between interoperating systems. Compact size is a current and vital requirement for applications in which imagery is passed across bandwidth-limited networks, or stored on identity credentials.

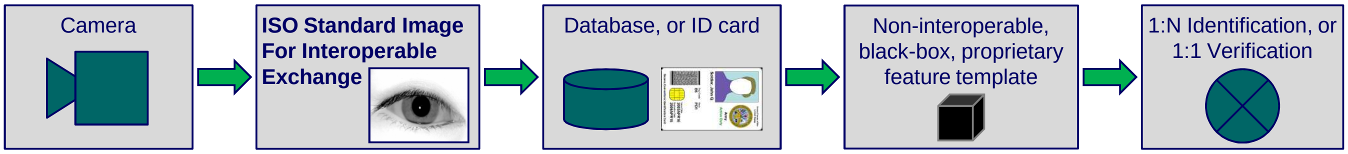
IREX I was initiated to give quantitative support to the revision of the ISO/IEC 19794-6 and ANSI/NIST TYPE 17 standards, and to form a multi-provider marketplace around those standards. As the largest independently administered test of iris recognition technology to date, IREX I includes a formal evaluation of the state-of-the-art of iris recognition algorithms from the following providers:

Cambridge University
L1 Identity Solutions
Cogent Systems
LG
Crossmatch Technologies
Neurotechnology
Honeywell
Retica Systems
Iritech
Sagem

Recognition algorithms from these organizations were evaluated in a three stage process. First, algorithms were applied to convert raw images from contemporary iris cameras to the standardized iris images (i.e. IREX records) depicted here:



Preparation of these records requires various detection, localization, cropping, sampling and masking operations. These operations are non-trivial. They precede the second stage of processing in which features are extracted from standard images to form a template. The IREX records are not iris templates; instead they are specialized interoperable images designed for efficient storage. Templates contain proprietary "black box" feature representations. Their content is non-standard, non-interoperable and not suitable for cross-agency exchange of iris data. The last stage, recognition, involves matching of templates to produce comparison scores. The role of standardized images is depicted as follows.


Camera
ISO Standard Image For Interoperable Exchange
Database, or ID card
Non-interoperable, black-box, proprietary feature template
1:N Identification, or 1:1 Verification

The primary impacts of IREX are listed below. These are followed, in the next section, by an extensive, technical, summary of the results.

- IREX has advanced iris recognition toward the level of technical maturity and interoperability of fingerprint biometrics and has affirmed the potential for using iris biometrics as a second modality for large-scale identity management applications. This result will support storage of iris biometrics on identity credentials such as the United States Government's PIV cards in support of Homeland Security Presidential Directive (HSPD) 12. In addition it will directly support the interoperability goals of Homeland Security Presidential Directives 6, 11 and 24.
- IREX quantified the core algorithmic capability of nineteen recent iris recognition software implementations from ten organizations. This represents an order of magnitude expansion in the number of providers over the last half decade.
- IREX required participating organizations to implement the image formats proposed for the ISO standard. The result is that each provider now has off-the-shelf, or readily portable, software to support creation, validation, and recognition of standard images.
- IREX I complements considerable activity in the area of iris camera development. This has occurred particularly in the stand-off capture (where iris images are acquired at a few meters) and mobile device arenas. These, and other, cameras are technically capable of producing images in conformance to formal standards. IREX recommends that users should require cameras to do so.
- Standard iris image records with size of approximately thirty kilobytes can be produced for large-scale identification applications. This represents a factor of ten reduction in size over the images captured using contemporary cameras.
- Standard iris image records with sizes around three kilobytes can be produced that are suitable for one-to-one authentication applications. This factor of one hundred reduction in size over the images captured using contemporary cameras makes the images suitable for storage on "smart card" credentials.
- There is an industry-wide accuracy versus speed tradespace: Large-scale identity management applications benefit from fast algorithms; Forensic and unconstrained-capture applications should leverage newer, more computationally expensive, iris recognition algorithms.

The authors are available to discuss and brief this report.

TECHNICAL SUMMARY

The significant results of the test are listed by subject-matter area below. These should be weighed in light of the caveats presented on page 13. Vendor comments on their IREX algorithms are included in the accompanying appendices[2]. These are accompanied, separately, by free vendor comments on IREX itself.

IMPACT ON THE MARKETPLACE

- In parallel with the revision of the ISO/IEC 19794-6 iris image interchange standard, the IREX study attracted ten organizations into implementing the standard. This entailed significant effort on the part of the providers with respect to the production of syntactically correct code and with respect to algorithmic functionality. The result is that each provider now has off-the-shelf, or readily portable, software to support creation, validation, and recognition of standard images. Sec. 4
- IREX quantified the core algorithmic capability of nineteen recent[3] iris recognition software implementations from ten organizations. This represents an order of magnitude expansion in the number of providers over the last half decade. There are at least two other commercial providers whose algorithms were not openly submitted to IREX. The availability of standards-compliant implementations from these providers is not known. Sec. 4
- The compact interoperable formats tested here are amenable to lossless compression to as little as 20 kilobytes. Lossless compression preserves imaging detail and ensures that iris recognition accuracy for large scale one-to-many applications is not compromised. Sec. 8.7
- Used with lossy compression, compact interoperable images occupy as little as two kilobytes. This makes them suitable for storage on ISO/IEC 7816 integrated circuit “smart card” identification tokens. In comparison to the other biometric interchange records currently used on such credentials, the IREX record is somewhat larger than standard fingerprint minutia templates (approx. 300 bytes[4]), but smaller than standard fingerprint images (typically 6 to 10 kilobytes[5]) and e-Passport face images (from about 15 to 20 kilobytes[6]). Sec. 8
- Compressed iris image sizes are similar to the template sizes measured for many algorithms submitted to the IREX evaluation. Such images can be matched without loss of recognition accuracy. There are no standards for iris recognition templates, and they are not interoperable. Their size advantage is small. Standard iris images should be exchanged instead. Sec. 7.5
- The ISO/IEC 19794-6 standard is application-neutral. Standard images are suitable for iris recognition applications embedding large-scale one-to-many identification searches (watchlist, deduplication, fraud detection), one-to-many token-less verification claims, and one-to-one verification claims with a credential.
- By executing a wide ranging study and reporting detailed results, IREX is expected to influence technical development of iris recognition algorithms in unforeseen ways.

INTEROPERABLE IMAGE FORMATS

[2] The appendices include: optional provider-supplied text describing each specific recognition algorithm; comments from the participants on the IREX activity; and detailed technical information for each specific IREX implementation. The appendices may be downloaded from http://iris.nist.gov/irex/irex_appendices.pdf.

[3] The implementations were sent to NIST in mid February 2009.

[4] A INCITS 378 minutia template from a flat index-finger impression containing 38 minutiae encoded in 6 byte format.

[5] A typical ISO/IEC 19794-4 single finger record, acquired at 500 pixels per inch and WSQ compressed at 15:1.

[6] A color ISO/IEC 19794-5 token face image, interocular distance 90 pixels, 15:1 compression and chrominance sub-sampling.

- For applications without size, transport, or communications-related throughput constraints the uncropped uncompressed KIND 1 [7]rectilinear record may be used. Such images, of size 640 x 480 pixels, can be losslessly compressed using the ISO/IEC 15948 Portable Network Graphics compression algorithm (PNG) by about a factor of two. The resulting records occupy a median size of 150 kilobytes. Sec. 8.7

- If the acquisition process can include a coarse iris detection operation, then the centered crop-only KIND 3 record almost always gives fewer false rejection errors than its un-cropped, uncompressed, and unconstrained KIND 1 parent. With lossless compression, KIND 3 instances require 50-80 kilobytes of storage. Instances may be further compressed using the lossy JPEG2000 algorithm. The crop-only KIND 3 image format should be retained in the ISO/IEC 19794-6 standard. Sec. 8.2

- The cropped-and-masked KIND 7 image format proposed for the ISO/IEC 19794-6 standard should be retained and advanced as the primary format for the exchange of compact iris images smaller than 3KB. At larger sizes or lower compression ratios, the KIND 3 format should be preferred: it is more easily and safely instantiated. The false rejection performance for some implementations exceeds that of the KIND 1 parent. The cropped-and-masked KIND 7 format is particularly amenable to lossless compression. This allows iris records to be produced in the 20-40 kilobyte range. This format should usually only be used in conjunction with the JPEG2000 and PNG compression algorithms. Sec. 8.2

- The KIND 16 unsegmented polar format proposed for the ISO/IEC 19794-6 standard should be rejected. The recognition error rates associated with the format are much larger than those attainable with rectilinear KIND 3 and KIND 7 records. This is true natively, when a single provider prepares and matches the records, and in the interoperable case also, when different providers do so. Sec. 8.2

- For some images, lossless compression will not be able to achieve a specific target size, and JPEG2000 should be applied at a specific targeted bit rate. For images below about 20KB, lossy JPEG2000 compression will usually be needed. Sec. 8

- Using false rejection error as a metric, the cropping operation used in preparation of KIND 3 and KIND 7 records should extend to no closer than 0.6 iris radii from the iris in the horizontal direction, and 0.2 radii in the vertical direction. Sec. 8.5

- Iris cameras should not internally apply compression to iris images, unless they are manufactured for a dedicated, profiled, application in which standardized compressed iris images are produced. An exception is lossless compression.

- There is a large academic literature addressing the iris localization problem, and a rich diversity of algorithms can be employed to effect the detection and localization steps necessary to instantiate KIND 7 and KIND 3 records. Detection and localization are non-trivial operations and are influential on recognition accuracy. Users should evaluate implementations accordingly. Sec. 1

RECOGNITION ACCURACY

- False rejection performance depends on the following (in decreasing order of importance): the recognition algorithm, the particular image dataset, the standard image format, and on the amount of compression applied. The observed error rate variations span at least an order of magnitude. Sec. 7.3.1

[7] For a visual description of the image KINDS evaluated in IREX, refer to Figure 3 on page 19.

- False match rate (FMR) calibration curves are computed. These confirm the low false match rates published for iris recognition algorithms at specific operational thresholds. By comparing images from two occupationally and geographically separated populations, of combined size $O(10^4)$, the FMR calibration is based on $O(10^9)$ comparisons. While this population size extends prior independent studies, larger operational corpora should be leveraged to refine FMR performance estimates and give statistical significance to FMR measurements on national-scale identification searches. Sec. 7.4

- The stability of the impostor distribution is measured and reported. At a specific threshold, the false match rate depends on the dataset. This applies across all formats and compression conditions, including uncompressed and unprocessed iris images. In addition, for most iris recognition algorithms, false match rates vary under compression. When severe compression is applied to damage the iris texture, some algorithms maintain low FMR; others do not. Sec. 7.4

Interoperable accuracy

- Given a standard for iris images, there are two separate tasks for iris recognition: Generation of the standard image, and matching of standard images (matching is implemented via by proprietary templates). These tasks will generally be executed by different providers' algorithms. The implementations that most accurately match standard IREX records are not generally the implementations that prepare the most matchable IREX records. That is, the error rates associated with the initial detection of the sclera-iris and iris-pupil boundaries are distinct from, but smaller than, the error rates associated with the end-stage fine-grained localization, feature extraction and matching. Sec. 8.6

- The interoperability of standardized iris images in IREX is better than that reported for standard fingerprint minutiae templates: There, the best accuracy was observed when the same provider generated and matched the enrollment and verification templates. In IREX, this native-mode bias is small. Instead the iris recognition algorithm (fine-grained localization, feature extraction, and matching) is most influential on outcome. Fingerprint minutia interoperability is degraded by idiosyncratic (i.e. algorithm-specific) minutia detection and selection[8, 25, 23]. For iris, the standard interchange medium is image data; for fingerprint minutiae, it is (x, y, θ) point data and the relative interoperability of the two is a product of the difficulty of consistently and uniformly instantiating the semantic content of the respective standards. Sec. 8.6

Computational efficiency

- The IREX test plan explicitly stated that algorithm timing estimates would be reported. Further, IREX encouraged the submission of slow-but-accurate vs. fast-but-less-accurate implementations within a generous timing budget. The result, across the nineteen IREX algorithms, is that there are two orders of magnitude for the time needed to prepare an IREX record and to generate a template from an IREX record, and *three* orders of magnitude for the time needed to execute a one-to-one match. Sec. 7.6

- The more computationally expensive algorithms give fewer recognition errors. This accuracy benefit applies on both the segmentation side, and the matching side. While the results demonstrate the existence of an industry-wide accuracy vs. time tradespace, not all providers demonstrate such a tradeoff between their primary and secondary SDKs. Sec. 7.6.3

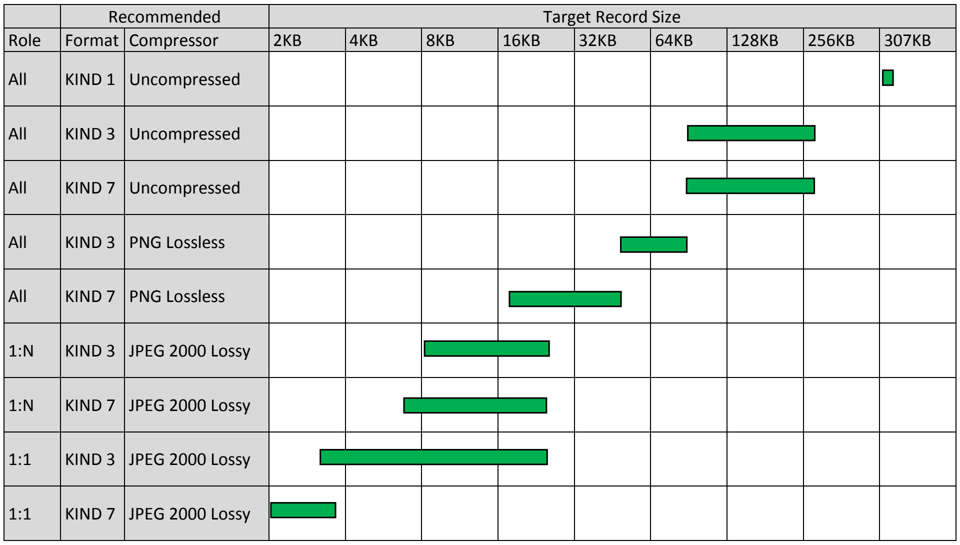


Figure 1: Application-specific recommendations on compression and format. The horizontal axis shows target file size in kilobytes on a logarithmic scale.

- ▷ Most prior published tests have ignored the tradeoff between computational expense and accuracy. Here more computationally expensive algorithms yield better accuracy. This indicates that difficult-to-match samples are amenable to more expensive algorithms. While IREX allowed computationally intensive algorithms, providers were given prior notice that this report would report execution speed in addition to accuracy. NIST biometric testing campaigns have historically emphasized matching accuracy. Sec. 7.6.4
- ▷ The speed of the fastest algorithms is in line with that reported in the academic literature. Sec. 7.6
- ▷ All of the implementations are fast enough for use in one-to-one applications. However, in one-to-many identification applications with even moderate enrolled population sizes, the viability of the slowest IREX implementations may rest on the availability of *fast search* (e.g., dataset partioning) algorithms. The degree to which the algorithms in the IREX SDKs can be expedited is not known. Sec. 7.6.2
- ▷ The fastest implementations are more than twice as fast as the slowest: They can compute and match pairs of templates from both eyes in less time than it takes the slowest algorithms to compute and match templates from a single eye. Sec. 7.6.1
- ▷ The more accurate implementations may be useful for forensic iris identification where computational expense is usually inconsequential. Sec. 7.6.1
- ▷ Important caveats apply to the measurement of computation time, and to the operational relevance of timing estimates. Sec. 7.6.4

SELECTION OF COMPRESSION ALGORITHMS

- ▷ The IREX study has supported refinement of the ISO/IEC 19794-6 international iris interchange standard. Particularly the IREX study supported exclusion of polar formats from the ISO/IEC 19794-6 standard. Sec. 8.2

- The IREX study has confirmed the findings of previous studies, namely that compression gives low, graduated, increases in false rejection errors. Prior work, published in the academic literature, has often considered single non-commercial algorithms running on smaller datasets. Sec. 8.1.1

- Lossy compression algorithm such as JPEG and JPEG2000 unrecoverably damage iris images. This has an adverse effect on false non-match error rates, and, for some iris recognition algorithms, on false match rates too. This latter aspect contraindicates application of lossy compression to images used in one-to-many searches. Sec. 8.1.3

- Compression should be applied at the minimum level needed to attain a storage or bandwidth requirement. The primary operational target variable is the size of the compressed image. This may be set directly using the JPEG2000 algorithm. Default guidance is given in Figure 1. However, implementers should quantify compression damage in terms of bits per pixel, and this will depend on iris radius - large irises should be compressed more lightly. Sec. 8.6

- The ISO/IEC 10918 JPEG compression algorithm should be deprecated. The presence of Discrete Cosine Transform blocking artifacts produces elevated false match rates. This recommendation applies particularly to compact iris images, but also to the cases where JPEG encoding is being used solely as a convenient container. It is recommended that the lossless PNG compressor be used in such cases because it is too easy to invoke JPEG with adverse parameters. Sec. 8.1

Effects of dilation, occlusion, center displacement

- Higher amounts of pupillary dilation increase image-specific false non-match and false-match rates. The effect diminishes when the enrollment images are JPEG2000 compressed. Sec. 8.9.2

- Images stored in KIND 16 format with constricted pupils produce higher image-specific false match rates. Perhaps high amounts of constriction make the features more difficult to localize. Sec. 8.9.2

- The difference between the dilations present in two iris images affects their comparison score. In particular, large disparities in dilation elevate false rejection error rates. False match rates are not changed. The magnitude of the effect is comparable to that of eyelid occlusion. Sec. 8.9.2

- The effect of dilation change on recognition accuracy diminishes when the enrollment images are JPEG2000 compressed. Sec. 8.9.2

- Changes in dilation tend to be smaller for intra-person comparisons than for inter-person comparisons. Under controlled illumination conditions, the amount of dilation may serve as an ancillary discriminating factor. Sec. 8.9.2

- Large amounts of eyelid occlusion increased the probability of a non-match for most algorithms, and increased the probability of a false match for some algorithms. Sec. 8.9.1

- Some IREX algorithms exhibit a small adverse dependence on the displacement of the pupil and limbus centers. Specifically, genuine matching scores are elevated. Sec. 8.9.3

Image quality assessment

▷ Three SDKs reported iris image quality scores on the standard range of [0-100]. A quality score, computed during the production of a KIND 3 record, is considered effective if they are quantitatively indicative of matching performance. Quality scores are considered interoperable if they are effective and uniformly interpretable. Two of three quality assessment algorithms, D1 and G1, are effective at assigning higher quality values to images with lower image-specific false match and false non-match rates. SDK D1 exhibits a larger stratification of the median quality scores assigned to four accuracy-categorized partitions of the ICE dataset. Sec. 8.8.1

▷ When images are excluded from testing on the basis of poor image quality assessments by SDKs G1 and D1, the false non-match rate of almost all matching algorithms improves. The best gain in accuracy is achieved for SDKs A1, I1, I2, G2, H2, H1. Exclusion of images judged to have low quality by SDK A1 does not appreciably improve observed false non-match rates. Sec. 8.8.2

▷ Quality scores are not interoperable: The distributions of the D1 and G1 quality scores are different and relate to different image error rates. This lack of interoperability implies a need for calibration. Sec. 8.8.2

▷ The IREX test plan did not require image quality assessment. While there is an increasing concensus that image quality estimates have greatest utility if they they are indicative of recognition accuracy, the IREX test plan did not mandate a semantic meaning for image quality. The new ISO/IEC 29794-6 iris image quality standard is intended to improve this situation. Academic and commercial research, and IREX activities will support this project. Sec. 8.8.2

EXPANDED TEST METHODS AND METRICS

▷ Detection error tradeoff characteristics are included that present lines joining points of fixed threshold. These reveal changes in *both* FNMR and FMR at a fixed threshold. These expose dependencies of FMR on dataset and compression parameters. Sec. 7.3

▷ Within-dataset and cross-dataset impostor comparisons were conducted. The latter guarantee the integrity of the impostor status of all image pairs. Matching of samples from different cameras is an implied aspect of applications based on standardized images. Sec. 7.4

▷ Conditional false non-match rates were defined and used to quantify errors induced by a change to a dataset (e.g., compression). This offers a more precise approach to failure analysis when the covariate is under the control of the experimenter. Sec. 8.2

▷ Template sizes were reported. Sec. 7.5

▷ Processing times were reported for IREX record preparation, template generation and one-to-one comparison. The computational cost vs. accuracy tradespace was documented. Sec. 7.6

▷ Image-specific false match and false non-match error rates were defined by inheriting concepts from the biometric zoo. These metrics support failure mode analyses by allowing association of a covariate (e.g., dilation) with a matching error rate without having to consider the covariate of a comparison image. Image-specific error rates are useful in detection of ground truth errors in datasets. Sec. 6.4

▷ Image-specific error rates were used to create four partitions of the ICE dataset: These are termed CLEAR ICE, BLUE GOATS, BLUE WOLVES, and BLACK ICE. The latter consists of those images that have pathological error rates on all SDKs. On request NIST will release the image partitions to interested parties. Sec. 8.10

- The IREX study defined a "C" API to support the various investigations. This API was implemented successfully by all the IREX participants. It is published[8], freely available, and easily re-usable by other testing programs. It could be used with its full IREX record functionality, or in a more generic raster-template-match mode.

NEXT STEPS AND SUPPORT FOR DEVELOPERS

- NIST will consider requests for additional quantitative feedback in support of algorithm development. As part of this process NIST will invite IREX participants to inspect some problematic images.
- NIST invites comment on what further work is needed in support of standardization of iris image interoperability. Further, NIST solicits input on how the IREX umbrella program might be extended to support iris image interoperability. Comments and inquiries are welcome via IREX@NIST.GOV.
- NIST will initiate a second activity under the IREX umbrella. The project, the *Iris Quality Evaluation and Calibration* (IQEC), focuses on evaluation and calibration of iris quality scores. IQEC is a large scale evaluation of iris quality scores and will commence in Fall 2009. IQEC aims to evaluate the effectiveness of image quality assessment algorithms IQAA in predicting recognition accuracy of particular comparison algorithms (from the supplier of the IQAA), and of others' algorithms. Given the IREX result that quality scores are not immediately interoperable, IQEC will establish a calibration procedure of IQAAS.

[8] See http://iris.nist.gov/irex/irex_api.pdf

CONTENTS

CAVEATS

As with all biometric evaluations, the results of this test must be carefully interpreted before any predictive conclusions can be made. Users should factor the following into policy, planning and operational decisions.

1. IREX I did not address evaluation or standardization of cameras, interfaces, and complete systems. It does not establish operational requirements, nor does it consider transmission protocols, and security issues such as algorithm vulnerabilities. These issues, which must be addressed operationally, may impact design tradeoffs based on IREX.

2. The absolute error rates quoted herein were measured by using the provided implementations on three large fixed corpora of operational and non-operational iris images. As with all offline biometric tests, the relevance of the results to operational reality must be considered in light of the fact that post-capture samples are used. Error rates observed in real-world applications are almost always strongly dependent on acquisition related factors. Generically these include

    - The degree to which the design compels, induces, or incentivizes the user to use the camera in a mode intended by its designers,
    - Cooperativeness of the user population (an uncooperative subject may be evade acquisition and be very hard to image, a non-cooperative user may similarly not look at, or properly present to, the camera);
    - Environment (e.g., low ambient light levels may impede detection);
    - The number of verification attempts allowed (typically more attempts lead to lower false rejection, and higher false acceptance);
    - The number of images used, and the fusion policy (if several images from a sequence are matched, accuracy can be improved);
    - Number of biometric objects (fingers, irides) used, and the fusion policy (two gives better accuracy than one)
    - Demographics (e.g., children and older adult populations may not present as quickly or as easily)
    - Habituation (Users who regularly interact with system often yield lower rejection rates);

3. The sensor, and the enrollment policy affect error rates. For example, iris cameras almost always compute quantitative quality criteria in an auto-capture loop either in the camera's firmware, or sometimes in a client-side application, or both. This may produce some failure-to-enroll occurrences, but will improve downstream matching error rates.

4. With respect to iris recognition specifically, the accuracy and speed of operational transactions will generally depend on a number of factors, including the following.

    - The template generation and matching algorithms are strongly influential on error rates.
    - The number of eyes imaged
    - The number of images available for matching
    - The quality of the enrollment procedure particularly whether a verification was done at time of original enrollment
    - The communications channel and interface, and the compression of the records it implies.

RELEASE NOTES

- The IREX evaluation was conducted in accordance with the IREX API and CONOPS which was developed over the period October 2007 through September 2008. It was developed in consultation with the public and received comments from seven different organizations. It imported content on the IREX image formats from documents submitted to SC 37 in support of ISO/IEC 19794-6 . It is archived at http://iris.nist.gov/irex.

- This document is accompanied by the IREX APPENDICES which present exhaustive results for each submitted iris recognition algorithm. These may be downloaded at http://iris.nist.gov/irex/irex_appendices.pdf.

- The IREX trial has been conducted in broad conformance to the ISO/IEC 19795 - Biometric Performance Testing and Reporting - Part 4: *Performance Interoperability Testing* standard.

- Throughout this report the submitted iris recognition algorithms are identified by a letter and a numeral of the form Nx. The letter identifies the company or university that submitted the algorithm. The numeral x takes the value 1 for the *primary* algorithm and 2 for *secondary*[a].

- The use of these codes is intended to conserve space in its many tables. For reference, the letters are associated with the providers' names in a running footnote.

- A glossary of terms and definitions is given on page 16.

- Much of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly typesettable LaTeX content. This improves timeliness, flexibility and maintainability, and reduces transcription errors.

- This PDF file is large. While it is likely to be of better quality in print than on-screen, it may print slowly.

- Readers are asked to direct any correspondence regarding this report to the IREX@NIST.GOV.

[a] While the intent in allowing submission of two algorithms was to explore the tradespace between speed and accuracy, IREX participants were free to vary the algorithmic functionality in any way they felt appropriate. The IREX SUPPLEMENTAL includes information volunteered by the IREX participants on their choices.

ERRATA AND CLARIFICATIONS

- 1. October 28, 2009: The green heatmaps appearing in section 7.2 (on the effect of iris radius) were incorrect. The heatmaps show the count of genuine comparisons broken out by radius of the enrollment and verification samples. The heatmaps erroneously showed a function of genuine scores. This applies to versions of the IREX report and its appendices issued on or before October 6, 2009. The iris radii themselves have not changed.

- 2. October 28, 2009: The red-yellow heatmaps appearing in section 7.2 (on the effect of iris radius) were incorrect. The heatmaps show the false non-match rate broken out by radius of the enrollment and verification samples. The heatmap computation sometimes incorrectly applied the iris radius of a sample that was not the enrollment sample. This applies to versions of the IREX report and appendices issued on or before October 6, 2009.

- 3. October 28, 2009: A line has been added to section 7.2 (on the effect of iris radius) to indicate that the iris radii are those estimated by the I1 implementation, as recorded in the header of its KIND 3 records.

- 4. October 30, 2009: The failure to acquire counts in Table 6 for the BATH images were double their true value. However, the FTE *rates* were and are correct. This applies to versions of the IREX report and appendices issued on or before October 6, 2009.

VERSION CONTROL

Study	IREX I
Report generated	Fri Oct 30 14:30:21 2009
Report name	tex/irex_report.tex
Report last modified	Fri Oct 30 14:07:34 2009
Report MD5 checksum	e54a635a43123e27f08cd118452b9565
NIST contact	irex@nist.gov

TERMS AND DEFINITIONS

Table 1 gives IREX-specific definitions to various words and acronyms found in this report.

No.	Term	Definition
Organizations		
1	ANSI	American National Standards Institute
2	DHS	U. S. Department of Homeland Security
3	DoD	U. S. Department of Defense
4	ISO	International Organization for Standardization
5	IEC	International Electrotechnical Commission
6	INCITS	International Committee for Information Technology Standards
7	NIST	National Institute of Standards and Technology
8	M1	The standards body that formulates comments toward SC 37 biometrics standards
9	SC 17	Subcommittee responsible for development of identification card standards
10	SC 37	Subcommittee responsible for development of biometrics standards
Programs		
11	IREX	Iris Exchange - NIST's umbrella program for supporting iris interoperability
12	MINEX	Minutiae Exchange - NIST's umbrella program supporting minutia interoperability
Standards		
13	INCITS 379:2004	U.S. standard for iris images
14	ISO/IEC 19794-6:2005	International variant of the INCITS 379 format, the focus of this report
Data elements		
15	Standard template	Standardized templates do not exist for iris recognition. IREX is concerned with interoperability based on standard images
16	Proprietary template	Usually unpublished feature representation of matchable iris data - comparable only with a template from the same vendor and product line
17	Enrollment template	Synonym for reference template
18	Reference template	Template, logically from the enrollment or first-encounter sample
19	Verification template	Template generated from a subsequent sample of a subject or from an un-enrolled unknown or impostor sample
20	BDB	Biometric Data Block (See SC37's *Harmonized Vocabulary*[3])
Function and process terms		
21	SDK	Software Development Kit
22	API	Application Programming Interface
23	Matcher	In IREX a matcher is logically a function that compares two IREX records and produces a dis-similarity score. Physically it compares two templates
24	Generator	Software function that accepts an image and produces a standard record
25	Native mode	Comparison by SDK X of IREX records prepared by SDK X
26	Interoperable mode	Comparison by SDK X of IREX records prepared by SDKs Y and Z
27	Genuine	Comparison of data from the same person
28	Impostor	Comparison of data from different individuals
29	Verification	One-to-one comparison
30	Authentication	Synonym for verification
31	Localization	Image processing operations to locate the iris or pupil boundaries
32	Segmentation	Synonym for localization
Metrics		
33	FAR	False accept rate (i.e. transactional outcome)
34	FRR	False reject rate (i.e. transactional outcome)
35	FMR	False match rate (i.e. 1:1 single sample comparison outcome)
36	FNMR	False non-match rate (i.e. 1:1 single sample comparison outcome)
37	DET	Detection Error Tradeoff characteristic

Table 1: Glossary of IREX related terms

Chronology of Iris Standardization

Table 2 lists the key dates in the evolution of IREX. The table includes items related to the standards.

No.	Period	Event
1	July 9, 2009	SC 37 Working Group 3 Meeting
2	July 3, 2009	Release of first public draft of this report
3	May, 2009	ISO/IEC 29794-6 - Iris Image Quality is approved
4	June 12, 2009	IREX Phase III report is submitted for release
5	March 30, 2009	NIST comments toward revision of ISO/IEC 19794-6
6	February 2, 2009	Release of first committee draft of ISO/IEC 19794-6 iris image interchange standard, document JTC 1 SC37 N3031
7	February 17, 2009	Deadline for delivery of Phase II SDKs to NIST
8	January 22, 2009	SC 37 Working Group 3 Meeting
9	October 20, 2008	Phase I evaluation begins
10	...	Acceptance testing of Phase I SDKs
11	September 17, 2008	Deadline for delivery of Phase I SDKs to NIST
12	August, 2008	Finalization of IREX CONOPS and API specification
13	November 16, 2008	Release of initial IREX CONOPS and API for comment
14	March 24, 2008	NIST comments toward revision of ISO/IEC 19794-6
15	February 1, 2006	NIST Special Publication 800-76-1 is released
16	September 15, 2005	Publication of ISO/IEC 19794-6 Biometric Data Interchange Format - Iris image data
17	August 27, 2004	Homeland Security Presidential Directive 12 is signed
18	March 8, 2004	INCITS 379 finalized

Table 2: IREX chronology and related events.

1. Introduction

NIST's iris interoperability program, IREX, was initiated to support an expanded marketplace of iris-based applications based on standardized interoperable imagery. The work is primarily conducted in support of the ISO/IEC 19794-6 standard, now under revision. It secondarily supports the upcoming revision of the ANSI/NIST ITL 1-2007 Type 17 standard[44].

Support for the ISO standard: The primary IREX motivation is to support a more robust, interoperable, useful and implementable ISO/IEC 19794-6 standard. The standard defines image formats for exchange of iris images. It allows migration from the single-provider template-based applications depicted in Figure 2 toward open, multi-provider, applications based on the standardized image formats presented in Figure 3.

IREX was initiated to give a quantitative basis for the inclusion and exclusion of image formats in the ISO standard. This work supports the SC 37 Working Group 3 (WG3) and M1.3 committees by specifically embedding conformance, performance and interoperability tests as part of the standards' development process. IREX was structured as an application-independent assessment of the core algorithmic performance of the localization and recognition components.

Toward compact iris images: The second motivation in executing IREX I was the establishment of a standardized accurate, interoperable and compact iris image format suitable for large-scale identity management applications. The IREX study was intended to avert the situation that arose with the ISO/IEC 19794-6 published in 2005. That standard defined a non-concentric polar format for compact storage of iris data. It was adopted operationally, and while interoperability problems were never formally demonstrated, two technical contributions to WG3 [Germany, N2059; Great Britain, N2124] asserted that accuracy of the polar format was critically sensitive to the consistency of the localization, and subject to sampling problems[53]. The two documents advocated removal of the polar format. A GB contribution[18] usefully suggested the polar format's size (around 2 kilobytes) can be achieved via cropping and compression of the rectilinear format. In addition, the US suggested a new polar variant termed the unsegmented polar format[36] with similarly low storage potential.

NIST is particularly interested in establishing a set of specifications for efficient transmission of iris imagery across a network, and for storage on an ISO/IEC 7816 crypto-token. Toward similar ends, iris compression studies have been conducted [34, 19, 54]. While the studies reported promising results, they explored only the case in which enrollment and verification data are processed by a lone supplier's localization and matching algorithms. The exception here is [54]

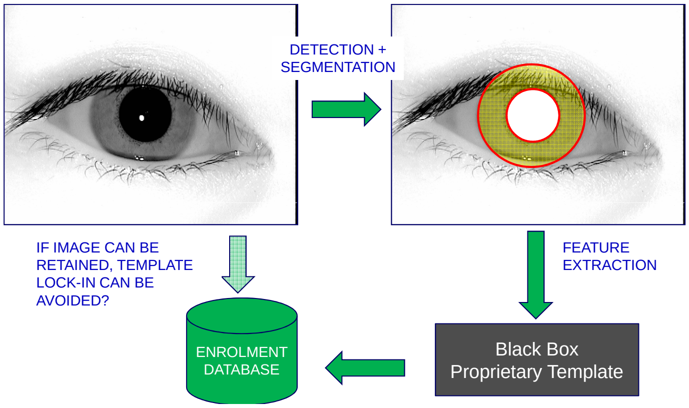


Figure 2: Non-interoperable biometrics: If an application does not retain the sample images, the system can only function with the matching algorithms of one provider, because the templates are non-standardized.

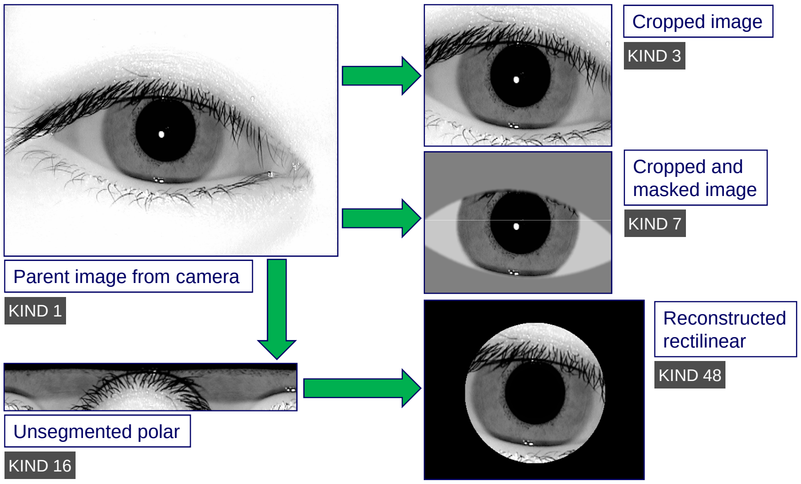


Figure 3: The standard image formats tested in IREX. Preparation of the KIND 3 record requires detection of the iris, and a crop operation to center the iris. The KIND 7 record requires detection of the iris-eyelid and iris-sclera boundaries and a pixel-replacement masking operation. Finally instantiation of the KIND 16 record requires location of concentric circles inside the pupil and outside the iris, followed by an rectilinear-to-polar mapping.

which showed similar compression sensitivities for two different matching algorithms. The ISO/IEC 19794-6 standard was published in 2005. It is almost identical to its precursor, the INCITS 379 standard published in the United States in 2004. As application-independent standards, neither document establishes normative requirements on compression. Instead the 2005 ISO/IEC 19794-6 standard's clause A.1.6 gives the following informative guidance " a compression factor of 6:1 or less is recommended". This is an order of magnitude smaller than compression ratios published in the last two years[19, 54].

IREX I was intended to support the ISO/IEC 19794-6 by requiring participants to produce conformant instances of the rectilinear format, and by evaluating the compact formats proposed by the GB and UK national bodies via submission of technical specifications[19, 18, 36] toward the revision of the original ISO/IEC 19794-6:2005 Iris image interchange standard. The revision is ongoing in Working Group 3 of the ISO committee on biometrics, SC 37. NIST supported the effort by defining a syntactic data record based on the 2005 standard and the technical contributions.

Examples of the images are shown in Figure 3 and their properties are summarized in Table 3.

1.1. Application scenarios and scope

Many systems operate under biometric data size constraints in their communications infrastructure, their storage requirements, and in the computational cost associated with larger data. In network-based applications the transmission times are linearly related to the data size. Similarly, in card-based applications the to-card and from-card transfer times influence the selection of modality (minutia, iris etc.) and the number of instances (two fingers, two irises etc). For IREX two notional operational scenarios are considered. These cover compression of one or both samples in a comparison.

- ▷ *Identity credential:* A compressed standard iris image is stored on, for example, a ISO/IEC 7816 smart card, and is compared during authentication against a newly collected uncompressed sample[9]. This scenario is representative of cooperative physical or logical access control situations in which the first sample is collected and prepared in an attended formal enrollment session, and the authentication sample exists only for the duration of the attempt. In

[9] The term uncompressed in IREX is used to indicate that both that the image is represented as a pixel raster and that it has never before been compressed.

Aspect	Cropped Rectilinear	Cropped + Masked Rectilinear	Unsegmented Polar
Identifier	KIND 3	KIND 7	KIND 16
Definition	The iris is cropped and centered.	The iris is cropped and centered, and the eyelids and sclera are masked with a uniform pixel value.	The region between concentric circles inside the pupil and outside the iris is unwrapped (rect → polar)
Defining citation	[19, 18]	[19, 18]	[36]
Standards compliance	As a rectilinear instance it is compliant to ISO/IEC 19794-6:2005 but that standard gave no size parameter guidance. Allowed in ISO/IEC 19794-6 revision project (estimated completion 2010). Allowed in ANSI/NIST ITL 2007 Type 17[44].	As a rectilinear instance it is compliant to ISO/IEC 19794-6:2005 but that standard gave no size parameter guidance. Allowed in ISO/IEC 19794-6 revision project (estimated completion 2010).	A non-concentric polar version was present in ISO/IEC 19794-6:2005. Withdrawn from ISO/IEC 19794-6 revision project (est. 2010).
Required localization	Coarse iris detection	Fine detection of sclera-iris boundary and eyelid-iris boundaries.	Coarse detection of pupil and iris.
Not intended or allowed (but not enforced)	Image enhancement	Image enhancement	Eyelash removal[67], eyelid masking

Table 3: Summary properties of the image formats tested in IREX.

IREX this scenario is addressed by compression of one sample (the enrollment sample) in one of the three compact formats, and execution of matching comparisons against uncompressed KIND 1 records.

- *Central matching facility:* Compressed standard samples are submitted to a central dataset. These are the first-encounter "enrollment" samples. Subsequently, compressed samples are transmitted to the central facility and are matched against the enrollments. This scenario would be typical in open-universe one-to-many applications such as visa fraud detection and watchlists. The need for compression is implied by operational network bandwidth constraints. In this scenario both images might be compressed in one of the standard formats.

  Figure 4 depicts a network-centric application in which iris image data is moved over network links (shown as green arrows) whose bandwidth may vary (as indicated by the width of the arrows).

- *Baseline:* To examine the effect of compression, a baseline is established by matching uncompressed unprocessed KIND 1 records against themselves.

With the applications defined, the formal IREX I scope was stated as follows.

- To quantify the performance and interoperability of rectilinear images, the UK-proposed ROI masked rectilinear images[18, 19], and the US-proposed unsegmented polar images[36].
- To measure the effect of JPEG and JPEG2000 compression on accuracy.
- To quantify the performance and interoperability of iris localization algorithms.
- To time the various operations.
- To formulate record structures and other content toward the revision of ISO/IEC 19794-6.
- TO check that suppliers can produce records conformant to the ISO/IEC 19794-6 standard.

The primary outputs of the test are statements of performance including:

- failure-to-segment rates for various compression levels;
- false non-match and false match error rates for various compression levels and operating thresholds;
- interoperable error rates for comparison of image records prepared by different providers' algorithms;

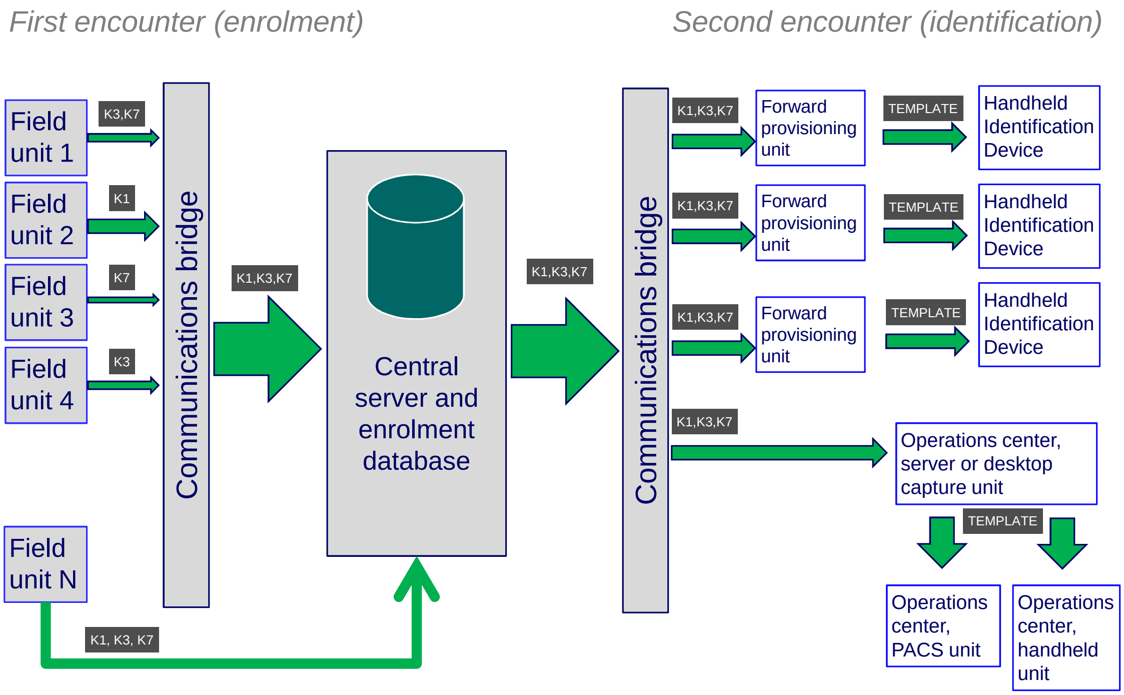

First encounter (enrolment)
Second encounter (identification)
Field unit 1
K3,K7
Field unit 2
K1
Field unit 3
K7
Field unit 4
K3
Field unit N
K1, K3, K7
Communications bridge
K1,K3,K7
Central server and enrolment database
K1,K3,K7
Communications bridge
K1,K3,K7
Forward provisioning unit
TEMPLATE
Handheld Identification Device
K1,K3,K7
Forward provisioning unit
TEMPLATE
Handheld Identification Device
K1,K3,K7
Forward provisioning unit
TEMPLATE
Handheld Identification Device
K1,K3,K7
Operations center, server or desktop capture unit
TEMPLATE
Operations center, PACS unit
Operations center, handheld unit

Figure 4: Example of a network-centric application in which standardized iris data is stored centrally and in the field. Note that compact iris images, in KIND 3 and KIND 7 formats, are passed across narrow bandwidth connections. The left-to-right data flow represents the temporal sequence of enrollment then identification. The storage of templates in the handheld units is a defined, localized use of non-interoperable data. In principle the units could be equipped with iris recognition software from different providers.

- ▷ time taken to prepare the various standard images;
- ▷ time taken to extract features from the various standard records;
- ▷ time taken to match feature-based templates.

As depicted in Figure 3, the IREX evaluation required conversion of:

- ▷ raw raster images into ISO/IEC 19794-6 rectilinear images;
- ▷ ISO/IEC 19794-6 rectilinear images into cropped iris-centered rectilinear images;
- ▷ ISO/IEC 19794-6 rectilinear images into ROI-masked rectilinear images;
- ▷ ISO/IEC 19794-6 rectilinear images into unsegmented polar images.

The following were specifically not within the current scope of this evaluation:

- ▷ predictions of operational performance;
- ▷ sensor usability or security evaluation (IREX was conducted with offline imagery);
- ▷ off-angle imagery (other than that incidentally present in the test corpora);
- ▷ conformance to already published standards (e.g., ISO/IEC 19794-6:2005).

The work was carried out with close conformance to ISO/IEC 19795-4:2008 Biometric Performance Testing and Reporting Part 4: Interoperability Performance Testing.

1.2. Relation to other NIST activities

Relation to PIV: The results of IREX may have implications for projects such as the US Government's Personal Identity Verification (PIV) program[10] which was initiated by Homeland Security Presidential Directive 12[11]. This mandated the establishment of a common identification standard for U.S. government employees and contractors. It required interoperable use of identity credentials to control physical and logical access to federal government facilities and systems. In response, NIST released FIPS 201[12] in February 2005, which includes the definition of an identity credential. It specified the inclusion of data from two fingerprints as a third authentication factor. The format for this information was finalized in February 2006, when NIST *Special Publication 800-76* specified the MINEX profile of the INCITS 378 standard for encoding and formatting of fingerprint minutiae[25].

NIST is considering including specifications for iris biometrics in an upcoming revision of SP 800-76-1. This would extend a second interoperable biometric authentication mechanism for U.S. government agencies.

Relation to ICE and MBGC: The IREX activities are distinct from NIST's prior Iris Challenge Evaluations (ICE)[50] and ongoing Multiple Biometric Grand Challenge activities which have more basic research goals.

2. Prior studies

IREX follows a number of studies on the effects of compression on iris recognition accuracy. These are detailed below. Such studies seem necessary for all biometric modalities, and as as noted by Daugman[19], this work has been done for fingerprints as far back as 1993 for 197pixels per centimeter WSQ certification, and more recently for JPEG2000 applied to 394ppcm scans.

University of Bath In 2008 Rakshit and Monro [54] applied the Monro[47, 46], Tan[39, 40] and Masek[42, 41] iris recognition algorithms to 2156 images of 308 different eyes contained in the extended CASIA dataset[12]. They used both JPEG and JPEG2000 (Kakadu) compressors at bits rates from 1.0 to as low as 0.1 bits per pixel. They reported DET characteristics with FMR as low 0.0001. Additionally they gave the dependence of "FMR at the first false rejection" on bits-per-pixel and advanced this as their preferred metric. Compression was applied to both the enrollment and verification images.

University of Cambridge In 2007 Daugman and Downing[19, 18][13] applied their iris recognition algorithms to 1425 images of 124 people from the ICE 05 dataset[50]. All images were cropped to 320x320, centered, and compressed with JPEG (IJG[14]) and with the JASPER implementation of JPEG2000 [15]. They executed 12214 genuine and 1002386 impostor comparisons, and reported performance using DET characteristics with FMR as low as 0.00001, and with decidability measures (see section 6.3). Compression was applied to both the enrollment and verification images.

U.S. Naval Academy In a pair of papers[34, 33] Ives et al. applied the Masek[42, 41] recognition algorithm to 756 images of 108 eyes present in the CASIA[12] dataset, and to 20 images of both eyes of 50 people in the University of Bath dataset[45]. The CASIA images were resized to 320x280. The CASIA images were compressed at ratios from 5:1 to 50:1. The BATH images, which had previously been compressed using JPEG2000 at 16:1 in their full 1280x960

[10] See http://csrc.nist.gov/piv-program/

[11] The text of HSPD 12 is here: http://www.whitehouse.gov/news/releases/2004/08/20040827-8.html

[12] See Federal Information Processing Standards Publication 201, *Personal Identity Verification for Federal Employees and Contractors* and related documents here: http://csrc.nist.gov/piv-program

[13] The authors are participants in IREX .

[14] The Independent JPEG Group's implementation is freely available from http://www.ijg.org/, downloaded June 22, 2009.

[15] The source code is available from http://www.ece.uvic.ca/m̃dadams/jasper/, downloaded June 22, 2009.

native format, were compressed from 20:1 to 50:1. They used both JPEG and JPEG2000 (Jasper) implementations, in both lossy and lossless mode. They conducted 2268 genuine and 283122 impostor comparisons with the CASIA images and, respectively, 19000 and 1980000 comparisons with the BATH set. The study reported first and second order statistics for the Hamming distance distributions and stated performance in terms of equal error rate and the true accept rate at the threshold that gave the lowest combined error count. The results show elevation of genuine scores but insensitivity of impostor scores under compression. For CASIA, this results in elevation of equal error rate, but this is not observed for BATH even at 50:1 compression.

International Biometric Group As part of their ITIRT study of camera interoperability[32], IBG applied JPEG2000 compression to subset of the collected images and executed approximately 4600 comparisons with 24 enrolled samples. They observed changes in more than 99% of the Iridian recognition engine's Hamming distances (HD) under compression, noted that HDs tended to decrease, and attributed this to loss in quality, changes in image localization, and format conversion. However within the 6:1 lossy compression limit established by the INCITS 379 there was little impact. The investigation did not describe the effects separately for genuine from impostor comparisons.

Carinthia Technology Institute In 2007 Matschitsch, Tschinder, and Uhl reported[43] on the performance of the Masek[42, 41] algorithm as applied to seven CASIA 1.0[12] images from 20 persons. The study considered five compression algorithms JPEG, JPEG 2000, SPIHT, PRVQ and a fractal method. The authors advocated peak signal-to-noise ratio (PSNR) as a measure of compression damage. The study reported the dependence of genuine and impostor scores on the number of bits per pixel, and concluded that FMR is unaffected by compression.

3. Compression algorithms

IREX made use of three compression algorithms. These were applied to images by NIST in a compression-decompression fashion so that the IREX SDKs were *never* exposed to a compressed data stream and were therefore not required to supply or call compression libraries. This removed any chance of compression-related interoperability problems.

Three compression algorithms were used. The first is lossless and has zero effect on accuracy. The remaining two are lossy algorithms and will effect accuracy.

PNG The PNG specification was finalized as a W3C Recommendation on 10 November 2003. This second edition was formally standardized as ISO/IEC 15948:2003. PNG is an extensible file format for the lossless, portable, well-compressed storage of raster images. It provides a patent-free mechanism to store greyscale (or color) iris images with pixel depths on $[1, 16]$ bits per component. PNG is allowed in the ISO/IEC 19794-6 revision. IREX used version 1.2.10 of libpng[16].

JPEG The venerable JPEG algorithm is widely used for non-biometric purposes. It is formally standardized as ISO/IEC 10918-1 and ITU-T Recommendation T.81. IREX used version 6b of the Independent JPEG Group's implementation. It is freely available[17] and very widely used.

JPEG 2000 The JPEG2000 standard is formally standardized as ISO/IEC 15444. It has many advanced features and is well suited for lossy compression of biometric data[18]. For IREX we used the JASPER implementation in its default configuration.

[16] See http://libpng.org.
[17] From http://www.ijg.org/, downloaded June 22, 2009.
[18] See http://www.jpeg.org/jpeg2000/.

Organization Name	Letter Code	KIND 1 (required)	KIND 3 (required)	KIND 7 (optional)	KIND 16 (optional)	Elliptical models iris + pupil	Scalar image quality	OS
Sagem	A	2	2	2	2	2	2	Win.
Cogent	B	2	2	2	2	0	0	Win.
Cross Match	C	2	2	2	0	0	0	Linux
Cambridge	D	2	2	2	0	0	2	Linux
L1 ID	E	2	2	2	0	0	0	Win.
Retica	F	1	1	1	1	0	0	Win.
LG	G	2	2	2	2	0	2	Win.
Honeywell	H	2	2	2	2	2	0	Linux
Iritech	I	2	2	2	2	2	0	Linux
Neurotechnology	J	2	2	2	2	2	0	Win.
No. Suppliers	10	10	10	10	7	4	3	
No. SDKs		19	19	19	13	8	6	

Table 4: IREX provider participation and functionality.

Dataset ID	Origin	Number of subjects	Number of images	Camera Citation
OPS	An operational set	8160	32640	Pier 2.3[19]
ICE	Extract of Notre Dame 2005-2006 ICE images	240	60000	LG IrisAccess 2200[20]
BATH	University of Bath UK	800	Dedicated	[45]

Table 5: Summary of the IREX datasets.

4. Participation

Participation in IREX was open worldwide, to individuals, universities, non-profits, technology providers and technology integrators. The deadline for a declaration of intent to participate was July 7, 2008. Ten organizations elected to participate, as shown in Table 4. There was no fee to participate. There were no qualification criteria for entry other than the ability to send an implementation conformant to the API specifications. There was no requirement to have any membership in any standards body. There was no requirement to sell or otherwise reproduce any of the IREX technology to other parties.

- ▷ The IREX participants are identified by their full name in Table 4, and by a letter code and an abbreviated name in the running footer of each page.
- ▷ The test was conducted in three phases. The first was an iterative validation and correctness phase intended to assure both NIST and the participant that the SDKs were operating correctly. The second was intended as the definitive IREX test. It was supplanted with the third and final phase when certain unrepresentative failures and conformance problems were resolved.

5. Datasets

The IREX study employed data from three collections of iris images. These are summarized in Table 5 and described in the following subsections.

5.1. The OPS dataset

The operational dataset consists of two captures of the left and right irises of 8160 individuals. This gives a total of 32640 distinct images. The images were collected using the PIER 2.3 camera from Securimetrics, now a division of L1 Identity Solutions. The files were extracted from a large multimodal dataset, according to a fixed criterion. This was applied by the provider of the data. The authors did not have any role in the selection process. The selection criteria were such that a person was included in the IREX partition if the following logical expression was TRUE.

RULE 1 **and** (RULE 2 **or** RULE 3 **or** (RULE 4 **and** RULE 5))

Where

RULE 1 the subjects ten-print fingerprints were matched by an operational AFIS system[21] at some threshold

RULE 2 the subject's pair of left eyes matched with iris recognition algorithm X at a score below threshold τ_1

RULE 3 the subject's pair of right eyes matched with iris recognition algorithm X at a score below threshold τ_1

RULE 4 the subject's pair of left eyes matched with iris recognition algorithm X at a score below threshold τ_2

RULE 5 the subject's pair of right eyes matched with iris recognition algorithm X at a score below threshold τ_2

It is known that $\tau_2 > \tau_1$. The provider of the OPS dataset stated that using X with the τ_1 and τ_2 threshold produced "zero FMR" in O(10^{10}) comparisons. The authors assume that

- this applied to person-pairings, i.e. a second (L,R) pair was bound to a first (L,R) pair with zero false pairings;
- ground truth was defined by the AFIS implementation; and
- single images may still come from different persons because of RULES 2 and 3.

The algorithm X was supplied by an IREX participant, Y. The identity of Y is not disclosed in this report. The image localization, feature representation and matching algorithms used in X are unknown. The nature of the IREX algorithms from provider Y (and all other providers) are also not known (IREX is a black box test). While the known use of an iris recognition algorithm in the construction of a test dataset is not best-practice, it is allowed by the ISO/IEC 19795 standards provided the practice is disclosed. Nevertheless, this use raises the follow questions:

- Did the use of X-selected data bias IREX toward Y? The answer is that no dominant bias is apparent. This answer is provided with no explanation beyond empirical observation that the performance measurements attributed here to Y are *never* amongst the best.
- Did the use of X-selected data bias IREX toward providers using similar algorithms and feature sets to those used in Y? The answer is unknown because the authors have no knowledge of any of the SDK feature representations. However, if we assume that some IREX implementations have an algorithmic lineage similar to that of X (via licensing, for example), and that failures are not matcher independent, then yes a bias may be introduced.
- Did the involvement of Y with the provider of the OPS dataset improve Y's capability? This in unknown to the authors because we know nothing of the relationship between Y and the OPS provider.

In summary, the application of the X algorithm to the selection process means that recognition accuracy is likely to be high - more specifically that the iris left and right eye *pairs* are matchable (at some threshold). Thus we anticipate that any L-R fusion procedure should give error rates of zero for iris recognition algorithms of similar capability to X. Critically, however, this does not hold for single *images*, and non-zero matching error rates should be expected and are, in fact, observed. Given these factors, the authors used the dataset because:

- The images are likely to be more representative of **enrollment** samples in which care had been taken to produce a pristine and matchable image. This makes the images suitable for the compression aspects of IREX because an identity credential would normally be populated with such images.

▷ The purpose of the IREX test is not prediction of operational performance - the IREX goals are to give quantitative support to the ISO/IEC 19794-6 standard, and to the implementers and adopters thereof. The dataset is suitable for demonstrating changes in performance (e.g., under compression, or across algorithms).

▷ The population size, at 8160, is a factor of 34 larger than the ICE dataset, and while this is invaluable to capture the natural variation between persons in order to better characterize false match performance, it remains too small to support robust quantitative estimation of false match performance in national-scale 1:N applications.

▷ The ICE dataset has its own detracting properties (see section 5.3).

▷ Images from the ICE dataset have been disseminated publicly and thus IREX providers perhaps anticipated and tuned to that kind of image. The IREX test plan hinted that these images might be used by including a code for the LG2200 camera in the published IREX API document. These disadvantages motivated use of the OPS dataset.

5.1.1. The all-failures partition

Given its provenance, the OPS dataset might be considered easy: many of the images will never be involved in a failed comparison. This affords an opportunity to produce a smaller dataset in which errors are concentrated. This subset was constructed to reduce the computational cost of the measurements of cross-supplier interoperability given in section 8.6. The data construction procedure was as follows. Given $N = 16320$ genuine comparisons, find the set B_i for SDK i of KIND 1 vs. KIND 1 comparisons that are falsely non-matched at the default threshold [22] of FMR = 10^{-4}. The resulting dataset is then the union $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 ... \cup \mathcal{B}_N$.

This is comprised of 1335 genuine image pairs from 1144 subjects. Unless otherwise stated, these are used by taking the first image of the pair and comparing with *all* members of the OPS dataset including the second mated member of the pair.

5.1.2. Operational relevance of the OPS dataset

The fact that the IREX set of the operational OPS dataset has already been matched (at some threshold) means the images are clean - false rejection will be less frequent than if no matcher had been used. That said, if the original OPS collection policy had embedded a matching phase[23] then localization failures on the resulting corpus would be much rarer and performance is likely to be better than for a collection that did no such thing.

5.2. The bath dataset

NIST was provided with images of individuals collected by the University of Bath in the United Kingdom. The images were collected[45] using a computer vision camera (not a commercial iris product) at a high resolution such that the uncompressed greyscale eight bit raster images had size 1280 x 640 pixels across the peri-ocular region. The main dataset is comprised of 29525 images from 800 individuals. This does not include the images held in directories labeled *NonIdeal* which were ignored throughout.

All of the raw images were downsampled to 640 x 480 via 2 x 2 neighborhood averaging. This made the images conformant to the IREX test plan specification. The images were then passed to the I1 SDK to prepare KIND 3 instances. The record headers included iris diameter estimates, the histogram of which is given in Figure 5. The distribution is clearly bimodal. Images with an iris diameter in excess of 340 pixels were omitted from the IREX sample. The effect of this operation reduced the number of images to 23055 and the number of subjects to 664. This is what is used for all IREX analyses.

[22] The default threshold is that computed using all OPS vs. OPS and OPS vs. ICE comparisons of uncompressed KIND 1 images.

[23] For example, consider an enrollment process that includes capture of three images of an eye. If these samples are immediately matched to produce comparison scores, s_{12}, s_{13}, s_{23}, then the sample with the lowest sum ($m_1 = s_{12} + s_{13}, m_2 = s_{12} + s_{23}$, and $m_3 = s_{13} + s_{23}$) could be retained.

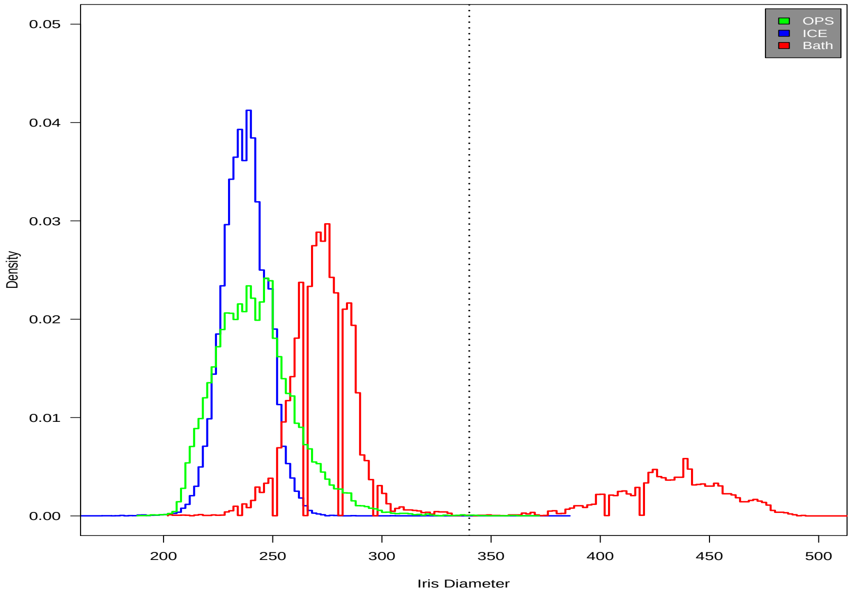


Figure 5: Histograms of the diameters of the irises in the images of the three IREX datasets. BATH images with iris diameter in excess of 340 pixels were not used in IREX.

Without the 340 pixel cap, some SDKs gave very high template generation failure rates because (we assume) the iris detection software was not configured to find large irises, or because insufficient margin surrounded the iris. Note that even if a template was produced, it may embed an incorrect segmentation which would usually lead to false non-match errors.

For the set actually used in IREX larger error rates were observed for atypically sized irises (see section 7.2).

5.3. The ICE dataset

This set of data was provided to the authors by the MBGC program[24]. Although the data was sequestered at the time IREX commenced, a representative disjoint set of images had been released under the ICE 05 development program.

The ICE corpus used in IREX consists of a left and right iris images collected from a university population over six semesters running from 2004 to 2006. The images were formally described[49] thus:

> The ICE 2006 images were acquired using an LG EOU 2200 iris scanner. The LG EOU 2200 is a complete acquisition system and has automatic image quality control checks. By agreement between U. of Notre Dame and Iridian, a modified version of the acquisition software was provided. The modified software allowed all images from the sensor to be saved under certain conditions, as explained below.
>
> The iris images are 480x640 in resolution, see Figure 2 [suppressed]. For most "good" iris images, the diameter of the iris in the image exceeds 200 pixels. The images are stored with 8 bits of intensity, but every third intensity level is unused. This is the result of a contrast stretching automatically applied within the LG EOU 2200 system. In our acquisitions, the subject was seated in front of the system. The system provides 32 recorded voice prompts to aid the subject to position their eye at the appropriate distance from the sensor. The system takes images in "shots" of three, with each image corresponding to illumination of one of the three infrared (IR) light emitting diodes (LED)s used to illuminate the iris.
>
> For a given subject at a given iris acquisition session, two "shots" of three images each are taken for each eye, for a total of 12 images. The system provides a feedback sound when an acceptable shot of images is taken. An acceptable shot has one or more images that pass the LG EOU 2200's built-in quality checks, but all three images are saved. If none of the three images pass the built-in quality checks, then none of the three images are saved. At least one third of the iris images do pass the Iridian quality control checks, and up to two thirds do not pass.
>
> A manual quality control step at Notre Dame was performed to remove images in which, for example, the eye was not visible at all due to the subject having turned their head.

The use of these images proved controversial in the ICE 2006 evaluation because the suppression of the camera's quality control apparatus caused operationally non-representative images (e.g., eyes closed, non-axial gaze, blur) to be present in the dataset. The presence of degraded images adversely affected iris recognition accuracy, and while larger error rates give better statistical significance to FNMR estimates, the test results are have less relevance to operational reality.

The authors' view is that whether or not to use ICE images depends on the purpose:

- If the purpose of the test is to determine the most capable fully automated "lights-out" processor of iris images then inclusion of such data is worthwhile and defensible. This might be more representative of applications that include a lightly constrained capture application.
- If the purpose is to assess the effect of compression on such images, or to assess the viability of algorithms on the core iris feature extraction and matching problem, then results from such images should be discounted.
- If, on the other hand, the aim is prediction or representation of operational performance then the images are suitable only to the degree that they quantitatively represent the geometric and photometric properties of those in the intended application. For contemporary mainstream identity management applications, that representativeness is very poor because of the subversion of the quality control apparatus, the presence of interlacing artifacts[24], the use of contrast stretching[17] and, not least, because the LG 2200 has been obsolete for several years[25].

5.4. On the role of technology testing

The following topics are advanced for consideration when reviewing the scope of the IREX results.

- **Operational relevance:** The predictive value of offline tests for operational performance is limited to cases where the actual operation is adequately reflected by the images themselves. Such is the case for latent fingerprint images

[24] The ICE images *do* exhibit interlacing artifacts - it is not known whether these arose in the camera itself or in ancillary sampling equipment. The magnitude of interlacing artifacts is presentation (and hence subject) dependent. It can reduce the vertical resolution by two times, and lead to segmentation difficulties.

[25] Note, however, that legacy databases may exist.

which are collected and used in an offline test exactly as they might be in an operational AFIS search. But a technology test is poorly suited for estimating the operational aspects of a cooperative access control application in which a user interactively presents one, or conditionally, two eyes (or fingers) to the system which, in turn, responds with prompts, feedback, and interim and final pass-fail decisions. The conventional approach to testing such a system is to run a scenario test[26] in which the intended application is modeled with a sufficiently large population of live subjects who attempt to use the systems in the intended manner. This was the practice in the DHS-funded ITIRT study[32].

▷ **Comparative testing:** Offline tests are, however, very well suited for comparing the algorithmic performance and properties of the core technologies embedded in systems. Offline tests proceed by running provider software on sequestered archived biometric data. This allows massive datasets and populations to be used - this affords statistical significance. More importantly it provides for a level playing field in which competing implementations are evaluated in an identical manner. Critically the tests are repeatable, as required by the scientific method.

▷ **Blind testing:** Tests like IREX, which work on sequestered sets of biometric samples, often proceed without any prior dissemination of training data to the test providers. This places developers in the invidious position of having to ship test SDKs which must be capable of processing images without having been specifically built, trained or parametrized to do so. This is a frequent refrain heard from prospective NIST test participants (for iris, and other modalities) and indeed at least one IREX provider addresses this issue in the respective IREX APPENDIX. The blind nature of testing undermines the representativeness of the test results to operational reality where, it is assumed, the software could be tailored in the design phase or in actual operation.

The counterargument, however, is that offline tests conducted in this manner constitute a forcing function for development of algorithms that are tolerant to known image variations. This aspect, universal interoperability and camera-matcher independence, is explicitly the aim of a standard interoperable iris image format, and the IREX test is well motivated in this respect. That said, the authors consider that the industry would be better served if standard reference datasets were available to developers.

▷ **Provider bias:** A further question on bias occurs because the manufacturer of the PIER camera, L1, used for the collection of the OPS dataset is also an IREX participant[27]. Similarly the manufacturer of the camera used for collection of the ICE images, LG Iris, is also an IREX participant.

In each case the camera manufacturer is theoretically afforded an advantage in that they would likely have had an opportunity to gain experience in developing algorithms for these images. However a set of LG 2200 images, from ICE 05, were released by NIST and in principle available to all IREX developers.

Note also the discussion, in section 5.1, on possible bias introduced for a specific provider (and a specific class of algorithms) by use of the OPS dataset.

▷ **Comparison to other modalities:** There is considerable interest in what modalities offer the best recognition performance - *"what is the best biometric"*? This is usually taken to mean which biometric is most discriminating between people (as needed in one-to-many applications, and in very low FMR one-to-one cases). Probably the best method for comparing modalities on this aspect is to apply contemporary algorithms from each modality to operationally collected or representative images or signals. The IREX-style technology tests are well suited to assessing performance on large populations.

[26] While the international standard for scenario testing ISO/IEC 19795-2 does not (and cannot) require retention of captured images, it would be clearly worthwhile in the case described.

[27] More specifically the camera is made by Securimetrics a division of L1 Identity solutions, and the IREX software was submitted by the Biometric Research division of L1.

6. Metrics

The direct and proper way to quantify accuracy and interoperability is in terms of false non-match and false match error rates, FNMR and FMR. The quantities are computed empirically. If d denotes a matcher dissimilarity score obtained by comparing two samples from the same person, and $M(\tau)$ is the number of such scores above threshold, τ,

$$M(\tau) = \sum_{d \in \mathcal{G}} H(d - \tau) \tag{1}$$

where $\mathcal{G}$ denotes the set of all genuine comparison scores. and $H(x)$ is the step function defined here as

$$H(x) = \begin{cases} 0 & x \le 0 \\ 1 & x > 0 \end{cases} \tag{2}$$

The inequality placement is unconventional (for the Heaviside step function) and is used so that scores equal to the threshold correspond to acceptance. FNMR is then the fraction of genuine comparisons for which the score is above the operating threshold:

$$\text{FNMR}(\tau) = \frac{M(\tau)}{M(-\infty)} \tag{3}$$

where $M(-\infty)$ is just the number of genuine comparisons considered. Likewise, when d denotes a score obtained by comparing samples from different persons, and $N(\tau)$ is the number of such scores below threshold, τ,

$$N(\tau) = \sum_{d \in \mathcal{I}} 1 - H(d - \tau) \tag{4}$$

where $\mathcal{I}$ denotes the set of all impostor scores. FMR is then the fraction of impostor comparisons resulting in a score less that or equal to the operating threshold:

$$\text{FMR}(\tau) = \frac{N(\tau)}{N(\infty)} \tag{5}$$

where $N(\infty)$ is the number of impostor comparisons conducted. In 1:N negative identification applications (e.g., watch-list, duplicate detection) FMR measures the rate at which a search sample is incorrectly associated with an enrolled sample. In 1:1 positive authentication applications FMR is regarded as a measure of security, i.e. the fraction of illegitimate matching attempts that result in success. In any case, these error rates must be understood as being *matching* error rates, not *transactional* rates. The ISO/IEC SC 37 Working Group 5 has established different terms for these rates: FMR and FNMR refer to comparisons of single samples, while FAR and FRR apply to the outcome of a human-system transaction in which a user might, for example, make multiple attempts with multiple eye presentations.

6.1. Treatment of failure to enroll and failure to acquire

The previous section defined error rates in terms of the logical comparison of samples, without considering the intermediate templates. The failure of an implementation to produce a template (or an intermediate image) from an input parent image may be elective (e.g., an image is assessed to have unusable quality) or otherwise (e.g., the software crashes). In any case, a testing laboratory, seeking to establish a level playing field in a comparative test, cannot ignore such events because downstream matching accuracy can be improved if they are used to exclude the worst images from the matching phase. The ISO/IEC 19795 standard, *Biometric Performance Testing and Reporting - Part 1: Principles and Framework* addresses this issue by observing, "Comparison of systems having different failure to enroll rates may require use of generalized false reject and false accept rates which combine enrollment, sample acquisition and matching errors". It

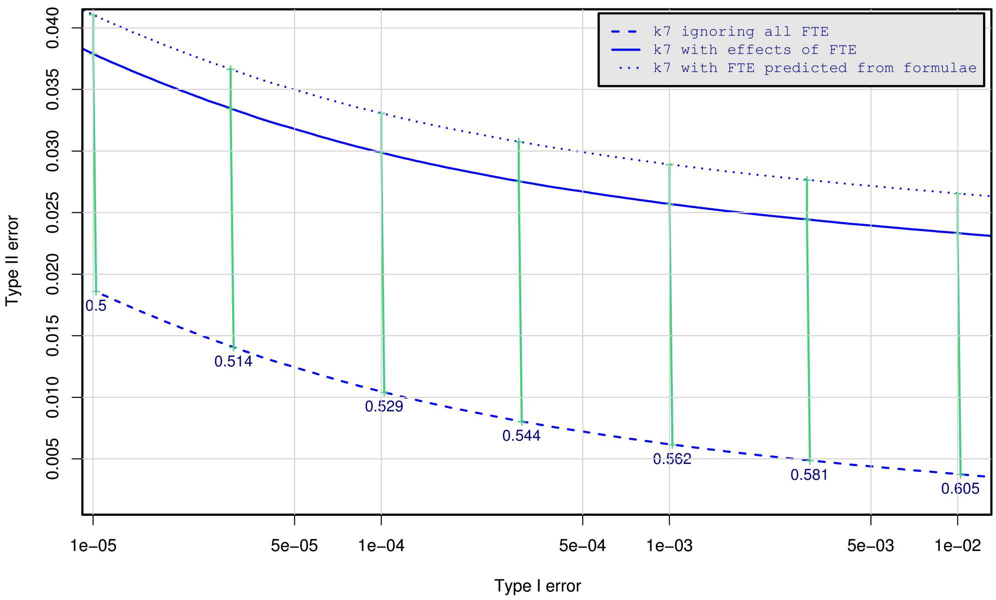


Figure 6: For preparation of KIND 7 records from ICE images, one IREX SDK fails to produce templates from verification images at rate FTE $= 0.0085$ and fails to produce templates from enrollment images at rate FTE $= 0.0145$. The figure shows: As the **middle solid** line, the DET that includes the effects of template generation failure; As the **bottom dashed** line, the DET obtained for matching of the properly generated templates *only*. The top dotted line is the result of apply equations 6 and 7 to bottom line with the scalar FTE and FTE values. The green links join points of constant threshold. A non-vertical link indicates a change in FMR. All results apply to native operation, and the effects of FTE are included.

continues by including a requirement to report how error rates are generalized in comparative testing.

Consider a full comparison of all samples in an enrollment set with all those in a verification set. If the fraction of missing enrollment instances is the failure-to-enroll rate, FTE, and the fraction of missing verification instances is the failure-to-acquire rate, FTA, the generalized Type I and II performance metrics are

$$\text{GFAR}(\tau) = (1 - \text{FTE})(1 - \text{FTA})\text{FMR}(\tau) \quad (6)$$

$$\text{GFRR}(\tau) = \text{FTA} \;+\; (1 - \text{FTA})\left[\text{FTE} \;+\; (1 - \text{FTE})\text{FNMR}(\tau)\right] \quad (7)$$

The second formula says that an individual is falsely rejected if either the verification sample resulted in an FTE or, if it didn't, that either the enrollment sample failed to enrolenroll or, that the two samples produced a false non-match. If, for example, an IREX algorithm gave FTE $=$ FTA $= 0.012$, then for an operating point (FMR , FNMR) $= (0.0001, 0.01)$, the generalized rates become (GFAR, GFRR) $= (0.000098, 0.0336)$. This is shown in Figure 6. The failure to produce 1.2% of the templates adds almost double that to the false rejection rate. This is because when $f =$ FTE $=$ FTE is small, equation 7 is approximately GFRR $= 2f+$FNMR reflecting the absence of enrollment and verification samples.

The formulae assume that samples are used in equal numbers of genuine comparisons. For OPS this is correct, but for the ICE and BATH datasets where subjects have differing numbers of images, samples are used in varying number of comparisons and the formulae do not hold. This is illustrated in Figure 6 which shows three DET characteristics. The center, solid, plot treats comparisons for which one or both templates were not produced as producing a high score which guarantees rejections). The lower, dashed, plot ignores such events; and the top dotted line is the result of applying equations 6 and 7. The green links join points of fixed threshold. The conclusions are that false acceptance varies little, but that false rejection varies considerably, and that the formulae overestimate the genuine error rates.

Given this discussion, IREX, unless stated otherwise, accounts for this tradeoff by regarding any comparison involving missing enrollment or verification templates as one that would produce a comparison score of, essentially, ∞. This maintains existing practice in NIST tests.

6.2. Thresholds for zero FMR

The attraction of iris recognition, as articulated in the literature, is that the impostor distribution is stable enough and sufficiently well characterized that a threshold may be set to give known, and very low, false match rates. While it is the case with all biometric algorithms and modalities that the operating threshold can be set to give FMR values arbitrarily close to zero, there are three practical consequences of doing so:

1. Elevated FNMR , such that for some modalities and applications, the false rejection rates may be so high that the threshold is utterly untenable. For some applications a high FNMR may be tenable - for example in a watchlist application[1] in which it is unacceptable to falsely detain someone, a 60% FNMR may still be attractive if the FMR can be maintained near zero. In contrast a physical access control system configured to give zero FMR might produce unacceptable FNMR and this may result in the decommissioning of the system. In iris recognition specifically, FNMR may be elevated if the sensing technology is insufficient to faithfully capture and digitize the iris. This has been demonstrated to occur

   ▷ when short wavelength light (e.g., visible) is absorbed by the melanin pigmentation of dark eyes such that the iris structure is not revealed,

   ▷ when subjects are imaged far from the focal plane,

   ▷ when subjects' gaze is not on the optical axis, or

   ▷ when the iris is occluded by glasses or patterned contact lenses.

   Commercial iris recognition cameras are specifically designed to mitigate all of these effects.

2. If the impostor distribution changes with respect to, for example, environmental changes or changes in the imaging system, the FMR may change.

3. Extreme value effects: Given long enough operation, the system sooner or later will report a false match. This depends on just how low (or high, for fingerprints) that the threshold is set. The problem exists in one-to-many systems in which new persons are continually enrolled without change of threshold. The threshold should generally be set depending on the size of the enrolled population, a step normally intended to render a fixed selectivity, i.e. a constant number of false matches (typically zero)[16].

The calibration results in section 7.4 are intended to assist users in setting thresholds for very low FMR .

The practice of setting FMR to 0.0001 is not intended as any indication of a suitable operating point. As with other biometric modalities the threshold should be set according to application requirements.

6.3. Decidability measures

A measure of the separation of the genuine and impostor distributions is the d-prime measure

$$d' = \frac{\mu_I - \mu_G}{\sqrt{\frac{\sigma_I^2+\sigma_G^2}{2}}} \qquad (8)$$

defined in terms of the mean μ and variance σ^2 of the impostor (I) and genuine (G) distributions. This is appropriate when the distributions are Normal but is less relevant otherwise because it does not capture the functional form at the overlap of the tails. Nevertheless, we use the following related quantity to quantify how compression causes genuine scores to degrade and move closer to the impostor distribution.

$$d'(C) = \frac{\mu_I - d_G(C)}{\sigma_I} \tag{9}$$

where $d_G(C)$ represents a genuine dissimilarity score involving an image lossy compressed with parameter C. This quantity is the number of standard deviations the genuine score lies from the impostor distribution.

6.4. IMAGE-SPECIFIC ERROR RATES

It is known that different users exhibit different levels of recognizability in biometric recognition systems. Some people are easy to recognize, while others can impersonate or be impersonated. The literature makes the analogy between the various Type I and Type II error rate heterogeneities and a biometric zoo.

The issue of performance variability among different users was first addressed by Campbell et al.[11]. Later, Doddington et al.[21] developed a statistical framework to identify four categories of speakers based on the recognition error of each speaker. Specifically, they introduced:

- **Sheep** - users who are recognized easily
- **Goats** - users who are particularly difficult to recognize
- **Lambs** - users who are particularly easy to imitate
- **Wolves** - who are users that are particularly successful in imitating others.

Others [62, 64, 59, 29] have investigated the existence of a biometric menagerie in face and fingerprint recognition systems. More recently Yager and Dunstone[66] introduced four new groups of animals. Recognizing the user-dependent performance variability, Poh et al.[52] ranked users based on the strength of their performance and used that information to do fusion on a per-user basis.

This non-uniform performance is of interest to the designers of biometric recognition systems. The difficult-to-recognize users are responsible for the major share of biometric errors. Goats contribute to FNMR but this poor performance in genuine comparisons does not necessarily elevate FMR. Goats are particularly problematic in access control systems where reliable, convenient, verification of users is the main interest (i.e. low FNMR is desirable). Wolves and lambs adversely affect the security of biometric systems by contributing to the FMR. Their biometric samples tend to match impostors, or be matched by impostors.

Similarly, different images of the same subject could exhibit different levels of matchability. Image performance variation is often ascribed to the capture device (e.g., different physical imaging properties of sensors), the environment (e.g., low light) or the user (e.g., squinting), and the thrust of research is therefore to make recognition algorithms more tolerant of such variations. Stated another way, algorithms that bound or constrain FNMR and FMR are more reliable and secure. To reduce false non-matches and improve reliability, it is a common policy to allow multiple acquisitions of the same biometric at the time of authentication (e.g., to re-acquire after a moistening of the finger). Dealing with false match occurrences, however, is a more difficult problem. In operational verification systems, false matches are likely to be undetected; in identification systems they lead to spurious entries on candidate lists and these elevate workload.

To examine performance variation among different images, we define the following image-specific error rates

- **Image false match rate**, iFMR - the proportion of comparisons for which an image produces false matches (i.e. non-match comparisons at or below the operating threshold).

- **Image false non-match rate**, iFNMR - the proportion of comparisons for which an image produces a false non-match (i.e. genuine comparisons above the operating threshold).

Specifically, if we define s^{ij}_{kl} to be the comparison score of the k-th image of subject i with the l-th image of subject j then the set of impostor scores of the k-th image of subject i is

$$\mathcal{I}(i,k) = \{\ s^{ij}_{kl}\ ,\ i \neq j\ ,\ j = 1 \dots J\ ,\ l = 1 \dots N_j \} \tag{10}$$

for comparison against all N_j images of all J persons in an enrolled set. The image false match rate is then defined as

$$\text{iFMR}\ (\tau, i, k) = \frac{\sum_{s \in \mathcal{I}(i,k)} \quad 1 - H(s - \tau)}{\sum_{s \in \mathcal{I}(i,k)} \quad 1} \tag{11}$$

where $H(s)$ is the step function of equation 2. If the threshold is set to τ in the conventional manner (i.e over some large cross comparison set) to give a global FMR of f, then the general case is that iFMR $\neq f$.

For the image false non-match rate, we use the set of non-self genuine scores of the k-th image of subject i

$$\mathcal{G}(i,k) = \{\ s^{ii}_{kl}, l = 1 \dots N_i, k \neq l\ \} \tag{12}$$

to compute

$$\text{iFNMR}\ (\tau, i, k) = \frac{\sum_{s \in \mathcal{G}(i,k)} \quad H(s - \tau)}{\sum_{s \in \mathcal{G}(i,k)} \quad 1} \tag{13}$$

where $H(x)$ is again the step function of equation 2.

Unless otherwise stated iFMR and iFNMR are computed for each comparison algorithm by substituting comparison scores of the algorithm in equations 11 and 13 above using the following datasets.

- The threshold can be set to any value. Here it is set over all OPS - to - OPS and OPS - to - ICE impostor comparisons to achieve FMR = 0.001.
- iFNMR is computed over the ICE dataset. It cannot be computed over the OPS dataset because only one genuine comparison is available per subject-eye.
- iFMR is computed for each ICE image by comparing it with 16320 OPS images.

Further we compute an *aggregate* iFMR as the arithmetic mean of image false match rates over comparison algorithms. Similarly the aggregate iFNMR is the arithmetic mean of image false non-match rates over the same algorithms.

7. Performance on full-size uncompressed images

This section gives the quantitative results for “traditional” rectilinear iris imagery, i.e. 640x480 greyscale images without application of compression. Such images are packaged as KIND 1 instances. This is advanced as a baseline ahead of the corresponding results for the cropped, cropped and masked, and polar IREX compact formats.

7.1. Failure to enroll

In the IREX context a *failure to enroll* refers to the failure of a SDK function invocation to produce the anticipated output. Two cases apply: The failure to convert a 640x480 input image to an IREX image record is counted as FTE. The failure to

SDK	OPS				ICE				BATH			
	k1	k3	k7	k16	k1	k3	k7	k16	k1	k3	k7	k16
A1	0.000 15 0	0.000 0 0	0.000 0 0	0.000 0 0	0.002 146 0	0.000 4 4	0.000 9 9	0.000 0 0	0.000 0 0	0.001 30 30	0.001 30 30	0.000 0 0
A2	0.000 15 0	0.000 0 0	0.000 0 0	0.000 0 0	0.002 146 0	0.000 4 4	0.000 9 9	0.000 0 0	0.000 0 0	0.001 30 30	0.001 30 30	0.000 0 0
B1	0.000 0 0	0.000 0 0	0.000 0 0	0.000 1 0	0.000 0 0	0.000 1 0	0.000 1 0	0.000 4 0	0.000 0 0	0.000 0 0	0.000 0 0	0.000 8 0
B2	0.000 0 0	0.000 0 0	0.000 0 0	0.000 1 0	0.000 0 0	0.000 1 0	0.000 1 0	0.000 4 0	0.000 0 0	0.000 0 0	0.000 0 0	0.000 8 0
C1	0.001 30 0	0.013 426 16	0.003 89 16	-	0.001 95 0	0.003 191 81	0.002 173 81	-	0.013 310 0	0.224 5163 291	0.049 1128 291	-
C2	0.001 43 0	0.013 420 4	0.002 73 2	-	0.002 127 0	0.004 295 43	0.003 214 37	-	0.011 263 0	0.225 5186 237	0.048 1111 218	-
D1	0.000 15 0	0.000 12 0	0.001 18 0	-	0.006 425 0	0.005 346 0	0.003 206 0	-	0.018 406 0	0.019 441 0	0.016 379 0	-
D2	0.000 16 0	0.001 18 0	0.001 37 0	-	0.008 602 0	0.007 505 0	0.007 477 0	-	0.019 446 0	0.020 471 0	0.019 440 0	-
E1	0.001 30 0	0.001 30 7	0.001 33 0	-	0.012 851 0	0.012 883 72	0.012 836 0	-	0.029 658 0	0.045 1043 721	0.030 690 0	-
E2	0.001 30 0	0.001 30 30	0.001 33 33	-	0.012 851 0	0.012 883 883	0.012 836 836	-	0.029 658 0	0.045 1043 1043	0.030 690 690	-
F1	0.000 12 0	0.000 16 5	0.000 11 5	0.001 20 5	0.003 228 0	0.003 242 33	0.003 181 33	0.015 1087 33	0.044 1003 0	0.051 1183 996	0.044 1008 996	0.046 1063 995
G1	0.001 22 0	0.001 22 22	0.002 76 76	0.003 104 104	0.005 392 0	0.005 364 364	0.007 508 508	0.010 742 742	0.043 1000 0	0.044 1009 1009	0.053 1219 1219	0.102 2353 2353
G2	0.001 28 0	0.001 28 28	0.002 70 70	0.003 107 107	0.003 248 0	0.003 248 248	0.005 348 348	0.009 627 627	0.024 556 0	0.025 585 585	0.034 778 778	0.092 2109 2109
H1	0.002 53 0	0.002 56 53	0.002 79 55	0.002 60 53	0.000 11 0	0.000 35 24	0.001 42 24	0.000 13 11	0.005 117 0	0.012 273 249	0.013 288 252	0.005 122 117
H2	0.002 53 0	0.002 56 53	0.002 79 55	0.002 59 53	0.000 11 0	0.000 35 24	0.001 42 24	0.000 13 11	0.005 117 0	0.012 273 249	0.013 288 252	0.005 122 117
I1	0.000 0 0	0.000 0 0	0.000 0 0	0.000 0 0	0.000 4 0	0.000 4 4	0.000 4 4	0.000 4 4	0.000 0 0	0.000 0 0	0.000 0 0	0.000 0 0
I2	0.000 0 0	0.000 0 0	0.000 0 0	0.000 0 0	0.000 5 0	0.000 6 6	0.000 6 6	0.000 6 6	0.000 0 0	0.000 0 0	0.000 0 0	0.000 0 0
J1	0.001 28 0	0.000 13 13	0.000 13 13	0.000 13 0	0.002 167 0	0.000 21 21	0.000 21 21	0.000 21 0	0.016 364 0	0.016 364 364	0.016 364 364	0.016 364 0
J2	0.001 28 0	0.001 17 8	0.000 15 8	0.001 21 0	0.002 167 0	0.000 24 17	0.000 30 17	0.001 46 0	0.016 364 0	0.022 497 351	0.021 479 351	0.027 613 1

Table 6: The numbers and proportions of failed template and IREX record creation attempts. In each cell there are three quantities: First is the fraction of templates missing, second is the number of templates missing, and third is the number of IREX records that were not produced. By definition, this last value is less than or equal to the second. All rates refer to native operation, and are summations over the enrolment and verification image sets. **Colors:** Cells are shaded dark red when the topline FTE is above 1.0%, light red above 0.5%, light green below 0.1% and dark green with exactly zero errors.

convert an IREX record to a template is also an FTE. A failure to acquire (FTA) is defined identically and only differs from FTE in that FTE refers on enrollment samples and FTE refers to verification samples. Note that even if an SDK produces an output the content may nevertheless be incorrect, e.g., the resulting image is of the endocanthion and not the iris. Such semantic faults will not count as FTE because we can't automatically detect such events. Instead they will most often lead to false rejection errors.

In any case, the occurrence of FTE events is shown in Table 6 which reports failure for all SDKs producing and operating on all KINDS of images from all three datasets. All input images were uncompressed. The notable results are below.

▷ Across all datasets the I1 and I2 SDKs give essentially zero failures to enroll. The only failures are limited to ICE images for which the eye is closed or, in one case, not present at all. Additionally the B1 and B2 SDKs are almost perfect with A1 and A2 failing only on very few ICE images.

▷ Elsewhere the BATH dataset presents more problems than ICE which, in turn, is much harder than the OPS set. The failures in the BATH dataset are for prosaic reasons of size as documented extensively in section 7.2. The images in the BATH set used here have iris diameters less than or equal to 340 pixels.

▷ Failures are often concentrated in specific stages of processing: For example, with ICE images the E2 SDK fails to make some KIND 3 records. In contrast, the failures for the D2 SDK occur with ICE imagery during the production of templates from the KIND 3 records. In some cases errors occur in both phases (e.g., SDK J2 making KIND 3 from ICE images). This result is explained by the fact that both steps involve significant image analysis operations.

7.2. EFFECT OF IRIS RADIUS

The question of whether performance is affected by iris size motivated the analyses of Figures 7 and 8 for the BATH and OPS datasets respectively. The radius values used in this analysis are those reported by the I1 SDK in the headers of its KIND 3 records. Note that the parent BATH dataset contains many images of large size. NIST downsampled all 1280x640 parents to the 640x480 norm. As described in section 5.2, all IREX analyses are restricted to those images for which the final iris radius $R \leq 170$ pixels. The histograms of iris size for these images are shown in Figure 5. The notable results are as follows. (Larger figures for all SDKs and all datasets appear in the appendices of the IREX SUPPLEMENT[28].)

▷ From the 2D comparison-count maps in the bottom right hand corner, the IREX partition of the BATH dataset contains a greater occurrence of atypical iris sizes than is present in the OPS dataset. Particularly the BATH dataset has a significant number of cases where the enrollment and verification iris pairs are either both large or both small.

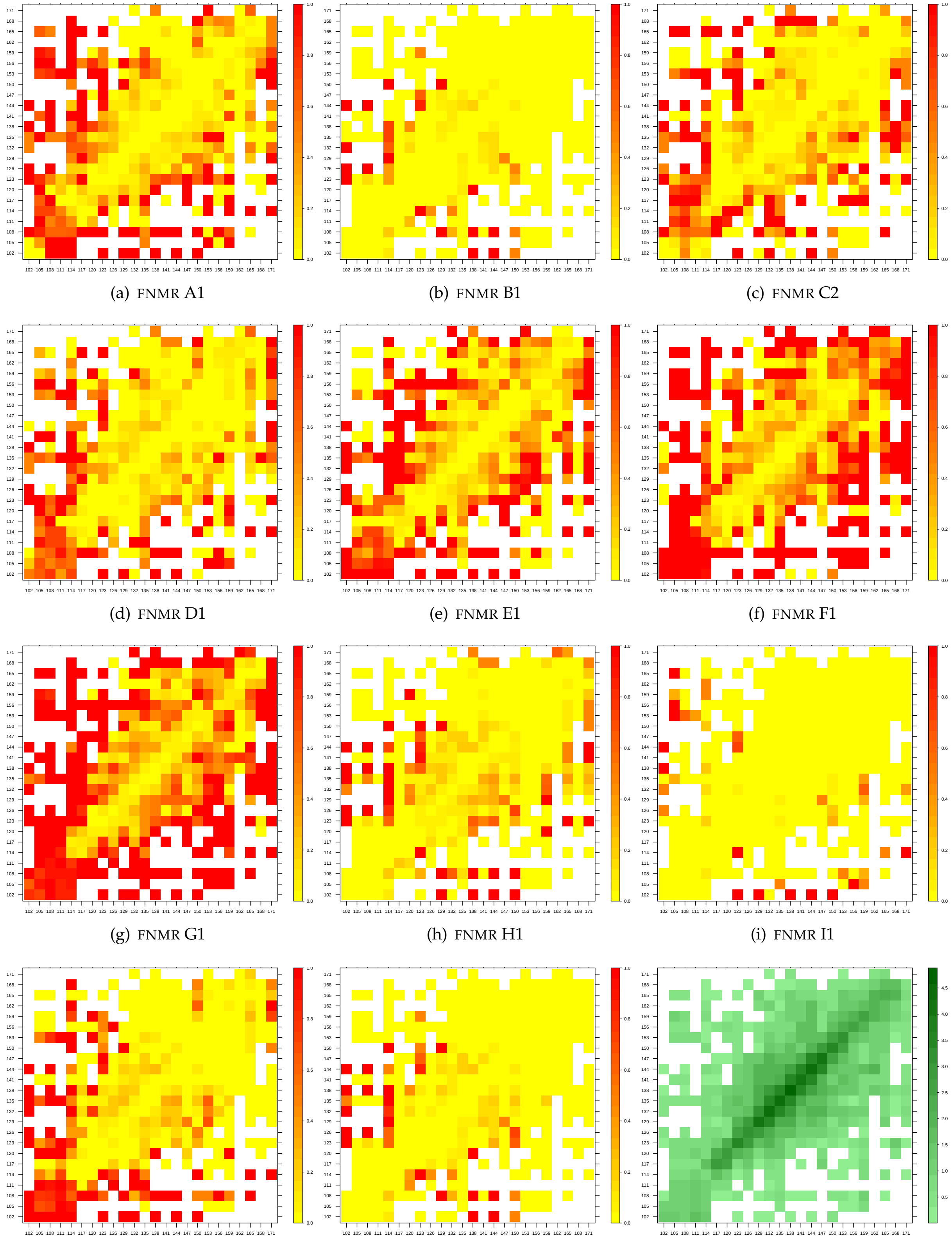
▷ The heat maps plot FNMR at a fixed FMR of 0.001 for KIND 1 vs. KIND 1 comparisons. Importantly, they treat template generation failures as false non-matches. The higher presence of red in Figure 7 vs. Figure 8 shows that there are fewer errors overall on the OPS dataset than on the BATH set.

▷ For the OPS images, the false rejections tend to occur when the radius of the iris is large. While it is possible (and likely for fixed focal-length cameras) that different sizes are symptomatic of other problems (e.g., out of focus images) a majority of the SDKs are immune to this.

▷ In the BATH dataset, several SDKs almost always fail on images with small iris radii. A larger number fail for the case when the two iris radii differ substantially (i.e. small vs. large).

These results have implications for the iris image standard. While 640x480 pixels is the de facto standard *image* size, the current ISO/IEC 19794-6 only guides on *iris* size in a best practice annex. This is likely to be remedied in the revised ISO standard, per IREX input. Performance-related image attributes will be further refined in the new ISO/IEC 29794-6 *Iris Image Quality* standard.

[28] http://iris.nist.gov/irex/irex_appendices.pdf.



(a) FNMR A1

(b) FNMR B1

(c) FNMR C2

(d) FNMR D1

(e) FNMR E1

(f) FNMR F1

(g) FNMR G1

(h) FNMR H1

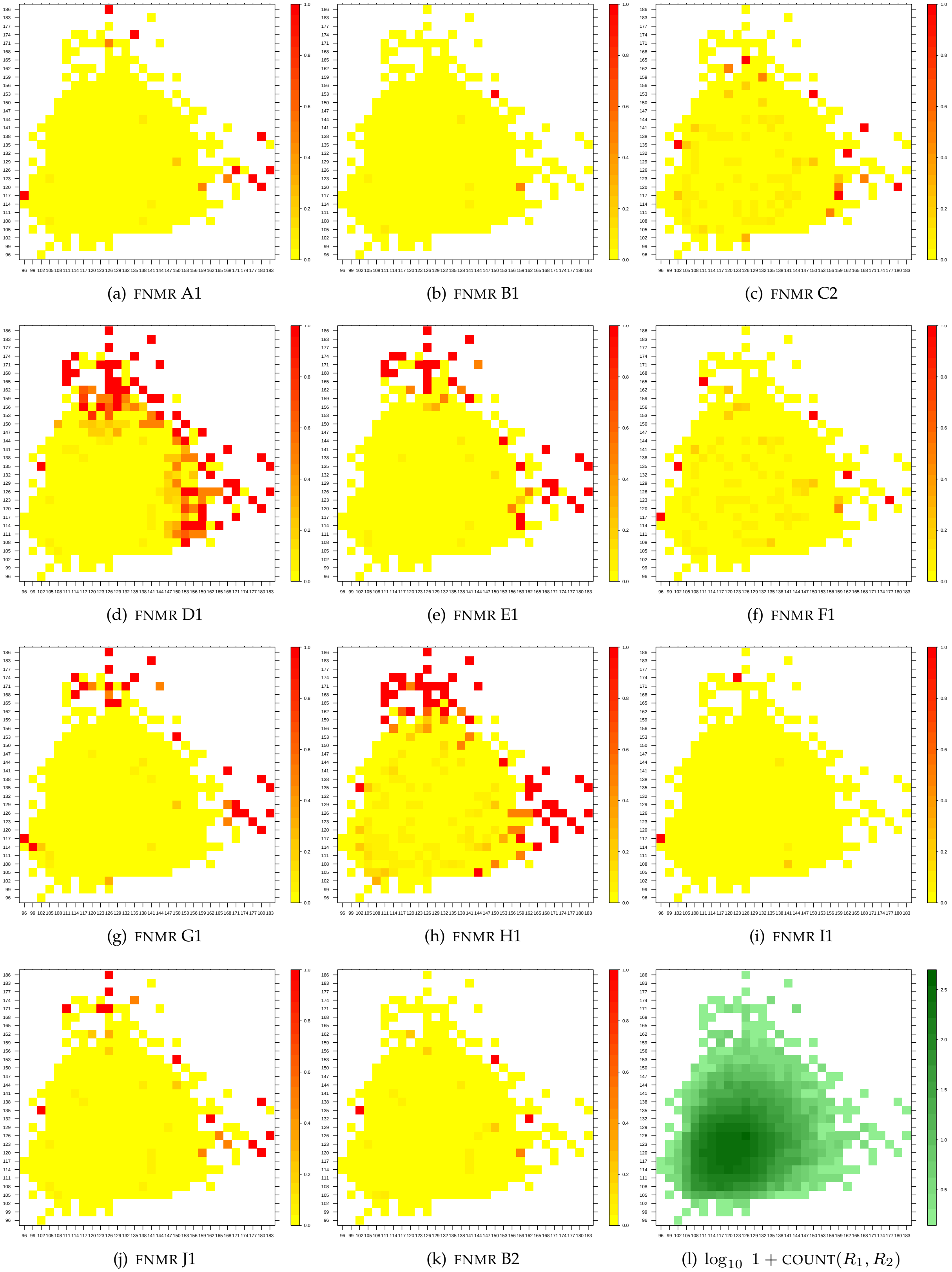
(i) FNMR I1

(j) FNMR J1

(k) FNMR B2

(l) $\log_{10}\ 1 + \text{COUNT}(R_1, R_2)$

Figure 7: For each primary SDK the figure shows the dependency of false non-match rate on iris radius for the BATH dataset. Each cell is the FNMR at FMR = 0.001 for enrollment samples on the y-axis and verification samples on the x-axis. The radii are quantized into three-pixel bins. The radii run $102 \leq r \leq 171$ pixels. Cells are uncolored when the dataset did not contain any comparisons with those radii. Elsewhere, particularly away from the diagonal and in the corners, the number of comparisons is sometimes small such that there is considerable error in the FNMR estimates. The number of the comparisons, on a log scale, is shown at bottom right. Larger images for each SDK on all three datasets appear in the respective appendices.



(a) FNMR A1 (b) FNMR B1 (c) FNMR C2

(d) FNMR D1 (e) FNMR E1 (f) FNMR F1

(g) FNMR G1 (h) FNMR H1 (i) FNMR I1

(j) FNMR J1 (k) FNMR B2 (l) $\log_{10}\ 1 + \text{COUNT}(R_1, R_2)$

Figure 8: For each primary SDK the figure shows the dependency of false non-match rate on iris radius for the OPS dataset. Each cell is the FNMR at FMR = 0.001 for enrollment samples on the y-axis and verification samples on the x-axis. The radii are quantized into three-pixel bins. The radii run $96 \leq r \leq 186$ pixels. Cells are uncolored when the dataset did not contain any comparisons with those radii. Elsewhere, particularly away from the diagonal and in the corners, the number of comparisons is sometimes small such that there is considerable error in the FNMR estimates. The number of the comparisons, on a log scale, is shown at bottom right. Larger images for each SDK on all three datasets appear in the respective appendices.

7.3. DET CHARACTERISTICS

As is typical in offline testing [4], this report computes the DET for all scores emitted by a tested algorithm[29]. This allows a survey over *all* operating points, τ, and contrasts with the operational situation in which the system is configured with a fixed operating threshold, against which a *decision* is rendered. The DET is a plot of FNMR(τ) against FMR(τ)[30] and, as the primary output of a biometric performance test, is vital in establishing the tradeoff between TYPE I and TYPE II errors.

7.3.1. RELATIVE DIFFICULTY OF DATASETS

For the purposes of comparing the three IREX datasets, Figure 9 shows, the DET characteristics for six primary SDKs [31] Note: The plots ignore comparisons where templates were missing (e.g., from FTE events), showing matching errors only. The intent is to show relative ordering of FNMR and slope. The notable observations are as follows:

- ▷ The ordering of the DET traces varies across SDKs. That is the implementations differ in which dataset gives the best accuracy. So while A1 finds OPS easier than ICE, which in turn is easier than BATH, B1 prefers BATH to ICE, and D1 prefers both of these to OPS. ICE is never best.
- ▷ At a fixed threshold, the observed FNMR values vary across the three datasets by as little as a factor of two (B1,H1,J1) to as much as an order of magnitude (C1).
- ▷ Similarly, at a fixed threshold, the observed FMR values vary across the three datasets by as little as a factor of two (A1) to as much as an order of magnitude (J1).
- ▷ The SDKs are much more consistent in the slopes of the DET characteristics. The slope is related to the separability of the genuine and impostor distributions, with a flatter DET being more desirable. The ordering for most SDKs, from flattest to steepest, is OPS, BATH, and ICE. The I1 and H1 SDKs prefer ICE to BATH.

7.3.2. RELATIVE PERFORMANCE OF ALGORITHMS

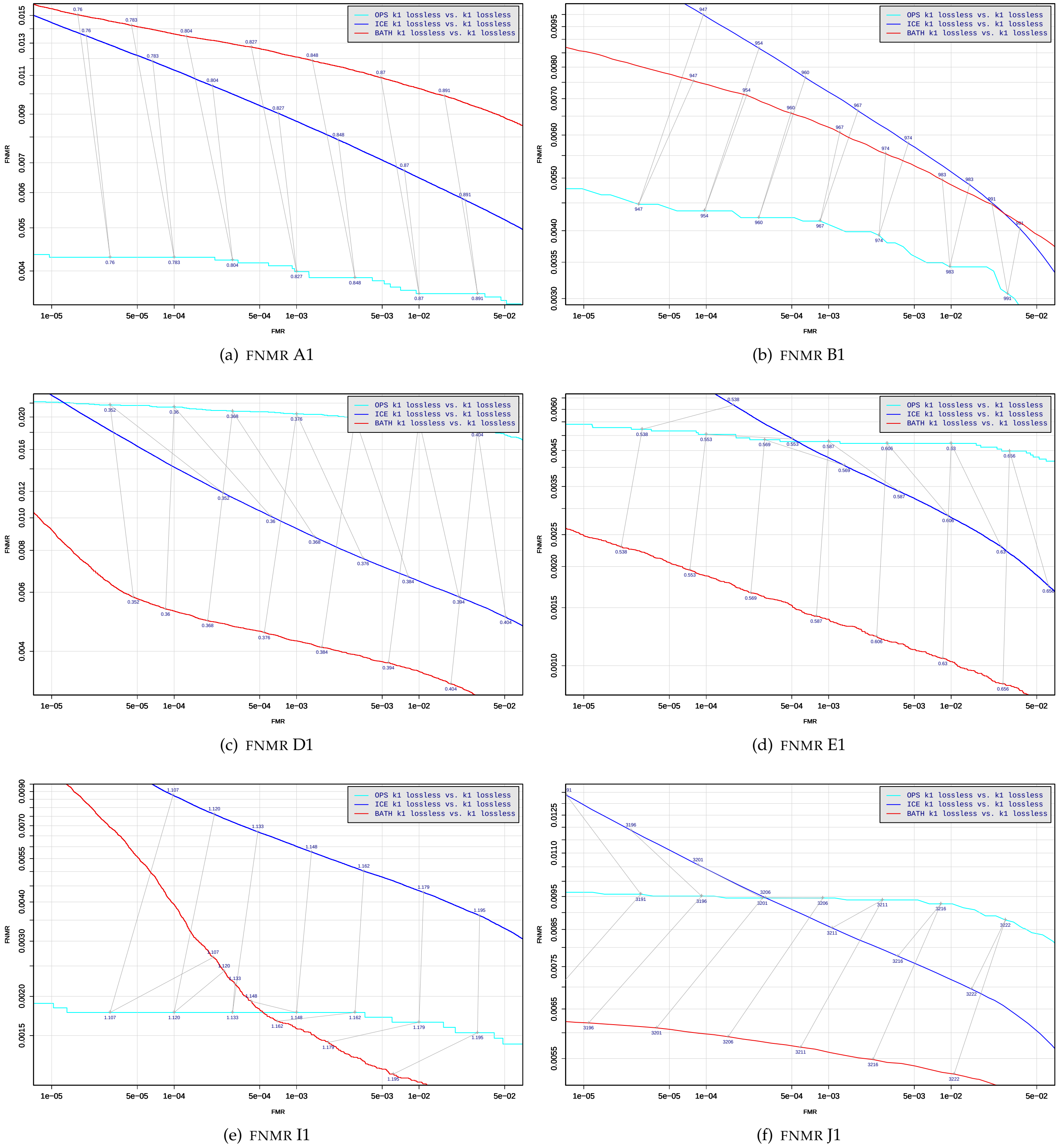
The DET characteristics of Figures 10, 11 and 12 allow comparison of the core algorithmic accuracy of the IREX algorithms. They are computed from comparison of uncompressed 640x480 images from the OPS, BATH and ICE datasets as described in section 5. While discussion in section 7.3.1 showed that different providers' algorithms work best on different datasets, the comparative DET plots in this section show that the relative ordering of algorithms is mostly consistent. For the main observations that follow, note that the DETs include the effects of FTE and FTE - see the discussion in section 7.3.3.

- ▷ On the OPS images, all SDKs produce reasonably flat DET characteristics. For most SDKs there is little variation in FNMR across the five decades of FMR plotted. On this dataset, the I1 and I2 SDKs are almost identical and give fewer than half the false non-matches of the next closest SDKs (B1 and A1).
- ▷ For all datasets, the false non-match rates of the SDKs vary over more than an order of magnitude. The DET characteristics tend not to cross, and they are almost always flat enough that a change in FMR operating point cannot compensate for choice of recognition algorithm. That is, a relaxation of the false acceptance criterion does not lower false rejection sufficiently to beat other algorithms.

[29] The IREX API defined comparison scores to have a floating point type. Although two of the participants, B and J, produced integer values, these are subject to identical analyses.

[30] DET characteristics sometimes plot Normal deviates, i.e. a plot in which the FNMR and FMR are (non-linearly) transformed by the inverse CDF of $N(0, 1)$. This is abandoned here because the score densities are often not Normal.

[31] All DET plots are included in the IREX SUPPLEMENT at http://iris.nist.gov/irex/irex_appendices.pdf.



(a) FNMR A1

(b) FNMR B1

(c) FNMR D1

(d) FNMR E1

(e) FNMR I1

(f) FNMR J1

Figure 9: DET characteristics for six implementations on three IREX datasets. All comparisons are with uncompressed KIND 1 vs. KIND 1 images. The lines join points corresponding to the a fixed threshold. All results apply to native operation. The vertical scales are different. Non-vertical links indicate a change in FMR when the dataset changes. The effect of failure to produce templates (i.e. FTE and FTA) is ignored, because the intent is to show relative ordering of FNMR and slope. This means the plots are not suitable for comparative testing of algorithms (see section 7.3.3).

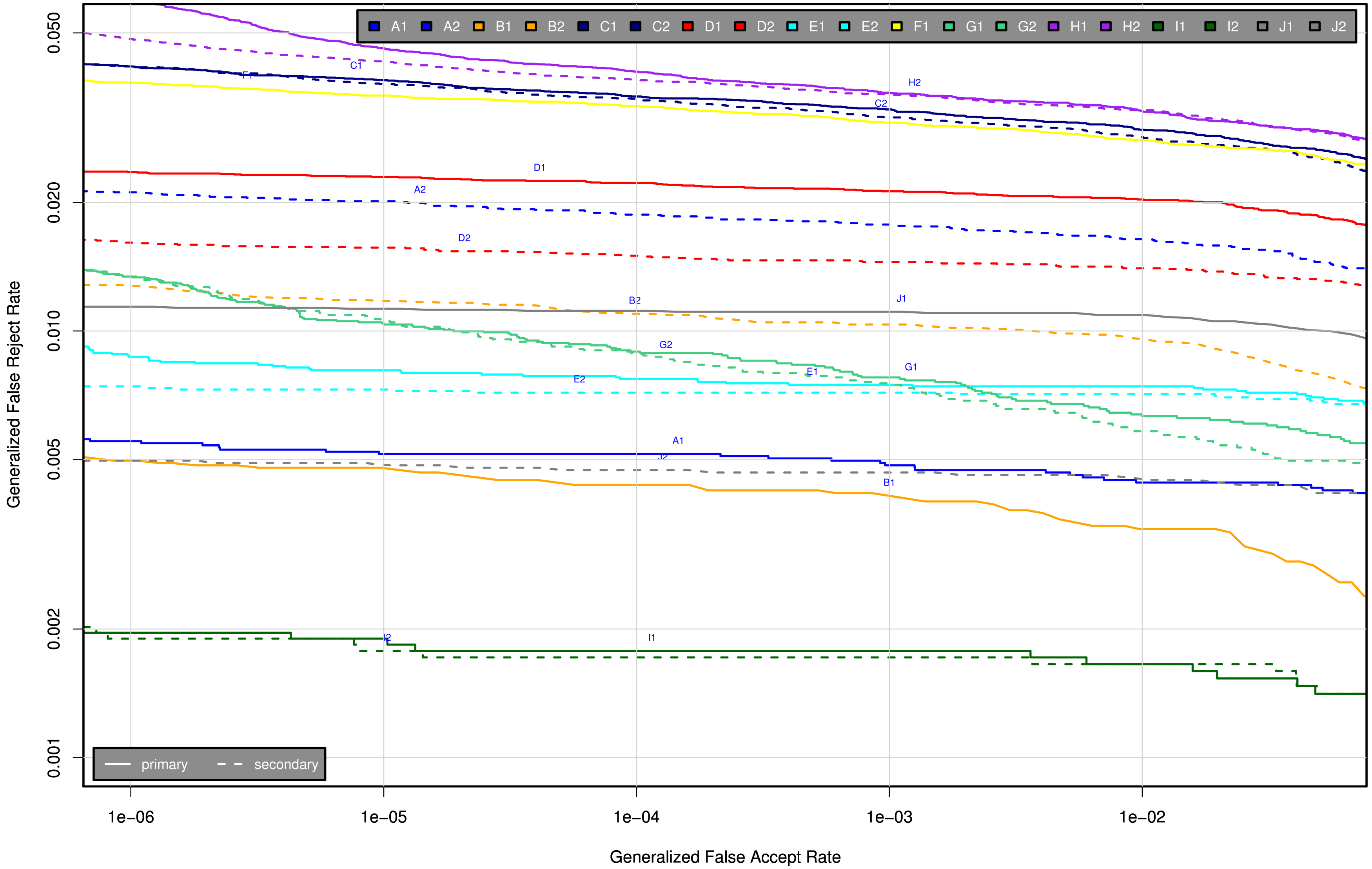


Figure 10: The DET characteristics for all IREX algorithms on the OPS dataset for uncompressed i.e. KIND 1 images. As documented in Table 6 of section 7.1, each SDK failed to produce templates from a fraction of the images and the effects of that are included in this plot. For the OPS dataset this fraction was often zero.

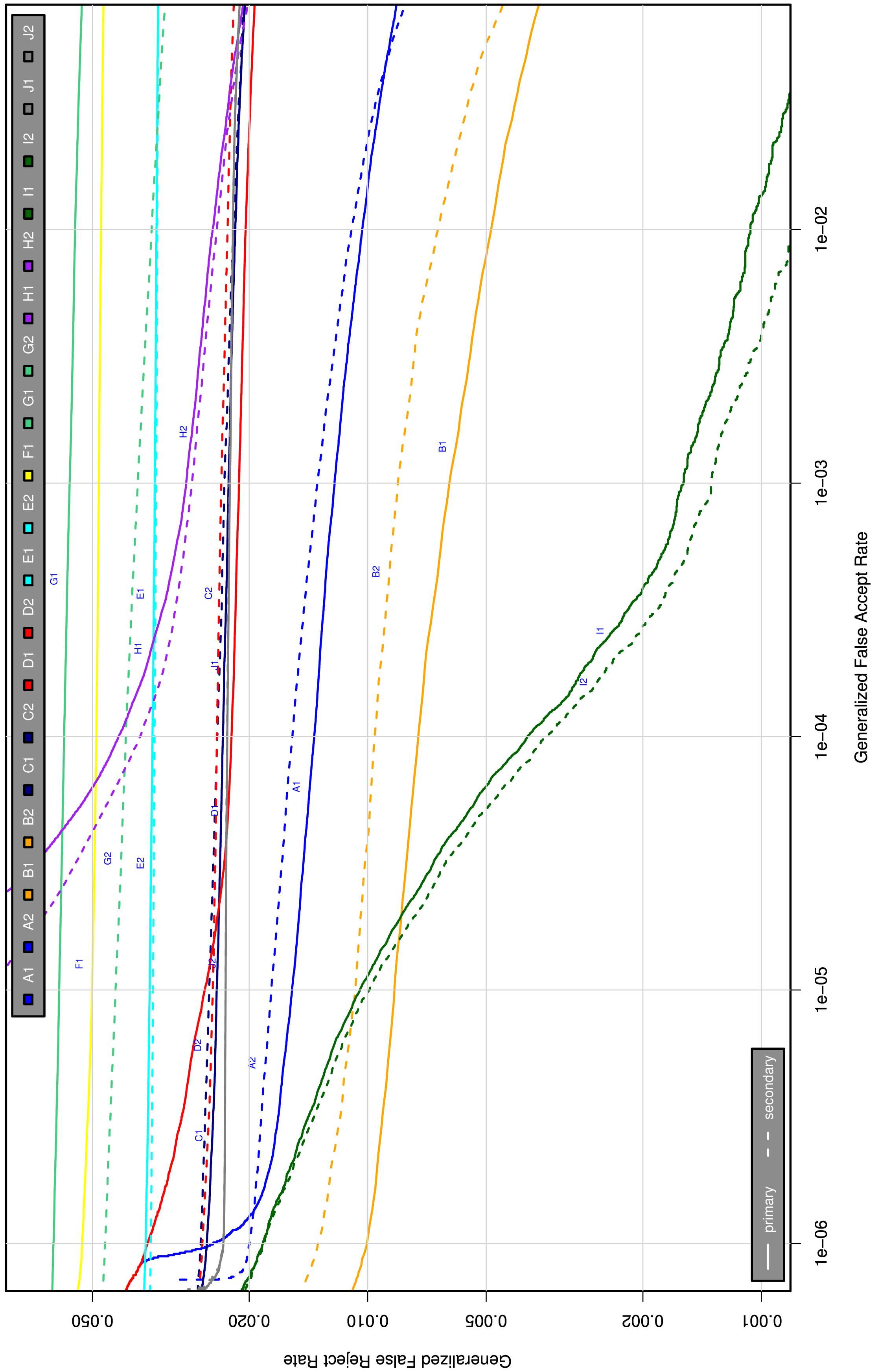


Figure 11: The DET characteristics for all IREX algorithms on the BATH dataset for uncompressed i.e. KIND 1 images. As documented in Table 6 of section 7.1, each SDK failed to produce templates from a fraction of the images and the effects of that are included in this plot.

A = SAGEM	B = COGENT	C = CROSSMATCH	D = CAMBRIDGE	E = L1	$x1$ = PRIMARY
F = RETICA	G = LG	H = HONEYWELL	I = IRITECH	J = NEUROTECHNOLOGY	$x2$ = SECONDARY

KIND 1 = RAW 640x480 KIND 3 = CROP KIND 7 = CROP+MASK KIND 16 = CONCENTRIC POLAR

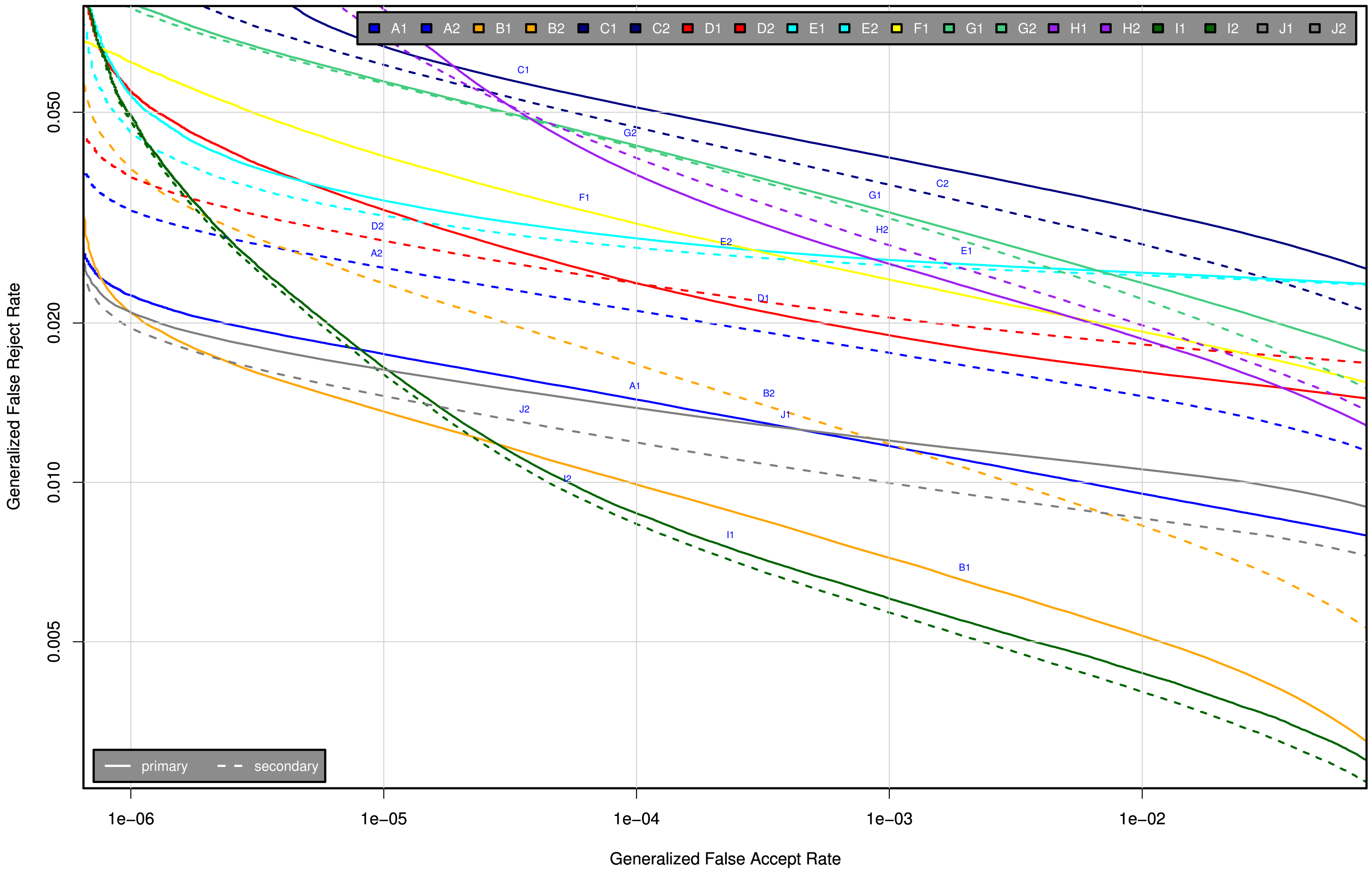


Figure 12: The DET characteristics for all IREX algorithms on the ICE dataset for uncompressed i.e. KIND 1 images. As documented in Table 6 of section 7.1, each SDK failed to produce templates from a fraction of the images and the effects of that are included in this plot.

- A notable exception applies to the ICE and BATH datasets: Figures 11 and 12 show that the I1 and I2 SDKs give very low error rates at high false match rates of 0.001, but this begins to degrade below about FMR = 10^{-4} such that the B1 and B2 SDKs give the best accuracy below FMR = 10^{-5}. On the BATH dataset, primary and secondary SDKs from each provider are usually very close.

- On the ICE dataset, the FNMR estimates appear more tightly clustered but this is an artifact of the log-scale. The FNMR values are uniformly higher, with the best FNMR values being an order of magnitude higher than those in the OPS dataset. This reflects the "clean" nature of the OPS dataset[32] and the operationally non-representative nature of the ICE images. Regarding the portability of these results to operational deployments readers should see the discussion on page 13 and in section 5.4.

7.3.3. The effect of trading FTE for FNMR

Section 6.1 introduced the failure to enroll and failure to acquire measurements and showed that they are influential on accuracy. Figure 13 shows two sets of DET characteristics for recognition of uncompressed KIND 1 ICE images. The first set of plots penalizes an SDK for failing to process enrollment and verification images. The second set does not. The relative accuracy depends on the template generation failure measurements reported in Table 6. For the A1, A2, B1, B2, I1 and I2 SDKs, the DET characteristics are identical because they always produce a template i.e. FTE = FTE = 0. For other SDKs these values are non-zero. For example, the E1 and E2 SDKs fail to make templates from 1.2% of the ICE images. This failure is often elective, i.e. the image processing algorithms of the SDKs determine that the input image is, in some internally defined sense, irregular or unsuitable (e.g., the image was blurred, or the eye was closed). Operationally this quality control function is common and valuable because failure to produce a template may trigger re-acquisition of a new (and hopefully better) image. However, in some applications where the subject cannot be imaged again, a failure is unrecoverable.

The efficacy of rejecting poor images can be seen by studying the results for SDKs E1 and E2 in Figure 13. With FTE and FTE removed from the error rates in the lower graph, they become the best performing algorithms. This is evidence that these SDKs are making good decisions when they declares FTE and FTE events i.e. that the failures are effective at reducing false non-match errors. This would not be the case if the failures were random.

This is clearly unfair for *comparative* testing because the zero-failure SDKs would potentially benefit from excluding those images and the providers of those SDKs deserve some credit for enrolling the image and trying to match them. However, note that comparative testing should also include the tradeoffs between accuracy and speed discussed in section 7.6, and in this respect the E SDKs are somewhat faster at template generation than the A and B implementations.

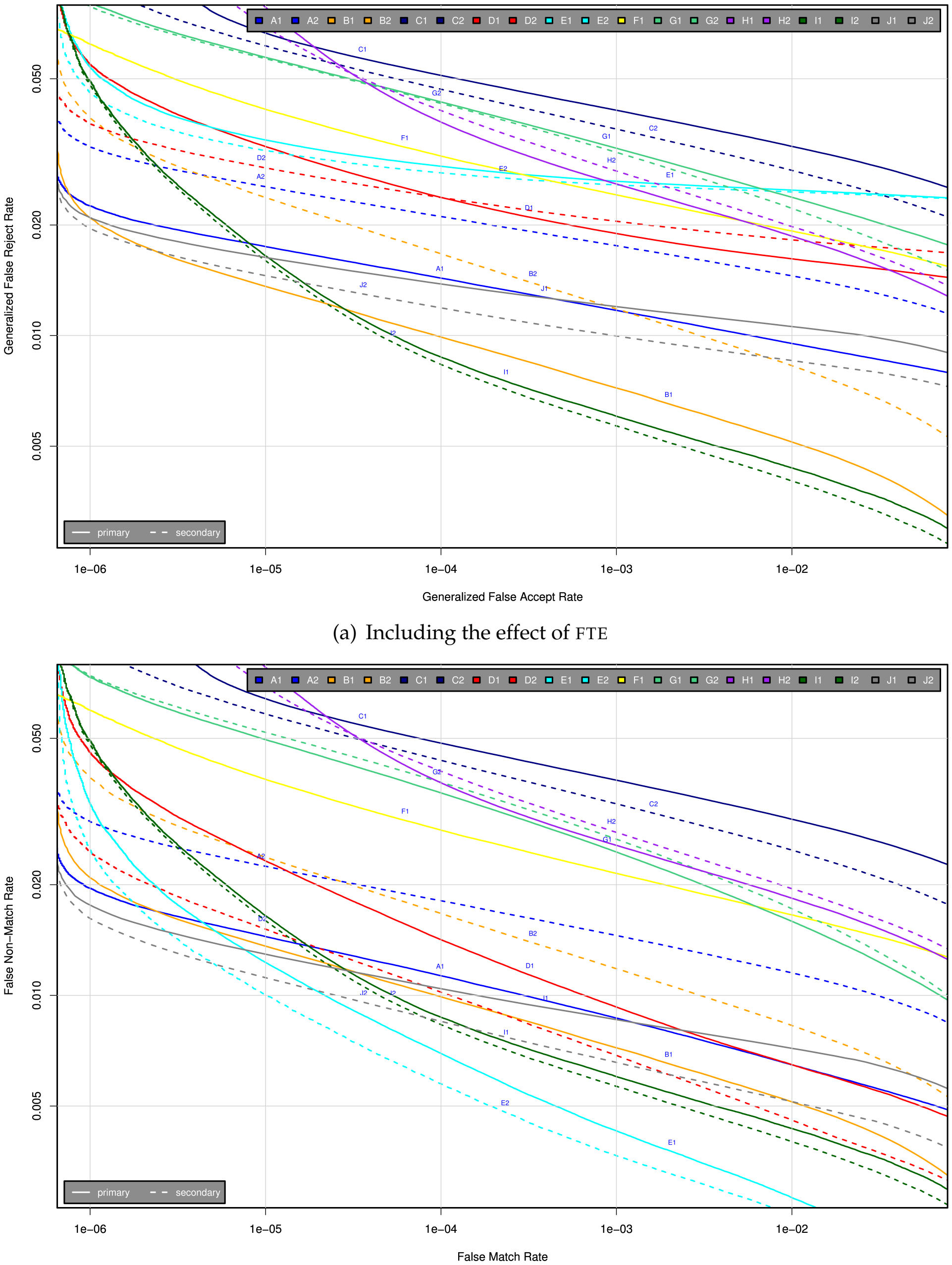
The overall interpretation of the results depends on the intended application: If FTE and FTE events are operationally present and tolerably infrequent then the algorithms may be viable; if on the other hand FTE and FTE events are unrecoverable then alternative algorithms seem necessary.

The same discussion above applies to the BATH dataset, although the causes of the FTE events are different. Note that FTE rates of the OPS dataset are very low.

7.4. False match rate calibration

The practice of using fixed FMR and threshold values elsewhere in this report is undertaken for the purpose of analysis only, and is not intended as any recommendation on operational FMR or threshold policies. However, this section supports threshold selection by giving the incidence of false matches as a function of threshold. This follows the ITIRT[32] study which exhaustively tabulated the false match rates as a function of threshold for the Iridian product used.

[32] The OPS dataset is easier to recognize because an iris recognition system was used in selection of the images - see section 5.1.



(a) Including the effect of FTE

(b) Excluding the effect of FTE

Figure 13: The DET characteristics for all IREX algorithms on the ICE dataset for uncompressed (i.e. KIND 1 images). **Above** are the plots that include the effects of FTE and FTE and **below** is the result of excluding missing templates from the FNMR and FMR computations. Each SDK failed to produce templates from a fraction of the images. For the ICE images used here, this fraction was often quite large (see Table 6).

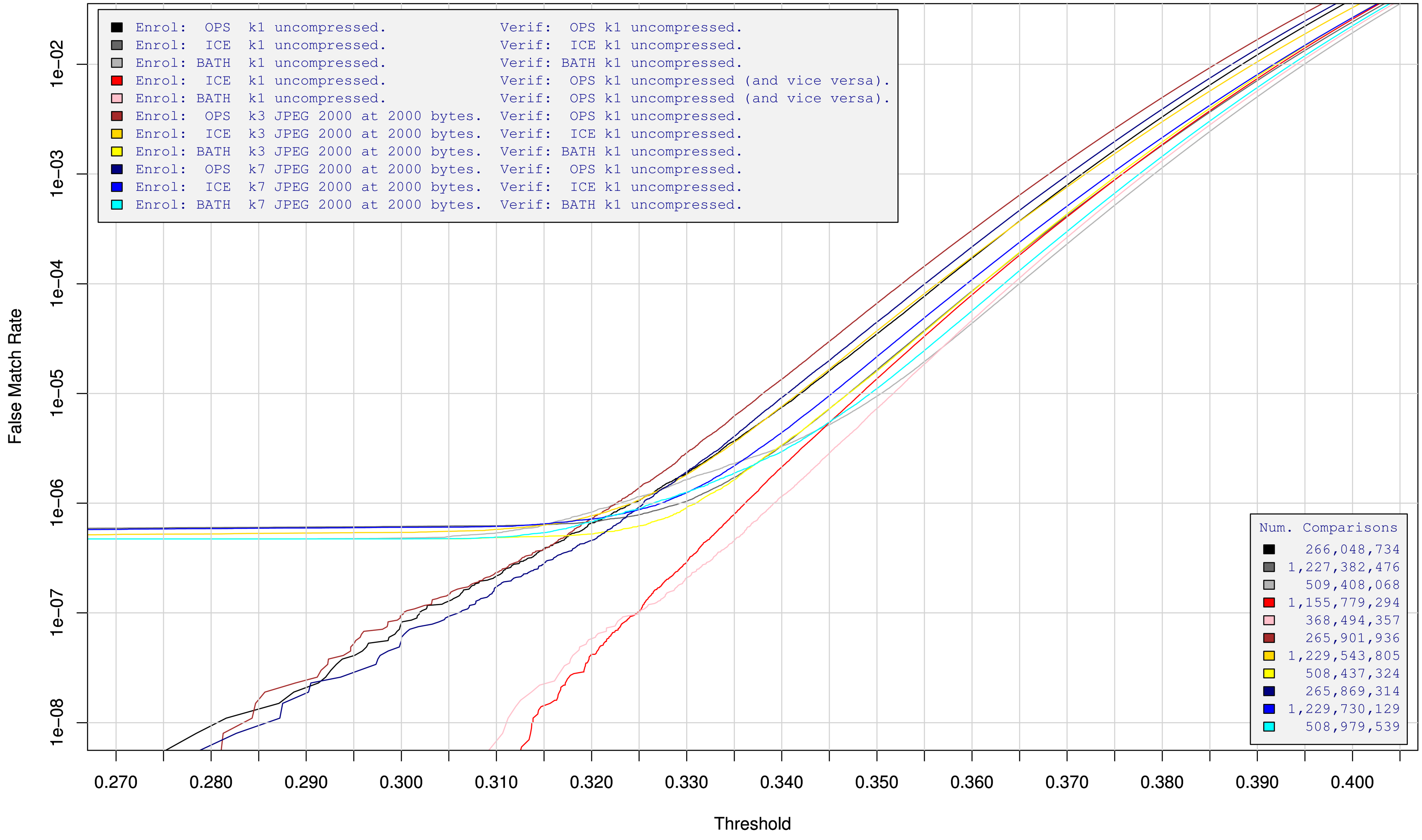


Figure 14: The plot of FMR vs. threshold for SDK D2 and various KINDS and sets of comparisons. Each curve is an empirical cumulative distribution function of a set of impostor scores with log-linear axes. Note that these curves are computed from images of O(10000) people for OPS, O(1000) people for BATH, and O(100) people for ICE. In addition, systematic variation beyond that observed here may occur. Threshold values only have meaning for specific algorithms; they are not interoperable across SDKs and providers.

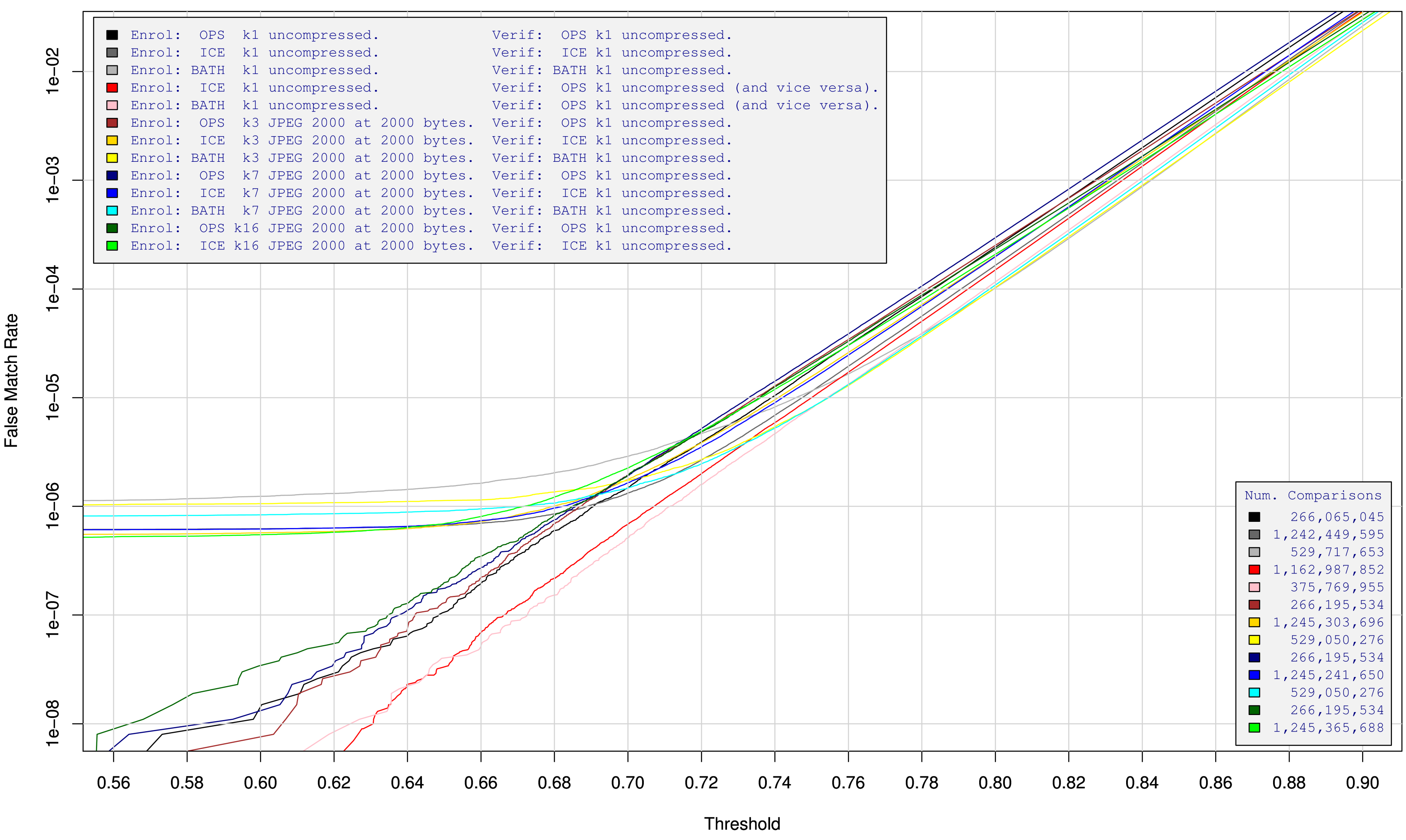


Figure 15: The plot of FMR vs. threshold for SDK A1 and various KINDS and sets of comparisons. Each curve is an empirical cumulative distribution function of a set of impostor scores with log-linear axes. Note that these curves are computed from images of O(10000) people for OPS, O(1000) people for BATH, and O(100) people for ICE. In addition, systematic variation beyond that observed here may occur. Threshold values only have meaning for specific algorithms; they are not interoperable across SDKs and providers.

The dependency of FMR on threshold has been the subject of much academic publication at least for one class of iris algorithms [15]. Indeed the claim of a known impostor distribution is an extremely attractive and valuable property for all biometrics and particularly for *one-to-many* applications where false match suppression is of paramount importance for large populations. Figures 14 and 15 gives this dependence for SDKs D2 and A1[33] as applied to various combinations of intra-dataset and inter-dataset, and compression image, comparisons. The corresponding graphs for other SDKs appear in the respective appendices.

The curves include use of inter-dataset comparisons (ICE vs. OPS) which afford very high assurance of ground truth integrity. This arises because the subject populations are very well separated geographically and occupationally, and the likelihood of co-membership is considered to be identically zero.

Note that while the number of impostor comparisons at $1.2\text{x}10^9$ is still short of the $1.7 \text{ x } 10^9$ performed in ITIRT[32], the number of persons present is somewhat larger ($8160 + 240 = 8400$ vs. 1224 in ITIRT). Both of these numbers are far short of the 632500 irises used in executing the approximately $2 \text{ x } 10^{11}$ impostor comparisons reported for the UAE data[16].

For these graphs the notable observations are as follows.

▷ For these sets, $\log$ FMR (τ) is usually a slightly sub-linear function of threshold (i.e. $\tau^a, a < 1$) reflecting only an approximately exponential form in the left tail of the impostor distribution. Exceptions to this are the H1 and H2 algorithms which exhibit a more nonlinear dependence.

▷ With intra-dataset comparisons, ground truth integrity is generally subject to unresolved consolidation errors. These appear in both the BATH - to - BATH and ICE - to - ICE FMR distribution functions as horizontal lines at low FMR because the genuine pairs that contaminate the impostor distributions are most often rejected only at much lower thresholds.

▷ Across datasets, formats, and compression levels there is often an order of magnitude variation in the observed false match rates. Specifically if the threshold that gives a FMR around 0.0001 is considered, there is a vertical spread of the various plots. For SDK A1, this variation is low, at about a factor of three. For others it can approach a factor of 20. This goes to the stability of the impostor distribution which is an important metric in its own right (see the discussion later, in section 9).

▷ Table 7 shows the correspondence of the impostor empirical cumulative distribution function computed here for KIND 1 ICE vs. KIND 1 OPS comparisons with values published by Daugman. Comparisons are provided for a) empirical estimates for the score-normalized algorithm[16], and b) theoretical predictions of false match occurrence for the original algorithm[15]. For D2, this proximity is noteworthy in that neither dataset existed at the time the

(a) Proc. IEEE[16]

Threshold HD	Published	D1
0.282	3.5 10^{-9}	3.8 10^{-8}
0.297	5.6 10^{-8}	1.9 10^{-7}
0.312	5 10^{-7}	9.0 10^{-7}
0.317	1 10^{-6}	1.6 10^{-6}

(b) IEEE Trans. CSVT[15]

Threshold HD	Published	D2
0.32	3.8 10^{-8}	4.2 10^{-8}
0.33	2.5 10^{-7}	2.9 10^{-7}
0.34	1.4 10^{-6}	2.1 10^{-6}
0.35	7.5 10^{-6}	1.4 10^{-5}

Table 7: Selected IREX estimates of FMR alongside previously published estimates. Cambridge indicates in their comments in the D1 and D2 appendices of the IREX SUPPLEMENT that score normalization is used for D2 but not for D1. The IREX estimates are for the 1,165,868,160. OPS - ICE impostor comparisons.

[33] See the IREX SUPPLEMENTAL appendices for the graphs for all other SDKs. The PDF of the appendices may be downloaded from http://iris.nist.gov/irex/irex_appendices.pdf.

Daugman paper was written. The full FMR (τ) plots, for SDKs D2 and A1, are presented in Figures 14 and 15 (the remainder are in the supplemental appendices).

- ▷ The D2 implementation's OPS - to - OPS FMR curves in Figure 14 are all significantly above the OPS - to - ICE plots.[34] This observation, that inter-dataset false matches are inherently less likely than intra-dataset, would occur if some aspect of the imaging environment imparted a degree of similarity to iris images within a dataset to which the feature extraction algorithms is not invariant. The imaging environment could include aspects intrinsic to the sensor (illuminants, for example) or extrinsic (manner of presentation, or ambient illumination, for example). The authors are not aware of published studies of this hypothesis. In any case, cross-database, cross-camera searches are a defining aspect of federated applications in which images are collected using cameras in a multi-vendor marketplace.

- ▷ For other implementations whose native score scale might *suggest* Daugman-like algorithmic function, the impostor distributions are also higher than expected. IREX and other NIST tests know nothing of the internals of algorithms submitted (i.e. they're black boxes), and thus the CDF plot cannot be held to be unusual or unexpected. We report these curves to assist practitioners in setting thresholds.

7.5. Template sizes

The template sizes are tabulated in Figure 16. While some observations are made below, see section 7.6 for additional discussion.

- ▷ There are two orders of magnitude between the largest and smallest template sizes: For SDK E2 the size of the verification template is 45080 bytes; for G1 the enrollment template is 512 bytes.

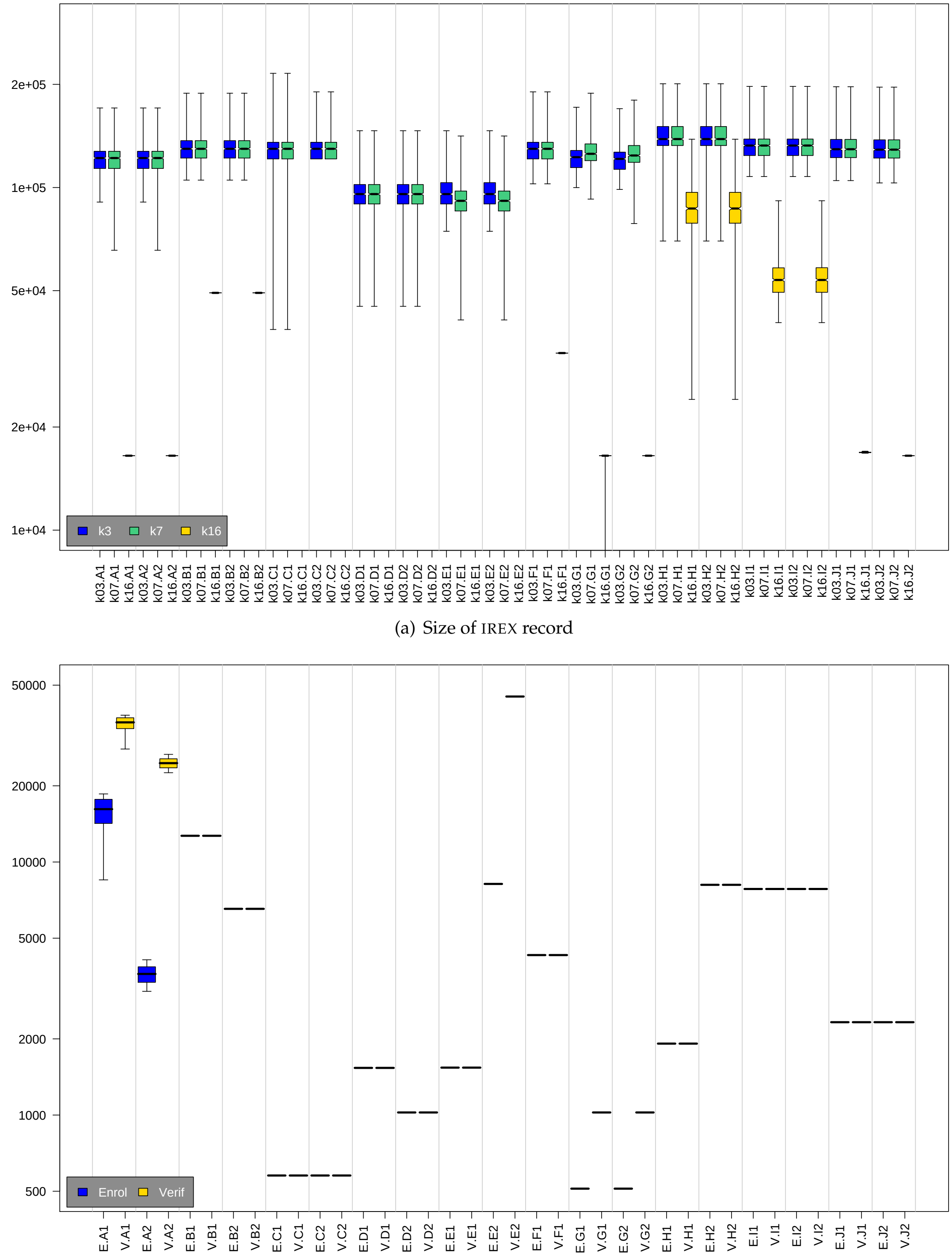
- ▷ Five SDKs from three providers (A, E, G) implemented asymmetric templates. The IREX API allowed *role-specific* templates where the template generation function is informed of the kind of template to be produced. In all those cases the verification template is larger than the enrollment template. This property is desirable for persistent enrollment datasets on cost-of-storage grounds.

- ▷ One provider submitted SDKs (A1 and A2) with variable length templates. While many iris recognition algorithms represent the iris in a fixed-dimensional vector space[47, 15] the templates allowed in IREX are fully proprietary and may include *any* kind of information, including the original image. Despite the extensive discussion of *iriscodes* in the academic literature there is no standard representation and no prospect of making one. In any case, NIST has no knowledge of the template contents.

- ▷ If templates are concatenated and compressed using the bzip2 lossless compressor some size benefits are realized for all SDKs. The size reductions range from around 12% through to 70% for the OPS dataset, and from 19% to 70% for the ICE templates. Many individual templates cannot be compressed in this manner (i.e. the size increases under compression).

- ▷ Some IREX templates are *larger* than the images as compressed under IREX . If an operation elected to compress images to 3KB, then the templates from A1, A2, B1, B2, E2, F1, H2, I1, and I2 would be larger than the image. If the image size was 2KB, J1 and J2 would be added to this list. This situation is atypical in biometrics. It occurs because the Daubechies wavelet representations used in compressed JPEG2000 streams are efficient texture-preserving, and

[34] In Figures 15 and 14 the ICE - to - ICE and BATH - to - BATH comparisons appear to be contaminated with residual ground truth errors, and are therefore not useful for examining the impostor distribution tails.



(a) Size of IREX record

(b) Size of template

Figure 16: **Top:** The distribution of the IREX record size (header + image), arranged by SDK and image KIND. The boxplot whiskers extend to the highest and lowest observed values except that failed, zero length, records are not included. The units are in bytes and a log-scale is used. **Bottom:** The template size arranged by SDK and the ROLE of the template (enrollment, then verification). The boxplot whiskers extend to the value in which 99.9% of Normal deviates would reside. In both cases, the vertical grey lines delimit the SDKs for ease of viewing.

	Linux SDKs							
	C1	C2	D1	D2	H1	H2	I1	I2
k1	166	238	21	17	2274	9086	1569	1557
k3	163	235	21	17	2276	9086	1571	1575
k7	166	239	21	17	2274	9086	1571	1550
k16	-	-	-	-	2273	9083	1569	1564

	Windows SDKs										
	A1	A2	B1	B2	E1	E2	F1	G1	G2	J1	J2
k1	1032	211	553	102	259	5115	233	51	53	14	20
k3	1051	211	540	101	259	5112	234	51	53	15	21
k7	1069	210	541	101	259	5276	234	51	53	15	21
k16	1111	211	540	101	-	-	-	51	53	15	21

Table 8: For LINUX SDKs (**left**) and Win32 SDKs (**right**) the time needed to execute a single template comparison by SDK and KIND. The units are microseconds (μs). The values are computed over a single timing of $C = N(N-1)/2$ comparisons. With $N = 600 + 300 = 900$, $C = 404550$. The LINUX SDKs run on Intel Xeon E5405 blades running at 1995MHz, with a 1333MHz bus and 2 x 6MB of cache, benchmarked at SpecInt of 17.6. The win32 SDKs run on Intel Xeon E5310 blades running at 1595MHz, with a 1066MHz bus and 2 x 4MB of cache, benchmarked at SpecInt of 12.6.

reversible, representations of the original iris raster. Daugman noted[19, 18] the utility of JPEG2000 for iris recognition may rest on commonalities between its wavelet representation and that used in the recognition algorithm[35].

7.6. Timing results and accuracy-time tradeoffs

This section gives results for the times of the elemental operations provided in the IREX API specification. These are

▷ Conversion of a raw 640x480 greyscale raster into the KIND 3, KIND 7 and KIND 16 records. This includes iris detection and initial localization, and in the case of unsegmented polar records, the rectilinear to polar interpolation.

▷ Conversion of an IREX record into a proprietary template. This requires final fine segmentation and feature extraction. For the unsegmented polar record this does not include polar to rectilinear conversion because that was performed by the NIST test harness which produces the KIND 48 instance passed to the template generator.

▷ One-to-one comparison of two proprietary templates.

7.6.1. IREX record and template generation times

The boxplots of Figure 17 show the times needed to generate IREX records and to produce templates from them. The plots include data for the various KINDS and all SDKs. The notable results on computational efficiency are as follows.

▷ The time taken to convert a parent 640x480 iris image into a IREX record varies by two orders of magnitude over the SDKs tested. For G1 and G2 the times are below 20 milliseconds (ms), except for KIND 7 which requires around 60 ms. For J2 the times are around 1200 ms on the same hardware.

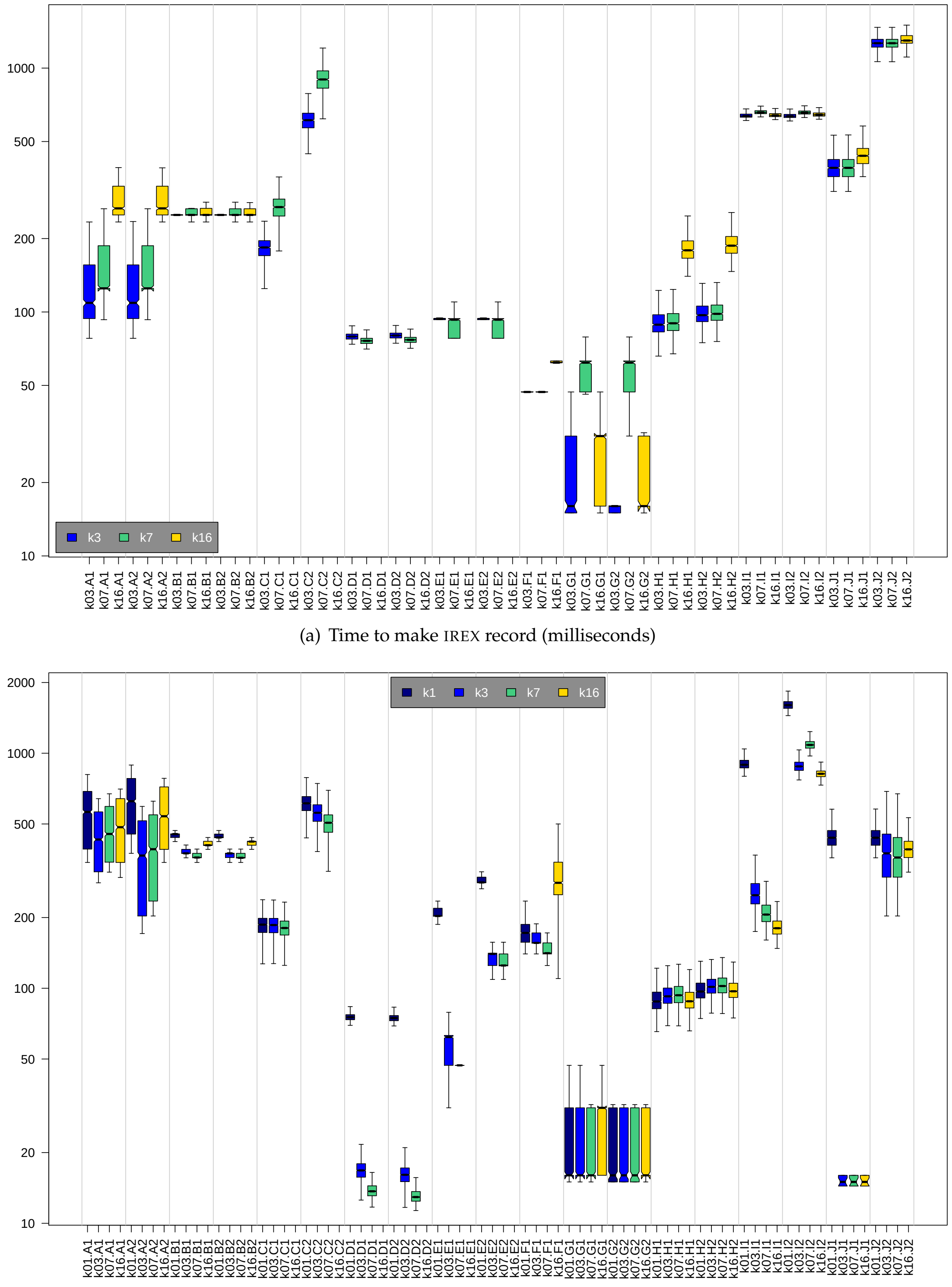
▷ For some SDKs the time needed to generate the various KINDS is essentially the same. This applies to B1, B2, D1, D2, E1, E2, F1, I1, I2 and J2. For the others, the KIND 3 record is usually the least expensive to produce, followed by KIND 7 and KIND 16. The increase in expense never exceeds a factor of about three (A1).

▷ The time needed to produce a template from an IREX record also varies by two orders of magnitude. For the KIND 7 record the time needed for D2 is 13 ms, while for I2, the figure is 1088 ms.

▷ The variation in template generation time across SDK is much larger than across the KIND of image. The template generation time usually depends on the KIND of the input record, the exceptions being for H1, H2, G1 and G2[36]. In almost all other cases the KIND 1 record is the most expensive, reflecting the cost of the iris detection in a large raster. In the case of J1, the time for KIND 1 is a factor of forty larger than for the other KINDS.

[35] Because the IREX test never passed compressed encoded data to the template generators it precluded the *direct* use of the JPEG2000 or JPEG stream. This possibility has been mooted because the wavelet representation of JPEG2000 is not dissimilar to the Gabor wavelets used in many published iris feature extractors. Given lack of research and unclear performance benefits NIST considers that this was not a significant limitation of the IREX design.

[36] The G1 and G2 implementations run under windows where the timing resolution is close to that of the measured time. The gives poor uncertainty estimates.



(a) Time to make IREX record (milliseconds)

(b) Time to make template (milliseconds)

Figure 17: At top, the distributions of the time needed to generate a standard IREX iris image record arranged by SDK and KIND. Below is the corresponding time to convert an IREX record into a proprietary template. In both cases the units are milliseconds (ms), and log-scales are used. The times are measured over 600 OPS and 300 ICE images. The vertical grey lines delimit the SDKs for ease of viewing. The timing resolution depends on the operating system. On windows platforms, for SDKs $\in \{A, B, E, F, G, J\}$ the resolution is 15ms. On the linux platform, for SDKs $\in \{C, D, H, I\}$ the resolution is 1μs. The faster SDKs operate near this resolution, and only the median values of the boxplots are meaningful. The boxplot whiskers extend to the value in which 99.9% of Normal deviates would reside.

- For the KIND 3, KIND 7 and KIND 16 instances, where the iris location is known, the template generation times are usually similar. While the notched boxplots do exhibit significant differences, the inter-quartile ranges often overlap. This indicates that cost of preparation is not innately derogatory to particular standard formats.

7.6.2. MATCHING TIMES

Table 8 shows the time needed to execute one-to-one template comparisons. All times are given in microseconds (μs) and are averages over 405550 comparisons. The notable observations are as follows.

- Across SDKs the time needed to compare two templates to produce a comparison score varies by three orders of magnitude. The times are as short as 14 microseconds for J1 to 9 milliseconds for H2[37]. While, for many implementations the comparison operation is fast (because the template represents the iris in a vector space[47] for which arithmetic operations can be as inexpensive as bitwise XOR [15]), the very large range observed here reflects a diversity in the algorithmic approaches.

- There is considerable range in speeds across the SDKs from a single IREX participant. SDK E2 is 20 times slower than E1. Similarly both A1 and A2, and B1 and B2, differ by a factor of five. This, too, is indicative of the use of different algorithms, representations or parametrization.

7.6.3. SPEED VS. ACCURACY TRADEOFFS

Given the wide ranges in processing times observed in the last subsections, the question arises whether accuracy can be traded for speed. This is addressed directly in the plots of Figure 18, each of which is a scatter plot of duration against accuracy, with each point corresponding to one SDK. In each case the accuracy is quantified by FNMR at a fixed global threshold giving FMR = 0.00001 computed over uncompressed and uncropped ICE and OPS images. The FNMR value is the simple average of those observed over the OPS and ICE corpora. The three durations are:

- The time taken to prepare KIND 3 and KIND 7 IREX records;

- The time taken to prepare templates from unprocessed 640x480 KIND 1 images;

- The time taken to compare two proprietary templates.

A fourth duration was added: This is a composite value representative of a physical access control application in which any of the N enrolled users cooperatively presents to the system without use of an identity credential. The duration of a first attempt can be modeled as the summation of the capture time, T_C, the template generation time, T_I, and N times the matching time, T_M. The whole transaction takes time

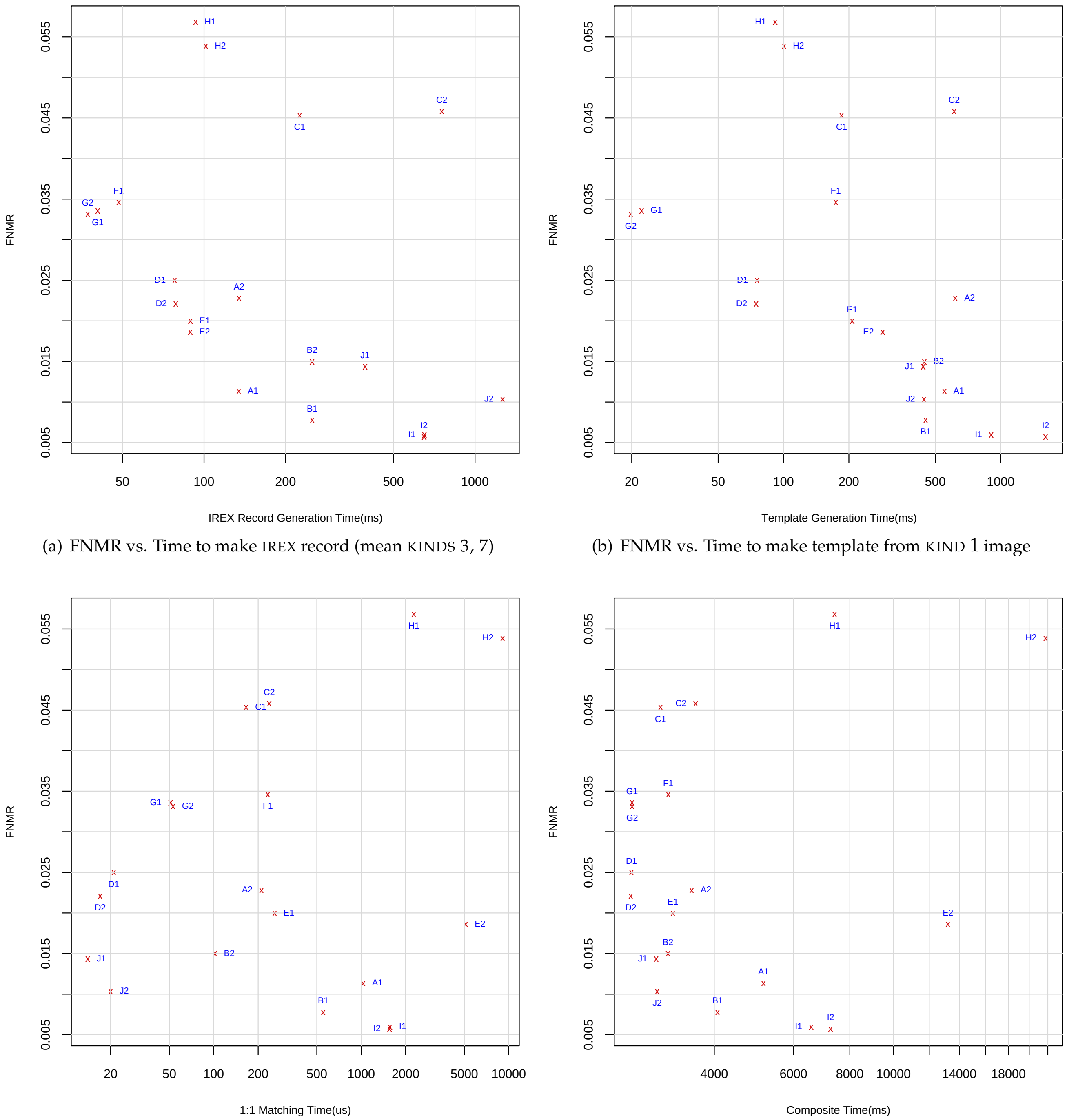
$$T = T_C + (T_I + NT_M)(1 + \text{FNMR}(\tau)) \tag{14}$$

where the possibility of false rejection at a first attempt is included by the FNMR term, and where it is assumed that the initial capture phase includes acquisition of at least one further image (e.g., via a two-eye camera). This result of applying this model with the measured FNMR values is shown in the last plot of Figure 18. The notable observations are:

- The inverse relation between accuracy and speed is even more evident than for the single-aspect time vs. accuracy plots. The four most accurate algorithms would take $T \geq 4$ seconds.

- The $N = 2000$ multiplier penalizes the expensive matchers, and such that the total time would be operationally untenable. Some providers have made comments on this aspect in their respective annexes of the IREX SUPPLEMENT.

[37] This range meant that a full cross-comparison of the IREX datasets would require from two hours to two months on a single CPU depending on the SDK. The latter figure is comparable with the time between outages in data-centers. The final IREX computations were parallelized over 64 CPUS.



(a) FNMR vs. Time to make IREX record (mean KINDS 3, 7)

(b) FNMR vs. Time to make template from KIND 1 image

(c) FNMR vs. Time to compare templates

(d) FNMR vs. Composite 1:2000 search time

Figure 18: Duration of the IREX functions versus FNMR at FMR = 0.00001 where the FNMR value is the simple average over the OPS and ICE dataset. At **top left** is the IREX record creation time measured in milliseconds; at **top right** is the template generation time also in milliseconds; at **bottom left** is the one-to-one comparison time in microseconds (μs); and finally at **bottom right** is the time for the hypothetical access control system of equation 14 with $N = 2000$ and $T_C = 2.5$ secs. All times are on log-scales.

7.6.4. Discussion

Some important caveats apply to the timing analysis:

- ▷ The IREX test plan established very liberal constraints on the expense of the algorithms[38] because recognition under compression was considered to be a harder and newer problem worthy of additional computation. Note that the IREX test plan did indicate that timing results would be reported.
- ▷ The IREX test plan only defined an API for one-to-one verification. One-to-many testing was considered unnecessary because the accuracy effects associated with formats and compression would be clearly evident in verification mode. One-to-many implementations may generally be expedited via several methods (e.g., incomplete distance computation, or fast search[27]).
- ▷ The IREX test plan did not prohibit threading and at least one SDK utilized this possibility. This can lead to considerable over-estimation of the core (i.e. single thread) speed of the algorithms. The IREX test plan should probably have prohibited threading since all *cpu*'s and cores were invoked simultaneously.
- ▷ One provider's SDKs required compilation with 64 bit compilation flags. Use of 64 bit executables does not necessarily expedite processing.
- ▷ One provider asserted that the template comparison operation would be faster if their template was used in consecutive comparisons. This condition held for the timing test but, for other reasons, was generally not true for the accuracy computations.
- ▷ *Fast* implementations have been reported[27]. These are intended give better-than-linear dependence on the size of the enrolled population in a one-to-many search.

7.6.5. CAN IREX ALGORITHMS BE EXPEDITED

While there is likely some opportunity for expediting the more expensive algorithms tested in IREX, prospective users of these algorithms should differentiate the following two classes of approach.

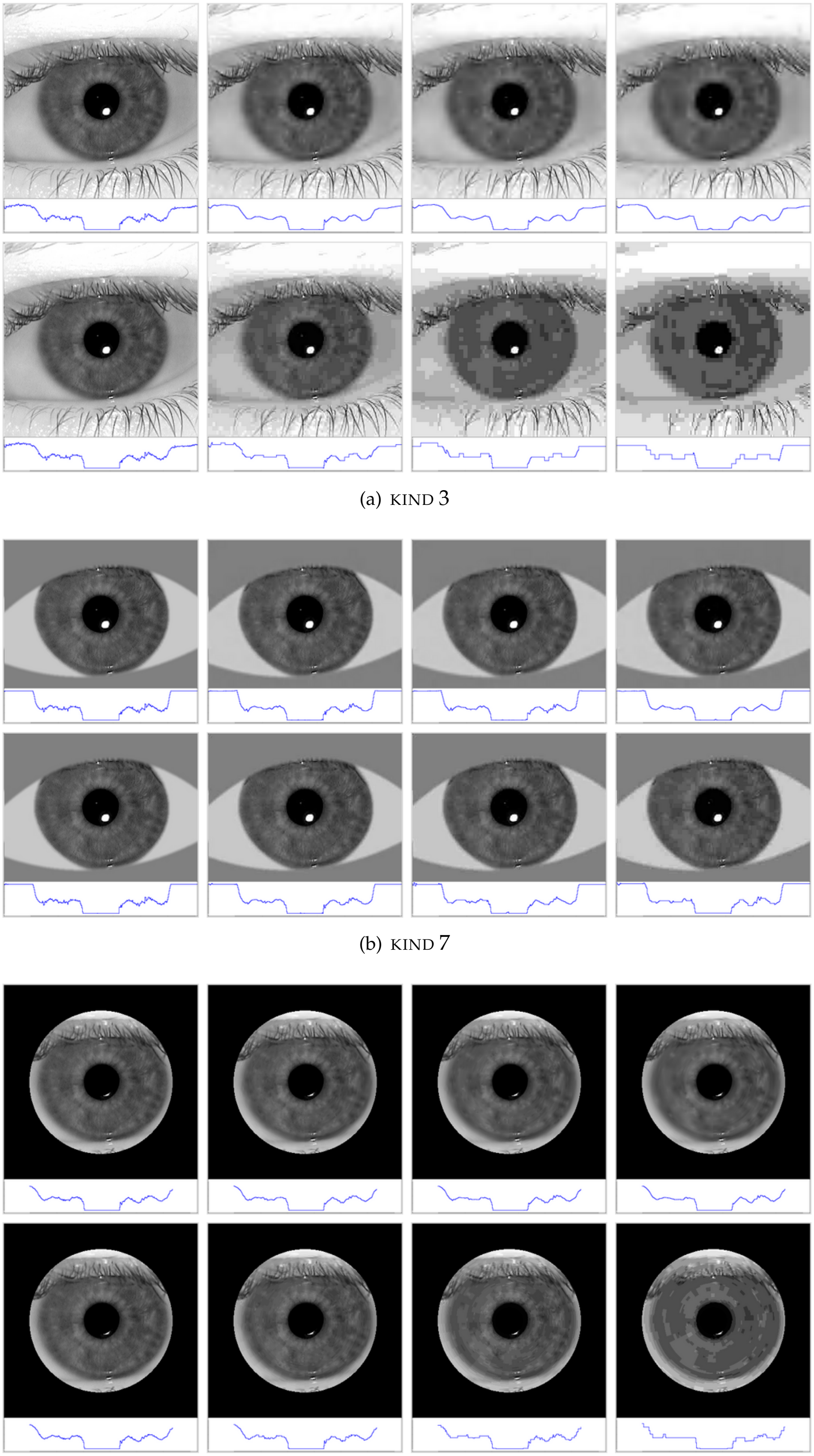
- ▷ *A - Computational optimization:* While this class of techniques is most simply achieved using more computationally powerful hardware, virtually all pieces of software can be optimized via compiler optimizations (loop unrolling, etc.), compiler changes (icc vs. gcc), recoding (assembly language vs. C vs. Java), and use of dedicated libraries (e.g., FFTW).
- ▷ *B - Algorithmic optimization:* This class includes alteration of the algorithms themselves with the possible consequence that different outputs are obtained for the same input. Examples here would be to reduce the extent of the search space, or to use Hough transforms vs. active contours.

It is important to note class A techniques do not change the actual algorithm and that class B techniques either change the algorithm completely or give heuristic approximations to it. In the latter case, the IREX accuracy results may no longer apply and performance could degrade).

8. Performance of compact formats

The prior section covered results for the traditional 640x480 iris image without compression. This section addresses the main focus of the IREX study - the performance of the compact image formats.

[38] Template comparison in 20ms, template and record creation in under 2500ms.



(a) KIND 3

(b) KIND 7

(c) KIND 48

Figure 19: An ICE image in each of the three IREX formats. From left to right, the sizes are uncompressed, 4KB, 3KB and 2KB. In each set, the top row uses JPEG2000 compression with JPEG below. The display/printing process may make it difficult to discern the increasing compression and sampling artifacts in the last three columns. Beneath each plot is a horizontal scanline across the center of each image. The KIND 3 images were produced by I1. The KIND 7 images were produced by D1. Note the tighter vertical cropping applied by D1.

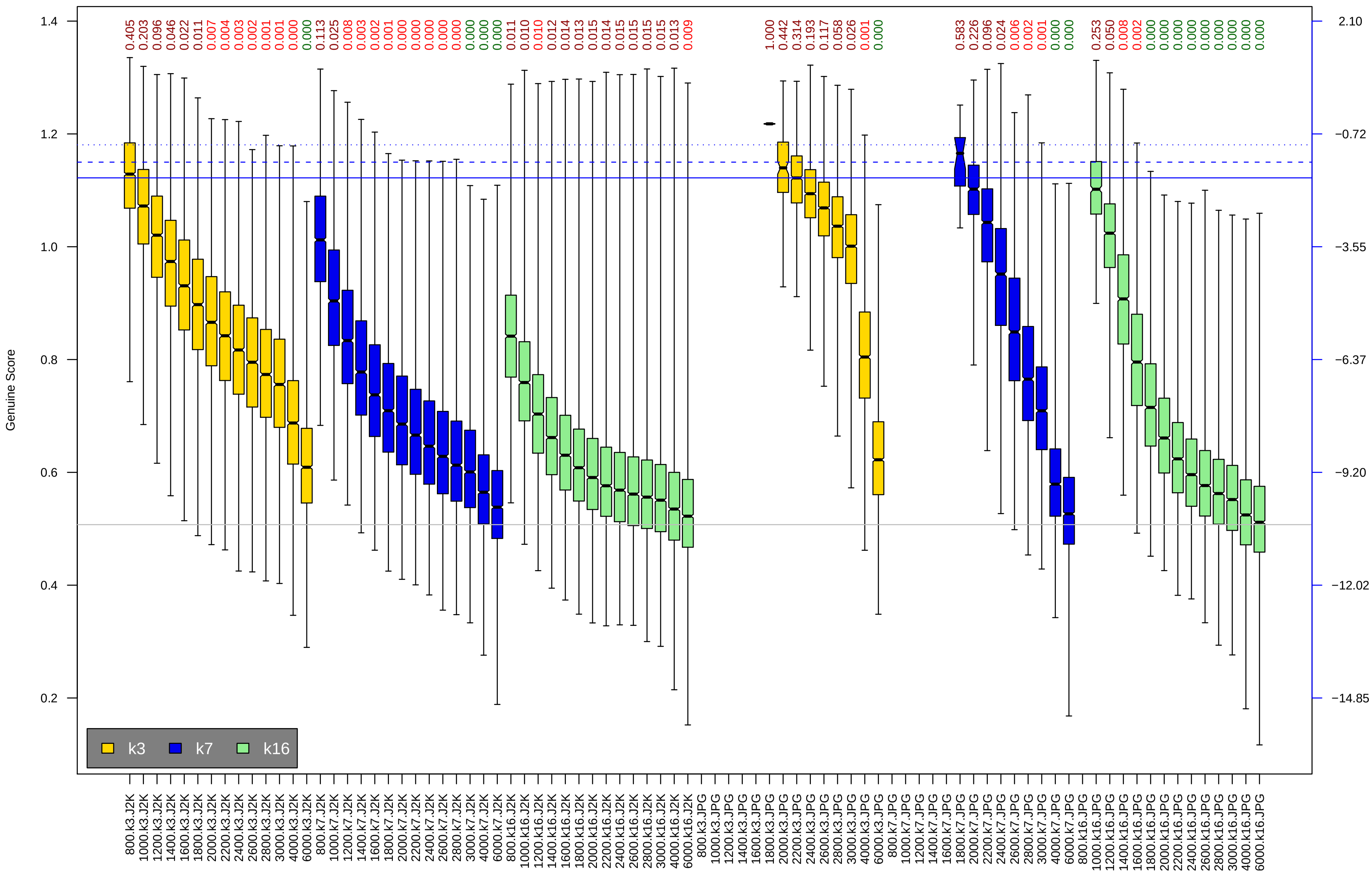


Figure 20: The distribution of I1 native genuine comparison scores by size of the compressed image, KIND and the compression algorithm. The images are from the OPS dataset. The right axis scale gives the corresponding value for $d' = (s - \mu_I)/\sqrt{0.5(\sigma_I^2 + \sigma_G^2)}$ for genuine score s. The boxplots only include comparison scores if the uncompressed version of the same image was matched below the FMR = 0.001 threshold. Above the boxplots are FNMR values at FMR = 10^{-3}. The three blue lines correspond, from the top, to FMR of $10^{\{-2,-3,-4\}}$. The lower grey line refers to the median score obtained from comparison of uncompressed KIND 3 images. Any comparison for which either template had not been generated is excluded. Note that the iris record size on the horizontal axis is not evenly spaced above 3000 bytes.

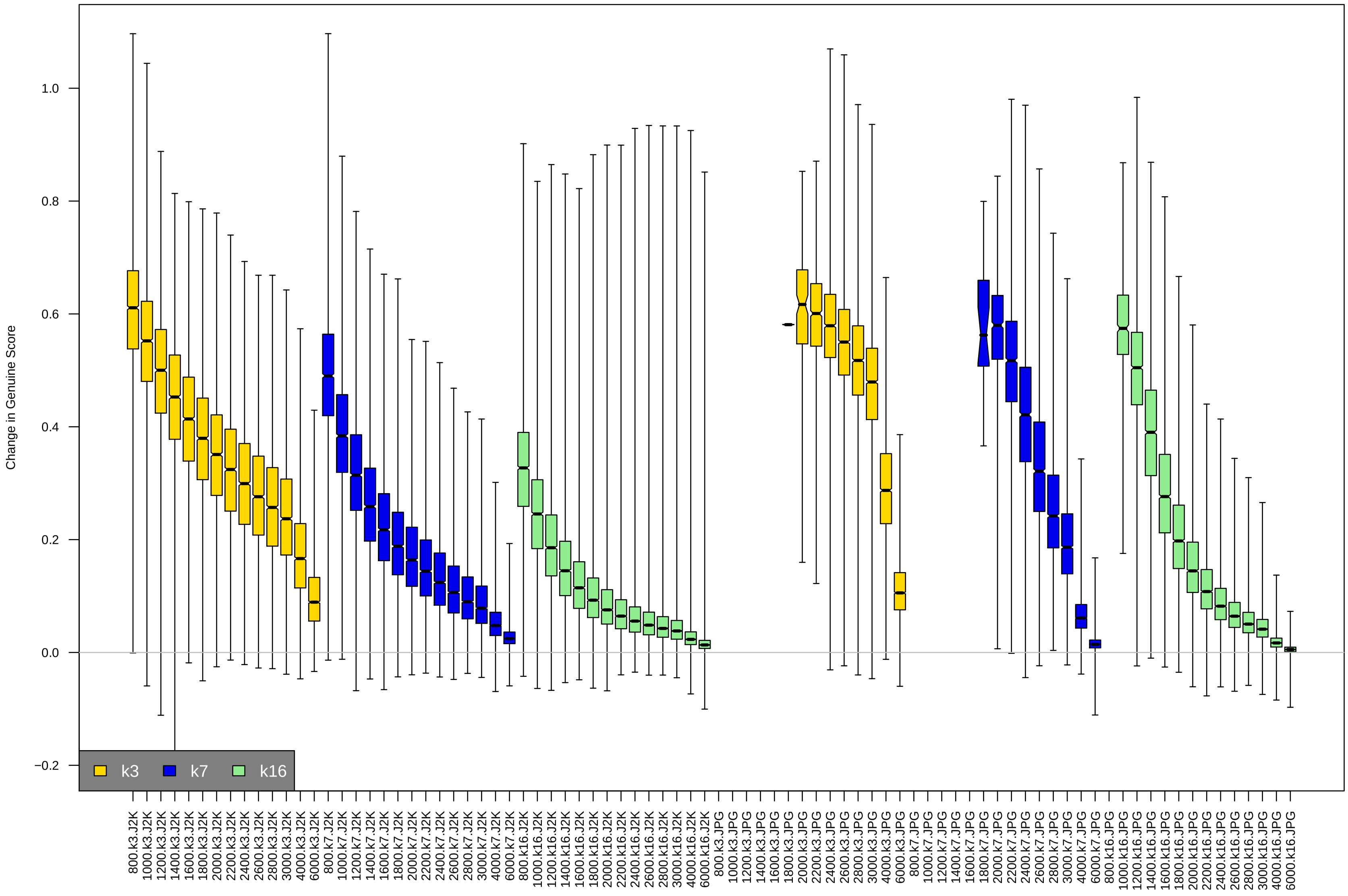


Figure 21: The distribution of the *increase* in I1 native genuine comparison scores between the uncompressed "parent" and the compressed image, arranged by size, KIND and the compression algorithm. The images are from the OPS dataset. Any comparison involving a failed template is excluded. Note that the iris record size on the horizontal axis is not evenly spaced above 3000 bytes.

8.1. Effect of compression

The application of lossy compression algorithms alters the pixel values in the iris region and in the limbic and pupillary boundary regions. When compression is low, these changes are inconsequential, but when compression is applied with sufficient severity the changes appear as visible artifacts and cause recognition errors. This is evident in Figure 19.

8.1.1. The genuine distribution under compression

Figure 20 shows the effect of JPEG and JPEG2000 compression on the genuine matching scores of the I1 SDK. Four thousand OPS images, in KIND 3, KIND 7, and KIND 16 format, are compressed to attain codestream sizes from 6000 bytes down to 800 bytes. These are compared with a second capture in uncompressed KIND 1 format. The results are displayed as groups of six boxplots: one for each of the three formats under JPEG2000 compression and an analogous three for JPEG. Each boxplot summarizes the distribution of the genuine scores. High scores lead to false rejection. The figure

- plots genuine score on the left axis as the primary response variable,
- plots a separability measure on the right axis,
- shows the median uncompressed score as a grey horizontal baseline,
- shows, as horizontal lines from top to bottom, thresholds corresponding to global FMR rates of 10^{-4}, 10^{-3} and 10^{-2} - scores above a horizontal line would be falsely rejected, and
- gives FNMR values in the top row.

Identical figures for all IREX SDKs are shown in the appendices of the IREX SUPPLEMENT[39]. The notable results are as follows.

- The genuine scores degrade much more rapidly with JPEG compression. This applies across all recognition algorithms and all formats. This contraindicates use of JPEG.
- The cropped and masked KIND 7 record can be compressed to about half the size of the KIND 3 record for the same increase in genuine comparison score.
- The boxplots themselves don't show the fraction of the genuine scores above the threshold lines. Those numbers are presented at the top of each plot. For KIND 16 some SDKs, particularly A1, G1 and J1, give fewer false rejections with JPEG than with JPEG2000. This arises because most of the size reduction for these SDKs comes from radial and circumferential sampling such that the compression itself is very light. Note also that any DCT tiling artifacts will not appear as such after reverse polar transformation.
- The number of scores in the tails is not directly dependent on the change in the scores. This arises because the upward drift of the scores is related to the gradual escalating compression damage to the iris texture, whereas outright false non-matches are the result of a failed localization of the relevant pupil and limbic boundaries.
- The KIND 3 and KIND 7 records tend to give fewer false rejections - than KIND 16. Comparing KIND 3 and KIND 7, some SDKs demonstrate lower FNMR on the former. This topic is addressed more fully in later subsections.

[39] Available at http://iris.nist.gov/irex/irex_appendices.pdf

8.1.2. CHANGE IN THE GENUINE DISTRIBUTION UNDER COMPRESSION

Figure 21 presents the *change* in genuine score under compression. It uses the same annotated boxplot format. Across all SDKs the key observations from the change boxplots are as follows.

- ▷ As more compression is applied, genuine scores rise asymptotically from zero. From uncompressed to 6000 bytes, the median change in genuine score is close to zero. In addition, the inter-quartile range depicted by the colored boxes is also very small. The IREX study has not established a functional model for the change in score $\Delta d(c)$ at compression ratio c, but the figure shows that any model will have smaller residuals if it incorporates the uncompressed score $d(1)$ vs. a fixed intercept.

- ▷ More important than the change in genuine score, however, is that it is long-tailed. The change in some cases is sufficient to cause false rejection and is evidence of a failed localization. This is evident in the second and third flights of boxplots in Figure 20 where the tails of the KIND 16 extend much higher than KIND 7 and include error even with the lightest compression (i.e. at 6000 bytes).

8.1.3. THE IMPOSTOR DISTRIBUTION UNDER COMPRESSION

Figures 22 and 23 are the impostor analogues of the prior two figures. They show, for the I1 SDK , the distributions of the impostor scores, and the change in impostor scores, under compression. Identical figures for all IREX SDKs are shown the appendices of the IREX SUPPLEMENT[40]. To show any unwelcome reduction in scores under compression (i.e. increased false match rate) the plots include horizontal lines which give the $\text{FMR} = 10^{-4}, 10^{-3}$ and 10^{-2} quantiles of the uncompressed impostor distributions. The key results across SDKs are as follows.

- ▷ Under JPEG compression some SDKs (I1, I2, D1, E1) give an increased FMR - the impostor distribution moves to the left. While this usually only happens at sizes below 2000 bytes it contraindicates use of JPEG compression, particularly in one-to-many applications.

- ▷ Under JPEG compression, some SDKs (B1, F1, G1, H1) show movement of the impostor distribution to the right (higher). This is benign. It is indicative that the feature based representation of the iris is not sensitive to the DCT tiling artifacts produced under JPEG compression.

- ▷ Under JPEG2000 compression, the impostor distributions are more stable. There is a slight increase in variance of the impostor scores under compression. Some SDKs (I1, G1, H1) exhibit a decrease in impostor scores which manifests itself as higher FMR in the DETs of the next section.

- ▷ The figures do not show that any of the compact formats is better or worse with respect to the properties of the impostor distribution. This aspect is discussed in later section.

8.1.4. THE EFFECT OF COMPRESSION ON FMR

The boxplots of the previous section were essentially qualitative in their presentation of the effect of compression. In addition, by conducting only 4000 impostor comparisons, the data did not support quantification of the effect of compression on the relevant metric, FMR. In this section a number of different figures and tables give more precise statements of the effect.

Figure 24 shows DET characteristics for four selected SDKs matching instances of the four standard IREX records. The compact formats, KIND 3, KIND 7, and KIND 16, were prepared by the I1 SDK and compressed to 3000 bytes using JPEG2000.

[40] Available at http://iris.nist.gov/irex/irex_appendices.pdf

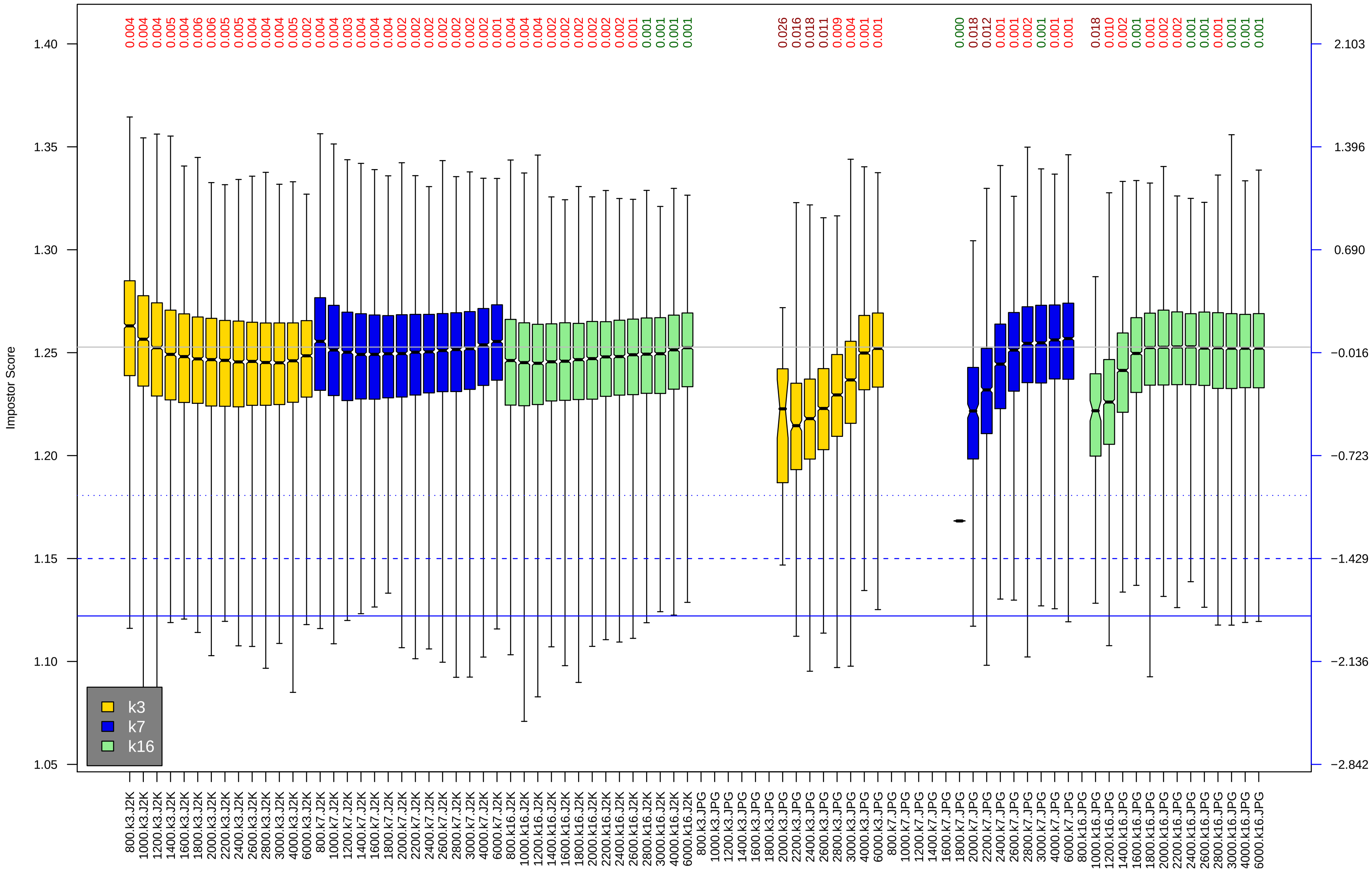


Figure 22: The distribution of I1 native impostor comparison scores by size of the compressed image, KIND and the compression algorithm. The right axis scale gives the corresponding value for $d' = (s - \mu_I)/\sqrt{0.5(\sigma_I^2 + \sigma_G^2)}$ for impostor score s. The three blue lines correspond, from the top, to FMR of $10^{\{-2,-3,-4\}}$. The lower grey line refers to the median score obtained from comparison of uncompressed KIND 3 images. Any comparison involving a failed template is excluded. Above the boxplots are FMR values at the threshold that gives FMR $= 10^{-3}$ on uncompressed images. These figures are computed from only 4000 comparisons so the FMR values and the tails of the impostor distribution are poorly characterized. Note that the iris record size on the horizontal axis is not evenly spaced above 3000 bytes.

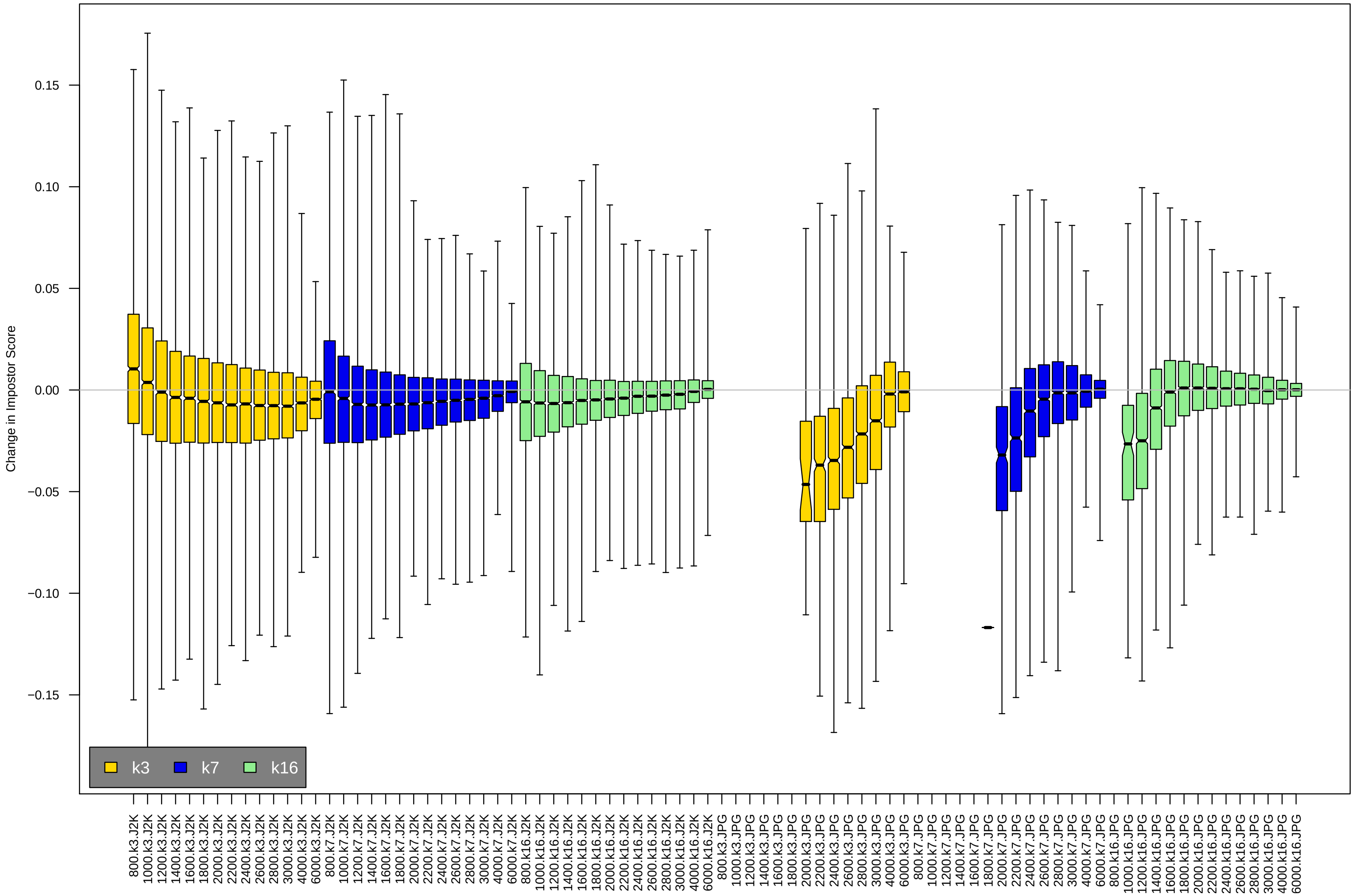


Figure 23: The distribution of the increase in I1 native impostor comparison scores between the uncompressed "parent" and the compressed image, arranged by size, KIND and the compression algorithm. The images are from the OPS dataset. Any comparison involving a failed template is excluded. Note that the iris record size on the horizontal axis is not evenly spaced above 3000 bytes.

All KIND 1 records were used without compression. The I1 SDK was adopted as a reference generator because it is one the implementations that never failed to produce IREX records, and because it gives good performance natively. The use of a common reference allows comparison of the resilience of the various matching algorithms under compression. It removes cropping margins as a confounding factor.

The Figure 24 DETs are computed for the ALL-FAILURES partition of the OPS dataset. As described in section 5.1.1, the subset is the union of all difficult comparisons, and therefore gives artificially high FNMR values. For comparative testing this is of no consequence. The notable results are as follows.

▷ For all SDKs and formats, FMR with compressed enrollment images differs from the baseline value obtained for uncompressed images. For most SDKs FMR increases as compression is applied. A notable exception is B1 for which FMR always decreases. The largest changes are observed for I1 and I2 where FMR changes by up to a factor of 10.

▷ The smallest changes are observed for the KIND 7 format, followed (usually) by KIND 16 and finally KIND 3. This order reflects the resilience of the format to compression: KIND 7 reserves more bits for encoding of the iris texture; KIND 3 the fewest.

▷ The variation is related to the amount of compression. For the dotted lines, the numbers of bits per pixel is 0.2; for the solid lines, bpp varies between 0.15 to 0.21 (the fifth and ninety-fifth percentiles) depending on the image size. It is generally the case that the change in FMR is larger for the solid lines.

8.2. SUFFICIENCY OF THE STANDARD FORMATS

The original and main goal of IREX is to establish formats and compression parameters for compact iris images. A necessary condition for a biometric data interchange format is that it should offer recognition performance comparable with that available from unconstrained, non-standard formats. This requirement is more fundamental than whether the format is interoperable. The word *sufficiency* refers to some statement of whether the data format is quantitatively sufficient, i.e. fit-for-purpose. This section gives quantitative statements of sufficiency.

The candidate formats, KIND 3, KIND 7 and KIND 16, were described and depicted in Figure 3. Images in these formats were subjected to PNG , JPEG and JPEG2000 compression. These algorithms were described in section 3. The viability of various combinations is assessed in terms of the core recognition error rates measured when samples are compared. Thus performance is stated by using DET characteristics and tabulating FNMR values at fixed FMR points. In addition we tabulate FNMR and FMR values at fixed threshold. Note that the most recent publications on the compression of iris data[19, 54] both quoted FNMR at FMR = 0.0001 or 0.001.

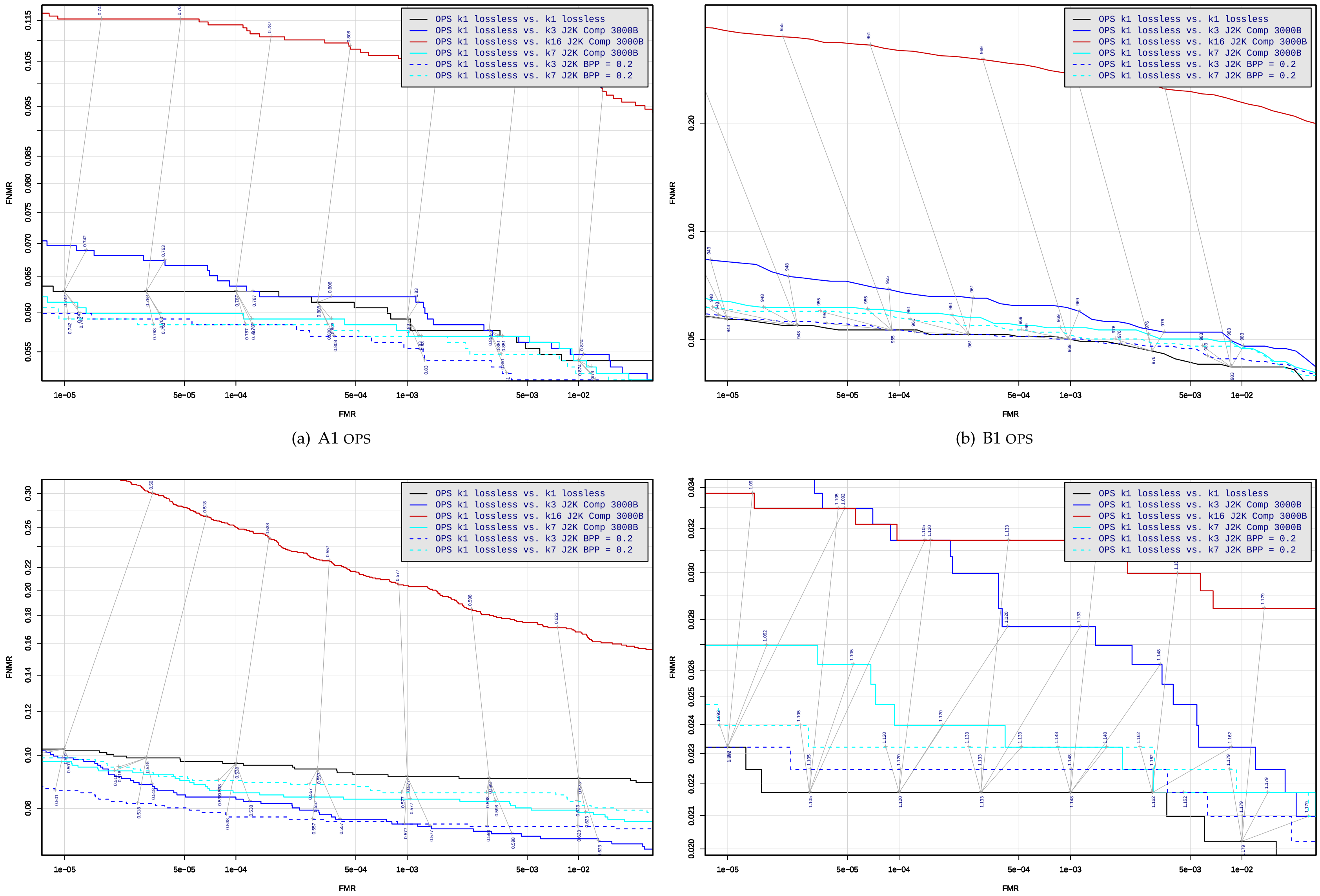
To establish the suitability of the three formats for compact representation of iris images, this section reports fundamental matching error rates for images from three datasets. It reports results for the application scenarios given in section 1.1.

Sufficiency results are tabulated in Tables 9, 10 and 11. Each table reports, for each IREX SDK, four different performance metrics. These are presented in three sub-tables (a), (b) and (c).

▷ Table(a): FNMR at the threshold, τ_b, fixed to give FMR = 0.001 over the approximately 10^9 KIND 1 vs. KIND 1 impostor comparisons present in the OPS - to - OPS and OPS - to - ICE comparison sets.

▷ Table(a): FNMR at a threshold fixed to give FMR = 0.001 on the specific combination of formats and compression sizes in use[41].

[41] This bullet and the preceding one represent two different approaches to setting the threshold: The first sets a global threshold whatever the kind and source of the data; the second tailors the threshold to the specific formatting of the data. If the iris recognition algorithm's impostor distribution was monolithic and invariant to format and compression, the threshold could be global.



(a) A1 OPS

(b) B1 OPS

(c) E1 OPS

(d) I1 OPS

Figure 24: DET characteristics for selected SDKs applied to records of the three KINDS compressed using JPEG2000 to 3000 bytes. The images are the pairs in the ALL-FAILURES partition of the OPS dataset: This leads to higher error rates than for the entire OPS dataset. Also shown is the DET for the same images without compression, in KIND 1 format. All the KIND 3, KIND 7 and KIND 16 images were produced by the I1 SDK which is adopted here as a reference for the purpose of normalizing away geometrical effects such as cropping margins and polar sampling rates. The grey lines link points of equal threshold. Note the FMR variation for some SDKs across KINDS.

(a) OPS FNMR

	Kind and Compression				Native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0047	0.0176	0.0040	0.0104	0.0331	0.0318	0.0215	0.0145	0.0075	0.0072	0.0319	0.0077	0.0074	0.0362	0.0356	0.0018	0.0017	0.0110	0.0047
2	k3	Uncomp	k1	Uncomp	0.0039	0.0167	0.0042	0.0104	0.0443	0.0445	0.0210	0.0138	0.0071	0.0068	0.0314	0.0076	0.0073	0.0341	0.0341	0.0018	0.0017	0.0109	0.0045
3	k7	Uncomp	k1	Uncomp	0.0042	0.0165	0.0041	0.0105	0.0342	0.0337	0.0224	0.0154	0.0111	0.0105	0.0325	0.0109	0.0097	0.0398	0.0382	0.0019	0.0018	0.0110	0.0045
4	k16	Uncomp	k1	Uncomp	0.0050	0.0173	0.0642	0.1094	-	-	-	-	-	-	0.0325	0.0140	0.0135	0.0457	0.0477	0.0019	0.0018	0.0110	0.0048
5	k3	J2K 2000B	k1	Uncomp	0.0054	0.0187	0.0070	0.0122	0.0420	0.0461	0.0215	0.0151	0.0070	0.0067	0.0307	0.0142	0.0135	0.0404	0.0376	0.0091	0.0091	0.0355	0.0328
6	k3	J2K 2000B	k1	Uncomp	0.0055	0.0197	0.0074	0.0116	-	0.0479	0.0215	0.0154	0.0070	0.0067	0.0319	0.0168	0.0159	0.0446	0.0436	0.0143	0.0143	0.0472	0.0437
7	k7	J2K 2000B	k1	Uncomp	0.0041	0.0167	0.0046	0.0104	0.1402	0.1398	0.0222	0.0156	0.0113	0.0107	0.0312	0.0126	0.0118	0.0420	0.0398	0.0020	0.0022	0.0108	0.0054
8	k7	J2K 2000B	k1	Uncomp	0.0042	0.0169	0.0046	0.0103	0.1411	0.1411	0.0222	0.0157	0.0112	0.0107	0.0314	0.0135	0.0127	0.0442	0.0434	0.0023	0.0024	0.0113	0.0058
9	k16	J2K 2000B	k1	Uncomp	0.0140	0.0292	0.1802	0.2493	-	-	-	-	-	-	0.1121	0.0570	0.0618	0.0549	0.0521	0.0137	0.0135	0.0545	0.0561
10	k16	J2K 2000B	k1	Uncomp	0.0144	0.0315	0.1802	0.2317	-	-	-	-	-	-	0.1134	0.0657	0.0701	0.0623	0.0638	0.0248	0.0245	0.0680	0.0776
11	k3	J2K 2000B	k3	J2K 2000B	0.0093	0.0237	0.0050	0.0080	0.0537	0.0589	0.0203	0.0148	0.0070	0.0070	0.0309	0.0254	0.0244	0.0507	0.0416	0.0055	0.0051	0.0419	0.0450
12	k7	J2K 2000B	k7	J2K 2000B	0.0035	0.0164	0.0039	0.0056	0.2322	0.2290	0.0229	0.0167	0.0142	0.0135	0.0309	0.0159	0.0156	0.0481	0.0403	0.0020	0.0023	0.0115	0.0061
13	k16	J2K 2000B	k16	J2K 2000B	0.0386	0.0555	0.2725	0.4031	-	-	-	-	-	-	0.1826	0.1173	0.1250	0.0841	0.0653	0.0064	0.0063	0.0594	0.0739
14	k3	J2K 2000B	k7	J2K 2000B	0.0051	0.0192	0.0042	0.0063	0.1491	0.1484	0.0222	0.0161	0.0104	0.0099	0.0317	0.0218	0.0209	0.0484	0.0405	0.0048	0.0051	0.0271	0.0241
15	k3	J2K 2000B	k16	J2K 2000B	0.0244	0.0420	0.1526	0.2471	-	-	-	-	-	-	0.1107	0.0757	0.0806	0.0705	0.0540	0.0064	0.0064	0.0538	0.0632
16	k7	J2K 2000B	k16	J2K 2000B	0.0173	0.0330	0.1539	0.2286	-	-	-	-	-	-	0.1126	0.0839	0.0863	0.0668	0.0540	0.0071	0.0076	0.0436	0.0483

(b) OPS cFNMR

	Kind and Compression				Conditional native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	k3	Uncomp	k1	Uncomp	0.0001	0.0001	0.0004	0.0002	0.0139	0.0151	0.0004	0.0004	0.0001	0.0001	0.0007	0.0001	0.0000	0.0010	0.0014	0.0001	0.0000	0.0006	0.0005
3	k7	Uncomp	k1	Uncomp	0.0002	0.0002	0.0004	0.0004	0.0040	0.0043	0.0016	0.0016	0.0047	0.0043	0.0012	0.0048	0.0036	0.0071	0.0058	0.0001	0.0001	0.0006	0.0004
4	k16	Uncomp	k1	Uncomp	0.0011	0.0013	0.0604	0.1002	-	-	-	-	-	-	0.0018	0.0071	0.0074	0.0112	0.0137	0.0001	0.0001	0.0005	0.0010
5	k3	J2K 2000B	k1	Uncomp	0.0017	0.0031	0.0030	0.0025	0.0141	0.0187	0.0013	0.0016	0.0005	0.0004	0.0029	0.0073	0.0071	0.0083	0.0061	0.0074	0.0074	0.0260	0.0288
6	k7	J2K 2000B	k1	Uncomp	0.0002	0.0009	0.0006	0.0005	0.1148	0.1151	0.0017	0.0019	0.0048	0.0044	0.0025	0.0061	0.0056	0.0097	0.0079	0.0003	0.0006	0.0007	0.0014
7	k16	J2K 2000B	k1	Uncomp	0.0100	0.0133	0.1769	0.2417	-	-	-	-	-	-	0.0866	0.0506	0.0556	0.0213	0.0187	0.0120	0.0118	0.0447	0.0525
8	k3	J2K 2000B	k3	J2K 2000B	0.0062	0.0093	0.0015	0.0020	0.0279	0.0337	0.0014	0.0024	0.0012	0.0013	0.0061	0.0191	0.0183	0.0211	0.0126	0.0038	0.0035	0.0331	0.0419
9	k7	J2K 2000B	k7	J2K 2000B	0.0004	0.0012	0.0005	0.0004	0.2110	0.2088	0.0029	0.0037	0.0087	0.0083	0.0044	0.0098	0.0098	0.0177	0.0111	0.0004	0.0007	0.0020	0.0029
10	k16	J2K 2000B	k16	J2K 2000B	0.0351	0.0411	0.2698	0.3978	-	-	-	-	-	-	0.1608	0.1113	0.1195	0.0519	0.0338	0.0050	0.0049	0.0512	0.0707
11	k3	J2K 2000B	k7	J2K 2000B	0.0023	0.0042	0.0008	0.0007	0.1258	0.1259	0.0024	0.0032	0.0048	0.0046	0.0061	0.0154	0.0151	0.0183	0.0114	0.0032	0.0035	0.0181	0.0209
12	k3	J2K 2000B	k16	J2K 2000B	0.0211	0.0271	0.1493	0.2396	-	-	-	-	-	-	0.0872	0.0695	0.0746	0.0395	0.0240	0.0047	0.0049	0.0454	0.0600
13	k7	J2K 2000B	k16	J2K 2000B	0.0139	0.0178	0.1504	0.2207	-	-	-	-	-	-	0.0884	0.0776	0.0802	0.0356	0.0236	0.0054	0.0059	0.0348	0.0449

(c) OPS FMR

	Kind and Compression				Native false match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0014	0.0012	0.0012	0.0010	0.0009	0.0009	0.0006	0.0016	0.0005	0.0006	0.0005	0.0012	0.0012	0.0009	0.0013	0.0012	0.0012	0.0014	0.0015
2	k3	J2K 2000B	k1	Uncomp	0.0013	0.0020	0.0013	0.0006	-	0.0016	0.0012	0.0025	0.0007	0.0007	0.0018	0.0023	0.0023	0.0022	0.0045	0.0049	0.0048	0.0026	0.0026
3	k7	J2K 2000B	k1	Uncomp	0.0016	0.0017	0.0011	0.0005	0.0016	0.0016	0.0008	0.0019	0.0006	0.0006	0.0011	0.0016	0.0017	0.0019	0.0036	0.0028	0.0028	0.0028	0.0028
4	k16	J2K 2000B	k1	Uncomp	0.0012	0.0020	0.0009	0.0005	-	-	-	-	-	-	0.0017	0.0022	0.0022	0.0022	0.0046	0.0052	0.0052	0.0022	0.0023

Table 9: Accuracy for OPS records stored in the KINDS identified in the rows. The compression applied to the enrollment and verification samples is identified in columns three and five. The remaining column headers identify the SDKs. All comparisons are native. **Top:** FNMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons *except* in rows with red text where FNMR is given at the threshold for which FMR = 0.001 on the kind of images identified in the row. Cells appear dark green when the error rate is less than the baseline on row 1, light green when it is less than the baseline times 1.1, light red when it exceeds three times the baseline, and dark red when the factor is more than ten times. **Center:** *Conditional* FNMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. FNMR is conditional on the same parent KIND 1 image being correctly matched at the same threshold. Cells appear dark green when less than or equal to 0.005, light green when the cFNMR is less than or equal to 0.01, light red when error rate is above 0.03, and dark red above 0.1. **Bottom:** FMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. The first row elements are not 0.001 because the global threshold was established across all OPS - to - OPS and OPS - to - ICE comparisons. Cells appear green when FMR is less than or equal to 0.001, light red when above 0.002, and dark red above 0.003.

(a) ICE FNMR

	Kind and Compression				Native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0116	0.0174	0.0063	0.0102	0.0335	0.0342	0.0176	0.0205	0.0256	0.0251	0.0197	0.0326	0.0317	0.0248	0.0276	0.0056	0.0053	0.0124	0.0103
2	k3	Uncomp	k1	Uncomp	0.0095	0.0152	0.0067	0.0110	0.0344	0.0379	0.0167	0.0194	0.0244	0.0239	0.0198	0.0307	0.0312	0.0255	0.0287	0.0056	0.0052	0.0118	0.0087
3	k7	Uncomp	k1	Uncomp	0.0094	0.0149	0.0069	0.0112	0.0339	0.0345	0.0161	0.0206	0.0244	0.0237	0.0201	0.0505	0.0459	0.0273	0.0301	0.0056	0.0055	0.0118	0.0086
4	k16	Uncomp	k1	Uncomp	0.0095	0.0150	0.1546	0.2148	-	-	-	-	-	-	0.0382	0.0357	0.0354	0.0261	0.0284	0.0055	0.0053	0.0111	0.0088
5	k3	J2K 2000B	k1	Uncomp	0.0240	0.0357	0.0213	0.0311	0.0643	0.0690	0.0220	0.0326	0.0306	0.0310	0.0335	0.0704	0.0699	0.0633	0.0606	0.0291	0.0287	0.0884	0.0938
6	k3	J2K 2000B	k1	Uncomp	0.0247	0.0399	0.0245	0.0291	0.0823	0.0870	0.0307	0.0352	0.0367	0.0365	0.0544	0.0785	0.0778	0.0829	0.0879	0.0509	0.0504	0.0990	0.1046
7	k7	J2K 2000B	k1	Uncomp	0.0121	0.0179	0.0099	0.0150	0.0461	0.0467	0.0167	0.0220	0.0245	0.0241	0.0264	0.0564	0.0561	0.0410	0.0406	0.0082	0.0087	0.0381	0.0392
8	k7	J2K 2000B	k1	Uncomp	0.0124	0.0187	0.0106	0.0145	0.0534	0.0541	0.0189	0.0223	0.0257	0.0251	0.0367	0.0583	0.0581	0.0499	0.0533	0.0111	0.0109	0.0428	0.0440
9	k16	J2K 2000B	k1	Uncomp	0.0522	0.0715	0.3360	0.4115	-	-	-	-	-	-	0.1076	0.1400	0.1434	0.0834	0.0786	0.0458	0.0456	0.1717	0.1711
10	k16	J2K 2000B	k1	Uncomp	0.0534	0.0833	0.3588	0.4024	-	-	-	-	-	-	0.1571	0.1611	0.1638	0.1146	0.1231	0.0961	0.0960	0.1925	0.1915
11	k3	J2K 2000B	k3	J2K 2000B	0.0454	0.0631	0.0136	0.0202	0.0856	0.0947	0.0246	0.0462	0.0398	0.0405	0.0499	0.1089	0.1106	0.1049	0.0740	0.0239	0.0238	0.1025	0.1181
12	k7	J2K 2000B	k7	J2K 2000B	0.0129	0.0188	0.0079	0.0103	0.0586	0.0587	0.0162	0.0235	0.0234	0.0229	0.0302	0.0622	0.0624	0.0593	0.0468	0.0088	0.0099	0.0543	0.0609
13	k16	J2K 2000B	k16	J2K 2000B	0.1272	0.1542	0.3955	0.5090	-	-	-	-	-	-	0.1801	0.2478	0.2550	0.1518	0.1004	0.0342	0.0343	0.1843	0.1982
14	k3	J2K 2000B	k7	J2K 2000B	0.0267	0.0386	0.0114	0.0153	0.0761	0.0797	0.0215	0.0358	0.0311	0.0316	0.0406	0.0932	0.0938	0.0876	0.0664	0.0216	0.0222	0.0862	0.0983
15	k3	J2K 2000B	k16	J2K 2000B	0.0878	0.1141	0.3142	0.4348	-	-	-	-	-	-	0.1308	0.1843	0.1890	0.1288	0.0878	0.0330	0.0331	0.1702	0.1920
16	k7	J2K 2000B	k16	J2K 2000B	0.0628	0.0825	0.3077	0.4076	-	-	-	-	-	-	0.1188	0.1909	0.1922	0.1052	0.0750	0.0337	0.0348	0.1559	0.1727

(b) ICE cFNMR

	Kind and Compression				Conditional native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	k3	Uncomp	k1	Uncomp	0.0009	0.0009	0.0007	0.0013	0.0063	0.0096	0.0019	0.0026	0.0008	0.0007	0.0027	0.0004	0.0004	0.0038	0.0050	0.0003	0.0002	0.0017	0.0012
3	k7	Uncomp	k1	Uncomp	0.0011	0.0012	0.0010	0.0017	0.0061	0.0068	0.0031	0.0052	0.0046	0.0044	0.0025	0.0245	0.0198	0.0065	0.0075	0.0007	0.0008	0.0016	0.0012
4	k16	Uncomp	k1	Uncomp	0.0016	0.0020	0.1508	0.2090	-	-	-	-	-	-	0.0227	0.0106	0.0116	0.0052	0.0059	0.0004	0.0005	0.0012	0.0013
5	k3	J2K 2000B	k1	Uncomp	0.0153	0.0217	0.0155	0.0217	0.0368	0.0409	0.0079	0.0164	0.0098	0.0105	0.0170	0.0462	0.0460	0.0428	0.0383	0.0243	0.0242	0.0799	0.0874
6	k7	J2K 2000B	k1	Uncomp	0.0034	0.0039	0.0040	0.0055	0.0183	0.0184	0.0038	0.0063	0.0050	0.0050	0.0091	0.0304	0.0305	0.0199	0.0181	0.0033	0.0039	0.0286	0.0319
7	k16	J2K 2000B	k1	Uncomp	0.0438	0.0584	0.3350	0.4094	-	-	-	-	-	-	0.0930	0.1159	0.1203	0.0633	0.0565	0.0411	0.0413	0.1647	0.1659
8	k3	J2K 2000B	k3	J2K 2000B	0.0386	0.0514	0.0086	0.0130	0.0609	0.0694	0.0127	0.0328	0.0221	0.0232	0.0348	0.0890	0.0902	0.0866	0.0544	0.0197	0.0198	0.0958	0.1136
9	k7	J2K 2000B	k7	J2K 2000B	0.0060	0.0069	0.0028	0.0034	0.0333	0.0330	0.0063	0.0108	0.0076	0.0075	0.0139	0.0397	0.0397	0.0399	0.0267	0.0043	0.0053	0.0466	0.0555
10	k16	J2K 2000B	k16	J2K 2000B	0.1216	0.1446	0.3964	0.5102	-	-	-	-	-	-	0.1683	0.2291	0.2375	0.1348	0.0816	0.0303	0.0307	0.1789	0.1947
11	k3	J2K 2000B	k7	J2K 2000B	0.0197	0.0267	0.0063	0.0080	0.0511	0.0542	0.0104	0.0226	0.0140	0.0148	0.0249	0.0721	0.0725	0.0688	0.0465	0.0172	0.0178	0.0792	0.0935
12	k3	J2K 2000B	k16	J2K 2000B	0.0816	0.1035	0.3129	0.4332	-	-	-	-	-	-	0.1175	0.1648	0.1699	0.1112	0.0685	0.0289	0.0292	0.1645	0.1884
13	k7	J2K 2000B	k16	J2K 2000B	0.0562	0.0714	0.3063	0.4055	-	-	-	-	-	-	0.1049	0.1699	0.1720	0.0868	0.0553	0.0293	0.0306	0.1499	0.1688

(c) ICE FMR

	Kind and Compression				Native false match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0010	0.0010	0.0023	0.0024	0.0017	0.0016	0.0024	0.0009	0.0021	0.0021	0.0035	0.0009	0.0009	0.0013	0.0009	0.0015	0.0015	0.0005	0.0005
2	k3	J2K 2000B	k1	Uncomp	0.0011	0.0016	0.0015	0.0007	0.0029	0.0029	0.0050	0.0015	0.0031	0.0027	0.0068	0.0017	0.0017	0.0025	0.0039	0.0064	0.0064	0.0014	0.0014
3	k7	J2K 2000B	k1	Uncomp	0.0011	0.0013	0.0012	0.0006	0.0026	0.0026	0.0029	0.0010	0.0024	0.0021	0.0056	0.0012	0.0012	0.0023	0.0032	0.0036	0.0028	0.0013	0.0013
4	k16	J2K 2000B	k1	Uncomp	0.0011	0.0018	0.0014	0.0007	-	-	-	-	-	-	0.0070	0.0020	0.0019	0.0027	0.0044	0.0085	0.0084	0.0014	0.0015

Table 10: Accuracy for ICE records stored of the KINDS identified in the rows. The compression applied to the enrollment and verification samples is identified in columns three and five. The remaining column headers identify the SDKs. All comparisons are native. **Top:** FNMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons *except* in rows with red text where FNMR is given at the threshold for which FMR = 0.001 on the kind of images identified in the row. Cells appear dark green when the error rate is less than the baseline on row 1, light green when it is less than the baseline times 1.1, light red when it exceeds three times the baseline, and dark red when the factor is more than ten times. **Center:** *Conditional* FNMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. FNMR is conditional on the same parent KIND 1 image being correctly matched at the same threshold. Cells appear dark green when less than or equal to 0.005, light green when the cFNMR is less than or equal to 0.01, light red when error rate is above 0.03, and dark red above 0.1. **Bottom:** FMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. The first row elements are not 0.001 because the global threshold was established across all OPS - to - OPS and OPS - to - ICE comparisons. Cells appear green when FMR is less than or equal to 0.001, light red when above 0.002, and dark red above 0.003.

(a) BATH FNMR

	Kind and Compression				Native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0119	0.0131	0.0056	0.0073	0.0222	0.0223	0.0206	0.0226	0.0332	0.0328	0.0459	0.0550	0.0374	0.0261	0.0258	0.0018	0.0016	0.0215	0.0215
2	k3	Uncomp	k1	Uncomp	0.0122	0.0138	0.0115	0.0156	0.2184	0.2277	0.0219	0.0239	0.0423	0.0422	0.0506	0.0546	0.0386	0.0365	0.0364	0.0017	0.0015	0.0215	0.7114
3	k7	Uncomp	k1	Uncomp	0.0123	0.0137	0.0122	0.0163	0.0600	0.0592	0.0224	0.0248	0.0351	0.0349	0.0463	0.0614	0.0459	0.0455	0.0417	0.0019	0.0021	0.0215	0.0236
4	k16	Uncomp	k1	Uncomp	0.0177	0.0193	0.4183	0.4897	-	-	-	-	-	-	0.0515	0.1113	0.1042	0.0497	0.0574	0.0019	0.0017	0.0216	0.0280
5	k3	J2K 2000B	k1	Uncomp	0.0218	0.0253	0.0156	0.0196	0.1595	0.1905	0.0231	0.0250	0.0425	0.0424	0.0538	0.0618	0.0509	0.0466	0.0456	0.0077	0.0072	0.0435	0.0452
6	k3	J2K 2000B	k1	Uncomp	0.0222	0.0270	0.0157	0.0193	-	0.2003	0.0240	0.0263	0.0446	0.0445	0.0573	0.0652	0.0536	0.0580	0.0570	0.0093	0.0088	0.0422	0.0424
7	k7	J2K 2000B	k1	Uncomp	0.0179	0.0203	0.0146	0.0183	0.1955	0.1946	0.0233	0.0258	0.0353	0.0351	0.0497	0.0645	0.0540	0.0553	0.0540	0.0052	0.0084	0.0232	0.0253
8	k7	J2K 2000B	k1	Uncomp	0.0183	0.0212	0.0147	0.0186	0.2058	0.2050	0.0242	0.0270	0.0370	0.0368	0.0527	0.0675	0.0563	0.0747	0.0713	0.0053	0.0082	0.0241	0.0264
9	k16	J2K 2000B	k1	Uncomp	0.0915	0.1060	0.6794	0.7277	-	-	-	-	-	-	0.1516	0.2459	0.2577	0.1250	0.1159	0.0720	0.0723	0.2170	0.2960
10	k16	J2K 2000B	k1	Uncomp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
11	k3	J2K 2000B	k3	J2K 2000B	0.0260	0.0303	0.0126	0.0161	0.1913	0.2201	0.0232	0.0254	0.0440	0.0440	0.0566	0.0689	0.0578	0.0501	0.0440	0.0036	0.0032	0.0357	0.0359
12	k7	J2K 2000B	k7	J2K 2000B	0.0195	0.0215	0.0127	0.0156	0.2787	0.2790	0.0235	0.0266	0.0298	0.0296	0.0511	0.0673	0.0583	0.0617	0.0543	0.0037	0.0049	0.0248	0.0268
13	k16	J2K 2000B	k16	J2K 2000B	0.1228	0.1585	0.1067	0.1354	-	-	-	-	-	-	0.2062	0.2801	0.2885	0.1449	0.1084	0.0180	0.0181	0.1264	0.1617
14	k3	J2K 2000B	k7	J2K 2000B	0.0240	0.0273	0.0134	0.0167	0.2906	0.3145	0.0244	0.0272	0.0427	0.0427	0.0566	0.0760	0.0677	0.0732	0.0638	0.0055	0.0101	0.0366	0.0375
15	k3	J2K 2000B	k16	J2K 2000B	0.1084	0.1400	0.6420	0.7304	-	-	-	-	-	-	0.1578	0.2743	0.2867	0.1633	0.1290	0.0223	0.0222	0.1813	0.2491
16	k7	J2K 2000B	k16	J2K 2000B	0.0981	0.1190	0.6628	0.7311	-	-	-	-	-	-	0.1589	0.3273	0.3406	0.1626	0.1302	0.0352	0.0496	0.1917	0.2600

(b) BATH cFNMR

	Kind and Compression				Conditional native false non-match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
2	k3	Uncomp	k1	Uncomp	0.0021	0.0023	0.0076	0.0097	0.2125	0.2228	0.0027	0.0030	0.0101	0.0103	0.0062	0.0017	0.0033	0.0153	0.0156	0.0002	0.0002	0.0002	0.7437
3	k7	Uncomp	k1	Uncomp	0.0024	0.0026	0.0082	0.0103	0.0418	0.0414	0.0028	0.0035	0.0066	0.0068	0.0010	0.0090	0.0115	0.0238	0.0205	0.0004	0.0008	0.0002	0.0030
4	k16	Uncomp	k1	Uncomp	0.0084	0.0087	0.4373	0.5120	-	-	-	-	-	-	0.0069	0.0642	0.0746	0.0286	0.0374	0.0004	0.0003	0.0003	0.0076
5	k3	J2K 2000B	k1	Uncomp	0.0122	0.0145	0.0114	0.0134	0.1491	0.1826	0.0042	0.0043	0.0110	0.0112	0.0097	0.0113	0.0181	0.0261	0.0254	0.0066	0.0063	0.0238	0.0261
6	k7	J2K 2000B	k1	Uncomp	0.0079	0.0091	0.0104	0.0121	0.1875	0.1868	0.0037	0.0044	0.0068	0.0070	0.0048	0.0121	0.0193	0.0342	0.0332	0.0039	0.0074	0.0019	0.0048
7	k16	J2K 2000B	k1	Uncomp	0.0862	0.1004	0.7139	0.7646	-	-	-	-	-	-	0.1174	0.2138	0.2419	0.1091	0.1001	0.0743	0.0748	0.2107	0.2960
8	k3	J2K 2000B	k3	J2K 2000B	0.0176	0.0208	0.0089	0.0105	0.1835	0.2148	0.0059	0.0065	0.0139	0.0141	0.0134	0.0226	0.0285	0.0298	0.0237	0.0026	0.0023	0.0154	0.0168
9	k7	J2K 2000B	k7	J2K 2000B	0.0104	0.0112	0.0089	0.0100	0.2773	0.2781	0.0046	0.0059	0.0085	0.0087	0.0067	0.0162	0.0247	0.0423	0.0348	0.0025	0.0038	0.0037	0.0070
10	k16	J2K 2000B	k16	J2K 2000B	0.1196	0.1565	0.1094	0.1396	-	-	-	-	-	-	0.1779	0.2526	0.2759	0.1307	0.0927	0.0176	0.0178	0.1132	0.1519
11	k3	J2K 2000B	k7	J2K 2000B	0.0151	0.0172	0.0096	0.0108	0.2903	0.3165	0.0062	0.0072	0.0140	0.0143	0.0131	0.0271	0.0367	0.0540	0.0445	0.0044	0.0093	0.0164	0.0185
12	k3	J2K 2000B	k16	J2K 2000B	0.1044	0.1369	0.6743	0.7675	-	-	-	-	-	-	0.1247	0.2468	0.2743	0.1504	0.1140	0.0222	0.0222	0.1723	0.2457
13	k7	J2K 2000B	k16	J2K 2000B	0.0934	0.1143	0.6964	0.7681	-	-	-	-	-	-	0.1256	0.3050	0.3327	0.1501	0.1157	0.0356	0.0509	0.1835	0.2574

(c) BATH FMR

	Kind and Compression				Native false match rates																		
	Enrollment		Verification		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2
1	k1	Uncomp	k1	Uncomp	0.0006	0.0006	0.0016	0.0025	0.0004	0.0004	0.0003	0.0005	0.0004	0.0004	0.0007	0.0005	0.0005	0.0021	0.0016	0.0004	0.0004	0.0003	0.0003
2	k3	J2K 2000B	k1	Uncomp	0.0006	0.0012	0.0004	0.0003	-	0.0009	0.0006	0.0009	0.0006	0.0005	0.0021	0.0010	0.0010	0.0032	0.0044	0.0014	0.0014	0.0005	0.0005
3	k7	J2K 2000B	k1	Uncomp	0.0007	0.0009	0.0003	0.0002	0.0010	0.0010	0.0004	0.0007	0.0005	0.0004	0.0016	0.0007	0.0007	0.0039	0.0046	0.0008	0.0007	0.0005	0.0005
4	k16	J2K 2000B	k1	Uncomp	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Table 11: Accuracy for BATH records stored of the KINDS identified in the rows. The compression applied to the enrollment and verification samples is identified in columns three and five. The remaining column headers identify the SDKs. All comparisons are native. **Top:** FNMR values at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons *except* in rows with red text where FNMR is given at the threshold that gives FMR = 0.001 on the kind of images identified in the row. Cells appear dark green when the error rate is less than the baseline on row 1, light green when it is less than the baseline times 1.1, light red when it exceeds three times the baseline, and dark red when the factor is more than ten times. **Center:** *Conditional* FNMR at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. FNMR is conditional on the same parent KIND 1 image being correctly matched at the same threshold. Cells appear light green when the cFNMR is less than or equal to 0.01, dark green when less than or equal to 0.005, light red when error rate is above 0.03, and are dark red above 0.1. **Bottom:** FMR values at the threshold for which FMR = 0.001 on uncompressed KIND 1 comparisons. The first row elements are not 0.001 because the global threshold was established across all OPS - to - OPS and OPS - to - ICE comparisons. Cells appear green when FMR is less than or equal to 0.001, light red when above 0.002, and dark red above 0.003.

- ▷ Table (b): To isolate the effect of the KIND on accuracy, the pure Table (a) FNMR results are restated only for the set of comparisons that are successful for raw unprocessed KIND 1 images. This corresponds to the conditional probability

$$\text{cFNMR}(\tau) = P(d_{jk} > \tau \mid d_{11} \leq \tau) = \frac{\sum_{d \in \mathcal{G}} H(d_{11} - \tau)(1 - H(d_{jk} - \tau))}{\sum_{d \in \mathcal{G}} H(d_{11} - \tau)} \qquad (15)$$

  where comparison score d_{jk} is the result of a genuine comparison of an enrollment sample of KIND j with a verification sample of KIND k. $\mathcal{G}$ represents the set of all genuine comparisons. This approach removes the effect of baseline SDK performance so that the values in row 1 of the conditional FNMR Tables 9, 10 and 11 are zero. The values in the remaining rows exclude comparisons that failed with unprocessed images and which might have then succeeded when, for example, the crop-only format was used.

- ▷ Table(c): The empirical FMR measured at threshold τ_b.

8.2.1. Core sufficiency

Referring to the sufficiency data for the three IREX datasets, i.e. Tables 9, 10 and 11, the notable results are as follows.

- ▷ When the compact formats are matched without compression (sub-table (b), rows 2 to 4) the error rates are often close to zero. This is true for many SDKs operating on OPS images, but less true for BATH . The observation of few format-related errors shows that the formats are not innately defective. For KIND 3 instances this shows that eye sockets are not necessary for accurate localization. Similarly for KIND 7 instances the presence of eyelid and sclera masking is not itself problematic. Finally, for KIND 16 instances, the rectilinear-to-polar and polar-to-rectilinear steps introduce essentially zero errors for I1, slight increases for others (A1, A2, J1, J2, F1), significant error (G1, G2, H1, H2) and catastrophic errors (B1, B2).
- ▷ Note however that for I1 the circumferential and radial sampling rates are markedly larger than those used by A1, A2, J1, J2 and F1, and while this might appear necessary to avoid initial incremental increases in FNMR, it is not true for H1 and H2 which use higher sampling rates still.
- ▷ The existence and prevalence of essentially zero error rates in rows 2 and 3 of each table supports the important conclusion that the KIND 3 and KIND 7 compact formats can be used with lossless PNG compression with very little elevation of error rate. From Figure 29, this allows iris records of sizes of approximately 60KB and 25KB, respectively, to be produced.
- ▷ This recommendation could reasonably extend to KIND 16 except for the elevation of FNMR on the BATH dataset.

Except as noted, these observations are consistent across all three IREX datasets.

With compression of the enrollment record

The results in the rows 5 to 10 of Tables 9, 10 and 11 correspond to the application of JPEG2000 to compress the compact enrollment record to 2000 bytes for comparison with an uncompressed KIND 1 instance. Error rates associated with less compressed records will be better - the general dependency on error rate is shown in section 8.1. The current purpose is to examine which formats are most suited for very compact storage. The following observations are drawn from rows 6, 8 and 10 of Tables 9, 10 and 11.

- ▷ The KIND 7 format is most resilient under compression. The prevalence of green-shaded elements for A1, A2, B1, B2, D1, D2 E1, E2, I1 and I2 show accuracy is well preserved under compression. This effect was described as *resilience* by Rakshit[54] and by Daugman who observed the formats to be *serviceable*[19] under heavy compression.

Note (Table 11(b)) that this finding is not really true for BATH images; compression itself does elevate error rates. The reason for this, versus OPS and ICE, is not known and worthy of further investigation particularly with respect to the photometric properties of the images.

- For KIND 3 the false non-match rates are generally worse than for KIND 7. However this result is both dataset and SDK-dependent. For BATH and ICE images the KIND 3 false rejection rates are usually much higher that for KIND 7. But, for OPS images, SDKs D1, D2, E1, E2 and F1 gives KIND 3 error rates are actually lower than those for KIND 7. See section 8.6.3 for further analysis.
- The KIND 16 error rates are substantially larger than those for KIND 7 with between a factor of two and six increase in error rate. This is true even for the I1 and I2 implementations from the provider that proposed the unsegmented polar format.

With compression of both the enrollment and verification records

The real strength of the KIND 7 record becomes apparent when both of the images involved in a comparison are severely compressed. Row 12 in Tables 9, 10 and 11 gives FNMR at FMR = 0.001 when JPEG2000 compression is applied to make KIND 7 records of size 2000 bytes. For OPS images, the performance of SDKs A1, A2, B1 and B2 is actually better than uncompressed KIND 1 instances. In addition, SDKs D1, D2, I1, I2, J1 and J2 are only slightly degraded. The situation for comparison of compressed KIND 3 pairs is usually considerably worse. Notable exceptions for OPS are D1, D2, E1, and E2 which tightly crop their KIND 3 and thereby elevate the bits-per-pixel. For ICE and BATH images these SDKs revert to KIND 7 substantially outperforming KIND 3.

Most SDKs perform worse with compression of both images in a comparison rather than with compression of only one. However, the B1 and I1 SDKs often prefer compression of both images (see rows 5 and 8 in the cFNMR entries of Tables 9, 10 and 11. This may be a result of competing effects: The compression degrades error rates but this is offset by reduced errors from using the centered formats vs. the parent KIND 1 . This begs the question: do the lowest overall error rates come from comparison of uncompresssed KIND 3 instances?

8.3. Conformance testing

Interoperability tests have been formally standardized by ISO/IEC 19795-4:2008 *Performance Interoperability Testing* which establishes procedures and reporting requirements for cross-provider tests and certification schemes. One of those requirements is that all samples should be *conformant* to the underlying standard. Conformance of a biometric component (in this case an IREX record generator) is most effectively assessed by executing defined test assertions on samples produced by that component. Such test assertions are formally prescribed by a conformance testing standard. For iris imagery that standard is ISO/IEC 29109-6 which checks instances of the ISO/IEC 19794-6 standard for conformance[42].

In any case, the conformance test assertions can be categorized into those addressing the syntax of the data (are the record values correct, and internally consistent?) and those governing the semantic content. This latter category would test, for example, that the KIND 3 iris really is centered, that the KIND 7 masks are indeed over the eyelids and sclera, and that the KIND 16 coordinate system is correct.

In the IREX trial itself, NIST checked all records for syntactic conformance. While the list of test assertions has not been formally documented, the "C" source code of the validating parser is available. Note that the binary structure of the revised ISO/IEC 19794-6 record has deviated (syntactically) from that finally adopted for the IREX trials.

[42] Unfortunately work in the SC 37 committee has thus far targeted the 2005 iris interchange standard and not the revision (which is now being completed) that IREX supports. A NIST comment to remedy this anachronism was submitted to the United States working group, M1, in April 2009 and approved on a 8-1-1 vote, but was not sent on to SC 37 due to a voting technicality - see M1.3 meeting minutes M1/09-0242. The authors will support actions to complete ISO/IEC 19794-6:201X and its conformance test as quickly as possible.

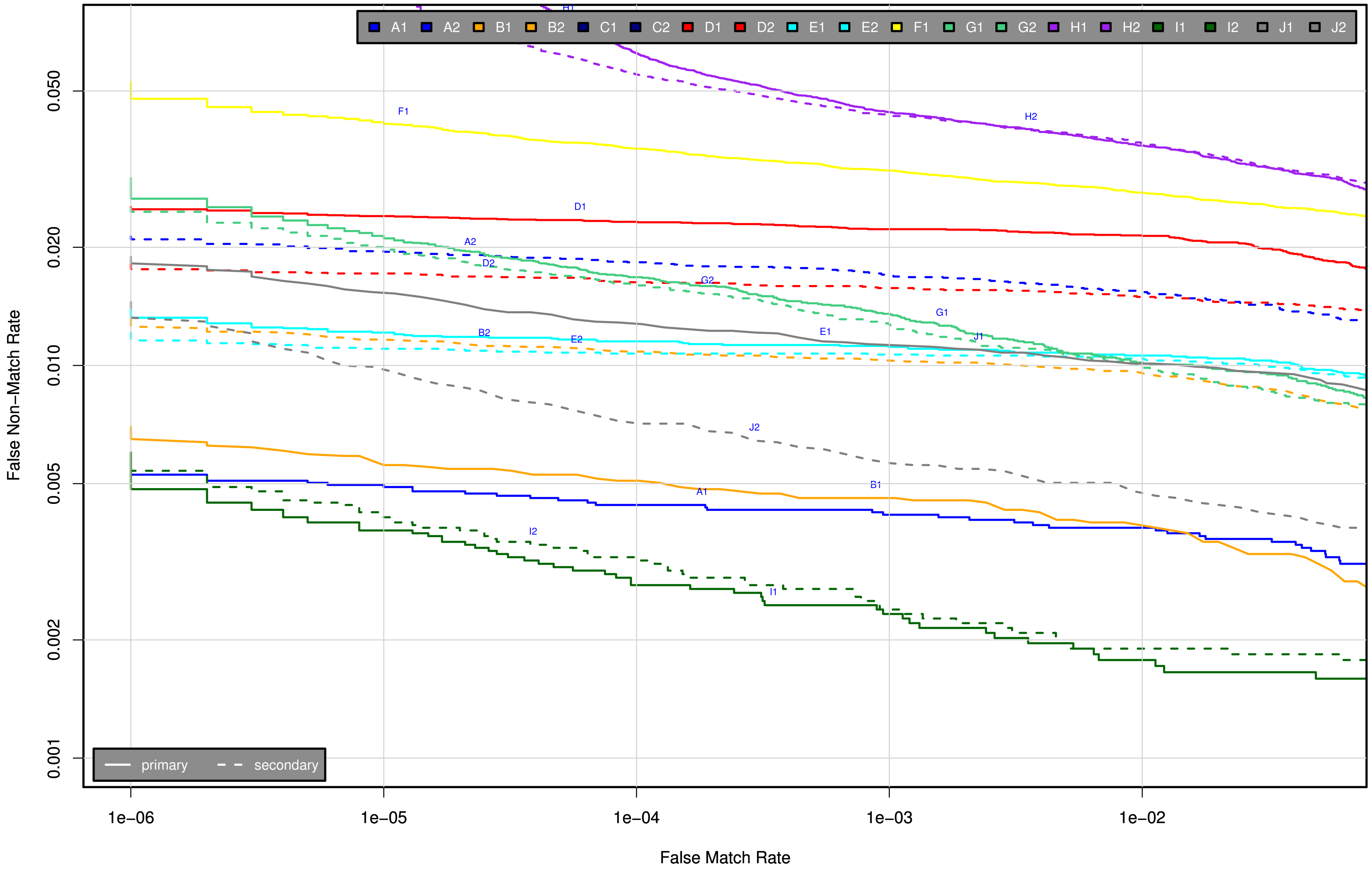


Figure 25: The DET characteristics for all IREX algorithms on the OPS dataset for comparison of 2000 byte JPEG2000 compressed KIND 7 images with raw KIND 1 images. Each SDK failed to produce templates from a fraction, $0 \leq \text{FTE} \leq 1$, of the images and the effects of that are included in this plot. The FTE rates appear in Table 6. For the OPS dataset this fraction was often zero (see Table 6).

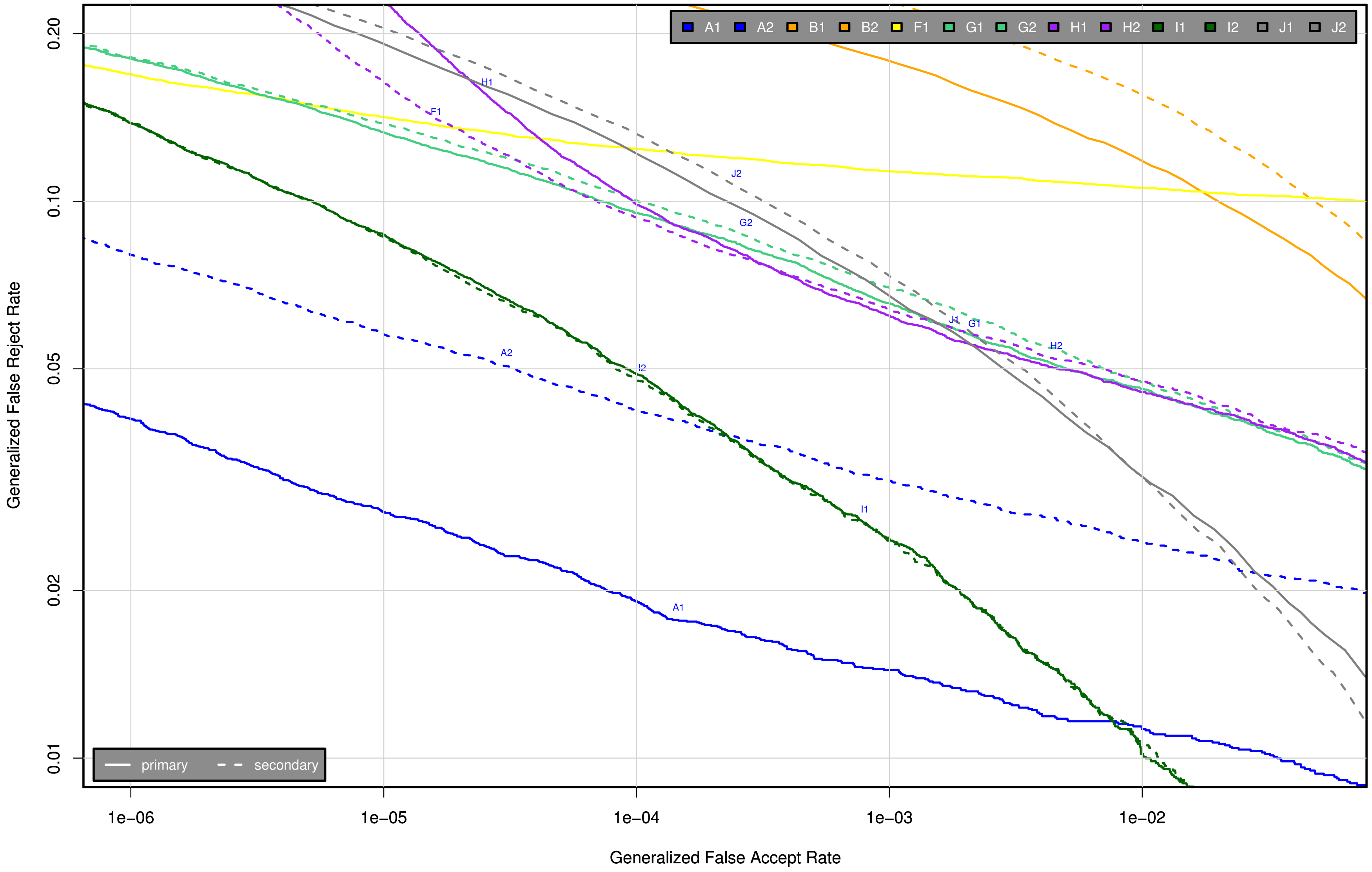


Figure 26: The DET characteristics for all IREX algorithms on the OPS dataset for comparison of 2000 byte JPEG2000 compressed KIND 16 images with raw KIND 1 images. Each SDK failed to produce templates from a fraction, $0 \leq \text{FTE} \leq 1$, of the images and the effects of that are included in this plot. The FTE rates appear in Table 6. For the OPS dataset this fraction was often zero (see Table 6).

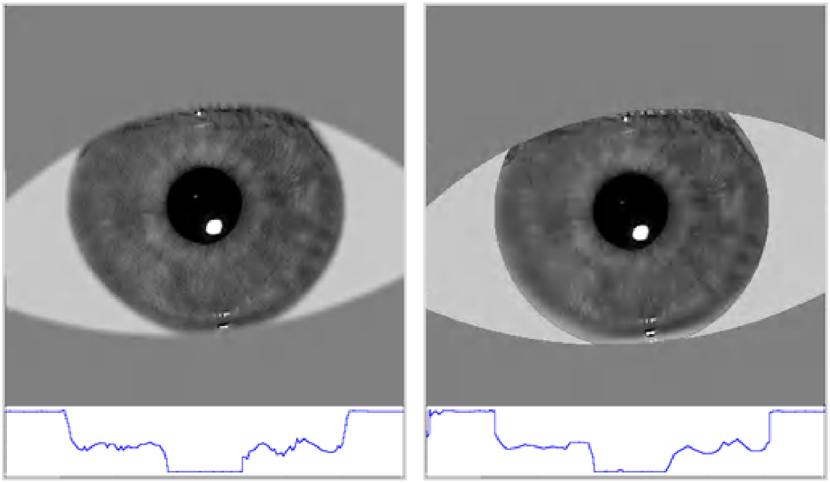


Figure 27: The figure shows the result of preparing the KIND 7 record with (left) and without (right) the mask boundary blurring operation. The images have been compressed to 2000 bytes using the JPEG2000 compressor and then reconstructed. The scan lines below each image reveal more detail of the iris texture is retained in the blurred edge image. This effect was part of the original Cambridge proposal for KIND 7 which found that blurring preserves the coding budget.

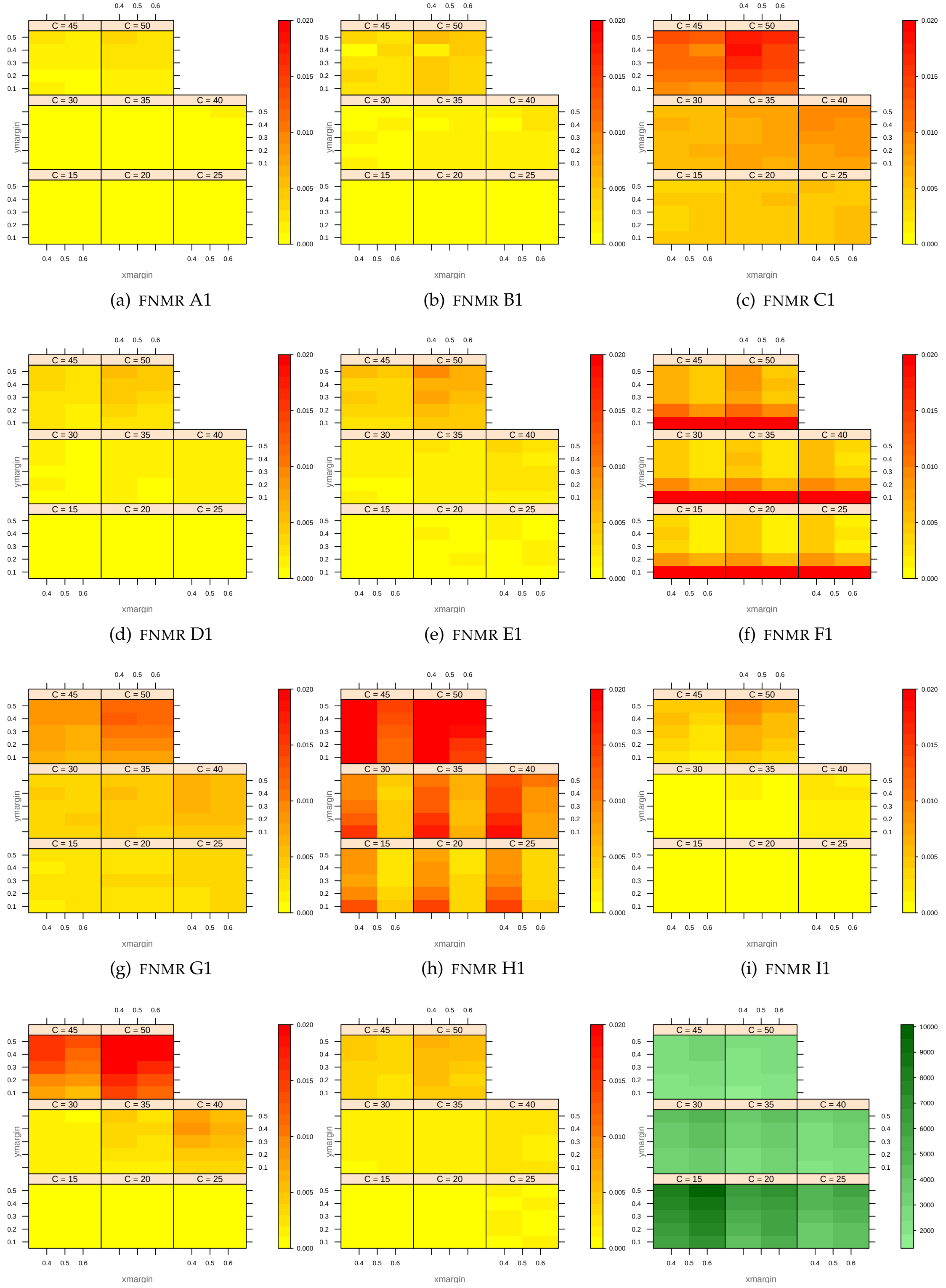
8.4. Variation of image records between SDKs

In preparing the cropped KIND 3, cropped-and-masked KIND 7, and unsegmented polar KIND 16 instances, the SDKs use a diversity of algorithms to localize the iris and eyelids. In doing so, they introduce some variation in the output image[43] as described here:

- ▷ **Crop-only** KIND **3:** Some SDKs crop more aggressively. Particularly the D1, D2, E1 and E2 SDKs leave a smaller margin above and below the iris than the other SDKs . This is discussed extensively in section 8.5 which quantifies the effect on matching performance.
- ▷ **Crop+mask** KIND **7:** As with KIND 3 there is also variation in the the amount of cropping used in the preparation of the KIND 7 instances. Also the aggressiveness with which eye lashes are masked (whether or not they're over the iris texture), and whether the lower eyelid was confused as being sclera and masked as such. A further variation is whether the SDK actually implemented the IREX test plan recommendation[44] to blur the masked edges. Figure 27 shows images with and without this blurring operation. The blur is intended to spare the coding budget.
- ▷ **Unsegmented polar** KIND **16:** The dominant source of variation across SDKs is the choice of radial and circumferential sampling rates. There's considerable disparity here between implementations: A1, A2, G1, G2, J1 and J2 all use 64 radial and 256 circumferential samples while I1 and I2 use roughly double (but variable) sampling rates. H1 and H2 were unique in using much larger circumferential rates. In addition, the degree to which the inner circle is smaller than the pupillary boundary, and the outer circle is beyond the limbic boundary, adds some variation.

[43] Whether the resulting image conforms to the text of the formal standard would be the subject of so-called Level 3 conformance test assertions. In its development of ISO/IEC 29109 - Conformance tests for biometric data interchange records, SC 37 Working Group 3 established three levels of conformance tests: Level 1 for syntax, Level 2 for internal consistency, and Level 3 for semantic image or signal properties (e.g., a frontal face is truly frontal).

[44] IREX followed the L1 suggestion to convolve with a binomial kernel $\mathbf{u} = (1, 6, 15, 20, 15, 6, 1)^T$. This gave an explicit value for the mask referred to in the original Cambridge definition of the KIND 7 format[19, 18].



(a) FNMR A1
(b) FNMR B1
(c) FNMR C1
(d) FNMR D1
(e) FNMR E1
(f) FNMR F1
(g) FNMR G1
(h) FNMR H1
(i) FNMR I1
(j) FNMR J1
(k) FNMR A2
(l) Image size (bytes)

Figure 28: For the entire partition of the OPS dataset the plots show the dependence of cFNMR on the vertical and horizontal iris cropping margins for various compression ratios. This applies only for KIND 3 records. The use of conditional FNMR excludes comparisons that were falsely rejected even before any compression was applied. All computations are driven by the bounding box coordinates reported by the I1 SDK. The native KIND 3 instances are **not** used in this analysis. The number of bits per pixel is $8/C$, where C is the compression ratio. The iris radius varies and, because the cropping margins are fixed multiples thereof, the image size varies. The last plot shows the compressed size, in bytes. This equals the width times height divided by C. The SDK-specific annexes include these figures in higher resolution alongside analogous display of the RMS change in genuine score.

8.5. The effect of cropping

Generation of the centered iris of the compact KIND 3 record requires cropping of the parent KIND 1 image in both the horizontal and vertical directions. The resulting iris is centered in a raster. It should have sufficient margin around it to allow iris detection algorithms to successfully fit the eyelid during segmentation. After consultation with prospective participants, the IREX test plan recommended that the crop operation should leave a margin $(\Delta x, \Delta y)$ around the iris where

$$\Delta x = \Delta y \geq \max(40, 0.6R) \tag{16}$$

There was lack of consensus on this matter and the recommendation was not adopted by some IREX implementations. Particularly implementations D1, D2, E1 and E2 elected to crop more aggressively in the vertical direction. This approach has the advantage of removing eye lashes.

This issue is important because, under compression, a smaller (more tightly cropped) image can be compressed at lower bit rates to achieve a size objective. Particularly, for given a compression ratio C, an iris of radius R will be compressed to a size S bytes where

$$S = \frac{4(R + \Delta x)(R + \Delta y)}{C} \tag{17}$$

and the number of bits per pixel B is

$$B = \frac{8}{C} \tag{18}$$

Thus for the KIND 3 crop-only formats the recognition performance will be a function of three independent variables: the vertical and horizontal cropping margins and the compression ratio. A user of the KIND 3 format can specify these and accept a variable size S:

$$S = \frac{4\gamma_x \gamma_y R}{C} \tag{19}$$

where the margins are now expressed as a multiple of the image radius i.e. $\Delta x = \gamma_x R$ and similarly for y.

Figure 28 gives the results of a survey over these parameters for the various SDKs. Each plot shows the conditional false non-match rate, cFNMR, for the entire OPS dataset as a heat map. It is plotted as a function of $\Delta x \Delta y$ at compression ratios of $C \in \{15, 20, 25, 30, 35, 40, 45, 50\}$ corresponding to bit rates $B \in \{0.53, 0.4, 0.32, 0.27, 0.23, 0.2, 0.18, 0.16\}$ respectively. The horizontal cropping margin takes on two values, $0.4R$ and $0.6R$. The vertical margin varies from $0.1R$ to 0.5 in steps of $0.1R$. The values currently adopted (July 2009) for the ISO/IEC 19794-6 are $\Delta x = 0.6R$ and $\Delta y = 0.2R$ so that image width and height are $w = 2(R + \Delta x)$ and $h = 2(R + \Delta y)$ respectively.

The notable results from the figures are as follows.

- ▷ **Horizontal margin:** Restriction to a horizontal margin of $0.4R$ is injurious in almost all cases. For the F1 and H1 SDK this is true even when compression is light (small C). For other SDKs the use of a narrow image only becomes problematic at high compression ratios. While SDK A1 is most tolerant of tighter horizontal cropping, enough false reject errors are observed at $\Delta x = 0.4R$ that the standard's requirement to use $\Delta x = 0.6R$ should be maintained.

- ▷ **Eye lash effects:** In the vertical direction the use of larger margins is injurious to FNMR. This would be a counter-intuitive result because retention of more of the image should allow better performance. Note that within any 10-element panel, the number of bits per pixel is fixed because it is solely dependent on C. The effect is explained by realizing that retention of the eye lashes in the image requires the JPEG 2000 compressor to dedicate more of its coding budget to the worthless high frequency information associated with eye lashes. The effect is particularly evident here because eye lashes in the OPS dataset are quite prominent; this is a property of the subject population.

- ▷ **Vertical margin:** This issue of how small the vertical margin can be is more straightforward. For all SDKs except

H1 and F1 the accuracy is best at a margin of 1.1 and virtually identical to that at 1.2. However, given the severity of the problem for some algorithms, and the possible difficulty for arbitrary providers to detect the iris boundary to within 10% of its radius with such a tight crop, the newly (July 2009) instituted ISO standard specification of $\Delta y = 0.2R$ should be maintained.

- **Algorithmic resistance to compression:** The matching algorithms exhibit different resistance to the iris texture-damaging effects of compression. For any given margin pair $(\Delta x, \Delta y)$, SDKs A1 and D1 give least growth in the error rate. Note that the quantity plotted is the *conditional* FNMR defined by equation 15. This measure is appropriate to represent compression of enrollment quality instances.

- **Attainable file sizes:** A more important result is that, for seven out of ten providers, compression at 25:1 (i.e. bits per pixel = 0.32) gives essentially undetectable increase in FNMR and this affords *mean* compressed raster sizes of fewer than 5 kilobytes[45]. This supports one of the main IREX conclusions, that cropping and JPEG 2000 compression to about 6 kilobytes is virtually error free.

8.6. INTEROPERABILITY OF THE STANDARD FORMATS

The power of formally standardizing iris image interchange records is that it supports a marketplace for products that either generate standard records or consume (match) them. Thus, for example, a product from provider X can be used to produce enrollment records which are then stored or transmitted and subsequently used in an identification search executed by provider Y. This cross-provider aspect requires *interoperability* and this necessitates that all marketplace providers have a uniform understanding and implementation of the relevant standard. NIST's status as an independent organization is well suited to cross-provider interoperability testing because single providers often do not have access to SDKs and biometric samples prepared by competing providers. In executing tests like IREX and MINEX[25][46], NIST is essentially a technical broker.

The IREX cross-provider interoperability tests have quadratic complexity reflecting all combinations of enroller X and matcher Y[47]. Interoperability is stated as the FNMR rate at a FMR = 0.0001. The threshold was set for each interoperable combination (X,Y) and this necessitated computation of the impostor distribution. This in turn necessitated use of the smaller (but more difficult) ALL-FAILURES partition of the OPS dataset. The result is nevertheless a computationally expensive operation involving nearly 8 billion comparisons[48] for each KIND.

8.6.1. TEST METHOD

The **enrollment source images** were the 1335 KIND 1 images that constitute the enrollments image of the ALL-FAILURES partition of the OPS dataset (see section 5.1.1). Each SDK was used to convert these to **enrollment** IREX **records** in each of the other IREX formats. While six SDKs did not support generation of KIND 16 records (see Table 4), they were required to generate templates and match them.

All 16320 **verification images** of 8160 subjects in the OPS dataset were converted by NIST into KIND 1 records - this operation is trivial and does not require localization.

JPEG2000 **compression** was applied in two separate ways. First, each enrollment IREX record was compressed to a target size of 3000 bytes. This differs from the 2000 bytes used elsewhere in this report. Second, the compressor was invoked

[45] For cropping margins of $\Delta x = 0.6R$ and $\Delta y = 0.6R$ the statistics on the number of bytes are: min = 2772, 25-th percentile = 4137, mean = 4501, median = 4574, 75-th percentile = 4937 and maximum = 7691.

[46] MINEX supports fingerprint minutia interoperability - see http://fingerprint.nist.gov/minex

[47] NIST did not consider the cubic case in which an enrollment record prepared by X and a verification sample prepared by Y are compared by provider Z.

[48] The number is 19 x 19 x 1335 x 16320 comparisons.

to produce a compression ratio of 40, corresponding to 0.2 bits per pixel. This mode produces a file size that depends on the initial size of IREX record's image.

Each SDK was used to convert each enrollment IREX record into a proprietary template. The SDK was applied to convert each verification IREX record into a proprietary template. The SDK was used to **match** all pairs of templates to produce N = 1335 x 16320 = 21,787,200 scores. This produces 1335 genuine scores and 21,784,530 impostor scores for each interoperating combination of enroler (X) and matcher (Y).

The comparison scores from each enroler-matcher pair are subject to the usual analyses, equations 5 and 3. The value τ that gives FMR (τ) is computed from the empirical distribution function of the impostor distribution and FNMR (τ) is computed from the genuine distribution.

8.6.2. Interoperability results

The results are tabulated in six tables, three for FNMR (Tables 12, 13 and 14) and three for FMR (Tables 15, 16 and 17), with each group of three referring to the three KINDS. The FNMR tables also include the proportion of enrollment images that could not be converted to templates by the matching SDK - this is the FTE rate. While the proportion of the uncompressed KIND 1 verification images that could not be converted to templates is not reported explicitly, the effect is included in the reported FNMR value. The notable observations are:

- The interoperability matrices show two performance aspects: IREX record production and IREX record consumption (matching). Thus, across all tables, the best overall producer (A1) is not the best matcher (I2). This is expected because the algorithmic functions needed to instantiate records are distinct from those needed to extract features and match.

- ANOVA analyses show the matcher to be more influential on accuracy measures than the record generator. This, however, is most true for KIND 3 but less so for KIND 7 and hardly at all for KIND 16. This suggests that the fine segmentation part of the template creation process is harder in KIND 3 images than in the others.

- From the "wins" column of Tables 12 and 13, the KIND 3 and KIND 7 instances from B1 and A1 tend to be preferred by any given matcher, i.e. to give the lowest error rates. That said the A2, B2, F1, I2, I2, and J2 generators also give mean FNMR values close to B1 and A1 (i.e FNMR < 0.08).

- From the "wins" row of Tables 12, 13 and 14 the I1 and I2 matchers are the most capable of matching IREX records regardless of their source. The false non-match rates are less than half of other matchers. The presence of green shading in the respective columns indicates these SDKs give lower error rates than the producers of the enrollment IREX records. A1, B1 and J2 also give lower mean error rates (FNMR < 0.07).

- Some matchers prefer other producers' IREX records over their own. This is indicated by red text in a column. For example, with KIND 3 records, the E2 matcher gives FNMR of 0.069 with A1 enrollments, versus 0.083 with its own. One possible explanation is better localization.

- The interoperability matrices of Tables 15, 16 and 17 show FMR when the threshold is fixed to the uncompressed KIND 1 default. The low variance observed in most columns shows that FMR has little dependence on which SDK prepared the IREX record. This is a very attractive property. Some SDKs for KIND 16 polar records do exhibit some dependency on source (e.g., row G2, column E1 in Table 17).

- The B1 and B2 implementations always produce low FMR under compression. This applies for all KINDS. A better metric would be the deviation from the nominal value. The "win" row and column refer to the worst case change in FMR from its nominal value of 0.0001. In this respect C1 excels for KIND 3 and KIND 16, while I2 is best for KIND 7.

8.6.3. COMPARING KINDS 3 AND 7

Comparing the interoperability matrices of Tables 12 and 13 there is little difference in the observed error rates. KIND 7 is expected to give fewer false rejection errors because the mask is highly compressible, and this preserves the iris texture for a given compression level. However, KIND 7 is often *less* accurate than KIND 3. Table 18 gives the simple difference of the FNMR (k3) - FNMR (k7) for each element of the interoperability matrix. This is done for two cases: enrollment images compressed to 3000 bytes, and enrollment images compressed to 0.2 bits per pixel (i.e. 40:1 compression ratio). The notable results are as follows:

- ▷ Cells are shaded green when the KIND 7 gives lower FNMR than KIND 3. The relative absence of green shading indicates that KIND 7 is often giving inferior false non-match rates than KIND 3. KIND 3 is superior at the lighter compression level (the top table, bits per pixel = 0.02) but its advantage is partially eroded with more compression (size = 3000 bytes).
- ▷ The magnitude of the difference is such that KIND 3 is often much better than KIND 7, and that KIND 7 is only ever marginally better than KIND 3. Again KIND 7 improves under stronger compression.
- ▷ Looking at the rows of the matrices in Table 18, some SDKs (G1, G2, D1, D2, E1, E2) make KIND 3 records that are almost always preferred by the matchers. Similarly, looking at the columns (C1, D1, D2, E1, and at, bpp = 0.2, B1, B2, H1, H2, I1, I2) the matchers prefer KIND 3 to KIND 7 .

These statements are consistent with two competing effects: The immunity of KIND 7 to severe compression is offset by errors in the eyelid detection operation which do not affect KIND 3. Masking errors result if the eyelid-iris and sclera-iris boundaries are difficult to find. These would cause elevated FNMR. Thus, at some compression level, where too much damage is done to the texture of a KIND 3 raster, KIND 7 accuracy begins to exceed that of KIND 3. Some implementations are better than others, so the cross-over point is variable. While a further, dedicated, experiment is needed to solidify this conclusion, it does weigh against KIND 7 in the *which-format-to-use* guidance given in section 10.

KIND 3	INTEROPERABLE FALSE NON-MATCH RATES																			Row	Num	Size
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Ave	Win	[5, 95] %
A1	0.053 0.000	0.219 0.000	0.054 0.000	0.130 0.000	0.379 0.010	- -	0.235 0.004	0.148 0.005	0.075 0.013	0.069 0.013	0.410 0.008	0.099 0.005	0.104 0.011	0.405 0.010	0.390 0.010	0.022 0.000	0.021 0.000	0.133 0.000	0.055 0.006	0.076 0.002	6	2673 - 3667
A2	0.053 0.000	0.219 0.000	0.054 0.000	0.130 0.000	0.379 0.010	- -	0.235 0.004	0.148 0.005	0.075 0.013	0.069 0.013	0.410 0.008	0.099 0.005	0.104 0.011	0.405 0.010	0.390 0.010	0.022 0.000	0.021 0.000	0.133 0.000	0.055 0.006	0.076 0.002	0	2673 - 3667
B1	0.058 0.000	0.223 0.000	0.055 0.000	0.132 0.000	0.375 0.004	0.400 0.004	0.229 0.004	0.141 0.004	0.078 0.015	0.072 0.015	0.412 0.006	0.097 0.007	0.104 0.012	0.402 0.006	0.386 0.006	0.023 0.000	0.022 0.000	0.136 0.004	0.054 0.004	0.077 0.002	8	2788 - 3920
B2	0.058 0.000	0.223 0.000	0.055 0.000	0.132 0.000	0.375 0.004	0.400 0.004	0.229 0.004	0.141 0.004	0.078 0.015	0.072 0.015	0.412 0.006	0.097 0.007	0.104 0.012	0.402 0.006	0.386 0.006	0.023 0.000	0.022 0.000	0.136 0.004	0.054 0.004	0.077 0.002	0	2788 - 3920
C1	0.075 0.008	0.240 0.008	0.064 0.002	0.141 0.002	0.400 0.041	0.426 0.044	0.251 0.013	0.159 0.015	0.095 0.022	0.089 0.022	0.422 0.019	0.113 0.025	0.120 0.026	0.434 0.013	0.413 0.013	0.034 0.004	0.035 0.004	0.145 0.011	0.062 0.011	0.092 0.009	0	
C2	0.088 0.019	0.253 0.019	0.073 0.012	0.150 0.012	0.424 0.055	0.443 0.057	0.260 0.025	0.171 0.025	0.100 0.034	0.094 0.034	0.438 0.033	0.119 0.031	0.127 0.036	0.451 0.025	0.432 0.025	0.040 0.013	0.038 0.013	0.156 0.020	0.073 0.020	0.100 0.020	0	2755 - 3920
D1	0.115 0.000	0.280 0.000	0.062 0.000	0.143 0.000	0.381 0.004	0.417 0.007	0.265 0.004	0.178 0.006	0.120 0.014	0.111 0.014	0.545 0.128	0.108 0.011	0.109 0.018	0.483 0.001	0.480 0.001	0.027 0.000	0.080 0.000	0.136 0.002	0.055 0.002	0.103 0.001	0	2029 - 2983
D2	0.115 0.000	0.280 0.000	0.062 0.000	0.142 0.000	0.381 0.004	0.417 0.007	0.265 0.004	0.178 0.006	0.120 0.014	0.111 0.014	0.545 0.128	0.108 0.011	0.109 0.018	0.483 0.001	0.480 0.001	0.027 0.000	0.080 0.000	0.136 0.002	0.055 0.002	0.103 0.001	0	2029 - 2983
E1	0.075 0.004	0.239 0.004	0.058 0.004	0.136 0.004	0.375 0.010	0.425 0.019	0.235 0.005	0.150 0.006	0.091 0.013	0.083 0.013	0.502 0.094	0.105 0.010	0.109 0.013	0.409 0.006	0.392 0.006	0.025 0.004	0.040 0.004	0.137 0.006	0.055 0.006	0.086 0.005	0	2029 - 2983
E2	0.075 0.004	0.239 0.004	0.058 0.004	0.136 0.004	0.375 0.010	0.425 0.019	0.235 0.005	0.150 0.006	0.091 0.013	0.083 0.013	0.502 0.094	0.105 0.010	0.109 0.013	0.409 0.006	0.392 0.006	0.025 0.004	0.040 0.004	0.137 0.006	0.055 0.006	0.086 0.005	0	2029 - 2983
F1	0.061 0.007	0.226 0.007	0.055 0.002	0.131 0.002	0.401 0.043	0.428 0.040	0.238 0.012	0.148 0.012	0.078 0.019	0.073 0.019	0.404 0.004	0.105 0.013	0.106 0.019	0.414 0.016	0.394 0.016	0.024 0.002	0.025 0.002	0.135 0.007	0.055 0.007	0.079 0.006	1	2822 - 3920
G1	0.066 0.004	0.231 0.004	0.062 0.004	0.142 0.004	0.390 0.017	0.413 0.020	0.239 0.008	0.151 0.008	0.084 0.021	0.076 0.021	0.420 0.018	0.107 0.008	0.107 0.017	0.425 0.010	0.409 0.010	0.031 0.004	0.031 0.004	0.139 0.007	0.058 0.007	0.085 0.006	0	2624 - 3610
G2	0.064 0.008	0.228 0.008	0.061 0.008	0.138 0.008	0.383 0.022	0.409 0.021	0.246 0.011	0.155 0.012	0.080 0.019	0.076 0.019	0.420 0.018	0.109 0.014	0.106 0.013	0.437 0.014	0.417 0.014	0.032 0.008	0.031 0.008	0.138 0.011	0.058 0.011	0.084 0.011	0	2656 - 3648
H1	0.163 0.021	0.324 0.021	0.146 0.021	0.225 0.021	0.453 0.065	0.488 0.064	0.336 0.037	0.231 0.051	0.176 0.073	0.170 0.073	0.518 0.103	0.196 0.082	0.193 0.082	1.000 0.021	0.712 0.021	0.112 0.024	0.118 0.022	0.898 0.192	0.140 0.042	0.182 0.028	0	2958 - 4243
H2	0.163 0.021	0.324 0.021	0.146 0.021	0.225 0.021	0.453 0.065	0.488 0.064	0.336 0.037	0.231 0.051	0.176 0.073	0.170 0.073	0.518 0.103	0.196 0.082	0.193 0.082	1.000 0.021	0.712 0.021	0.112 0.024	0.118 0.022	0.898 0.192	0.140 0.042	0.182 0.028	0	2958 - 4243
I1	0.058 0.006	0.220 0.006	0.053 0.000	0.131 0.000	0.396 0.040	0.426 0.039	0.235 0.010	0.145 0.010	0.077 0.017	0.072 0.017	0.407 0.006	0.104 0.014	0.108 0.016	0.413 0.012	0.393 0.012	0.022 0.000	0.022 0.000	0.132 0.000	0.055 0.006	0.078 0.004	2	2890 - 4000
I2	0.059 0.006	0.221 0.006	0.055 0.000	0.133 0.000	0.397 0.040	0.426 0.038	0.236 0.010	0.147 0.010	0.076 0.016	0.071 0.016	0.404 0.004	0.105 0.015	0.107 0.016	0.414 0.012	0.394 0.012	0.022 0.000	0.021 0.000	0.133 0.000	0.055 0.006	0.078 0.004	1	2890 - 4000
J1	0.058 0.007	0.221 0.007	0.060 0.007	0.134 0.007	0.409 0.041	- -	0.236 0.012	0.145 0.013	0.079 0.021	0.073 0.021	0.414 0.012	0.115 0.019	0.109 0.021	0.954 0.062	0.948 0.062	0.028 0.007	0.028 0.007	0.134 0.007	0.055 0.007	0.082 0.009	0	2805 - 3980
J2	0.056 0.004	0.219 0.004	0.057 0.004	0.132 0.004	0.396 0.035	- -	0.234 0.009	0.143 0.010	0.078 0.019	0.071 0.019	0.413 0.009	0.102 0.013	0.103 0.016	0.413 0.014	0.393 0.014	0.025 0.004	0.025 0.004	0.133 0.004	0.055 0.007	0.078 0.006	1	2805 - 3980
Ave	0.064	0.247	0.061	0.147	0.421	0.462	0.260	0.162	0.086	0.080	0.456	0.113	0.117	0.454	0.434	0.026	0.026	0.149	0.061			
Wins	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	10	0	0			

Table 12: Cross-provider FNMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 3 records were compressed to 0.2 bits per pixel (CR = 40) using JPEG 2000. The row label identifies the producer of the IREX enrollment KIND 3 record. The column label identifies the SDK that compares (proprietary templates of) the KIND 3 enrollment record against a KIND 1 verification record. **Error rates**: In each cell two numbers appear: At top is FNMR at the threshold that gives FMR = 0.0001 for that combination of algorithm. Below is FTE covering failures of the column-identified SDK to make an enrollment template from the row-identified KIND 3 instance and of the row-identified SDK to make an IREX record. The FTE is included in the FNMR . **Wins**: The next to last column gives the number of columns in which the row-identified KIND 3 generator gives the best accuracy. The last row gives the number of rows for which the matcher gives the best accuracy. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells with red text indicate better accuracy than the matcher's native performance. Cells filled green indicate better accuracy than when the producer of the enrollment record executes the match. All means are taken over the best ten elements - as a robustness measure to remove poor performances.

KIND 7	INTEROPERABLE FALSE NON-MATCH RATES																			Row	Num	Size
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Ave	Win	[5, 95] %
A1	0.052 0.000	0.216 0.000	0.054 0.000	0.133 0.000	0.379 0.006	0.390 0.007	0.239 0.004	0.148 0.004	0.079 0.011	0.067 0.011	0.403 0.008	0.095 0.009	0.097 0.015	0.420 0.008	0.397 0.008	0.023 0.000	0.023 0.000	0.130 0.000	0.058 0.004	0.076 0.001	8	2673 - 3667
A2	0.052 0.000	0.216 0.000	0.054 0.000	0.133 0.000	0.379 0.006	0.390 0.007	0.239 0.004	0.148 0.004	0.079 0.011	0.067 0.011	0.403 0.008	0.095 0.009	0.097 0.015	0.420 0.008	0.397 0.008	0.023 0.000	0.023 0.000	0.130 0.000	0.058 0.004	0.076 0.001	0	2673 - 3667
B1	0.055 0.000	0.221 0.000	0.055 0.000	0.131 0.000	0.374 0.003	0.393 0.003	0.234 0.004	0.142 0.005	0.079 0.015	0.072 0.015	0.399 0.006	0.097 0.010	0.094 0.012	0.414 0.006	0.401 0.006	0.025 0.000	0.024 0.000	0.133 0.002	0.052 0.002	0.076 0.001	6	2788 - 3920
B2	0.055 0.000	0.221 0.000	0.055 0.000	0.131 0.000	0.374 0.003	0.393 0.003	0.234 0.004	0.142 0.005	0.079 0.015	0.072 0.015	0.399 0.006	0.097 0.010	0.094 0.012	0.414 0.006	0.401 0.006	0.025 0.000	0.024 0.000	0.133 0.002	0.052 0.002	0.076 0.001	0	2788 - 3920
C1	0.082 0.008	0.252 0.008	0.071 0.002	0.153 0.002	0.390 0.019	0.411 0.019	0.263 0.011	0.176 0.013	0.100 0.026	0.093 0.026	0.428 0.010	0.109 0.019	0.109 0.020	0.443 0.007	0.427 0.007	0.041 0.004	0.041 0.003	0.148 0.017	0.069 0.017	0.096 0.007	0	
C2	0.097 0.019	0.268 0.019	0.081 0.012	0.163 0.012	0.408 0.028	0.425 0.028	0.276 0.022	0.191 0.022	0.115 0.035	0.102 0.035	0.448 0.022	0.128 0.033	0.123 0.032	0.494 0.019	0.460 0.019	0.058 0.013	0.059 0.013	0.161 0.029	0.084 0.029	0.112 0.019	0	2755 - 3920
D1	0.145 0.043	0.299 0.043	0.079 0.000	0.195 0.000	0.400 0.007	0.419 0.007	0.270 0.004	0.182 0.010	0.138 0.014	0.127 0.014	0.467 0.025	0.142 0.034	0.135 0.044	0.515 0.000	0.510 0.000	0.072 0.001	0.090 0.005	0.156 0.011	0.073 0.011	0.128 0.004	0	2029 - 2983
D2	0.145 0.043	0.299 0.043	0.079 0.000	0.196 0.000	0.400 0.007	0.419 0.007	0.270 0.004	0.182 0.010	0.138 0.014	0.127 0.014	0.467 0.025	0.142 0.034	0.135 0.044	0.515 0.000	0.510 0.000	0.072 0.001	0.090 0.005	0.156 0.011	0.073 0.011	0.128 0.004	0	2029 - 2983
E1	0.122 0.000	0.285 0.000	0.075 0.000	0.193 0.000	0.404 0.006	0.416 0.006	0.267 0.002	0.182 0.005	0.137 0.011	0.126 0.011	0.470 0.029	0.150 0.040	0.140 0.050	0.455 0.000	0.440 0.000	0.073 0.001	0.091 0.001	0.147 0.006	0.066 0.006	0.125 0.001	0	2029 - 2983
E2	0.122 0.000	0.285 0.000	0.076 0.000	0.195 0.000	0.404 0.006	0.416 0.006	0.267 0.002	0.182 0.005	0.137 0.011	0.126 0.011	0.470 0.029	0.150 0.040	0.140 0.050	0.455 0.000	0.440 0.000	0.073 0.001	0.091 0.001	0.147 0.006	0.066 0.006	0.125 0.001	0	2029 - 2983
F1	0.065 0.007	0.233 0.007	0.061 0.002	0.147 0.002	0.426 0.045	0.433 0.045	0.258 0.010	0.170 0.011	0.086 0.020	0.078 0.020	0.414 0.004	0.110 0.016	0.102 0.019	0.425 0.010	0.416 0.010	0.030 0.003	0.028 0.004	0.144 0.007	0.060 0.007	0.085 0.006	0	2822 - 3920
G1	0.077 0.011	0.249 0.011	0.074 0.011	0.152 0.011	0.408 0.028	0.426 0.028	0.266 0.016	0.172 0.018	0.096 0.025	0.088 0.025	0.432 0.017	0.118 0.017	0.107 0.023	0.495 0.014	0.482 0.014	0.045 0.011	0.042 0.011	0.144 0.019	0.067 0.019	0.095 0.014	0	2624 - 3610
G2	0.079 0.017	0.250 0.017	0.073 0.017	0.149 0.017	0.407 0.030	0.427 0.030	0.264 0.020	0.175 0.021	0.098 0.025	0.088 0.025	0.422 0.023	0.122 0.023	0.115 0.021	0.467 0.021	0.457 0.021	0.042 0.017	0.040 0.017	0.148 0.019	0.069 0.019	0.097 0.020	0	2656 - 3648
H1	0.184 0.022	0.345 0.022	0.177 0.022	0.276 0.022	0.506 0.085	0.520 0.084	0.379 0.030	0.265 0.043	0.210 0.094	0.192 0.094	0.595 0.052	0.240 0.086	0.232 0.106	0.658 0.028	1.000 0.028	0.157 0.030	0.154 0.023	0.941 0.192	0.170 0.067	0.220 0.030	0	2958 - 4243
H2	0.184 0.022	0.345 0.022	0.177 0.022	0.276 0.022	0.506 0.085	0.520 0.084	0.379 0.030	0.265 0.043	0.210 0.094	0.192 0.094	0.595 0.052	0.240 0.086	0.232 0.106	0.658 0.028	1.000 0.028	0.157 0.030	0.154 0.023	0.941 0.192	0.170 0.067	0.220 0.030	0	2958 - 4243
I1	0.058 0.006	0.228 0.006	0.058 0.000	0.139 0.000	0.404 0.042	0.422 0.042	0.255 0.010	0.165 0.013	0.090 0.021	0.082 0.021	0.404 0.002	0.100 0.013	0.085 0.015	0.407 0.009	0.396 0.009	0.023 0.000	0.022 0.000	0.133 0.000	0.053 0.005	0.078 0.003	2	2890 - 4000
I2	0.058 0.006	0.228 0.006	0.058 0.000	0.139 0.000	0.404 0.043	0.422 0.043	0.256 0.010	0.166 0.013	0.091 0.022	0.083 0.022	0.403 0.003	0.101 0.013	0.086 0.015	0.407 0.009	0.397 0.009	0.023 0.000	0.022 0.000	0.132 0.000	0.055 0.005	0.079 0.003	0	2890 - 4000
J1	0.060 0.007	0.228 0.007	0.058 0.007	0.136 0.007	0.380 0.012	0.402 0.011	0.237 0.012	0.148 0.013	0.079 0.021	0.073 0.021	0.406 0.012	0.115 0.017	0.109 0.022	0.968 0.060	0.951 0.060	0.030 0.007	0.028 0.007	0.135 0.007	0.056 0.007	0.083 0.009	0	2805 - 3980
J2	0.058 0.004	0.224 0.004	0.056 0.004	0.134 0.004	0.375 0.005	0.393 0.005	0.237 0.008	0.147 0.010	0.078 0.018	0.070 0.018	0.407 0.010	0.103 0.013	0.101 0.014	0.401 0.008	0.384 0.008	0.028 0.004	0.027 0.004	0.133 0.004	0.055 0.007	0.079 0.005	3	2805 - 3980
Ave	0.066	0.252	0.064	0.152	0.428	0.447	0.273	0.172	0.093	0.084	0.451	0.114	0.108	0.467	0.451	0.030	0.029	0.150	0.063			
Wins	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6	11	0	2			

Table 13: Cross-provider FNMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 7 records were compressed to 0.2 bits per pixel (CR = 40) using JPEG 2000. The row label identifies the producer of the IREX enrollment KIND 7 record. The column label identifies the SDK that compares (proprietary templates of) the KIND 7 enrollment record against a KIND 1 verification record. **Error rates**: In each cell two numbers appear: At top is FNMR at the threshold that gives FMR = 0.0001 for that combination of algorithm. Below is FTE covering failures of the column-identified SDK to make an enrollment template from the row-identified KIND 7 instance and of the row-identified SDK to make an IREX record. The FTE is included in the FNMR . **Wins**: The next to last column gives the number of columns in which the row-identified KIND 7 generator gives the best accuracy. The last row gives the number of rows for which the matcher gives the best accuracy. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells with red text indicate better accuracy than the matcher's native performance. Cells filled green indicate better accuracy than when the producer of the enrollment record executes the match. All means are taken over the best ten elements - as a robustness measure to remove poor performances.

KIND 16	INTEROPERABLE FALSE NON-MATCH RATES																			Row	Num	BPP
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Ave	Win	[5, 95] %
A1	0.064 0.000	0.231 0.000	0.182 0.000	0.309 0.000	0.360 0.007	0.376 0.007	0.286 0.001	0.181 0.004	0.127 0.025	0.094 0.025	0.396 0.002	0.103 0.010	0.100 0.018	0.343 0.000	0.336 0.000	0.028 0.000	0.027 0.000	0.133 0.005	0.056 0.005	0.102 0.000	17	0.16 - 0.22
A2	0.064 0.000	0.231 0.000	0.183 0.000	0.309 0.000	0.360 0.007	0.376 0.007	0.286 0.001	0.181 0.004	0.127 0.025	0.094 0.025	0.396 0.002	0.103 0.010	0.100 0.018	0.343 0.000	0.336 0.000	0.028 0.000	0.027 0.000	0.133 0.005	0.056 0.005	0.102 0.000	0	0.16 - 0.22
B1	0.144 0.000	0.354 0.000	0.244 0.000	0.355 0.000	0.845 0.231	0.847 0.231	0.476 0.001	0.351 0.003	0.316 0.029	0.224 0.029	0.710 0.013	0.270 0.018	0.265 0.013	0.709 0.000	0.716 0.000	0.042 0.000	0.040 0.000	0.249 0.002	0.164 0.002	0.218 0.000	0	0.15 - 0.22
B2	0.144 0.000	0.354 0.000	0.244 0.000	0.355 0.000	0.845 0.231	0.847 0.231	0.476 0.001	0.351 0.003	0.316 0.029	0.224 0.029	0.710 0.013	0.270 0.018	0.265 0.013	0.709 0.000	0.716 0.000	0.042 0.000	0.040 0.000	0.249 0.002	0.164 0.002	0.218 0.000	0	0.15 - 0.22
F1	0.110 0.002	0.277 0.002	0.333 0.002	0.467 0.002	0.484 0.022	0.487 0.022	0.453 0.004	0.279 0.007	0.285 0.075	0.198 0.075	0.407 0.004	0.188 0.025	0.192 0.033	0.420 0.002	0.418 0.002	0.031 0.002	0.030 0.002	0.167 0.008	0.081 0.008	0.172 0.003	0	0.15 - 0.21
G1	0.104 0.006	0.270 0.006	0.277 0.006	0.428 0.006	0.428 0.013	0.440 0.013	0.335 0.012	0.213 0.013	0.324 0.018	0.176 0.018	0.414 0.008	0.142 0.019	0.145 0.016	0.392 0.006	0.392 0.006	0.046 0.006	0.046 0.006	0.180 0.014	0.088 0.014	0.157 0.008	0	0.17 - 0.23
G2	0.106 0.010	0.272 0.010	0.292 0.010	0.429 0.010	0.439 0.016	0.450 0.016	0.369 0.014	0.231 0.016	0.538 0.019	0.425 0.019	0.422 0.010	0.157 0.025	0.154 0.024	0.399 0.010	0.402 0.010	0.044 0.010	0.043 0.010	0.183 0.012	0.094 0.012	0.175 0.011	0	0.16 - 0.23
H1	0.222 0.021	0.393 0.021	0.407 0.023	0.518 0.023	0.610 0.100	0.623 0.100	0.531 0.079	0.369 0.081	0.390 0.115	0.293 0.115	0.543 0.072	0.285 0.102	0.282 0.099	0.580 0.024	0.583 0.024	0.144 0.063	0.148 0.058	0.936 0.192	0.195 0.076	0.302 0.045	0	0.14 - 0.20
H2	0.222 0.021	0.393 0.021	0.407 0.023	0.518 0.023	0.610 0.100	0.623 0.100	0.531 0.079	0.369 0.081	0.390 0.115	0.293 0.115	0.543 0.072	0.285 0.102	0.282 0.099	0.580 0.024	0.583 0.024	0.144 0.063	0.148 0.058	0.936 0.192	0.195 0.076	0.302 0.045	0	0.14 - 0.20
I1	0.114 0.000	0.281 0.000	0.318 0.000	0.449 0.000	0.521 0.030	0.527 0.030	0.419 0.002	0.273 0.004	0.260 0.038	0.175 0.038	0.407 0.003	0.184 0.014	0.181 0.009	0.416 0.000	0.414 0.000	0.031 0.000	0.031 0.000	0.165 0.000	0.084 0.003	0.166 0.000	0	0.15 - 0.21
I2	0.115 0.000	0.280 0.000	0.319 0.000	0.451 0.000	0.523 0.031	0.528 0.031	0.418 0.002	0.276 0.004	0.263 0.038	0.175 0.038	0.408 0.002	0.183 0.015	0.183 0.009	0.418 0.000	0.414 0.000	0.032 0.000	0.031 0.000	0.167 0.000	0.086 0.003	0.168 0.000	0	0.15 - 0.21
J1	0.110 0.000	0.273 0.000	0.338 0.000	0.477 0.000	0.990 0.977	0.990 0.977	0.413 0.002	0.256 0.004	0.294 0.023	0.163 0.023	0.394 0.002	0.980 0.956	0.961 0.937	0.000 0.303	0.999 0.303	0.037 0.000	0.037 0.000	0.171 0.004	0.094 0.007	0.197 0.001	2	0.15 - 0.21
J2	0.113 0.000	0.271 0.000	0.339 0.000	0.479 0.000	0.450 0.006	0.455 0.006	0.434 0.004	0.255 0.005	0.254 0.028	0.154 0.028	0.395 0.004	0.172 0.014	0.169 0.015	0.405 0.000	0.414 0.000	0.038 0.000	0.037 0.000	0.167 0.001	0.091 0.004	0.161 0.001	0	0.15 - 0.21
Ave	0.116	0.305	0.303	0.448	0.532	0.543	0.432	0.277	0.285	0.186	0.465	0.197	0.195	0.477	0.477	0.039	0.038	0.191	0.099			
Wins	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	10	0	0			

Table 14: Cross-provider FNMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 16 records were compressed to a fixed target of 3000 bytes using JPEG 2000. The row label identifies the producer of the IREX enrollment KIND 16 record. The column label identifies the SDK that compares (proprietary templates of) the KIND 16 enrollment record against a KIND 1 verification record. **Error rates**: In each cell two numbers appear: At top is FNMR at the threshold that gives FMR = 0.0001 for that combination of algorithm. Below is FTE covering failures of the column-identified SDK to make an enrollment template from the row-identified KIND 16 instance and of the row-identified SDK to make an IREX record. The FTE is included in the FNMR . **Wins**: The next to last column gives the number of columns in which the row-identified KIND 16 generator gives the best accuracy. The last row gives the number of rows for which the matcher gives the best accuracy. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells with red text indicate better accuracy than the matcher's native performance. Cells filled green indicate better accuracy than when the producer of the enrollment record executes the match. All means are taken over the best ten elements - as a robustness measure to remove poor performances.

KIND 3	INTEROPERABLE FALSE MATCH RATES																			Row	Num	Size
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Worst	Win	[5, 95] %
A1	0.00013	0.00013	0.00004	0.00001	0.00010	-	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00014	0.00019	0.00012	0.00012	0.00021	0.00015	0.00011	4	2673 - 3667
A2	0.00013	0.00013	0.00004	0.00001	0.00010	-	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00014	0.00019	0.00012	0.00012	0.00021	0.00015	0.00011	0	2673 - 3667
B1	0.00015	0.00015	0.00004	0.00002	0.00010	0.00010	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00016	0.00021	0.00012	0.00012	0.00016	0.00014	0.00011	1	2788 - 3920
B2	0.00015	0.00015	0.00004	0.00002	0.00010	0.00010	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00016	0.00021	0.00012	0.00012	0.00016	0.00014	0.00011	0	2788 - 3920
C1	0.00014	0.00015	0.00004	0.00002	0.00009	0.00009	0.00011	0.00018	0.00011	0.00012	0.00006	0.00012	0.00012	0.00016	0.00020	0.00013	0.00087	0.00016	0.00014	0.00077	1	
C2	0.00014	0.00015	0.00004	0.00002	0.00009	0.00009	0.00008	0.00017	0.00011	0.00011	0.00006	0.00011	0.00012	0.00023	0.00027	0.00013	0.00013	0.00016	0.00014	0.00017	2	2755 - 3920
D1	0.00014	0.00014	0.00003	0.00001	0.00010	0.00010	0.00009	0.00021	0.00014	0.00013	0.00006	0.00012	0.00012	0.00049	0.00064	0.00013	0.00013	0.00015	0.00014	0.00054	1	2029 - 2983
D2	0.00014	0.00014	0.00003	0.00001	0.00010	0.00010	0.00009	0.00021	0.00014	0.00013	0.00006	0.00012	0.00012	0.00049	0.00064	0.00013	0.00013	0.00015	0.00014	0.00054	0	2029 - 2983
E1	0.00015	0.00015	0.00003	0.00001	0.00010	0.00010	0.00007	0.00019	0.00012	0.00012	0.00007	0.00012	0.00012	0.00023	0.00024	0.00013	0.00013	0.00016	0.00015	0.00014	0	2029 - 2983
E2	0.00015	0.00015	0.00003	0.00001	0.00010	0.00010	0.00007	0.00019	0.00012	0.00012	0.00007	0.00012	0.00012	0.00023	0.00024	0.00013	0.00013	0.00016	0.00015	0.00014	0	2029 - 2983
F1	0.00015	0.00015	0.00004	0.00001	0.00009	0.00009	0.00008	0.00017	0.00012	0.00012	0.00006	0.00012	0.00012	0.00016	0.00017	0.00012	0.00012	0.00016	0.00015	0.00007	1	2822 - 3920
G1	0.00015	0.00015	0.00004	0.00001	0.00010	0.00010	0.00008	0.00018	0.00011	0.00012	0.00006	0.00012	0.00012	0.00014	0.00017	0.00012	0.00012	0.00017	0.00015	0.00008	1	2624 - 3610
G2	0.00014	0.00015	0.00004	0.00001	0.00010	0.00009	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00015	0.00019	0.00012	0.00012	0.00016	0.00014	0.00009	0	2656 - 3648
H1	0.00016	0.00015	0.00003	0.00003	0.00010	0.00010	0.00040	0.00017	0.00016	0.00016	0.00006	0.00010	0.00011	0.00175	0.00163	0.00014	0.00016	0.00000	0.00013	0.00165	4	2958 - 4243
H2	0.00016	0.00015	0.00003	0.00004	0.00010	0.00010	0.00040	0.00017	0.00016	0.00016	0.00006	0.00010	0.00011	0.00175	0.00163	0.00014	0.00016	0.00000	0.00013	0.00165	1	2958 - 4243
I1	0.00015	0.00015	0.00004	0.00001	0.00009	0.00009	0.00008	0.00018	0.00011	0.00012	0.00006	0.00012	0.00012	0.00018	0.00018	0.00013	0.00013	0.00012	0.00015	0.00008	1	2890 - 4000
I2	0.00015	0.00015	0.00007	0.00003	0.00009	0.00009	0.00008	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00017	0.00018	0.00013	0.00013	0.00022	0.00028	0.00018	1	2890 - 4000
J1	0.00015	0.00015	0.00004	0.00001	0.00009	-	0.00008	0.00017	0.00012	0.00012	0.00006	0.00012	0.00012	0.00362	0.00354	0.00012	0.00012	0.00014	0.00015	0.00352	0	2805 - 3980
J2	0.00015	0.00015	0.00004	0.00001	0.00009	-	0.00008	0.00017	0.00012	0.00012	0.00007	0.00012	0.00012	0.00018	0.00020	0.00012	0.00012	0.00014	0.00015	0.00010	1	2805 - 3980
W-C	0.00006	0.00005	-0.00003	-0.00006	-0.00000	-0.00000	0.00030	0.00011	0.00006	0.00006	-0.00003	0.00002	0.00002	0.00352	0.00344	0.00004	0.00077	0.00012	0.00018			
Wins	0	0	0	0	13	6	0	0	0	0	0	0	0	0	0	0	0	0	0			

Table 15: Cross-provider FMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 3 records were compressed to 0.2 bits per pixel (CR = 40) using JPEG2000. The row label identifies the producer of the IREX KIND 3 enrollment record. The column label identifies the SDK that compares (proprietary templates of) the KIND 3 enrollment record against a KIND 1 verification record. **Error rate**: In each cell is the FMR at the threshold that gives FMR = 0.0001 when matching uncompressed KIND 1 vs. KIND 1 instances. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells filled green indicate FMR better than 0.0001. Cells with red text indicate FMR is better than when the producer of the template executes the match. The second to last row and third to last column give worst case values of FMR - 0.0001. The "wins" are the number of times the generator (matcher) gives the smallest absolute change in FMR , i.e. |FMR - 0.0001|.

KIND 7	INTEROPERABLE FALSE MATCH RATES																			Row	Num	Size
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Worst	Win	[5, 95] %
A1	0.00015	0.00014	0.00004	0.00002	0.00014	0.00014	0.00005	0.00017	0.00014	0.00014	0.00003	0.00012	0.00012	0.00040	0.00034	0.00009	0.00009	0.00015	0.00009	0.00030	0	2673 - 3667
A2	0.00015	0.00014	0.00004	0.00002	0.00014	0.00014	0.00005	0.00017	0.00014	0.00014	0.00003	0.00012	0.00012	0.00040	0.00034	0.00009	0.00009	0.00015	0.00009	0.00030	0	2673 - 3667
B1	0.00016	0.00015	0.00004	0.00002	0.00013	0.00013	0.00006	0.00017	0.00013	0.00013	0.00003	0.00012	0.00012	0.00030	0.00035	0.00009	0.00010	0.00011	0.00010	0.00025	1	2788 - 3920
B2	0.00016	0.00015	0.00004	0.00002	0.00013	0.00013	0.00006	0.00017	0.00013	0.00013	0.00003	0.00012	0.00012	0.00030	0.00035	0.00009	0.00010	0.00011	0.00010	0.00025	0	2788 - 3920
C1	0.00015	0.00015	0.00003	0.00001	0.00013	0.00014	0.00006	0.00018	0.00012	0.00012	0.00006	0.00012	0.00012	0.00033	0.00037	0.00010	0.00011	0.00013	0.00011	0.00027	0	
C2	0.00015	0.00015	0.00003	0.00001	0.00013	0.00013	0.00006	0.00017	0.00012	0.00012	0.00006	0.00012	0.00012	0.00053	0.00051	0.00010	0.00010	0.00013	0.00011	0.00043	0	2755 - 3920
D1	0.00015	0.00014	0.00002	0.00001	0.00013	0.00014	0.00010	0.00019	0.00010	0.00010	0.00005	0.00011	0.00012	0.00068	0.00084	0.00009	0.00010	0.00012	0.00011	0.00074	4	2029 - 2983
D2	0.00015	0.00014	0.00002	0.00001	0.00013	0.00014	0.00010	0.00019	0.00010	0.00010	0.00005	0.00011	0.00012	0.00068	0.00084	0.00009	0.00010	0.00012	0.00011	0.00074	0	2029 - 2983
E1	0.00016	0.00015	0.00002	0.00001	0.00013	0.00013	0.00007	0.00019	0.00011	0.00012	0.00005	0.00011	0.00012	0.00027	0.00033	0.00009	0.00010	0.00012	0.00011	0.00023	0	2029 - 2983
E2	0.00016	0.00015	0.00002	0.00001	0.00013	0.00013	0.00007	0.00019	0.00011	0.00012	0.00005	0.00011	0.00012	0.00027	0.00033	0.00009	0.00010	0.00012	0.00011	0.00023	0	2029 - 2983
F1	0.00016	0.00016	0.00004	0.00002	0.00019	0.00019	0.00006	0.00018	0.00011	0.00012	0.00009	0.00012	0.00013	0.00020	0.00031	0.00010	0.00010	0.00012	0.00011	0.00021	1	2822 - 3920
G1	0.00012	0.00012	0.00004	0.00002	0.00016	0.00016	0.00008	0.00015	0.00010	0.00011	0.00007	0.00010	0.00010	0.00044	0.00048	0.00008	0.00008	0.00008	0.00007	0.00038	3	2624 - 3610
G2	0.00012	0.00013	0.00004	0.00002	0.00015	0.00015	0.00007	0.00016	0.00012	0.00012	0.00005	0.00010	0.00011	0.00027	0.00030	0.00008	0.00008	0.00008	0.00007	0.00020	1	2656 - 3648
H1	0.00014	0.00013	0.00003	0.00002	0.00011	0.00012	0.00077	0.00013	0.00013	0.00012	0.00016	0.00011	0.00011	0.00389	0.00641	0.00076	0.00010	0.00000	0.00009	0.00631	1	2958 - 4243
H2	0.00014	0.00013	0.00003	0.00002	0.00011	0.00012	0.00077	0.00013	0.00013	0.00012	0.00016	0.00011	0.00011	0.00389	0.00641	0.00076	0.00010	0.00000	0.00009	0.00631	0	2958 - 4243
I1	0.00014	0.00014	0.00003	0.00001	0.00012	0.00012	0.00011	0.00017	0.00010	0.00011	0.00006	0.00011	0.00012	0.00019	0.00024	0.00010	0.00012	0.00010	0.00011	0.00014	2	2890 - 4000
I2	0.00014	0.00014	0.00006	0.00003	0.00012	0.00012	0.00011	0.00017	0.00010	0.00011	0.00006	0.00011	0.00012	0.00020	0.00024	0.00010	0.00012	0.00018	0.00020	0.00015	2	2890 - 4000
J1	0.00014	0.00015	0.00004	0.00002	0.00010	0.00010	0.00006	0.00017	0.00012	0.00012	0.00004	0.00012	0.00012	0.01222	0.00920	0.00010	0.00012	0.00014	0.00015	0.01213	0	2805 - 3980
J2	0.00014	0.00015	0.00004	0.00002	0.00010	0.00010	0.00006	0.00017	0.00012	0.00012	0.00004	0.00012	0.00012	0.00018	0.00022	0.00009	0.00012	0.00014	0.00015	0.00012	4	2805 - 3980
W-C	0.00006	0.00006	-0.00004	-0.00007	0.00009	0.00009	0.00067	0.00009	0.00004	0.00004	0.00006	0.00002	0.00003	0.01213	0.00910	0.00066	0.00002	0.00008	0.00010			
Wins	0	0	0	0	0	2	0	0	0	0	0	1	1	0	0	4	11	0	0			

Table 16: Cross-provider FMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 7 records were compressed to 0.2 bits per pixel (CR = 40) using JPEG2000. The row label identifies the producer of the IREX KIND 7 enrollment record. The column label identifies the SDK that compares (proprietary templates of) the KIND 7 enrollment record against a KIND 1 verification record. **Error rate**: In each cell is the FMR at the threshold that gives FMR = 0.0001 when matching uncompressed KIND 1 vs. KIND 1 instances. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells filled green indicate FMR better than 0.0001. Cells with red text indicate FMR is better than when the producer of the template executes the match. The second to last row and third to last column give worst case values of FMR - 0.0001. The "wins" are the number of times the generator (matcher) gives the smallest absolute change in FMR , i.e. |FMR - 0.0001|.

KIND 16	INTEROPERABLE FALSE MATCH RATES																			Row	Num	BPP
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	Worst	Win	[5, 95] %
A1	0.00020	0.00020	0.00002	0.00001	0.00010	0.00011	0.00076	0.00015	0.00085	0.00076	0.00007	0.00012	0.00012	0.00011	0.00017	0.00013	0.00014	0.00020	0.00017	0.00075	3	0.16 - 0.22
A2	0.00020	0.00020	0.00002	0.00001	0.00010	0.00011	0.00076	0.00015	0.00085	0.00076	0.00007	0.00012	0.00012	0.00011	0.00017	0.00013	0.00014	0.00020	0.00017	0.00075	1	0.16 - 0.22
B1	0.00019	0.00018	0.00003	0.00001	0.00006	0.00006	0.00008	0.00016	0.00012	0.00012	0.00002	0.00011	0.00011	0.00114	0.00069	0.00016	0.00015	0.00021	0.00019	0.00104	4	0.15 - 0.22
B2	0.00019	0.00018	0.00003	0.00001	0.00006	0.00006	0.00008	0.00016	0.00012	0.00012	0.00002	0.00011	0.00011	0.00114	0.00069	0.00016	0.00015	0.00021	0.00019	0.00104	0	0.15 - 0.22
F1	0.00019	0.00019	0.00002	0.00001	0.00009	0.00009	0.00038	0.00016	0.00014	0.00014	0.00006	0.00010	0.00010	0.00021	0.00024	0.00014	0.00015	0.00021	0.00018	0.00028	2	0.15 - 0.21
G1	0.00021	0.00020	0.00002	0.00001	0.00010	0.00010	0.00041	0.00016	0.00121	0.00116	0.00006	0.00010	0.00011	0.00014	0.00016	0.00014	0.00014	0.00019	0.00017	0.00111	2	0.17 - 0.23
G2	0.00019	0.00018	0.00002	0.00001	0.00010	0.00010	0.00057	0.00017	0.00148	0.00141	0.00006	0.00010	0.00010	0.00011	0.00014	0.00014	0.00014	0.00020	0.00018	0.00138	2	0.16 - 0.23
H1	0.00017	0.00018	0.00002	0.00001	0.00008	0.00009	0.00027	0.00011	0.00028	0.00026	0.00009	0.00009	0.00009	0.00062	0.00067	0.00031	0.00132	0.00000	0.00016	0.00122	5	0.14 - 0.20
H2	0.00017	0.00018	0.00002	0.00001	0.00008	0.00009	0.00027	0.00011	0.00028	0.00026	0.00009	0.00009	0.00009	0.00062	0.00067	0.00031	0.00132	0.00000	0.00016	0.00122	0	0.14 - 0.20
I1	0.00019	0.00020	0.00002	0.00001	0.00011	0.00011	0.00021	0.00018	0.00018	0.00018	0.00006	0.00012	0.00011	0.00018	0.00023	0.00018	0.00018	0.00023	0.00022	0.00013	0	0.15 - 0.21
I2	0.00019	0.00020	0.00002	0.00001	0.00011	0.00011	0.00021	0.00018	0.00018	0.00018	0.00006	0.00012	0.00011	0.00018	0.00023	0.00019	0.00019	0.00023	0.00022	0.00013	0	0.15 - 0.21
J1	0.00020	0.00020	0.00002	0.00001	0.00000	0.00000	0.00048	0.00021	0.00045	0.00042	0.00007	0.00001	0.00001	0.03124	0.02539	0.00017	0.00017	0.00023	0.00021	0.03114	0	0.15 - 0.21
J2	0.00021	0.00020	0.00002	0.00001	0.00011	0.00011	0.00070	0.00019	0.00032	0.00030	0.00006	0.00012	0.00011	0.00015	0.00022	0.00017	0.00017	0.00023	0.00021	0.00060	0	0.15 - 0.21
W-C	0.00011	0.00010	-0.00007	-0.00009	0.00001	0.00001	0.00066	0.00011	0.00138	0.00131	-0.00001	0.00002	0.00002	0.03114	0.02529	0.00021	0.00122	0.00013	0.00012			
Wins	0	0	0	0	4	3	0	0	0	0	3	2	1	0	0	0	0	0	0			

Table 17: Cross-provider FMR interoperability measured over the ALL-FAILURES partition of the OPS database (see section 5.1.1). All KIND 16 records were compressed to a fixed target of 3000 bytes using JPEG 2000. The row label identifies the producer of the IREX KIND 16 enrollment record. The column label identifies the SDK that compares (proprietary templates of) the KIND 16 enrollment record against a KIND 1 verification record. **Error rate**: In each cell is the FMR at the threshold that gives FMR = 0.0001 when matching uncompressed KIND 1 vs. KIND 1 instances. **Colors**: The on-diagonal within-provider elements are colored in yellow. Cells filled green indicate FMR better than 0.0001. Cells with red text indicate FMR is better than when the producer of the template executes the match. The second to last row and third to last column give worst case values of FMR - 0.0001. The "wins" are the number of times the generator (matcher) gives the smallest absolute change in FMR , i.e. |FMR - 0.0001|.

	INTEROPERABLE DIFFERENCE: FNMR (k3) - FNMR (k7) FOR JPEG 2000 AT CR = 40, BPP = 0.2																			Row		Bytes
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	MAD	Npos	[5, 95] %
A1	0.001	0.002	0.000	-0.002	0.000	-	-0.004	0.000	-0.004	0.001	0.007	0.004	0.007	-0.015	-0.007	-0.001	-0.002	0.002	-0.004	0.000	7	2673 - 3667
A2	0.001	0.002	0.000	-0.002	0.000	-	-0.004	0.000	-0.004	0.001	0.007	0.004	0.007	-0.015	-0.007	-0.001	-0.002	0.002	-0.004	0.000	7	2673 - 3667
B1	0.002	0.002	0.000	0.001	0.001	0.007	-0.004	-0.001	-0.001	0.000	0.013	0.001	0.010	-0.012	-0.016	-0.001	-0.002	0.003	0.001	0.001	10	2788 - 3920
B2	0.002	0.002	0.000	0.001	0.001	0.007	-0.004	-0.001	-0.001	0.000	0.013	0.001	0.010	-0.012	-0.016	-0.001	-0.002	0.003	0.001	0.001	10	2788 - 3920
C1	-0.007	-0.012	-0.007	-0.012	0.010	0.015	-0.012	-0.017	-0.005	-0.004	-0.006	0.004	0.011	-0.009	-0.013	-0.007	-0.006	-0.003	-0.007	-0.007	4	
C2	-0.008	-0.015	-0.007	-0.013	0.016	0.018	-0.016	-0.020	-0.015	-0.007	-0.010	-0.009	0.004	-0.043	-0.028	-0.018	-0.021	-0.005	-0.011	-0.011	3	2755 - 3920
D1	-0.030	-0.019	-0.016	-0.052	-0.019	-0.001	-0.005	-0.004	-0.018	-0.016	0.077	-0.034	-0.026	-0.031	-0.030	-0.045	-0.010	-0.019	-0.018	-0.019	1	2029 - 2983
D2	-0.030	-0.019	-0.016	-0.053	-0.019	-0.001	-0.005	-0.004	-0.018	-0.016	0.077	-0.034	-0.026	-0.031	-0.030	-0.045	-0.010	-0.019	-0.018	-0.019	1	2029 - 2983
E1	-0.047	-0.046	-0.017	-0.057	-0.029	0.009	-0.032	-0.032	-0.046	-0.043	0.032	-0.045	-0.031	-0.046	-0.049	-0.048	-0.051	-0.010	-0.010	-0.043	2	2029 - 2983
E2	-0.047	-0.046	-0.018	-0.059	-0.029	0.009	-0.032	-0.032	-0.046	-0.043	0.032	-0.045	-0.031	-0.046	-0.049	-0.048	-0.051	-0.010	-0.010	-0.043	2	2029 - 2983
F1	-0.004	-0.007	-0.006	-0.016	-0.025	-0.005	-0.020	-0.022	-0.008	-0.004	-0.010	-0.005	0.004	-0.011	-0.022	-0.006	-0.003	-0.009	-0.005	-0.007	1	2822 - 3920
G1	-0.011	-0.018	-0.012	-0.010	-0.019	-0.013	-0.027	-0.020	-0.012	-0.012	-0.012	-0.011	0.000	-0.070	-0.073	-0.014	-0.010	-0.004	-0.010	-0.012	0	2624 - 3610
G2	-0.015	-0.022	-0.012	-0.011	-0.025	-0.018	-0.018	-0.020	-0.018	-0.013	-0.001	-0.013	-0.009	-0.029	-0.040	-0.010	-0.010	-0.010	-0.011	-0.013	0	2656 - 3648
H1	-0.021	-0.021	-0.031	-0.051	-0.053	-0.031	-0.043	-0.034	-0.034	-0.022	-0.076	-0.044	-0.039	0.342	-0.288	-0.044	-0.037	-0.043	-0.030	-0.037	1	2958 - 4243
H2	-0.021	-0.021	-0.031	-0.051	-0.053	-0.031	-0.043	-0.034	-0.034	-0.022	-0.076	-0.044	-0.039	0.342	-0.288	-0.044	-0.037	-0.043	-0.030	-0.037	1	2958 - 4243
I1	0.000	-0.008	-0.004	-0.007	-0.009	0.004	-0.020	-0.019	-0.013	-0.010	0.003	0.004	0.023	0.006	-0.003	-0.001	-0.001	-0.001	0.001	-0.001	6	2890 - 4000
I2	0.001	-0.007	-0.004	-0.006	-0.007	0.004	-0.020	-0.019	-0.016	-0.012	0.001	0.004	0.021	0.007	-0.003	-0.001	-0.001	0.001	0.000	-0.001	7	2890 - 4000
J1	-0.001	-0.007	0.001	-0.002	0.029	-	-0.001	-0.004	0.000	0.001	0.008	0.000	0.001	-0.014	-0.004	-0.001	0.000	-0.001	-0.001	-0.001	5	2805 - 3980
J2	-0.001	-0.005	0.001	-0.002	0.022	-	-0.004	-0.004	0.000	0.001	0.006	-0.001	0.002	0.012	0.008	-0.002	-0.002	0.000	0.000	0.000	7	2805 - 3980
MAD	-0.007	-0.012	-0.007	-0.011	-0.009	0.004	-0.016	-0.019	-0.013	-0.010	0.006	-0.005	0.002	-0.014	-0.022	-0.007	-0.006	-0.004	-0.007			
Npos	5	4	2	2	6	8	0	0	0	4	12	7	11	5	1	0	0	5	3			

	INTEROPERABLE DIFFERENCE: FNMR (k3) - FNMR (k7) FOR JPEG 2000 AT 3000B																			Row		BPP
	A1	A2	B1	B2	C1	C2	D1	D2	E1	E2	F1	G1	G2	H1	H2	I1	I2	J1	J2	MAD	Npos	[5, 95] %
A1	0.003	0.007	0.008	0.000	-0.001	-	-0.007	-0.001	-0.001	0.004	0.001	0.014	0.009	-0.013	0.006	0.003	0.004	0.032	0.013	0.003	12	0.16 - 0.22
A2	0.003	0.007	0.008	0.000	-0.001	-	-0.007	-0.001	-0.001	0.004	0.001	0.014	0.009	-0.013	0.000	0.003	0.004	0.032	0.013	0.003	12	0.16 - 0.22
B1	0.005	0.009	0.013	0.000	-0.001	0.008	0.001	0.000	-0.003	0.001	-0.003	0.018	0.010	0.019	0.016	0.004	0.004	0.016	0.016	0.005	14	0.15 - 0.22
B2	0.005	0.009	0.013	0.000	-0.001	0.008	0.001	0.000	-0.003	0.001	-0.003	0.018	0.010	0.019	0.016	0.004	0.004	0.016	0.016	0.005	14	0.15 - 0.22
C1	0.001	0.003	0.002	-0.003	-0.037	-0.019	-0.015	-0.017	-0.003	-0.001	-0.013	0.007	0.010	0.061	0.021	0.001	-0.002	0.010	0.010	0.001	10	
C2	-0.004	-0.005	0.002	-0.005	-0.045	-0.027	-0.021	-0.016	-0.011	-0.007	-0.007	0.004	0.004	-0.018	-0.016	-0.008	-0.012	0.004	0.006	-0.007	5	0.15 - 0.22
D1	-0.028	-0.011	-0.011	-0.050	-0.022	-0.003	-0.002	-0.005	-0.017	-0.015	0.074	-0.019	-0.019	-0.043	-0.061	-0.041	-0.006	-0.018	-0.015	-0.017	1	0.20 - 0.30
D2	-0.028	-0.011	-0.011	-0.050	-0.022	-0.003	-0.002	-0.005	-0.017	-0.015	0.074	-0.019	-0.019	-0.043	-0.060	-0.041	-0.006	-0.018	-0.015	-0.017	1	0.20 - 0.30
E1	-0.048	-0.036	-0.016	-0.055	-0.042	-0.019	-0.034	-0.033	-0.051	-0.046	0.027	-0.032	-0.028	-0.022	-0.032	-0.046	-0.051	-0.015	-0.013	-0.033	1	0.20 - 0.30
E2	-0.048	-0.036	-0.017	-0.055	-0.042	-0.019	-0.034	-0.033	-0.051	-0.046	0.027	-0.032	-0.028	-0.022	-0.032	-0.046	-0.051	-0.015	-0.013	-0.033	1	0.20 - 0.30
F1	-0.002	0.005	-0.001	-0.010	-0.082	-0.053	-0.021	-0.018	-0.008	-0.004	-0.010	0.005	0.006	0.055	0.003	0.001	-0.001	0.004	0.009	-0.001	8	0.15 - 0.21
G1	-0.008	-0.010	-0.006	-0.014	-0.035	-0.026	-0.022	-0.014	-0.011	-0.013	-0.010	-0.011	0.008	-0.043	-0.046	-0.010	-0.009	-0.001	0.004	-0.011	2	0.17 - 0.23
G2	-0.010	-0.013	-0.006	-0.010	-0.047	-0.028	-0.018	-0.016	-0.019	-0.019	-0.005	-0.014	0.004	-0.016	-0.019	-0.008	-0.005	-0.001	0.004	-0.013	2	0.16 - 0.23
H1	-0.023	-0.017	-0.019	-0.037	-0.023	-0.025	-0.008	-0.030	-0.031	-0.019	-0.082	-0.037	-0.027	0.711	0.049	-0.033	-0.033	0.006	-0.014	-0.023	3	0.14 - 0.20
H2	-0.023	-0.017	-0.019	-0.037	-0.023	-0.025	-0.008	-0.030	-0.031	-0.019	-0.082	-0.037	-0.027	0.711	0.049	-0.033	-0.033	0.006	-0.014	-0.023	3	0.14 - 0.20
I1	0.004	0.009	0.009	0.001	-0.019	-0.007	-0.018	-0.010	-0.002	-0.004	0.001	0.023	0.020	0.010	0.001	0.007	0.009	0.036	0.023	0.004	13	0.15 - 0.21
I2	0.004	0.006	0.011	-0.001	-0.017	-0.006	-0.017	-0.010	-0.004	-0.006	0.001	0.022	0.022	-0.006	0.002	0.009	0.009	0.045	0.027	0.002	11	0.15 - 0.21
J1	0.003	-0.002	0.011	0.004	-0.014	0.569	-0.001	0.004	0.010	0.010	-0.010	0.016	0.025	0.028	0.036	0.006	0.005	0.017	0.019	0.010	15	0.15 - 0.21
J2	0.002	-0.003	0.009	0.001	-0.001	0.581	-0.004	0.004	0.010	0.010	-0.005	0.007	0.020	0.029	0.017	0.005	0.006	0.016	0.017	0.007	15	0.15 - 0.21
MAD	-0.002	-0.003	0.002	-0.005	-0.022	-0.019	-0.008	-0.010	-0.008	-0.006	-0.003	0.005	0.008	-0.006	0.002	0.001	-0.002	0.006	0.009			
Npos	9	8	10	3	0	4	2	2	2	6	8	11	13	9	12	10	8	13	13			

Table 18: Cross-provider FNMR (KIND 3) - FNMR (KIND 7) measured over the ALL-FAILURES partition of the OPS dataset (see section 5.1.1). The row label identifies the producer of the IREX enrollment record. The column label identifies the SDK that compares (proprietary templates of) the record against a KIND 1 verification record. **Error rates**: In each cell is the difference in FNMR for KIND 3 and KIND 7 at the threshold that gives FMR = 0.0001 for that combination of algorithm. The top table applies to images compressed at 40:1. The bottom table applies to images compressed to 3000 bytes. JPEG2000 was used throughout. **Colors**: The on-diagonal within-provider elements are colored yellow. Cells with red text indicate |ΔFNMR | > 0.02. Cells filled green indicate ΔFNMR > 0 i.e. KIND 7 outperforms KIND 3 i.e. the "expected" result, under compression. **MAD**: These rows and columns give the median absolute deviation. **Npos**: These rows and columns give the number of occasions KIND 7 outperforms KIND 3.

8.7. Effect of lossless compression

In section 8.2 the three IREX formats were matched without compression to assess whether they were *sufficient* for the intended purpose. It was shown that they invariably offer excellent accuracy relative to the parent 640x480 KIND 1 records. However, as shown in section 7.4, under (severe) lossy compression the impostor distribution moves, often to the left. This will have negative implications for identification systems and this motivates the study of how far the three IREX records can be compressed without *any* change in the pixel values. The ISO/IEC 19794-6 standard allows the use of two lossless compressors: JPEG2000 and PNG . The result of applying the PNG option to greyscale raster data is presented here for the four IREX formats. Lossless compression does not alter the original image and therefore preserves the iris texture exactly as it was rendered by the camera.

The notable observations are drawn from Figure 29 which uses boxplots to show the distributions of the sizes obtained by application of the PNG compressor to all images in all three datasets. Compressed size is also broken out by SDK. Variation across SDKs is observed only to the extent that the SDKs prepared the various KINDS differently. The modes of variation are described in section 8.4.

The notable results for lossless compression are as follows.

- ▷ The median size for the most interoperable KIND 3 records (see Table 12), from producers A1, B1 and I1 are around 65KB. Smaller sizes are available around 50KB for the E1 SDK which crops an image to three quarters of the height that A1 and B1 do. By more tightly cropping the iris, the number of bits per pixel is larger when compressing to a given file size (here 3000 bytes) and D1 and E1 KIND 3 instances are accordingly among the most matchable instances *for a fixed target file size* (see the interoperable error rates in Table 12). Tight vertical margins are sustainable (see section 8.5) so the size of losslessly compressed KIND 3 records is 50KB (as attained by SDK D1).
- ▷ The best interoperable error rates for KIND 7 in Table 13 are again for producers B1 and A1. These images can be compressed to a median size of about 30KB. This is at least a factor of two smaller than KIND 3, the improvement obviously due to the compressibility of the mask regions.
- ▷ For KIND 16 records, the best interoperable error rates (see Table 14) occur for the A1 algorithm, and are losslessly compressed down to 10KB. The low size is attributable to the low circumferential and radial sampling rates. The variance also is very small because the polar image size is fixed. The low mean and variance *are* distinct advantages of the KIND 16 record. Whether losslessly compressed KIND 16 records outperform JPEG2000 compressed KIND 7 records at 10KB has not been studied but is worthy of further investigation.

8.8. Predictive power of image quality scores

Three SDKs, A1, D1, and G1 reported iris image quality scores. These are computed during the preparation of the IREX records. The quality values are stored in the record header. This section examines the association between those values and the observed accuracy measures. The analysis is restricted to the quality scores reported for KIND 3 records.

Correlation with quality	A1	D1	G1
iFMR	-0.051	-0.403	-0.321
iFNMR	-0.104	-0.236	-0.334

Table 19: Spearman correlation coefficients for quality scores of SDKs A1, D1 and G1 and image-specific errors. iFMR and iFNMR are computed at overall threshold of iFMR =0.001 for KIND 1 uncompressed comparisons. Quality scores are computed for KIND 3 record types. A1 quality scores show little correlation with image error, while D1 and G1 show some correlation.

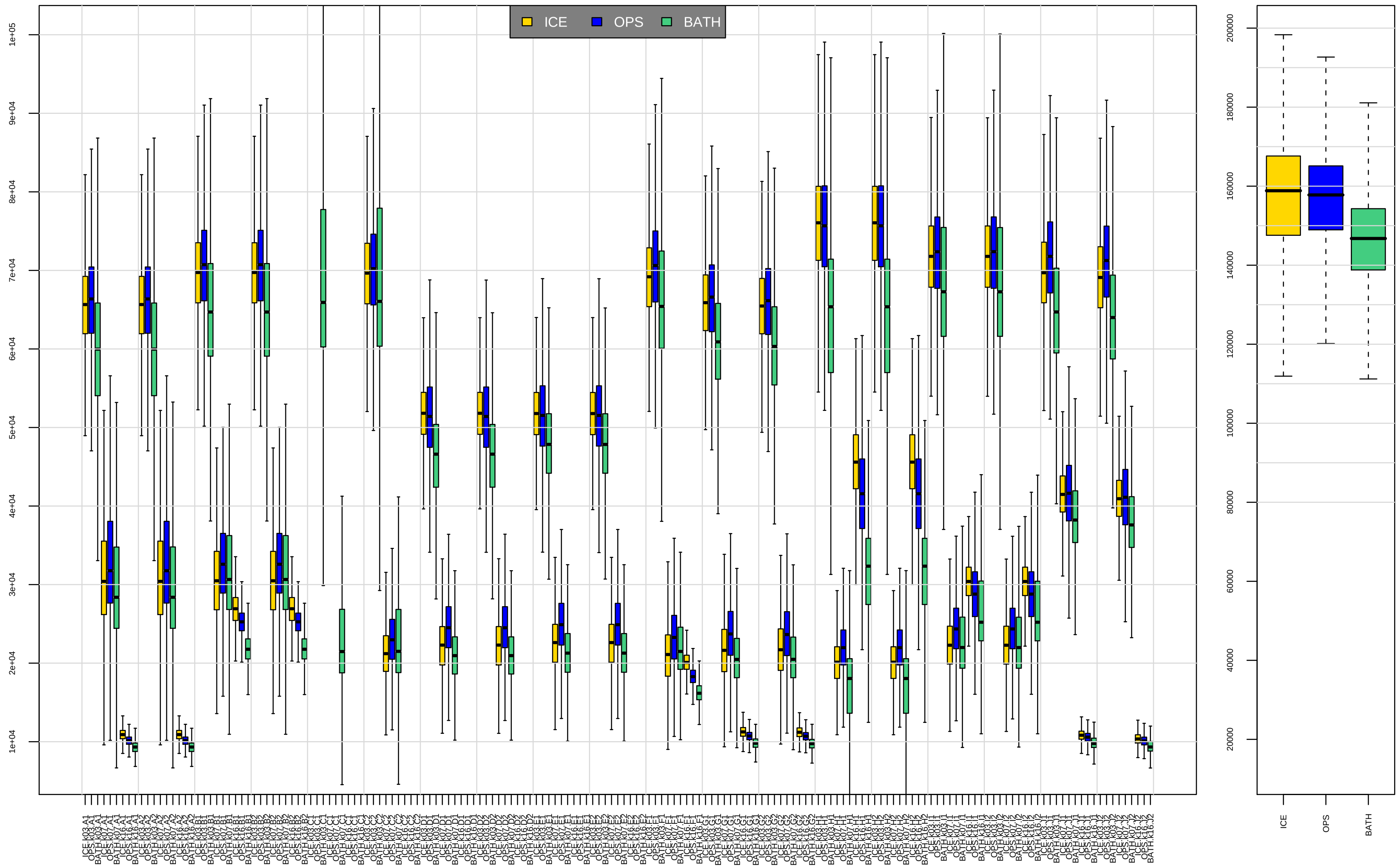
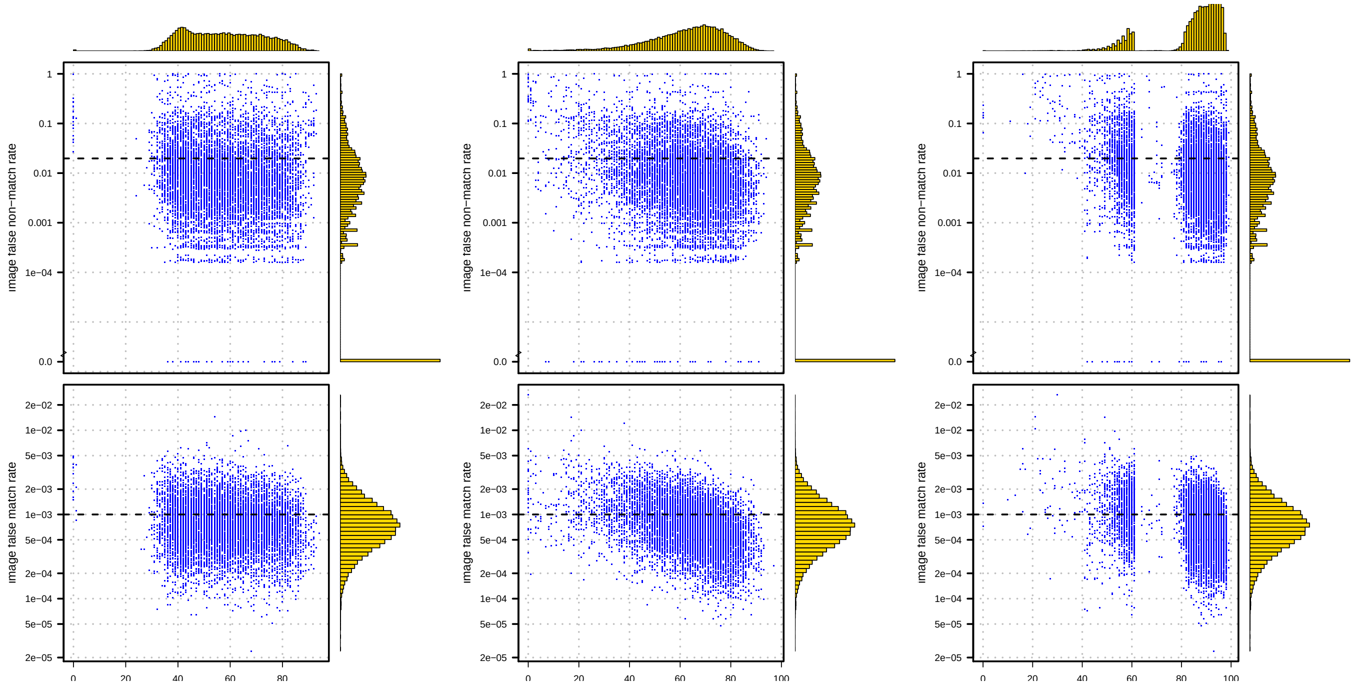


Figure 29: The distribution of losslessly compressed image size in bytes by DATASET, KIND and SDK. The whiskers extend to the value expected to give 99.9% coverage of a Normal deviate.

The excellent compression of the KIND 16 records is due to the the radial and circumferential sampling of the parent image. This is not a lossless operation in the sense that polar-to-rectilinear reconstruction yields pixel values different from those in the parent.



(a) Quality scores computed by A1

(b) Quality scores computed by D1

(c) Quality scores computed by G1

Figure 30: Quality score vs. image FMR and image FNMR for images in the ICE dataset. The quality scores are computed by the algorithm identified in the figure label as applied to uncompressed KIND 3 images. Image error rates are computed using comparison scores of uncompressed KIND 1 records and aggregated over all SDKs. Threshold was set to give global FMR = 0.001 on uncompressed KIND 1 instances.

8.8.1. Relationship of quality to image error rates

A quality algorithm is effective if it assigns higher quality scores to images with low image false match rate and image false non-match rates. Figure 30 shows the scatter plot of quality vs. aggregate iFMR and iFNMR. (equations 11 and 13). Histograms of image error rates and quality appear on the side of the plots. For all three quality algorithms, there are negative correlation between quality scores and image error rates. However, SDKs D1 and G1 quality scores exhibit a higher correlation with image error rates (see Table 19). Image errors are computed at the threshold of FMR $= 0.001$ and is aggregated over all comparison SDKs. This threshold is chosen to assure there are few errors per image. Image false match rates are computed over 16,320 impostor comparisons, and the average number of genuine comparisons per image is 245.

Figure 31 shows the boxplots of quality scores for images of each KIND and for the four ICE partitions CLEAR ICE, BLUE GOATS, BLUE WOLVES and BLACK ICE [49]. An effective quality algorithm should assign the highest scores to CLEAR ICE images and the lowest to BLACK ICE images. BLUE GOATS and BLUE WOLVES should have quality scores in between. Any other result is undesirable. The notable results are as follows.

- ▷ Quality scores computed by algorithms D1 and G1 trend in the expected direction (highest quality with CLEAR ICE progressing down to lowest quality with BLACK ICE. D1 exhibiting a bigger difference in the quality score median of the four partitions.
- ▷ Another observation is the median of G1 quality scores is much higher than those of A1 and D1. That means the

[49] The partitioning was computed using all comparison algorithms with the threshold set to give a global FMR = 0.001. Recall that CLEAR ICE images are those with image FMR and image FNMR lower than the operational FMR and FNMR and BLACK ICE are images with both image FMR and FNMR higher than the operational error rates.

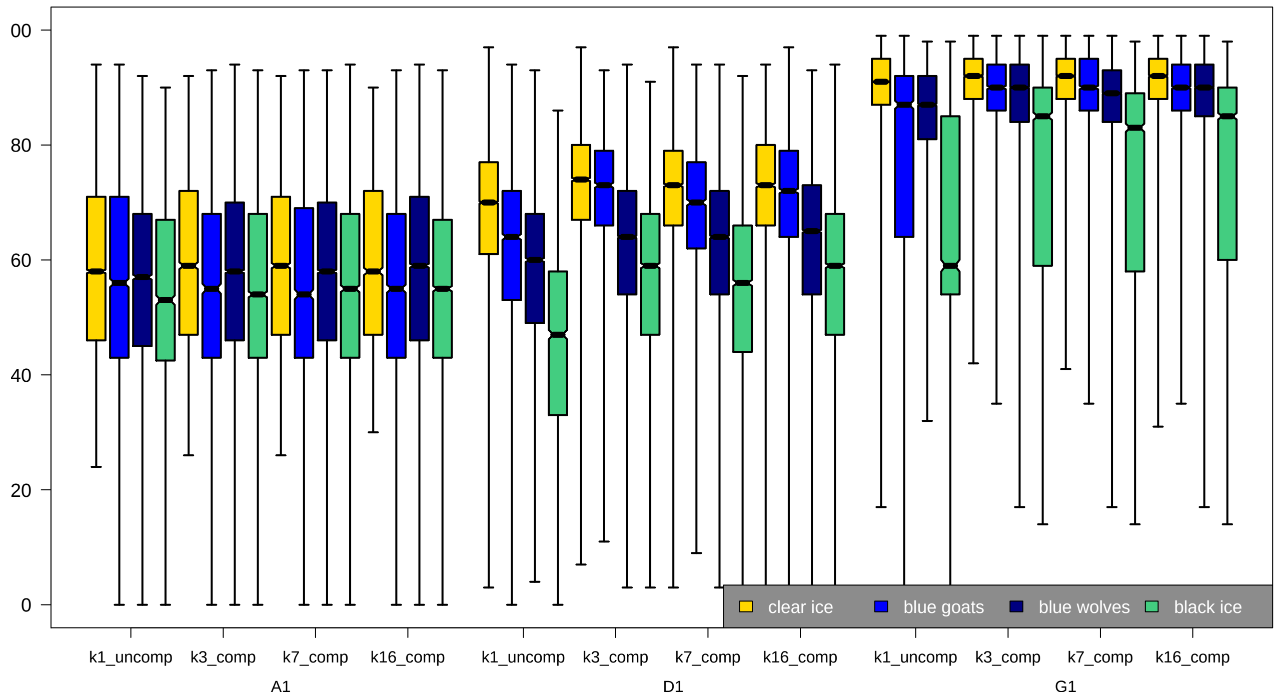


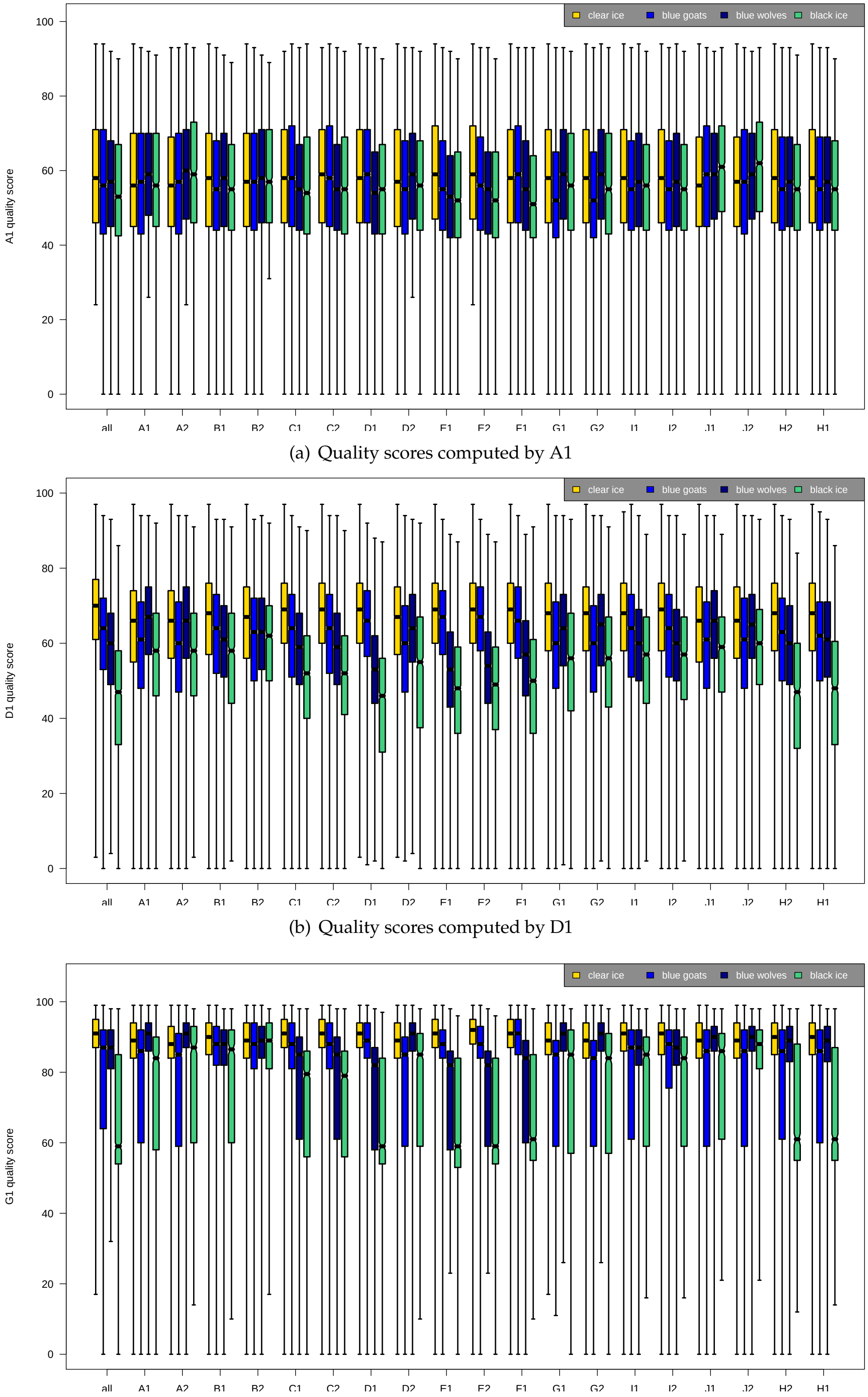
Figure 31: Boxplots of quality scores for the four ICE partitions: CLEAR ICE, BLUE GOATS, BLUE WOLVES and BLACK ICE. All three image quality algorithms are applied to images of each KIND. The KIND 1 records are uncompressed and all others are JPEG2000 compressed to 2000 bytes. The SDKs are present on the horizontal axis because each was used independently to partition the ICE images into the four categories. In all cases a global threshold was used, that which gives FMR = 0.001 on uncompressed KIND 1 instances.

same quality score value for G1 and D1 (or A1) will relate to different image error rates. This indicates raw quality scores are not directly interoperable and some calibration is needed.

▷ Table 19 shows correlation coefficients between aggregated image error and quality scores of SDKs A1, D1 and G1. The latter two assign higher quality scores to images with lower image error rate, which is exhibited by a modest negative correlation. The weak correlation between A1 quality scores and image-specific error rates indicates the lack of a strong relationship between quality algorithm A1 and image-specific error rates. SDK A1 assigns lower quality scores to BLACK ICE images, but only slightly higher quality scores to CLEAR ICE than BLUE GOATS or BLUE WOLVES.

▷ Quality algorithms might be expected to be most effective in the native case, when the quality computation and the matcher are from the same provider. Nonetheless, it could be useful if quality scores of one SDK could correlate with performance of other SDKs. Effectiveness of each SDK in predicting performance of different comparison algorithms are shown in Figure 32. Each plot corresponds to one of the three quality SDKs that computed quality scores (A1, D1, and G1). Partition has been done across all comparison algorithms (aggregate iFMR and iFNMR) at operational threshold FMR = 0.001 (shown in the leftmost set of boxes) as well as each comparison algorithms.

The plots of Figure 32 are useful in examining association of the quality scores with different comparison algorithms, that is whether the quality algorithm could be generalized across comparison algorithms. Quality scores generated by SDKs D1 and G1 are reasonably generalizable to other comparison algorithms: CLEAR ICE images are given higher quality scores than BLACK ICE images for each comparison algorithm. The median A1 quality scores of BLACK ICE images is higher than median quality score of CLEAR ICE images for comparison algorithms A2, J1 and J2. Furthermore, A1 quality



(a) Quality scores computed by A1

(b) Quality scores computed by D1

(c) Quality scores computed by G1

Figure 32: Boxplots of quality score for the CLEAR, BLUE GOATS, BLUE WOLVES, and BLACK image partitions of the ICE dataset. Quality scores are computed using the quality algorithms identified in the figure labels as applied to uncompressed KIND 3 records. ICE partition is done at threshold FMR = 0.001. The set of boxes on the left shows quality scores vs. ICE partition where image errors are aggregated over all SDKs. The remaining boxes show the quality score distribution for ICE 2006 partition using each comparison algorithms. Quality algorithm D1 seems to be most effective followed by quality algorithm G1.

scores of BLUE WOLVES and BLUE GOATS are higher than CLEAR ICE quality scores for comparison algorithms A1, A2, J1 and J2. These are not the expected behavior and suggests that A1 quality scores cannot be used as a predictor of performance for A2, J1 or J2 SDKs. For all other SDKs, the difference in A1 median quality scores of images in the four partitions of ICE dataset, is much less then the difference in median quality scores of D1 and G1.

8.8.2. ERROR VS. REJECT CURVES

Figures 33 and 34 show the error vs. reject curves[26] for images in the OPS dataset for all SDKs. Similar results for ICE images are shown in Figure 35. The goal is to state how efficiently rejection of low quality samples results in improved performance. This models the operational case in which quality is maintained by reacquisition after a low quality sample is detected. Consider that a pair of samples (k, l) from subject i, with qualities $q_i^{(k)}$ and $q_i^{(l)}$, are compared to produce a score $s_{ii}^{(kl)}$, and this is repeated for N such pairs.

We introduce thresholds u and v that define levels of acceptable quality and define the set of low-quality entries as

$$R(u, v) = \left\{(i, k, l) \mid q_i^{(k)} < u, \quad q_i^{(l)} < v\right\} \tag{20}$$

We compute FNMR as the fraction of genuine scores below threshold computed for those samples *not* in this set

$$\text{FNMR}(t, u, v) = \frac{\left|\left\{s_{ii}^{kl} \mid s_{ii}^{kl} \geq \tau, (i, k, l) \notin R(u, v)\right\}\right|}{\left|\left\{s_{ii}^{kl} \mid s_{ii}^{kl} < \infty, (i, k, l) \notin R(u, v)\right\}\right|} \tag{21}$$

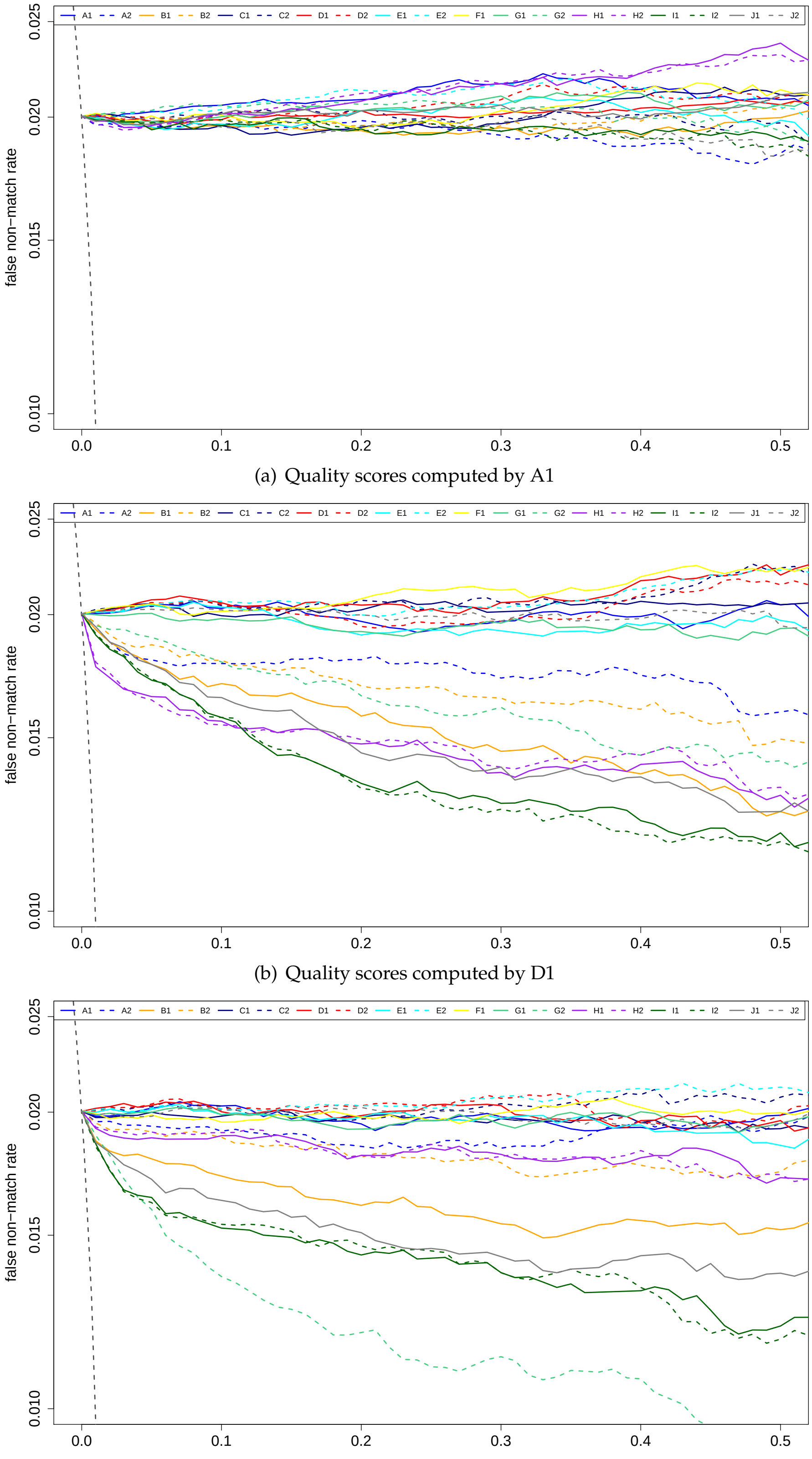
If the quality values are perfectly correlated with the genuine comparison scores, setting threshold τ to give an overall FNMR of x and then rejecting x percent with the lowest qualities should result in FNMR of zero after recomputing FNMR.

In Figure 33 we measured FNMR improvements as poor quality samples are rejected. We set the value of τ to give a false non-match rate of two percent, u and v are varied to show the dependence of FNMR on quality. Likewise, in Figure 34, to measure FMR improvement as poor quality samples are rejected, we set threshold to give an overall FMR of two percent and then reject proportion x with the lowest qualities and re-compute FMR. Pair wise quality is computed using the minimum of the two images being compared.

The most operationally relevant part of the error vs. reject curves is usually on the left side where a small fraction, x, of low-quality rejections would be tolerable from the perspective of forcing a second enrollment attempt. However, for the ICE database used here, the appropriate fraction is probably larger because the camera's own quality measurement apparatus was suppressed (see section 5.3) such that many of the ICE images wouldn't ordinarily be available.

For a good quality algorithm, FNMR should decrease quickly with the fraction rejected. An almost flat curve (as is the case with A1) suggests that the quality algorithm is not effective in prediction of performance. Another desirable feature of a quality algorithm is how well it can be generalized to predict performance of other SDKs. Figure 33 indicates that SDK G1 is most effective in prediction of it's own performance, but also does well with I1, I2 and J1. Rejection of samples deemed to have low quality by the D1 quality measure improved the FNMR of H1, H2, I1, I2, and J1 more than other SDKs including D1 or D2.

It is important to note that, because the way images in OPS dataset have been selected, this dataset consists of *matchable* images (see section 5.1). The small percentage of images that cause recognition error, are not really problematic images and so the quality algorithms might be unable to detect and quantify image imperfections to the fine degree needed here to make a difference in error vs. reject curves. That might be the reason no big improvement in error rates is observed as poor quality samples are rejected. We conclude that effectiveness of quality scores in predicting the positive or negative contribution of the image to overall false match and false non-match error rates are more appropriately quantified by quality score correlation with image-specific error rates.

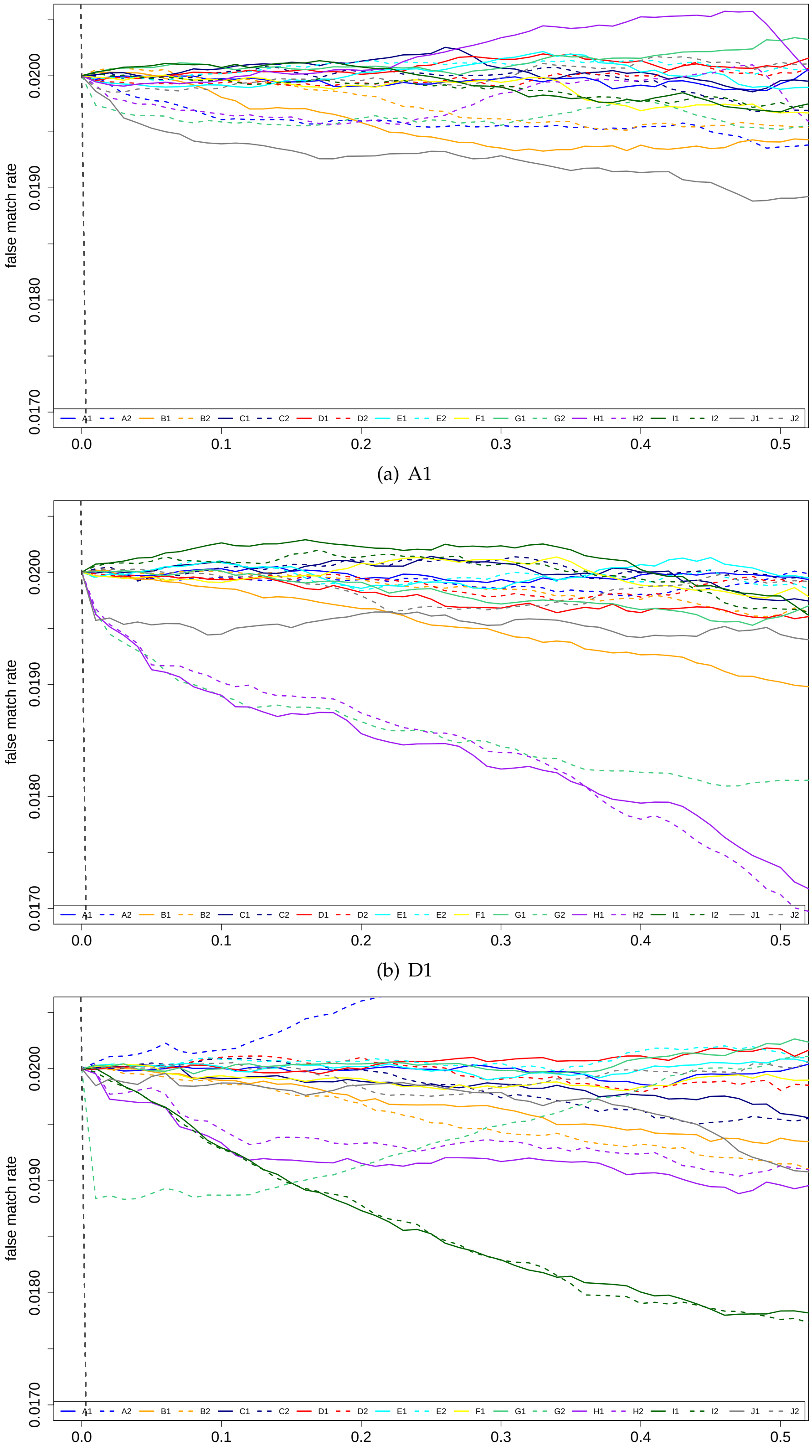


(a) Quality scores computed by A1

(b) Quality scores computed by D1

(c) Quality scores computed by G1

Figure 33: FNMR vs. reject curves for quality scores from SDKs A1, D1 and G1 on dataset OPS . The threshold is set to give an initial FNMR = 0.02. The gray dotted line shows the ideal case where rejection of two percent of lowest quality results in zero FNMR. While 50% rejection is not operationally feasible, it is shown here to display any unexpected behavior of the quality assessment algorithm. Quality scores of SDK G1 is most effective in predicting performance of SDK G2. Both SDKs D1 and G1 reasonably predict performance of SDKs I1 and I2. Quality scores of SDK D1 more effectively predicts the FNMR of the SDK I2 than D1.



(a) A1

(b) D1

(c) G1

Figure 34: FMR vs. reject curves for quality scores from SDKs A1, D1 and G1 on dataset OPS . The threshold is set to give overall FMR of 0.02. There is no strong relation between quality scores and FMR except for scoring SDK G2 which shows a slight decrease in FMR when rejecting low quality images assessed by SDK D1 and bigger improvement in FMR when rejecting images assessed to be the lowest quality by SDK G1. The gray dotted line shows the ideal case where rejection of two percent of lowest quality results in zero false match rate. While 50% rejection is not operationally feasible, itis shown here to display any unexpected behavior of the quality assessment algorithm.

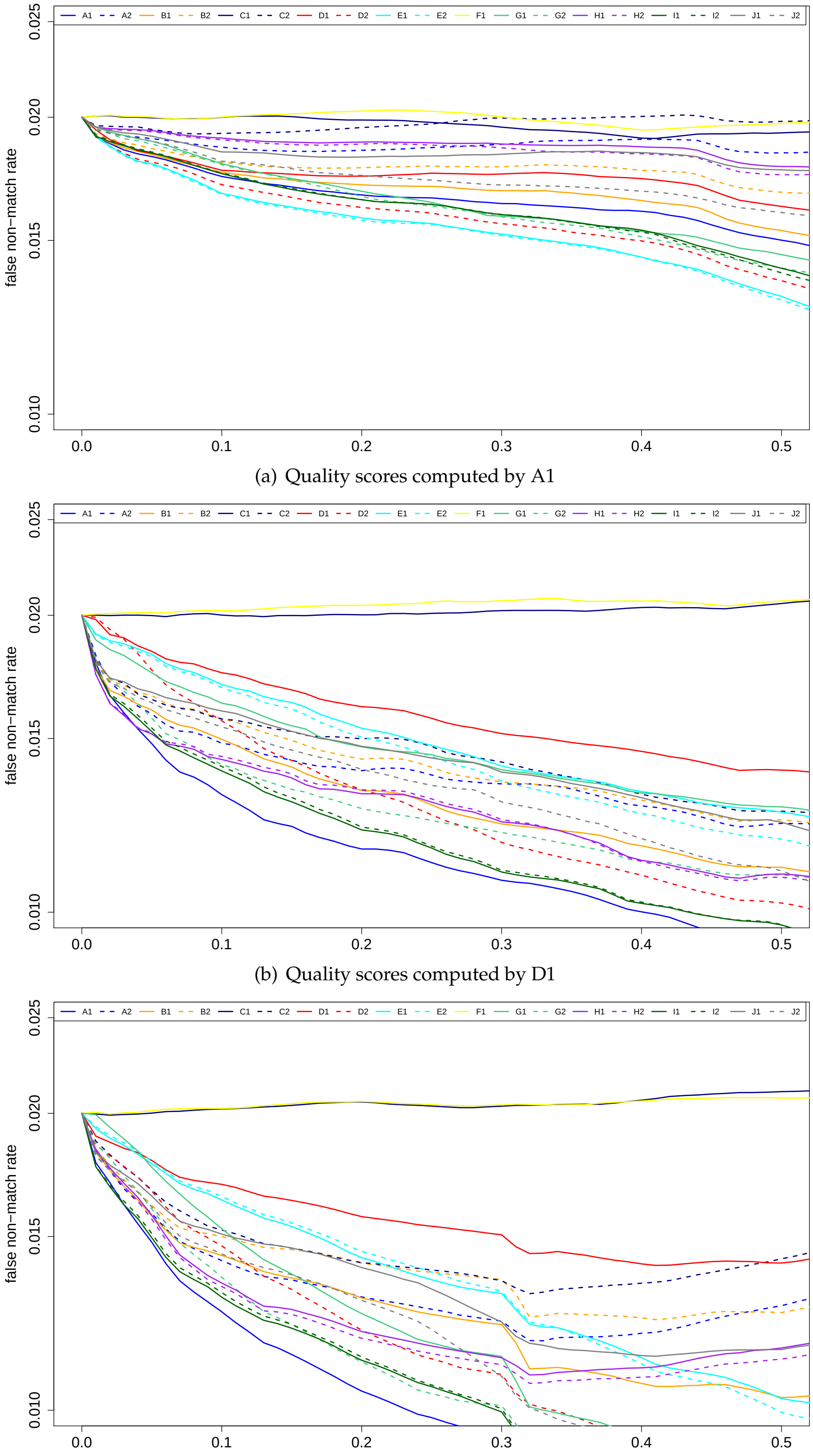
FNMR vs. reject curves for ICE dataset are shown in Figure 35. Results are similar to that observed for the OPS dataset. Rejection of images judged to have poor quality by SDKs G1 or D1 improves the FNMR of most SDKs except for C1 and F1. The greatest reduction in FNMR was achieved for A1, I1, I2, G2, H2, H1. In general the FNMR vs. reject curves for SDKs D1 and G1 are similar. Rejection of samples that SDK A1 considered of low quality scores did not significantly improve the false non-match rate. One caveat here is that ICE contains multiple images per subject, and the number of images varies by subject. Therefore rejection of a single image usually results in exclusion of many genuine comparison scores.

To compare effectiveness of quality scores in predicting performance of each SDK, we plotted error vs. reject curves for each comparison SDK. Figure 36 shows FNMR vs. reject curves for SDKs C1, H1 and G2. Each plot has three curves corresponding to three quality assessment algorithms A1, D1 and G1. Threshold is set at overall FMR = 0.0001 to ensure there are enough false non-matches. Quality scores of SDK D1 works reasonably for SDKs H1, H2, G2, and J1. Quality scores of SDK G1 works well for G2, G1, H1, H2, and I1.

Note that the main difference in plots of Figures 33 and 36 is the choice of threshold at which false nonmatch rates are computed. For Figure 33 threshold is fixed at per SDK FNMR = 0.02. Threshold value is different for each SDK, but there is same fraction of false reject for each of them. In Figure 36, the same threshold is applied to all SDKs, therefore different FNMR results for each SDK. While setting threshold at fixed FNMR per SDK is more reasonable in evaluating effectiveness of quality algorithms in improving FNMR when rejecting lowest quality samples, fixing the threshold value models an operational scenario where different comparison or quality algorithms could be deployed at a threshold fixed across the application.

There is no strong relation between quality scores and FMR . Rejection of samples considered low quality by SDK G1, improved FMR of G2 slightly, and FMR of H1, H2, J1, I1, and I2 barely. SDK D1 shows a slight decrease in FMR for G2, H1, H2, and J1. While a decrease in FMR with rejection of bad quality samples would be beneficial, impostor dissimilarity scores should be high regardless of the underlying quality: False matches should occur only when samples are biometrically similar (with regard to a matcher) as for example when identical twins' faces are matched.

Quality evaluation is a subject of ongoing research, and further work in this area is underway.

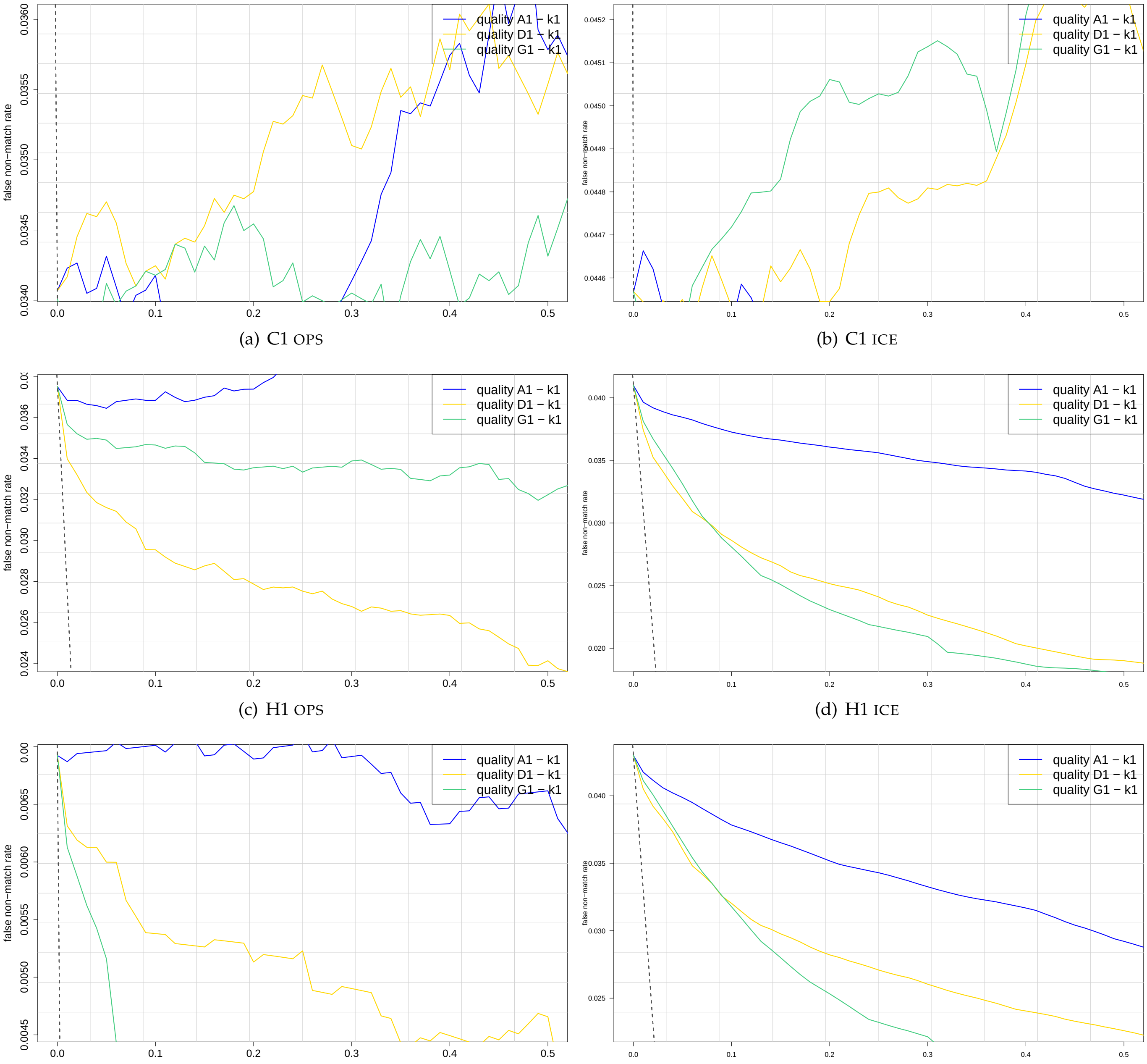


(a) Quality scores computed by A1

(b) Quality scores computed by D1

(c) Quality scores computed by G1

Figure 35: FNMR vs. reject curves for quality scores from SDKs A1, D1 and G1 on dataset ICE . Threshold is set at overall false non-match rate of two percent. While 50% rejection is not operationally feasible, it is shown here to display any unexpected behavior of the quality assessment algorithm. Both SDKS D1 and G1 reasonably predict performance of other SDKs except C1 and F1. Quality scores of SDK A1 does not effectively predicts the FNMR of any SDKs.



(a) C1 OPS

(b) C1 ICE

(c) H1 OPS

(d) H1 ICE

(e) G2OPS

(f) G2 ICE

Figure 36: FNMR vs. reject curves for images matched by selected SDKs and quality assessed by SDKs A1, D1, and G1. The graphs in the left column apply to OPS images, with ICE images on the right. The threshold is set to give overall FMR of 0.0001, which results in different initial FNMR for each SDK. For OPS none of the three quality score algorithms is effective in predicting performance of SDK C1. Quality scores of D1 and G1 seem to be effective for H1. Quality scores of G1 works best for SDK G2. While 50% rejection is not operationally feasible, it is shown here to display any unexpected behavior of the quality assessment algorithm.

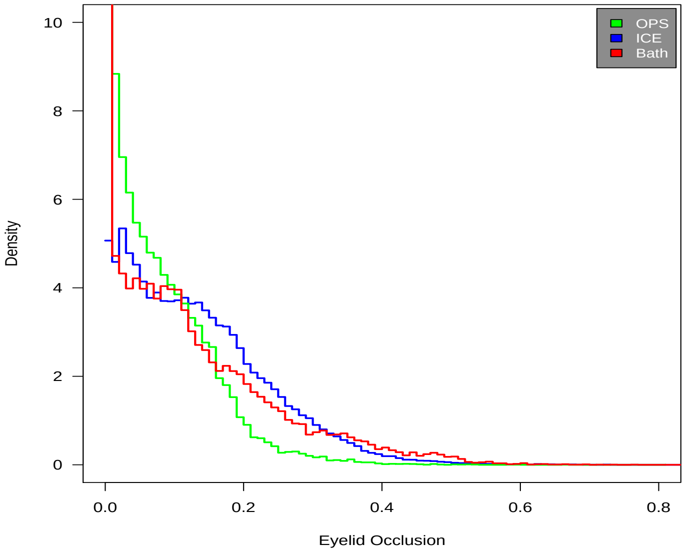


Figure 37: For the three IREX datasets, the density of the occlusion metric, Θ, as computed from the KIND 3 and KIND 7 instances generated by the A1 implementation applied to OPS images.

8.9. Effect of geometric properties

The following subsections quantify the effects of eyelid occlusion, dilation and iris-pupil center displacement on recognition accuracy. These studies are made possible by utilizing the standardized metadata present in the IREX records.

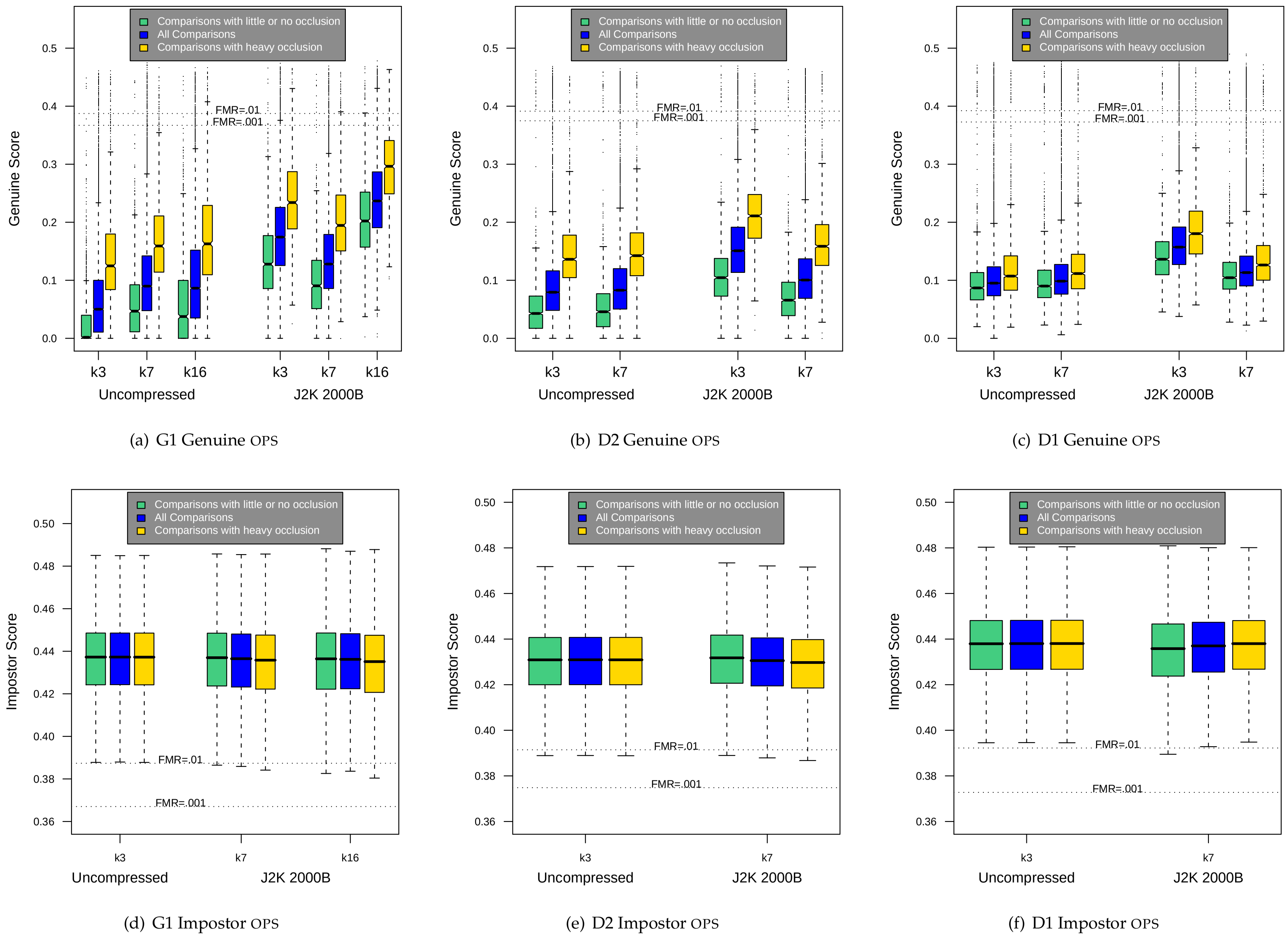
8.9.1. Eyelid occlusion of the iris

Eyelid occlusion is a well known problem for iris matchers [57, 38, 37]. When an eyelid occludes part of the iris, it reduces the information available for matching. The upper eyelid tends to be much closer to the iris and is responsible for a greater amount of occlusion than the lower eyelid. The resting position of the upper eyelid is usually between 0.5 and 2.0 millimeters below the top of the iris[5], but can depend on several factors. For example, when a person is tired, the upper eyelids droop, and several studies [30, 61, 13] have shown that eyelids lose tension as a person ages. This causes both the upper and lower eyelids to rest further down on the eye.

A person's gaze direction can also affect the position of the upper and lower eyelids relative to the pupil center. Read [55] determined that when a person looks down by 40 degrees, the distance from the pupil center to the lower eyelid reduces by an average of 2.7mm, and the distance from the pupil center to the upper eyelid reduces by 0.5mm. How an upward gaze affects the eyelid positions is less well known. The same study found that asians and females tend to have a smaller distance between the upper and lower eyelids, although Hollingsworth [31] found no statistical difference in the amount of eyelid occlusion between males and females. Becker et al. [9] noted that the upper eyelid can undergo idiosyncratic movements of up to 5 degrees while the eye itself is completely stationary. This suggests there may be some benefit in capturing a sequence of images and retaining only the one with smallest occlusion.

Since eyelid occlusion degrades matching accuracy, a person may attempt to purposefully occlude his iris to elicit a false match or false non-match. Daugman [14] described a method of compensating for occlusion by normalizing the score according to how many bits are available for comparison between two iris images. This should transform all samples of impostor scores to the same binomial distribution. When the impostor distribution is stabilized, the performance of the iris samples is reflected in the genuine distribution.

Most iris matchers mask the areas of the iris that are occluded by the eyelids before performing the comparison. The mask typically has a horizontal or parabolic edge. IREX required participants to replace the upper and lower eyelids with



(a) G1 Genuine OPS

(b) D2 Genuine OPS

(c) D1 Genuine OPS

(d) G1 Impostor OPS

(e) D2 Impostor OPS

(f) D1 Impostor OPS

Figure 38: The effect of eyelid occlusion on the the genuine (top) and impostor (bottom) distributions for three iris recognition algorithms applied to images from the OPS dataset.

a solid color for KIND 7 records, which makes the sample more compressible. In the analyses that follow, the difference between the cropped KIND 3 and cropped-and-masked KIND 7 records, is exploited to estimate eyelid occlusion. It is computed as the fraction of pixels within the annular iris region that were replaced by the mask in the KIND 7 record.

$$\Theta \;=\; \frac{\sum_{(x,y)\in\mathcal{R}} \;\; \delta(I_{k3}(x,y) - I_{k7}(x,y))}{\sum_{(x,y)\in\mathcal{R}} \;\; 1} \tag{22}$$

where $\delta(x)$ is the Delta function, and $\mathcal{R}$ denotes the iris region as approximated by the annulus defined by the best-fit limbic and pupillary circles reported by the I1 generator. Figure 37 plots the distribution of the measured occlusion for iris images from each dataset according to A1.

The effect of eyelid occlusion on the comparison score is assumed to be proportional to the maximum occlusion over the enrollment and verification samples (i.e. $\Theta = \max(\Theta_1, \Theta_2)$). To measure the effect of eyelid occlusion on recognition accuracy, comparisons were separated into three groups: those with Θ above the 90th percentile among genuine comparisons; and those with Θ below the 10th percentile; and those in the middle. In this case, the Θ values were 0.1700 and 0.0014 respectively. Figure 38 compares the score distributions for both sets for various algorithms and record formats. Figures for all SDKs are available in the IREX SUPPLEMENTAL appendices.

Several conclusions can be drawn from the figures:

- ▷ Large amounts of occlusion lead to higher genuine scores and lower impostor scores for most algorithms. Some algorithms produced higher impostor scores when the occlusion amount was large.
- ▷ The occlusion effect is not reduced for compressed images. The genuine score distributions for the high and low occlusion sets appear equally separated regardless of whether the images are JPEG2000 compressed or not.
- ▷ All of the record formats appear equally affected by eyelid occlusion.
- ▷ Large amounts of eyelid occlusion increase the probability of a false match for D2 but not for D1. Furthermore, large amounts of occlusion actually reduce the probability of a false match for D1 when the images are JPEG2000

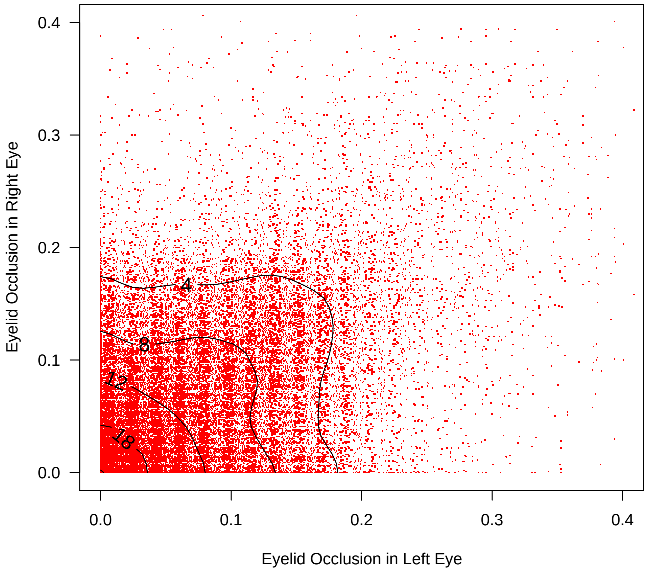


Figure 39: Density plot of eyelid occlusion for left-right eye pairs from the same person. Each point corresponds to an eye pair from the OPS dataset. If no correlation existed, the contours would be straight lines of slope negative one.

compressed.

The fact that the eyelid occlusion effect is as pronounced for KIND 7 compressed images as any other is interesting. The eyelid regions in a KIND 7 record are masked to make the image easier to compress, and one might expect that images with less visible iris texture would compress better, since less information needs to be retained. It may be that the variations in the amount of occlusion are not enough to produce a significant effect on how well an image compresses.

When both irises of a person are captured concurrently, the amount of eyelid occlusion presented is expected to be similar for both images. The OPS images were captured sequentially, but Figure 39 shows that a correlation is still present. This would offer one explanation for why cross-eye comparisons might be more likely to false match for some algorithms.

8.9.2. Dilation of the iris

The primary anatomical function of the iris is to control the amount of light that enters the eye by varying the size of the pupil. This is accomplished by two complementary muscles groups 1) the sphincter muscles, which constrict the pupil, and 2) the dilator muscles, which expand the size of the pupil. The sphincter muscle fibers are oriented circumferentially and are located behind the pupillary region of the iris. The dilator muscles are oriented radially and are located behind the stroma. Although the stroma is nearly flat, it cannot be entirely regarded as a two-dimensional surface since it consists of a meshwork of interlacing fibrous tissue. When the pupil size changes, these fibers can change both in direction and shape, altering the apparent texture of the iris surface. For example, Phang [48] noted that crypts close during mydrasis, but open during miosis, sometimes to reveal underlying iris vessels. This section investigates how variations in pupil size affect the accuracy of the iris matchers.

The most common method of compensating for deformations caused by changes in pupil size is to assume uniform elasticity of the iris surface[15]. However, Hollingsworth et al.[31] demonstrated that this results in a pupillary dilation effect. In particular, it was shown that large differences in pupillary dilation are more likely to produce false non-matches. Furthermore, Phang[48] noted that the precise surface deformations vary from one person to another, and studies have demonstrated that non-linear methods of deformation correction are better at reducing error rates[63, 60]. In general, the iris features near the pupillary region stretch more in the radial direction than those near the periphery, although the

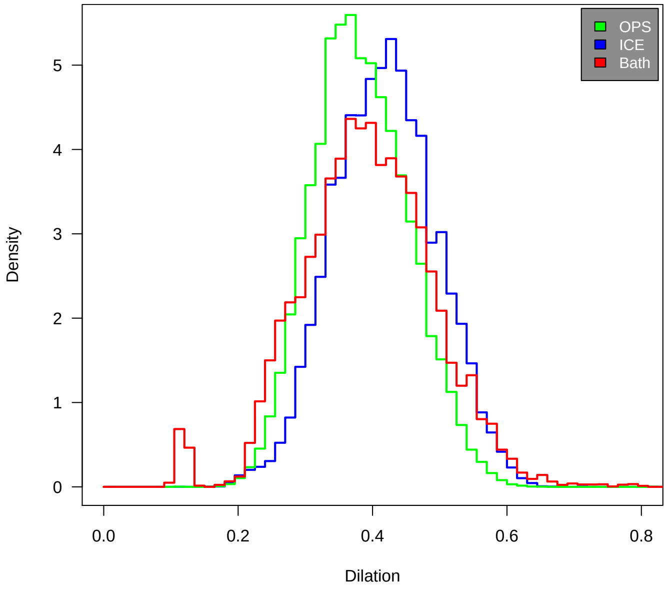


Figure 40: Estimated dilation densities for the three IREX datasets.

pupillary region itself stretches less than the immediately surrounding area in the cilliary region.

Pupil size is very difficult to control at the time of capture. It depends primarily on the pupillary light reflex (infrared light is not in the visible spectrum and does not elicit a pupillary response), and full control of lighting conditions at the time of capture is rarely guaranteed. Furthermore, several other factors are known to affect pupil size. For example, extreme emotional states (e.g. fear and pain)[28] cause pupillary dilation, as do certain drugs (e.g. alcohol, opiates)[51]. Drugs that cause pupillary constriction are known as miotics. Persons can be trained to vary the size of their pupils[22]. Thus, it is unrealistic to assume that changes in dilation can simply be eliminated at the time of capture.

The amount of measured pupillary dilation can be quantified as a simple ratio

$$D = \frac{R_p}{R_i} \tag{23}$$

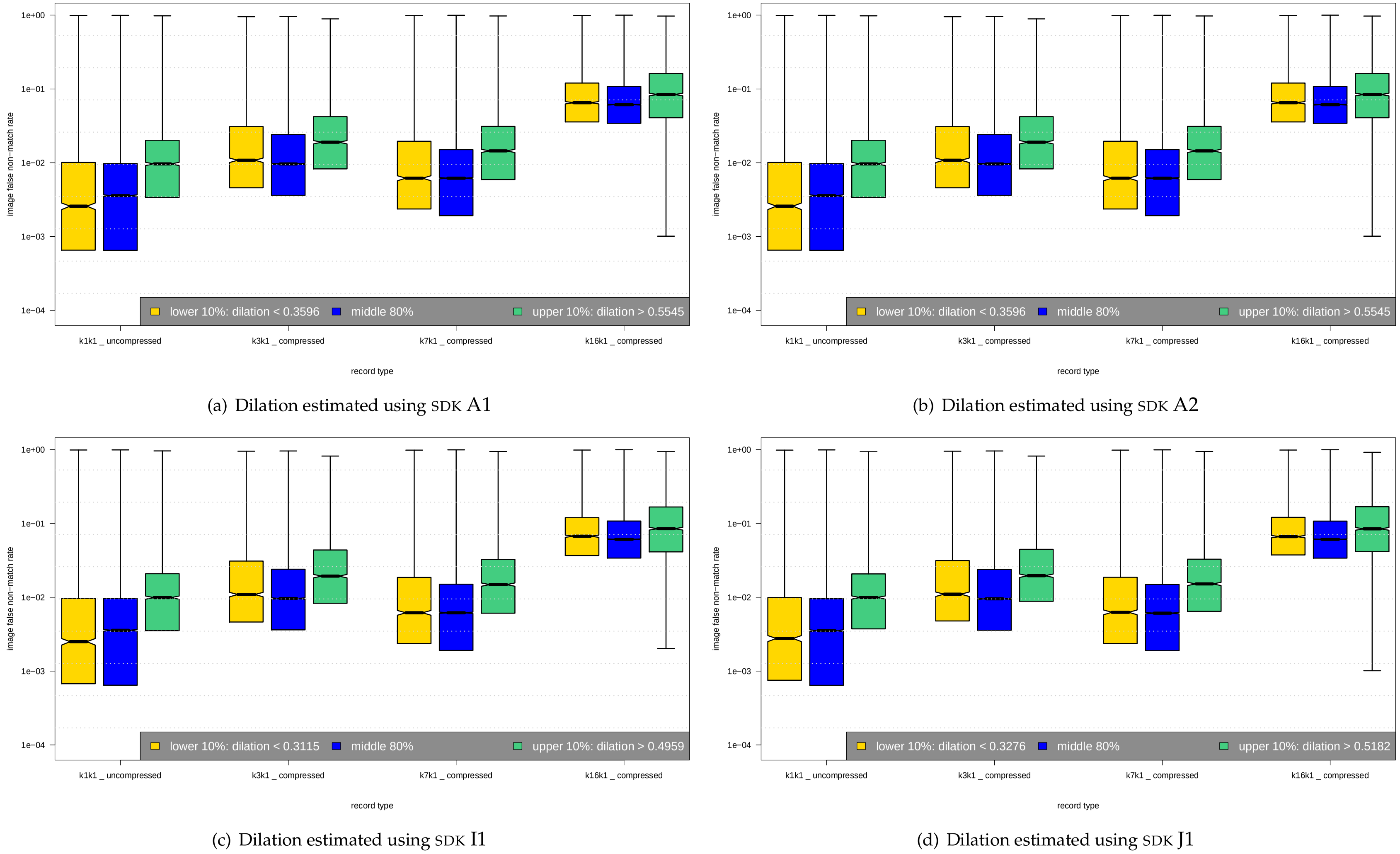
where R_p and R_i are circular estimates of the pupil and iris radii respectively. The distribution of pupillary dilations for each dataset is presented in Figure 40. These are computed from information encoded in the IREX record headers. The I1 SDK was used for this purpose. The amount of pupillary dilation ranged from .10 (lowest) to .84 (highest) among all three datasets. Values outside this range are likely to be artificially induced by drugs or other means, assuming lighting conditions are otherwise normal.

To investigate the effect of pupillary dilation on recognition accuracy, the relationship to iFMR and iFNMR (equations 11 and 13) was investigated. The ICE images were partitioned into three sets according to dilation. The high dilation set contains the 10% of images that have the dilation above 0.5227. The low dilation set contains the 10% of images with the dilation below 0.3329. The remaining 80% were placed into a third set. These boundary criteria are computed across dilation estimates from the A1, A2, I1, and J1 SDKs applied to all ICE images.

Effect on FNMR : Figure 41 shows boxplots of aggregate iFNMR for each of the three dilation partitions using shape information provided by A1, A2, I1, and J1 in the IREX record headers. The distribution of iFNMR is further broken down by record type and compression state. Figures presenting boxplots of iFNMR for each individual algorithm are available in the SUPPLEMENTAL APPENDICES. The figure shows a higher distribution of iFNMR for the high dilation set, possibly as a result of less iris area being available for comparison. This behavior diminishes when the images are JPEG2000 compressed to 2000 bytes. The difference in iFNMR is much less pronounced between the low and medium dilation sets, suggesting that perhaps only high amounts of pupillary dilation are problematic for matchers. In some cases, the distribution of iFNMR is actually lowest for the medium dilation set. It may be that the pupil border is sometimes more difficult to localize when the pupil is highly constricted.

Effect on FMR : Figure 42 presents the same information for iFMR. As with the previous figure, error rates are generally higher for the high dilation set when the iris images are uncompressed, and the separation between the three sets diminishes when the images are JPEG2000 compressed. The medium dilation set often produces a lower iFMR distribution than the low dilation set when the images are stored as compressed KIND 16 records. The variation in iFMR increases when the images are JPEG2000 compressed, as evidenced by the greater inter-quartile ranges.

The reliability of any conclusions drawn from the figures depends on the accuracy of the computed dilation ratios, which in turn depend on the accuracy of the shape information provided by the SDKs. Nevertheless, it should be noted that the dilation ratios computed for the different SDKs were generally in agreement. The low dilation sets for each of the four SDKs share 83% of the same images in common. Similarly, all four high dilation sets share 85% of the same images in common. Moreover, the most significant observation, that high pupillary dilation increases both iFMR and iFNMR (and thus the probability of getting either type of error), is true across all four SDKs.



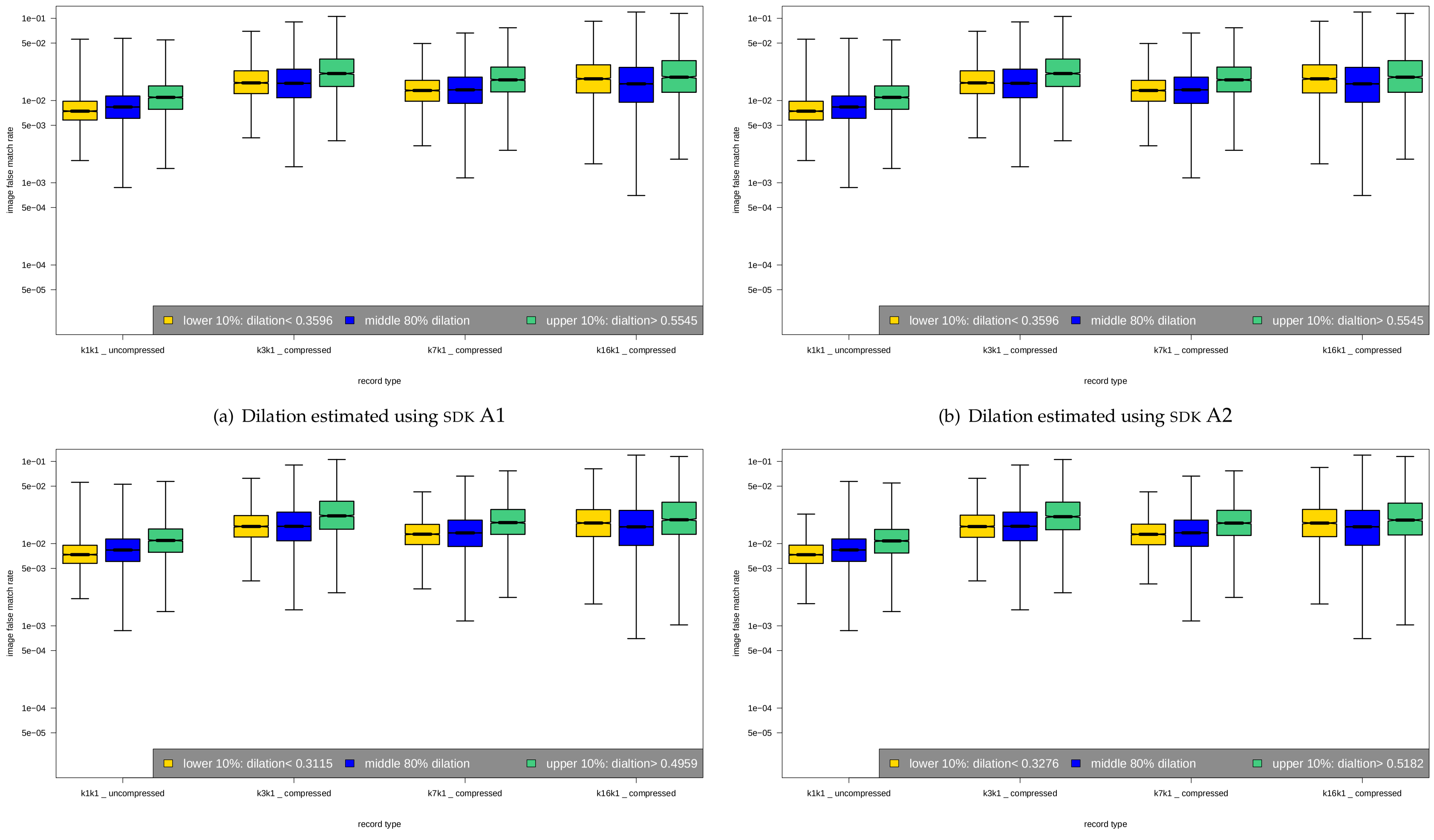
(a) Dilation estimated using SDK A1

(b) Dilation estimated using SDK A2

(c) Dilation estimated using SDK I1

(d) Dilation estimated using SDK J1

Figure 41: Distribution of aggregate iFNMR for ICE images divided into three sets according to pupillary dilation amount. The amount of dilation was computed using equation 23 and the pupil and limbus radii as reported in the IREX record headers generated by A1 (top-left), A2 (top-right), I1 (bottom-left), and J1 (bottom-right). Aggregate error rates are computed over all SDKs using an SDK dependent threshold that produces an FMR of 0.01 for comparisons between uncompressed images. iFNMR are shown on a log-scale. 10 percentile and 90 percentile dilation are shown in the legend.



(a) Dilation estimated using SDK A1

(b) Dilation estimated using SDK A2

(c) Dilation estimated using SDK I1

(d) Dilation estimated using SDK J1

Figure 42: Distribution of aggregate iFMR for ICE images divided into three sets according to pupil dilation amount. The amount of dilation was computed using eq. 23 and the pupil and limbus radii as reported in the IREX record headers generated by A1 (top-left), A2 (top-right), I1 (bottom-left), and J1 (bottom-right). Aggregate error rates, shown on a log-scale, are computed over all SDKs using an SDK-dependent threshold that produces an FMR of 0.01 for comparisons between uncompressed images. The 10-th and 90-th percentile dilations are shown in the legend.

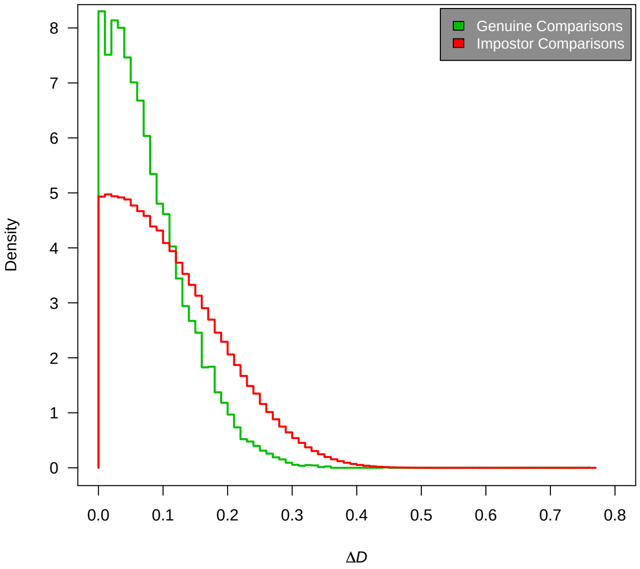


Figure 43: Estimated probability density of dilation change, ΔD, as computed using equation 24, for genuine and impostor comparisons on images in the OPS database.

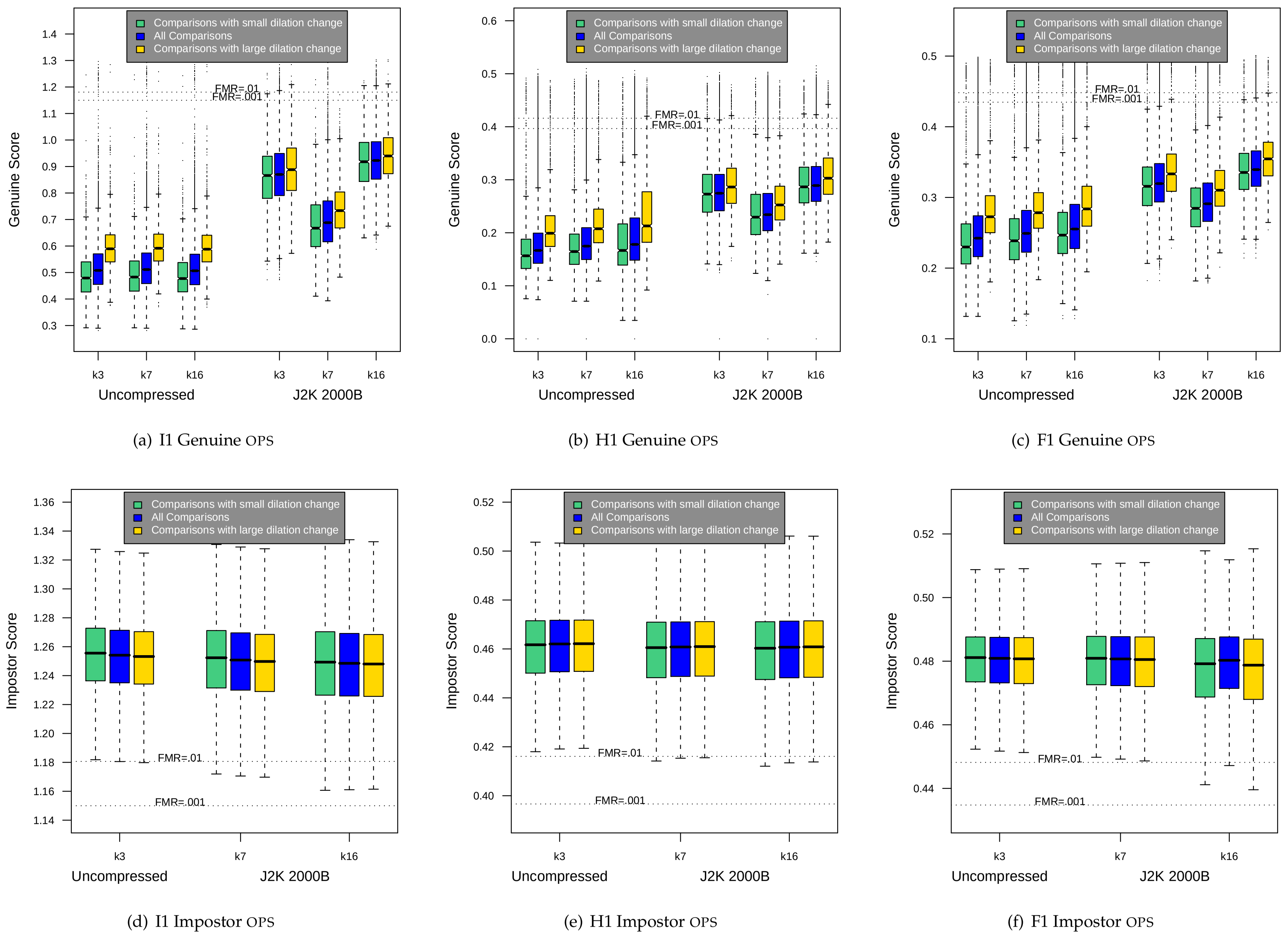
Dilation Change

When considering only a single iris image, a constricted pupil may be more preferable since it maximizes the area of the iris texture. However, when two iris images are compared, the relative change in dilation between the two is expected to be more relevant. We define the relative change in dilation between the two images being compared as:

$$\Delta D = 1 - \left(\frac{R_i^{(2)}}{R_i^{(1)}}\right)\left(\frac{R_i^{(1)} - R_p^{(1)}}{R_i^{(2)} - R_p^{(2)}}\right) = 1 - \frac{1 - D^{(1)}}{1 - D^{(2)}} \tag{24}$$

where $D^{(1)} > D^{(2)}$ is assumed without loss of generality. The first form of the equation consists of two ratios. The first is a scaling factor to compensate for possible differences in magnification. The second divides the annular width of the first iris by that of the second, which provides a measurement of the amount of radial stretching in the iris. Figure 43 shows that the probability distribution for ΔD for genuine and impostor comparisons. Note that the two distributions differ. In particular, the likelihood of the comparison being between samples of the same iris decrease as ΔD increases. This is significant because it offers additional discriminating information on the comparison that could be integrated into an iris matching algorithm. To understand this, assume a given comparison between OPS images, and assume that it is equally likely to be either a genuine or impostor comparison. If the comparison produces a ΔD of 0.1, then the odds are approximately 2 to 1 that it is an impostor comparison. If the comparison produces a ΔD of 0.4, then the odds are 10 to 1 that it is a genuine comparison. In this respect, the amount of pupillary dilation could be referred to as a *soft-biometric* [35] since it offers some useful information about an individual, but lacks the distinctiveness to sufficiently differentiate any two individuals.

A large change in dilation is likely to make two images of the same iris appear less similar. A matching algorithm that is unable to compensate will produce a higher comparison score. To determine if this effect is present, comparisons were separated into two partitions: those with $\Delta D > 0.16$ (i.e. the 90-th percentile); and those with $\Delta D < 0.02$ (i.e. the 10-th percentile). Figure 44 compares the score distributions for the two subsets for various algorithms and record types. The figure also shows results when the enrollment images are JPEG2000 compressed to 2000 bytes. Figures for all IREX algorithms are available in the IREX SUPPLEMENTAL appendices.



(a) I1 Genuine OPS

(b) H1 Genuine OPS

(c) F1 Genuine OPS

(d) I1 Impostor OPS

(e) H1 Impostor OPS

(f) F1 Impostor OPS

Figure 44: Effect of dilation change on the genuine distribution (top) and impostor distribution (bottom) for three IREX algorithms: I1 (left), H1 (middle), and F1 (right). Results are stated over the 16320 image pairs of the OPS dataset.

Some common trends were observed. Most notably:

- Genuine scores were increased for the high ΔD set and decreased for the low ΔD set. The trend occurred for all record formats, and also when the enrollment images were JPEG2000 compressed to 2000B.
- Impostor scores were not appreciably affected by dilation change.
- The dilation effect has less of an effect on the genuine score when the enrollment images are JPEG2000 compressed. This conclusion is based on the observation that the three dilation partitions tend to be more separated when the enrollment images are uncompressed.
- The dilation effect was noticeably more pronounced for H1's KIND 16 records. Error rates were, in general, slightly higher for H1's KIND 16 records, indicating that the effect may be due to image processing that occurred during the conversion to polar format.

That the dilation effect is less pronounced when the enrollment images are JPEG2000 compressed (as evidenced by the separation of the three partitions, which is less with enrollment image compression) is interesting. A possible explanation is that the changes in dilation primarily affect the finer details on the iris texture, which get blurred when an image is JPEG2000 compressed. In other words, the alterations in the texture due to dilation change might get thrown away when the image is compressed. In the current analysis, only the enrollment images were compressed; if both images were compressed, the dilation effect might be further reduced.

Previous studies on how variations in pupil size affect comparison scores have restricted their analyses to un-occluded iris images. This is to eliminate eyelid occlusion as a possible confounding factor. When the pupil is dilated, more of the iris texture is pushed to the periphery, where it is more likely to be occluded by eyelids or eyelashes. Removing occlusion is ideal when the goal is to analyze the precise deformations that occur on the iris surface. However, acquiring un-occluded images requires full cooperation from the subjects, which is atypical in operational systems. Most operational images contain some amount of occlusion, since subjects are usually not instructed to open their eyes wide and stare directly at the iris camera. If the goal is to determine how dilation change affects operational images, perhaps as part of a quality measure, then the effect of dilation change on matching accuracy should be measured in the presence of occlusion.

Since eyelid occlusion and dilation change both affect the comparison score, and both are correlated with each other, it may be that some of the dilation effect is just a hidden occlusion effect, or vice versa. Therefore, we use logistic regression to determine the effect that each has on the comparison score independently of the other

$$P(Y(d)=1) = (1+e^{-z})^{-1} \tag{25}$$

where Y is the response variable related to the comparison score, d, and z is the linear predictor defined as

$$z = \beta_0 + \beta_1\,\Delta D_i \;+\; \beta_2\,\Theta_i \tag{26}$$

where β_i are the regression coefficients, and Θ_i and ΔD_i are the independent variables representing the amount of occlusion and dilation change respectively. The coefficients β_1 and β_2 determine the respective contributions to the response variable. In our analysis, we relate occlusion and dilation change to the genuine distribution by using FNMR as the response variable.

$$Y(s) = H(d-\tau) \tag{27}$$

	FNMR=.10			FNMR=.05		
	β_0 (intercept)	β_1 (dilation)	β_2 (occlusion)	β_0 (intercept)	β_1 (dilation)	β_2 (occlusion)
A1				-5.91	3.62	10.13
A1 (j2k 2000B)				-5.19	4.99	9.91
A2				-4.08	2.10 (p=0.0058)	3.14
A2 (j2k 2000B)				-4.14	1.93 (p=0.0064)	5.10
B1	-4.03	9.18	8.17	-4.64	8.15	7.61
B1 (j2k 2000B)	-3.75	5.64	8.6	-4.48	4.60	8.78
B2	-3.98	8.58	8.29	-4.51	7.01	7.53
B2 (j2k 2000B)	-3.78	5.65	8.74	-4.43	4.70	8.41
C1	-3.04	4.91	3.79	-3.51	3.04	2.81
C1 (j2k 2000B)	-3.02	2.71	5.21	-3.44	1.61 (p=0.0100)	3.30
D1	-3.51	8.67	4.65	-3.9	5.96	3.81
D1 (j2k 2000B)	-3.39	7.76	4.41	-3.89	5.77	3.84
D2	-4.34	6.43	12.25	-4.60	4.29	9.84
D2 (j2k 2000B)	-4.29	5.75	12.31	-4.61	4.19	9.96
E1	-3.47	7.84	5.01	-4.16	6.85	5.12
E1 (j2k 2000B)	-3.37	7.23	4.71	-4.08	6.35	4.95
E2	-3.5	8.12	5.04	-4.20	7.08	5.30
E2 (j2k 2000B)	-3.43	7.44	4.97	-4.17	6.69	5.36
F1	-3.02	4.73	3.75	-3.27	2.08	1.47 (p=0.0016)
F1 (j2k 2000B)	-3.05	3.82	4.65	-3.40	2.10	2.56
G1	-3.94	4.31	10.83	-4.43	2.91	9.54
G1 (j2k 2000B)	-3.75	2.89	10.35	-4.39	2.35	9.56
G2	-3.96	4.01	11.14	-4.49	2.69	10.04
G2 (j2k 2000B)	-3.79	3.16	10.45	-4.51	2.99	9.99
H1	-3.25	5.43	5.11	-3.56	3.43	2.95
H1 (j2k 2000B)	-3.14	2.38	6.31	-3.68	2.42	4.62
H2	-3.36	6.16	5.43	-3.62	3.69	3.22
H2 (j2k 2000B)	-3.17	2.46	6.47	-3.72	2.30	5.00
I1	-4.15	10.99	7.73	-5.05	10.29	8.92
I1 (j2k 2000B)	-3.23	3.83	6.07	-4.14	4.25	6.79
I2	-4.18	11.17	7.79	-5.06	10.53	8.8
I2 (j2k 2000B)	-3.26	4.00	6.16	-4.12	4.05	6.78
J1	-3.71	7.08	7.32	-4.37	6.30	7.06
J1 (j2k 2000B)	-3.55	2.49	9.10	-4.28	2.26	8.84
J2	-3.83	7.73	7.82	-4.64	7.31	8.15
J2 (j2k 2000B)	-3.70	2.82	9.91	-4.45	2.11	9.87

Table 20: Results of fitting the logistic regression model, which show the effect of dilation change and eyelid occlusion on the probability of a false non-match. Results are presented for each algorithm, and at two operating points. P-values are listed when $p > .001$.

where d is the genuine score, τ is a user-defined score threshold, and $H(.)$ is the step function of equation 2. The regression models the probability of a false non-match given the explanatory dilation and occlusion variables. The response variable was chosen because it focuses on the upper tail of the genuine distribution (i.e. the probability of a false non-match at a threshold that is closer to what might be used by operational systems). A logit model was chosen because it assumes a binomial response variable and correctly restricts probabilities to the range $(0, 1)$.

If dilation change exerts only a small independent effect on the genuine score, then its coefficient will be small relative to the occlusion coefficient. The regression model was implemented in *R*[2]. P-values were computed for each predictor using the chi-squared test, which is analogous to the ANOVA test for linear regression. Results of fitting the model to the OPS database are presented in Table 20. Occlusion values were estimated from C1 records using the method described in the previous section. Cells for the A1 SDK are left blank when the desired FNMR could not be attained at any threshold.

The following conclusions are made based on the results of fitting the model:

- ▷ Eyelid occlusion and changes in dilation both have an adverse effect, but the amount varies depending on the algorithm. The occlusion amount tends to be a stronger predictor of the probability of a false non-match, but in many cases dilation change is more significant. In particular, eyelid occlusion is a stronger predictor for A1, A2, D2, G1, and G2, while dilation change is a stronger predictor for D1, E1, and E2.

- ▷ Compressing the enrollment image reduces the effect of dilation change. At an FNMR of 0.05, using JPEG2000 compressed enrollment images reduced the dilation change coefficient for 15 of the 18 algorithms. This behavior was noted previously.

- ▷ The effect of eyelid occlusion and dilation change is significant at both thresholds. This establishes that the two factors affect the upper tail of the genuine distribution as well as the median and quartiles.
- ▷ The genuine distribution for D2 is significantly more affected by eyelid occlusion than that for D1.
- ▷ Eyelid occlusion and dilation change always had a statistically significant effect on the probability of getting a false non-match (assuming a significance level of .01).

Research into pupillary dilation as a predictor of performance is relatively new. Most studies on the subject of iris sample quality have focused on iris occlusion, gaze direction, and blur. This section confirms that changes in pupillary dilation do affect recognition accuracy. In fact, pupillary dilation affects the genuine score as much as eyelid occlusion in some cases. Deformation models are unlikely to entirely eliminate the dilation effect since the iris is multi-layered, and the precise deformations of the iris surface vary for different people. Therefore, there may be a benefit to storing multiple images of the same iris at different pupillary dilations. Another alternative is to store additional information describing the specific deformations that occur. Error rates could be reduced by exploiting the fact that a large dilation change indicates a decreased likelihood of the two iris images originating from the same person.

8.9.3. Displacement of the iris and pupil centers

Proper iris feature extraction requires an accurate localization of the pupillary boundary, which is known to be neither perfectly circular, nor precisely concentric with the limbus boundary (see Figure 45). Although the simplest method of specifying the pupil boundary is to assume a circular shape, studies have shown that fitting with the slightly more flexible ellipse can improve matching accuracy[58, 56]. More elaborate methods of delimiting the pupillary boundary include contour-based methods such as snakes[16, 7], and shape-fitting models such as the circular and elliptical Fourier Series[54]. The IREX specification supported encoding of circular and elliptical boundaries, and the use of Freeman Chain Codes (FCC) to define arbitrary closed paths.

The precise shape of the pupillary boundary varies for different individuals, but it is unlikely to offer potential as a discriminator because its shape can vary over time, and in response to dilation. In particular, Wyatt[65] discovered that the pupil tends to become more circular as it dilates, and less circular as a person ages. The same study discovered that the boundary shape is fairly consistent during a single session, but can differ across sessions.

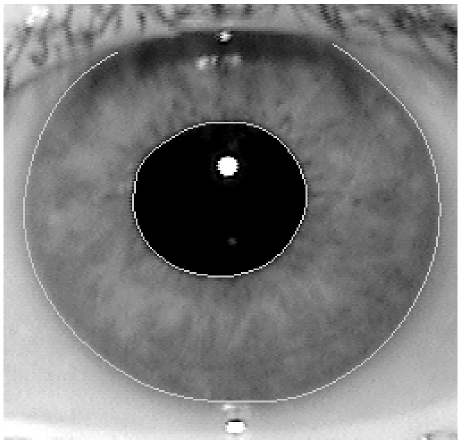


Figure 45: ICE image demonstrating that the pupil boundary is neither perfectly circular nor precisely concentric with the limbus boundary. The center of the pupil is shifted up and to the left relative to the center of the limbus boundary. The pupil and limbus boundaries are outlined in white.

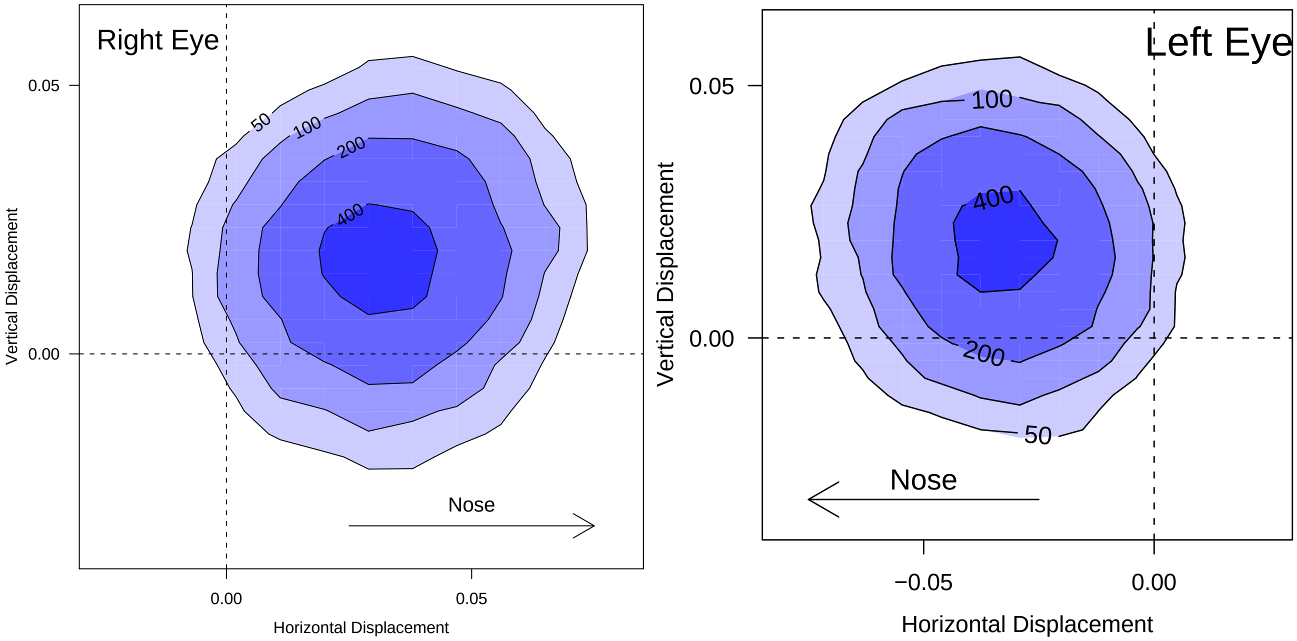


Figure 46: Displacement of pupil center relative to limbus center for left and right eyes. The right figure shows displacement for the left eye. Vertical and horizontal displacement was measured as the pixel disparity between the pupil and limbus centers divided by the pixel radius of the limbus boundary.

The pupil center (i.e. center of mass) tends to be above and nasal (i.e. closer to the nasal cavity) relative to the center of the limbus boundary. Figure 46 shows the displacements according to *I1*, where displacement is measured as the absolute difference of the pupil and limbus centers divided by the limbus radius. According to *I1*, the average displacement is 0.04 (± 0.02) above and 0.02 (± 0.01) nasal. Barry et al.[6] estimated the average at 0.4 mm above, and 0.25 mm nasal, which is consistent with current results if we assume a constant iris diameter of 11.8 mm. The density plots also indicate that deviations from the mean location are approximately radially symmetric.

The amount of pupil displacement is correlated for left and right eyes from the same person when geometry information produced by I1 is used. The Pearson correlation coefficient is 0.411 for OPS images, where the paired data is the distance between the pupil and limbus centers for each pair of eyes. The correlation coefficient implies only a weak linear relationship, although a significance test using 16318 degrees of freedom produces a two-tailed p-value less than 10^{-6}. Figure 47 shows the density plot for the difference in pupil displacement for left and right eyes from the same person. Positive values indicate the left eye pupil is more displaced. For simplicity, the figure does not separate vertical and horizontal displacement and only considers the absolute amount of displacement between the pupil and limbus centers. The density plot concentrates about the origin, indicating that neither pupil center has a propensity to be more displaced than the other. The large amount of variation suggests that the amount of pupil displacement in one eye is not very indicative of the amount of pupil displacement in the other eye.

To determine if pupil displacement affects the comparison score, comparisons were separated into two sets: 1) those with a low amount of displacement, and 2) those with a high amount displacement. We define the amount of displacement for a comparison as the average displacement for both images involved in the comparison. The displacement amount was considered high if it was above the 90th percentile for genuine comparisons and low if it was below the 10th percentile. Figure 48 compares the score distributions for both sets for various algorithms and record formats. Figures for all eighteen algorithms are available in the IREX SUPPLEMENTAL appendices.

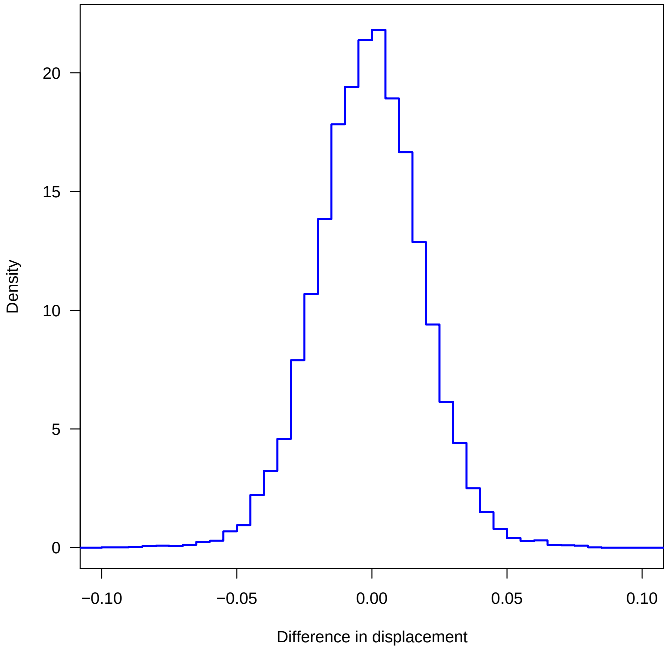


Figure 47: Distribution of the difference in pupil displacement between left and right eyes from the same person. Positive values indicate the left eye pupil has greater displacement.

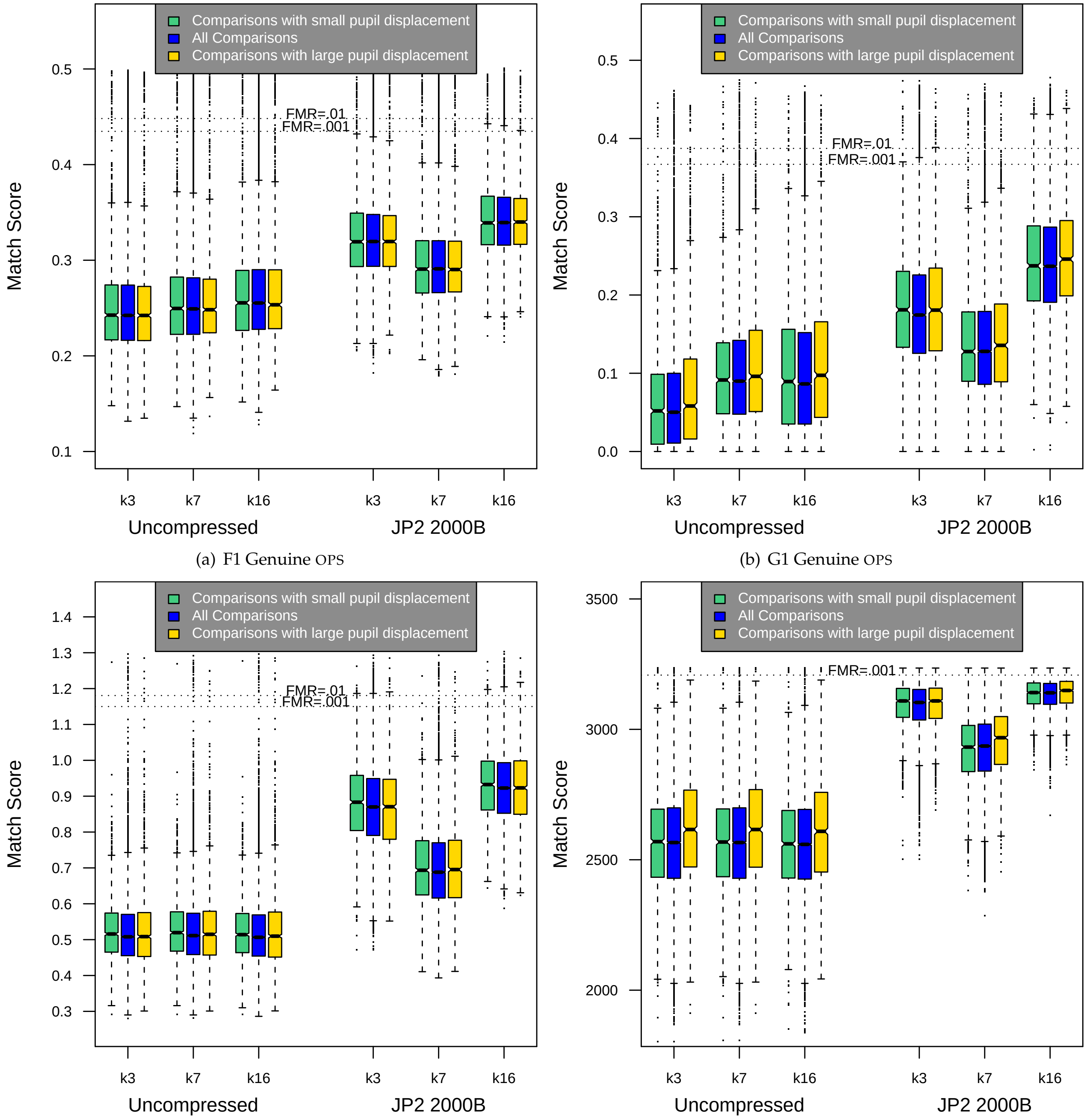
The amount of pupil displacement has a small but discernible effect on the genuine score for some algorithms, most notably J1, where higher displacements tend to generate higher genuine scores. There are several possible explanations for this effect. An algorithm that assumes concentricity of the pupil and iris centers would fail to produce an accurate segmentation of the iris features when the pupil center is significantly displaced from the iris center. Large pupil displacements might also make it more difficult for algorithms to compensate for deformations of the iris due to dilation change. The most likely reason, however, is that large displacements are indicative of an iris image that is ill-suited for matching due to some other reason. For example, a contact lens can distort the appearance of the iris and translate the apparent position of the pupil. Further investigation into the cause of the pupil displacement effect is recommended.

8.10. IMAGE ERROR RATE ESTIMATES

For any given image, its image-specific error rates will vary depending on the algorithm. That is iFMR and iFNMR will be expected to vary when computed using comparison scores of different algorithms. This variation is addressed here.

Figure 49 shows the scatter plot of aggregate iFNMR vs. iFMR for ICE images. Each plot corresponds to one of the record types: uncompressed KIND 1, JPEG2000 compressed (to 2000 bytes) KIND 3, JPEG2000 compressed (to 2000 bytes) KIND 7, and JPEG2000 compressed (to 2000 bytes) KIND 16. The dotted black lines mark the FMR and FNMR at the operating threshold of FMR = 0.001. The spread of points demonstrates image performance variability. Images can be categorized according to their level of difficulty.

- ▷ CLEAR ICE These are image for which the aggregate iFMR is less than the nominal FMR, and aggregate iFNMR is less than the nominal FNMR. These images occupy the lower left quadrant of the plots and may be considered *easy* to recognize.
- ▷ BLACK ICE These images occupy the upper right quadrant. They are the most challenging images of the ICE dataset since their image error rates are higher than the nominal error rates indicated by the dotted black line.
- ▷ BLUE GOATS Images in the top left quadrant have iFNMR $>$ FNMR (τ) and iFMR $\leq$ FMR (τ). These are more



(a) F1 Genuine OPS

(b) G1 Genuine OPS

(c) I1 Genuine OPS

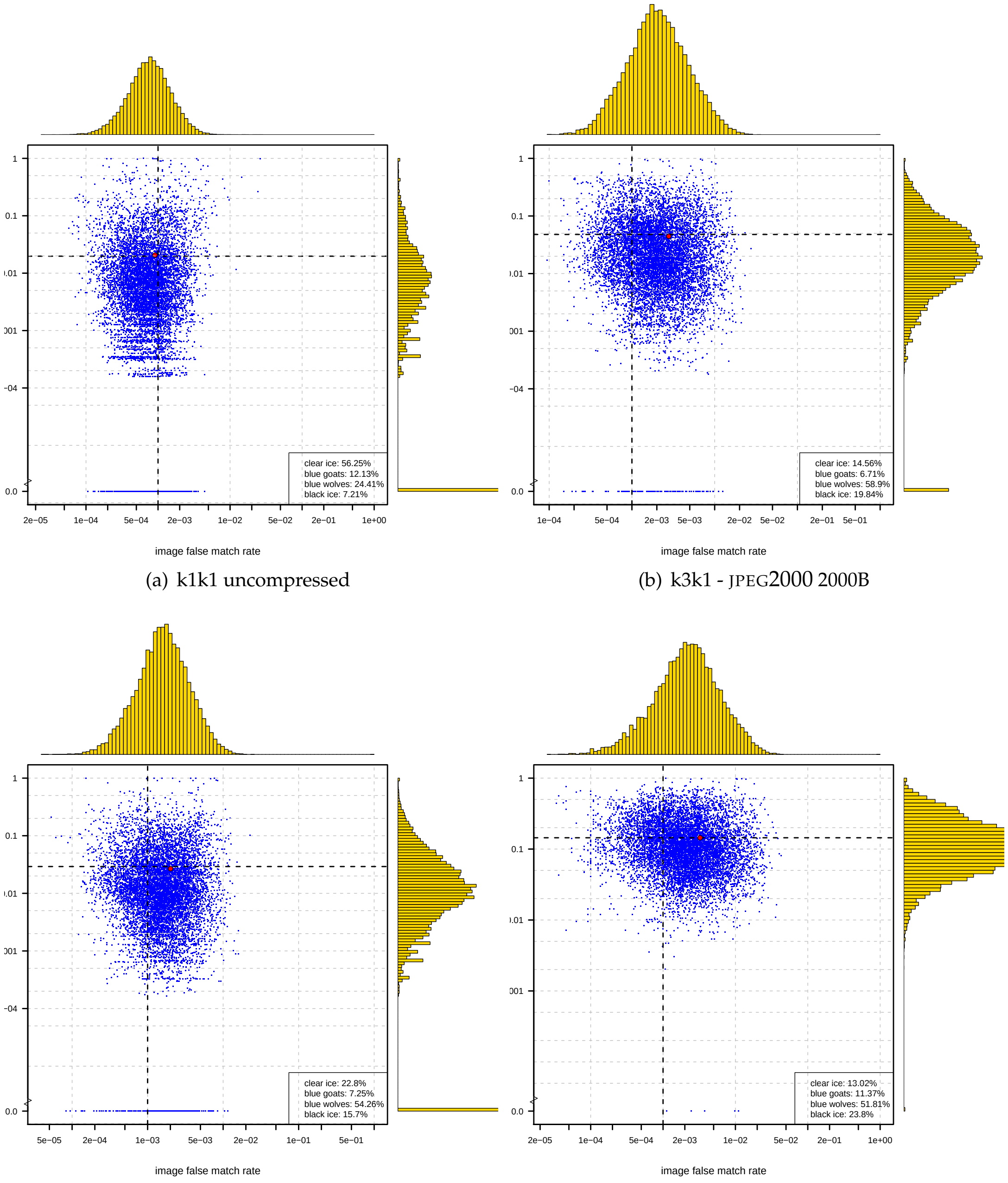
(d) J1 Genuine OPS

Figure 48: Effect of pupil displacement on the genuine distribution for various algorithms. Results are stated over the 16320 image pairs of the OPS dataset.

frequently falsely rejected but do not attract false matches.

▷ BLUE WOLVES Images residing in the bottom right quadrant have iFMR $> \text{FMR}(\tau)$ and iFNMR $\leq \text{FNMR}(\tau)$. These images are implicated in more false matches and are generally easy to match.

An interesting observation is variation in aggregate image error rates across the four plots of figure 49. Aggregate image



(a) k1k1 uncompressed

(b) k3k1 - JPEG2000 2000B

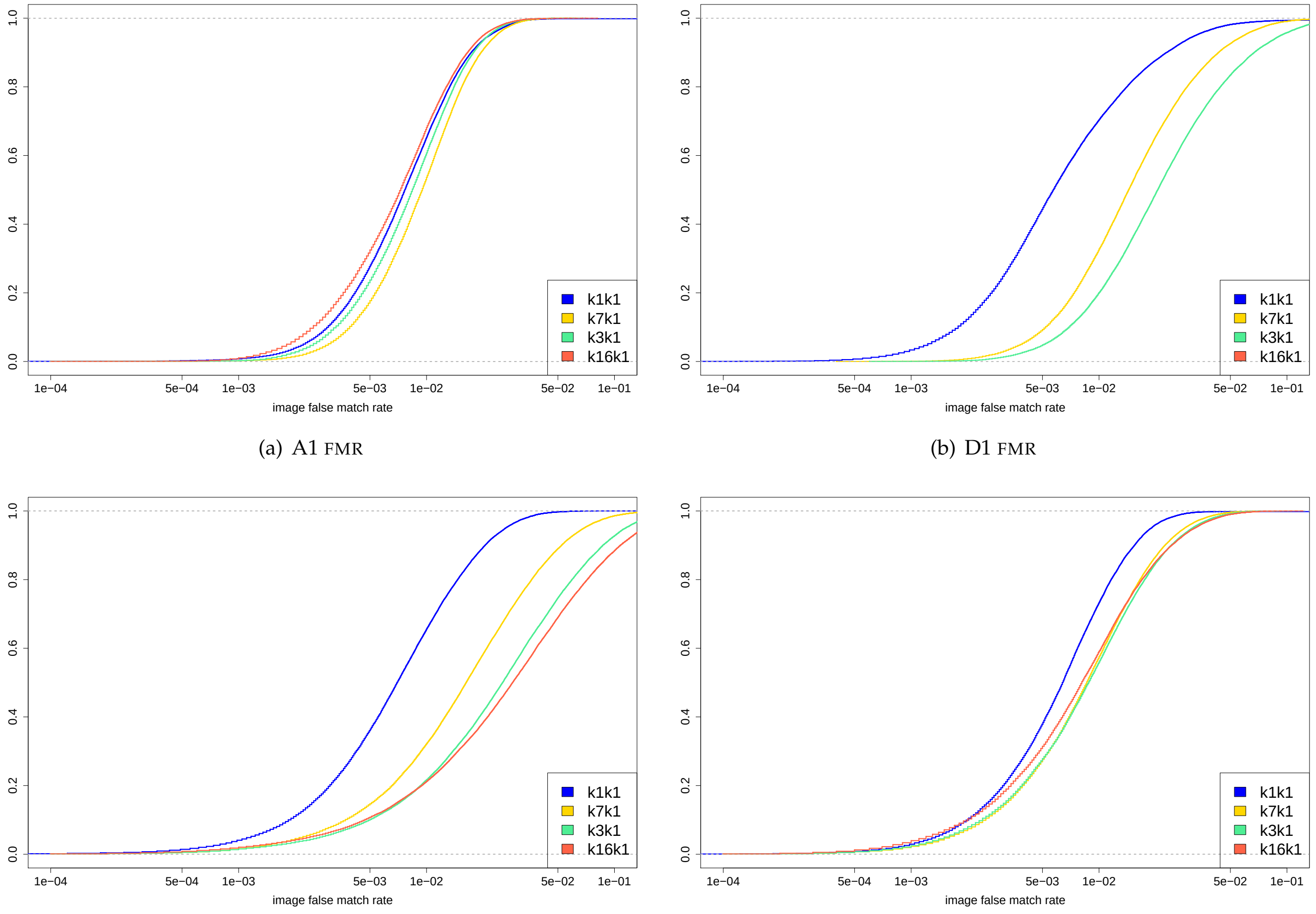
(c) k7k1 - JPEG2000 2000B

(d) k16k1 - JPEG2000 2000B

Figure 49: Image FNMR vs. image FMR for 31415 images of the ICE dataset for various KINDS. The image errors are computed at score threshold corresponding to global FMR = 0.001 and aggregated across all SDKs. The red dot in each plot is the center of mass (computed as the mean value) of the image errors. The black dotted lines corresponds to aggregate error rate of the system. Image false non-match (and false match) probability densities are plotted on the top (and left side). The legend shows percentage of CLEAR ICE, BLUE GOAT, BLUE WOLF and BLACK ICE images.

errors for JPEG2000 compressed KIND 16 show a wider spread in image FMR than the image FNMR , while for JPEG2000 compressed KIND 7 the spread is wider in image FNMR.

Figure 50 shows the cumulative distribution functions (CDFs) of aggregate iFMR for four leading implementations. Ideally, all images have equal likelihoods for error, in which case the CDFs would be step functions. In practice some images



(a) A1 FMR

(b) D1 FMR

(c) I1 FMR

(d) J1 FMR

Figure 50: Cumulative distribution functions of the image false match rates for selected SDKs. Each CDF of iFMR is computed for each KIND, at thresholds corresponding to overall FMR = 0.01 (on uncompressed KIND 1 instances). The KIND 1 records are uncompressed, all others are JPEG2000 compressed to 2000 bytes. The CDFs of both image error rates are included in the IREX SUPPLEMENTAL appendices for all SDKs.

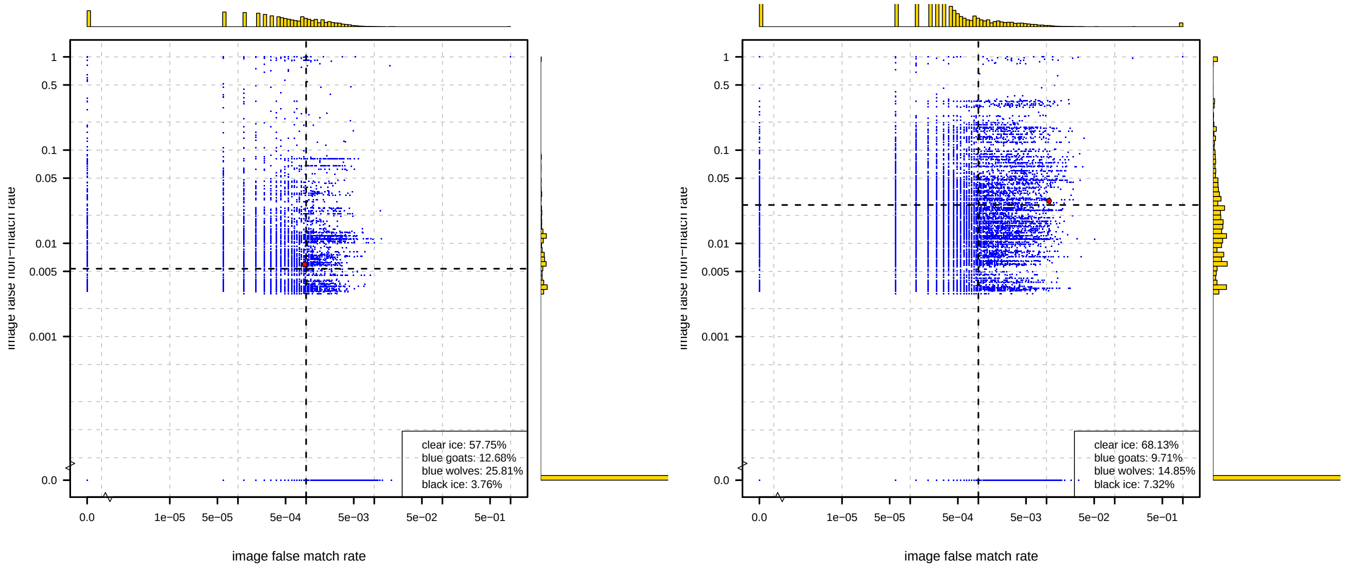
produce fewer errors, and some more. A matching algorithm that constrains the worst-case behavior is preferable. The notable observations are:

- ▷ While the threshold is set to FMR = 0.01[50], it is generally the case that more than 50% of images have image false match rates less than 0.01.
- ▷ For some SDKs, the application of lossy compression causes iFMR distributions to move to the right *and* to broaden. The problematic area is the upper tails of the iFMR distributions where, especially for KIND 16 and KIND 3 images, there are higher numbers of false matches.
- ▷ The CDFs of the A1 implementation are most similar, and are more bounded at the top end.

8.10.1. Dependence on comparison algorithm

Figure 49 is useful in assessing the difficulty of an image (and so the dataset if proper summarization is performed) because image error rates are computed across a diverse set of comparison algorithms. However, it does not reveal

[50] The iFMR is computed over 16320 comparisons, so it is necessary to use a high threshold to get observable errors.



(a) I2 k1k1 uncompressed

(b) E1 k1k1 uncompressed

Figure 51: Image FNMR vs. image FMR for 31415 images of the ICE dataset for two SDKs. Image errors are computed at the threshold that gives global FMR = 0.001 for each SDK . The red dot in each plot is the center of the mass (computed as the mean value) of the image errors. The black dotted lines correspond to aggregate error rate of the system. Image false non-match (and false match probability) density is plotted on the top (left side). The relative spread of the image errors suggests comparison algorithm I1 is more robust to image variation than comparison algorithm E1. The legend shows percentage of clear ice, blue goat, blue wolf and black ice images.

image error variations among comparators. An image with a high FMR or FNMR for one comparison algorithm, might result in low image error rates when a different comparison algorithm is applied. In other words, a difficult to process and recognize image for one algorithm, might be easy for another algorithm. To examine the dependence of image errors on comparison algorithms, we computed image error for each comparison algorithm. Figure 51 shows image error rate for SDK I2 and SDK E1. The ideal case is when a matching algorithm produces constant false matches and false non-match (with a very small spread) for any image regardless of its underlying properties and quality. However, the relatively wide spread and heavy tails of the distributions in Figure 51(b) suggest that it is not the case in the real world. It can be seen that the spread of the blue cloud (and the histogram of image errors) is different for the two algorithms. That means an image that presents wolf-like behavior to one algorithm (i.e. its iFMR is larger than overall FMR of the system at the operating threshold) might be a sheep when matched by a different comparison algorithm (i.e. not causing any false matches). This is a strong argument for fusion of matching algorithms. Image error plots for each SDK are in the IREX SUPPLEMENTAL appendices.

The cause of the performance variations is interesting. The extent to which combinations of image (or user) covariates and/or matching algorithm might cause this is, to our knowledge, unreported. Another interesting problem is whether, regardless of the matching algorithm, there are wolves (or goats) at large (i.e. users that account for disproportional share of the overall FMR or FNMR). We intend to pursue both subjects in the future. For now, we focus on investigating a) how to quantify the level of difficulty of an image (and so a dataset) and b) the ability of matching algorithms to produce comparison scores that are robust to variation in image (or user) covariate. In other words, if an algorithm operating at a fixed threshold could maintain a relatively constant false match rate or false non-match rate regardless of image (or user) properties.

9. CONSIDERATIONS FOR OVERALL PERFORMANCE

The performance data contained in this report and its accompanying IREX SUPPLEMENT is intended to guide developers, standards makers, and users. This last category might include policy makers, system integrators, designers and analysts. This discussion suggests that no single best algorithm exists because the definition of *best* is multiply defined and should, in any case, be considered in response to operational requirements.

The following text lists criteria that will have more or less importance in a given application.

- **False rejection** In many performance tests, particularly offline technology tests, there is an explicit emphasis on false non-match rate as the primary performance measure. Thus, reports will summarize a DET characteristic by tabulating FNMR at a fixed FMR [51]. This is clearly a valid and valuable discriminator between algorithms or technologies for access control-like applications in which false rejection performance is the primary performance indicator. This is reasonable because enrolled users of the system are inconvenienced by high FRR, and because impostors are often rare (i.e. prior probability is small).

  But in non-access control applications, the primacy of FRR is less justified. For example, in a watchlist open-set identification scenario[1] the enrolled persons may very rarely appear and the primary performance metric is the rate at which un-enrolled passers-by are falsely matched. This too necessitates control and stability of FMR. On the "FNMR at FMR = 0.0001" metric the best three IREX providers are:

  - For uncompressed KIND 1 images: I, B, J (In Figures 10, 11, and 12, for the OPS, BATH and ICE datasets respectively, the leading two are I and B. Thereafter J is is in third place except for the BATH dataset, where A and D perform better).
  - For KIND 3 images JPEG2000 compressed to 0.2 bits per pixel: I, B, J (see column means in the penultimate row of Table 12).
  - For KIND 7 images JPEG2000 compressed to 0.2 bits per pixel: I, B, J (see column means in the penultimate row of Table 13).
  - For KIND 16 images JPEG2000 compressed to 2000 bytes: I, J, A (see column means in the penultimate row of Table 14).

  However, the above assumes that a security objective (expressed by some low FMR requirement) can be met via appropriate setting of a threshold. While this can always be achieved by setting a conservative threshold, it requires some calibration of the false match rate (either by the manufacturer or by a testing laboratory) and some assurance that the false match rate is stable across, at least, time, environmental conditions, and population. This advocates for the second performance criterion below, *impostor distribution stability*.

- **Impostor distribution stability** The stability of the impostor distribution is a very valuable property for a biometric system that is used out-of-the-box with a fixed operating threshold because it gives predictable occurrence of false matches. In one-to-many identification applications it means that the expected number of false matches produced in a search, a metric called *selectivity*[10] will be stable.

  Figures 24 and 14 (and the corresponding Figures in the IREX SUPPLEMENTAL APPENDICES) show that some algorithms produce somewhat varying impostor distributions across datasets and compression levels. Note that in advocating this as an important algorithm performance metric:

[51] Many studies refer to Type I and Type II error rates as false reject rate, FRR and false accept rate, FAR and they do so with identical meanings to the FNMR and FMR quantities used here. However, the SC37 Working Group 5 committee usefully gave different meanings to these terms to express the difference between matching error rates and transactional rates, the former being the result of comparison of two images or templates and the latter being the result of an attempt by a user to authenticate to a system (for example by retrying, or by presenting a second eye).

- We are not denying the existence of wolves and lambs in the biometric zoo[20] but instead requiring their prevalence and behavior to be stable.
- We acknowledge that while score normalization can be applied on samples of the operational data in order to stabilize the distribution, there are two problems with this approach: First is that the capability must actually be installed and maintained; and second that it is most effective in a closed operational environment in which the cameras and population are known. In federated applications in which standardized data records are collected remotely (e.g., worldwide) in a variety of circumstances (cameras, compression regimes) the ability to control impostor distributions is more difficult.

The SDKs from the following IREX providers demonstrate generally small variation in the impostor distributions: B, A, E and D (see the lines of constant-threshold in the four sets of DETs in Figure 24).

▷ **Resistance to compression** The ability of a recognition system to function after compression has been applied is vital in some applications. Compression makes both the localization and the feature extraction more difficult, because it blurs boundaries and alters the iris texture. Referring to the cropping survey results in Figure 28, the algorithms giving the smallest increases in FNMR values (i.e. conditional FNMR) are from providers A, D, B.

▷ **Localization** Instantiation of the KIND 3 record requires only an initial detection and cropping of the iris. Referring to Figure 12 the lowest average FNMR at FMR = 0.0001 values are available from the KIND 3 IREX records produced by SDKs from providers A, B and I.

For KIND 7 there is the additional task of finding the iris-sclera and iris-eyelid boundaries. In this respect the list of best providers is identical to that for KIND 3 A, B, and I.

For KIND 16 polar records, the implementation performs the pupil and iris detection tasks, choses radial and circumferential sampling rates and performs rectlinear-to-polar interpolation. For this operation providers A, J and I instantiate the most interoperable instances.

▷ **High speed** Computational cost is clearly important as a performance related variable. The relative contributions of the durations of the various functions (preparation of standard images, template generation, and matching) need to be factored into the particular application (e.g., one-to-one vs. one-to-many in a large population.) Referring to the figures of section 7.6, the SDKs from the following IREX providers have lowest computational cost: D, G, J.

10. Proposed changes to the standards

The following comments are made toward the current commmittee draft of the ISO/IEC 19794-6

▷ The KIND 16 unsegmented polar format should be deleted from the draft standard.

▷ Some version of the guidance given in Figure 52 should be inserted into the revised ISO/IEC 19794-6 standard. This content should be advanced as a informative annex and accompanied by a note that there may be some dependency on the iris camera in use.

▷ JPEG should not be used for compact records. For 1:N applications lossy compression should be discouraged and JPEG should be entirely prohibited.

▷ A conformance testing Annex should be developed as a normative Annex of ISO/IEC 19794-6 - The current project ISO/IEC 29109-6 is targeting the 2005 standard - NIST expects the new revised standard to be adopted vs. the 2005 one.

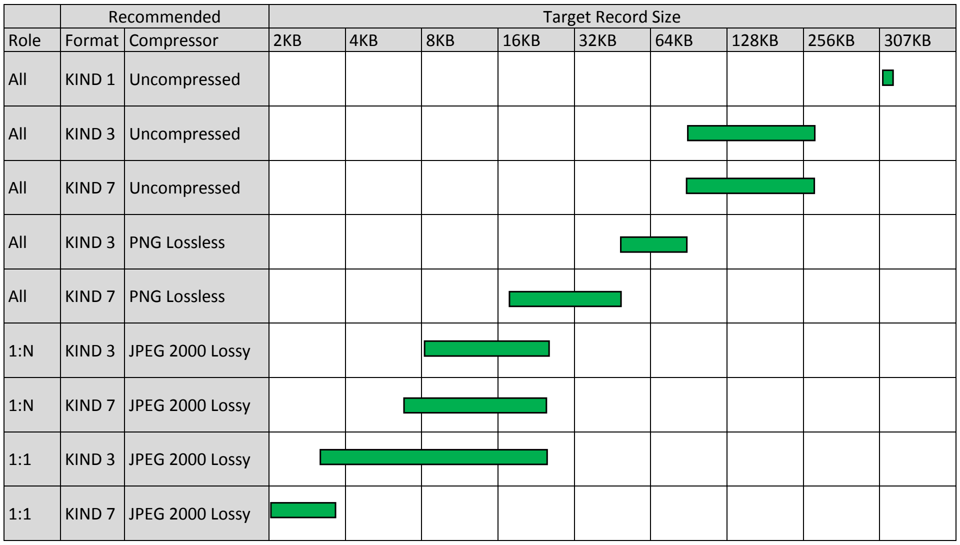


Figure 52: Provisional recommendations on the application of compression to standard image formats. The horizontal axis shows target file size in kilobytes on a logarithmic scale. Note the dependence on the intended application: Identification applications should avoid lossy compression. This guidance is provisional.

- ▷ Spectral information should be added to the standard.
- ▷ Lossless compression should be recommended as the default container for iris imagery. High Q-value JPEG compression should not be deprecated.
- ▷ The standard should acknowledge 640x480 as a de facto, but allow the possiblity that larger images may become common, and useful. The standard needs refined guidance on iris size.
- ▷ A field to encode pupil dilation should be added.
- ▷ The WG3 committee should consider whether a numerical field could be added to the standard that quantifies compression damage. The purpose is to augment score normalization techniques. Without this field a blind estimate of compression is needed.

References

[1] Face recognition as a search tool foto-fahndung. Technical report, Bundeskriminamt (BKA), Thaerstrasse 11, 65193, Wiesbaden, Germany, February 2007.

[2] The r project for statistical computing, July 2009. `http://www.r-project.org/`.

[3] Working Group 1. Standing Document 2 Harmonized Biometric Vocabulary. Technical report, ISO/IEC JTC1 SC37 N1248, November 2005.

[4] Working Group 5. *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*. JTC1 :: SC37, international standard edition, August 2005. http://isotc.iso.org/isotcportal.

[5] P. Akaishi, L. Neves, C. Polegato, and A. Cruz. The effect of darkness on the upper eyelid position. *Investigative Ophthalmology and Visual Science*, 43, 2002.

[6] J. Barry and A. Backes. Limbus versus pupil center for ocular alignment measurement with corneal reflexes. *Investigative Ophthalmology and Visual Science*, 38(12):2597–2607, November 1997.

[7] N. Barzegar and M. Moin. A new approach to iris localization in iris recognition systems. In *IEEE/ACS International Conference on Computer Systems and Applications, 2008, AICCSA 2008*, pages 516–523, March 2008.

[8] A. Bazin and T. Mansfield. An investigation of minutiae interoperability. In *Proc. Fifth IEEE Workhop on Automated Identification Advanced Technologies*, June 2007. AUTO-ID 2007, Alghero Italy.

[9] W. Becker and A. Fuchs. Lid-eye coordination during vertical gaze changes in man and monkey. *Journal of Neurophysiology*, 60(4):1227–12252, October 1988.

[10] Ruud Bolle, Jonathan H. Connell, Sharath Pankanti, Nalini K. Ratha, and Andrew W. Senior. *Guide to Biometrics*, pages 94–96. Springer, 2004.

[11] J. P. Campbell. Speaker recognition: A tutorial. In *Proc. of the IEEE*, volume 85, 1997.

[12] CASIA. Specification of casia iris image database - version 1.0. Technical report, Chinese Academy of Sciences, March 2007. http://www.nlpr.ia.ac.cn/english/irdsirisdatabase.htm.

[13] C. Cavallotti and L. Cerulli, editors. *Age-Related Changes of the Human Eye*. Humana Press, 1 edition, April 2007.

[14] J. Daugman. New methods in iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(5):1167–1175, October 2007.

[15] John Daugman. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):21–30, January 2004.

[16] John Daugman. Probing the uniqueness and randomness of iriscodes: Results from 200 billion iris pair comparisons. *Proc. of the IEEE*, 94(11):1927–1935, Nov 2006.

[17] John Daugman. Response to nistir-7440 and frvt/ice2006 reports. Technical report, University of Cambridge, 2008. http://www.cl.cam.ac.uk/users/jgd1000/Response_2_NIST_7440.pdf [on June 22, 2009].

[18] John Daugman and Cathryn Downing. Effect of severe image compression on iris recognition performance. Technical report, University of Cambridge, Computer Laboratory, May 2007. Submitted as UK contribution to SC37 WG3, doc. N2125.

[19] John Daugman and Cathryn Downing. Effect of severe image compression on iris recognition performance. *IEEE Transactions on Information Forensics and Security*, 3(1):52–61, October 2008.

[20] 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 *Proceedings of 5th International Conference of Spoken Language Processing*, ICSLP 98, Sydney, Australia, 1998. Paper 608 on CD-ROM.

[21] G.R. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance. In *Proc. Fifth Int'l Conf. Spoken Language Processing (ICSLP)*, pages 1351–1354, Sydney, Australia, 1998.

[22] I. Ekman, A. Poikola, M. Mäkärainen, T. Takala, and P. Hämäläinen. Voluntary pupil size change as control in eyes only interaction. In *Proceedings of the 2008 Symposium on Eye Tracking Research and Applications*, pages 115–118. Association for Computing Machinery, 2008.

[23] J. Campbell et al. *ILO Seafarers' Identity Documents Biometric Testing Campaign Report*. International Labour Organization, Geneva, 2005. http://www.ilo.org/public/english/dialogue/sector/papers/maritime/sid-test-report2.pdf.

[24] P. J. Phillips et al. Overview of the multiple biometrics grand challenge. Technical report, National Institute of Standards and Technology, www.nd.edu/ kwb/PhillipsEtAlICB_2009.pdf [on June 24, 2009], 2008.

[25] P. Grother, M. McCabe, C. Watson, M. Indovina, W. Salamon, P. Flanagan, E. Tabassi, E. Newton, and C. Wilson. Performance and interoperability of the incits 378 fingerprint template. Technical report, National Institute of Standards and Technology, March 2006. Published as NIST Interagency Report 7296.

[26] P.J. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. Pattern Anal. Mach. Intelligence (PAMI)*, 29(4):531–543, April 2007.

[27] Feng Hao, John Daugman, and Piotr Zielinski. A fast search algorithm for a large fuzzy database. *IEEE Transactions on Information Forensics and Security*, 3(2):203–212, 2008.

[28] E. Hess. Attitude and pupil size. *Scientific American*, April 1965.

[29] A. Hicklin, C. Watson, and B. Ulery. The myth of goats: How many people have fingerprints that are hard to match. Technical report, National Institute of Standards and Technology, 2005. Interagency Report 7271.

[30] J. Hill. Analysis of senile changes in the palpebral fissure. *Transactions of the American Ophthalmological Society*, 95(1):49–53, 1975.

[31] K. Hollingsworth. Sources of error in iris biometrics. Master's thesis, University of Notre Dame, 2008.

[32] International Biometric Group. *Independent Testing of Iris Recognition Technology*, May 2005.

[33] R. W. Ives, R. P. Broussard, L. R. Kennell, and D.L. Soldan. Effects of image compression on iris recognition system performance. *Journal of Electronic Imaging*, 17, 2008. http://link.aip.org/link/?JEIME5/17/011015/1.

[34] Robert W. Ives, Bradford L. Bonney, and Delores M. Etter. Effect of image compression on iris recognition. In *Instrumentation and Measurement Technology Conference (IMTC)*, Ottawa, Canada, May 2005.

[35] A. Jain, S. Dass, and K. Nandakumar. Soft biometric traits for personal recognition systems. In *Proceedings of Int. Conf. on Biometric Authentication*, pages 731–738, July 2004.

[36] D. Kim. The unsegmented polar format. Technical report, Iritech Corp, November 2007. JTC001-SC37-N-2296 US NB Contribution on Compact Iris Format.

[37] X. Liu, K. Boyer, and P. Flynn. Experimental evaluation of iris recognition. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, 2005*, June 2005.

[38] L Ma, T. Tan, Y. Wang, and D. Zhang. Efficient iris recognition by characterizing key local variations. *IEEE Transactions on Image Processing*, 13(6):739–745, June 2004.

[39] Li Ma, Tieniu Tan, Senior Member, Yunhong Wang, and Dexin Zhang. Personal identification based on iris texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:1519–1533, 2003.

[40] Li Ma, Tieniu Tan, Yunhong Wang, and Dexin Zhang. Efficient iris recognition by characterizing key local variations. *IEEE Transactions on Image Processing*, 13(6):739–750, 2004.

[41] L. Masek. Recognition of human iris patterns for biometric identification. Master's thesis, The University of Western Australia, 2003. www.csse.uwa.edu.au/ pk/studentprojects/libor/LiborMasekThesis.pdf.

[42] Libor Masek and Peter Kovesi. Matlab source code for a biometric identification system based on iris patterns. Technical report, The School of Computer Science and Software Engineering, The University of Western Australia, 2003.

[43] Stefan Matschitsch1, Martin Tschinder1, and Andreas Uhl. *Comparison of Compression Algorithms Impact on Iris Recognition Accuracy*, pages 232–241. Lecture Notes in Computer Science. Springer, Berlin / Heidelberg, August 2007.

[44] R. Michael McCabe. Nist special publication 500-271: American national standard for information systems data format for the interchange of fingerprint, facial, and other biometric information part 1. Technical report, April 2007. ANSI/NIST ITL 1-2007.

[45] D. M. Monro. University of bath iris image database. Technical report, University of Bath, 2008. http://www.bath.ac.uk/elec-eng/research/sipg/irisweb/ [on June 22, 2009].

[46] Donald M. Monro, Soumyadip Rakshit, and Dexin Zhang. Correction to "dct-based iris recognition". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):Not in print, 2007.

[47] Donald M. Monro, Soumyadip Rakshit, and Dexin Zhang. Dct-based iris recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):586–595, 2007.

[48] S. S. Phang. *Investigating and developing a model for iris changes under varied lighting conditions*. PhD thesis, Queensland University of Technology, 2007.

[49] P. Jonathon Phillips, Kevin W. Bowyer, and Patrick J. Flynn. Comments on the casia version 1.0 iris data set. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1869–1870, 2007.

[50] P. Jonathon Phillips, W. Todd Scruggs, Alice J. O'Toole, Patrick J. Flynn, Kevin W. Bowyer, Cathy L. Schott, and Matthew Sharpe. Frvt 2006 and ice 2006 large-scale experimental results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(1), 2009.

[51] W. Pickworth, P. Welch, J. Henningfield, and E. Cone. Opiate-induced pupillary effects in humans. *Methods and Findings in Experimental and Clinical Pharmacology*, 11(12):759–763, December 1989.

[52] N. Poh, S. Bengio, and A. Ross. Revisiting doddington's zoo: A systematic method to assess user-dependent variables". *Multimodal User Authentication (MMU)*, 13(1):234–778, 2003.

[53] Hugo Proena and Lus A. Alex. Iris recognition: An analysis of the aliasing problem in the iris normalization stage. In *International Conference on Computational Intelligence and Security*, volume 2, pages 1771–1774, November 2006.

[54] Soumyadip Rakshit and Donald M. Monro. An evaluation of image sampling and compression for human iris recognition. *IEEE Transactions on Information Forensics and Security*, 2(3-2):605–612, 2007.

[55] S. Read. *Corneal topography and the morphology of the palpebral fissure*. PhD thesis, Queensland University of Technology, 2006.

[56] W. Ryan, D. Woodard, A. Duchowski, and S. Birchfield. Adapting *Stadburst* new techinque for elliptical iris segmentation. In *IEEE Second International Conference on Biometrics: Theory, Applications, and Systems*, 2008.

[57] S. Schuckers, N. Schmid, A. Abhyankar, V. Dorairaj, C. Boyce, and L. Hornak. On techniques for angle compensation in nonideal iris recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 37(5):1176–1190, October 2007.

[58] C. Sreecholpech and S. Thainimit. Circular and elliptical modeling for pupil boundary in closed-up human eye images. In *Proceedings of ECTI-CON 2008*, pages 445–448, 2008.

[59] R. Sutton. Doddington's zoo and fingerprint system security, 2007. Noblis Lecture .

[60] Jason Thornton, Marios Savvides, and B. V. K. Vijaya Kumar. A bayesian approach to deformed pattern matching of iris images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):596–606, 2007.

[61] F. Vihlen and G. Wilson. The relation between eyelid tension, corneal toricity, and age. *Investigative Ophthalmology and Visual Science*, 24(10):1367–1373, 1983.

[62] J. L. Wayman. Multifinger penetration rate and roc variability for automatic fingerprint identification systems, 1999. National Biometric Test Center.

[63] Z. Wei, T. Tan, and Z. Sun. *Lecture Notes in Computer Science*, volume 4642/2007, chapter Nonlinear Iris Deformation Correction Based on the Gaussian Model, pages 780–789. Springer Berlin / Heidelberg, 2007.

[64] M. Wittman, P. Davis, and P. J. Flynn. Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus. In *CVPRW '06: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop*, page 33, Washington, DC, USA, 2006. IEEE Computer Society.

[65] H. Wyatt. The form of the human pupil. *Vision Research*, 35(14):2021–2036, July 1995.

[66] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. Accepted for future publication.

[67] D. Zhang, D.M. Monro, and S. Rakshit. Eyelash removal method for human iris recognition. In *ICIP06*, pages 285–288, 2006.