
IREX VI

Temporal Stability of Iris Recognition Accuracy

NIST Interagency Report 7948

P. Grother J. R. Matey E. Tabassi G. W. Quinn M. Chumakov

<http://dx.doi.org/10.6028/NIST.IR.7948>

Information Access Division
National Institute of Standards and Technology



July 11, 2013

Executive Summary

Purpose: This work was conducted determine whether iris recognition accuracy decreases with the time lapsed between collection of initial enrollment and recognition images. More specifically, it seeks to quantify accuracy changes associated with any permanent changes to the iris and its proximal anatomy. This study is intended to quantify natural ageing effects in a healthy population; medical conditions and injuries can rapidly and severely affect recognition, so these are out of scope.

Background: Stability is a required definitional property for a biometric to be useful. Quantitative statements of stability are operationally important as they dictate re-enrollment schedules e.g. of a face on a passport. The ophthalmologists who filed initial patents on iris recognition posited the iris to be “extremely stable” over “many years” but that “features which do develop” do so “rather slowly”[31]. A further patent held that irises have “texture of high complexity, that prove to be immutable over a person’s life”[21]. This view held until several recent empirical studies suggested otherwise. Those studies, and ours, were motivated to check the veracity of the 1994 patent’s assertion that an enrolled iris can be viable over decades. Two studies, using separate iris image collections from the University of Notre Dame, reported a large increase in false rejection rates[8, 29]. The studies made attempts to account for several possible causes of the observed ageing, but could not conclude that the iris texture itself was changing. Their results, however, were widely reported[59, 24, 3] with statements such as “irises, rather than being stable over a lifetime, are susceptible to ageing effects that steadily change the appearance over time”[33]. A further study, however, identified pupil-dilation[27] as the primary causal variable. Operational iris systems have identified individuals over periods up to 10 years[5] and 7 years[6].

Conclusions: Using two large operational datasets, we find no evidence of a widespread iris ageing effect. Specifically, the population statistics (mean and variance) are constant over periods of up to nine years. This is consistent with the ability to enroll most individuals and see no degradation in overall recognition accuracy. Furthermore, we compute an ageing rate for how quickly recognition degrades with changes in the iris anatomy; this estimate suggests that iris recognition of average individuals will remain viable over decades. However, given the large population sizes, we identify a small percentage of individuals whose recognition scores do degrade consistent with disease or an ageing effect. These results are confined to adult populations. Additionally, we show that the template ageing reported in the Notre Dame studies is largely due to systematic dilation change over the collection period. Pupil dilation varies under environmental and several biological influences, with variations occurring on timescales ranging from below one second up to several decades. Our data suggests that the natural constriction of pupil size over decades does not necessitate re-enrollment of a well enrolled iris. The ISO/IEC 19795-1 testing standard defines ageing as any increase in error rates with time. This definition is imprecise because temporary changes due to environment (e.g. lighting) or user behavior (e.g. blinking) might yield elevated error rates without any change in the biometric source itself. Dilation has been suggested to be part of ageing under that definition. Instead we assert that ageing stems from irreversible changes to the anatomy, primarily the iris texture. Dilation should not be considered part of ageing because it varies stochastically and can be mitigated - some iris cameras normalize dilation by shielding or by active illumination. Corneal shape changes have been suggested as influential on iris recognition too, but their effect has never been quantified.

Technical Summary

Approach: Given observations of the kind shown in Figure 1, we consider ageing to be a time-series problem[16] requiring analysis of dissimilarity scores between enrollment and recognition images as a function of elapsed time. This derives from Czajka’s suggestion[16] that permanent changes in the iris texture would give a non-stationary genuine score distribution, such that ageing would generally produce non-linear and even non-monotonic changes to score statistics. We restrict our analysis to the linear case, using longitudinal mixed-effects regression models to compute rates-of-change of each eye’s scores, and the population-average thereof.

Detection of long-term trends is complicated by short-term stochastic variations inherent in acquiring digital images from a live analog anatomic source. Here

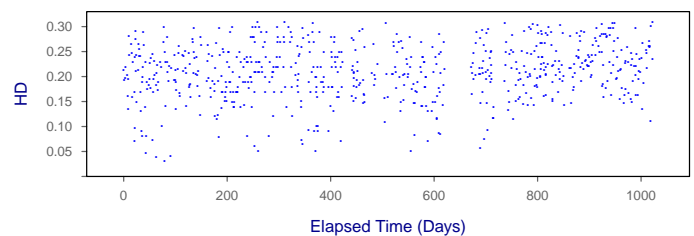


Figure 1: **Varying recognition outcomes:** Hamming distances (HD) from comparison of an enrolled iris image with images collected in a physical access control system over three years. Small HD values indicate high similarity between iris images. Note that high values are often followed by low values, and a trend is difficult to discern.

variations in quantities such as pupil dilation, gaze angle, focus, and eyelid position manifest themselves as the “measurement noise” evident in Figure 1.

Also, we apply eight contemporary iris recognition algorithms to the images used in the Notre Dame studies. We relate pupil dilation and exposed iris area measures to recognition outcomes. Additionally, we review other published studies, formulate recommendations for conduct of biometric ageing studies and for the mitigation of ageing effects in operational systems.

Results: The primary results of this investigation are as follows.

- ▷ **Rate of accuracy change:** Our best estimate of iris recognition ageing is derived from a 7876 person subset of an operational registered traveler deployment[5] who have used the system on forty or more occasions over at least four years and up to nine years. The estimate is a population-average from a linear mixed-effects regression model. It states Hamming distances increase at a rate of $(8 \pm 2) \times 10^{-7}$ per day. [Sec. 4.2](#)

We put this rate into context in three ways. First, this rate, if sustained, would mean that an average person would exhibit an increase $\Delta HD = 0.003$ over a ten year period. Such an increase is an order of magnitude smaller than the variation ($\sigma = 0.04$) that an average individual exhibits in routine usage of an iris camera. If sustained this rate would cause consistent identification failure only over time periods longer than a human lifetime. Second, this age-related change is far smaller than the change measured for the use of different cameras for enrollment and search ($\Delta HD = 0.05$). Third, this change is smaller than that observed between frequent and infrequent users of a border-crossing system ($\Delta HD = 0.02$). Our rate-of-change estimate must be considered provisional pending application of refined statistical techniques to larger and richer data sets, generalization to other recognition algorithms, better modelling of dilation and inter-camera comparisons, and consideration of bilateral ageing in both eyes.

A low population-average ageing rate does not necessarily mean that some persons do not age more quickly. Our individual-specific measurements on this dataset show symmetric variation around the population mean - roughly equal numbers of subjects have increasing and decreasing score trajectories. The rate-of-change estimate varies with the camera used, and with the subset of the population used. Additionally, given the population size, we expect some instances of ocular disease to cause some variation. [Sec. 4.2.1.](#)

- ▷ **Field operational data:** When state-of-the-art recognition algorithms are applied to 3.5 million images collected over a period of about 6 years from 622,464 subjects the genuine score distributions are stable and show no evidence of an widespread upward template-ageing trend. [Sec. 4.4.](#)
- ▷ **Re-analysis of Notre Dame collections:** We use eight recent commercial recognition algorithms to broadly reiterate the empirical time dependence reported for both the 2004-2008 and 2008-2010 Notre Dame (ND) collections. However, three observations in the second collection motivate an analysis that explains the observed variation. We note a) heterogeneity in the false non-match errors, particularly the errors are concentrated in fewer than one third of the individuals, b) that pupil dilation varies, particularly across the three collection semesters, with widespread pupil constriction in 2010, and c) that accuracies over consecutive one year intervals differ. The dilation change is key. When amounts proportional to the dilation difference between two images are subtracted from the observed iris dissimilarities, the ageing effect substantially disappears or is at least difficult to detect given the residual presence of other influential factors in the images. [Sec. 4.3.1](#)
- ▷ **New methodologies for biometric ageing studies:** In the domain of biometric performance testing, this study innovates in at least two ways. a) It includes extensive visualizations of individual-specific measurements. These are advanced as exploratory analysis[93] precursors to more quantitative methods. b) The adoption of mixed-effects regression models from the medical literature should be directly relevant to other longitudinal biometric studies, particularly face ageing where time dependence is clearly evident. This arises from the models’ ability to handle imbalanced, irregularly sampled, individual-specific observations, with temporal correlations. Additionally, the random-effects afford modeling of “biometric zoo”[25] heterogeneities across individuals. [Sec. 4.2.1](#)
- ▷ **Guidance on ageing measurement:** The report lists technical considerations for organizations engaged in biometric ageing studies. These include a recommendation that ageing studies should adopt the tight image acquisition controls used in many ophthalmological studies, and in some iris recognition studies. [Sec. 5.1](#)

Guidance on ageing mitigation in biometric systems: We advance a set of considerations for mitigation of temporal effects in operational systems. Among these are that system owners should log all pertinent performance data - scores, qualities, timings - to support retroactive analysis. More directly, while ageing can often be mitigated by collection of a replacement enrollment sample, ee caution against doing this in an un-attended process. [Sec. 5.2](#)

Classification of biometric ageing effects: We propose a classification for effects leading to longitudinal variation in biometric accuracy. CLASS A variations are caused by systematic effects that can be remediated by the system operator; these effects, e.g. failed illuminators in cameras, would typically affect many users. CLASS B variations are subject-specific and can be remediated by modifying the behavior, skill, cooperation or physical condition of the subject. For example, a subject could be asked to open his eyes more widely. CLASS C variations are subject-specific and cannot be remediated without a substantial design change to the system. An example would be to use shorter exposure times to mitigate motion symptoms of Parkinson's disease. These changes are separate from the anatomical biometric information source itself. CLASS D is similarly subject-specific and not easily remediated, and which are related to the biometric information source, in this case the iris texture. Variations of this type are irreversible and violate the permanence requirement for a biometric. This paper attempts to quantify the effects of CLASS D changes on iris recognition. [Sec. 5.3.](#)

Acknowledgements

The authors would like to thank the sponsors of this activity. These are the Criminal Justice Information Systems division of the Federal Bureau of Investigation, the U.S. VISIT office in the Department of Homeland Security (DHS), and the Science and Technology Directorate of DHS.

The NIST authors are especially indebted the Canada Border Services Agency (CBSA) whose invaluable dataset was provided to NIST and comprises the main contribution of this report. Author Michael Chumakov is employed by CBSA. Additional thanks go to those individuals with foresight enough to log the internal details of the operational use of, by some measures, the world's largest iris recognition deployment.

Our gratitude goes to James Wayman and Vince Stanford who reviewed this paper, to Kevin Bowyer of Notre Dame for his encouragement in starting this work, and to Jonathan Phillips at NIST for describing and facilitating access to the Notre Dame datasets. The authors are grateful also to Dan Potter of Scitor Corporation for perspectives and insights, and to Terry Waters for ophthalmological advice. Thanks go to John Daugman for noting the existence of arcus senilis. Similarly to Svetlana Shchegrova of AOptix Inc. for candid discussions on the topic. Thanks go to John Yap of the FDA for directing our attention to longitudinal analysis techniques.

The authors would also like to thank the United States Department of Defense for their support and collaboration.

Finally, the authors are grateful to the iris recognition vendors who provided their state-of-the-art recognition engine prototypes for use in this effort. Use of high accuracy and operationally deployable algorithms is essential to the relevance and robustness of the results.

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.

Release Notes

All IREX related reports, drafts, announcements and news items may be found on the homepage <http://iris.nist.gov/irex>.

Appendices: This report is accompanied by a number of appendices which present exhaustive results on a per-algorithm basis. These are machine-generated and are included because the authors believe that visualization of such data is broadly informative and vital to understanding the context of the report.

Typesetting: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable L^AT_EX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.

Graphics: Many of the Figures in this report were produced using Deepayan Sarkar's Lattice package[83] running under R, the capabilities of which extend beyond those evident in this document.

Contact: Correspondence regarding this report should be directed to PGROTHER at NIST dot GOV or JAMES dot MATEY at NIST dot GOV.

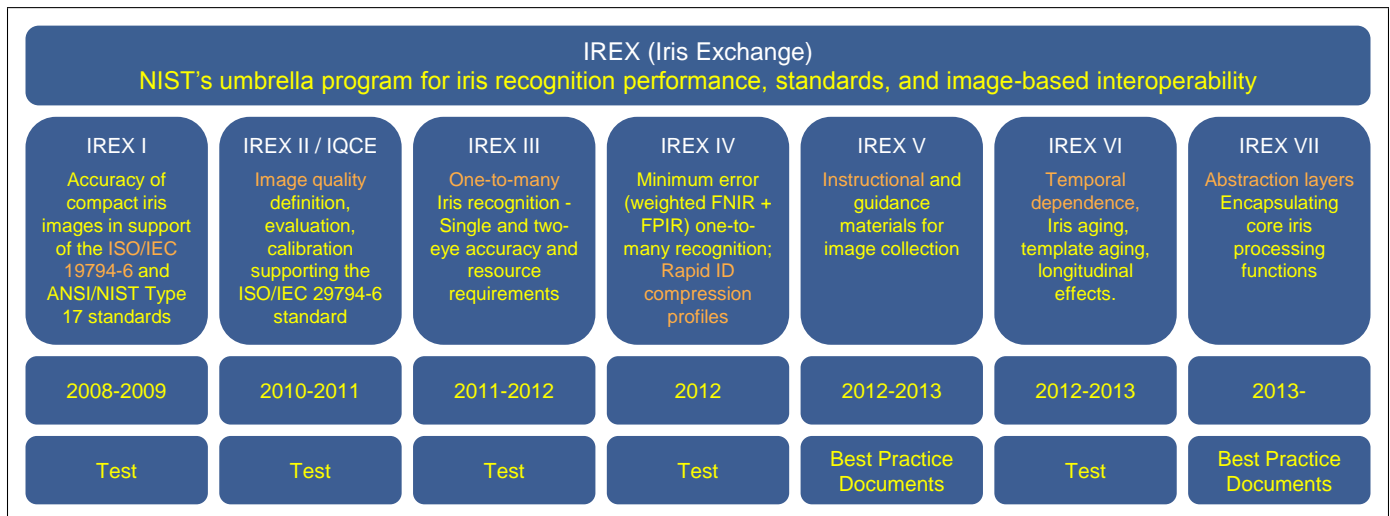
Call for data: The authors are very enthusiastic to extend the ageing analyses presented in this paper, particularly by considering other datasets. We therefore explicitly request members of the community to make available images or recognition results from those images. NIST would be a willing recipient for appropriate data. The academic community would offer considerable analytic insights if such data were openly provided to them.



The Iris Exchange (IREX) Program

In 2008 NIST established the IREX program to give quantitative support to iris recognition standardization, development and deployment. The activities that have been conducted under IREX so far are described below. All IREX-related reports, drafts, announcements and news items may be found on the homepage <http://iris.nist.gov/irex>.

- ▷ **IREX I:** The 2009 IREX I evaluation, which tested the efficacy of leading commercial and university algorithms on the specialized image formats proposed for the ISO/IEC 19794-6[22] iris image data interchange standard. IREX I also established viable limits for standardized image compression algorithms applied to iris images. Accuracy was measured over one-to-one comparisons.
- ▷ **IREX II:** The 2010-2011 Iris Quality Calibration and Evaluation (IQCE), which assessed the capabilities of iris image quality assessment algorithms and supported the ISO/IEC 29794-6[64] iris image quality standard by establishing metrics, reference thresholds, and ranges for various appearance, geometric and photometric properties of iris images. Accuracy was measured using one-to-one comparisons operating separately from the image quality assessment algorithm.
- ▷ **IREX III:** The 2011 IREX III activity executed the first public independent comparative evaluation of one-to-many iris recognition algorithms running on enrolled populations up to 3.9 million. It considered one- and two-eye recognition to validate results published in the academic literature that iris is a very powerful biometric.
- ▷ **IREX IV:** The 2012-2013 IREX IV activity, proposed as a direct follow-on to the IREX III study, applies recent one-to-many recognition algorithms to uncompressed iris images. This work supports development of definitive JPEG 2000 compression profiles for iris identification. This extends the IREX I work by considering the false positive demands of one-to-many, and by refining JPEG 2000's parameters. The compression profiles will be contributed to formal iris image standards[22, 97].
- ▷ **IREX V:** The 2012-2013 IREX V activity leverages lessons-learned in IREX-II/III in crafting best practice recommendations for avoiding the collection of poor iris images. These are primarily directed at operators of iris camera equipment.
- ▷ **IREX VI:** Starting with this report, this activity considers any appropriate longitudinal aspects of iris recognition.
- ▷ **IREX VII:** The forthcoming 2012 IREX VII activity is dedicated to definition of abstraction layers around iris cameras and algorithms.



Contents

EXECUTIVE SUMMARY	1
TECHNICAL SUMMARY	1
ACKNOWLEDGEMENTS	4
DISCLAIMER	4
RELEASE NOTES	4
THE IREX PROGRAM	5
1 INTRODUCTION	8
1.1 DEFINITION OF IRIS AGEING	9
1.2 SOURCES OF LONGITUDINAL CHANGE IN BIOMETRICS	10
1.3 OVERVIEW OF PRIOR IRIS AGEING STUDIES	11
2 EXPERIMENTAL DATASETS	14
2.1 BORDER CROSSING DATA OPS-XING	14
2.2 NOTRE DAME DATA	18
2.3 FIELD-COLLECTED DATA OPS-FIELD	18
3 LONGITUDINAL VARIATION IN IRIS RECOGNITION	18
3.1 OVERVIEW	20
3.2 EFFECT OF DILATION AND DILATION-CHANGE	20
4 RESULTS	23
4.1 OVERVIEW	23
4.2 RESULTS FOR OPS-XING TRANSACTIONS	23
4.3 RESULTS FOR ND RECOGNITION	33
4.4 RESULTS FOR OPS-FIELD	55
5 DISCUSSION AND CONCLUSIONS	59
5.1 CONSIDERATIONS FOR TESTS OF BIOMETRIC AGEING	59
5.2 CONSIDERATIONS FOR OPERATIONAL MITIGATION OF BIOMETRIC AGEING	60
5.3 HIERARCHY OF LONGITUDINAL EFFECTS IN BIOMETRICS	62
5.4 FUTURE WORK	62

List of Figures

1	EXAMPLE TIME SERIES	1
2	EXAMPLE OF FACE AGEING	8
3	OPS-XING SUMMARY STATISTICS	15
4	HAMMING DISTANCE VS. TIME, BY INDIVIDUAL	16
5	HAMMING DISTANCE VS. TIME, BY INDIVIDUAL AND DATASET	19
6	HD AS A FUNCTION OF SEARCH AND ENROLLMENT PUPIL DILATION	21
7	HD AS FUNCTION OF RADIAL THICKNESS CHANGE, AND DILATION	22
8	OPS-XING TIME EVOLUTION OF HD DISTRIBUTION	23
9	OPS-XING TIME EVOLUTION OF POPULATION FNMR	24
10	COMPARING LONG VS. ALL OTHER HDS	25
11	HABITUATION	26
12	OPS-XING GRADIENT PREDICTORS	31
13	ND FALSE NON-MATCH CONCENTRATION IN EYES	32
14	ND04-08 COMPARISON SCORE DISTRIBUTIONS VS. TIME	34
15	ND08-10 COMPARISON SCORE DISTRIBUTIONS VS. TIME	35
16	ND04-08 INDIVIDUAL TIME-SERIES IO2P	36
17	ND04-08 INDIVIDUAL TIME-SERIES D03P	37

18	ND08-10 INDIVIDUAL TIME-SERIES I02P	38
19	ND08-10 INDIVIDUAL TIME-SERIES D03P	39
20	ND08-10 RAW FNMR VS. THRESHOLD	41
21	COVARIATE PROGRESSION IN ND08-10	42
22	ND08-10 DILATION VS. TIME	43
23	ND04-08 DILATION VS. TIME	44
24	IMAGES FROM ND08-10 INDIVIDUAL 05455L	46
25	EFFECT OF PUPIL DILATION	47
26	IMAGES FROM ND08-10 INDIVIDUAL 05456L	48
27	ND08-10 RAW AND ADJUSTED FNMR VS. THRESHOLD	49
28	ND08-10 RAW AND ADJUSTED FNMR VS. THRESHOLD	50
29	ND08-10 RAW AND ADJUSTED FNMR VS. THRESHOLD	51
30	ND08-10 DILATION, AREA ADJUSTED TIMESERIES I02P	53
31	ND08-10 DILATION, AREA ADJUSTED TIMESERIES D03P	54
32	ND04-08 RAW AND ADJUSTED FNMR VS. THRESHOLD	56
33	TIME EVOLUTION OF FALSE NEGATIVE IDENTIFICATION RATE ON THE OPS-FIELD PARTITIONS	57
34	TIME EVOLUTION OF DISSIMILARITY ON THE OPS-FIELD PARTITIONS	58

List of Tables

1	IREX VI DATASETS	14
2	OPS-XING SUMMARY STATISTICS	24
3	OPS-XING REGRESSION RESULTS BY EYE	28



Figure 2: **Face ageing:** Images separated by 25 and 7 years respectively, show clear change in facial appearance with age. Reproduced with permission[47].

1 Introduction

The ISO/IEC 2382-37 vocabulary standard[4] requires that ‘repeatable biometric features’ can be extracted from a “biometric characteristic”, reflecting the practical need for recognition accuracy to be maintained over useful time periods. Jain et al.[44] identified the more idealized property of permanence as a definitional property. In any case, the archetype here is the face image stored in identity documents (e.g. passports). Humans understand that faces are recognizable over useful time spans - passports are most commonly issued for 5 or 10 years - and that over extended periods, the face is subject to irreversible changes that render recognition more difficult - see Figure 2. As physiologically and biochemically active systems, all biometric traits are potentially time-dependent¹ and this is a legitimate area for empirical study. Ageing itself may depend on known covariates such as age, gender, race, and on time-varying environmental factors.

Longitudinal variation of biometric comparison scores is expected because capture of biometric traits is an analog to digital conversion process that is not exactly repeatable. Samples captured on any two occasions differ because of variations in how the human presents to the system, differences due to sensor noise, differences in the imaging environment, and any change in the expressed biometric trait itself. This last aspect includes temporary and permanent changes that can occur on timescales ranging from less than one second up to decades.

The biometrics research literature is replete with studies where biometric data is collected from an individual on as few as two occasions. The first instance is regarded as an enrollment instance for incorporation into a gallery. The second instance, the probe, is compared with the first (gallery) instance to produce a comparison score. These are aggregated over a population to produce a statement of authentication accuracy of legitimate users². Figure 1 shows dissimilarity scores, in this case Hamming distances (HDs), produced when a person’s enrollment sample is compared with samples collected over a period of about three years. The considerable HD variation includes random effects measurement error inherent in the non-repeatability of the biometric capture process, and the effects of any systematic changes to the biometric trait itself. The size of such source-specific variation has rarely been reported in the scientific literature because frequent data collection is expensive and difficult to sustain. Unquantified score variation is a problem in many biometric studies which rely on population averaging over uncorrelated random effects to allow the investigator to draw conclusions about whatever systematic change has been applied (e.g. to the sensor or algorithm).

Ageing is primarily of interest because changing appearance eventually means that a person cannot be recognized against prior samples, i.e. false non-match. This effect is recognized in the ISO-IEC 19795-1 biometric testing standard [53] which says “Of particular importance when planning the test is the time interval between enrollment and the collection of verification or

¹As Benjamin Franklin noted, “When you’re finished changing, you’re finished”.

²When comparing samples, formal testing standards [53, 90] state accuracy as the false non-match rate (FNMR). It is computed from scores with lower-is-more-similar semantics, e.g. Hamming distances, as the fraction of genuine comparison scores, s_j , $j = 1 \dots N$, that are worse than (i.e. above) any threshold, τ .

$$\text{FNMR}(\tau) = \frac{1}{N} \sum_{j=1}^N H(s_j - \tau) \quad (1)$$

where H is the Heaviside step function. FNMR is the complement of the empirical distribution function. Further, it is sometimes very useful to compute the image-specific false non-match rate iFNMR as the number of occasions a single image is involved in failed comparisons[88].

identification data. Longer time intervals generally make it more difficult to match samples to templates due to the phenomenon known as *template ageing*". The standard then defines "template-ageing" as an "increase in error rates caused by time-related changes in the biometric pattern, its presentation, and the sensor". This definition in terms of recognition error rates rather than the core comparison scores reflects the operational relevance of error rates.

The standard notes that for "some modalities, performance a short time after enrollment, when the user appearance and behaviour has changed very little, is far better than that obtained weeks or months later". The ISO standard then requires "that genuine transaction data shall therefore be separated in time from enrollment by an interval commensurate with the target application" ... "at least by the general time of healing of the body part". A standardized test of access control systems[52] requires enrollment and authentication samples to be separated in time by a minimum of 7 days³.

Jain[44] also identified ubiquity and uniqueness as required properties of a biometric trait. Together these allow (most) members of a population to be accurately differentiated from all others, i.e. that false matches occur with calibrated rarity. In ageing studies, false match rate has been of secondary interest because it is defined over a population - i.e. the comparison of images from a single source against those of N people of generally different age. However, it is possible that as an individual ages, their likelihood of matching other individuals will change. This possibility appears in face recognition where younger persons are more likely to false match than older ones. The iris ageing literature has noted stability of the impostor distribution [8, 28, 29]. For these reasons this paper, rather unusually in biometrics, does not cite false match rates (FMR). This reflects the narrow focus of this paper to same-person effects.

1.1 Defintition of iris ageing

This report operates under the following definition.

iris ageing

irreversible changes to the healthy iris or neighboring anatomy that yield mated dissimilarity scores that increase monotonically with time-separation of the compared images

NOTE 1: The restriction to healthy individuals is practical. It is made because while injury and disease can lead to arbitrarily rapid and serious changes⁴, its prevalence and incidence are low or confined to certain small sub-populations⁵. The operational utility of all biometric characteristic rest on their stability in healthy individuals.

- ▷ NOTE 2: The monotonic qualifier is included to separate temporary and reversible changes from permanent and irreversible ones. Reversible changes include dilation of constricted pupils (e.g. by reducing light), or drooping of eyelids. Irreversible changes would include those to the arrangement of stroma in the iris, a change to the shape of the cornea, or a change in the limbus (as discussed in the next section).
- ▷ NOTE 3: The inclusion of the phrase "or neighboring anatomy" acknowledges that changes to the limbus might impede segmentation, or that corneal shape changes can alter iris appearance via optical refraction.
- ▷ NOTE 4: The scores "increase" because changes to the iris are expected to lead to reduced similarity (because of non-negativity of distance metrics).
- ▷ NOTE 5: This definition implies the use of a recognition algorithm. This is done because recognition outcomes are what matter operationally. The algorithm here is assumed not to change because it is assumed that, given sufficient understanding of ageing processes, algorithm developers could mitigate their effects.
- ▷ NOTE 6: Aspects of this definition may be applicable to other biometric modes.

³The standard, for its purposes, also imposes a maximum of 90 days for revisit transaction.

⁴For iris, see for example results pre- and post cataract-surgery[23], and diseases affecting segmentation[58, 92]. Disease affects other modalities too: For face, Bells palsy[65]; For fingerprints, palmar eczema, dishidrosis, amputation, arthritis, and acute damage due to manual labor[26].

⁵For example, pterygium, an iris-occluding growth of blood vessels across the iris, has higher prevalence in agricultural workers[89].

The assumption that iris ageing processes will give steadily increasing scores implies that it will not immediately yield outright false non-match failure. This assumption has correctly motivated ageing studies to remove poor images from their analyses[8]. Accordingly, ageing studies should include tight enough capture controls that outright recognition failures should not occur, or, secondarily, that failures should be discarded entirely during the analysis. Here a failure should be suspected if the genuine score is high e.g. comparable with an impostor score.

1.2 Sources of longitudinal change in biometrics

Biometric comparison scores vary systematically and stochastically. Iris scores vary with at least the following.

Sensor changes: Sequential captures of an inanimate test target will yield images that differ due to read- and shot-noise effects in the sensor electronics[56]. In most situations, these will usually have little effect on iris recognition performance. If, however, a camera's LED illuminators degrade[72, 73], recognition can obviously be impeded.

Environmental changes: Changes in either the ambient or infrared illumination incident on the eye, or in the light reflected toward the sensor (e.g. due to atmospheric effects), will yield differences in the captured images. Sensors running with a automatic gain control alter their response depending on incident illumination.

Behavioral changes: It is well known that individuals using a biometric system on a regular basis become acclimated to it. This is most often known as habituation[51] and is known to be more prominent if feedback is provided e.g. as a yes/no decision. The OPS-XING data used in this report reveals evidence of habituation - the persons who use the system most frequently produce lower (i.e. more similar) average comparison scores - see the discussion later around Figure 11.

Changes in the eye itself: In healthy individuals the following are either known, or understood to be, influential on iris recognition over various time scales.

- **Pupil size:** Pupil size varies over timescales below one second up to several decades. The Hippus phenomenon, irregular onset oscillations at 0.05 to 0.3 Hz[11] with amplitude around 1mm, has unknown cause and can be chaotic[80]. Size is influenced by cognitive effort[66] and old/new memory recall[38]. Fatigue causes modulation of pupil size[95], cycling with periods above 5 seconds with amplitude 0 to 1mm[60]. On longer timescales, pupil dilation is affected by many pharmacologic agents, beyond those used in routine ophthalmology. Particularly, stimulants and sedatives act to broadly increase and decrease pupil size respectively[82] but the physiological mechanism matters: drugs acting on the parasympathetic and sympathetic nervous systems differentially innervate the iris sphincter and pupillary dilator muscles respectively[41]⁶. Smoking gives an acute increase in pupil size by suppressing parasympathetic activity[61], but not on longer timescales[86]. Over variable timescales, pupil dilation changes with environmental illumination levels, such as those in the workplace, at home, and places in between, perhaps on a diurnal timescale. Over very long timescales pupil size decreases in healthy adults by about 0.4mm per decade[94, 13]⁷ due to fibrosis and an increase in rigidity of the sphincter muscle, and not due to a change in retinal sensitivity[17]. While iris recognition algorithms are designed

⁶Disruption of the parasympathetic, or stimulation of the sympathetic, nervous systems produce the same effect, mydriasis (pupil dilation), because the iris sphincter muscle acts to constrict the pupil, the dilator muscle to dilate it.

⁷This is a dark-adapted value in which the iris is radially narrow. Over eight decades and 263 individuals, the study [13] reports that for subjects 18 to 19 years (n=6), the mean dark-adapted pupil diameter was 6.85 mm (range: 5.6 to 7.5 mm); 20 to 29 years (n=66), 7.33 mm (range: 5.7 to 8.8 mm); 30 to 39 years (n=50), 6.64 mm (range: 5.3 to 8.7 mm); 40 to 49 years (n=51), 6.15 mm (range: 4.5 to 8.2 mm); 50 to 59 years (n=50), 5.77 mm (range: 4.4 to 7.2 mm); 60 to 69 years (n=30), 5.58 mm (range: 3.5 to 7.5 mm); 70 to 79 years (n=6), 5.17 mm (range: 4.6 to 6.0 mm); and 80 years (n=4), 4.85 mm (range: 4.1 to 5.3 mm). This is a reduction of about one third over 6 decades.

to have some invariance to pupil dilation[19], the various technical approaches rely on models of iris texture change and these, empirically, do not remove all the dependence of iris comparison scores on the dilation present in the images[36, 23, 40, 87, 34] - see section 3.2 for a large-population result.

Under our section 1.1 definition of iris-ageing, we do not regard pupil dilation as a component because much of its variation is stochastic. The decades-long constriction process would be a source of iris ageing but its effects a) may nevertheless be small⁸, b) are tolerated by current algorithms, and c) compensated for by certain cameras. As a slow process, it will have had negligible influence on published studies.

Nevertheless short term pupil dilation effects are influential on iris recognition, and should be compensated for in iris ageing studies.

- **Eyelid occlusion:** Eyelid occlusion degrades iris recognition accuracy[87] because segmentation is impeded, and information available from the iris increases non-linearly with the height of the palpebrae fissure (PF), the distance from the upper to lower eyelid margins, normally 9 to 12 mm. The eyelid position is under voluntary and autonomic control via the tarsal muscles[82]. It is most obviously manifest on millisecond timescales as blinking, but on longer timescales, fatigue and pharmaceuticals consumption are influential. The medical condition ptosis is characterized by drooping of the upper eyelid as far as the pupil. Ptosis occurs in later years. Head position also affects eyelid position - down gaze angles to 40 degrees produce $\Delta PF = -0.024 \text{ mm deg}^{-1}$ [37] or higher[78].
- **Corneal shape:** The cornea covers the iris and its optical refractive properties mean that changes in shape may alter an iris image. Shape changes do occur during accommodation[69], with head position[49], with surgery, and over decades[48]. The effects on iris recognition have not been quantified although some modeling has been done[71, 46].
- **Arcus senilis:** This condition[15] manifests as an opaque white to gray colored ring at the periphery of the cornea. It is caused by deposits of cholesterol in the cornea or hyaline degeneration. It occurs in older persons. Albredo changes could impede accurate segmentation of the outer boundary[18].

1.3 Overview of prior iris ageing studies

The core issue of whether the iris texture⁹ ages has been the subject of the following prior studies. These aim to quantify iris ageing and are motivated to check the Daugman patent assertion[21] that iris texture is “essentially immutable over a person’s life”.

Baker et al. used 6797 images of 23 persons to report an adverse, statistically significant, increase in FNMR from comparisons of images collected more than 1200 days apart vs. those from short term collections fewer than 120 days apart[8]. This shift in the genuine distribution was noted “across a broad range of threshold values”.

Notably the researchers manually screened all images for poor quality, excluding those with “out-of-focus irises, major portions of the iris occluded, obvious interlace artifacts etc”. This step was necessary because the LG2200 camera had been modified to produce images that would normally have been subject to “built-in quality control checks”. Additionally images for which the IrisBEE algorithm gave a “noticeably poor” segmentation were also excluded.

⁸ For example, if a person with an 11mm diameter iris is enrolled at age 20 with pupil diameter 7mm, and this reduces to 5mm at age 70, this would correspond to $\Delta D = 1 - (1 - 0.64)/(1 - 0.45) = 0.33$. By looking at Figure 7, or by drawing vertical lines on Figure 6, we estimate Hamming distance changes $\Delta HD < 0.1$. This would be tolerable if the initial enrollment was free of image quality problems.

⁹ The term texture is used in this document to refer to the visible parts of the iris notably the collarette, Fuch’s crypts, base crypts, the stroma, and Schwalbe’s contraction and radial folds, and the pupillary ruff.

The study dismissed pupil dilation as responsible for the ageing effect based on a small rank correlation of average dilation difference and average dissimilarity scores. The study also dismissed exposed iris area as causal concluding that “there is no substantial correlation between the number of non-occluded bits and elapsed time” and that “change in the amount of iris occluded does not account for the increase in false reject rate”. It is not clear whether the correlation was against absolute time, or time between captures i.e. whether the analysis was done for single images, or pairs. The correlation of the number of bits with comparison score was not reported. The team first reported on ageing in 2009[7] using a smaller population, and a non-commercial recognition algorithm.

Fenker et al. reported that recognition error rates increase by 153% over a three year period, and by 82% over a two year period.¹⁰ These results were the primary motivators of this NIST study. A number of media stories¹¹ included this figure without noting that this result was the ratio of two numbers (short term FNMR of 0.096 (i.e. nearly 10%) and long term FNMR of 0.243, i.e. nearly 25%). This 153% increase in FNMR is noisy. The application of a bootstrap uncertainty estimation procedure gave an interval of [80%,307%] around this estimate but it was implemented to first sample subjects (with replacement) and then scores - it did not additionally sample images as indicated in standards[53]. This would have been reasonable if subjects were hypothesized to cause recognition failure, but not reasonable if particular images were responsible.

This study[29] made no attempt to compensate for effects of pupil dilation. That aspect had been addressed in a prior study[28] which used images in 2008-2010. The study found significant¹² ageing effects which remained even after measures were taken to remove the effects of dilation change. This issue is discussed at length in section 4.3.1.

Tome-Gonzalez et al. used 8128 images of 254 individuals from the BioSecurID database to investigate ageing[91]. The data was collected in four sessions “separated typically by one to four weeks” to show that inter-collection-session false non-match rates were higher than intra-session rates. However: a) The recognition error rates are very high (FNMR \approx 0.1 intra-session and FNMR \approx 0.25 inter-session, at FMR = 0.01) despite culling 1905 incorrectly segmented images. High FNMR is indicative of poor images or a poor algorithm relative to the commercial mainstream[34]. b) The dependence on time is not monotonic. The authors note that “once a minimum time between samples has passed, error rates are not apparently increased” noting that inter-session 1-3 is worse than 1-4 which may be due to statistical significance problems. c) The authors do not discuss the high session 1 pupil sizes that are evident in Figure 9 in [91] as being influential. d) The total duration of the study is so short (16 weeks) that any anatomical ageing effect would be smaller than in other studies.

Sazanova et al. used 7628 images of 244 subjects (46 over more than one year) to produce formal rate-of-change-of-comparison-score statements of iris ageing for the Masek[55] and Neurotechnology algorithms[84]. They acknowledged the stochastic time-varying nature of contrast, occlusion, illumination and blur as factors that can undermine

¹⁰ This results appear in Table 2 of [29] for the period 2008-2011. The number for the 2008-2010 interval is 82%. NIST was unable to conduct studies with the 2011 images, pending legal issues. The ND results were obtained with the commercial Neurotechnology recognition engine. The company has submitted competitive recognition prototypes to the NIST IREX-III and IREX-IV evaluations. The threshold was set to 580 per a false match calibration of 1 in 2 million. This threshold corresponds to a threshold of 0.0017 on the B02P, as the developer uses a $1/x$ mechanism to convert between dissimilarity and similarity scores. For the 2008-2011 set, FNMR increased from 0.096 to 0.243 corresponding to a mean percent increase of 153%, with confidence interval of [85, 307]%. This corresponds to $0.096 \times 5244 = 503$ within semester errors, and $0.243 \times 20888 = 5076$ in Spring 08 to Spring 11 comparisons. This involves 32 people and 2338 images.

¹¹ These papers have been widely, and with less than complete accuracy, quoted in the press with titles, *Ageing Eyes Hinder Biometric Scans*[33], *Accuracy of Iris Recognition Systems Degrades over Time*[59], *Iris aging raises issues about recognition accuracy*[96], *Aging process confounds iris recognition biometrics*[2], *Future Eye Scanners Must Combat Aging Eyes*[24], *Researchers question long-term reliability of iris recognition*[3, 12], *Aged eyes prevent iris recognition*[1], *Researchers question long-term reliability of iris recognition*[3], and *Ageing irises could confound biometric checks*[32]. Inevitably this meme has propagated internationally to government and other procurement officials tasked with determining whether iris is a viable biometric.

¹² This and other studies use full cross-comparison of images. While this gives the best estimate of the mean, the scores are not independent such that uncertainty estimates are governed more closely by the number of images $O(N)$ rather than scores $O(N^2)$. This issue can lead to optimistic uncertainty estimates.

recognition and used a robust linear regression to estimate coefficients for these and for time between captures. This regression was conducted once, over the entire imbalanced population - it did not account for correlated observations, nor inter-person heterogeneity. Notably the study did not include dilation in the regression, and it proposed to model this and other (unspecified) “changes in data acquisition procedure” in future work. The work quantified false rejection rates over time at fixed false match rates (FMR), rather than fixed threshold. This only matters to the extent that FMR is invariant with ageing and that the use of the $FMR = 0$ operating point - an extreme value - is repeatable across partitions. The publication does include time-series - but not for specific individuals - and these reveal noise (i.e. variation in comparison scores) far in excess of the computed age-related change.

Fairhurst et al. used images of 632 images of 79 users (158 eyes) with a Masek[55] implementation modified to reduce segmentation errors[27]. The authors note the role of dilation, the age-related effect of ambient illumination on dilation, and conclude that “dilation decreases with age and consequently Hamming distances decrease”. They note the inconsistency of Baker’s[8] result, that dilation change is not responsible for the noted ageing effect, with other Notre Dame research that dilation change does change recognition scores[40]. They “affirm the importance of a reliable and robust segmentation” in iris recognition and conclude that “after eliminating this segmentation factor” that “physical ageing effects ... are primarily the result of physiology of pupil dilation mechanisms”.

Rankin et al. exposed 456 visible light images from 76 subjects imaged on 3 occasions each 3 months apart to several variants of their own iris recognition algorithms to conclude that ageing does occur. Their paper bins genuine HD scores in the intervals $[0,0.09)$, $[0.09,0.19)$, $[0.19,0.29)$, $[0.29,0.39)$, $[0.39,0.49)$, and adopts a threshold of $HD \geq 0.39$ to declare false rejection. The paper notes lower FNMR at time lapses of 6 months than at 3 months (Table 1, $FNMR(3) = 0.212$ vs. $FNMR(6) = 0.205$), and that scores (Table 2) shift lower (better) after 6 months vs. 3. These error rates are much higher than those published in large scale tests[34] at such thresholds either because the images are visible-light or because the algorithm is poor. Daugman asserted as much[20] suspecting the existence of segmentation errors. This was denied[77] by the study authors¹³.

Czajka reports results for 571 images of 58 eyes imaged on two or more occasions ranging from 30 days up to 2960 days. They applied three recognition algorithms, two commercial, and one self-developed. Noting zero recognition errors from one algorithm, the paper notes the explicit time-series nature of ageing, suggesting that studies of biometric ageing should address the stationarity of the dissimilarity scores. This view requires all statistical moments of the score distribution to be stable rather than just whatever summary statistic is used in routine performance assessment. The outcome is a 14% deterioration in the mean of 2432 Neurotechnology comparison scores in the 5 to 9 year period vs. the 1588 scores in the up-to-2-year period. This is consistent with ageing. While the paper establishes strong significance via a t-test, it ignores correlation of asymmetric match scores, and of scores from cross-comparison of images. The study observed that more accurate algorithms are more susceptible to ageing. Czajka uses this observation to claim the existence of iris texture ageing effects because a) better segmenters give better alignment of iris regions being compared, and b) better algorithms have higher sensitivity to individual features of the iris texture. The study suggests that its finding of iris ageing should be confirmed by exploiting “big, heterogeneous datasets”.

¹³ While all ageing studies have acknowledged that correct segmentation is critical, there may be definitional differences of what is “correct”. One step is to visually check that the detected boundaries (e.g. circles) align with the limbal and pupillary boundaries. However this should be done for pairs of images to ensure consistency because it is known that polar representations of iris features are sensitive to the *precise* location of, particularly, the pupil boundary. One study[57] showed fractional HD values increase by 0.023 per linear pixel misplacement of the pupil. Thus a pair of images with fine differences in boundaries will produce degraded similarity scores.

	A	B	C	D
Quantity	ND04-08	ND08-10	OPS-XING	OPS-FIELD
Source type	Image matching	Image matching	Archived logs	Image matching
Computation	2012	2012	2003-2012	2012
Num rec. algs	9	8	1	5
Source algs	ICE/IREX-IV/VI	IREX-IV/VI	Iridian[19]	IREX-IV
How Genuines Calc.	1:1 Cross-Comp	1:1 Cross-Comp	1:N Fixed-enrol	1:N Cross-Comp
Capture Dates	2004-2008	2008-2010	2003-2012	2004-2012
Capture Setting	Univ. lab	Univ. lab	Airports	Field
Num images	6802	11776	1042948	3544068
Num eyes	46	434	521474	NA
Num people	23	217	350566	622464
Num people 1YR	23	199	214731	286387
Num people 2YR	23	36	146104	138064
Num people 3YR	23	0	96588	64554
Num people 4YR	18	0	53880	21238
Num genuine	1321236	356022	5710434	2301246
Num genuine 1YR	350622	48136	3749774	1049422
Num genuine 2YR	253112	3750	2515911	491246
Num genuine 3YR	165314	0	1597711	210564
Num genuine 4YR	46520	0	896368	59266
Citations	[8, 7]	[28, 29]		[34, 74]

Table 1: **Data:** Summaries of the four IREX VI data sets used in this paper.

2 Experimental datasets

The four datasets used in this study are summarized in Table 1, and described in the sections below.

Table 1 summarizes the properties of the datasets. Column A represents the application of 2012 commercial recognition algorithms to the images collected by Baker et al. and a re-analysis of the scores from the CAM-2 algorithm that was used in their study. Columns B and D summarize the application of the new algorithms to different sets of images. Column C refers to textual data from the logs of operational systems.

The various datasets are differentiated by the number of subjects, the frequency and regularity of image collection, the duration of collection (in years), whether or not the recognition algorithms were applied retro-actively to generate all possible genuine scores (columns A, B, D) or only a single score against an enrollment sample during collection (C). The details appear below.

Note that one of the datasets is comprised only of text log files, instead of images most often used in biometrics research.

Note that all of the data sets were collected without inquiries or tests of general or ophthalmological health.

2.1 Border crossing data OPS-XING

OPS-XING is the set of time-stamped genuine scores recorded in the logs of an operational border crossing system[5]. The scores come from the Iridian implementation of the Daugman algorithm[19]. The scores are Hamming distances (HDs) from a 1:N search of a live image against the enrolled database. Neither the enrollment nor recognition images were provided to NIST. This precluded re-matching with contemporary algorithms. The camera used in border crossing transactions was a Panasonic BM-ET 330 which, at the time it was deployed in 2007, was a leading iris camera. It remains in effective use today. Enrollment images were collected from 2003 to mid 2007 using an LG 2200 camera, and, from mid 2007 onwards using the Panasonic camera. The LG 2200 camera is identified in this report by the letter “L” in the report. The Panasonic cameras carries the letter “B”. Neither camera is marketed commercially today. Border crossing

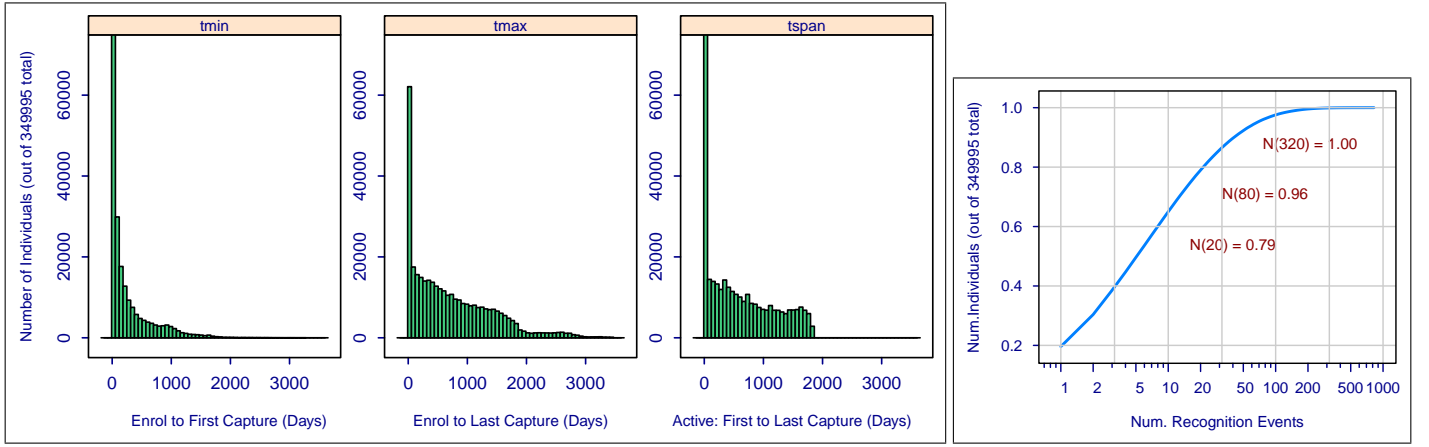


Figure 3: **Population sizes:** For dataset OPS-XING, the three panels show counts of individuals by the time elapsed from their enrollment to first capture (T_{\min}), time from enrollment to last capture (T_{\max}), and the duration of the active period ($T_A = T_{\max} - T_{\min}$). The spikes at zero days are due to a training passage that each enrollee performs shortly after iris enrolment. The center panel, T_{\max} , is most germane to ageing studies. The rightmost graph shows the empirical distribution function of the number of recognition events per individual.

transactions were logged from November 2007 onwards.

Table 1 includes a summary of the OPS-XING collection. The data is advantaged by the size of its population, its extent (up to nine years), its controlled nature - fixed installations, a single and capable recognition algorithm that did not change, and only two camera models are used and logged. The population is a set of frequent travelers who paid money to enroll (currently \$50), and who execute multiple transactions at irregular times with differing frequencies. Figures 3 show transaction volume statistics. The system remains operational and in heavy use.

The logs for this data include iris and pupil radius estimates from which dilation can be computed - see section 3.2. The authors discarded some data using the following criteria. First enrolled persons without recognition transactions were not analyzed - and their numbers are not reported here. Second, individual eyes for which only one recognition transaction was logged were discarded because of their limited usefulness in analysis. Similarly individual eyes for which transactions only occurred on one day were discarded. Finally, eyes for which the maximum time between enrollment and recognition was less than one day were ignored.

2.1.1 Camera interoperability

Each transaction is accompanied by an L or B label indicating the camera used for initial enrollment. All recognition images were collected with instances of camera B, and thus any given comparison involves (L,B) or (B,B) camera pairs. As shown later, (L,B) Hamming distances are higher than (B,B). This systematic effect is due to imaging-system differences, particularly with respect to wavelength of the infra-red illuminant¹⁴ the angle of incident infrared light due to light emitting diode (LED) placement, and the difference in the way subjects interact with the single-eye “L” and dual-eye “B” cameras. Particularly, when camera B was adopted for recognition transactions, and logging commenced, the L-enrollees had to switch to using camera B without any prior experience. Those enrolled on camera B were specifically trained.

¹⁴Images of irises collected at different wavelengths have different appearance[42, 81] and this may be influential on feature extraction algorithms.

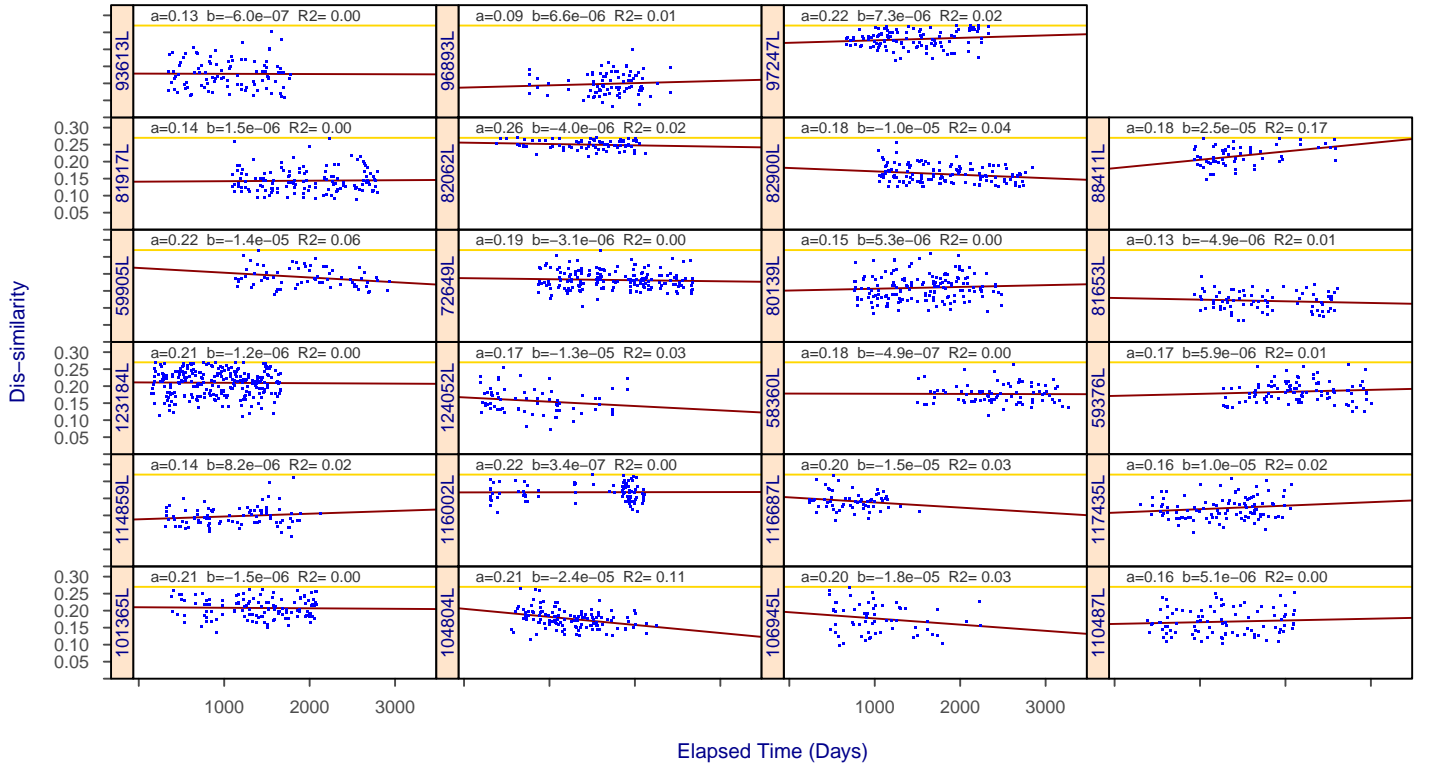
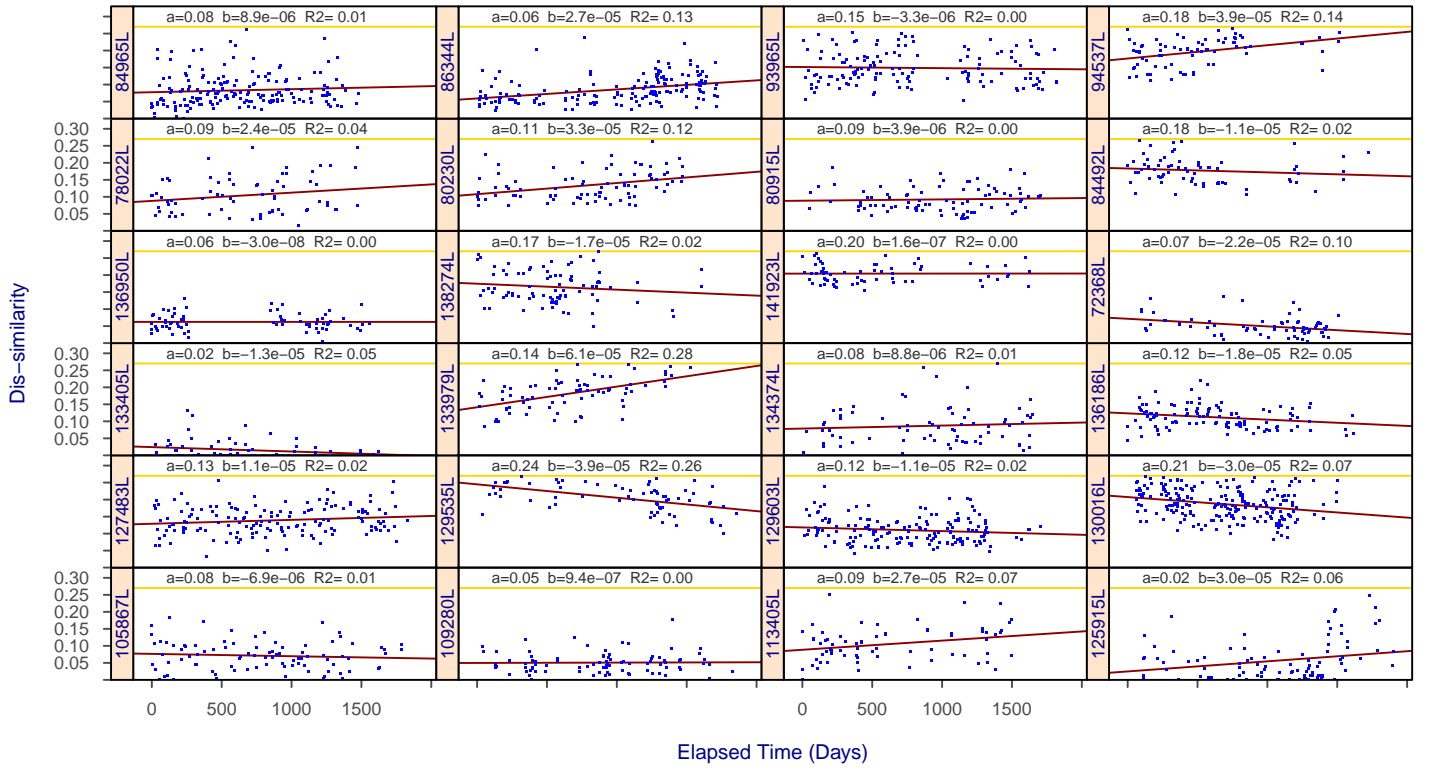


Figure 4: Score trajectories: For dataset OPS-XING, the panels show Hamming distance scores for 24 randomly selected individuals executing identification attempts at least 60 times in five years. The horz. axis is time since enrollment. The gold line indicates the 0.27 threshold of section 2.1.2. The blue dots indicate a single logged score and the grey text gives the intercept, gradient and R_2 of the linear regression line shown in red. The regression is computed with time as the only covariate (dilation is not included). Note the varying intercepts, the intra-person and inter-person variance, negative gradients, and cases where positive gradients are nevertheless accompanied by low HDs at the end of each period. Logging was enabled in 2007, so persons enrolled previously on camera L have no score data initially.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

2.1.2 The use of thresholded data

Subjects in the OPS-XING application are invested in an immediate (successful) outcome (i.e. border crossing). They make attempts until successfully recognized, or a timeout occurs. This *photo-and-match* approach yields a corpus of images that are matchable essentially by-definition, and in this case the system logs only contain scores from successful comparisons, at or below a dissimilarity threshold¹⁵. Rejections, false or otherwise, do not appear in the logs and their absence imparts a censoring of the right tail of the underlying score distribution. However, the data points that would be present in that tail, had they been logged, would typically occur because a camera's quality assessment apparatus allowed a poor quality image (e.g. of a closed blinking eye, or of an off-angle gaze) to pass to the recognition algorithm. This assertion ignores the possibility that an underlying ageing effect causes false rejections. However, assuming such ageing would be permanent, the individual would then have persistent rejections such that low Hamming distances would not be observed subsequently - this is often *not* the case.

This operational system reduced the threshold as enrolled population size increased in order to maintain a low calibrated false match rate. This complicates our analysis because the distribution is truncated differently earlier in the period. CBSA confirmed the use of a threshold of 0.2797 in the earliest events (in 2007), reducing to 0.2704 at the end of the period (in 2012). We therefore *re-thresholded* the data removing all points with Hamming distances above 0.27. The original logs contain 5710434 entries from 521474 eyes of 350566 people. The post-thresholded logs contain 5667363 entries from 519945 eyes of 349995 people. The rethresholding affects 28900 eyes of 25610 individuals, with 1529 eyes of 571 people being dropped from the analysis entirely. Approximately two thirds of these were enrolled with camera L. The mean number of scores removed is 1.49 per affected eye. The re-thresholding establishes a time-independent censoring of the right tail of the genuine distribution, and this supports unbiased detection of trends in the data. Without this step, some transactions logged early in the period (e.g. 2007) would be expected in the interval [0.27,0.28] and these would not be present in more recent events. Note that most users' HDs are clustered well below 0.27 (see Figure 4 and Table 2).

The absence of the right-tail of the genuine distribution does not indict this dataset for detection of iris texture ageing if we assume that there are no individuals who age so rapidly that their HDs degrade and they can no longer be recognized. Our assumption is supported by noting that while some individuals have positive gradients they do not quickly approach the threshold. Note that in Figure 4 eye 133979L does increase over a 1500 day span, and the last transaction is high. This could be the result of the CLASS B, C or D ageing processes defined later in section 5.3, or due to progressive ill health.

2.1.3 Summary

In conclusion, we assert that operational logs of successful recognition attempts are an invaluable resource, not least because of their large size. The use of previously-matched images in an ageing study is not biased toward a no-ageing result because the physical manifestation of iris-ageing is not (at least immediately) the occurrence of poor genuine scores in the tail of the genuine distribution (i.e. ones that would lead to outright recognition failure), but rather the increase in the broad dissimilarity distribution.

However, by analyzing logs of successful transactions, defective images are explicitly omitted. This is not the case later in this report (Figure 16), where above threshold scores do occur and the genuine distribution is bimodal - this is characteristic of recognition (particularly segmentation) failure. We recognize that complete logs that include failed recognition attempts offer extended possibilities for other analyses.

¹⁵One-to-many identification algorithms can realize speed gains by short-circuiting a distance calculation when a partial match is already above a distance threshold

2.2 Notre Dame data

The University of Notre Dame has collected multimodal biometric data at least over the interval 2004 through 2011. The iris images of two partitions were made available to the authors. Sizes of these sets appear in Table 1.

The first partition, ND04-08, was used in template ageing studies reported by Baker et al.[7, 8]. The images were collected with a modified LG2200 camera that is no longer marketed¹⁶. Importantly the images that “did not pass the normal built-in quality control checks” of the camera were excluded because they degraded recognition accuracy[68]. For the ageing studies, Notre Dame manually screened these images, correctly excluding those with “noticeably poor quality” particularly those with an “out-of-focus iris, major portions of the iris occluded, interlace artifacts etc”¹⁷. The result is a set of 6802 images from the approx. 60000 parent set[68].

The second partition, ND08-10, was used in template ageing studies reported by Fenker et al.[28, 29]. Collected with a more modern camera, the LG4000¹⁸, it has a larger population than the ND04-08 set, but shorter temporal extent. This dataset is the focus of the NIST work reported in this paper because it was widely cited as giving rise to serious concerns regarding ageing effects [33].

These datasets are available to researchers. As such, the algorithms used later in this report could have been developed to process these images with better than normal capability.

2.3 Field-collected data OPS-FIELD

The OPS-FIELD set is the largest in terms of subjects and numbers of images, but not number of genuine comparisons. The set has two partitions. The first is one termed DETAINEE corresponds to 390,119 persons interdicted in military encounters - these subjects are imaged very few times on an irregular schedule. They produce 1,199,052 genuine scores. The second is labelled PACS and is comprised of enrollment images of 232,345 persons who subsequently use a physical access control system. These people were re-enrolled on a regular schedule so there are multiple images per person. This partition yields 1,102,194 genuine scores.

The available images were matched using several very recent recognition algorithms. These are the latest commercial competitors submitted to the IREX IV evaluation. The genuine scores are extracted from the $L = 20$ candidates returned from 1:N searches. The images were collected using several cameras, and several versions of each of those. The Securimetrics/L1 HIIDE and PIER and Crossmatch iSCAN/SEEK cameras comprise most of the collection. A number of other cameras are used, some of which are unknown.

3 Longitudinal variation in iris recognition

This section commences with an overview of the problem, showing how measured biometric outcomes can vary. This is followed by a discussion of sources of change in iris recognition. This supports a proposed hierarchy of ageing types. The final section is dedicated to pupil dilation.

¹⁶The modification was a bypass of the internal image quality checks.

¹⁷These and other avoidable image quality problems have been catalogued[73].

¹⁸Available still from the manufacturer, now named IrisID.

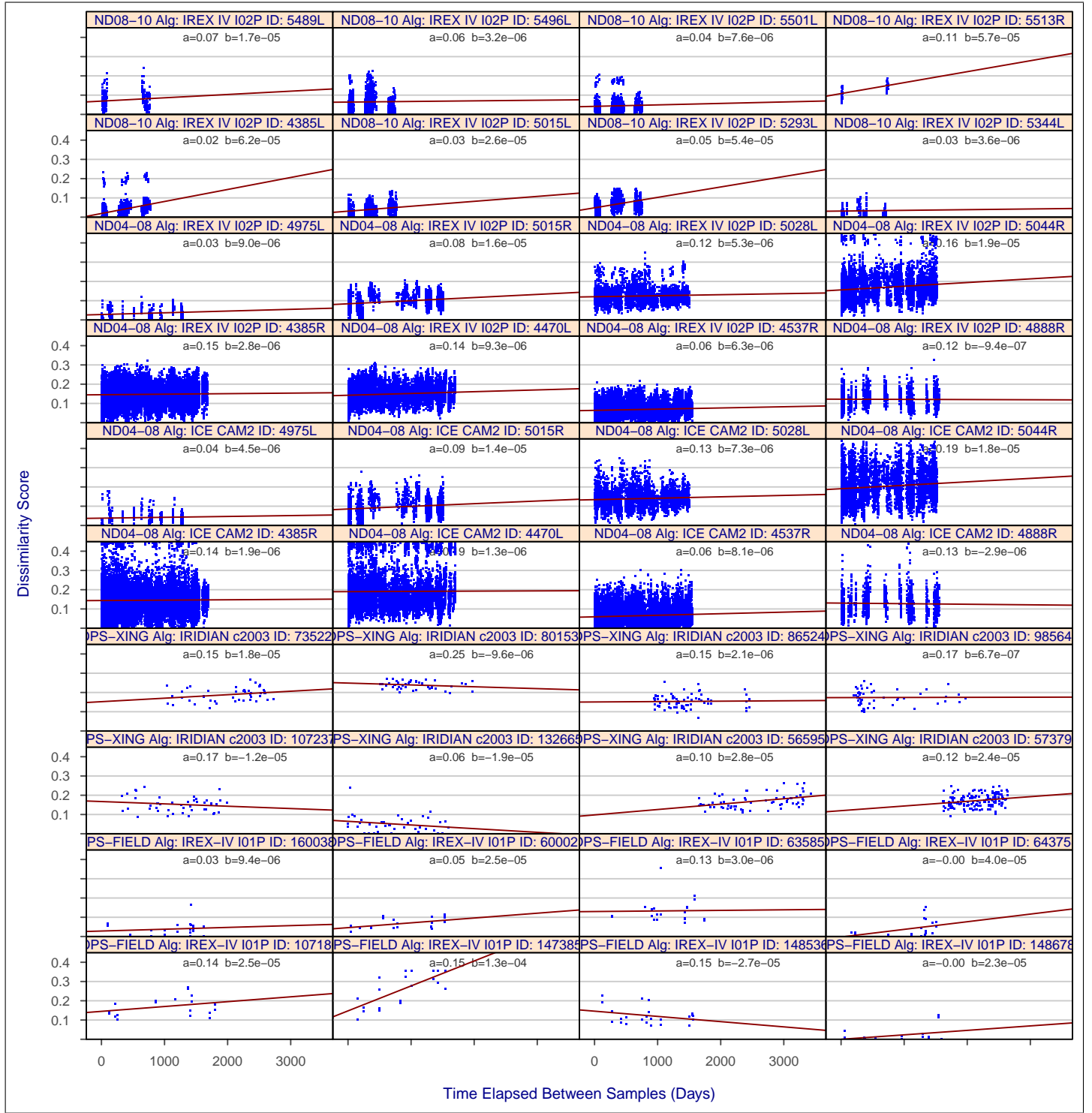


Figure 5: Score trajectories by dataset: Hamming distances over time for eight selected individuals in each of the four datasets considered in this paper. The matching algorithms are identified in the panel header and all originate with Daugman or his commercial partners. The same individuals from the ND04-08 set are shown for the c. 2006 CAM2 and c. 2012 I02P University of Cambridge algorithms. The density of points in the ND panels is due to the use of $O(N^2)$ full cross comparison. The individuals were randomly selected from the subset who participated in the respective collections more than N times over at least T_A days as follows: OPS-FIELD, $N \geq 20$, $T_A \geq 1460$; OPS-XING, $N \geq 40$, $T_A \geq 1460$; ND04-08, $N \geq 20$, $T_A \geq 1095$; and ND80-10, $N \geq 20$, $T_A \geq 600$. The red lines are the result of fitting a linear model to $HD = HD(t)$ and are useful as an exploratory indicator.

3.1 Overview

Figure 5 shows recognition scores as a function of time for the four datasets considered in this paper. The following points are made to note the varied characteristics of time-series biometric data.

Sampling rates: In most cases, subjects were sampled irregularly in time, with the exception that the Notre Dame protocol called for regular collection. Note that the graphs show time *between* collections so a pair of proximal points will sometimes correspond to images collected years apart (e.g. 2004-2005 and 2007-2008). The ND plots include many points as a result of full-cross comparison of images, while the OPS-XING points represent recognition against singular enrolled images.

Averages vary: The mean HD values, and intercepts, vary between person, and between dataset. This variation exists as product of equipment, cooperation, habituation and recognition algorithm¹⁹. For the OPS-XING set, the mean HD will be dependent on the quality of the initial enrollment image - any collected defect there will elevate scores in perpetuity.

Variances vary: The variation in HD is evident across datasets. Larger variation would occur if the collection was not controlled, if the camera did not enforce quality criteria, or if the subject was not cooperative. The laboratory set ND04-08 gives the highest ranges while the two large operational sets, OPS-FIELD and OPS-XING, give somewhat lower variance.

Gradients vary: The slope of the simple regression lines vary between persons, some increase, some decrease. The lines are strictly for exploratory purposes, more powerful models are included later. If CLASS D ageing effects associated with the iris texture are present in these time-series, the problem of quantifying an ageing rate involves detection of signal within noise.

3.2 Effect of dilation and dilation-change

Given pupil and iris radius estimates, R_P and R_I , dilation is stated as the pupil-iris radii ratio, $D = R_P/R_I$ where for the OPS-XING data the radius estimates come from the native operational algorithm, and for the ND sets radii are consensus values estimated as the arithmetic mean of values reported by the F02P, I02P and D03P algorithms submitted to IREX IV:

$$R = \frac{1}{3}(R_{F02P} + R_{I02P} + R_{D03P}) \quad (2)$$

Radius estimates usually differ only by very small amounts. We did not compute radii for the OPS-FIELD images.

Let a mated pair of images have dilation estimates, D_1 and D_2 . If, without loss of generality $D_1 \geq D_2$, then a dilation difference measure can be stated as:

$$\Delta D = 1 - \frac{1 - D_1}{1 - D_2} \quad (3)$$

which, with $D = R_P/R_I$, can be restated in terms of the radial iris thicknesses, t , as

$$\Delta D = \left(\frac{R_{I1}}{R_{I2}} \right) \left(\frac{R_{I2} - R_{P2}}{R_{I1} - R_{P1}} \right) = \gamma \left(\frac{t_2}{t_1} \right) \quad (4)$$

¹⁹The Figures all show HDs from algorithms traceable to John Daugman at Cambridge but may differ in material details such as their image processing and feature extraction methods.

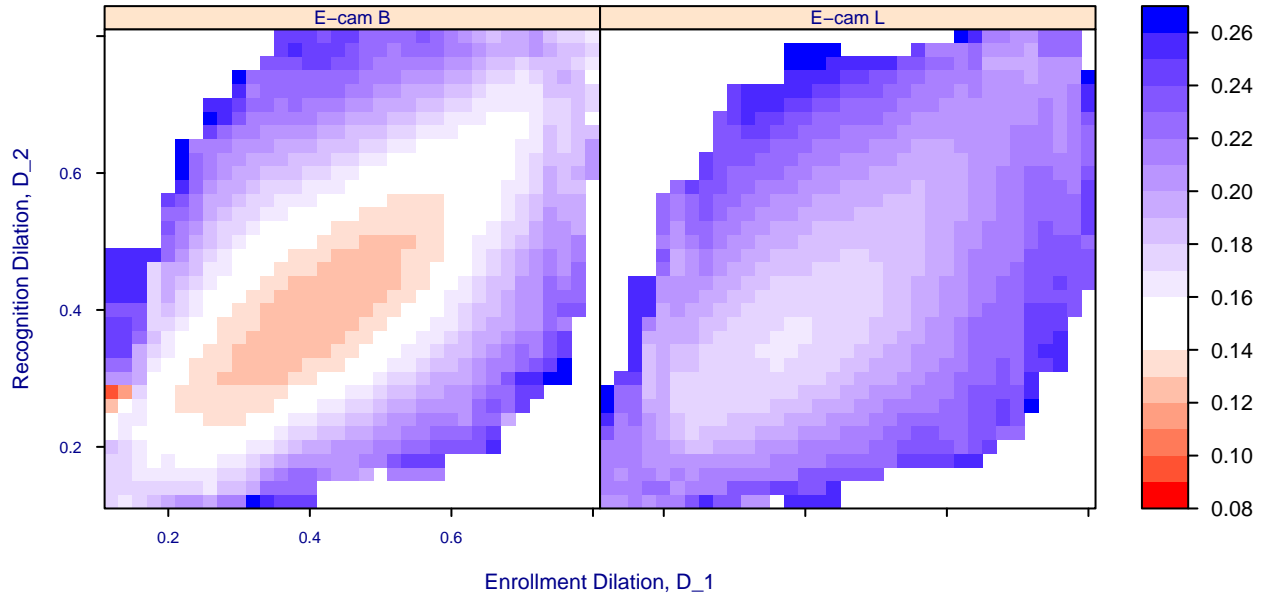


Figure 6: **Score depends on paired dilation:** For the OPS-XING dataset, HD as a function of the dilation of the search and enrollment images. The heatmap shows mean HD computed over $[D - 0.02, D + 0.02]$ intervals value, incrementing D in steps of 0.02. Note that difference in dilation is very influential on score. Note also that low or high dilation in both images is also deleterious. The left panel corresponds to enrollment and recognition camera pair (B,B), the right panel to (L,B). The latter scores are higher due to the use of the L camera, and the camera interoperability effects discussed in section 2.1.1.

which has the more physical interpretation as the product of a camera magnification term, γ assuming the anatomical iris has constant size, and an iris radial thickness ratio t_2/t_1 . This measure was used previously to assess sensitivity of algorithms to dilation change[36]. Other metrics have been reported in the literature[28, 39, 40] particularly

$$\Delta N = D_2 - D_1 \quad (5)$$

where a difference symbol is used for reference later. This metric lacks the physical intuitiveness of the thickness measure. The dependence on ΔD appears in Figure 7 and on D_1 and D_2 independently in the heatmap of Figure 6. For recent algorithms, IREX III showed dependence of both false negative and positive identification rates on D_1 and D_2 [34].

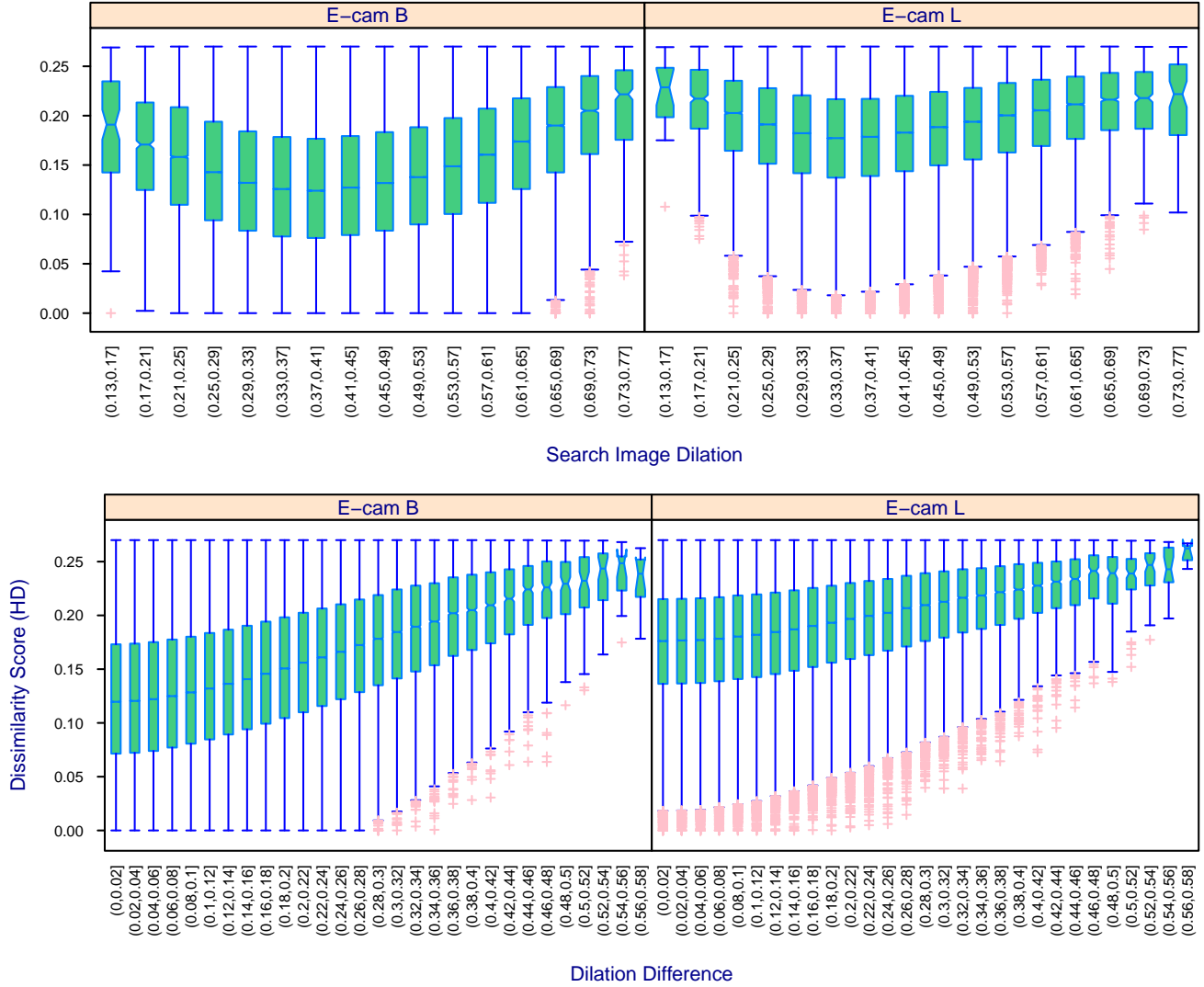


Figure 7: **Score depends on pupil dilation and dilation difference:** For the OPS-XING dataset and its two enrollment camera models, plots of Hamming distance HD as a function of the absolute dilation of the search image (above) and the dilation change measure of equation 4 (below). Poor HD values are produced when the search image dilation is absolutely low or high. Poor HD values are produced when pupils have different dilations.

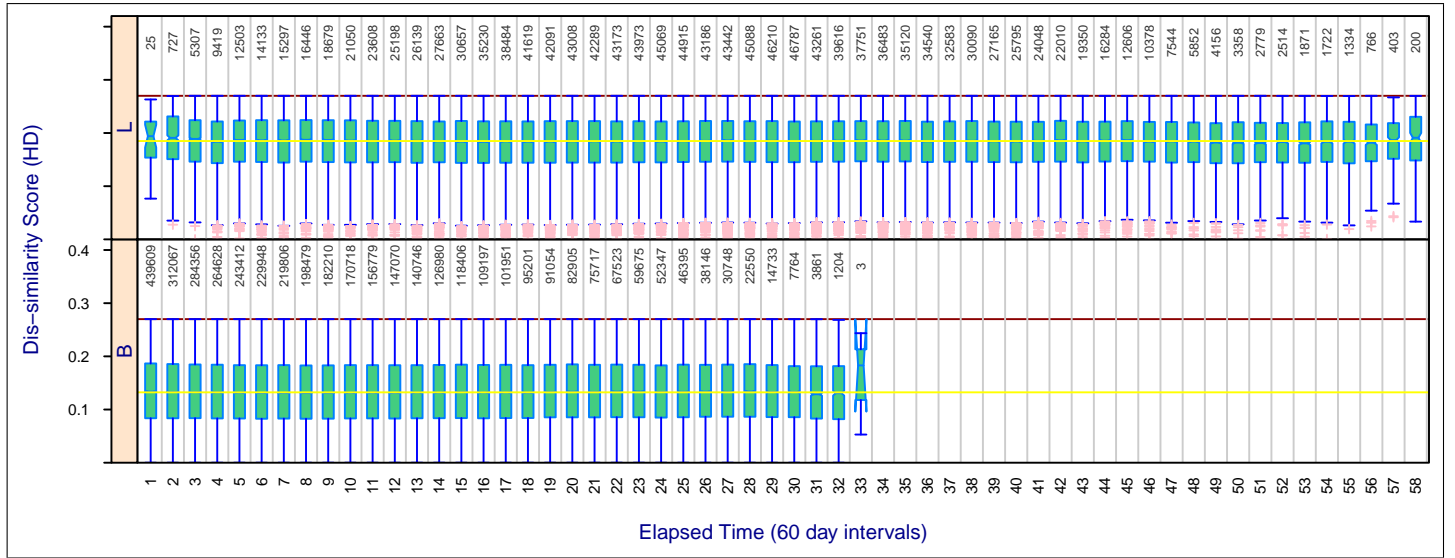


Figure 8: **Stability of score distributions:** For the OPS-XING data, the time evolution of the HD distribution computed over all individuals enrolled on camera L (the LG2200), above, or camera B (the Panasonic BM-ET 330), below. All border crossing captures use camera B. The horizontal axis is divided into 60-day bins, spanning 3480 days. The matching algorithm is the c. 2003 Daugman variant used by Iridian. The distributions are shown as box plots whose whiskers extend to 1.5 times the interquartile range which is shown by the green box. The red crosses below the whiskers are outliers under this definition. The median is indicated by the horizontal line, and the notch size in the side of the box is an uncertainty statement inversely related to the number of transactions which appears as grey text above each box. There are no outliers indicated above the whiskers because the data was thresholded at $HD = 0.27$ as discussed in section 2.1.2. The first three boxplots for camera L are higher due to the effect of transition (from late 2007 until early 2008) of subjects who had previously only used camera L, to recognition with camera B.

4 Results

4.1 Overview

The results are organized as follows: Section 4.2 reports results for the OPS-XING set of border crossing logs. The results include regression approaches supplemented by exploratory figures. Section 4.3 gives recognition results for the images of the University of Notre Dame (ND) datasets. This includes exploratory analyses, consideration of particular images, and quantitative cause-and-effect results. Section 4.4 gives exploratory results for the OPS-FIELD images.

4.2 Results for OPS-XING transactions

Figure 8 presents the time evolution of the Hamming distance distribution as a function of the time elapsed between enrollment and recognition. These are computed over the numbers of recognition events given above each box - some 60-day intervals see more than 100,000 transactions. For an explanation of the two panels, labelled “L” and “B”, see section 2.1.1. The centers of the boxplots show the median score has small deviations around the global median shown in yellow. Moreover, there is no evidence of the upward trend that would indicate ageing. In and of itself, this is an important result obtained from a population 1200 times larger than the largest prior study[29], and over time intervals about twice as long as those at Notre Dame[29] and as long as those in Warsaw University of Technology[16].

Nevertheless, this result is inadequate in two ways. First, it does not show whether more scores are appearing in the upper tail. To examine that, Figure 9 shows the proportion of scores above various thresholds, τ , i.e. the false non-

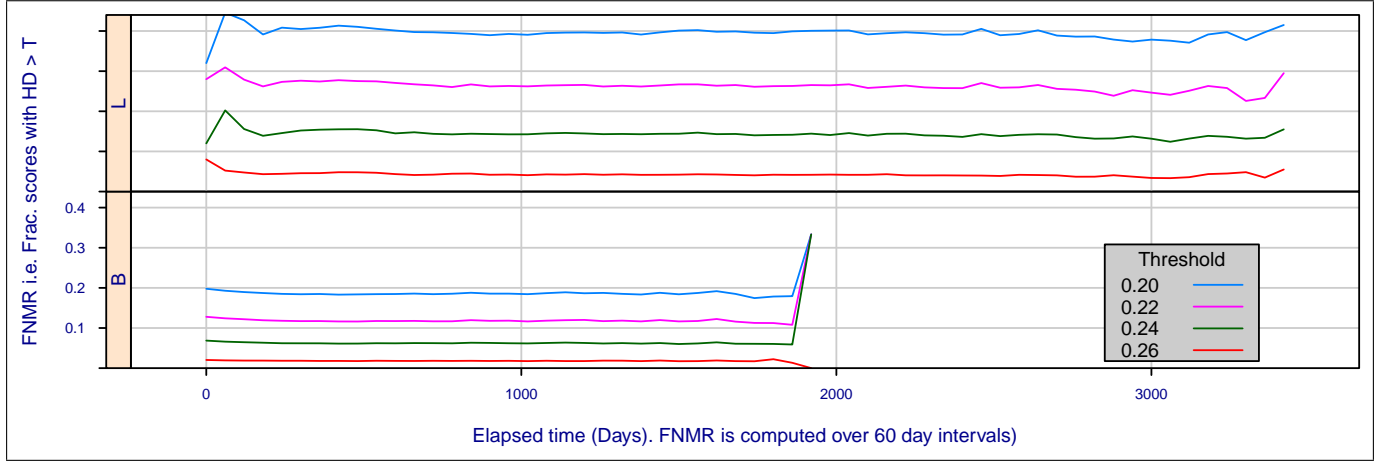


Figure 9: **Stability of error rates:** Time evolution of FNMR for four given thresholds applied to the OPS-XING data. The estimates are computed for all individuals enrolled on the LG2200 (camera L), above, and the others on the Panasonic BM-ET 330 (camera B), below. All verification captures use camera B. The matching algorithm is the c. 2003 Daugman variant used by Iridian. FNMR is computed over contiguous 60-day intervals. Thresholds values above 0.27 are not possible, per section 2.1.2. The spikes in the lower right and upper left panels are due to very small sample sizes.

Quantity	Enrollment on LG 2200						Enrollment on Panasonic BM-ET 330					
	Q01	Q25	Med	Mean	Q75	Q99	Q01	Q25	Med	Mean	Q75	Q99
T_{min}	126.0	502.0	901.0	987.6	1364.0	2685.0	0.0	0.0	41.0	194.7	241.0	1393.6
T_{max}	295.0	1648.0	2032.0	1987.3	2476.0	3306.0	0.0	156.0	506.0	614.1	1003.0	1765.0
T_A	0.0	350.0	1181.0	999.7	1612.0	1808.0	0.0	0.0	235.0	419.4	716.0	1689.0
n_i	1.0	3.0	8.0	23.5	27.0	192.0	1.0	1.0	3.0	9.3	8.0	96.0
HD mean	0.081	0.162	0.197	0.191	0.225	0.264	0.014	0.104	0.150	0.147	0.193	0.262
HD sd	0.003	0.022	0.029	0.030	0.036	0.067	0.002	0.027	0.039	0.040	0.051	0.102
D_2 mean	0.270	0.372	0.415	0.420	0.463	0.604	0.279	0.390	0.440	0.444	0.494	0.634
D_2 sd	0.004	0.031	0.042	0.044	0.055	0.107	0.002	0.026	0.039	0.041	0.053	0.112
D_1 mean	0.272	0.377	0.435	0.442	0.500	0.655	0.258	0.350	0.398	0.407	0.456	0.618
ΔD mean	0.010	0.061	0.092	0.107	0.139	0.307	0.006	0.058	0.092	0.104	0.138	0.299
ΔD sd	0.005	0.036	0.051	0.053	0.067	0.131	0.002	0.032	0.048	0.051	0.066	0.138
ΔN mean	0.006	0.037	0.056	0.065	0.084	0.185	0.003	0.035	0.055	0.063	0.083	0.181
ΔN sd	0.003	0.024	0.034	0.035	0.045	0.088	0.001	0.020	0.030	0.032	0.041	0.086

Table 2: **Population statistics:** For the OPS-XING dataset, population summary statistics (1, 25, 50, 75, 99 quantiles) for, in the left six columns those eyes enrolled on the LG2200, and on the right the BM-ET 330. The rows give, in order, T_{min} , the time from enrollment until first use; T_{max} , the time from enrollment until last use; T_A , the time from first to last use; n_i , the number of recognition results for a particular eye; the eye-specific mean HD; its standard deviation; the eye-specific mean dilation during recognition; its standard deviation; the dilation of the enrollment image; the mean dilation thickness difference (eq. 4); its standard deviation; the mean dilation difference (eq. 5); and its standard deviation.

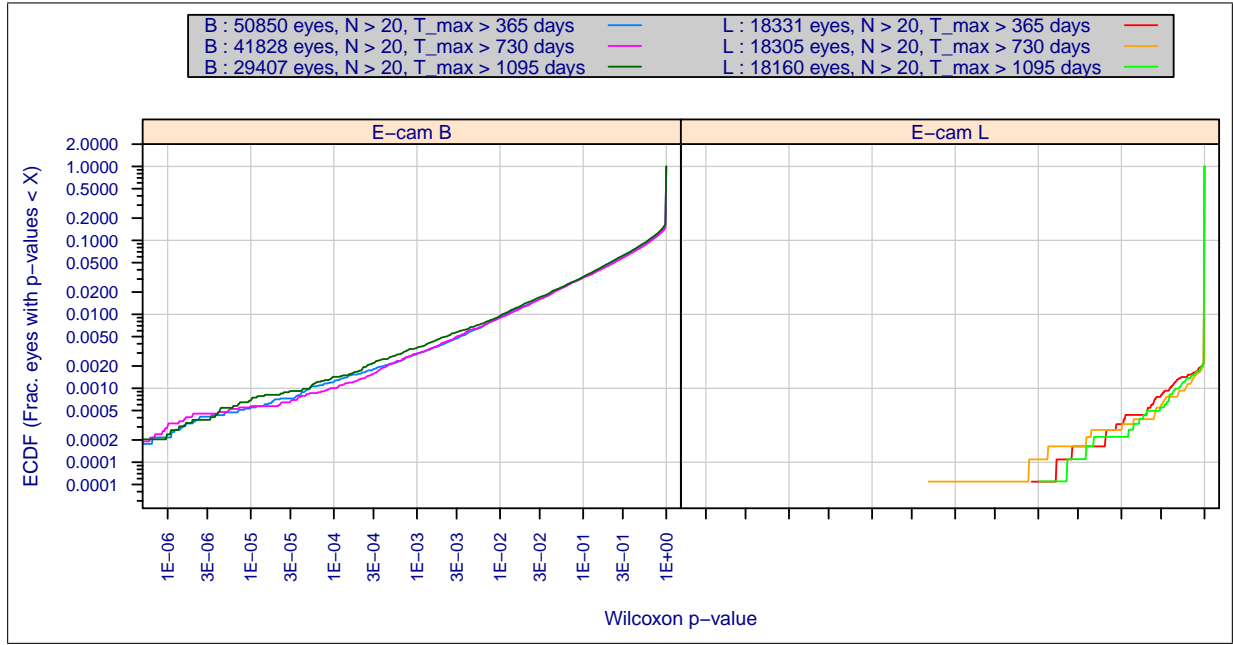


Figure 10: **Long vs. short time lapse:** For the OPS-XING dataset, the graph shows the cumulative distribution function of p-values produced in a non-paired Wilcoxon test of HD increase. The test is applied to eyes for which more than $N \geq 20$ comparison scores are available. It tests whether the median of the later scores is higher than that of the early scores. A low p-value indicates that the median is significantly greater. For those enrolled on camera B, the plots indicates that about 1.1% of eyes give significantly higher late scores vs. early scores if $p = 0.01$ is held to be significant. For those enrolled on camera L the fraction is smaller.

match rate $\text{FNMR}(\tau)$. Four different thresholds are used. For enrollment camera B, FNMR is flat, and for camera L, very slightly decreasing²⁰. A decrease would be consistent with subjects presenting better images to the camera, as would occur with gradual habituation to the camera as explained in Figure 11. The overall lack of trend is consistent with absence of population wide iris-ageing. The second inadequacy is that there could be some individuals (a minority) who are actually ageing quite rapidly even though the population apparently is not. Two approaches are used to detect such possibilities. First is a simple test for increasing scores. Second is mixed-effect regression approach which additionally includes the effect of pupil dilation.

Figure 10 shows the fraction of the population producing higher second half scores than first half. Specifically if recognition data is available on the interval $[T, T + T_A]$, we use a non-paired one-sided Wilcoxon rank sum test of whether the scores on $[T + T_A/2, T + T_A]$ have higher median than those on $[T, T + T_A/2]$. A significant result would indicate an upward trend in an individual trajectory.

The Figure shows the cumulative distribution function of p-values from this test. For eyes enrolled on camera B, about 1% of the 50,850 eyes for which at least 20 recognition transaction exist over at least 1 year exhibit a significant increase, if $p < 0.01$ is held as significant. This high figure is similar when T_{\max} is extended to include only long-term users. On camera L the figures are lower possibly because personal standard deviations are lower - Table 2 shows the population average $\sigma = 0.030$ for camera L and $\sigma = 0.040$ for camera B - and this makes increase-detection more difficult.

²⁰The early and late fluctuations in FNMR are due to significance - see the transaction counts in Figure 8

4.2.1 Individual-specific results for OPS-XING

This section applies mixed-effects regression to time-series data of the kind shown in Figure 4. This is intended to detect iris ageing effects in the presence of measurement noise by regarding observed iris recognition dissimilarity scores from each eye as samples drawn from a potentially non-stationary process, i.e. one whose statistics vary. This approach has been noted previously[16]. We consider only wide-sense stationarity examining only the mean and variance.

Figure 4 shows eyes for which 60 or more transactions succeed over five years. It is clear that these distributions are sometimes not stationary. The possible causes of this include changes in the illumination environment, in the anatomical source (iris and pupil), and in behavioral interaction with the camera (i.e. habituation[51]²¹). Figure 11 is consistent with human acclimation to the system. It shows the genuine distribution as a function of the frequency of use. For those using the system once for year or less, the mean HD is 0.154; for those using it more frequently than once a month the average HD is 0.129. The default conclusion from this is that frequent travelers have a better conscious or unconscious understanding of what it takes to produce a high quality iris image. These subjects may, for example, remove their eye-glasses, open their eyes widely, reduce head motion, or intuit a better idea of where the focal plane lies. While the observed score effect could have a correlation with age, sex, or profession, the leading hypothesis, given quantitative[51] and anecdotal evidence, is that habituation is responsible.

Longitudinal analysis addresses the case where many response variables (e.g. patients' responses to a drug) are modeled given (many) fewer repeated measurements. It is advanced here for biometric ageing because it is capable of handling multiple responses (i.e. subjects or eyes) that are not balanced (subjects are imaged irregularly over time), and potentially correlated through time (a Hamming distance today may be correlated with later ones) and across units (e.g. between left-and-right eyes). In these respects, the generalized linear mixed effects model (GLMM)[30] is appropriate. The "generalized" part is needed for non-Normal responses (e.g. Bernoulli recognition decisions). The "mixed" part of the model refers to "fixed" and "random" effects, the former capture the population-wide variation, while the latter models individual-specifics. Examples of fixed covariates in iris recognition would be camera and type of ambient illumination. The random effects part gives subject-specific regression effects and these, when combined with the fixed effects give the mean response of an individual.

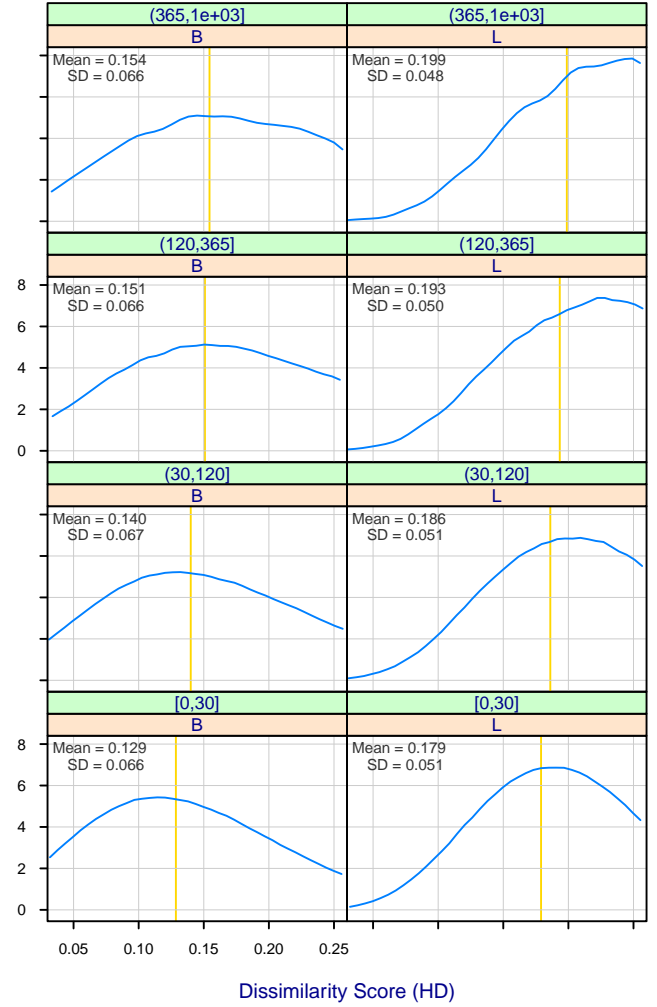


Figure 11: **Frequent travellers, better scores:** Hamming distance distributions by frequency of use, and by enrollment camera. The vertical line is the mean. The $[a,b]$ intervals indicate the average time, in days, between uses of the system - frequent travelers are at the bottom. Note the leftward shift with frequency.

²¹Kukula et al.[51] chart the time-evolution of comparison scores for a hand geometry authentication system, and conclude "that repeated use of the [hand geometry] device yields some increase in the performance".

Additionally covariates may be constant e.g. iris color or gender, or time varying. Particularly the time since enrollment will take on different values at each capture. These covariates might be either under experimental control (e.g. a different camera is used) or stochastic (governed by a random process). Pupil dilation falls into this category and the dilation difference (eq. (4)) is therefore also stochastic²². The use of stochastic covariates as one of the explanatory variables can lead to difficulty in interpretation of regression results because the mean iris comparison score, conditioned on all the covariates, should depend only on the covariates at that time. This would not be true for a stochastic covariate where the response variable causes a change in the covariate (e.g. a high cholesterol measurement causes an altered fat intake) or if the response is dependent on historical values of the covariate. When the covariate and the response are confounded, regression coefficients (β) can be biased and lose their implied causal role. A time varying covariate is called external when its value is not predicted by past values of the response variable[30] - for dilation this would mean that the act of sensing an iris (and comparing it to a template) does not influence future dilation values. This assumption is not true on a timescale of seconds because some cameras specifically illuminate the iris with visible light. But at longer timescales we do not expect repeated use of an iris camera to induce changes in dilation, nor do we expect humans to exercise conscious (or otherwise) control over pupil dilation during camera usage.

We apply mixed-effects approaches to the longitudinal evolution of the dissimilarity between the enrollment and search images. Mixed-effects models have been used in biometrics previously for failure analysis[9] but not longitudinal change.

Given n_i dissimilarity scores for the i -th eye, we model the j -th score

$$d_{ij} = \beta_1 + \beta_2 T_{ij} + \beta_3 \Delta D_{ij} + \beta_4 \Delta D_{ij}^2 + \beta_5 D_{ij} + \beta_6 D_{ij}^2 + \beta_7 C(r) + b_{i1} + b_{i2} T_{ij} + b_{i3} \Delta D_{ij} + b_{i4} D_{ij} + b_{i5} D_{ij}^2 + e_{ij} \quad (6)$$

where

The β_k values are fixed over the population, and the b_{ik} values are eye-specific.

the sum of the fixed and random effects, $\beta_1 + b_{i1}$, is an eye-specific intercept corresponding to initial status after zero elapsed time.

Given n_i times *between* captures, T_{ij} , the sum, $\beta_2 + b_{i2}$, is an eye-specific gradient expressing rate-of-change of dissimilarity with time. The value β_2 is the population average rate that is the primary focus of this paper. This model uses a linear form for three reasons. First, we expect iris ageing to be a continuous monotonic process. Second, we have no prior (anatomical or physiological) justification for a more elaborate functional form. Third, model-parsimony dictates a simple intercept and gradient model.

Dilation differences, ΔD , are included with linear and quadratic terms²³ in the fixed part of the model, and with just a linear term in the random-effects part. This allows the dilation dependence of Figure 7 to be captured across the population with an eye-specific linear adjustment.

The dilation of the search image, D , is represented by linear and quadratic terms. The dilation of the enrollment image is included only via the ΔD term. An alternative, to incorporate a model of the joint dilation shown in Figure 6, was not considered, but may yield benefits in future work.

$C(r)$ is the contrast for the eye label, 0 for left and 1 for right.

²²Dilation is computed from consensus radius estimates for the first and second image (eq. 2). Dilation could itself be the response variable in a regression analysis, but without covariates such as ambient light, recent proximity to sunshine, recent pharmaceutical consumption, this cannot be pursued here.

²³These terms are actually included using orthonormal polynomials, which satisfy the requirement that covariates are linearly independent. The coefficients represented in eq. 6 and in the results tables are converted to a raw polynomial form $a + bx + bx^2$ per [45].

ΔT	T_A	Nmin	Cam	β_1	$\beta_2 = \text{Coef}(\Delta T)$	$\text{Coef}(D_2)$	$\text{Coef}(D_2^2)$	$\text{Coef}(\Delta D)$	$\text{Coef}(\Delta D^2)$	$\text{Coef}(\text{Eye})$	Nscr	Neyes
5	182	5	B	0.13	-4.0E-07 \pm 1.0E-07	-0.37	0.43	0.08	0.31	0.028 \pm 0.000	3583505	150442
5	182	7	B	0.13	-3.4E-07 \pm 1.1E-07	-0.37	0.44	0.08	0.32	0.030 \pm 0.000	3402362	122357
365	182	20	B	0.13	-4.9E-07 \pm 1.3E-07	-0.43	0.50	0.07	0.36	0.038 \pm 0.001	2448914	48448
730	182	20	B	0.13	-2.6E-07 \pm 1.4E-07	-0.43	0.51	0.07	0.36	0.038 \pm 0.001	2124322	39909
1095	182	20	B	0.13	1.2E-07 \pm 1.4E-07	-0.45	0.53	0.07	0.37	0.039 \pm 0.001	1609932	28240
1460	182	20	B	0.13	7.5E-07 \pm 1.8E-07	-0.47	0.55	0.07	0.38	0.038 \pm 0.002	925737	14601
1460	182	40	B	0.13	7.2E-07 \pm 2.2E-07	-0.53	0.62	0.06	0.41	0.048 \pm 0.003	741702	8271
1460	182	60	B	0.12	7.2E-07 \pm 2.6E-07	-0.57	0.67	0.06	0.42	0.045 \pm 0.005	584651	5104
1460	182	80	B	0.12	7.4E-07 \pm 3.2E-07	-0.61	0.71	0.06	0.43	0.043 \pm 0.008	461792	3337
1460	182	100	B	0.12	8.0E-07 \pm 3.9E-07	-0.61	0.73	0.05	0.45	0.037 \pm 0.009	360694	2213
1460	1460	40	B	0.13	7.8E-07 \pm 2.2E-07	-0.53	0.63	0.06	0.41	0.048 \pm 0.004	715612	7876
1460	1460	60	B	0.12	7.3E-07 \pm 2.7E-07	-0.57	0.67	0.06	0.42	0.041 \pm 0.005	569897	4945
1460	1460	80	B	0.12	7.6E-07 \pm 3.3E-07	-0.61	0.71	0.06	0.43	0.037 \pm 0.008	452292	3255
5	182	5	L	0.18	1.5E-06 \pm 9.5E-08	-0.37	0.52	0.03	0.28	0.011 \pm 0.000	1318689	35064
5	182	7	L	0.18	1.7E-06 \pm 9.7E-08	-0.38	0.53	0.03	0.28	0.012 \pm 0.001	1291124	30800
365	182	20	L	0.18	2.4E-06 \pm 1.1E-07	-0.42	0.58	0.03	0.30	0.020 \pm 0.001	1123839	17812
1095	182	20	L	0.18	2.5E-06 \pm 1.0E-07	-0.42	0.58	0.03	0.30	0.024 \pm 0.001	1115230	17598
1825	182	20	L	0.18	2.4E-06 \pm 1.2E-07	-0.41	0.58	0.03	0.30	0.020 \pm 0.001	1022768	15577
2190	182	20	L	0.18	2.7E-06 \pm 1.4E-07	-0.42	0.59	0.02	0.30	0.020 \pm 0.001	750745	10952
2555	182	20	L	0.18	3.0E-06 \pm 1.8E-07	-0.44	0.61	0.02	0.30	0.018 \pm 0.001	448513	6042
2920	182	20	L	0.17	3.0E-06 \pm 3.4E-07	-0.48	0.67	0.02	0.32	0.024 \pm 0.004	94201	1360
1460	182	40	L	0.17	2.6E-06 \pm 1.4E-07	-0.47	0.66	0.02	0.31	0.025 \pm 0.002	883057	9847
1460	182	60	L	0.17	2.9E-06 \pm 1.6E-07	-0.53	0.72	0.02	0.33	0.026 \pm 0.002	699788	6145
1460	182	80	L	0.17	2.8E-06 \pm 2.0E-07	-0.57	0.78	0.02	0.33	0.024 \pm 0.003	550440	3999
1460	182	100	L	0.17	2.8E-06 \pm 2.4E-07	-0.61	0.83	0.02	0.34	0.021 \pm 0.004	439132	2756
1460	1095	100	L	0.17	2.8E-06 \pm 2.4E-07	-0.62	0.85	0.01	0.34	0.020 \pm 0.004	435002	2725
1460	1460	100	L	0.17	2.8E-06 \pm 2.3E-07	-0.61	0.83	0.02	0.34	0.025 \pm 0.005	423363	2644

Table 3: **Population average ageing rates:** Regression results for subsets of the OPS-XING dataset. Each row specifies a particular set of eyes (i.e. unique combinations of person and left-or-right eye) for which there are Nmin transactions spread over at least T_A days, and for which the last transaction occurs at least (ΔT) days after enrollment. Each row shows mixed-effects regression coefficients for the eq. (6) model, particularly the rate coefficient for the time between enrollment and search, ΔT , dilation change, ΔD (eq. 4), and for the right eye (left is zero). The comparison scores were produced by comparing recognition samples against a singular initial enrollment image. The key result, highlighted in yellow, is that population averaged changes in Hamming distance are small - the largest β_2 value, 3×10^{-6} corresponds to $\Delta HD = 0.011$ over 10 years. The uncertainty estimates on these rates are square roots of the variance of eye-specific ageing rates, b_i .

e_{ij} are residuals.

The model represented by eq. (6) was applied separately for enrollment cameras L and B, because the evident complex interaction between HD, dilation, and camera in Figure 7. A large number of models were considered, this one was selected on the basis of the significance of the various terms, parsimony, and the Akaike Information Criterion[70] summary. We also considered auto-correlation of the response by using the exponential model of auto-covariance. This gave a change in the standard deviation of random effects in the fourth decimal place. It is computationally expensive so an uncorrelated within-subject response was used for all results with this dataset. This decision was supported by inspection of individual auto-correlation functions, which do not exhibit lagged correlations. Note that while ordinary least squares (OLS) estimators are unbiased predictors of true longitudinal change, the mixed-effects models estimates tend to shrink from the OLS values, when the observations are correlated. The models were fit with version 3.1 of the NLME[70] package running under version 2.15.1 of R. Restricted maximum likelihood estimation was used.

If this analysis is restricted only to left eyes, omitting the eye as a covariate in eq. (6), the intercepts and dilation coefficients change in the last decimal place. The population means for rate-of-change $\beta_2 = \text{Coef } \Delta T$ are higher by no more than 0.5×10^{-7} .

Table 3 gives population means produced by applying the mixed-effects model of equation (6) to various subsets of the OPS-XING logs. The notable observations are:

Rate of change: This paper seeks to estimate the time-dependence of iris recognition due to changes in the iris texture. This is quantified as a rate of change of dissimilarity score. Its calculation has been based on averaging results over as large a population as possible on the basis that random effects, not otherwise accounted for, will be unbiased. The sixth column of Table 3 tabulates β_2 population means varying between -4×10^{-7} to 8×10^{-7} for persons enrolled on camera B and 1.5×10^{-6} to 3×10^{-6} for those using camera L. The values quantify change in Hamming distance per day. They vary slightly with the subset of the data used. Per equation (8), wide temporal distribution datasets offer better precision in the measurement, and thus the latter rows (with large T_A values) in the Table are more interesting. However, the last columns show that fewer individuals eyes are engaged for those durations. Longer engagements of persons enrolled with camera L produce more precise estimates.

For the two cameras, the rates differ by a factor between three and four: 0.8×10^{-6} for camera B vs. 3.0×10^{-6} for camera L. The reason for this is not obvious; it is not a degradation in the camera as the L/B label applies to initial enrollment and camera B is used for all recognition transactions. Instead the rate difference may be related to the higher overall HD values for camera L vs. B. Subjects enrolled on camera L have longer average time durations, T_{\max} , but this should not effect the rate estimate if ageing is a continuous linear process. Further investigation is warranted.

Camera effect: The β_1 values are HD-intercepts treated as a fixed effect. The estimate for the B camera is about 0.12, and that for the L camera is 0.17. Recall that these labels indicate the camera used for the initial enrollment image, and all recognition images were collected using camera B. Thus, the difference here is most likely the direct result of cross camera (B,L) interoperability effects - see section 2.1.1. This systematic difference in Hamming distances, $\Delta HD = 0.05$, is an order of magnitude larger than any ageing-related effects projected over a decade.

Dilation difference: The dependence of HD on pupil dilation differences is modelled as an individual eye-specific effect, using eq. (4). The population means are camera specific: the coefficient for camera B is 0.18, and for camera L, 0.11. This means that if the enrollment and recognition images differ in pupil dilation by 0.1 (say $D = 0.25$ vs. 0.35), then this model indicates Hamming distance changes of $\Delta HD = 0.018$ and 0.011 , for cameras B and L respectively.

Search image dilation: While dilation difference is the primary contributor to increases in Hamming distance, Figure 7 shows that low and high values in both images will also produce an increase. The regressions over the various OPS-XING subsets produce somewhat varied dependences on dilation which limits the precision of the model. This part of the model is approximately $\Delta HD = -0.5D + 0.7D^2$ (via inspection of rows in the Table).

Left vs. right eye: The eye itself, left or right, is influential: right eyes give higher HD values - over the whole dataset this is 0.03 for camera B, and 0.02 for camera L. This occurs because the right eye is used only if the left eye failed or was not acquired. The number of left eye events in the OPS-XING database is 4,920,638 vs. 725,300 for the right eye.

We now address variation in eye-specific gradients. Figure 4 shows instances of upward and downward trend in HD values. Moreover, that Figure applies to people with at least 60 data points spanning $T_A \geq 1460$ days and, on that basis, the gradients are “robust”. Those plots are estimated using ordinary least squares (OLS) regression (for exploratory purposes) without dilation. OLS assumes independence and homoscedasticity of the residuals. However when these assumptions are violated the intercept and gradient estimates are still unbiased but have worse efficiency i.e. they’re noisy estimators of the underlying true values[85].

However it is known that if the residual e_{ij} is the difference between the HD for the j -th observation of the i -th individual and the value predicted by the OLS fit at that time, then the variability of HD around the OLS-estimate is summarized for

the i -th eye as

$$\sigma_{\epsilon_i}^2 = \frac{1}{N} \sum_{j=1}^{n_i} e_{ij}^2 \quad (7)$$

whence the precision of the OLS-estimated gradient for person i is given by the sampling variance for that statistic[85]:

$$V_i(m) = \frac{\sigma_{\epsilon_i}^2}{\sum_{j=1}^{n_i} (t_{ij} - \bar{t}_i)^2} \quad (8)$$

Thus the precision of the OLS-estimate will be increased if either the numerator's "measurement" error can be reduced, or the denominator's statement of temporal extent can be increased. The term "measurement" error here is the residual difference between the measured HD and an ideal HD which would arise if a flawless iris was presented consistently every time (no eyelashes, frontal gaze, constant dilation, etc).

The nature of equation (8) applies to the mixed-effects results tabulated in Table 3, i.e. the long-term data gives a narrower range of iris rate-of-change estimates. Figure 12 shows the distribution of the $\beta_2 + b_{i2}$, i.e. the population-mean plus the best linear unbiased predictor (BLUP) for each eye. The notable observations are:

Variance: The gradients for individuals enrolled on camera B have larger variance than camera L. The distributions for camera L are narrower because a) scores tend to vary less (mean personal $\sigma_L = 0.03$ vs. $\sigma_B = 0.04$ in Table 2), and b) the subjects are present in the dataset over longer time periods, i.e. average T_A is larger than that for camera B. If the minimum active duration T_A is reduced below the 1460 days shown in the Figure, the distributions become broader (following eq. (8)).

Worst case upward trends: Some individuals age quickly with rates $\beta + b_i \geq 2 \times 10^{-5} \text{ day}^{-1}$. This is not a significance artefact - instances of ageing at these rates are present in the data - see Figure 4²⁴.

The reasons for ageing on these timescales in populations this large will inevitably include some cases of disease. Other causes would include contact lens presence, changes in contact lens prescription, changes in use of eye-lash cosmetics and eye surgery. In addition, the regression may yield these gradients simply on the basis of the (noisy) data. But note that in many cases the individual is capable of producing a low HD score at the end of the interval. This indicates that the iris itself has not aged to the point that successful recognition is impossible.

Downward trends: Note also that some individuals have downward trends. This, by definition, is consistent with a collection of an image similar to the initial enrolment, which would occur under habituation, e.g. the subject learns to open eyes widely.

Normality: The QQ-plots of Figure 12 indicate the gradient predictors are approximately Normally distributed through $\pm 2\sigma$ but with heavier-than-Normal tails. Beyond the suggested causes above, we cannot further explain such outcomes because images are unavailable.

²⁴See individual 133979L but note there that the gradient value appearing above the points does not account for dilation changes. The mixed effects model does include dilation effects.

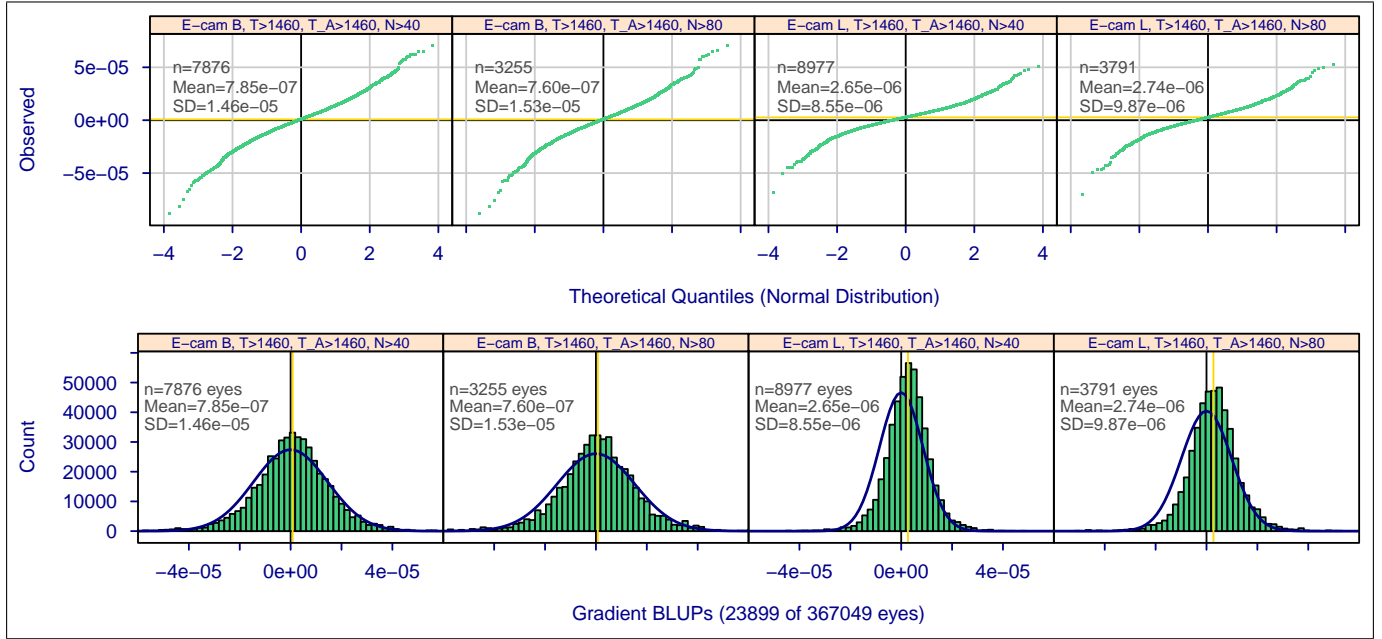


Figure 12: **Variable ageing rates:** For various subsets of the OPS-XING data, the lower figure shows the distribution of the best linear unbiased predictors (BLUP) rates-of-change $\beta_2 + b_{i2}$ as computed for each eye in the model of eq. (6). The left side applies to the Panasonic BM-ET 330 camera (B), the right to the LG-2200 (L). Only those eyes for which at least the given N recognition transactions span at least $T_A > 1460$ days are included. The vertical yellow and black lines are separated by the value of β_2 since $E[b_{i2}]$ is zero. The upper figures are QQ plots of the empirical values vs. quantiles of a Normal distribution. These probably constitute the best estimates given this dataset. The distributions of L are narrower than B because the T_A values are longer on average.

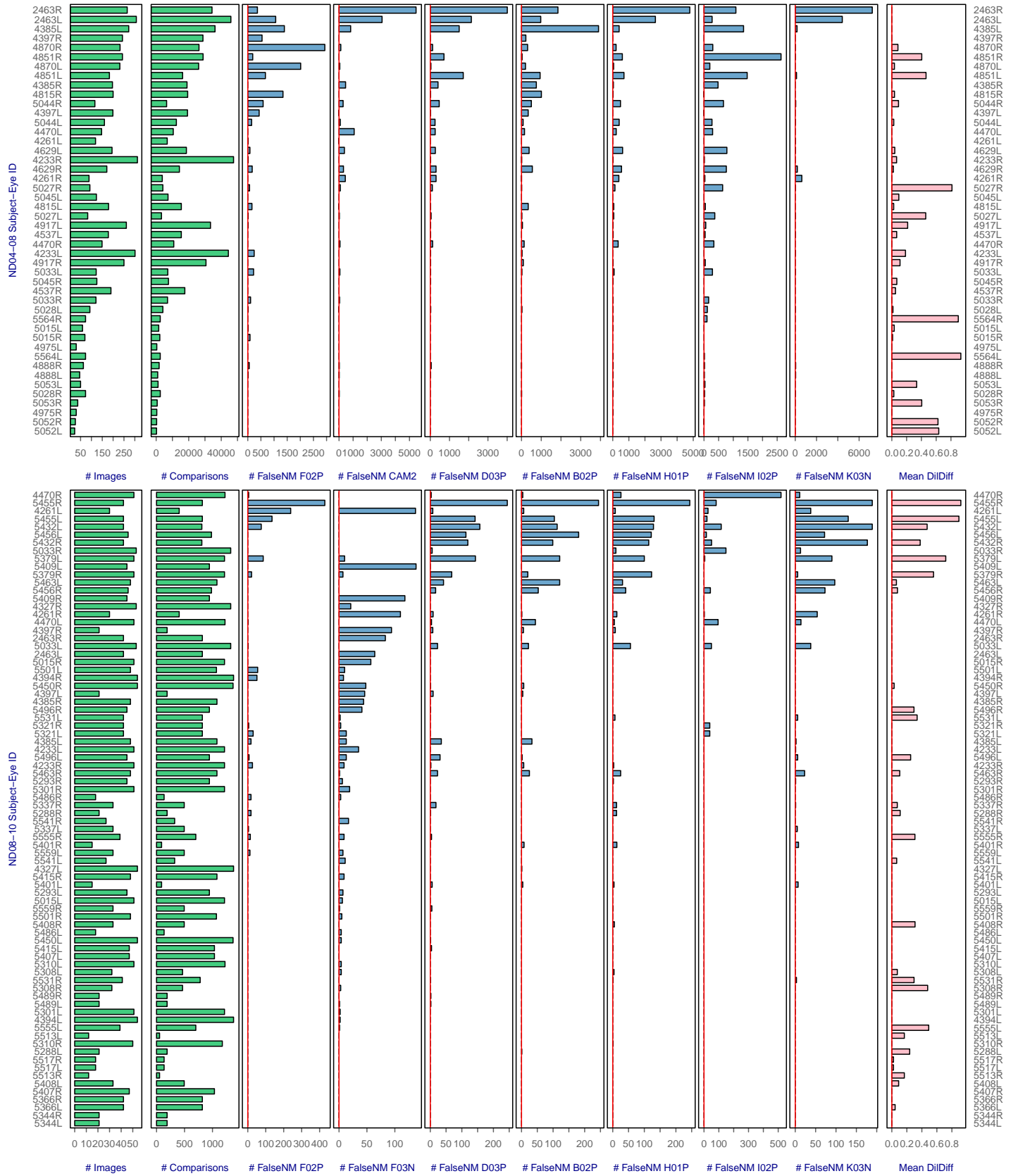


Figure 13: **Goats in the biometric zoo:** For all eyes in the ND04-08 dataset (above) and the 3-semester subjects of the ND08-10 dataset (below), the barcharts show, from left to right: the numbers of images of the eye; the number of non-self comparisons involving those images; the FNMR values for various matchers whose thresholds have been set to give FNMR= 0.02 over the whole dataset; and the mean dilation difference between image pairs, per eq (4). The concentration of errors in individuals indicates “goats” in a “biometric zoo”[25, 99]. As shown by the vertical red lines, FNMR = 0 for the majority of subjects.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

4.3 Results for ND recognition

This section aims to confirm and explain the observations reported by the researchers at Notre Dame for image sets collected in their laboratory[8, 7, 28, 29]. The analysis proceeds by applying c. 2012 commercial recognition algorithms to the archived ND images. In addition, we analyze archived results from the application of the CAM-2 algorithm submitted by University of Cambridge to the 2006 Iris Challenge Evaluation (ICE)[68].

Figure 13, is exploratory in nature. Its first two columns indicate an imbalance in the number of images collected from each individual. This has a quadratic influence on the number of comparison scores available for each individual, due to cross-comparison. This method of generating genuine scores represents an efficient use of the data but is not usually an operational use-case. The OPS-XING collection matches an initial enrollment instance with multiple recognition captures captured over many years. If this yields n_i genuine scores for the i -th individual, the total number is $\sum_i^N n_i$. For cases where images are retained and matched in a laboratory, as with ND, the standard experiment makes efficient use of the data by executing all genuine comparisons, yielding $\sum_i (n_i^2 - n_i)/2$ scores. This has implications for significance tests where different components of uncertainty scale as the number of comparisons, images, eyes, or people. This contrasts with many operational cases where specific care, review and re-collection are key aspects of the enrollment process.

By plotting the count of false non-matches for each eye when thresholds are set to give a false non-match rate FNMR of 0.02 over the entire dataset, columns 3 through 9 of Figure 13 show that multiple recognition algorithms produce false non-match errors but that these errors are concentrated in the eyes of relatively few individuals. For example, in the 2004-2008 collection, algorithms K02P, H01P and CAM-2 give false non-matches essentially on only a single individual, 02463. Concentration of false non-matches in certain individuals renders them “goats” in the biometric zoo[25]. The presence of such subjects in the ND sets has not been previously published, and while this analysis is not explicitly time dependent, it adds a qualifier to the ND findings, particularly that FNMR increases by 82% in the period 2008-2010²⁵.

The two plots of Figure 13 also reveal a correlation between the individuals that participated most and their FNMR, particularly in the first 2004-2008 collection. The explanation of this is not known. One hypothesis is that repetition engenders ennui or blasé behavior and this manifests itself in the images, for example as poor axial gaze, or poor focus due to sub-optimal positioning relative to the depth of field.

Time Evolution: Figures 14 and 15 show the evolution of the genuine score distributions as a function of time. Note that the scores in the 2004-2008 set are actually lower than in the later set²⁶. The lack of a clear trend away from the global median is consistent with a no-ageing result, but is insufficient to conclude that no ageing is occurring because it could be confined only to a minority of individuals. This prompts the production of four full-page figures 16- 19 showing eye-specific time-series for all individuals in the two ND partitions.

Figures 16 and 17 show the time evolution of comparison scores for two recent IREX IV algorithms (I02P and D03P) applied to the ND images collected 2004-2008. Similarly Figures 18 and 19 show the time evolution of comparison scores for two recent IREX IV algorithms (I02P and D03P) for the 40 individuals who were imaged at Notre Dame over the two year period, Spring 2008 to Spring 2010, many of whom where also captured in Spring 2009. This is exactly one partition reported in Table 1 of Fenker[29]. These figures, and those for six other algorithms included in the Appendices²⁷, include annotations:

²⁵The 82% figure is taken from Table 2 in [29] for a Neurotechnology algorithm. The paper notes a 95% confidence interval for FNMR of [38,150]%. Note that this confidence interval used the bootstrap method, sampling individuals and scores. It should also[53] have sampled *images* which are the causative agent of recognition failure, and thereby did not confer the robustness against outliers that the bootstrap usually affords.

²⁶ The early ND data included many poor images because the LG2200 camera’s internal quality controls had been disabled. Noting this was inappropriate in an ageing study, Baker et al.[8] produced a dedicated subset of images that “were manually screened for image quality” and this is the set used in this report. Additionally, because the ND04-08 data has been made available to researchers it is likely to have been used by all providers whose algorithms are used here.

²⁷See the PDF file linked from <http://iris.nist.gov/irex>

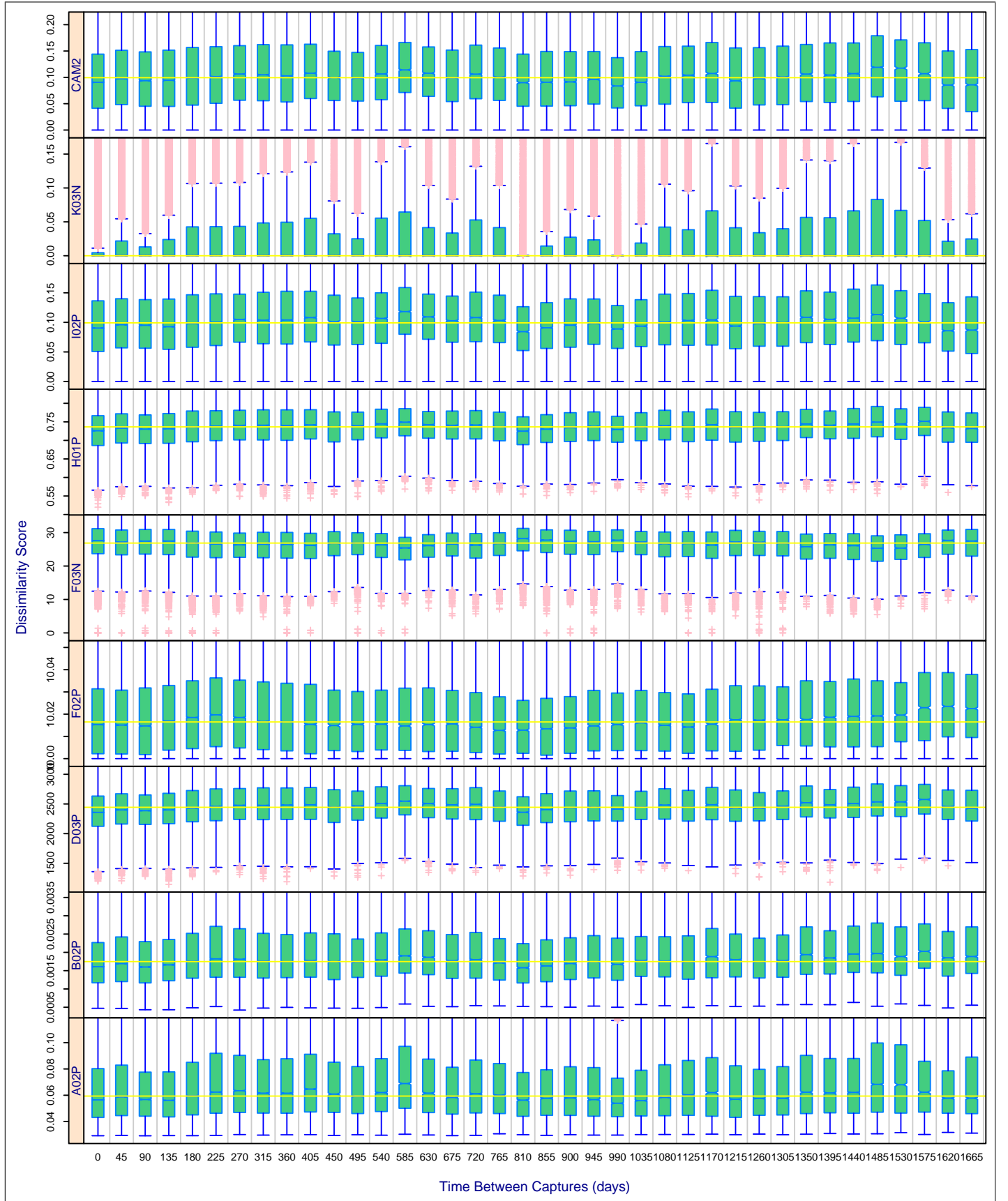


Figure 14: **Stability of score distributions over time:** Time-evolution of genuine comparison scores for the ND04-08 images. The nine panels correspond to eight c. 2012 algorithms submitted to NIST's IREX IV evaluation [74], and one, CAM-2 (top), submitted in 2006 to ICE[68]. The green boxes denote the interquartile range, their center lines indicate the median. The yellow line is the all-time median. Outliers, in the distribution tails, are shown as pink crosses. The K03N algorithm emits most genuine scores with value exactly 0.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

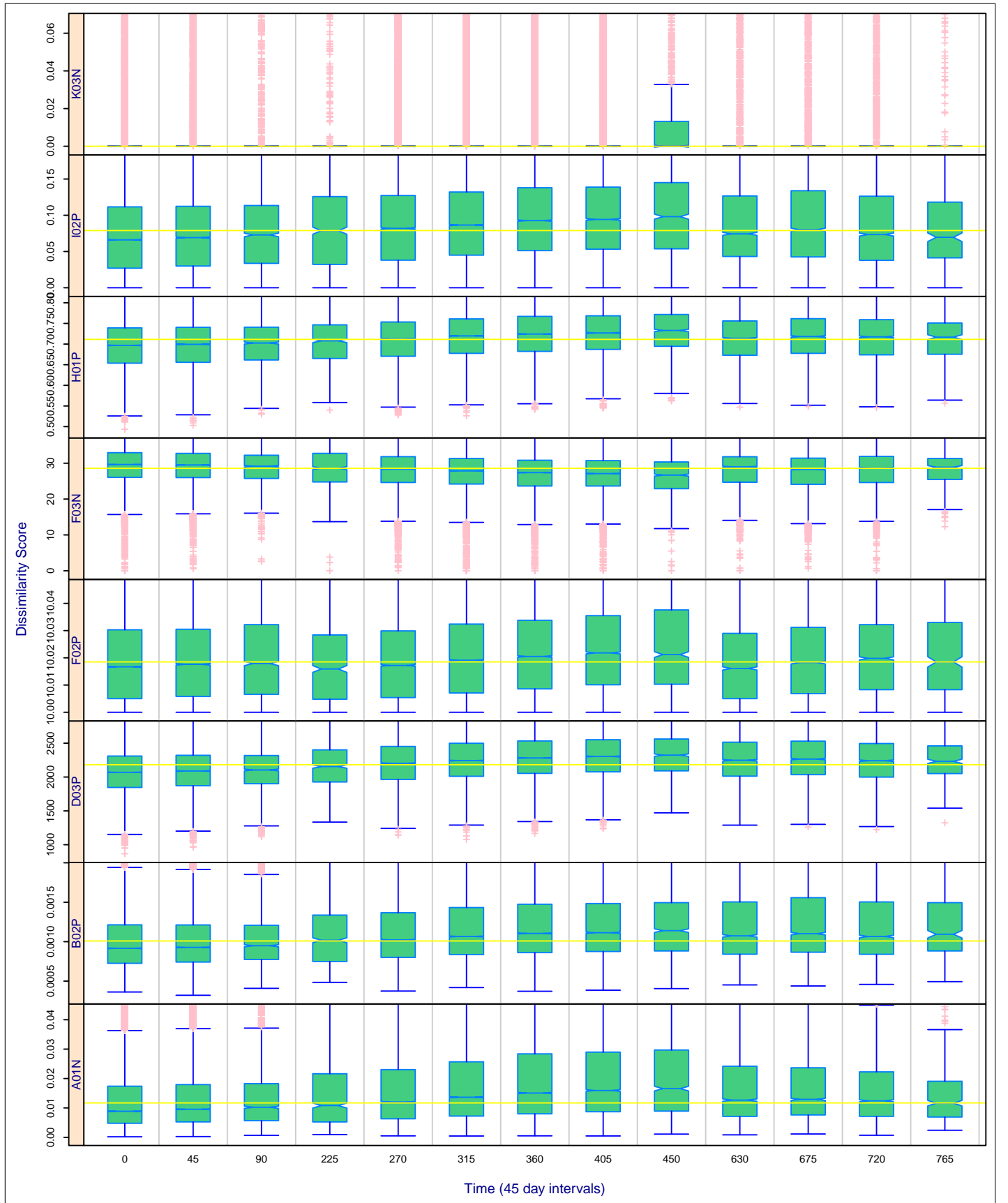


Figure 15: **Stability of score distributions over time:** Time-evolution of genuine comparison scores for the ND08-10 images. The nine panels correspond to eight c. 2012 algorithms submitted to NIST's IREX IV evaluation [74], and one, CAM-2 (top), submitted in 2006 to ICE[68]. The green boxes denote the interquartile range, their center lines indicate the median. The yellow line is the all-time median. Outliers, in the distribution tails, are shown as pink crosses. The K03N algorithm emits most genuine scores with value exactly 0.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

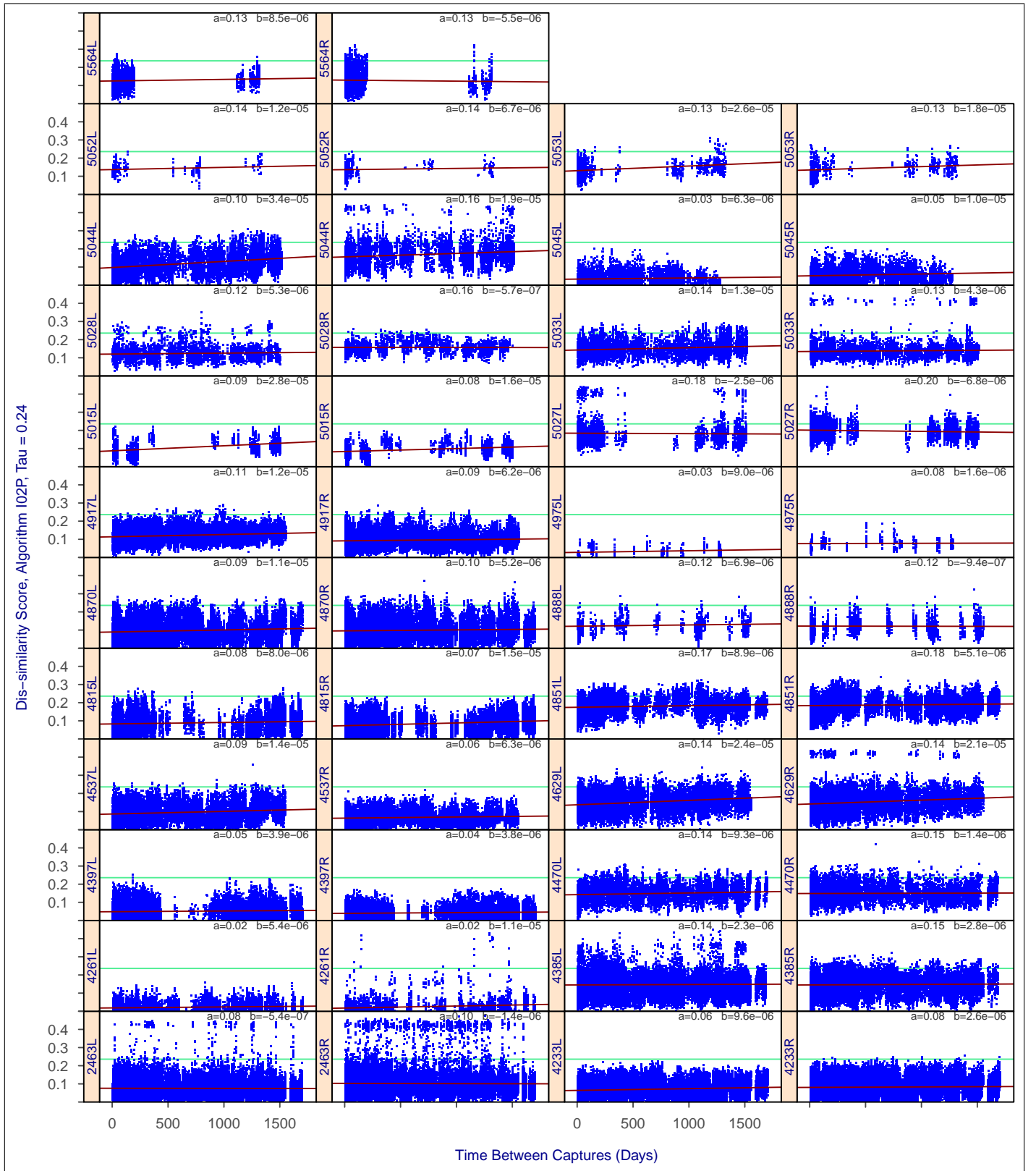


Figure 16: Individual score trajectories: For dataset ND04-08, the plots show individuals' dissimilarity scores for the I02P algorithm. The x-axis indicates time between two samples. Each dot corresponds to a pair of images. Any given image, i , is involved in $n_i - 1$ comparisons, i.e. one fewer than the number of images of the given eye. The green horizontal line shows the threshold that gives FNMR = 0.02 globally. The straight red line is the result of regressing $d = d(T)$, the intercept and gradient of which appear as text. Note the bimodal score distribution in panels 5044R, 5027L, 02463L+R i.e. the occurrence of HDs near 0.4. These are consistent with outright segmentation failure. Their use in ageing rate computations is not recommended irrespective of the time-lapse at which they occur.

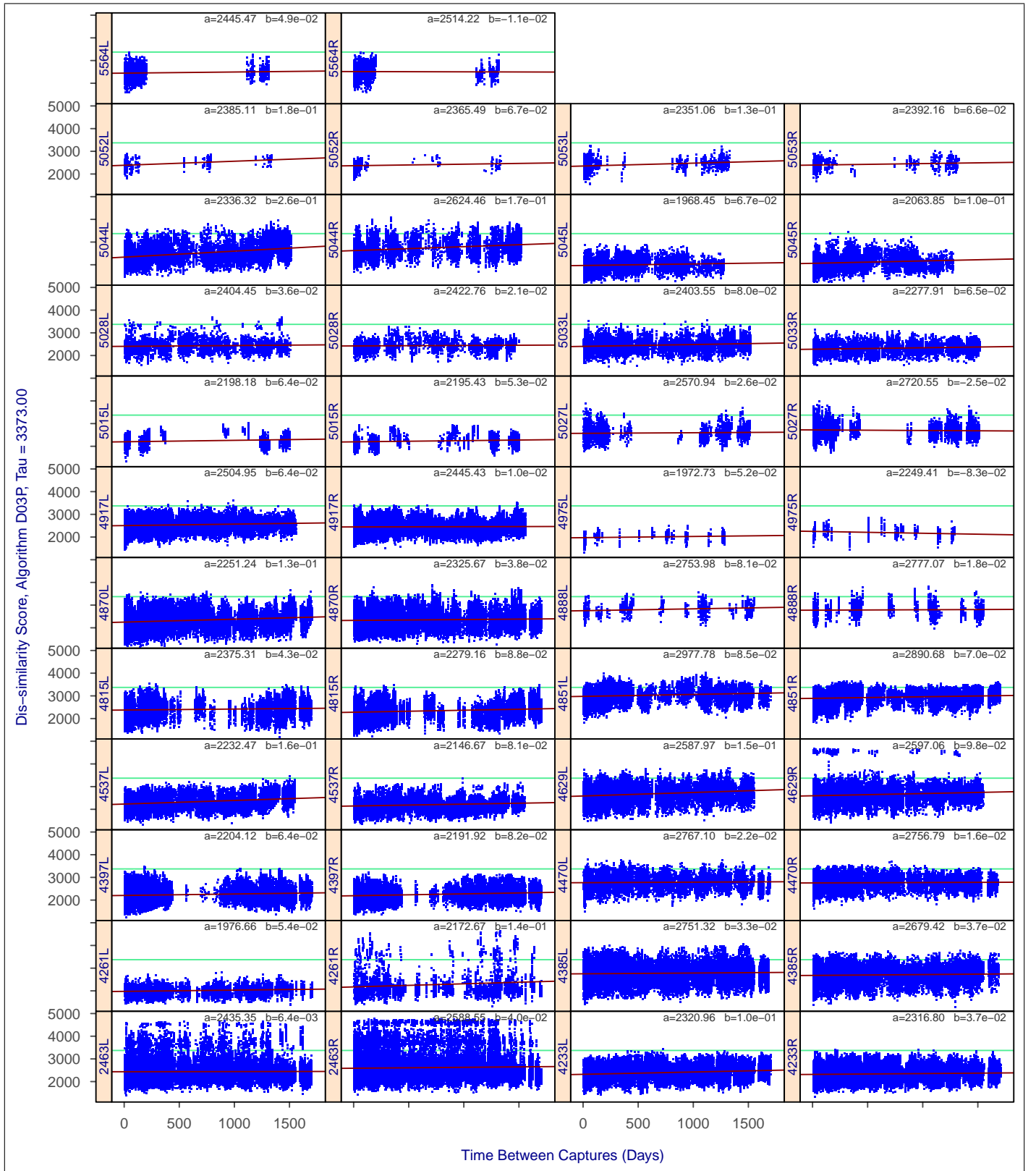


Figure 17: **Individual score trajectories:** For dataset ND04-08, the plots show individuals' dissimilarity scores for the D03P algorithm. The x-axis indicates time between two samples. Each dot corresponds to a pair of images. Any given image, i , is involved in $n_i - 1$ comparisons, i.e. one fewer than the number of images of the given eye. The green horizontal line shows the threshold that gives FNMR = 0.02 globally. The straight red line is the result of regressing $d = d(T)$, the intercept and gradient of which appear as text. Note the concentration of false non-matches in individual 02463, 05044R, 04385L.

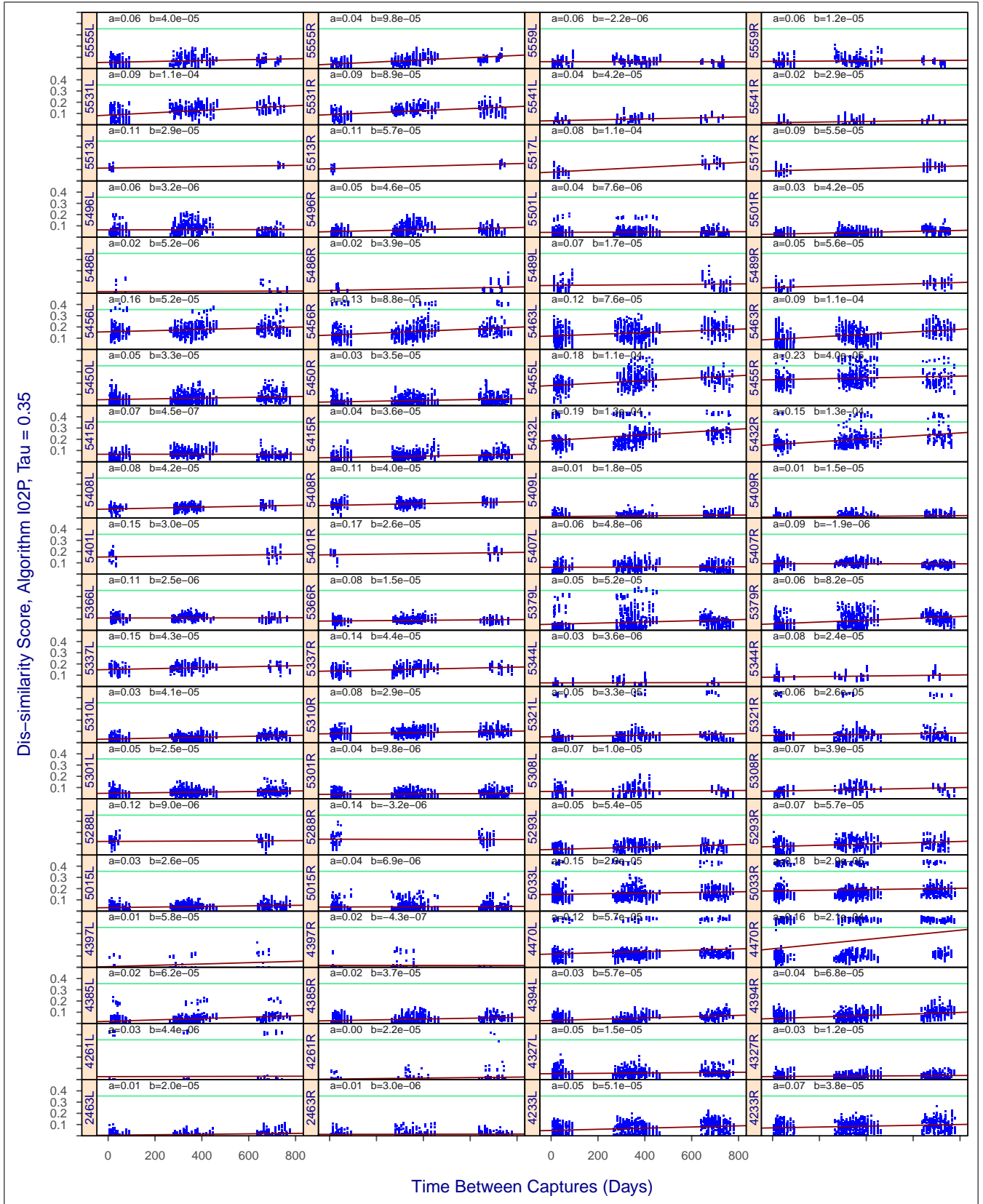


Figure 18: **Individual score trajectories:** For dataset ND08-10, the plots show individual dissimilarity scores for the I02P algorithm. The x-axis indicates time between two samples. Each dot corresponds to a pair of images. Any given image, i , is involved in $n_i - 1$ comparisons, i.e. one fewer than the number of images of the given eye. The green horizontal line shows the threshold that gives FNMR = 0.02 globally. The straight red line is the result of regressing $d = d(T)$, the intercept and gradient of which appear as text. Note the bimodal score distribution in panels 5432L+R, 4470L+R and 5033L+R i.e. the occurrence of HDs near 0.4. Their use in ageing rate computations is not recommended irrespective of the time-lapse at which they occur.

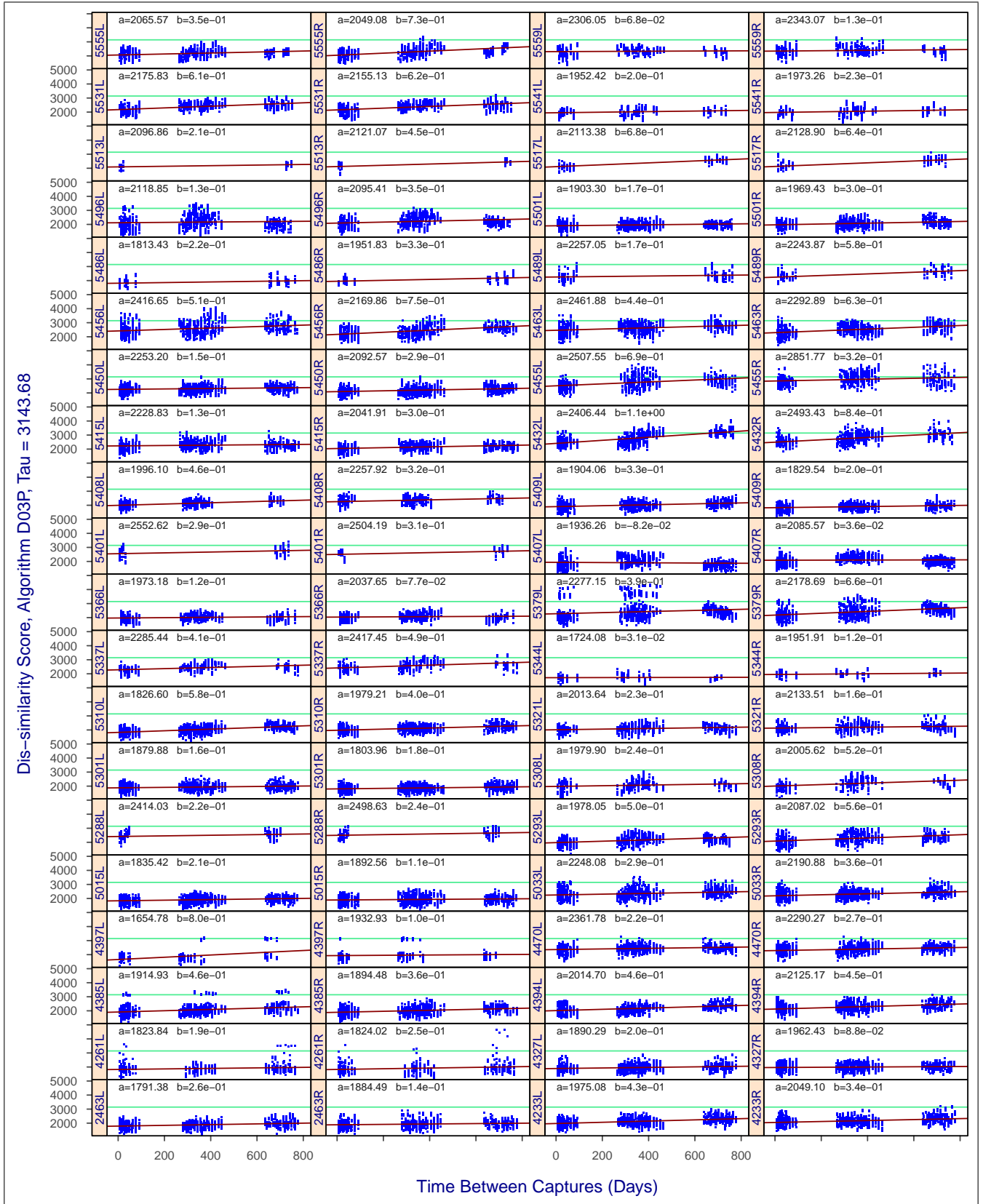


Figure 19: **Individual score trajectories:** For dataset ND08-10, the plots show individuals' dissimilarity scores for the D03P algorithm. The x-axis indicates time between two samples. Each dot corresponds to a pair of images. Any given image, i , is involved in $n_i - 1$ comparisons, i.e. one fewer than the number of images of the given eye. The green horizontal line shows the threshold that gives FNMR= 0.02 globally. The straight red line is the result of regressing $d = d(T)$, the intercept and gradient of which appear as text.

Threshold: The fixed green horizontal lines indicate thresholds set to give FNMR = 0.02 over the entire dataset (including individuals for which less than two year data was available). The exact value is not important, the goal being to look for trends about that line. The threshold value appears in the y-axis label.

Regression: The straight red line indicates the result of a simple regression of dissimilarity score against elapsed time. It is included here for visualization and is not intended as a definitive regression model. The text shows the intercept, a , and gradient, b .

Inspection of the Figures 16, 17, 18 and 19 reveals several notable effects.

A biometric zoo: As noted previously in Figure 13 false non-match errors are concentrated in fewer than about 10 individuals for the algorithms. Some individuals' time-series show that scores are always low (e.g. 04233, 04379, 05015, 05045 in ND04-08). Others have large numbers of false rejections. The term "biometric zoo"[25] and "menagerie"[99] are used to describe error heterogeneities in biometric errors, here the FNMR concentration is described by the label "goat". Further, note the users specific means and variance which will be associated with variables such as eye-openness, eye-lash prominence, pupil dilation, and human-camera interaction factors.

Bimodal score distributions: Further some subjects (04470, 05432 for algorithm I02P and 04261, 05379L for algorithm D03P in the ND08-10 set) exhibit a bimodal dissimilarity score distribution. The first, lower, distribution corresponds to correct matches, the second higher distribution is centered where the impostor distribution would appear. The cause of this could be either gross segmentation failure due to non-ideal images, or erroneous ground truth (e.g. mis-labeled left and right eyes). A key contention of this (NIST) paper is that progressive ageing of the iris texture would manifest itself in continuously increasing scores, not catastrophic failure. This is consistent with prior studies that show aberrant images cause outright failure[73, 92]. We recommend such images not be included in ageing analyses.

Ageing candidates: Many individual eyes show evidence of elevated scores after one and two year time lapses. Notable examples are 5555R, 5517L, 5455L, 5456R, 5310R and 5379R, and these occur even without any false rejections at the nominal threshold. This score dependence on time without clear false non-match signifies ageing of unknown cause. On the other hand, 5432L+R both show increases in scores that *do* culminate in false non-matches.

The first column of Figure 20 confirms the published ND08-10 result[29] that the distribution of genuine scores undergoes an adverse shift with time between captures. It plots $\text{FNMR}(\tau)$ against threshold, τ , for eight algorithms comparing short, intermediate, and long separation images. These curves have the same form as those reported by Fenker et al. (Figures 2-7 in [29]) for a Neurotechnology algorithm, and Baker et al. (Figure 2 in [8]) for a Cambridge University algorithm. Two exceptions must be noted. The first, for algorithm F03N shows a *reversal* of the ageing effect. The F02P algorithm includes score normalization (across candidates), F03N does not. This algorithm exhibits the ageing effect only in the right tail of the distribution.

Three additional algorithms, A01N, D03P, K03N, show that the one-year-separated distribution (M) is nearly indistinguishable from the two-year(L), implying that ageing is limited only to the first year between captures. This non-physical result prompts the following re-plotting of the data. The second column Figure 20 plots the same data for intra- and inter-semester sets of comparisons, 2008-2008, 2009-2009, 2010-2010, 2008-2009, 2009-2010, and 2008-2010. The plots reveal that the intra-semester genuine distributions differ as would occur from an ageing effect *during* a semester, or due to statistical fluctuation. Further the 2008-2009 distribution is much closer to that of the intra-semester comparisons than is the 2009-2010 set. This would be consistent with a material change to the collection (camera, population, procedure) between the 2009 and 2010 collections. Possible causes of the observed ageing effects follow in the next section.

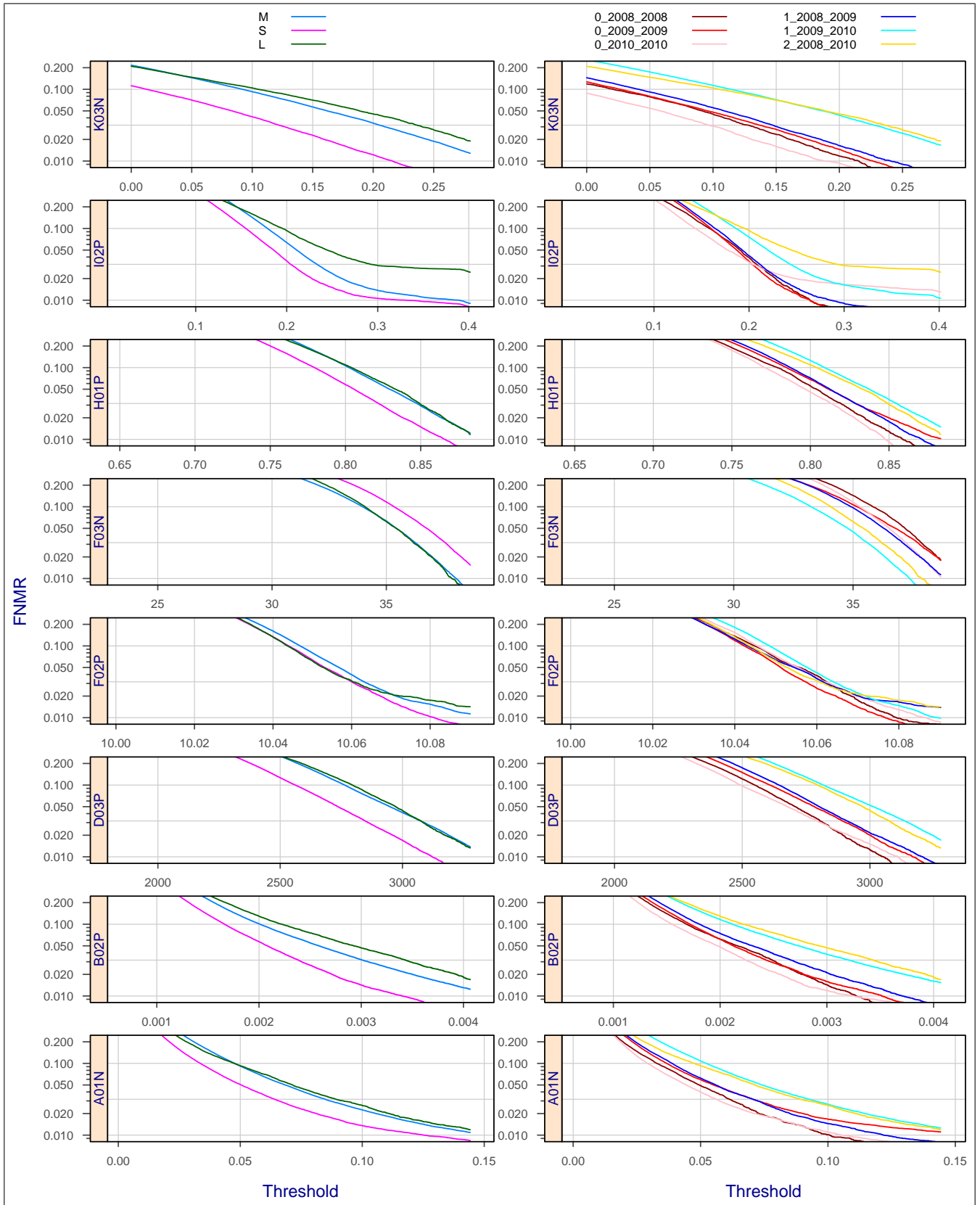


Figure 20: Score distributions by time-difference and by time: For dataset ND08-10, the eight rows plot $\text{FNMR}(\tau)$ for eight IREX IV algorithms. On the left side, each panel shows three traces for S, short (fewer than 120 days), , L long (greater than 600 days) and M, intermediate time lapses between comparisons. Short term corresponds to within semester (2008,2009,2010), intermediate to 2008-2009 and 2009-2010 comparisons, and long term refers to 2008-2010. The right side panels make this explicit, plotting FNMR for intra- and inter-year comparisons. Broadly, the left side repeats the published ND result ($\text{FNMR}_S < \text{FNMR}_M < \text{FNMR}_L$). The right side reveals that one-year comparisons of 2009 with 2010 images give false non-matches at about the same rate as over the two years 2008 to 2010.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

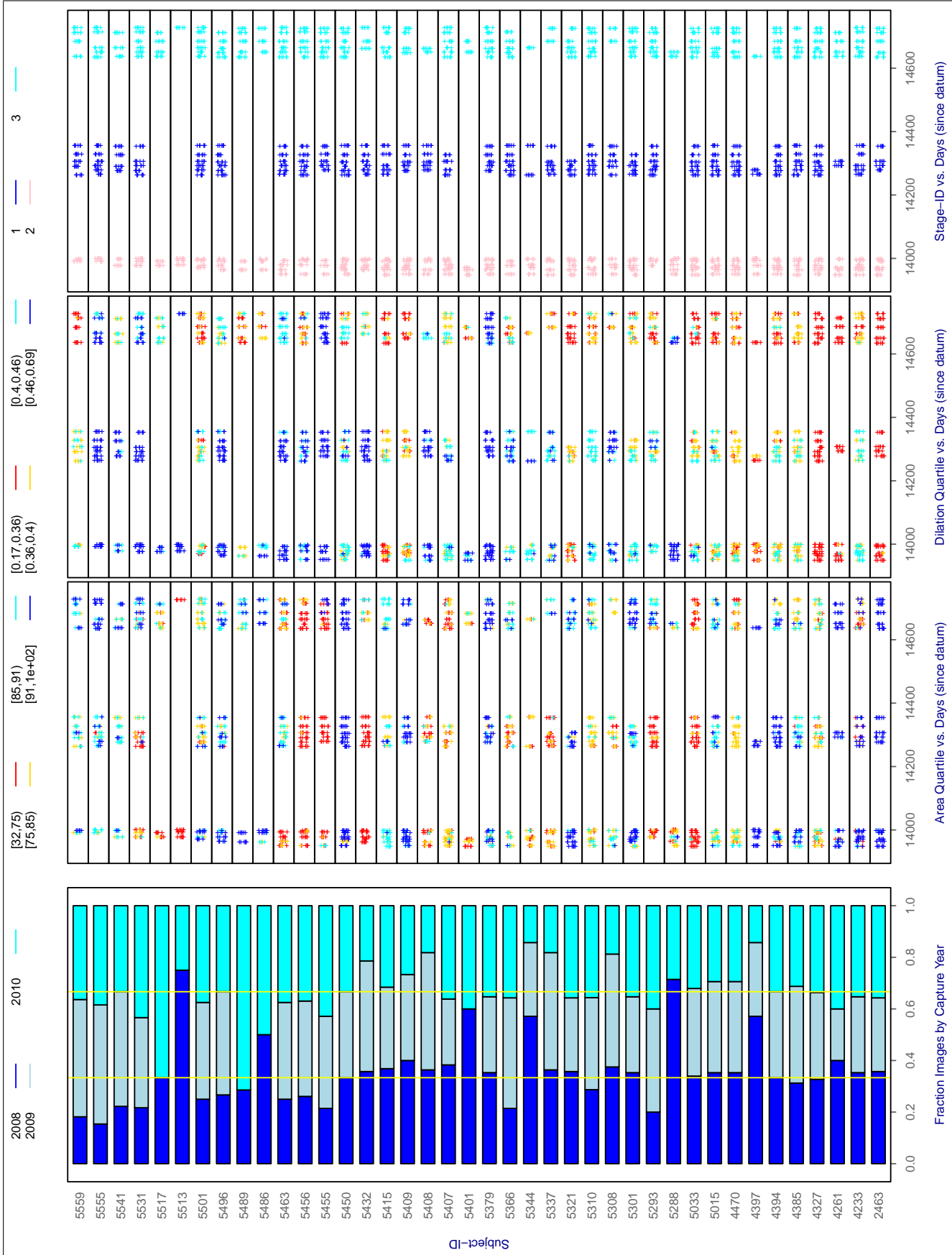


Figure 21: **Dilation and area by year:** From left to right the four panels give a) the fraction of a subject's images collected in each year 2008-2010, b) quartiles for each image's exposed iris area quality estimate, c) quartiles for pupil dilation, and d) the STAGE-ID indicator. In panels 2-4, the x-axis is absolute time since 1/1/1970 corresponding to Spring 2008, 2009, and 2010, and y-coordinates are jittered slightly to separate the points. Increased amounts of red in panel 3 shows that pupil constriction is more common in 2010, an aspect shown more exhaustively in Figure 22. The un-documented STAGE-ID value is different in 2008, 2009, and 2010 and is thus a proxy for time.

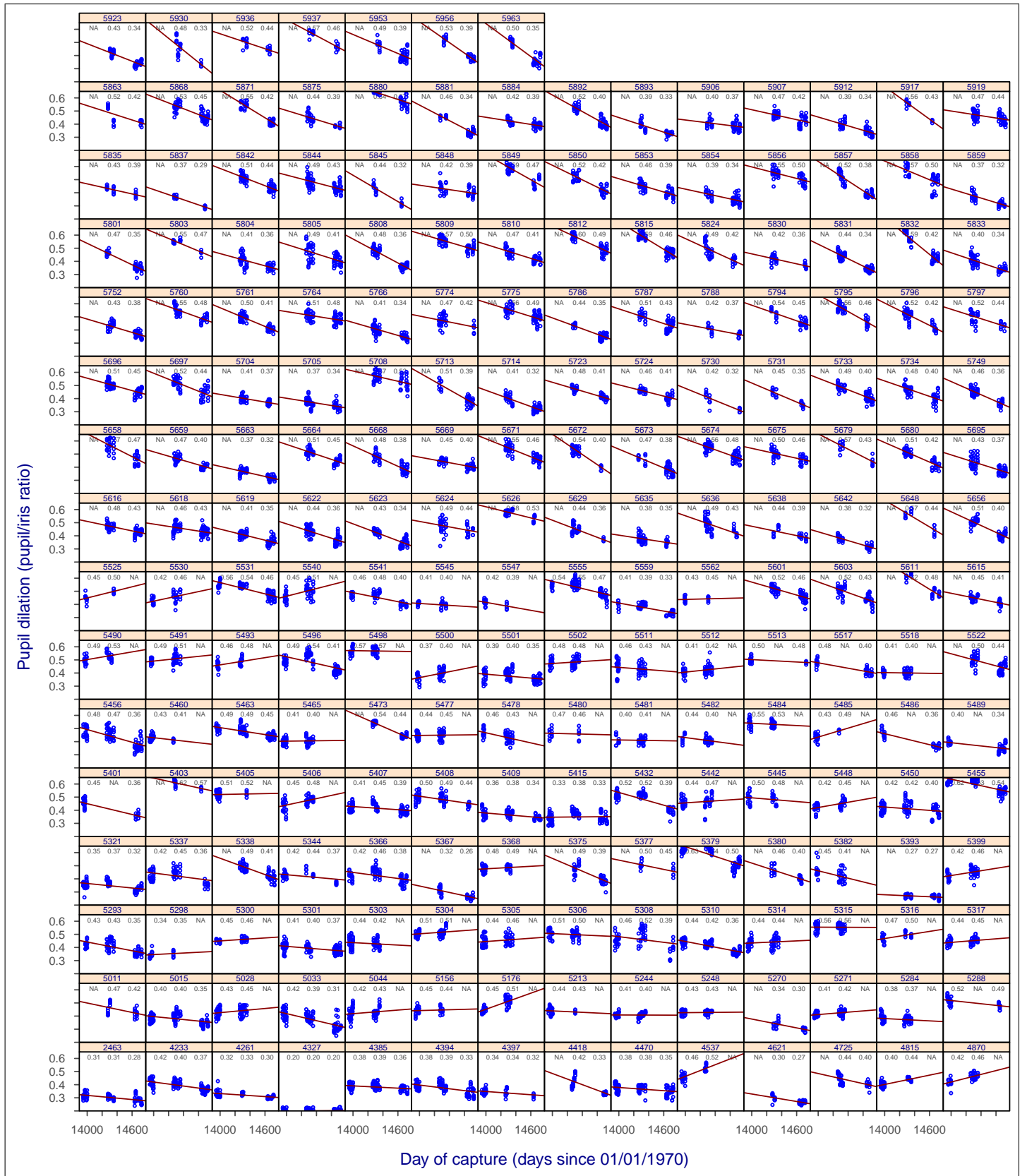


Figure 22: Dilation over time: (This figure is best viewed on a computer screen with software magnification). For all 217 individuals in dataset ND08-10, the panels show absolute pupil dilation for all images of each person against date of capture. The clusters correspond to collections in the Spring of 2008, 2009 and 2010. Some individuals only appear in two of those years. The dilations are consensus estimates per equation 2. The straight red line is the result of regressing $D = D(T)$. The text in each box gives the median dilation over all images collected from an individual in that calendar year. Note both inter- and intra-personal variation. Population medians are highest in 2009 and lowest in 2010. The reasons for these systematic changes in dilation are unknown.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

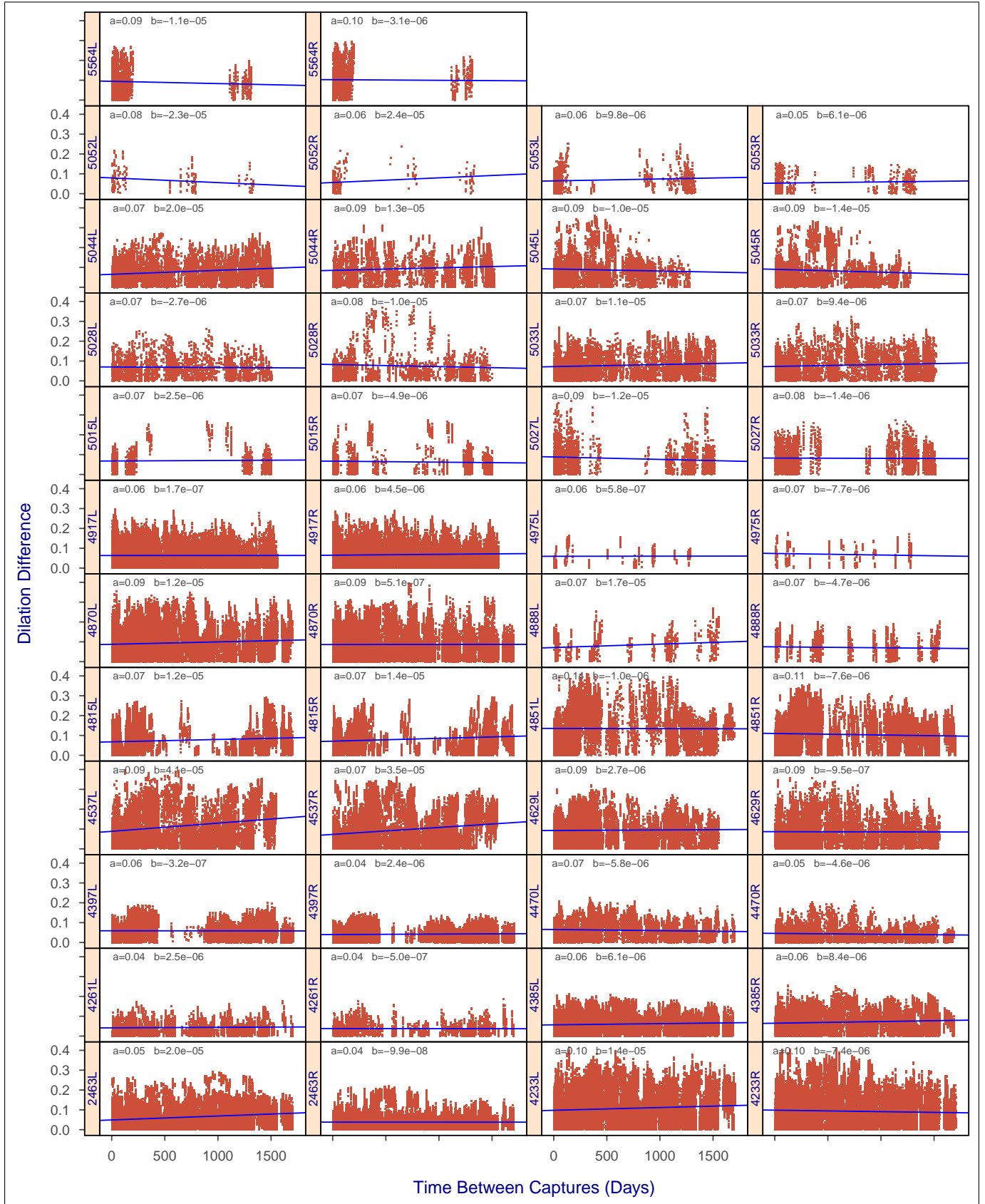


Figure 23: **Dilation differences over time:** For dataset ND04-08, each panel shows dilation differences for all pairs of images of one eye vs. time between two samples. The dilations are consensus estimates per equation 2, and the dilation difference is the radial iris texture thickness ratio of equation 4. The straight blue line is the result of regressing $\Delta D = \Delta D(T)$, the intercept and gradient of which appear as text.

4.3.1 Causes of longitudinal change in ND08-10

For the 40 subjects in the ND08-10 population that participated in both 2008 and 2010, Figures 22 and 23 plot time-series of the eq. (2) dilation difference estimates. These data suggest an explanation of the trends observed in the dissimilarity score time-series: namely that changes in dilation cause changes in scores. The third panel of Figure 21 depicts dilation as a function of date, and this reveals that some individuals have consistently different dilations across the three year collection. For example, some pupils of individual 05033 constrict from $D > 0.46$ to $D < 0.36$. Moreover the pupils of most persons in 2010 are less dilated than those in 2008. This is evident in Figure 24 which shows all the images of the left eye of individual 05455; note the irises are all well centered with specular reflections located consistently inside the pupil. However, high dilation in 2008 and 2009 reduces markedly in 2010 (yet still remains in the top quartile in Figure 21).

Another covariate affecting recognition is the STAGE-ID taken from the ND metadata provided with the images. This value is not documented specifically but is used in a larger metadata framework for documenting biometric data collections to denote location or setting. However, we do not further consider this annotation because, as revealed in Figure 21, it is a perfect proxy for the year of capture.

Finally we use quantitative statements of exposed iris area. These are values produced by the C4X iris image quality assessment algorithm²⁸ submitted to IREX II[87]. This algorithm, from University of Cambridge, takes a single iris image and renders a quality score on $[0, 100]$ with bigger-is-better semantics. While its internal definition is proprietary, IREX II and Figure 25 show this to have high correlation with observed dissimilarity scores and false rejection errors. We consider it to be monotonically related to actual exposed iris area. Given values, A_1 and A_2 , for a mated pair of images we form a summary measure,

$$\Delta A = \frac{\min(A_1, A_2)}{100} \quad (9)$$

as the minimum of the two values scaled to $[0, 1]$. This is done on the basis that successful genuine comparison will be subverted by low exposed area in either of the two samples. This measure does not capture the *overlap* area common to both images. Referring to the images of individual 05456 in Figure 26, the irises are all well centered with specular reflections located consistently inside the pupil. However, area differences are evident for example in images 1001, 997, 971, 900, and 60. Note that while Baker[8] excluded poor images from the 2004-2008 set by “manually screening for image quality ... e.g. out-of-focus irises, major portions of the iris occluded ...”, Fenker did not consider this necessary in 2008-2010 given the LG4000 superiority. It is clear from the image-specific FNMR values above each image that errors and non-idealities²⁹ do occur.

We therefore consider three quantitative adjustments to the scores to remove the effects of exposed iris area, dilation difference, and both. Specifically, given an observed dissimilarity score, d_{ij} , from the j -th comparison of the i -th individual,

²⁸Quality algorithms map an input image into either a summary scalar quality value, or into a vector of image quality measures. Operationally the summary measure is useful because it is interpreted as measure of suitability of the image for recognition by a specific or generic unspecified recognition algorithm. If quality is low, a recapture is initiated. A quality vector is comprised of specified image properties each of which is taken as a measure of recognition accuracy. The elements of an iris quality vector are being formally standardized as ISO/IEC 29794-6 along with prescriptions for their computation. Examples of such elements are axial gaze angle, exposed iris area, and focus. The precision and efficacy of quality assessment algorithms has been assessed in IREX II[87]. Quality measures are of interest here because they operate on single images and therefore do not make a statement of dissimilarity between time-separated irises. Their application can reveal longitudinal changes in salient image properties. Commercial quality assessment algorithms exist for face, fingerprint and iris images.

²⁹04233L: This eye appears to be wide open in 2008 and 2009, but with the upper eyelid approaching the pupil in 2010. 04261L: This individual has contact lenses, and an overall 16.6% FNMR. Image 996 (2008) exhibits ptosis. Images 1030 (2008) and 1601 (2010) are rotated. 04385L: Prominent eye lashes, but low FNMR everywhere except image 2245 (2010) which has mild motion blur. 04397L: One image 2059 (2010) has off-axial gaze. 04870L: Variable, though not extreme, dilation throughout. 05028L: Has eye lash occlusion throughout, some off-axis, but low error rates. 05033L: Has lower pupil dilation in 2010 than in 2008 and 2009. High error rates due to eyelash occlusion. 05044L: Exhibits variable dilation and eye openness. Eyelid motion in image 1259. 05293L: Eyes more open in 2010. Reduced dilation in 2010. 05300L: Image 393 (2009) is far off axis with iFNMR is 0.655. 05310L: Lower dilation in 2010 than in prior years. But low FNMR. Amount of light reflected from skin below eye is less in 2010. 05321L: Amount of light reflected from skin below eye is less in 2010. 05337L: Reduced dilation in 2010 05366L: Reduced dilation in 2010 05379L: Very high dilation in 2008 and 2009. Reduced in 2010.

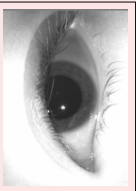
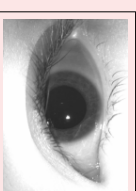
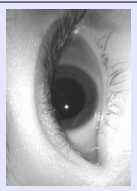
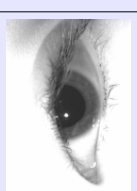
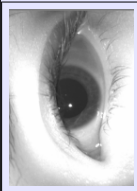
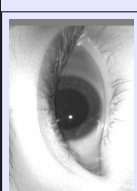
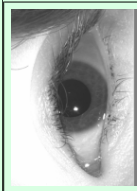
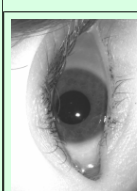
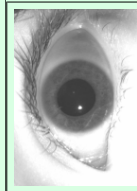
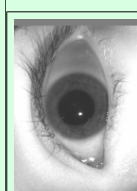


05455L : Global FNMR 0.173									
K03N 0.187	A01N 0.291	F03N 0.000	I02P 0.231	F02P 0.165	H01P 0.160	D03P 0.188	B02P 0.165		
52 0.106	54 0.096	56 0.160	57 0.096	59 0.048	61 0.054	75 0.106	77 0.247		
79 0.311	457 0.071	459 0.071	461 0.042	487 0.199	489 0.205	491 0.545	517 0.186		
519 0.397	521 0.179	570 0.170	572 0.035	574 0.122	612 0.163	614 0.163	616 0.167		
835 0.324	837 0.131	839 0.157	865 0.071	867 0.122	869 0.138	893 0.407	895 0.160		
897 0.109	923 0.391	925 0.090	927 0.436	953 0.096	955 0.042	957 0.173	983 0.141		
985 0.167	987 0.183								

Figure 24: **One eye over time:** Images of the left eye of person 05455 in the ND08-10 dataset. The pink shading indicates 2008, blue 2009 and green 2010. The images are ordered in time according to the sequence number which appears above each image. Three classes of false non-match rate FNMR are given, all computed when the threshold is set to give FNMR = 0.02 over the entire dataset. The first row states FNMR for images of this eye (i.e. mean FNMR over eight recognition algorithms). The second row gives eye-specific FNMR by algorithm. Above each image is image-specific error rate (iFNMR - the proportion of genuine comparisons involving an image that result in a false non-match[88]).

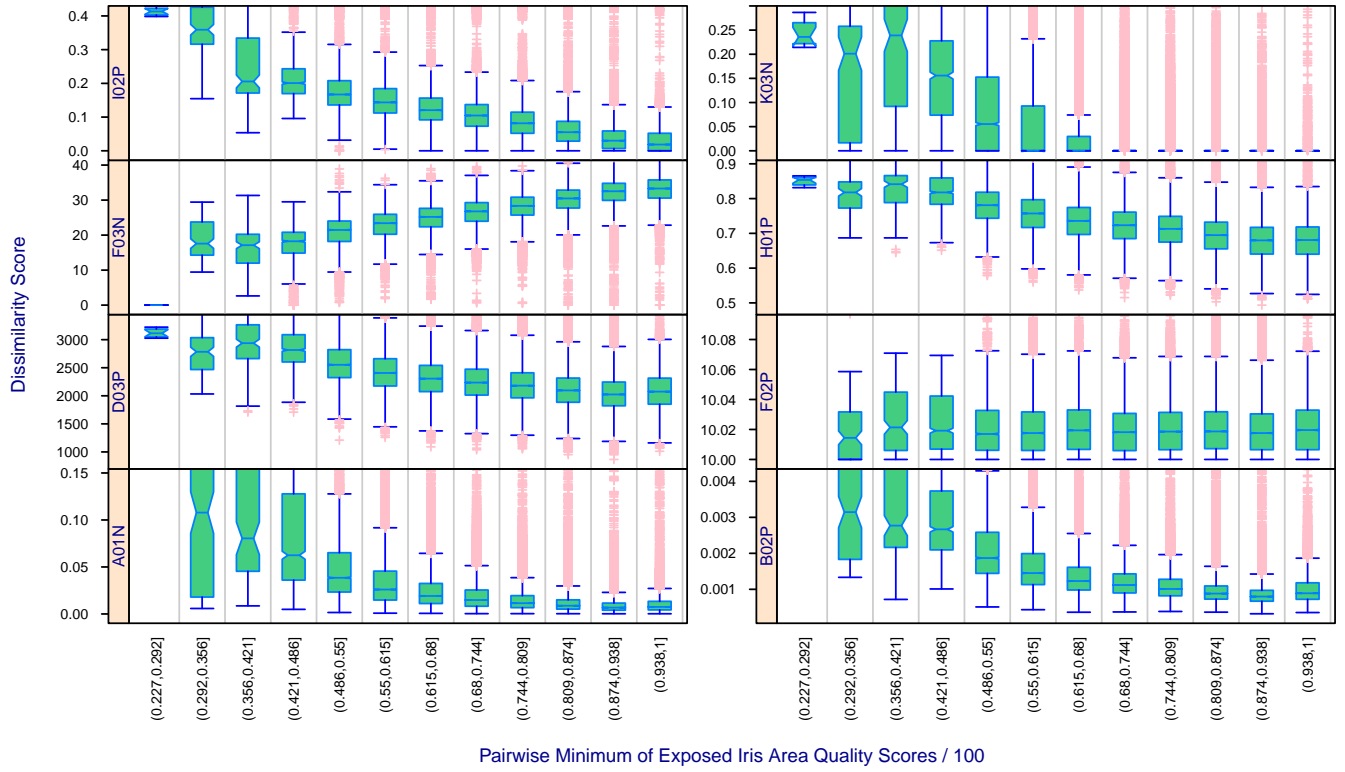


Figure 25: **Influence of exposed iris area:** For the ND08-10 dataset, the dependence of algorithms' comparison scores on iris area. Specifically the x-axis gives binned values of the lesser of Daugman's exposed iris area quality scores computed from the two images using the C4x implementation submitted to IREX II IQCE[87]. These quality scores are proprietary but are assumed to be a monotonic function of the exposed iris texture area as a fraction of its maximum. Another variable, area of overlap of the two irides, is not available.

the adjustments subtract fixed proportions of the dilation change

$$d'_{ij} = d_{ij} - \beta_{i2}\Delta D_{ij} \quad (10)$$

and the exposed iris area minimum ΔA ,

$$d'_{ij} = d_{ij} - \beta_{i3}\Delta A_{ij} \quad (11)$$

and their joint effect

$$d'_{ij} = d_{ij} - \beta_{i2}\Delta D_{ij} - \beta_{i3}\Delta A_{ij} - \beta_{i4}\Delta D_{ij}\Delta A_{ij} \quad (12)$$

where the coefficients β could come from many sources (e.g. first principles, anatomic models, exhaustive search), or in this case from ordinary least squares (OLS) regression using the models

$$d_{ij} = \beta_{i1} + \beta_{i2}\Delta D_{ij} + e_{ij} \quad (13)$$

$$d_{ij} = \beta_{i1} + \beta_{i2}\Delta A_{ij} + e_{ij} \quad (14)$$

$$d_{ij} = \beta_{i1} + \beta_{i2}\Delta D_{ij} + \beta_{i3}\Delta A_{ij} + \beta_{i4}\Delta D_{ij}\Delta A_{ij} + e_{ij} \quad (15)$$

where e_{ij} is an error term representing our failure to measure all causal reasons for comparison score variation. Here OLS is applied to each eye individually such that the β_i are specific to one eye. Note the adjusted d' values are not predictions from regression models because they include the residuals. Rather, the approach is equivalent to using the constructed

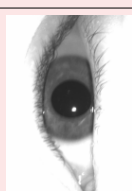
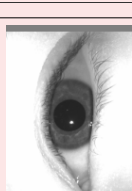
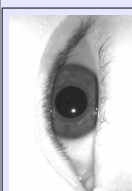
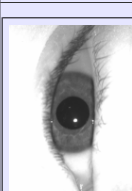
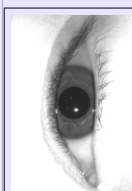
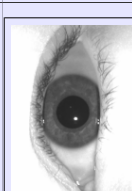
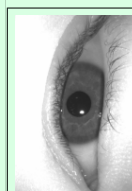
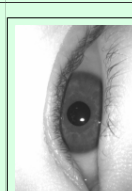


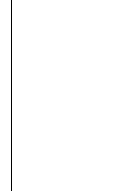
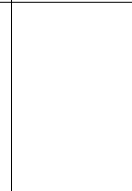
05456L : Global FNMR 0.088									
K03N 0.090	A01N 0.064	F03N 0.000	I02P 0.062	F02P 0.000	H01P 0.121	D03P 0.127	B02P 0.239		
101 0.028	103 0.028	60 0.049	62 0.038	64 0.047	65 0.035	67 0.032	69 0.355		
81 0.050	83 0.016	85 0.044	99 0.034	497 0.195	499 0.262	501 0.140	504 0.029		
506 0.047	508 0.026	557 0.031	559 0.048	562 0.038	564 0.020	566 0.032	592 0.142		
594 0.049	596 0.047	645 0.029	647 0.032	649 0.023	1001 0.070	843 0.154	845 0.108		
847 0.064	870 0.071	872 0.074	898 0.067	900 0.198	902 0.084	937 0.070	939 0.259		
941 0.113	967 0.049	969 0.073	971 0.541	997 0.035	999 0.052				

Figure 26: **One eye over time:** Images of the left eye of person 05456 in the ND08-10 dataset. The pink shading indicates 2008, blue 2009 and green 2010. The images are ordered in time according to the sequence number which appears above each image. Three classes of false non-match rate FNMR are given, all computed when the threshold is set to give FNMR = 0.02 over the entire dataset. The first row indicates FNMR for images of this eye over all eight recognition algorithms. The second row gives eye-specific FNMR by algorithm. Above each image is image-specific error rate (iFNMR) - the proportion of genuine comparisons involving an image that result in a false non-match[88].

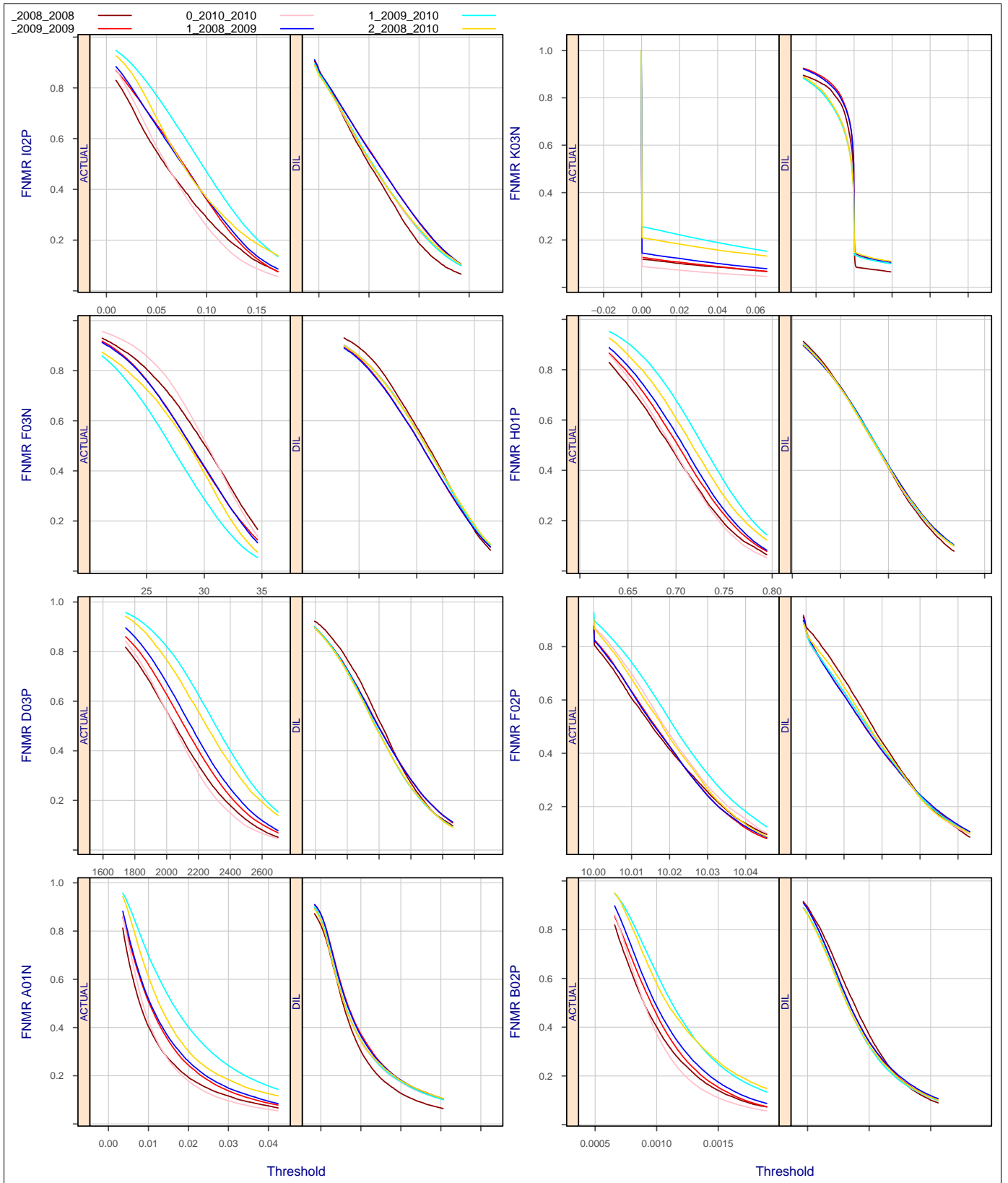


Figure 27: **Score distributions with and without dilation effects:** For dataset ND08-10, the eight panel pairs plot FNMR corresponding to the entire the genuine distribution, $0 \leq \text{FNMR}(\tau) \leq 1$ for eight IREX IV algorithms applied to the images of all eyes. The pairs show fractions of raw dissimilarities and the dilation-adjusted values (per eq. 10) that are above threshold. In each panel the six traces correspond to within semester (2008,2009,2010) (in pink-red), one year separated 2008-2009 + 2009-2010 (in blues), and two year separated 2008-2010 (in yellow) comparisons. Clustering of lines indicates that the genuine distribution is stable over 0, 1 and 2 year intervals, consistent with a no-ageing effect. Column 2 shows dilation to be the major source of time variation. The K03N distribution is unusual because the algorithm emits many 0 scores.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

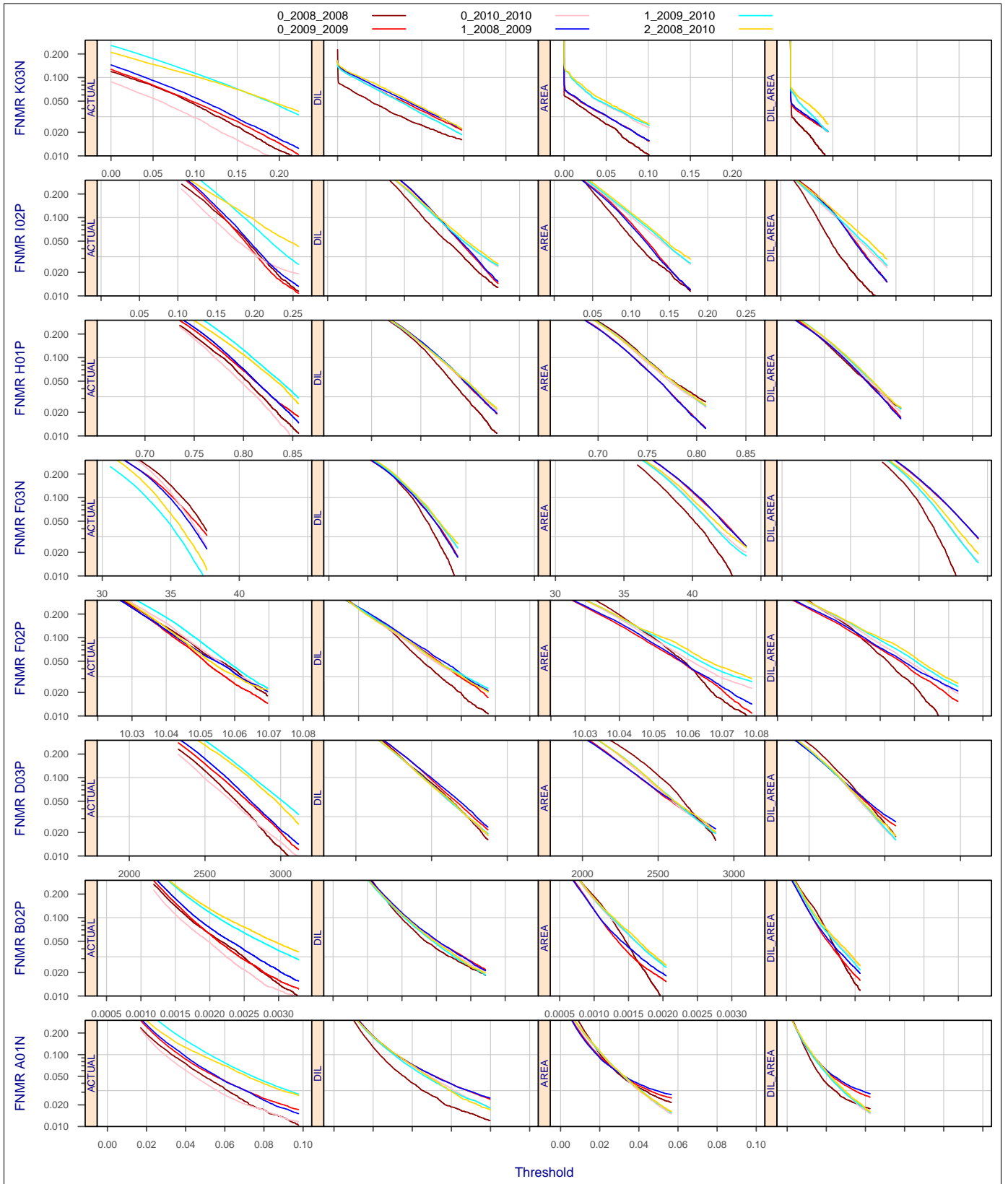


Figure 28: **Score distributions with and without dilation and area effects:** For dataset ND08-10, the rows plot $\text{FNMR}(\tau)$, corresponding to the right tail (top 20%) of the genuine distribution. for eight IREX IV algorithms applied to images of all eyes. The four columns show: the raw dissimilarities from the algorithms; the dilation-adjusted values per eq. 10; the iris area adjusted values per eq. 11; and the dilation-and-area adjusted values per eq. 12. In each panel the six traces correspond to within semester (2008,2009,2010) (in pink-red), one year separated 2008-2009 + 2009-2010 (in blues), and two year separated 2008-2010 (in yellow) comparisons. Clustering of lines indicates that the genuine distribution is stable over 0, 1 and 2 year intervals, consistent with a no-ageing effect. Column 2 shows dilation to be the major source of time variation.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

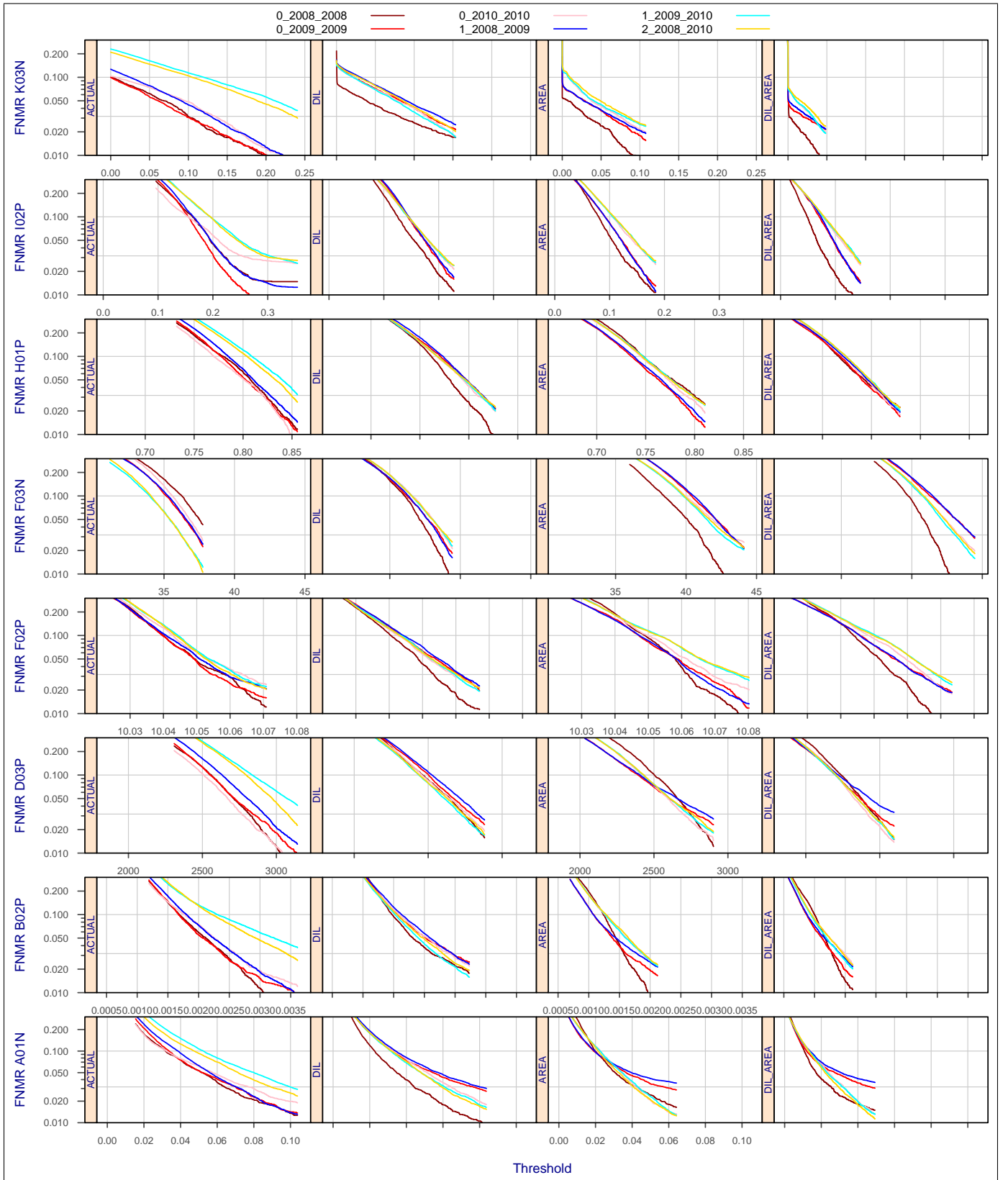


Figure 29: **Score distributions with and without dilation and area effects:** As Fig. 28 but with just those 80 eyes imaged in 2008 and 2010. For dataset ND08-10, the eight rows plot $\text{FNMR}(\tau)$ for eight IREX IV algorithms, corresponding to the right tail (top 20%) of the genuine distribution. The four columns show: the raw dissimilarities from the algorithms; the dilation-adjusted values per eq. 10; the iris area adjusted values per eq. 11; and the dilation-and-area adjusted values per eq. 12. In each panel the six traces correspond to within semester (2008,2009,2010) (in red), one year separated 2008-2009 + 2009-2010 (in blue), and two year separated 2008-2010 (in yellow) comparisons. Clustering of lines indicates that the genuine distribution is stable over 0, 1 and 2 year intervals, consistent with a no-ageing effect. Column 2 shows dilation to be the major source of time variation.

FNMR = False non-match rate
ND = Uni. of Notre Dame

A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last

OLS models with dilation differences (eq. 4) to be 0, all “stages” to be the same, and area minima (eq. 9) set to 1, *and then adding back in the residuals*. Prior applications of regression to iris-ageing[84] computed a model for the whole population not for individuals - this did not allow eye-specific intercepts for example.

Importantly, note that time between image captures is *not* used in computing the adjusted d' values. Thus the method simply subtracts functions of image-pair area and dilation from each eye sequence. These quantities vary stochastically through time as described in section 3.

The results of applying these corrections are shown in Figure 27. The curves correspond to the center of the genuine distribution by plotting $\text{FNMR}(\tau)$ against threshold, τ , for eight algorithms. These curves have the same form as in Fenker et al. (Figures 2-7 in [29] with $0 \leq \text{FNMR}(\tau) \leq 0.7$) and Baker et al. (Figure 2[8] with $0 \leq \text{FNMR}(\tau) \leq 0.07$). The first column, for the raw observed dissimilarity scores, suggests the same conclusion that dissimilarity does indeed increase with elapsed time. However, in subsequent columns, with the adjustments of equations (10) - (12), this time dependence is substantially reduced.

Two additional visualizations of this are plotted. First, Figure 28 corresponds to just the right tail of the genuine distribution function, with $0.01 \leq \text{FNMR}(\tau) \leq 0.2$. Figure 29 shows the result just for the 40 individuals imaged in 2008 *and* 2010. Second, all adjusted scores are plotted exhaustively, as time-series, for algorithms I02P and D03P in Figures 30 and 31 with other algorithms' plots in the accompanying Appendix. The notable observations are:

The figures show that many individual score trajectories have lower gradient after adjustment for dilation and area. For example, in the top row Figure 30, the first four panels show that the left and right eyes of individual 5455 give lower gradients. This holds in most cases. One exception is in the 12th row, for eye 5344R, where the plotted line has increased gradient.

The scores overall tend to decrease because the transformations of eqs. 13-15 do not conserve means. Indeed, some individual values become negative. This is immaterial because only trend is of concern.

Some scores increase. This occurs because the OLS coefficients can be negative, usually because ΔA is counter-intuitively positively correlated with dissimilarity score. This arises when the eq. (9) area metric takes on a high value but the area of the overlapping texture in a pair of images is small. This would occur when upper and lower eyelids occlude different parts of the two irises.

The conclusion is that dilation, and also area changes, account for many of the observed increases in dissimilarity. This begs two questions: Why does dilation undergo a step change in the ND data? And why did the two published ND treatments of dilation difference not yield the same conclusion? The first question may have a random component, and is largely un-answerable absent more detailed knowledge of causal factors particularly environmental illumination, medication[41] and fatigue levels. The second question is answered separately for Baker et al.[8] and Fenker et al.[28] as follows.

ND04-08 Baker et al. quantified dilation differences as the difference in the mean of dilation differences between the sets of images used in long term and short term comparisons. The averaging of dilation values hides pairs of images with markedly different dilations, because the mean of dilation differences is the difference of the mean dilations. However it is pairs of images for which dilations are different that produce poor dissimilarity scores.

Their analysis proceeds by computing IrisBEE Hamming distances for long-separation image pairs, and their mean, μ_L . They likewise compute the mean HD for short separation pairs, μ_S . They conclude that “changes in pupil dilation are not an appreciable factor” in the “observed [ageing] result” on the basis that the Kendall tau correlation

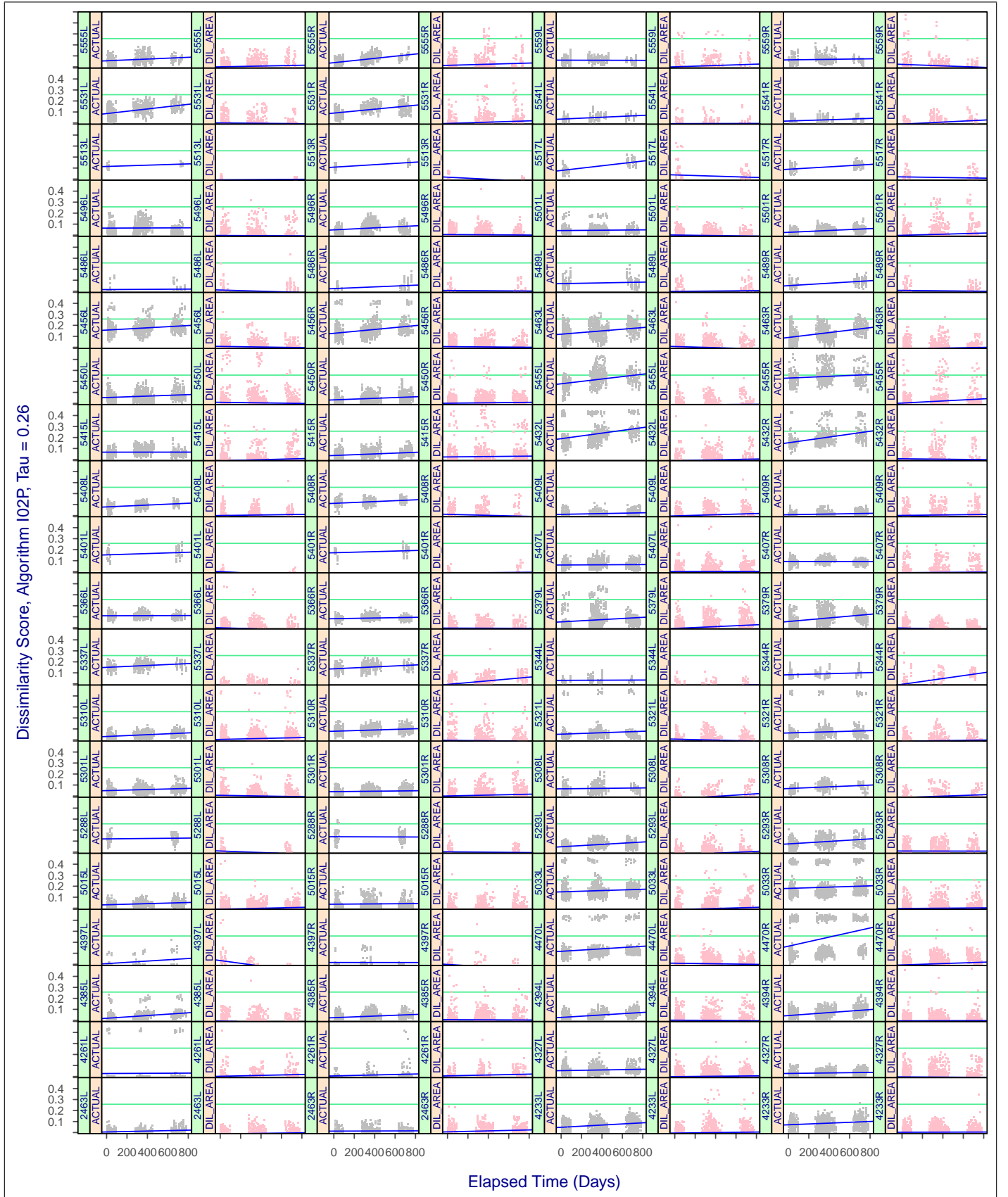


Figure 30: **Score trajectories pre and post dilation and area adjustment:** For dataset ND08-10, the plots show individuals' dissimilarity scores for the I02P algorithm vs. time between two samples. Each dot corresponds to a pair of images, grey indicates raw scores, pink indicates dilation-and-area adjusted scores per eq. (12). The green horizontal line shows the threshold that gives FNMR = 0.02 globally. The straight blue line is the result of regressing $d = d(T)$. A reduced slope in adjacent panels indicates that dilation and area variations explain any ageing effect in the left panel.

FNMR = False non-match rate
ND = Uni. of Notre Dame

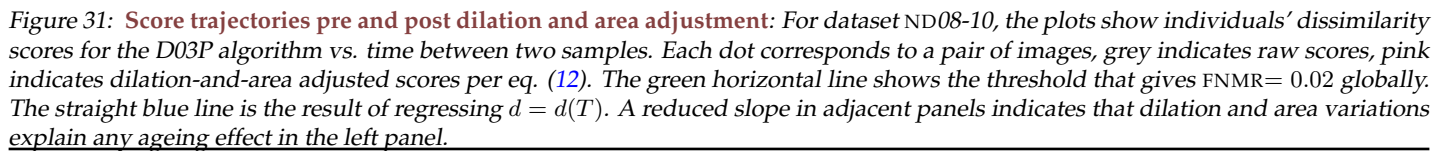
A = Uni. Bath
B = Neurotech.

D = 3M Cogent
F = MorphoTrust

H = Delta ID
I = Uni. Cambridge

K = Morpho

T_{\min} = Time enroll to first
 T_{\max} = Time enroll to last
 T_A = Active, first to last



coefficient of $(\rho_L - \rho_S)$ with $(\mu_L - \mu_S)$, computed over the 46 individual eyes, is 0.217 (see [8], Figure 3). Whether 0.2 is small is not clear, but the method is clearly undermined by the use of averaging - the penultimate row of Table 2 indicates that an average person has a mean $|D_1 - D_2|$ of 0.06, corresponding to a change in radial thickness above 0.1 and an increase in HD of at least 0.03 (via Table 2 and Figure 7 in this paper). This would be higher for small or high values of D .

Our analysis of dilation in the ND04-08 set follows the score-adjustment approach of equation (10). The result, in Figure 32, shows that the increase in FNMR over short time periods (i.e. the blue line) is reduced once dilation and iris area are factored in. However, the adjustment is not as effective at stabilizing the genuine distribution for ND 2004-2008 as it was for 2008-2010 suggesting other variations remain. But note also that this dataset was produced from its parent corpus by removing 90% of the poorest quality images.

ND08-10 Fenker's approach to dilation was to retain comparisons in the FNMR estimate if the dilation difference $\Delta D = |D_2 - D_1|$ did not exceed 0.1. However, this rule for removal of image-pair comparisons does not exclude any individual from the analysis, and removes only a few images: For the 40 individuals appearing in both 2008 and 2010, it removes 1 of 1035 images involved in 2008-2008 comparisons, 0 of 1033 in 2009-2009, 1 of 1044 in 2010-2010, 6 of 1966 in 2008-2009, 93 of 1993 in 2009-2010, and 65 of 2097 in 2008-2010. Specifically, the number of images involved in short-term comparisons reduces from 3112 to 3110, medium term from 2920 to 2879, and 2097-2032 long term.

Similarly, for all 217 the individuals appearing in 2008, 2009 or 2010, the rule removes 3 of 2300 images involved in 2008-2008 comparisons, 10 of 5396 in 2009-2009, 1 of 3926 in 2010-2010, 1 of 4553 in 2008-2009, 575 of 8046 in 2009-2010, and 65 of 2097 in 2008-2010. Specifically, the number of images involved in short-term comparisons reduces from 3112 to 3110, medium term from 2920 to 2879, and 2097-2032 long term. If ΔD had been capped to lower values a larger effect would have been observed.

4.3.2 Discussion of the ND result

The ND team's collection protocol (off the shelf cameras, regular scheduled capture sessions, over several years) was intended to support a number of iris recognition investigations, among them identification of longitudinal changes in iris recognition. The ND08-10 collection protocol benefitted from the camera's internal quality control mechanisms and from human review. This review discarded, for example, egregiously blurred images. However, as evidenced by measureable false non-matches (Fig. 13), elevated scores (Fig. 19), and changes in dilation and eyelid occlusion (Figs. 21 and 24), the collection did not attain the level of control needed to separate any iris texture changes from all others.

The ND analyses attempted to address this by retroactively accounting for some variations. However, as described previously the two treatments of dilation were insufficient to suppress its influence. The end-result is that the ND studies do not specifically claim an iris-texture ageing effect changes in the core emitting feature - they suggest only that the observed longitudinal increase in recognition error rates and recognition dissimilarity scores is due to iris texture changes, or to some unknown factor.

4.4 Results for OPS-FIELD

Despite its large population size (622464 subjects), the OPS-FIELD database is of limited utility for detecting ageing effects. It was collected over several years but without any notion that it would be used to estimate iris ageing effects. The database is actually comprised of two partitions, roughly equal in size. The first is the set of images collected during enrollment of

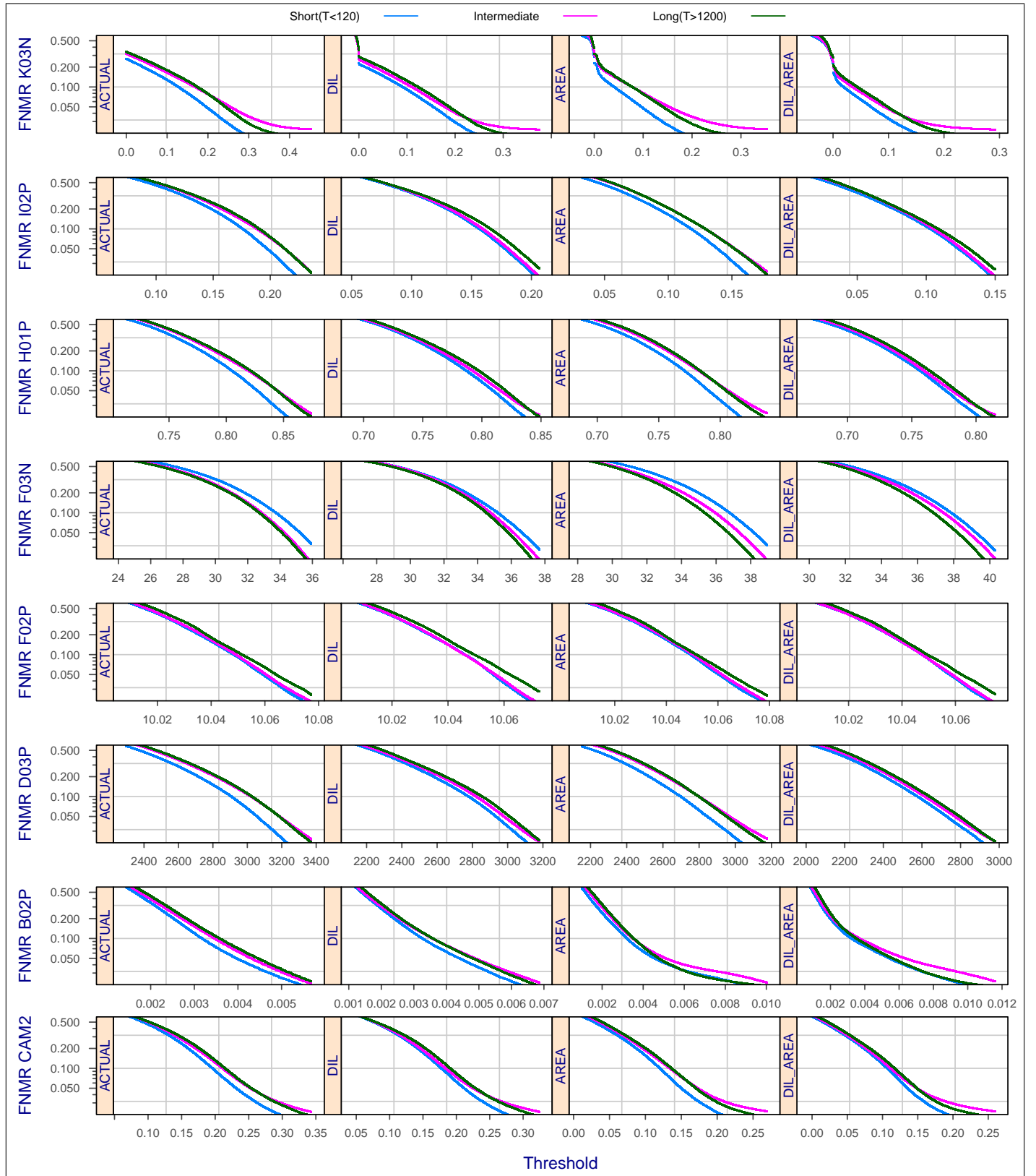


Figure 32: **Score distributions with and without dilation and area effects:** For dataset ND04-08, the eight rows plot $\text{FNMR}(\tau)$ for eight IREX IV algorithms applied to the images. In each panel the three traces correspond to three sets of elapsed time bins, short is fewer than 120 days, long is greater than 1200 days, and intermediate is for anything in between. Thus short term corresponds to within semester (2004,2005,2006,2007,2008), long term refers to 2004/05 - 2007/08 and intermediate corresponds to everything in between. The four panels show from left: the raw dissimilarities from the algorithms; the dilation-adjusted values per eq. 10; the iris area adjusted values per eq. 11; and the dilation-and-area adjusted values per eq. 12. Clustering of lines indicates that the genuine distribution is stable over 0, 1 and 2 year intervals, consistent with a no-ageing effect.

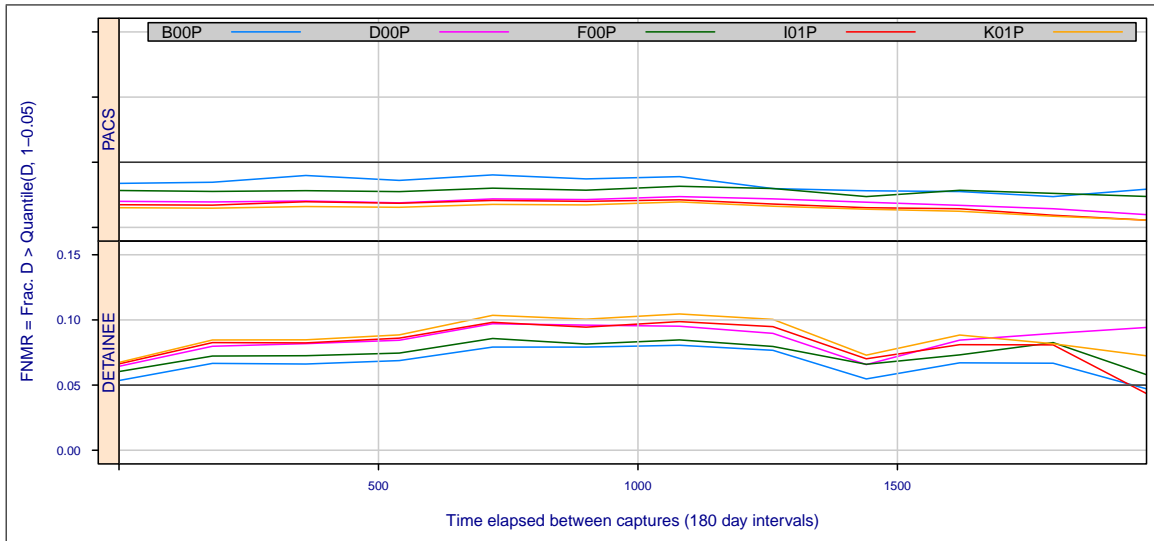


Figure 33: **Stability of error rate:** For dataset OPS-FIELD, the fraction of mates not identified at better than a fixed threshold vs. binned time between captures. Each trace corresponds to an algorithm submitted to NIST’s IREX IV evaluation. The threshold for each is set to achieve FNIR= 0.05 over all mated searches using the entire OPS-FIELD dataset. The upper panel refers to the cooperative enrollments for physical access control (PACS); the lower to variably non-cooperative detainees. The figures include results for images acquired with cameras from one manufacturer, but with some variety of models. These amount to about 80% of the corpus. Temporal changes here will depend on un-reported systematic changes to camera design, to standard operating procedure and collection environment and geography, and possibly to ethnic composition and demography.

subjects for physical access control PACS. The second consists of images collected from subjects detained during military operations DETAINEE. The two partitions will clearly differ in the cooperativeness of the subjects. Less obviously they differ as a consequence of the capture environment. Particularly the detainee set includes many images collected in non-ideal circumstances sometimes in improvised structures in close proximity to strong reflected sunlight[73]. Additionally, the data does not include large numbers of individuals with multiple encounters.

Note, however, that the database does include many excellent images: In IREX III identification trials with enrolled populations of 3.9 million subjects, false negative identification rates (“miss rates”) are below 1.5% at thresholds set so that false matches occur at rates below 10^{-13} [34].

Given the above discussion and the lack of control on any given capture, the only results we report are population aggregates. Figure 33 shows that none of the algorithms exhibit an increase in false negative error rates over the 1900 day (5 year) interval. The difference between the PACS and DETAINEE partitions manifest themselves in separation of false negative identification rates, as depicted in Figure 33: The more controlled cooperative-subject PACS images produce about half as many errors as the DETAINEE images. Note that the figure cannot be used for algorithm comparison because the thresholds used to attain FNMR= 0.05 are widely varying such that false positive rates are very different.

Figure 34 shows that for all algorithms the dissimilarity score distributions are quite stable - there is no obvious trend in the median and the interquartile range is not increasing. The extreme values, indicated by the boxplot whiskers decline due to reductions in the sample size.

In conclusion, despite its inherent limits, the OPS-FIELD dataset reveals no evidence of an ageing effect.

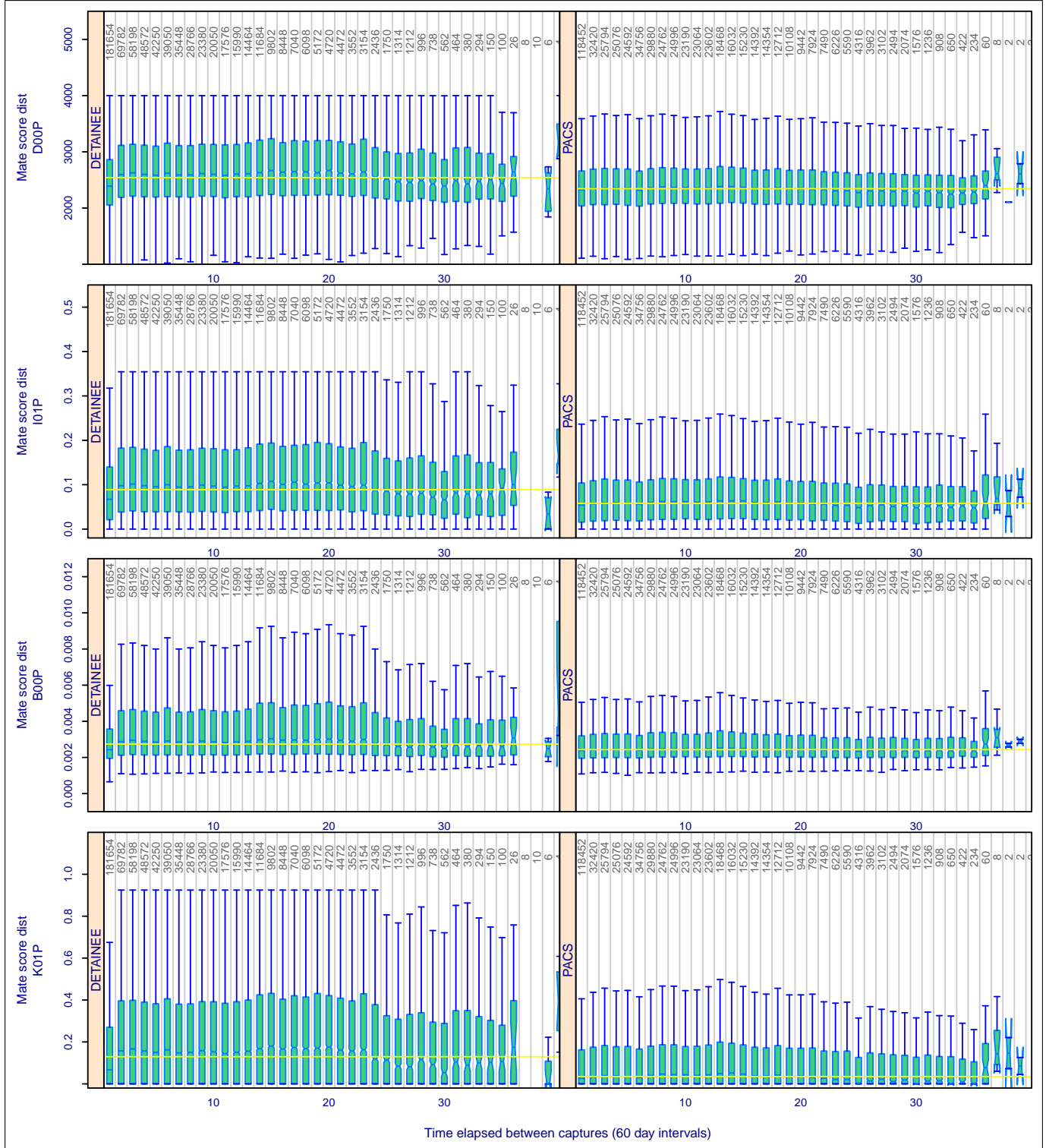


Figure 34: Stability of score distributions: For dataset OPS-FIELD, the panels show the distribution of scores from each of five algorithms applied to all mated pairs of images, plotted as a function of the time elapsed between samples. The x-axis is labelled in multiples of 60 days - the plots extend to 2160 days (nearly 6 years). Thereafter the number of samples (shown as grey text) are far too small to be significant.

5 Discussion and conclusions

The following sections list a) considerations for future ageing tests, b) considerations for operational mitigation of ageing effects, c) a classification of temporal effects, and d) some areas of future work.

5.1 Considerations for tests of biometric ageing

We advance the following guidance for consideration by experimenters tasked with measuring biometric ageing effects. The content may be applicable to all modalities, not just iris.

Comparison scores are primary response variable: While biometric system owners are primarily concerned with adverse recognition outcomes (i.e. false rejections and false acceptances), ageing effects should be quantified in terms of comparison scores. Under the definition that ageing in healthy individuals is a continuous monotonic process, its effects will manifest as decreases in biometric similarity before they result in categorical recognition failure. Recognition failures are defined by very high dissimilarity scores and these often occur as a result of a poor presentation to the capture device (e.g. motion blur, non-frontal imaging, misplacement of a finger), or poor character (e.g. face occluded by hair, irises occluded by eye lashes). For this reason, comparison scores should be adopted as the response variable in biometric ageing studies rather than binary recognition outcomes. Score distributions should be reported as cumulative distribution functions (i.e. false non-match rate, or its complement) over a wide range (e.g. $0 \leq \text{FNMR} \leq 0.7$) rather than in just the extreme tail.

Chart individual responses: Biometric ageing studies should include inspection of individual-specific effects, particularly time-series plots. Visualizations, with appropriate response variables - primarily comparison scores and other quantitative measures of causal factors (e.g. iris area, facial head pose[62]) - can reveal magnitudes, trends, biometrics zoo heterogeneities, and anomalies. Pertinent variables should be plotted against *absolute* time. For example, in a fingerprint acquisition a plot of humidity against time could be generated. Paired data can be plotted against time *difference*, the prime example being comparison scores.

Improved capture controls: Detection of iris texture ageing would best be supported by datasets exhibiting minimal variation in exposed iris area, pupil dilation, gaze angle, and other covariates. This can be achieved by design, and by review:

Iris collection and design: Some prior iris recognition studies have used tight image acquisition controls[63, 81, 76]. These used adjustable height cameras, chin and forehead rests, and calibrated LEDs to control illumination, centering, magnification, focus, motion blur, and specular reflection position. Quantitative ophthalmological studies[41], such as those for measurement of intraocular pressure (tonometry), routinely use: a) a head positioning brace that at least fixes head pitch, b) a fixed calibrated lighting environment, e.g. to scotopic (dark-light) or photopic adjustment, c) instructions to subjects to avoid stimulant and sedative intake in the prior (24) hours, d) recording of such, e) removal of contact lenses, and f) steps to ensure that the eyes are wide open.

Some current commercial cameras control pupil size via application of visible light; while this is may not be under closed-loop control³⁰, it is intended to reduce dilation differences by pushing pupil size down to a canonical person-specific value. A passive form of pupil dilation control applies a nominally fixed level of visible illumination. If this is moderately higher than the expected ambient illumination, the dilation will be stabilized against modest variations in the ambient. When the ambient has large variations, a shroud may be employed to block the ambient illumination so that the only visible illumination comes from the camera and is nominally constant. The binocular Crossmatch SEEK, Cogent Fusion and IrisID T10 are extant examples of this approach.

Some commercial iris recognition algorithms recognize eyelid occlusion and compensate for its effects on the impostor distribution (eq. (4) in [43]). However, a non-occluded image or minimally occluded image is preferable. It is possible to incorporate the eyelid occlusion algorithms into the image capture

³⁰Pupil diameter can be measured in an video stream of iris images[14], in real time, with latency of a few frames. Using size information to control visible-light irradiance would afford active control of pupil dilation. One COTS camera, the IrisGuard IG-AD100 claims to implement essentially this approach - see <https://www.prbuzz.com/technology/63941-irisguard-acknowledges-but-refutes.html> retrieved 2013-03-04.

loop and to only accept iris images that have some minimum level of exposed iris area. Both the PIER and the IRIS-ID-4000 can be demonstrated to provide such control by noting that attempts to capture images with intentional eyelid occlusion will cause the cameras to refuse to capture if the occlusion level is above some internal threshold.

Online (capture-time) sample review: Aberrant images can be detected by requiring images to pass quality tests[87] or one-to-one recognition against a reference with a moderately high threshold. Many biometric studies proceed without such checks and without any feedback to the subject. Temporal separation of collection and experimentation limits the detection of unexpected or aberrant data. Checks could be manual or automated, but should in any case be incorporated into the collection protocol. A protocol that simply captures and stores images without review is not recommended, even with nominally cooperative subjects.

Maintaining participation and motivation: Experimenters should establish mechanisms to maintain a level of engagement, investment or reward for continued participation in a biometric ageing collection. This recommendation is made to maintain adherence to the protocol over extended time periods by mitigating the effects of ennui or blasé presentation, i.e. casual compliance to the imaging protocol. For example, a subject might keep talking during collection and this could alter facial expression in a face study, or change eyelid occlusion in an iris study.

Operators, involved in continued and orderly biometric collection should similarly be incentivized or invested in some eventual outcome. In an academic study, the operator might be charged with analysis during collection, for example, on a weekly basis and this might reveal avertible artifacts.

Considerations peculiar to university-collected data: Studies confined to university populations are often characterized by small populations, due to constraints associated with duration of the study, funding, catchment area, incentives and the voluntary nature of participation. University populations are often different in their ethnic and demographic composition from typical operational populations. Finally, specific protection against, or consideration of, inter-subject interaction may be necessary. Correlated behaviour may appear between subjects who spend time together - e.g. how to use a fingerprint sensor.

Cost benefit of ageing studies: As age-related effects can be avoided by periodic re-enrollment, dedicated studies of ageing should be conducted after review of potentially high costs. These include the costs associated with: retention of a dedicated large test population over a period extending most usefully up to a decade; maintaining obsolescent equipment; employment of necessary staff; and diversion of resources from other more beneficial tests.

Health Condition: A longitudinal study should be scoped with clear goals. For example, a study intended to look at the effects of eye disease would state a delineation between the target population and any healthy control group it might use. Similarly studies intended to quantify ageing of the core biometric source in healthy individuals would only include individuals for which medical conditions are absent, or for which the biometric trait is considered to be unaffected by known medical conditions. Such studies would report their scope and procedures used to establish health condition. Alternatively, a protocol that records health condition at each capture might explicitly include surgical and acute-disease onset events in analyses that model step changes[30].

5.2 Considerations for operational mitigation of biometric ageing

Pupil dilation: The adverse effects of pupil dilation differences remain evident in state-of-the-art recognition algorithms. While these algorithms are proprietary black-boxes, linear “rubber sheet” scaling to the fixed polar representation is not a correct model of reality for sufficiently large dilation differences. It remains important operationally to enroll iris images without unusual dilation values.

Instrumenting operational systems: Operators of biometric systems should instrument their collection and recognition equipment in order to log pertinent performance-specific variables. The variables would include capture duration, all processing times, template sizes, sample qualities, comparison scores and candidate lists, and recognition outcomes. While this measurement recommendation supports logging of information beyond that

needed for operational use, it supports higher level surveys, including of possible ageing effects. Operational systems often enroll large numbers of individuals and these afford enhanced analytical opportunities.

Measure individual responses: Operators of biometric systems will almost always care whether some fraction of the user population is subject to processes which lead to recognition failure. For this reason, biometric ageing studies should include subject-specific and/or image-specific analyses. This recommendation is applicable to more general performance analysis, e.g. for investigation of effects due to age or ethnicity. The corollary of this recommendation is that population aggregated metrics should be treated with caution and are not, by themselves sufficient for quantifying ageing effects.

Operational mitigation of ageing: In long-term applications, suspected biometric ageing effects can be mitigated by re-enrollment of subjects. For example passport expiration enforces 5 or 10 year re-capture of facial images³¹. Requirements to re-enroll should be specifically planned and documented, and these should be based on longitudinal data and change measures. In applications where a credential is issued (e.g. a passport, or a driver's licence), the re-enrollment can be scheduled by instituting an expiry date for the credential. In applications where biometric data is retained on a central server, additional metadata should include collection dates and expiration dates. Old biometric data might be retained for investigation purposes.

We do not recommended so-called *automated* template-update or template-replacement strategies without a thorough consideration of the security context. The default guidance is to re-enroll only in the presence of a trained operator who should inspect for the presence of contact lenses, artificial fingerprints, and face masks.

Negative identification systems: In some applications, particularly negative identification ones such as detection of visa or benefits fraud, mitigation of ageing is not automatically possible because opportunities for second captures are not under control of the system operator. While in some cases, second encounters will be rare (e.g. counter-terrorism), other applications see second encounters more frequently - for example recidivism rates in criminal law enforcement can exceed 50%[10]. In negative identification applications, all collected data should be retained and annotated with capture date, and used in search applications. System operators should refrain from matching against only recent samples because any given sample can be affected by sample quality problems whose effects dominate ageing-related degradation. For example, face recognition against the full historical capture record has been shown to be superior to just recognizing the most recent image [35].

Use of contemporary and capable algorithms: Ageing studies conducted with algorithms that are not competitive with likely fielded capability have limited relevance to operational use. This recommendation is made for all modalities for two reasons. First, is to capture the latest developments against ageing. For example in face recognition there is ongoing research to model age progression[75] and to achieve age-invariant representations[67]. Second, because biometric samples are variously noisy or degraded it is beneficial to use high performance localization algorithms to effectively separate the computer vision problem (e.g. for finding faces or eyes or iris boundaries) from the problem of measuring change in the core biometric trait. In iris recognition, the commercial and academic consensus is that segmentation is both important and hard. It is a very often-studied problem[50] with confounding effects on detection of ageing[27].

Disease: Acute and chronic medical conditions can have severe effects on biometric recognition. Most conditions are rare, but the effect will depend on the population incidence and prevalence, and the severity of the condition. Dedicated studies have been conducted for fingerprints[26] and iris[92, 79, 23] recognition, and there are many more general publications in the medical literature.

Effects on Impostor Accuracy: Ageing studies are primarily of interest in assessing how dis-similar persons are to themselves after various periods of time. This intra-person effect is most effectively quantified in terms of a genuine matching score emanating from a recognition algorithm. The question of whether individuals become more or less distinctive (with respect to other individuals) as time passes is relevant if the demographic age structure changes. It should only be examined if resources are available because, to first order, false matches occur when two samples are biologically similar, by definition, and the expectation over a population is that subjects who are similar to other persons in the population now, will a) not become more similar to those persons, and b) will not increase in similarity to others. In cases where false match probabilities depend on

³¹ Modern e-Passports are increasingly populated with other biometric modalities - the ICAO 9303 specification defines containers for face, fingerprint, iris and a number of other modalities.

age, the time-dependence of false-match may require modeling of the changing demographic composition of the enrolled population. For example, if a biometric trait was characterized by its gradual appearance or disappearance in a population over time (e.g. if all babies' faces initially appeared identical, and then became individualized), then this would warrant quantification.

Considerations peculiar to operational log files: Operationally collected data may represent very large populations with wide ethnic and demographic variation. Additionally operational systems may be used continuously, and at any time, such that diurnal or seasonal effects are present. Another distinguishing characteristic of some operational data is that the subjects do not interact with each other, for example in a border crossing system, vs. a school population. Operational settings may not support use of the experimental controls necessary to relate cause and effect. Note a formal standard dedicated to testing of operational systems has been published[54].

5.3 Hierarchy of longitudinal effects in biometrics

Given the need in this paper to separate core iris-ageing effects from extraneous variation, we propose the following categorization.

CLASS A - *Time variation in biometric accuracy associated with the system.* Example causes would be gross dirt accumulation on a fingerprint imaging platen, failure of an infra-red LED in an iris camera, or focusing mechanism in a face camera. This definition also includes environmental variations (e.g. in illumination) that are not handled by the system; for example diurnal variation in lighting levels in a face imaging system.

CLASS B - *Temporary or remediable subject-specific variations.* These changes are due to the presentation of the subject to the system. Examples are: occlusion of the face by hair, scarf, heavy glasses; occlusion of an iris by mascara or eyelid; dryness of, or injury to, a finger;

CLASS C - *Permanent and not-remediable subject-specific variations unrelated to the biometric source.* Examples would be arthritic fingers that impede presentation of plain impressions on a fingerprint sensor, the increase in fingerprint scars associated with periods of manual labor, the growth of a full beard altering facial appearance, and pupil size reduction in iris recognition. Examples would be an increase in the incidence of broken ridges in fingerprints (inducing false minutiae), the loss of hair in males can defeat face finding algorithms, and in iris the condition arcus senilis.

CLASS D - *Time variation associated with the anatomical, physiological or behavioral aspects of the human subject.* Examples are the growth of the volar pad in childrens' fingers, changes in the facial anatomy[98], and material change in the iris texture itself.

5.4 Future work

This work, and ageing studies more broadly, can be progressed as follows.

Bilateral analysis This study treats each eye as if it were from a separate person. This loses the potential power of identifying bilateral ageing processes, i.e those that would affect both eyes. The exploratory and quantitative analyses should be augmented to bind eyes to people. For example, the mixed effects regression technique can be augmented to allow the eye label to be nested within a subject, and to capture correlations between left and right.

Other data sets The authors advocate for the use of larger populations, and these are almost only ever found in operational circumstances. They balance drawback of potential lack of experimental control, with the benefit of large populations.

Improved modelling Future work should better model the effects of known or measureable covariates. Particularly the dependence of dissimilarity scores on dilation and dilation change is non-linear and strongly influential on accuracy. The limited data in this paper and in IREX III[34] is the largest multi-algorithm set available to support construction of dilation models.

References

- [1] Aged eyes prevent iris recognition. *Healthy Seniors*, 3/7/2012. <http://www.healthyolderpersons.org/news/aged-eyes-prevent-iris-rec>.
- [2] Aging process confounds iris recognition biometrics. *Homeland Security Newswire*, 5/31/2012. <http://www.homelandsecuritynewswire.com/dr20120531-aging-process-confounds-iris-recognition-biometrics>.
- [3] Researchers question long-term reliability of iris recognition. *Third Factor*, 7/17/2012. <http://www.thirdfactor.com/2012/07/17/researchers-question-long-term-reliability-of-iris-recognition>.
- [4] Working Group 1. *ISO/IEC 2382-37 Information Technology - Vocabulary - Part 37: Biometrics*. JTC1 :: SC37, international standard edition, December 2012. <http://webstore.ansi.org>.
- [5] Canada Border Services Agency. Nexus. 2004-2013, <http://www.cbsa-asfc.gc.ca/prog/nexus/menu-eng.html>.
- [6] Home Office UK Border Agency. Uk iris. 2005-2012, <http://www.ukba.homeoffice.gov.uk/customs-travel/EnteringtheUK/usingiris>.
- [7] Sarah Baker, Kevin W. Bowyer, and Patrick J. Flynn. Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches. In *Proc. of International Conference on Biometrics*, pages 1170–1179, 2009.
- [8] Sarah Baker, Kevin W. Bowyer, Patrick J. Flynn, and P. Jonathon Phillips. *Template Aging in Iris Biometrics: Evidence of Increased False Reject Rate in ICE 2006*, chapter 11, pages 205–218. Springer, 2013.
- [9] J. Ross Beveridge, Geof H. Givens, P. Jonathon Phillips, and Bruce A. Draper. Factors that influence algorithm performance in the face recognition grand challenge. *Computer Vision and Image Understanding*, 113(6):750–762, 2009.
- [10] Thomas P. Bonczar and Lauren E. Glaze. Probation and parole in the united statesm 2007, statistical tables. Technical report, Bureau of Justice Statistics, December 2008.
- [11] H. Bouma and L. C. J. Baghuis. Hippus of the pupil: Periods of slow oscillation of unknown origin. *Vision Research*, 11:1345–1351, November 1971.
- [12] Kevin Bowyer. Research raises questions about iris recognition systems. *The Cutting Edge*. 7/12/2012.
- [13] J.C. Bradley, K.C. Bentley, A. Mughal, H. Bodhiredy, and S. M. Brown. Dark-adapted pupil diameter as a function of age measured with the neurooptics pupillometer. *Journal of Refractive Surgery*, 27(3):202–207, March 2011.
- [14] T.A. Camus and R. Wildes. Reliable and fast eye finding in close-up images. In *Proceedings of the 16th International Conference on Pattern Recognition*, volume 1, pages 389–394, 2002.
- [15] D. G. Cogan and Kuwabara T. Arcus senilis: Its pathology and histochemistry. *A.M.A. Archives of Ophthalmology*, 61(4):553–560, April 1959.
- [16] Adam Czajka. Template ageing in iris recognition. In *Biosignals - Conf. on 6th International Conference on Bio-Inspired Systems and Signal Processing*, number 73, February 2013.
- [17] V. Daneault, G. Vandewalle, M. Hbert, P. Teikari, L.S. Mure, J. Doyon, C. Gronfier, H.M. Cooper, M. Dumont, and J. Carrier. Does pupil constriction under blue and green monochromatic light exposure change with age? *J Biol Rhythms*, 27(3):257–64, 2012.
- [18] J. Daugman, November 2012. Personal communication.
- [19] John Daugman. How iris recognition works. *IEEE Transactions on Circuits and Systems for Video Technology*, 14:21–30, 2002.
- [20] John Daugman and Cathryn Downing. No change over time is shown in rankin et al. “iris recognition failure over time: The effects of texture. *Pattern Recognition*, 46:609–610, 2013.
- [21] John G. Daugman. Biometric personal identification based on iris analysis. *U. S. Patent*, 5,291,560 A, March 1994. Filed 1991/07/15.
- [22] Working Group 3. Ed. J. Daugman. *ISO/IEC 19794-6 Information Technology - Biometric Data Interchange Formats - Part 2: Iris image data*. JTC1 :: SC37, international standard edition, 2011. <http://webstore.ansi.org>.
- [23] L Dhir, N E Habib, D M Monro, and S Rakshit. Effect of cataract surgery and pupil dilation on iris pattern recognition for personal authentication. *Eye*, 24(6):1006–1010, November 2009. <http://www.nature.com/eye/journal/v24/n6/full/eye2009275a.html>.
- [24] Francie Diep. Future eye scanners must combat aging eyes. *Tech News Daily*. 7/17/2012.
- [25] 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.

- [26] Michal Dolezel, Martin Drahansky, Jaroslav Urbanek, Eva Brezinova, and Tai hoon Kim. *Influence of skin diseases on fingerprint quality and recognition*, chapter 12. November 2012.
- [27] M. Fairhurst and M. Erbilek. Analysis of physical ageing effects in iris biometrics. *IET Computer Vision*, 5(6):358–366, 2011. www.ietdl.org.
- [28] Sam Fenker and Kevin W. Bowyer. Experimental evidence of a template aging effect in iris biometrics. In *IEEE Computer Society Workshop on Applications of Computer Vision*, November 2012.
- [29] Samuel P. Fenker and Kevin W. Bowyer. Analysis of template aging in iris biometrics. In *Proc. CVPR Biometrics Workshop*, June 2012.
- [30] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. Wiley Series in Probability and Statistics. Wiley, 2011.
- [31] L. Flom and A. Safir. Iris recognition system, u.s. patent 4,641,349. 1987.
- [32] Duncan Graham-Rowe. Ageing eyes could confound biometric checks. *New Scientist Tech*, August 2010. 8/4/2010.
- [33] Duncan Graham-Rowe. Ageing eyes hinder biometric scans. *Nature/News*, May 2012. 5/25/2012.
- [34] P. Grother, G. W. Quinn, J. Matey, M. Ngan, W. Salamon, G. Fiumara, and C. Watson. Irex iii performance of iris identification algorithms. Technical Report NIST Interagency Report 7836, National Institute of Standards and Technology, <http://iris.nist.gov/irex/>, April 2012.
- [35] P. Grother, G. W. Quinn, and P. J. Phillips. Evaluation of 2d still-image face recognition algorithms. NIST Interagency Report 7709, National Institute of Standards and Technology, 2010. <http://face.nist.gov/mbe> as MBE2010 FRVT2010.
- [36] P. Grother, E. Tabassi, G. W. Quinn, and W. Salamon. Irex i: Performance of iris recognition algorithms on standard images. Technical Report NIST Interagency Report 7629, National Institute of Standards and Technology, <http://iris.nist.gov/irex/>, October 2009.
- [37] F. C. Guimaraes and A. A. Cruz. Palpebral fissure height and downgaze in patients with graves upper eyelid retraction and congenital blepharoptosis. *Ophthalmology*, 102(8):1218–1222, August 1995.
- [38] B Heaver and S. B. Hutton. Keeping an eye on the truth? pupil size changes associated with recognition memory. *Memory*, 19(4):398–405, May 2011.
- [39] K. P. Hollingsworth, K. W. Bowyer, and P. J. Flynn. The importance of small pupils: A study of how pupil dilation affects iris biometrics. In *Proceedings of the Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1–6, 2008.
- [40] Karen Hollingsworth, Kevin W. Bowyer, and Patrick J. Flynn. Pupil dilation degrades iris biometric performance. *Computer Vision and Image Understanding*, 113(1):150–157, January 2009.
- [41] Ruihua H. Hou, Jessica Scaife, Clare Freeman, Rob W. Langley, Elemer Szabadi, and Chris M. Bradshaw. Relationship between sedation and pupillary function: comparison between diazepam and diphenhydramine. *Journal of Clinical Pharmacology*, 61(6):752–760, June 2006.
- [42] R. W. Ives, H. T. Ngo, S. D. Winchell, and J. R. Matey. Preliminary evaluation of multispectral iris imagery. In *Proc. IET Conference of Image Processing (IPR)*, January 2012.
- [43] Daugman J. New methods in iris recognition. *IEEE Trans. Systems, Man, Cybernetics B*, 37(5):1167–1175, 2007.
- [44] Anil K. Jain, Arun Ross, and Sharath Pankanti. Biometrics: A tool for information security. *IEEE Transactions on Information Forensics and Security*, 1(2):125–143, June 2006.
- [45] W. J. Kennedy and J. E. Gentle. *Statistical Computing*, pages 343–344. Marcel Dekker, 1980.
- [46] Lauren R. Kennell, Randy P. Broussard, Robert W. Ives, and James R. Matey. Preprocessing of off-axis iris images for recognition. In *Proc. of SPIE Conference on Optics and Photonics for Counterterrorism and Crime Fighting IV*, volume 7119, 2008.
- [47] Peter Komarinski. Automated fingerprint identification systems (afis). page 312, 2005.
- [48] Aachal Kotecha, Ahmed Elsheikh, Cynthia R Roberts, Haogang Zhu, and David F Garway-Heath. Corneal thickness-and age-related biomechanical properties of the cornea measured with the ocular response analyzer. *Investigative ophthalmology & visual science*, 47(12):5337–5347, 2006.
- [49] Malgorzata A. Kowalska, Henryk T. Kasprzak, D. Robert Iskander, Monika Danielewska, and David Mas. Ultrasonic in-vivo measurement of ocular surface expansion. *IEEE. Transactions on Biomedical Engineering*, 58:674–680, March 2011.
- [50] Emine Krichen. Lef3a: Pupil segmentation using viterbi search algorithm. In *Proc. of 5th IAPR International Conference on Biometrics (ICB)*, pages 323–329, March 2012.

- [51] Eric P. Kukula, Stephen J. Elliott, Bryan P. Gresock, and Nathan W. Dunning. Defining habituation using hand geometry. In *Proc. IEEE Workshop on Automatic Identification Advanced Technologies*, pages 242–246, June 2007.
- [52] R. L. Lazarick. *ISO/IEC 19795-5 Biometric Performance Testing and Reporting: Access control scenario and grading scheme*. JTC1 :: SC37 :: Working Group 5, first edition, March 2011. <http://webstore.ansi.org>.
- [53] Working Group 5. Ed. T. Mansfield. *ISO/IEC 19795-1 Biometric Performance Testing and Reporting: Principles and Framework*. JTC1 :: SC37, international standard edition, August 2005. <http://webstore.ansi.org>.
- [54] Working Group 5. Ed. T. Mansfield. *ISO/IEC 19795-6 Biometric Performance Testing and Reporting: Testing Methodologies for Operational Evaluation*. JTC1 :: SC37, international standard edition, January 2012. <http://webstore.ansi.org>.
- [55] 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.
- [56] James Matey, David Ackerman, James Bergen, and Michael Tinker. Iris recognition in less constrained environments. *Advances in Biometrics*, pages 107–131, 2008.
- [57] James R. Matey, Randy P. Broussard, and Lauren R. Kennell. Iris image segmentation and sub-optimal images. *Image Vision Computing*, 28(2):215–222, 2010.
- [58] G. McConnon, F. Deravi, Hoque S., K. Sirlantzis, and G. Howells. Impact of common ophthalmic disorders on iris segmentation. In *Proc. of IAPR International Conference on Biometrics (ICB)*, pages 277–282, March 2012.
- [59] Shane McGlaun. Accuracy of iris recognition systems degrades over time according to new study. *DailyTech*, (25168), July 2012. 7/13/2012.
- [60] Stephen D. Miller and H. Stanley Thompson. Edge-light pupil cycle time. *British Journal of Ophthalmology*, 62:495–500, 1978.
- [61] A. Morte, L. Benito, E. Grasa, S. Clos, J. Riba, and M. J. Barbanj. Effects of tobacco smoking on the kinetics of the pupillary light reflex: A comparison between smokers and non-smokers. *Neuropsychobiology*, 52:169–175, 2005.
- [62] Erik Murphy-Chutorian and Mohan M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 31(4):607–626, 2009.
- [63] Hau T. Ngo, Robert W. Ives, James R. Matey, Jeff Dormo, Michael Rhoads, and Debbie Choi. Design and implementation of a multispectral iris capture system. In *Proceedings of the 43rd Asilomar conference on Signals, systems and computers*, Asilomar'09, pages 380–384, Piscataway, NJ, USA, 2009. IEEE Press.
- [64] Working Group 3. Ed. E. Tabassi (NIST). *ISO/IEC 29794-6 Biometric Sample Quality Standard : Iris image*. JTC1 :: SC37, international standard edition. Expected completion, 2014.
- [65] National Institute of Neurological Disorders and Stroke. Bell's palsy fact sheet. Technical Report 03-5114, NINDS/NIH, 2003. NIH Publication, retrieved 2003-03-11.
- [66] M. H. Papesh, S. D. Goldinger, and M. C. Hout. Memory strength and specificity revealed by pupillometry. *International Journal Psychophysiology*, 83(1):56–64, January 2012. Epub 2011 Oct 20.
- [67] Unsang Park, Yiyong Tong, and Anil K. Jain. *IEEE. Trans on Pattern Analysis and Machine Intelligence*, 32(5):947–954, May 2010.
- [68] 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 Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 32(5):831–846, 2010.
- [69] B. K. Pierscioneck, Popiotek-Masaiada A., and Kasprzak H. Corneal shape change during accommodation. *Eye*, 15(6):766–769, December 2001.
- [70] José Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-Plus*. Springer Verlag, New York, 2000.
- [71] Jeffrey R. Price, Timothy F. Gee, Vincent Paquit, and Kenneth W. Tobin Jr. On the efficacy of correcting for refractive effects in iris recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6, June 2007.
- [72] Buildings Technology Program. Lifetime of white leds. Technical Report SA-50957, Pacific Northwest National Laboratory, U. S. Department of Energy, <http://www.eere.energy.gov>, September 2009. EERE Information Center.
- [73] G. W. Quinn and P. Grother. Irex iii supplement 1: Failure analysis. Technical Report Interagency Report 7853, National Institute of Standards and Technology, <http://iris.nist.gov/irex/>, April 2012.
- [74] G. W. Quinn, P. Grother, J. Matey, and M. Ngan. Irex iv performance of iris identification algorithms. Technical Report NIST Interagency Report 7836, National Institute of Standards and Technology, <http://iris.nist.gov/irex/>, March 2013.

- [75] N. Ramanathan and R. Chellappa. Modeling age progression in young faces. In *Proc. of IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 387–394, June 2006.
- [76] D. M. Rankin, B. W. Scotney, P. J. Morrow, and B. K. Pierscioneck. Iris recognition failure over time: The effects of texture. *Pattern Recognition*, 45:145–150, 2012.
- [77] D. M. Rankin, B. W. Scotney, P. J. Morrow, and B. K. Pierscioneck. Iris recognition - the need to recognise the iris as a dynamic biologic system: Response to daugman and downing. *Pattern Recognition*, 46:611–612, 2013.
- [78] S. A. Reid, M. J. Collins, L. G. Carney, and D. R. Iskander. The morphology of the palpebral fissure in different directions of vertical gaze. *Optometry and Vision Science*, 83(10):715–722, October 2006.
- [79] Roberto Roizenblatt, Paulo Schorr, Fabio Dante, Jaime Roizenblatt, and Rubens Belfort. Iris recognition as a biometric method after cataract surgery. *BioMedical Engineering Online*, 3(1), January 2004.
- [80] M. L. Rosenberg and Martin H. Kroll. Pupillary hippus: An unrecognized example of biologic chaos. *Journal of Biological Systems*, 7:85–94, 1999.
- [81] Arun Ross, Raghunandan Pasula, and Lawrence Hornak. Exploring multispectral iris recognition beyond 900nm. In *Proceedings of the 3rd IEEE international conference on Biometrics: Theory, applications and systems*, BTAS'09, pages 1–8, Piscataway, NJ, USA, 2009. IEEE Press.
- [82] Richard D. Sanders. Cranial nerves ii, iv, vi - oculomotor function. *Psychiatry*, 6(11):34–39, November 2009.
- [83] Deepayan Sarkar. *Lattice: Multivariate Data Visualization with R*. Springer, New York, 2008. ISBN 978-0-387-75968-5.
- [84] Nadezha Sazanava, Fang Hua, Xuan Liu, Jeremiah Remus, Arun Ross, Lawrence Hornak, and Stephanie Schuckers. A study on quality-adjusted impact of time lapse on iris recognition. In *Proc. SPIE Biometric Technology for Human Identification IX*, volume 8371B, April 2012.
- [85] Judith D. Singer and John B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, New York, March 2003.
- [86] G. Sobaci, U. Erdem, F. C. Gundogan, and S. Musayev. The effect of chronic smoking on the pupil and photostress recovery time. *Ophthalmic Research*, 49:167–170, 2013.
- [87] E. Tabassi, P. Grother, and W. Salamon. Irex ii : Iqce - performance evaluation of iris quality measures. Technical Report NIST Interagency Report 7820, National Institute of Standards and Technology, <http://iris.nist.gov/irex/>, September 2011.
- [88] Elham Tabassi. Image specific error rate: A biometric performance metric. In *Proc. 20th International Conference on Pattern Recognition, ICPR*, pages 1124–1127. IEEE, August 2010. Istanbul, Turkey.
- [89] Sarah L. Taylor, Michael L. Coates, Quirina Vallejos, Steven R. Feldman, Mark R. Schulz, Sara A. Quandt, Jr Alan B. Fleischer, and Thomas A. Arcury. Pterygium among latino migrant farmworkers in north carolina. *Archives of Environmental and Occupational Health*, 61:27–32, 2006.
- [90] M. Thieme. *ISO/IEC 19795-2 Biometric Performance Testing and Reporting: Scenario Testing*. JTC1 :: SC37 :: Working Group 5, international standard edition, February 2007. <http://isotc.iso.org/isotcportal>.
- [91] P. Tomé-Gonzalez, F. Alonso-Fernandez, and J. Ortega-Garcia. On the effects of time variability in iris recognition. In *Proceedings of the Biometrics: Theory, Applications, and Systems (BTAS)*, September 2008.
- [92] Immaculada Tomeo-Reyes, Judith Liu-Jimenez, Ivan Rubio-Polo, Jorge Redondo-Justo, and Raul Sanchez-Reillo. Input images in iris recognition systems: A case study. In *Proc. of IEEE Systems Conference (SysCon)*, pages 501–505, April 2011.
- [93] John W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
- [94] Andrew B. Watson and John I. Yellott. A unified formula for light-adapted pupil size. *Journal of Vision*, 12(10), September 2012.
- [95] Barbara Wilhelm, Henner Giedke, Holger Ludtke, Evelyn Bittner, Anna Hofmann, and Helmut Wilhelm. Daytime variations in central nervous system activation measured by pupillographic sleepiness test. *Journal of Sleep Research*, 10:1–7, 2001.
- [96] Dave Wilson. Iris aging raises issues about recognition accuracy. <http://www.vision-systems.com/articles/2012/07/iris-aging-raises-issues-about-recognition-accuracy.html>.
- [97] Bradford Wing and 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, September 2011. ANSI/NIST ITL 1-2011.
- [98] Allan Wulc, Pooja Sharma, and Craig N. Czyz. *The Anatomic Basis of Midfacial Aging*, chapter 2, pages 15–29. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2012.
- [99] Neil Yager and Ted Dunstone. The biometric menagerie. *IEEE. Trans on Pattern Analysis and Machine Intelligence*, 32(2):220–230, February 2010.