

Iris Recognition Performance in Children: A Longitudinal Study

Priyanka Das^{1b}, *Graduate Student Member, IEEE*, Laura Holsopple, Dan Rissacher,
Michael Schuckers, and Stephanie Schuckers^{1b}, *Senior Member, IEEE*

Abstract—There is uncertainty around the effect of aging of children on biometric characteristics impacting applications relying on biometric recognition, particularly as the time between enrollment and query increases. Though there have been studies of such effects for iris recognition in adults, there have been few studies evaluating impact in children. This article presents longitudinal analysis from 209 subjects aged 4 to 11 years at enrollment and six additional sessions over a period of 3 years. The influence of time, dilation and enrollment age on iris recognition have been analyzed and their statistical importance has been evaluated. A minor aging effect is noted which is statistically significant, but practically insignificant and is comparatively less important than other variability factors. Practical biometric applications of iris recognition in children are feasible for a time frame of at least 3 years between samples, for ages 4 to 11 years, even in presence of aging, though we note practical difficulties in enrolling young children with cameras not designed for the purpose. To the best of our knowledge, the database used in this study is the only dataset of longitudinal iris images from children for this age group and time period that is available for research.

Index Terms—Biometrics, iris, children, linear mixed effects modelling, longitudinal, aging.

I. INTRODUCTION

DIGITAL identity and the need for convenient secure authentication have contributed to the proliferation of automated biometric recognition in the government and consumer space from the late 1990s until today. Applications include government sector (border security, criminal investigation, national security, citizenship registration, benefit distribution, human trafficking) and commercial sector (security, banking, personal identification, healthcare). ‘Persistence’ (permanence) and ‘distinctiveness, (individuality) [1] inherent to individuals are core characteristics of biometric recognition.

The possible individuality and permanence of the iris led the ophthalmologist Adler in 1965 to consider it for biometric recognition - “The markings on the iris are so distinctive that

it has been proposed to use photographs of iris as a means of identification instead of fingerprints” [2]. It is a general assumption that the iris is ‘highly stable’ [3], [4], [5], [6], [7] throughout a person’s lifetime. However, datasets supporting this hypothesis are limited, challenging to obtain and largely limited to adults. As children grow and mature, one might conjecture that biometric traits, including the iris, might change. Hence, results for adults may not apply to children. If we wish to use iris biometric recognition in applications involving children (e.g., keeping vaccination records in developing countries, investigation of human trafficking), it is important to determine how well iris recognition works for children.

The present pressing questions from a technical and practical viewpoint are-

- Does aging change the iris structure to a point that it impacts the use of iris for biometric recognition in children?
- If growth in children impacts iris recognition performance, is there an age at which these impacts are no longer seen?

This article works towards answering issues related to iris aging in children through analyzing the iris of the same children collected every 6 months, for over 3 years from 209 children in the age group of 4 to 11 years at enrollment.

International Organization for Standardization (ISO) defines biometric ‘reference aging’ as “change in error rates with respect to fixed reference caused by time-related changes in the biometric characteristics” [8]. The National Institute of Standard and Technology (NIST) defines iris aging as “irreversible changes to the anatomy, primarily the iris texture” [9]. In this work we investigate whether iris recognition accuracy decreases with the time lapsed between collection of initial enrollment and subsequent recognition images in children. Our analysis includes investigating the match scores for false match and false non-match errors, identifying the causes of the errors, and investigating the factors contributing to match score variation like enrollment age, dilation, and dilation constancy in addition to aging. A linear mixed effects model has been designed and used to understand the effect of different factors on match score variability.

A. State of Art

Though the idea that the iris could be used to recognize people goes back to at least Bertillon and McClaughry [10], modern automated, iris recognition began in 1987 with a patent by Flom and Safir [11], followed by Daugman’s development

Manuscript received September 22, 2020; revised December 18, 2020; accepted December 23, 2020. Date of publication January 13, 2021; date of current version February 12, 2021. This work was supported in part by the Center for Identification Technology Research (CITeR) and in part by the National Science Foundation (NSF) under Grant 1650503. This article was recommended for publication by Associate Editor K. W. Bowyer upon evaluation of the reviewers’ comments. (*Corresponding author: Priyanka Das.*)

Priyanka Das, Laura Holsopple, and Stephanie Schuckers are with the Department of Electrical and Computer Engineering, Clarkson University, Potsdam, NY 13699 USA (e-mail: prdas@clarkson.edu).

Dan Rissacher is with CRIA Corporation, Huntington, VT 05462 USA.

Michael Schuckers is with the Department of Mathematics, Computer Science and Statistics, St. Lawrence University, Canton, NY 13617 USA.

Digital Object Identifier 10.1109/TBIOM.2021.3050094

of the algorithm of iris pattern coding for automated visual recognition [4], [5]. Most research has concentrated on the technical development and improvement of iris recognition systems. The effect of aging on iris for biometric recognition has been a minimally studied area with most researchers initially echoing the assumption of high temporal stability of iris with little evidence to support this claim. Studies began emerging from 2008.

Between 2008 and 2014 multiple research groups investigated the temporal stability of iris as a factor of aging [12], [13], [14], [15], [16], [17], [18] and noted increased False Non-Match Rates (FNMR), increased Hamming distance between images or decay in genuine match scores (MS) in longitudinal scenario. Researches in [17] and [18] drew notice to the importance of quality factors in iris recognition and found a lack of evidence to conclude aging of iris texture as pupil dilation, iris diameter and occlusion of iris can impact performance.

In the wake of these studies, in 2013 NIST re-evaluated the studies by Baker *et al.* [13] and [14] and published their report - IREX VI, [9] which concluded that the increased FNMR in [13], [14] was the result of variation in dilation over the collection period. The report also analyzed and modelled a large match score log from a border crossing application (NEXUS) and concluded minimal ageing over a 10 year time-span, much smaller than typical day to day variations. These results were primarily from adults. NIST was unaware of a small portion of the children population enrolled in NEXUS [19] and did not take into account in their analysis [20]. Change in dilation is excluded from consideration by their definition of aging as dilation varies stochastically on a “*timescale ranging from below one second up to several decades*”, impacted by factors including environmental factors or disease and it can be mitigated by external illumination and other hardware or software solutions. This motivated discussion about the very definition of ‘iris aging’, approaches for statistical regression modelling characterizing relationship between variables [21], [22] and impact of retrospective dataset structure on modelling. The NEXUS dataset was influenced by ‘truncation’ and ‘censoring’ [23] which might impact the modelled estimate of aging. The difficulties around usage of retrospective large operational datasets for studying age related impacts on biometric recognition, specifically iris, was explored and analyzed in [24]. Research results may be impacted due to heterogeneity in the datasets because of data collected in multiple locations, different times, uncontrolled environment, seasonal impact and impacted by variability in lighting. With the exception noted, the prior works described above were specific to adults. Our own research attempts to perform similar analyses on children. Prior work related to children is described later in this section.

Various research has found that pupil size/dilation varies as a function of age [2], [25], [26], [27], [28] where pupil size is small for newborns until the first year of life, reaching its maximum size in childhood and adolescence and then gradually becoming smaller with advancing age [2] and the decay is linear (age: 17 and above) with increasing age at different luminance levels [26]. A measurable degradation in MS due to dilation differences at different ages has been noted in adults

with a corresponding increased FNMR [27] by applying a linear regression model. In the statistical analysis for our paper, we consider the effect of dilation on performance and factors including age and environment.

Research on aging effects on biometric iris recognition performance in children has been limited primarily due to the scarcity of longitudinal iris data on children. Three factors that contribute to this scarcity are - (1) the recent introduction of iris recognition for scenarios including children; (2) the Institutional Review Board (IRB) regulations for children [29]; and (3) the lack of iris capture systems specifically designed for children. In 2017 Basak *et al.* [30] created the first public multimodal dataset [31], including iris modality, from approximately 100 children aged between 18 months and four years, and evaluated the feasibility of data collection from toddlers and pre-school children with the equipment available in the market. The study concluded (i) though iris data capture is challenging in younger children, it yields best performance –99.82% single iris accuracy (98.95 left) for iris verification and (ii) 100% recognition accuracy with multiple iris images. In 2019 Nelufule *et al.* worked on iris quality assessment algorithm, using a proprietary dataset of 103 subjects in the age group of six weeks to five years [32], to identify and eliminate noise data like occluded iris, light variation, off angle, pupil dilation. No information is provided on the availability of the database. The assessment concluded that child data produces similar quality distribution as that of an adult if the unusable images are removed from the child dataset.

In real-world applications, the Aadhaar program in India, which creates a unique ID using fingerprint and iris for its population, limited enrollment age at 5yrs with re-enrollment at the age of 15yrs, and updating the biometric data every 10yrs [33]. No supporting reason has been offered by UIDAI for the age limits. Report on the Nexus program by the Canadian Defence Research Organization [34], which uses iris biometric recognition with no age limit, mentions a child was enrolled at an age of 8 months and verified successfully at the age of 9 months, 12 months and 14 months. Based on a study of the transaction logs, the report pronounced - (1) the size of the iris changes for children is not stable until 6 yrs to 8yrs of age; (2) enrollment of both eyes is much more difficult for children under 14. The subjects below 14 had a higher percentage of enrollment with only one eye; (3) successful enrollment in the younger age group (under 14 years) are substantially low compared to the middle age group, which can be attributed to less frequent younger age travellers using the system. However, no information is provided for unsuccessful enrollments in any age group. The report concludes that iris recognition is less reliable but still useful for younger age groups.

In 2018, our group issued a preliminary report focused on iris recognition in children [35] with three collections spaced by 6 months over one year - a subset of the dataset that we use in this article. The article reported on inter-session match score variation and age group analysis and concluded biometric characteristic stability over a period of one year from 123 children aged four and above. As a continuation of this work [35], we extended the analysis of the same data, now seven collections

over three years, to study the longitudinal effect of aging in iris recognition of children. Studies before Johnson *et al.* [35] in 2018 were not longitudinal, i.e., the same subjects were not followed up for more than six months, or the subjects that were followed were adults. The focus age group of our research is 4 to 14 years. In addition, many of the prior studies concentrated on FNMR performance and did not explore the root causes for the match score effects that drive the FNMR. We believe that understanding root causes is essential in making reliable decisions on temporal stability of biometric recognition systems. To the best of our knowledge this is one of the first longitudinal studies exploring the aging effect on iris as a biometric in children. The rest of this article is organized in the following order: Section II details the dataset, collection protocol, data statistics and provides an outline of the data processing. Section III details the Linear Mixed Effects (LME) model that we use in our analysis. Section IV provides a discussion on overall image quality and dilation constancy. Section V reports our results. Section VI discusses the limitations of our analysis and suggests possible future work. Section VII summarizes answers to the questions posted in Section I.

II. METHODOLOGY

The research in this article is part of a larger longitudinal study of biometric characteristics in children. The age of participants in the study spans 4yrs to 14yrs. The dataset consists of six modalities including iris. The research team collaborates with the local elementary and middle school to identify and enroll subjects for voluntary participation, in accordance with an approved IRB protocol.

The collection equipment is set up in an isolated room provided by the school for the entire duration of the collection week. However, the rooms may vary with availability at each session. Thus it is expected that there may be changes in the collection environment due to factors such as lighting and noise which may impact biometric measurements. Measures to mitigate the impact of environmental factors are taken, including drawing the blinds in the room to prevent exposure from external daylight and providing participants some leisure time to get accommodated to the room environment like lighting, body temperature, and humidity. The protocol allows participants voluntary participation on a given day of collection with the option to abstain from participating; however, any personal emotional state remains unaccounted. The same equipment, Iris Guard IG-AD100 Dual Iris Camera, has been used for all collections.

Starting in January 2016, the study covers three years, with an approximate time interval of 6 months, with seven sessions included as of November 2019. 239 subjects were enrolled; however, not all subjects participated in all sessions. The variation in the number of participants for each session is impacted by newly enrolled participants each school year, absentees, those unwilling to participate on a given day, and participants whose families have moved out of the school district. 209 subjects have participated in more than one session. Enrollment age varies from 4yrs to 11 yrs, with one subject with an enrollment age of 3yrs. After the first year, enrollment was offered to only pre-kindergarten class, which are primarily 4 years

TABLE I
ENROLLMENT AGE STATISTICS

Age	4	5	6	7	8	9	10	11	NA
Count	39	20	31	28	26	24	29	10	2

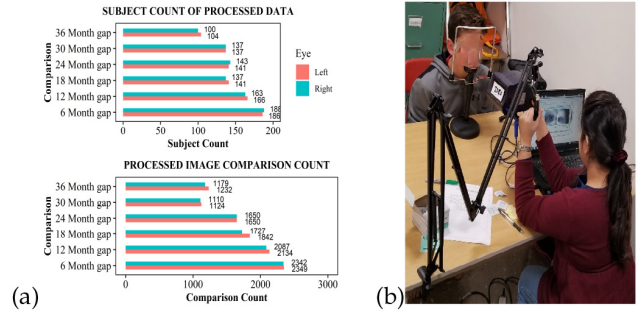


Fig. 1. (a) Summary of subject count (top) and comparison count (bottom) for right and left iris by time between sessions(6,12,18,24,30 and 36 months) in this study; (b) An articulating arm mount for the Iris ID camera.

old. To maintain privacy, we use only the birth-year of the participants. Thus, the ages are approximate. The age of three subjects were approximated based on their grade at enrollment as their birth-year was unavailable. The enrollment age breakdown is shown in Table I.

A. Iris Data Collection

The Iris Guard IG-AD100 Dual Iris Camera [36] provides ISO/IEC 29794-6 [8] compliant iris images in the NIR wavelength. The sensor has auto-focus and a mirror which assists participants to self-adjust their position relative to the sensor. Once the subjects are correctly positioned, both the eyes are captured, with a delay of a few seconds between the left and right iris capture. In addition to the NIR illuminators, the sensor also uses a flashing white light intended to stabilize the subject's pupil dilation. Capturing the iris was more difficult in younger subjects (4 - 5 yrs), as they have difficulty holding their eyes still, opening their eyes wide enough for the sensor to sense and capture the data, and not tilting their head relative to the camera. Participants were asked to place their chin on an eye-examination grade chin rest for stability, and collectors adjusted the height, distance and positioning of the camera. Some subjects still had difficulty with iris capture resulting in poor quality images as well as failure to capture. We included an articulating mount for the camera, shown in Figure 1(b), from the 6th session to provide more flexible control on part of the collector. With this, the collector had control of the position of the camera instead of asking the subjects to adjust their position with respect to the camera. This reduced the issue of failure to capture by a large degree. A second commercially available iris sensor, iCAM T10 by Iris ID, was introduced from the 4th Collection due to its binocular form factor. However, this article analyzes the data collected only from IG-AD100.

B. Failure to Acquire (FTA)

Data from approximately 25 subjects were not acquired at one or more sessions, even though they participated for other modalities. This includes failure to acquire (FTA), refusal

to participate, health issues or time-constraint for collection. More details are provided in the Discussion section and this is the focus of future work.

C. Iris Matcher

A commercially available SDK, VeriEye [37], by Neurotechnology, was used for image comparison and deriving all attributes used in this study. The iris recognition software and all metrics considered for analyses follows ISO/IEC 29794-6 [8]. Based on the IREX X [38] report, Neurotechnology is one of the top two contenders in every field of analysis. Matching was done on VeriEye version 10.0 and all other attributes were obtained from VeriEye version 11.1. VeriEye computes a similarity score or match score (MS) in the range of 0 and 1556 where a higher score indicates greater similarity; it also computes the pupil and iris radii, expressed in pixel units. VeriEye provides a threshold calibration of match score vs. FAR; for an FAR of 0.1% the corresponding match score threshold is 36 [37]. All analysis has been done at 0.1% False Match Rate (FMR) as suggested in IREX VI [8]. Additional analyses were performed in MATLAB R2018b and RStudio.

D. Statistics of the Data

418 irides from 209 subjects were processed. There is a variation in the capture between right iris (RI) and left iris (LI), due to enrollment failure.

Two images were captured for each eye at the first and sixth session; for all other sessions, four images were captured per eye. Images collected before the seventh session had high intra-session correlation due to the internal setup of the camera which captures multiple images within seconds. The collection protocol was modified in the seventh session. Four images were captured per eye in two different sets spaced by a time gap of approximately two minutes. Participation count of number of subjects and the number of comparisons for different time-frames are summarized in Figure 1(a).

Six images from two different subjects with clearly observable artifacts were discarded from the database based on a manual cleaning. Two different errors were identified in those images - four images were blurred with iris obstructed by hair and two images had an obscured iris with reflecting light from eye glasses. Examples are shown in Figure 2.

The first session that a subject participated is considered as the enrollment session. Images collected at the enrollment session are mated with all images from subsequent sessions. Intra-session image comparison is not considered in this study due to:

- intra-session image comparison does not contribute to a real world application scenario
- the measures have no contribution towards how increased time between enrollment and probe impact performance.

The data has been sectioned and analyzed in three ways: G1: analysis of all 209 subjects who participated in more than one session; G2: analysis of a subgroup of 105 (RI:101) subjects who participated in at least 1st and 7th session with intermittent absentees; and G3: analysis of a subgroup of 63 (RI:62) subjects who participated in all seven sessions. LI and RI irides

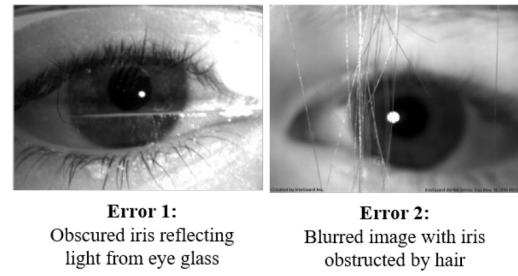


Fig. 2. Examples of images of two subjects removed from the dataset on manual cleaning. Obvious errors were identified (obscured iris and blurred iris). A total of six images were removed. Two images had Error 1 and four images had Error 2.

were processed separately. The total number of comparisons per subject are non-uniform. A total of 20428 (right: 10096, left: 10332) comparisons were performed; 6421 (LI: 3216; RI: 3205) probe images were compared against 1301 (LI: 658; RI: 643) enrollment images from 209 subjects.

III. LINEAR MIXED EFFECTS MODEL

Ordinary Least Square (OLS) regression modelling gives exploratory insight into the match score (MS) as time increases between enrollment and query. A linear functional form is often used for individual growth modelling based on visual inspection of the individual patterns, constricted time-duration of the data (3 years) and referencing the literature [9], [39]. However, modelling longitudinal data of sampled data sets for constricted time duration to explore changes in a population is a challenge due to the following reasons:

- Different subjects will have a different trajectory of the empirical growth;
- A simple linear model might not be a proper representation of accumulative intra-individual and inter-individual variation for longitudinal data; and
- Variation in MS may reflect other factors in addition to ageing like random fluctuations of different factors (error or quantitatively unaccountable factors like environment, medication, illumination, emotion etc).

Linear Mixed Effects Models address the disparity between within-individual variation, and inter-individual differences. “Mixed Effects” refers to a combination of *fixed effects* and *random effects* on the response variable. *Fixed effects* quantify the response variable with respect to the effects of predictors on the inter-subject variation and *random effects* quantify any possible variation in the response that could be accounted for by predictors on the intra-subject variation. The basic model structure is defined as:

$$\text{Response} \sim F_{\text{expression}} + (R_{\text{expression}}|Factor) \quad (1)$$

Here, $F_{\text{expression}}$ represents the fixed effects model matrix. Each random effect is represented by: $(R_{\text{expression}}|Factor)$ where, $R_{\text{expression}}$ is the random effect predictor having different effects on the *Response* variable for each level of grouping factor, i.e., *Factor*.

All modelling and analysis has been computed on the R platform using the package ‘lme4’ [40]. When a restricted portion of lifespan is being analyzed, as in our case, the model

does not predict or conclude anything beyond the time-frame of research.

A. LMER Model and Predictors

The variation in MS is modelled using the predictors - time difference between collection of enrollment and probe image (TD), enrollment age (EA), probe dilation (PD) and dilation difference between probe and enrollment image (ΔD) as described below.

- 1) *Time Difference (TD)*: The difference in time between the enrollment and probe image measured in “months”.
- 2) *Enrollment Age (EA)*: Age when the subject is enrolled in the study, i.e., the age of their first collection in “years”.
- 3) *Dilation (D)*: Dilation or pupil dilation is a dimensionless quantity measuring the degree to which the pupil is dilated or constricted, measured as a ratio of pupil radius and iris radius. The measure follows ISO/IEC 29794-6 [8] as defined below.

$$\text{Dilation}(D) = \frac{\text{Pupil radius}}{\text{Iris radius}} \times 100 \quad (2)$$

The dilation of the probe image is considered for the model denoted by PD.

- 4) *Delta Dilation (ΔD)*: Difference in the pupil dilation between a mated pair of iris images. The measure follows NIST work in [9] as below.

$$\text{Delta Dilation}(\Delta D) = 1 - \frac{1 - \frac{D1}{100}}{1 - \frac{D2}{100}} \quad (3)$$

considering, $D1 \geq D2$, where, $D1$ and $D2$ are the pupil dilation of the first and the second iris images as estimated by equation (2).

The designed model to predict the match score is represented in equation (4). The model considers linear change in the match score for up to 36 months between enrollment and query. Though the modelling of the MS is considered linear statistically, non-linear changes in the predictors are taken into account by including second order terms. To the best of our knowledge, there is no prior indication or evidence of non-linear change in the iris. We have tested different configurations of the model. The model in Equation (4) was concluded to be the best fit. This model has been used to interpret results and is extensively discussed in Section V-A1.

$$\begin{aligned} \text{MS} \sim & \beta_0 + \beta_1 \text{TD} + \beta_2 \text{EA} + \beta_3 \text{PD} + \beta_4 \Delta D \\ & + \beta_5 \text{EA}^2 + \beta_6 \Delta \text{PD}^2 + \beta_7 \Delta D^2 + b_{0i} \\ & + b_{1i} \text{TD} + b_{3i} \text{PD} + b_{4i} \Delta D + b_{6i} \text{PD}^2 \end{aligned} \quad (4)$$

where,

- β_k is the fixed regression coefficient for corresponding parameter, k .
- b_{ki} is the random regression coefficient for corresponding parameter, k , for subject, i .
- $\beta_0 + b_{0i}$ is the sum of fixed and subject specific random intercept corresponding to the initial state.
- $\beta_k + b_{ki}$ is the subject specific (i) gradient for the corresponding parameter k .

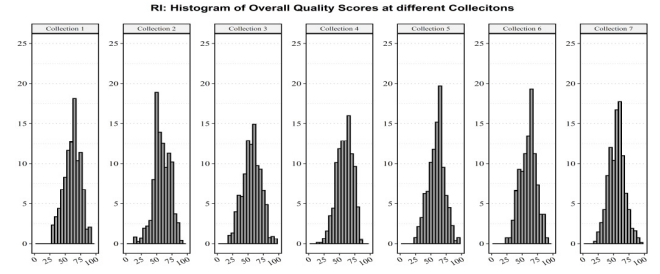


Fig. 3. Distribution of overall quality score of all images from Collection 1 through Collection 7 (left to right).

- Time Difference (TD) between enrollment and subsequent verification has been considered for both fixed and random effects to account for possible variability in MS due to both intra-subjects and inter-subject effects of TD.
- Enrollment Age (EA) has been considered only for fixed effect. All other predictors are considered for both fixed and random effect.
- Both linear and quadratic factors are introduced for variables: dilation difference (ΔD), probe dilation (PD) and enrollment age (EA) for fixed effects. We assume the non-linearity in modelling with respect to match score (MS) for inter-subject data.
- Both linear and quadratic factors are considered for probe dilation in random effects for intra-subject data, to account for any non-linearity in dilation for each subject over time.
- We consider correlation between all coefficients associated with the same random effect term, i.e., for each subject we consider correlation between coefficients of TD, PD, ΔD and PD^2 .

IV. OVERALL IMAGE QUALITY AND DILATION CONSTANCY (DC)

ISO/IEC 29794-2 [8] defines quality measures for a single image, which can be combined to a single scalar overall iris quality score ranging from 0 to 100; higher value is better quality. The score reflects the expected performance of the iris image. The overall image quality scores of all images collected in the seven sessions are derived from VeriEye [37] and the distribution is shown in Figure 3. The t-test statistics show there is a statistically significant ($p < 0.001$) difference in the mean of the overall quality scores between collection sessions when compared to Collection 1 except for Collection 6 ($p > 0.5$). However, as the mean values are $-57.79, 51.72, 54.53, 55.97, 54.95, 55.97, 54.95$ for Collection 1 through Collection 7, the difference in the mean quality scores are practically not significant. Pearson’s correlation between image quality score and age at collection of the image spanning 4 to 14 years is weak with correlation coefficient of 0.14.

Dilation Constancy (DC) evaluates the similarity in pupil dilation in a mated pair of iris images. It is a dimensionless quantity measured as below.

$$\text{Dilation Constancy}(\text{DC}) = \frac{100 - \max(D1, D2)}{100 - \min(D1, D2)} \quad (5)$$

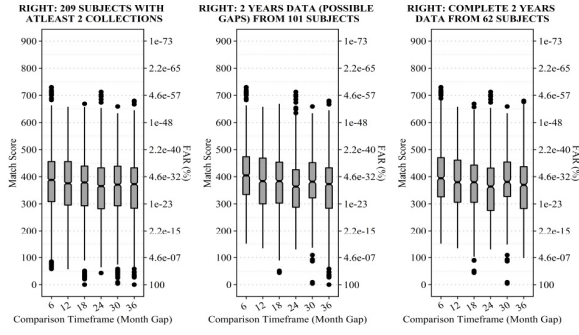


Fig. 4. Box-plots of all scores from individual subjects for right iris grouped based on participation: G1 (left) : all subjects have at least one mated pair data; G2 (middle): all subjects have mated pairs from at least 6 and 36 month TF; G3 (right): all subjects have mated pairs from all TFs (6, 12, 18, 24, 30, 36 months). Relative False Accept Rate (FAR) w.r.t the MS is shown in the right y-axis.

where D1 and D2 are the pupil dilation of the first and the second iris images as estimated in equation (2) as detailed in Section III. Delta Dilation depends on the variability of both D1 and D2. If either of D1 or D2 changes, delta dilation changes. If the delta dilation increases, the dilation constancy decreases and vice-versa. DC has been used to understand the match score variability with DC and any relation between false non-match and DC. DC and Delta Dilation are related: $\Delta D = 1 - DC$.

V. RESULTS & DISCUSSION

This section reports the analysis of the performance of biometric recognition of iris in children with increasing time difference between enrollment and probe in presence of variable factors - enrollment age, dilation, and dilation difference between enrollment and probe. The next sections cover modelling of the match scores using linear mixed effect model for groups based on participation and age, dilation and dilation constancy, false non-match rate, detailed examination of the false non-match cases and ROC of the dataset. The overall goal is to assess the viability of biometric applications of iris recognition in children as they age.

A. Iris Match Score Analysis: Participation Based Groups

Based on participation we have three groups: G1, G2 and G3 (refer Section II-D). We have analyzed MS following the model described in Section III-A. A visual summary of the data is shown in Figure 4. Extensive analysis of the results from the model for G1-LI addressing the impact of aging on iris recognition performance in children are detailed in the following section. The same conclusions are true for all three groups with only a small variation in coefficients.

1) *Iris Match Score Modelling*: The Linear Mixed Effects Regression Model (LMER) addresses the questions - Are the explanatory/predictor variables - time difference (TD), enrollment age (EA), pupil dilation (PD), change in dilation between mating pair of images (ΔD) a significant predictor of the response variable (match score)? Is there a significant decrease in the iris match scores in children as time between enrollment and query increases? The LMER model with random slope for each subject is designed as shown in Equation (4) and

TABLE II
FIXED EFFECTS FOR LEFT AND RIGHT IRIS

Variable	Parameter	Left Iris (Est \pm SE)	Right Iris (Est \pm SE)
Intercept	β_0	387.0 \pm 12.7 ***	401.5 \pm 11.6 ***
TD	β_1	-1.09 \pm 0.33 **	-1.39 \pm 0.31 ***
EA	β_2	10.88 \pm 3.4 **	13.3 \pm 3.1 ***
PD	β_3	-0.74 \pm 1.3 NS	-1.4 \pm 1.3 NS
ΔD	β_4	-394.9 \pm 78.9 ***	-269.5 \pm 61.86 ***
EA^2	β_5	1.4 \pm 1.7 NS	-1.9 \pm 1.5 NS
PD^2	β_6	-0.22 \pm 0.20 NS	-168.2 \pm 0.16 NS
ΔD^2	β_7	-1540 \pm 334.9 ***	-1360.2 \pm 355.7 ***

Significance Code: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ; *** indicates p-value between 0 and 0.001 with significance level 0.001 and so on.
Est.: Estimate, SE: Standard Error, NS: Not significant

TABLE III
RANDOM EFFECTS LEFT AND RIGHT IRIS

Groups	Parameter	Standard Deviation	
		Left Iris	Right Iris
Intercept	b_{0i}	129.7	118.9
TD	b_{1i}	4.2	3.9
PD	b_{3i}	15.7	15.3
ΔD	b_{4i}	886.38	636.08
PD^2	b_{6i}	2.4	1.8
Residual		40.7	40.6

detailed in Section III. We have referenced the model used by NIST in IREX VI [9] as a base model and augmented it with additional terms like enrollment age as a quadratic factor. Enrollment age is added to account for the variable age of the subjects in our dataset. The right and left iris are modelled separately. The multi-collinearity between independent variables are eliminated by re-centering the following variables - PD, EA and ΔD , following discussion in [21]. The fixed effect of the designed model considers the overall inter-subject variability of TD, EA, PD, ΔD and impact of these variabilities on the MS of the entire dataset representing the population. Approximately 5.54% (4.03% for RI) variation in the MS is explained by the fixed effects. The intra-subject variability in the intercept and the gradient is induced in the model by the random effects of TD, PD, ΔD factored by individual subjects. The marginal and combined r-squared values are obtained using the 'r2glmm' package. Both fixed and random effects combined accounted for approximately 96.34% (95.91% for RI) of variability in the match score. The 3.6 - 4.1% (approx.) variability which is not explained by any covariates in the model can be directed to the miscellaneous factors like illumination and physiological condition which is beyond the scope of this study. Restricted or Residual Maximum Likelihood (REML) method is applied for model fitting. The model is created based on 10096 observations from RI and 10332 observations from LI grouped into 209 distinct individuals. The fixed and random effects are summarized in Table II and Table III respectively.

The analytic interpretation of the output of the model from G1-LI is noted below. G1-RI has the same interpretation, but is not included due to space limitations.

- *Null Hypothesis (H_0)*: There is no correlation between the individual predictors and the response variable (MS).
- *Alternative Hypothesis (H_A)*: There is correlation between the individual predictors and the response variable.

TABLE IV
NUMBER OF SUBJECTS FOR EACH AGE GROUP

Groups	4 to 6 years (LI/RI)	7 to 8 years (LI/RI)	9 to 11 years (LI/RI)	Total
G1	90/90	54/54	63/63	207/207
G2	25/25	37/34	43/42	105/101
G3	13/15	22/21	28/26	63/62

- Estimate (Est.) provides a measure of change in MS from the intercept for each unit change in each variable.
- Very high similarity is observed in both estimate and variability of TD coefficient (β_1) between both RI and LI. The estimated decrease in MS, β_0 , due to TD estimate, β_1 , is 1.09 ± 0.33 (approx) with each month increase in the time gap between mated pair of images, given that the null hypothesis is true. With $p < 0.01$, the null hypothesis is rejected. This signifies there is a statistically significant decaying relationship between time gap and match score. However, the practical significance of the estimate is trivial considering the range of the scores (0 to 1557), the fixed effect intercept being 387.0 (401.5 for RI), and the threshold score being 36 at FAR 0.1%. For example, a MS estimate of 387.0 would reduce to 347.76 in three years considering the effects predicted by the model. This score would still result in a match and does not impact biometric recognition over three years.
- A positive baseline shift is noted for MS estimate, β_0 , due to EA estimate, β_2 . With an increase of EA by 1 year, the MS increases by 10.8 (13.3 in RI) ± 3 (approx) with $p < 0.01$. This implies that EA has a significant effect on MS. So, with higher EA, a higher MS can be noted. It is expected that with increased age the interaction of the children with the data acquisition system improves. Older kids are able to hold still and have higher stability by keeping their eyes open and directed towards the sensor until the sensor auto-captures. The effect of EA on MS is linear as the quadratic factor is statistically insignificant.
- Pupil dilation of the probe image (PD), β_3 affects the MS negatively; however the impact is statistically insignificant.
- Difference in pupil dilation (ΔD) between two mated pair of images, β_4 , has negative correlation with MS. ΔD in our dataset varies in the range of 0.1 to 0.4. Thus the MS may decrease in the range 39.4 to 157.6 due to ΔD . It is important to note that the effect on MS due to ΔD is higher by approximately 29 to 45 times than that of TD (LI). The effect of ΔD is non-linear.
- Ordering covariates based on adding random variability to the response in the model while considering intercept for each individual subject (random effect) as seen in Table III:
 $\Delta D > \text{Subject} > \text{Dilation of the probe image} > \text{Time Difference}$
 Interestingly, we can infer from the model that dilation difference between two comparing images (in our case from 2 different TFs) adds in the most variability to the match score than any other significant factor. Aging effect contributes to the least variability in MS.

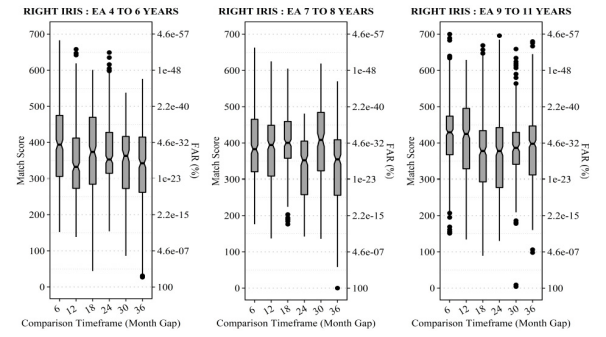


Fig. 5. Box-plots of all scores from subjects grouped based on age for G2 (all subjects have mated pairs from at least 6 and 36 month TF) for different TFs for RI. G1 and G3 plots are similar to this and thus are not included to accommodate space and reduce redundancy. Relative False Accept Rate (FAR) w.r.t the MS is shown in the right y-axis.

B. Iris Match Score Analysis: Age Based Groups

The dataset has subjects aged between four to 11 years at enrollment. For this analysis, the entire dataset is grouped into three sections based on enrollment age - 4 to 6 years, 7 to 8 years and 9 to 11 years for analysis. The count of subjects for each age group are non-uniform; bin sizes are summarized in Table IV. Each age group has been analyzed for each participation based group - G1, G2 and G3 following the LMER model represented in equation (4). We do not have age information from 2 subjects in G1 and thus were removed. Thus, instead of 209, 207 subjects were analyzed in this section for G1. Boxplots of MS as a factor of TF for different age groups are shown in Figure 5. The attributes of the boxplots for G1, G2 and G3 are similar; thus only plots from G2 are presented to accommodate space. Conclusion from the LMER model for each age group are noted below:

- Estimated TD has a similar impact on MS on different age groups as our previous model. With the estimated decay of MS β_1 by < 1.02 ($p < 0.01$) for each month difference between the capture of the mated pair of images, the impact is practically insignificant. For the age group of 4 to 6 years, TD is not a statistically significant predictor of MS.
- ΔD has similar correlation with MS as we noted in the previous model analysis. The impact is linear for the 4 to 6 years age group and is non-linear for the other age groups.
- In contrast to our observation from group-based analysis, where all age groups were modeled as a whole, EA is not statistically significant when considering shorter age ranges compared to a larger age range of the full analysis.

We also did t-test analysis to test the difference in mean MS between different age groups, if any. The test concluded that the mean MS of the age group 9 to 11 is higher than 4 to 6 years ($p < 0.001$) and 7 to 8 years ($p < 0.01$). The difference in mean MS between age groups 4 to 6 years and 7 to 8 years is not statistically significant ($p > 0.05$). However, we do not notice any impact on the tails of the distribution, i.e., the iris performance is not impacted as a result of age or aging as will be discussed in future sections. However to make a substantial conclusion on this, a more intensive study for each age is needed.

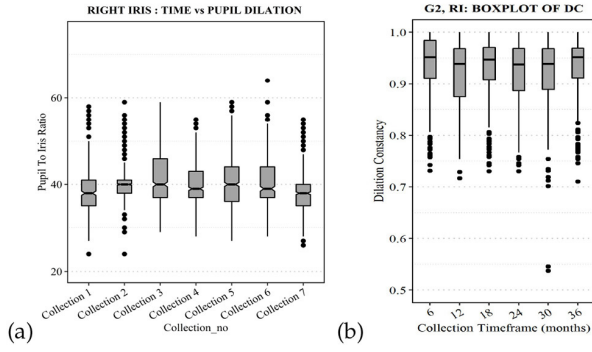


Fig. 6. (a) Boxplot of pupil dilation from all images captured in each session for RI; (b) Summarized iris dilation constancy compared to the enrollment image at different TFs (6, 12, 18, 24, 30, 36 months from enrollment) for right iris.

C. Dilation and Its Relationship With Match Score

Dilation is an important factor to be considered in iris recognition. Studies have confirmed change in ocular biometric parameters with change in dilation [41]. These changes might affect the performance of algorithms used to generate a match score. Variable dilation is considered as an explanatory variable in the match score variability with time in recent research concerning evaluation of aging effect [9]. Physiologically, studies have indicated that pupil size varies with age [2]. Pupil size is small for the first year after birth and gradually attains its maximum size in childhood and adolescence and again gradually becomes smaller with time. However, the pattern may not be reflected in this study due to the limited TF of 3 yrs and having no infants in the dataset. Dilation is also a response variable to environmental factors like illumination, weather, subject's medical history and fight and flight response of the Central Nervous System in response to a stressful environment. In this section, we study dilation as an independent factor impacting performance of iris recognition. Dilation is derived as defined in Equation (2).

Figure 6 shows dilation for each collection which includes all participating children at a fixed time point (e.g., Collection 2 in November 2017). All statistical tests were compared to Collection 1. The t-test of the mean dilation for the six collections compared to Collection 1 shows that the mean dilation is relatively constant ($p > 0.05$) except for Collection 3 ($p < 0.01$). Chi-square analysis of dilation concludes that there is statistically significant ($p < 0.05$) difference in the dilation variance of Collection 3, 5, 6 and 7 when compared to Collection 1. Measures are taken during collection to minimize impact of illumination and weather as detailed in Section II and II-A. Although measures are taken to control data collection environment, the dataset is not immune to impact from miscellaneous variability factors.

Dilation constancy (DC) is the measure of similarity in pupil dilation between enrollment and probe image as detailed in Section IV. Figure 6(b) summarizes the DC between enrollment image and respective probe images collected at different TFs. Only the graph from G2-RI is included in the paper to accommodate space. All three groups based on participation for both LI and RI show similar attributes. In the majority of the TFs across different groups, the upper 75% of the mated

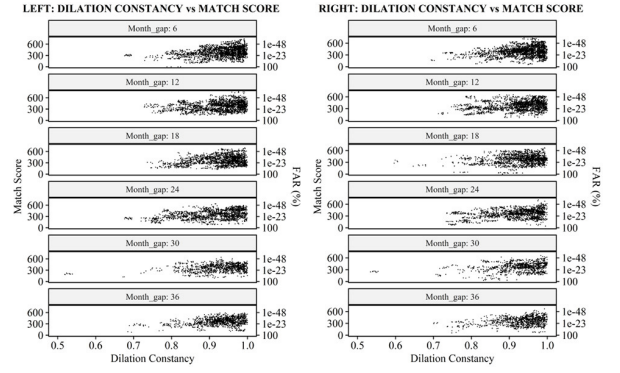


Fig. 7. Graphical presentation of match scores (y-axis) and dilation constancy (x-axis) for different time-periods (top to bottom: 6, 12, 18, 24, 30, 36) for left and right iris (left and right).

pairs have DC above 0.9. The median DC is greater than 0.9 across all sessions and all TFs. In our dataset, the DC ranges from 0.55 to 1 and we observe more outliers in the lower quartile of the plots. We implemented a LMER model as described in Equation (6) to predict DC employing TF and EA as predictors. From the model we conclude that the impact of TF and EA on DC is statistically insignificant ($p > 0.05$). Increased time difference between probe and gallery does not influence the DC.

$$DC \sim \beta_0 + \beta_1 TD + \beta_2 EA + \beta_5 EA^2 + b_{0i} + b_{1i} TD \quad (6)$$

where,

- β_k and b_{ki} are the fixed regression coefficient for corresponding parameter, k and the random regression coefficient for corresponding parameter, k , for subject, i , respectively.
- $\beta_0 + b_{0i}$ is the sum of fixed and subject specific random intercept corresponding to the initial state.

Figure 7 illustrates the relationship between MS and DC. A triangular pattern is noted. Lower dilation constancy translates to lower match scores; however, low MS can be seen even with high dilation constancy likely due to other factors such as usable iris area, noise, or other quality factors. DC, as has been previously reported, impacts the MS. However, in this study, the Pearson's correlation between DC and MS of all matches performed is weak with correlation coefficient of 0.28. DC in this dataset, with substantial outlier variability has not impacted the tails of the distribution, i.e., DC did not impact practical application of iris recognition in children to a point of false non-match. Saying that, iris recognition performance is sensitive to dilation and thus the data has been collected in a semi-controlled environment to minimize impact on dilation.

D. Iris Recognition Performance: False Non-Match Rate

False Non-Match Rate (FNMR) is an evaluation metric used in prior aging studies as discussed in Section I-A and defined in ISO/IEC standard 2382-37 [42]; it is the fraction of mated (from the same individual) images that fail to match; it has operational impact. We calculated FNMR with a MS threshold value of 36, equivalent to False Match Rate (FMR) of 0.1% using the VeriEye FMR calibration [37]. Images were

TABLE V
SUMMARIZING THE FALSE NON-MATCHES IN THE STUDY

Coded ID	Collection Number for Rejection	Time-frame of Rejection	Number of Rejections	Nature of Failure	LI/RI/Both
20181012017	Collection 7	6 month	4 of 8 mated pairs	Angle of Presentation w.r.t. camera	LI
20160104994	Collection 5 and 7	18 months; 30 months	18 months: 14 of 16 mated pairs 30 month: 2 of 16 mated pairs	Poor Enrollment	RI

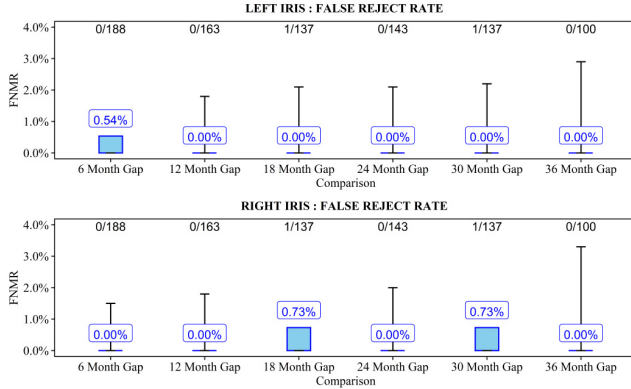


Fig. 8. FNMR at different TFs (6, 12, 18, 24, 30 and 36 months from enrollment) for left and right iris. Values marked in blue indicate FNMR based on our observation at 0.1% FMR. The black lines show the error-bars calculated based on ‘Rule of 3’. X/Y above each bar indicates the number of subjects rejected/total number of subjects for this TF.

rejected when the MS were below the set threshold. A chronic increase in FNMR with time may be correlated with degrading biometric performance with increased time duration between enrollment and probe.

Variation in FNMR is shown in Figure 8. Rejection is observed only in G1. No rejections were seen from G2 and G3. Single cases of rejections are observed at 6, 18 and 30 month TF. The three occurrences of FNM cases are summarized in Table V. At 6 month TF, RI of one of the 186 participating subjects is rejected; thus, the FNMR at 6 months is 0.54%. At each of 18 and 30 month TFs, the LI of one of the 137 participating subjects were rejected; The FNMR at both TF is 0.73%. The slight increase in the FNMR from 0.54% to 0.76% is not an increasing trend in the FNMR but a reflection of decreased number of subjects.

Out of 20428 mated pairs of images from 418 different irides in the study 20408 mated pairs were above the threshold and results in a match. 20 mated pairs from two different irides from two subjects failed the set criteria and were falsely rejected. The failures are summarized in Table V and detailed analysis and understanding of the root causes are elaborated in Section V-E. Subject 20181012017 had rejection in the LI in one session and subject 20160104994 in the RI in two different sessions. No subjects were rejected in both irides. When a subject is rejected, not all images from that session are rejected (refer to Table V - Number of Rejections). The average DC of the rejected images is substantially high (0.8 to 0.9). Thus the FNM are not a result of a large dilation difference between the mated pair of images in the dataset. Since multiple images from a single triggered capture were collected on each visit (except seventh session), the images have high correlation. Due to the intrinsic setup of the sensor, the sensor provides multiple images for each triggered capture, all captured within a few

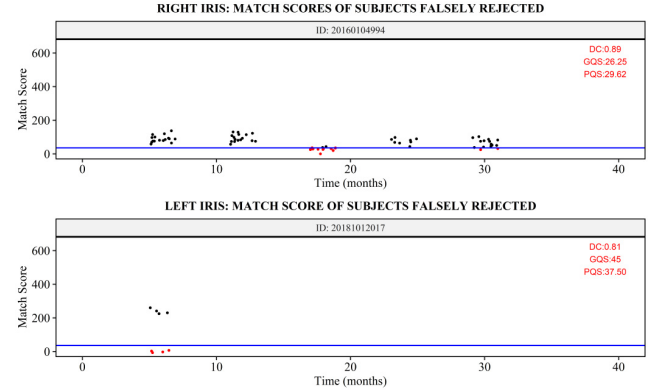


Fig. 9. Match scores for all cases of rejections in the study. Each graph shows the match scores (y-axis) of individual subjects at different time-frames (x-axis). The rejected comparisons MS are marked by red dots. The blue line indicates the set threshold (MS 36 at 0.1% FAR) for the study. ‘DC’, ‘GQS’ and ‘PQS’ denotes the average dilation constancy, average quality score of gallery and average quality score of probe of all rejected images of that subject respectively.

seconds. However, with change in protocol in the 7th session, multiple images were collected in two sets, decreasing the high correlation between images collected in different sets. Two of the three occurrences of rejections were collected in the 7th session (refer Table V). Figure 9 depicts the individual exploratory view of the MS from the two subjects for all the TFs they participated. Subject 20181012017 has recorded comparisons at 6 month TF and showed rejections in four of the eight mated pairs. Since all of the images were not rejected it is highly unlikely that the rejection is due to irreversible change in the iris texture. Subject 20160104994 had matching data from five TFs (6, 12, 18, 24 and 30 month TF) in which rejection was observed in two sessions - 18 and 30 months. At 18 month TF, 14 of the 16 mated pairs were rejected with the other two having scores just above the threshold. At 30 month TF only two of the 16 mated pairs were rejected. However the MS of the matched pairs of comparison exceed the set threshold by a large margin (refer Figure 9). In between 18 and 36 month TF, the MS “returns” to above threshold at 24 month TF. These sporadic false non-matches and a trending low MS all through the five sessions indicate that the false non-matches are a result of other factors than aging. The underlying causes of rejections are explored in more detail in the next section.

Since the occurrences of false matches at different TFs are episodic, the results of the FNMR as marked in ‘blue’ in Figure 8 are statistically analyzed by the ‘Rule of 3’ and the projected FNMR are marked by the error bars. This rule has been adapted in practice in biometrics and is described in ISO/IEC 19795-1 [43]. The rate of occurrence of a particular event in a population which does not occur in an experiment is calculated by the ‘Rule of 3’. In our study, at particular

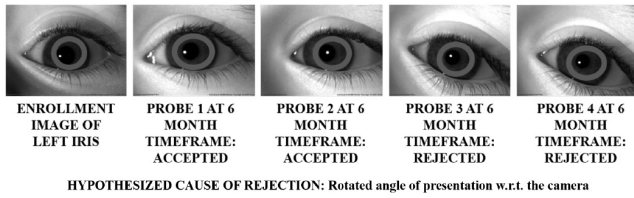


Fig. 10. False Non Match: Case 1: Iris images of subject 20181012017 at enrollment (left) and the probe images matched and falsely non-matched at 6 month TF. Iris patterns are partly obscured to maintain privacy.

TFs (12, 24 and 36) there is no occurrence of FNM. The rule states that $3/N$ is the upper bound of such an occurrence at 95% confidence interval, where N is the number of people in the experiment. We have 209 subjects in the study. Thus based on this rule the range of FNMR in this study is 0% to 1.4%. However, since our experiment is longitudinal, having different numbers of subjects at different TFs and between LI and RI, the fragmented projected maximum FNMR is shown in Figure 8. The maximum upper bound FNMR is 3.3% for RI at 36 months. Instances where the projected error is not shown are cases where we have recorded false non-matches and do not qualify the criteria for the ‘Rule of 3’. The increase of the upper limit of FNMR is a reflection of the decreasing number of subjects, i.e., the upper bound is higher when there are fewer subjects.

E. Detailed Examination of FNM Images

We did a root cause analysis for false non-matches of the two subjects (refer Table V) individually.

1) *Case 1 (Coded ID - 20181012017)*: The subject was enrolled in the 6th session, having two collections in the dataset, i.e., one recorded TF at 6 month. Rejection was observed only in the LI. Four out of eight comparisons (2 of 4 images from the 7th session) are rejected. The 4 comparisons from the truly accepted mated pair had an average MS of 238. The rejected 4 comparisons had an average MS of 0. The MS at different TFs are plotted in Figure 9. The average DC of the four rejected pair of mated images is 0.81. The average overall image quality of the rejected images of the gallery images (GQS) is 45 and the probe images (PQS) is 37.5 which are below the mean overall quality score distribution. The images are displayed in Figure 10.

Visual inspection revealed that in the two rejected iris images (4th and 5th images in Figure 10), the eye is captured at a rotated angle of 15 degree w.r.t the camera. VeriEye being a black box, we are not aware of the compensation adapted by the algorithm to accommodate angle of rotation considered in their algorithm. However, manually rotating the images by -1 degree, the VeriEye software is able to match the images. This is a case where the issue is the presentation of the iris to the camera which is challenging in children and has been observed in previous studies [30], [32]. We conclude that the root cause of rejection has no relation to age/aging of the iris biometric. However, errors such as these lead to false non-matches. We observe and suggest that when working with children there is a need for a tool to detect the canthi and rotate the image

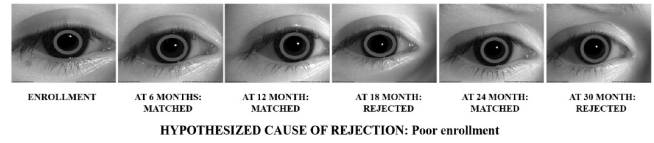


Fig. 11. False Non Match: Case 2: Example of images collected at different TF for subject 20160104994. The subject matched with the enrollment image at 6 month and 24 month TF but falsely non-matched at 12 and 30 month TF.

to bring the canthi onto a horizontal line for cases where the image has significant rotation.

2) *Case 2 (Coded ID - 20160104994)*: The subject participated in six of the seven collections. We have matching information from five TFs (6, 12, 18, 24 and 30 months). The MS at different TFs are plotted in Figure 9. The subject was falsely rejected at 18 month and 30 month TF only in the RI. At 18 month TF, 12 of the 16 comparisons performed are rejected. The average MS of the four matched image-pairs is 38.2, which is just above the set threshold. At 30 month TF, 2 of 16 comparisons performed are rejected. The average MS of the 14 matched pairs is 66.2. As noted, the subject has a very low MS for all true matches across different TFs. The average DC of the rejected mated pairs of images is 0.89.

Visual inspection and analyzing the quality factors of the images from the subject renders the following observations:

- Visually, the enrollment image has “poor quality” in terms of the exposed iris area, occluded iris with eyelash and low contrast between eyelash, pupil and iris.
- Visually, the contrast between the iris and the pupil is substantially low even with NIR illumination for this subject across all TFs, giving a salient notion of a “dark iris”.
- The average overall image quality of the rejected images of the gallery images (GQS) is 26.25 and the rejected probe images (PQS) is 29.62 on a range of 0 to 100. The average usable iris area of the 4 enrollment images is 59.75% and the average pupil to iris contrast is 52.7. All 3 quality measures - overall image quality, usable iris area and pupil to iris contrast of the rejected images falls in the lowest quartile of the distribution.

Since the MS across all TFs for this subject are substantially low, we are inclined to believe that this is a case of poor enrollment. To further explore this, we set aside the images from the original enrollment session and considered the subsequent collection (the subject’s second session) as the enrollment and match it against all subsequent collections. All mated pairs match. Thus we conclude that the rejections observed in this case is due to poor image quality of the enrolled image and is not affected by time-frame.

F. FMR and FNMR

This section provides a more detailed analysis of ROC, False Match Rate (FMR) and False Non-Match Rate (FNMR) for our dataset. All reports are for RI. LI shows a similar pattern and is thus not included here. The ROC curve is shown in Figure 12. Table VI compiles the FMR and FNMR over the 36 month TF at different threshold values. VeriEye claimed a FAR of 0.1% at threshold value 36. At that threshold, for our

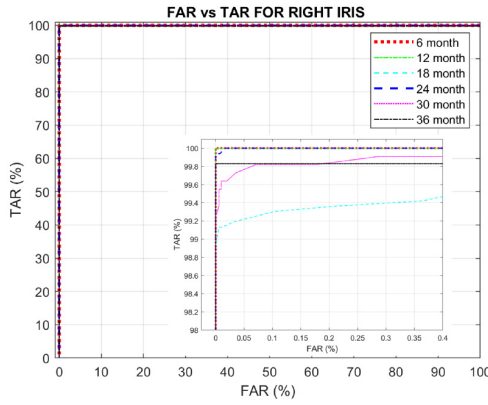


Fig. 12. ROC curves for different TFs; Inset image is a magnified representation.

TABLE VI
FMR AND FNMR IN % AT FIXED THRESHOLDS

Time Frame	Threshold 36		Threshold 50		Threshold 100	
	FMR	FNMR	FMR	FNMR	FMR	FNMR
6 mos.	0.05	0	0	0	0	0.68
12 mos.	0.07	0	0	0	0	0.47
18 mos.	0.03	0.81	0	1.16	0	1.4
24 mos.	0.06	0	0.00	0.06	0	1.4
30 mos.	0.07	0.18	0.01	0.63	0	1.8
36 mos.	0.08	0.17	0.00	0.17	0	1.4

dataset the FMR varies between 0.03% to 0.08% at different TFs.

VI. DISCUSSION, LIMITATIONS, AND FUTURE WORK

Impact on iris biometric performance due to increasing time difference between gallery and probe of iris image in children is the focus of this work. Addressing this, iris images from 209 subjects in the age group of 3 to 14 years were collected in 7 sessions, spaced approximately six months over 3 years from the same subjects and were analyzed. In addition to the effect of time window between collections on match scores (MS), variability factors like dilation, dilation difference between enrollment and probe image, and enrollment age have been taken into consideration for analysis. The performance is assessed with false non-match rate (FNMR) and Linear Mixed Effects Modelling (LMER).

Statistical conclusions are drawn from the LMER (refer Section III) which modelled the match score (MS) variability as a function of increasing time difference between gallery and probe of iris image (TF) in presence of inter-subject and intra subject variability factors of enrollment age (EA), dilation (PD), and dilation difference between enrollment and probe (ΔD). The model considers that the MS varies as a linear function of the predictor variables. However, non-linear impact of the predictors are taken into account by adding second order terms to the variability factors. With respect to the design of the model, a next step might be to consider other models to include additional order terms to fit this data. As Pinheiro and Bates [44] mentions “By increasing the order of a polynomial model, one can get increasingly accurate approximations to the true, usually nonlinear, regression function,

within the observed range of the data”. Additionally, non-linear models which also consider underlying mechanisms like asymptotes and monotonicity producing the data, while modelling response variables, could also be considered as future work. The same model is implemented for three groups based on participation - G1, G2 and G3. All groups render similar conclusions. The results can be summarized as: (i) a slight decay in MS (approx: 16.8/year and avg MS is 399) with increased TF in a three year period for the enrollment age of 4 to 14 years; this small decay has no practical impact on performance of iris recognition; (ii) strong negative correlation between ΔD and MS; (iii) no significant effect of PD on MS; (iv) linear significant effect of EA on MS ($p < 0.01$). Even with EA rounded off to the nearest year (age calculated from birth year), we see that EA is significant and thus believe important to consider it as a predictor in the model.

In addition to aging, i.e., increased time between enrollment and probe, studying the impact of age on MS is important to understand any underlying effect. We did a preliminary statistical assessment of the difference in mean MS between different age groups - 4 to 6 years, 7 to 8 years and 9 to 11 years. t-test statistics concluded that the mean MS of age group of 9 to 11 years is significantly ($p < 0.01$) higher than other age groups. Additionally, we employed the same LMER model for 3 age based groups. The conclusions in all age based groups are mostly similar to the conclusions from LMER implementation on the participation based groups with few exceptions: (i) For 4 to 6 years age group, the time difference has statistically insignificant impact on MS and ΔD has linear impact on MS; (ii) Enrollment age is not a statistically significant predictor of MS for smaller fragmented age groups. Addressing age based issues is challenging due to limited data for each age group. The small, non-uniform count of subjects in each age for the age spread of 4-11 years from approximately 209 subjects poses a challenge to make conclusions. An age focused data set needs to be created to meet this challenge.

Dilation and dilation constancy (DC) between enrollment and probe is evaluated. DC varied between 0.55 and 1 in our dataset. Low DC necessarily translates to low MS; for high DC, MS varies from low to high (refer Figure 7), where low MS are likely due to other factors. The DC of the rejected images are substantially high (all above 0.8). Dilation difference between two mated pairs of images has a major contribution to errors in recognition performance. However, no case of rejection is recorded due to low DC in our database. In our study we note that the MS is weakly correlated to DC with Pearson’s coefficient of 0.28. Dilation and DC impact the iris recognition on the MS independently. Having very high/very low dilation in both enrollment and probe leads to a high DC. Very high dilation in both gallery and probe, though leads to a high DC, leaves a very low available iris area for analysis. A very low dilation may lead to segmentation error. We observe that both dilation and DC contribute to the MS with substantial effect. TD and EA have no statistical impact on DC. Dilation is a crucial factor and has multiple causes including age, illumination, medical causes and environmental factors. Of the studies reviewed in literature in Section I-A, the closest to the age range of our study, were by Adler in [2] which mentioned

dilation reaching its maximum in the childhood and adolescence. Thus, a constructive physical relation between age and dilation remains unavailable. A comprehensive understanding of the relationship between age and aging on dilation in children based on our dataset was performed [28]. We were unable to acquire exact same dilation at each session from the same subject. However, even with this knowledge we are unable to have controlled constant illumination due to our dependency on the school setup who permits us to use their facility for collections. Measures including closing blinds and giving subjects time to adapt to the room, were taken to minimize the effect and variation of illumination and environmental factors. The Iris Guard camera [36] used for data collection, has a blinking white LED light with the purpose to provide external stimuli to accommodate pupil to an uniform dilation for each subject across sessions. Technology has been adapted by some software to accommodate dilation to minimize the dilation difference between collections. Statistical tests showed significant variability in dilation in collection sessions (Collection 3, 5, 6, 7) when compared to Collection 1. However, the mean dilation in different sessions remained relatively constant (exception: Collection 3 ($p < 0.01$)) when compared to Collection 1. Illumination heavily impacts dilation. Thus it is essential to account for change in dilation and area of iris available for analysis while calculating the MS.

False Non-Match Rate (FNMR) in this study ranges between 0% to 0.73%. Two subjects were falsely rejected in the three sessions (6, 18 and 30 month TF). The FNMR at 6 month is 0.54% and 0.73% at 18 and 30 months. The increase is not an indication of increased FNMR with time but a reflection of variable subject participation in different sessions. The FNMR being episodic with most sessions having 0% FNMR, the upper bound (UB) of the confidence interval for FNMR is statistically calculated with the 'Rule of 3'. The UB for FNMR in 36 month is projected at 3.3%. FNMR in this study does not indicate aging effect on the iris recognition performance in children in the age group of 4 to 11 years at enrollment for a period of three years. Each individual case of rejection is studied in detail to understand the underlying cause as detailed in Section V-E. All cases of false non-matches were accounted for by factors encompassing poor enrollment and angle of rotation of the captured image w.r.t. to the camera. This observation goes to the heart of template aging questions. *Even though changes in FNMR are seen, is this aging?* Indeed, false non-matches do not necessarily mean an irreversible change in the iris itself. In this study we verified that all rejected subjects returned to above threshold match scores, and identified the root causes for the rejections.

This work opens some major areas for future work based on the challenges and limitations faced. Further quality assessment is needed to understand the errors. Additional quality assessment of the iris images during capture would be a step towards accurate assessment of the data. Increased FNMR are primarily caused by poor quality images. Factors may include partial occlusion of the area of interest, motion blur, unfocused image, high dilation or other quality issues. Standard Image Quality for iris is set in the report ISO/IEC 19794-6 by ISO (International Organization for Standardization) and IEC

(International Electrotechnical Commission) [8]. We plan to study quality in more detail in future study.

The biggest challenge in research is the limited availability of data, especially in the fields involving children. The dataset used in the study is an in-house dataset, which is part of an ongoing study. To the best of our knowledge this is the only longitudinal database in the focus age group of 4 to 14 years directed to study biometrics in children. Presently the study has enrolled 239 subjects. We believe analyzing longitudinal data from 209 subjects has statistical significance as we have reported in this article. High correlation between images from the same time-frame is noted due to the collection protocol. All images were collected within seconds of each other based on the internal setup of the sensor for multiple image collection. The issue was identified and the protocol was modified from the 7th session. Multiple images are collected in 2 sets with approximately one minute in between. LI and RI images are captured sequentially with a short time delay in between, the only factor that might influence any difference between LI and RI analysis.

Any data that holds identifiable information of an individual is sensitive; more so when it involves children. A two tiered privacy protocol is adapted to protect the identity of the subjects. Due to privacy reasons we do not share any images in this article unless absolutely needed (falsely rejected image samples and image removed from the database upon manual cleaning) to convey our message. For the images that are included, the iris patterns are obscured. Protecting the privacy of participants in research study is the responsibility of the researchers. We are aware of the lack of children biometric data in the scientific community and the importance of data in advancing research related to child biometrics. The scope of such research has the potential for world-wide applications. We are sharing [45] our dataset through BEAT [46] platform for algorithm testing for research purposes. Protecting the privacy of the data and the individuals, the platform will restrict public access to the data, while giving feedback on the statistics of the performance of the algorithms. In addition to BEAT, we are directly sharing [45] the VeriEye match scores, dilation information and quality scores associated with EA and TF that are used in all of the analyses in this article. We are also sharing the data via BEAT and providing VeriEye quality scores associated with each image/pair of images.

Further scrutiny on the causes of failure to acquire (FTA) could assist hardware development and application of biometrics for children. FTA is a concern in this study. We noted approximately 25 subjects failed to be captured in one or more sessions. To improve upon the issue, we incorporated an assistive aid from 6th session in the form of a flexible articulating arm for the iris camera mount with a stabilizer handle allowing the collector to adjust the optimal distance required by the camera from the target (refer Figure 1) and thus essentially eliminating FTA in the 6th session. We also included a second binocular sensor to support the data collection. However, those data are not part of this study. On average, biometrics are presented to the sensor thrice before concluding on FTA. Pushing too hard to collect data from children from whom we failed to acquire, would be harsh, keeping in mind the younger

age of the children, which may induce discomfort and possible withdrawal from voluntary participation. Our protocol restricts aggressive data collection from children. The hardware available and used in this study are primarily designed for adults with inbuilt quality assessment setup for enrollment with no manual control. We believe if the systems were designed specifically for children, FTA would be improved.

Different studies employ different algorithms to extract features. Commercial algorithms are black-boxes with no public information on the techniques used for image to feature domain transformation, what (if any) methods are adopted to address different quality deformations (like dilation, rotation etc.) or matching techniques. Environmental factors may vary between sessions, like illumination and weather, affecting the captured image. Induced factors like medication, can also affect iris. Thus, the features under investigation are an essential factor in concluding the effect of aging on iris biometric. This reminds us of the need for further research on invariant features for iris recognition. In this study we have investigated the impact of difference in time between enrollment and probe on the iris recognition performance in children under a semi-controlled environment using a commercial software VeriEye [37].

VII. CONCLUSION

We conclude by summarizing the answers to the questions, that this work majorly focused on -

Does aging change the iris structure to a point that impacts the use of iris for biometric recognition in children? From the point of view of biometric applications, this study indicates the viability of the use of iris as a biometric modality in a 3 year TF on the population with the enrollment age group of 4 to 11 years when the quality of iris image is maintained and FTA are addressed. Statistical analysis demonstrated a minor decay in match score over time, i.e., a very small aging effect, much smaller than other typical sources of variation; the impact is practically insignificant. The statistical analysis accounts only for a time-frame of 3 years. The effect on match score is negligible in the 3-year TF of this study.

If growth in children impacts iris recognition performance, is there an age at which these impacts are no longer seen? Our study concluded that iris recognition is effective in children for the age group studied, i.e., between 4 to 11 years. We noted no impact in the iris recognition performance based on age. A positive correlation between enrollment age and match score is indicated in the statistical model. This is expected considering more control and stability with age when presenting the biometric to the system, as well as experience with usage over years. The current study cannot set the minimal age below which iris recognition is ineffective due to three major factors. First, ages below four which were not included in our current dataset needs to be studied. Second, our conclusion excludes consideration of causes of Failure to Acquire (FTA). Thus, FTA needs to be studied. Third, a larger bin size at each age is needed to make any conclusions about a specific age group.

Finally, our conclusions are based on research bound by age between 4 yrs to 14 years over a span of three years, and does not translate to life-time conclusions. Some answers are limited by the availability of required research data necessary for analysis. We plan to continue appending our dataset and report the analysis in future towards the unaddressed queries.

ACKNOWLEDGMENT

The authors extend their gratitude to the Potsdam Elementary and Middle School administration, staff, students and the parents of the participants for supporting their research and the greater goal of scientific contribution to society. This project would not have been successful without the efforts and the valuable time put in by all data collectors including David Yambay, Ganghee Jang, Keivan Bahmani, Sandip Purnapatra and Richard Plesh.

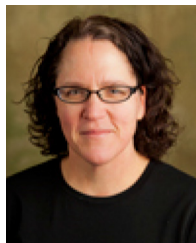
REFERENCES

- [1] A. K. Jain, B. Klare, and A. Ross, "Guidelines for best practices in biometrics research," in *Proc. Int. Conf. Biometrics (ICB)*, Phuket, Thailand, 2015, pp. 541–545.
- [2] F. H. Adler, *Physiology of the Eye: Clinical Application*. St. Louis, MO, USA: Mosby, 1965.
- [3] L. Ma, T. Tan, Y. Wang, and D. Zhang, "Efficient iris recognition by characterizing key local variations," *IEEE Trans. Image Process.*, vol. 13, pp. 739–750, 2004.
- [4] J. G. Daugman, "Biometric personal identification system based on iris analysis," U.S. Patent 5 291 560, Mar. 1994.
- [5] J. G. Daugman, "High confidence visual recognition of persons by a test of statistical independence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1148–1161, Nov. 1993.
- [6] R. P. Wildes, "Iris recognition: An emerging biometric technology," *Proc. IEEE*, vol. 85, no. 9, pp. 1348–1363, Sep. 1997.
- [7] L. Flom and A. Safir, "Iris recognition system," U.S. Patent 4 641 349, Feb. 1987.
- [8] *Information Technology—Biometric Sample Quality—Part 6: Iris Image Data*, Standard ISO/IEC 29794-6:2015, 2015.
- [9] P. J. Grother, J. R. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov, "IREX VI—Temporal stability of iris recognition accuracy," NIST, Gaithersburg, MD, USA, Rep. 7948, 2013.
- [10] A. Bertillon and R. W. McClaughry, *Signaletic Instructions Including the Theory and Practice of Anthropometrical Identification*. New York, NY, USA: Werner Company, 1896.
- [11] L. Flom and A. Safir, *U.S. Patent# 4 641 349*, U.S. Government Printing Office, Washington, DC, USA, 1987.
- [12] P. Tome-Gonzalez, F. Alonso-Fernandez, and J. Ortega-Garcia, "On the effects of time variability in iris recognition," in *Proc. 2nd IEEE Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Arlington, VA, USA, 2008, pp. 411–416.
- [13] S. E. Baker, K. W. Bowyer, and P. J. Flynn, "Empirical evidence for correct iris match score degradation with increased time-lapse between gallery and probe matches," in *Proc. Int. Conf. Biometrics*, 2009, pp. 1170–1179.
- [14] S. E. Baker, K. W. Bowyer, P. J. Flynn, and P. J. Phillips, "Template aging in iris biometrics," in *Handbook of Iris Recognition*. London, U.K.: Springer, 2013, pp. 205–218.
- [15] S. P. Fenker and K. W. Bowyer, "Analysis of template aging in iris biometrics," in *Proc. IEEE Comput. Sco. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Providence, RI, USA, 2012, pp. 45–51.
- [16] D. M. Rankin, B. W. Scotney, P. J. Morrow, and B. K. Pierscionek, "Iris recognition failure over time: The effects of texture," *Pattern Recognit.*, vol. 45, no. 1, pp. 145–150, 2012.
- [17] N. Sazonova *et al.*, "A study on quality-adjusted impact of time lapse on iris recognition," in *Proc. Sens. Technol. Global Health Military Med. Disaster Response Environ. Monitor. II Biometric Technol. Hum. Identif. IX*, vol. 8371, 2012, p. 83711W.
- [18] A. Czajka, "Template ageing in iris recognition," in *Proc. Int. Conf. Bio-Inspired Syst. Signal Process. (BioSignals)*, 2013, pp. 70–78.
- [19] D. O. Gorodnichy and M. P. Chumakov, "Analysis of the effect of ageing, age, and other factors on iris recognition performance using nexus scores dataset," *IET Biometrics*, vol. 8, no. 1, pp. 29–39, Jan. 2019.

- [20] J. Matey, *Private Communication*, NIST, Gaithersburg, MD, USA, Oct. 2019.
- [21] K. W. Bowyer and E. Ortiz, "Critical examination of the irex VI results," *IET Biometrics*, vol. 4, no. 4, pp. 192–199, Nov. 2015.
- [22] P. Grother, J. R. Matey, and G. W. Quinn, "IREX VI: Mixed-effects longitudinal models for iris ageing: Response to bowyer and ortiz," *IET Biometrics*, vol. 4, no. 4, pp. 200–205, Dec. 2015.
- [23] E. Ortiz and K. W. Bowyer, "Exploratory analysis of an operational iris recognition dataset from a CBSA border-crossing application," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Boston, MA, USA, 2015, pp. 34–41.
- [24] E. Ortiz and K. W. Bowyer, "Pitfalls in studying 'big data' from operational scenarios," in *Proc. IEEE 8th Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Niagara Falls, NY, USA, 2016, pp. 1–7.
- [25] J. E. Birren, R. C. Casperson, and J. Botwinick, "Age changes in pupil size," *J. Gerontol.*, vol. 5, no. 3, pp. 216–221, 1950.
- [26] B. Winn, D. Whitaker, D. B. Elliott, and N. J. Phillips, "Factors affecting light-adapted pupil size in normal human subjects," *Invest. Ophthalmol. Vis. Sci.*, vol. 35, no. 3, pp. 1132–1137, 1994.
- [27] E. Ortiz, K. W. Bowyer, and P. J. Flynn, "A linear regression analysis of the effects of age related pupil dilation change in iris biometrics," in *Proc. IEEE 6th Int. Conf. Biometrics Theory Appl. Syst. (BTAS)*, Arlington, VA, USA, 2013, pp. 1–6.
- [28] P. Das, L. Holsopple, M. Schuckers, and S. Schuckers, "Analysis of dilation in children and its impact on iris recognition," in *Proc. Forthcoming Int. Joint Conf. Biometrics (IJCB)*, Houston, TX, USA, 2020, pp. 1–9.
- [29] *Office of Human Research Protection*. Accessed: Oct. 17, 2019. [Online]. Available: <https://www.hhs.gov/ohrp/regulations-and-policy/guidance/faq/children-research/index.html>
- [30] P. Basak, S. De, M. Agarwal, A. Malhotra, M. Vatsa, and R. Singh, "Multimodal biometric recognition for toddlers and pre-school children," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Denver, CO, USA, 2017, pp. 627–633.
- [31] *Children Biometric Multimodal Database (CMBD)*. Accessed: Jul. 30, 2020. [Online]. Available: <http://iab-rubric.org/resources.html>
- [32] N. Nelufule, A. de Kock, G. Mabuza-Hocquet, and Y. Moolla, "Image quality assessment for iris biometrics for minors," in *Proc. Conf. Inf. Commun. Technol. Soc. (ICTAS)*, Durban, South Africa, 2019, pp. 1–6.
- [33] *Unique Identification Authority of India*. Accessed: Oct. 16, 2018. [Online]. Available: <https://uidai.gov.in/>
- [34] D. J. W. Campbell, *Impact of Demographic Factors on Performance of Biometric Systems*, Defence Res. Develop. Canada, Ottawa, ON, Canada, 2018. [Online]. Available: http://cradpdf.drdc-rddc.gc.ca/PDFS/unc308/p806591_A1b.pdf
- [35] M. Johnson, D. Yambay, D. Rissacher, L. Holsopple, and S. Schuckers, "A longitudinal study of iris recognition in children," in *Proc. IEEE 4th Int. Conf. Identity Security Behav. Anal. (ISBA)*, Singapore, 2018, pp. 1–7.
- [36] *IrisGuard*. Accessed: Nov. 14, 2019. [Online]. Available: <https://www.irisguard.com>
- [37] *MegaMatcher 10.0, VeriFinger 10.0, VeriLook 10.0, VeriEye 10.0 and VeriSpeak 10.0 SDK: Developer's Guide*, Neurotechnology, Vilnius, Lithuania, 2017.
- [38] *IREX 10: Identification Track*. Accessed: Jan. 21, 2021. [Online]. Available: <https://pages.nist.gov/IREX10/>
- [39] J. D. Singer and J. B. Willett, *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford, U.K.: Oxford Univ. Press, 2003.
- [40] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," 2014. [Online]. Available: [arXiv:1406.5823](https://arxiv.org/abs/1406.5823).
- [41] X. Wang, J. Dong, M. Tang, X. Wang, H. Wang, and S. Zhang, "Effect of pupil dilation on biometric measurements and intraocular lens power calculations in schoolchildren," *PLoS ONE*, vol. 13, no. 9, 2018, Art. no. e0203677.
- [42] *Information technology—Vocabulary—Part 37: Biometrics*, Standard ISO/IEC 2382-37:2017(en), 2017.
- [43] *Information Technology—Biometric Performance Testing and Reporting—Part 1: Principles and Framework*, Standard ISO/IEC 19795-1:2006, 2006.
- [44] J. C. Pinheiro and D. M. Bates, *Mixed-Effects Models in S and S-PLUS*. New York, NY, USA: Springer, 2006.
- [45] *Clarkson Longitudinal Iris in Children Database (CLIC)—Matching and Quality Scores*. Accessed: Dec. 7, 2020. [Online]. Available: <https://citer.clarkson.edu/clarkson-longitudinal-iris-in-children-database-clic-matching-and-quality-scores/>
- [46] A. Anjos, L. El-Shafey, and S. Marcel, "BEAT: An open-source web-based open-science platform," 2017. [Online]. Available: [arXiv:1704.02319](https://arxiv.org/abs/1704.02319).



Priyanka Das (Graduate Student Member, IEEE) received the bachelor's and M.E. degrees in biomedical engineering from Jadavpur University, India, in 2014 and 2016, respectively. She is currently pursuing the Ph.D. degree in iris biometrics with the Department of Electrical and Computer Engineering, Clarkson University, where she is a Research Assistant.



Laura Holsopple is the Managing Director of the Center for Identification Technology Research (CITeR), an NSF funded Industry University Cooperative Research Center (IUCRC), Clarkson University, focusing on the science of biometrics. Drawing on her over 15 years from Industry Manufacturing and Development, she has worked with CITeR/Clarkson University over the last ten years to promote the Center's research and bring continual improvement to the CITeR community and the affiliate members served by membership in the

center.



Dan Rissacher received the M.S. degree from the Georgia Institute of Technology and the Ph.D. degree from Clarkson University, with a dissertation demonstrating automated detection of human pain using neural networks on EEG data, where he was an Assistant Research Professor, when he founded the study for biometric aging in children. He is currently an Independent Researcher with his company CRIA Corporation, a Pilot for American Airlines, and a Cyber Warfare Operator for the Vermont Air National Guard. Since then, his research has focused

on pattern recognition, data analytics, and machine learning.



Michael Schuckers received the Doctorate degree in statistics from Iowa State University, his academic work has focused on developing statistical methodology for biometric authentication and applications in sports, particularly ice hockey. He is the Charles A. Dana Professor of Statistics with St. Lawrence University, Canton, NY, USA, where he also serves as the Director of the Peterson Quantitative Resource Center. He has also consulted for professional teams in Major League Baseball and the National Hockey League among other professional sports organizations in his role as a Co-Founder of Statistical Sports Consulting, LLC.



Stephanie Schuckers (Senior Member, IEEE) received the Doctoral degree in electrical engineering from the University of Michigan. She is the Paynter-Krigman Endowed Professor of Engineering Science with the Department of Electrical and Computer Engineering, Clarkson University, and serves as the Director of the Center for Identification Technology Research (CITeR), a National Science Foundation Industry/University Cooperative Research Center. Her work is funded from various sources, including the National Science Foundation, the Department of Homeland Security, and private industry among others. She has started her own business, testified for the U.S. Congress, and has over 100 other academic publications. Her research focuses on processing and interpreting signals which arise from the human body.