Longitudinal Study of Automatic Face Recognition

Lacey Best-Rowden, *Student Member, IEEE,* and Anil K. Jain, *Fellow IEEE*

***Abstract*—With the deployment of automatic face recognition systems for many large-scale applications, it is crucial that we gain a thorough understanding of how facial aging affects the recognition performance, particularly across a large population. Because aging is a complex process involving genetic and environmental factors, some faces "age well" while the appearance of others can change drastically over time. This heterogeneity (between-subject variability) suggests the need for a subject-specific aging analysis. In this paper, we conduct such an analysis using two different longitudinal face databases of operational mugshots of repeat criminal offenders. Each of the 5,636 and 18,007 subjects in the two databases has at least four face images acquired over a minimum of five years. By fitting mixed-effects statistical models to genuine similarity scores from state-of-the-art commercial-off-the-shelf (COTS) face matcher, we quantify (i) the population average rate of change in genuine scores with respect to the elapsed time between two face images, and (ii) how closely the subject-specific rates of change follow the population average. Longitudinal analysis shows that despite decreasing genuine scores over time, as expected, the average subject can still be correctly verified at a false accept rate (FAR) of 0.01% across all 8 and 16 years of maximum elapsed time in the two face databases. We also investigate the effects of several other covariates (age, sex, race, face quality). We find that differences in subject age at enrollment are significant but marginal, females generally have lower scores than males, and interpupillary distance and a measure of frontalness explain a significant amount of variation in genuine scores that is not accounted for by elapsed time alone.**

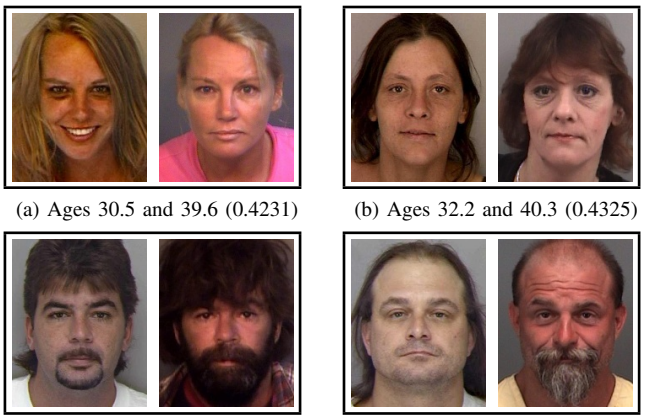
***Index Terms*—face recognition, longitudinal study, mixed-effects, facial aging.**

I. Introduction

FACIAL recognition technology has rapidly matured over the last two decades, to the point where it is now utilized in many commercial and law enforcement applications for person recognition (*e.g.* mobile face unlock and de-duplication of driver's licenses). Automatic face recognition systems operating on face images acquired in controlled conditions, such as mugshots or driver's license photos, have achieved high accuracies (99% TAR at 0.1% FAR) in large-scale evaluations conducted by the National Institute of Standards and Technology (NIST) [1].

Technological advancements in automatic face recognition have progressively tackled challenges caused by variations in facial pose, illumination, and expression (collectively called PIE variations). Current efforts (*e.g.* [2], [3]) are breaking ground on robustness to "faces in the wild" to account for PIE, occlusion, and partial face images (*e.g.* images posted on the web). Comparatively, aging variations (*i.e.* large time lapse between pairs of images being compared) have received considerably less attention in the face recognition community.

L. Best-Rowden and Anil K. Jain are with the Department of Computer Science and Engineering, Michigan State University, East Lansing, MI, 48824 USA e-mail: {bestrow1, jain}@msu.edu).



(a) Ages 30.5 and 39.6 (0.4231) (b) Ages 32.2 and 40.3 (0.4325)

(c) Ages 29.5 and 38.3 (0.4984) (d) Ages 39.2 and 48.6 (0.5002)

Fig. 1: Face image pairs of four different subjects from the PCSO_LS mugshot database which are age-separated by eight to ten years. Similarity scores from a state-of-the-art face matcher (COTS-A) for each pair are shown in parentheses (score range is [0.0, 1.0]). The thresholds for COTS-A scores at 0.01% and 0.1% FAR are 0.5331 and 0.4542, respectively. Hence, all of these genuine pairs would be falsely rejected at 0.01% FAR, while the two female subjects would also be rejected at 0.1% FAR.

Published studies on facial aging in the context of automatic face recognition have primarily employed *cross-sectional* techniques where a population of individuals who differ in age are analyzed according to differences between age groups [1], [4], [5], [6], [7]. However, cross-sectional analysis cannot adequately explore age-related effects on face recognition because of the assumption that the individuals were sampled at a single point in time; past and future measurements are not considered so the trends of individuals over time are not analyzed. Hypotheses about facial aging are, instead, *longitudinal* by nature and require tracking the comparison scores of the same individuals typically over several years.

While longitudinal studies for automatic iris recognition [8] and fingerprint recognition [9] have been published, to our knowledge, no large-scale longitudinal study of automatic face recognition performance has been reported in the literature. We aim to fill this gap by fitting multi-level statistical models to a longitudinal face dataset to address the following question: *How robust are state-of-the-art automatic face recognition systems to facial aging?*

Aging effects on the performance of automatic face recognition systems are of more than mere theoretical concern. Because the appearance of the face changes throughout a person's life, most identity documents containing face images expire after a designated period of time; U.S. passports are only valid for 5 years for minors and 10 years for adults,

while U.S. driver's licenses typically require renewal every 5 years. Validity periods of such identity documents may be too long if these photos are to be used with state-of-the-art face matching systems. Figure 1 shows that elapsed times of eight to ten years between two face images can cause false non-match errors. Additionally, to our knowledge, ensuring that a new photo has been submitted for renewal is not verified, especially for renewals by mail or online. Studying how the actual comparison scores change over time is important for understanding the implications of operating with a global threshold[1] (*e.g.* de-duplication and other open-set scenarios) on face recognition performance degradation due to aging.

Aging of the human body naturally causes changes in facial appearance over time. During years of adolescence, facial changes are predominantly due to the maturation of the shape of the head; whereas in later stages of life, an adult face may experience additional changes affecting skin texture and elasticity. However, a teenager can also experience skin texture changes due to acne scarring and/or growth of facial hair, and the shape of an older person's face can be severely altered by rapid decline in health (*e.g.* weight loss or gain). In addition to anatomical factors, environmental and/or lifestyle factors also have a significant impact on facial appearance over time. Smoking, sun exposure, stress levels, and drug abuse, for examples, can drastically alter a person's face, sometimes over just a short period of time. Due to the cumulative effects of both biological and environmental factors, facial aging is a complex process that affects each individual differently. In this paper, we only look at the aging issue as reflected in the visual appearance of the subject's face in the photograph.

A. Contributions

We conduct a longitudinal analysis of the performance of a state-of-the-art COTS face recognition system on two longitudinal face image databases consisting of mugshots of repeat criminal offenders from two different law enforcement agencies (see Table I). The COTS matchers used here are among the top-ranked performers in the FRVT 2013 face recognition evaluation [1]. Mixed-effects statistical models, which are appropriate for longitudinal data, are used to analyze the variation in genuine comparison scores over time from the COTS matchers. The contributions of this paper can be summarized as follows:

1) Conduct a multilevel statistical analysis of the longitudinal effects of facial aging on automatic face recognition for the two largest longitudinal databases studied to date.
2) Quantification of the average rate of change in face comparison scores per one year increase in the average age of a subject and one year increase in time since enrollment.
3) Quantification of the variance in subject-specific temporal trends in genuine face comparison scores; a representation of how genuine comparison scores for each individual change over time for a large population of subjects.

[1]A biometric system operating with a global threshold applies the same threshold to all subjects and all comparisons.

TABLE I: Face Aging Databases

Database	Num. Subjects	Num. Images	Num. Images per Subject	Age Range (years)
MORPH-II [11][a]	13,000	55,134	2–53 (avg. 4)	16–77 (avg. 42)
MORPH-II commercial [11][b]	20,569	78,207	1–76 (avg. 4)	15–77 (avg. 33)
FG-NET [12]	82	1,002	6–18 (avg. 12)	0–69 (avg. 16)
LEO_LS[c]	5,636	31,852	4–20 (avg. 6)	12–69 (avg. 31)
PCSO_LS[c]	18,007	147,784	5–60 (avg. 8)	18–83 (avg. 35)

[a]MORPH-II is "Album 2" of the MORPH database. There is also an earlier released version ("Album 1") that contains only 612 subjects and less than 2,000 images.
[b]This largest version of MORPH only has 317 subjects with at least 5 images acquired over at least 5 years.
[c]The longitudinal face databases used in this study (details in Sec. III-A).

This work extends, as well as refines, our previous longitudinal analysis of automatic face recognition first published in [10]. The primary differences are stated below:

1) Analysis of both elapsed time and absolute age ([10] only studied elapsed time).
2) Genuine scores are now computed assuming that the youngest image of each subject is enrolled in a gallery; if a subject has n images total, we compute $n-1$ comparison scores, whereas [10] computed all $\binom{n}{2}$ genuine scores. Although the total number of genuine scores being analyzed is lower than in [10], comparisons are made to a fixed point in time which simplifies the complex correlation structure that is present for all pairwise comparisons.
3) Analysis of an additional longitudinal database (LEO_LS) of face images from a different law enforcement agency. LEO_LS database has different characteristics than the PCSO_LS database studied in [10] (*e.g.* shorter elapsed times), and comparison scores are obtained from a different COTS matcher. Still, the longitudinal analysis reveals similar trends for both databases.

The remainder of this paper is organized as follows. Section II briefly highlights related work on facial aging as it pertains to automatic face recognition. Section III details the two longitudinal face databases used in this study and computation of comparison scores. Section IV explains the methodology behind the mixed-effects statistical models used for longitudinal analysis. Section V gives results from fitting mixed-effects models to genuine comparison scores from both the PCSO_LS and LEO_LS face databases. Section VI summarizes our observations and findings about the performance of automatic face recognition over time.

II. Related Work

Almost all of the published studies that invesitgate the effects of facial aging on automatic face recognition performance adopt the following approach: (i) divide the database (face pairs) into partitions depending on age group or time lapse, (ii) report summary performance measures (*e.g.* TAR at fixed FAR) for each partition independently, and then (iii) draw

TABLE II: Summary of Related Work

Study	Database	Age or Elapsed Time Partitions	Findings
Klare and Jain [4]	PCSO (200,000 mugshots, 64,000 subjects)	0-1, 1-5, 5-10, 10+ years	96.3%, 94.3%, 88.6%, and 80.5% TAR at 1% FAR
Otto *et al.* [5]	MORPH	0–1, 1–5 years	97% and 95% TAR at 1% FAR
Ling *et al.* [6]	Passports (private)	4–11 years	EER degradation saturates after 4 years elapsed time.
Ling *et al.* [6]	FG-NET	0–8, 8–18, and 18+ years	Verification gets easier with increasing age.
NIST FRVT [1]	Visa images (19,972 subjects)	baby, kid, pre-teen, teen, young, parents, older	Error rates higher for younger age groups when the same threshold is used for all age groups.
Bereta *et al.* [7]	FG-NET	0–5, 6–10, 11–15, 16–20, 21–30, $>$ 30 years; 23–30, 31–40, 41–50, and $>$ 50 years	Performance of local descriptors varies across absolute age and age gap groups, but when combined with Gabor filters, most local descriptors become relatively robust to ages and age gaps.

conclusions from the differences in performance across the partitions. Such an approach has led to the following general conjectures [13]:

(i) Face recognition performance decreases as the time elapsed between two images of the same person increases (*e.g.* [4], [5], [6]).

(ii) Faces of older individuals are easier to recognize/discriminate than faces of younger individuals (*e.g.* [1], [6]).

See Table II for a summary of these studies.[2]

Partitioning of data (images or subjects) based on age group or time lapse is often arbitrary and varies from one study to another. For example, Erbilek and Fairhurst show that different age group partitionings result in different performance trends for both iris and signature modalities [16]. Furthermore, this cohort-based analysis with summary statistics cannot address whether age-related performance trends are due to changes in genuine (same subject) comparison scores, impostor (different subjects) comparison scores, or both.

Multilevel (hierarchical or mixed-effects) statistical models have been used for determining important factors (covariates) to explain the performance of face recognition systems. Beveridge *et al.* [17] apply generalized linear mixed models to verification decisions made by three algorithms in the FRGC Exp. 4 evaluation. In addition to eight levels of FAR as a covariate, they analyze gender, race, image focus, eye distances, age, and elapsed time. The limitations of this study include (i) the maximum elapsed time between face images of the same subject is less than one year, and (ii) it only involves 351 subjects. The longitudinal study on face recognition in this paper follows the general methodology of linear mixed-effects statistical models outlined in [8] for iris recognition and [9] for fingerprint recognition.

While the FG-NET [12] and MORPH[3] databases have contributed to studies on facial aging, they are not suitable for longitudinal analysis because (i) FG-NET contains only 82 subjects in total, and (ii) MORPH contains only a small number of subjects with multiple images over time (only 317 subjects have at least 5 images over at least a 5 year time span). For these reasons, we compiled two new longitudinal databases of face images, detailed in Section III-A.

[2]Studies that address age-invariant face recognition (*e.g.* [14], [15]) are not in the scope of this paper.

[3]http://www.faceaginggroup.com/morph/

(a) (b) (c)

Fig. 3: Three examples of labeling errors in the PCSO_LS face database. All pairs show different subjects who are labeled with the same subject ID number in the database.

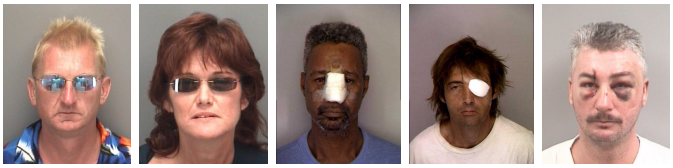


Fig. 4: Examples of facial occlusions (sunglasses, bandages, and bruises) in the PCSO_LS face database.

III. Materials

A. Longitudinal Face Databases

Operational face datasets maintained by government and law enforcement agencies can offer sources of longitudinal records of individuals of magnitudes (*e.g.* over 10 years) that are infeasible to collect in laboratory settings. These agencies routinely collect face images of the same individuals over time and have been doing so for relatively long durations, primarily for applications involving driver's licenses, visa and passport applications/renewals, frequent travelers, and multiple arrests of the same persons. While such databases often contain a large number of subjects, they may or may not contain a large number of images per subject and most are restricted to individuals typically over the age of 18 years (adult populations). For privacy reasons, it is also extremely difficult to access these longitudinal face datasets from government and law enforcement agencies.

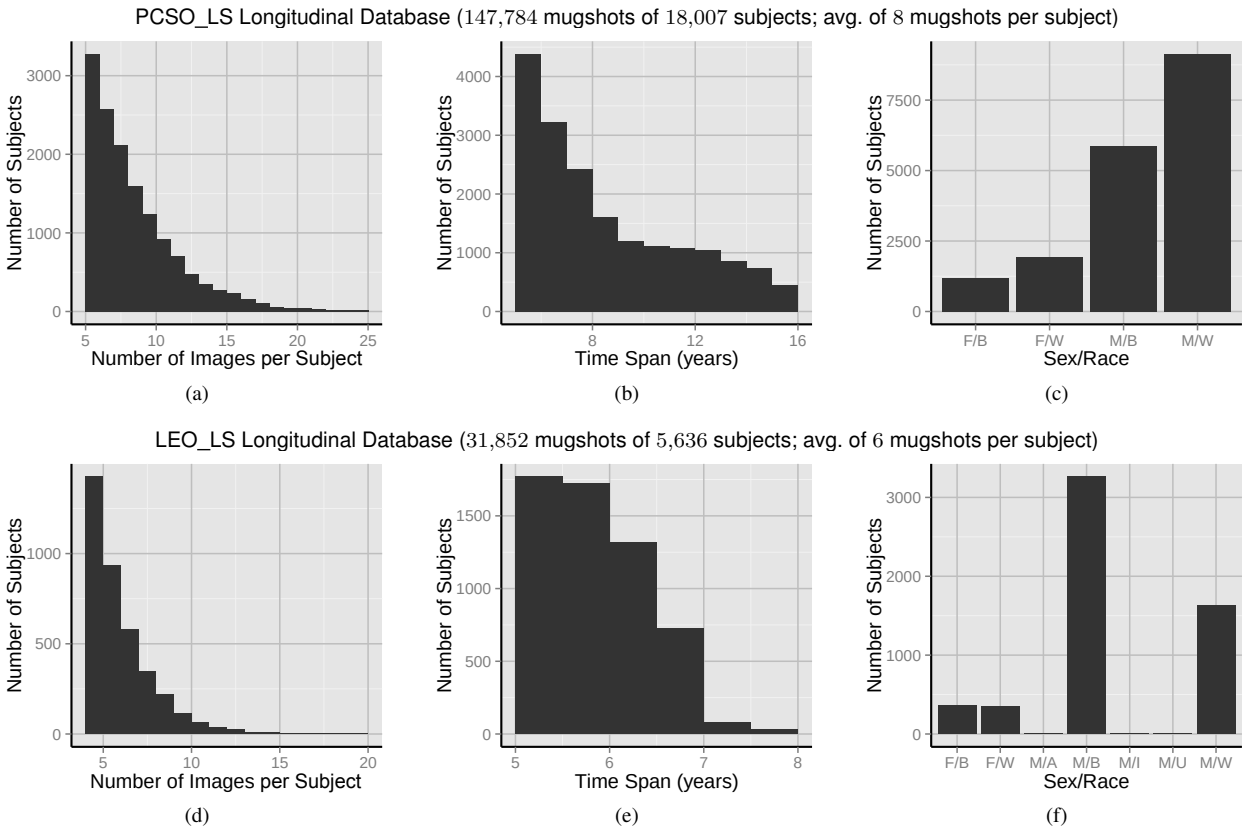


Fig. 2: Statistics of the two longitudinal face image databases used in this study: top and bottom row figures are for the PCSO_LS and LEO_LS databases, respectively. (a) and (d) Number of face images per subject, (b) and (e) the time span of each subject (*i.e.* the number of years between a subject's first and last face image acquisitions), and (c) and (f) demographic distributions of sex (male, female) and race (white, black, Asian, Indian, unknown).

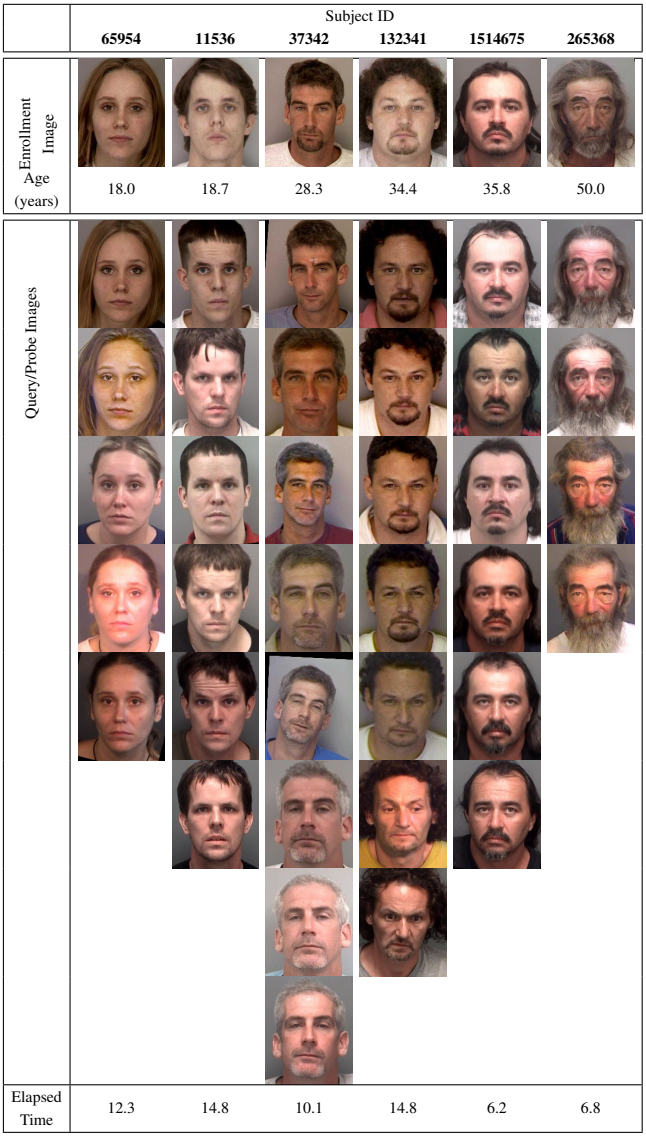
The sources of face images in our longitudinal analysis are mugshot bookings from two different law enforcement agencies. (Note that the MORPH database in Table I is also a mugshot database). While we acknowledge that lifestyle factors (*e.g.* drug and alcohol use, etc.) likely increase aging rates for this population, we have not been able to access any other longitudinal face data. We did attempt to obtain longitudinal face images from the State Department visa databases, but discovered that roughly 5% of genuine face image pairs were duplicate photo submissions. This can happen when an individual reuses the same photo for a visa renewal application; hence, age information for this database is not reliable for longitudinal study.

1) LEO_LS Longitudinal Face Database: The LEO_LS is a subset of face images from an operational dataset of over 3 million law enforcement images. The subset, LEO_LS, contains 31,852 images of 5,636 individuals, where each individual has at least 4 face images acquired over at least 5 years, with at least a one month separation between consecutive images of an individual. In order to focus on longitudinal effects of facial aging, webcam images and profile mugshot images were removed, so LEO_LS contains nearly frontal face images, most being mugshots. The LEO_LS database does contain 656 images of 369 subjects that are younger than 18 years-old; these may be juvenile[4] arrests or they could be data entry errors (it is difficult to tell based on visual examination whether an individual is actually a juvenile versus whether they are 18 or older). We only have access to the comparison scores (both genuine and impostor), so we cannot show face images from this database.

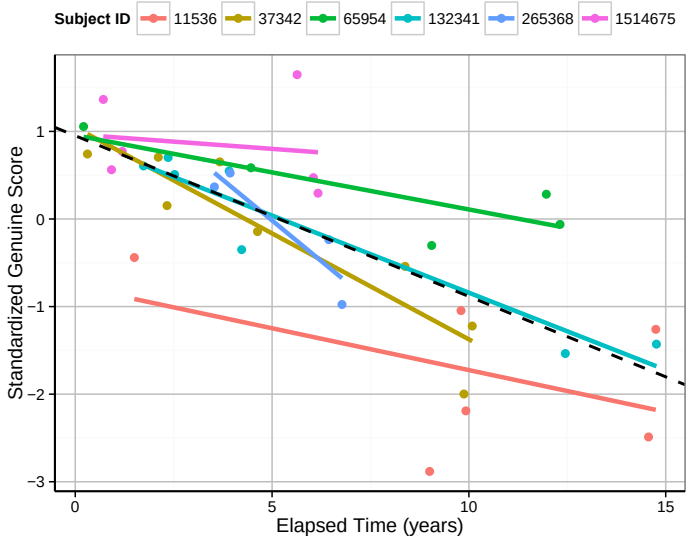
2) PCSO_LS Longitudinal Face Database: The PCSO_LS database consists of 147,784 operational mugshots of 18,007 repeat criminal offenders booked by the Pinellas County Sheriff's Office (PCSO) from 1994 to 2010. This subset of images was selected from a larger database consisting of 1.5 million images of 450,000 subjects using the following criteria. Each subject has at least 5 face images that were acquired over at least a 5 year time span, where each pair of consecutive images is time-separated by at least one month. The database statistics are shown in Fig. 2. Example face images from PCSO_LS are shown in Fig. 5. Each booking record, in addition to the face image, also includes ancillary information (*e.g.* gender, race, date of birth, date of arrest).

For both databases, we only include white and black race subjects in this study because there are too few subjects of other races to do a meaningful statistical analysis (see Fig. 2). Human labeling errors of demographic attributes, as well as subject ID, are typical of large-scale legacy databases. Identi-

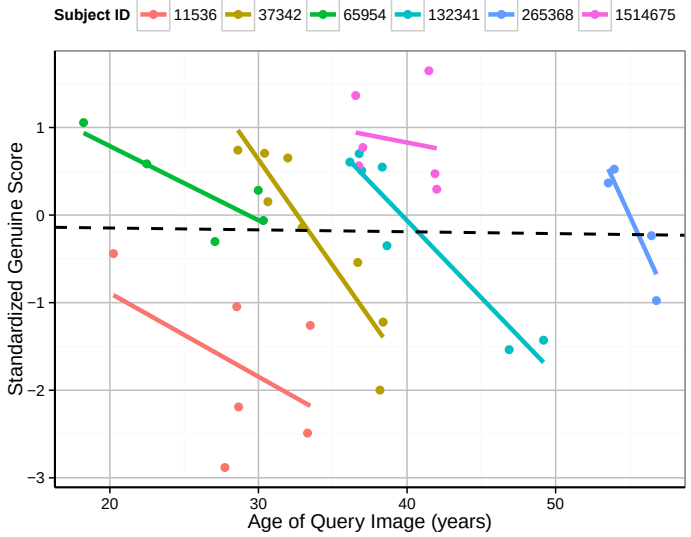
[4]In the United States, a juvenile is typically under the age of 17.



(a)



(b)



(c)

Fig. 5: (a) Face images of six example subjects from the PCSO_LS database. The enrollment face image in the top row is the youngest image of each subject; all other images are in order of increasing age. Genuine scores are computed by comparing the query/probe images of each subject to his/her enrollment image. The maximum elapsed time between a subject's enrollment image and oldest image is given in the bottom row of (a). Genuine scores and example trend lines for each subject are plotted against (b) elapsed time and (c) age of query image. The differences in intercepts and slopes of the trend lines for the six subjects are subject-specific variations (random effects). The deviations of each subject's own scores around their own trend line are the residual variations.

fying all such errors is not feasible due to the large size of both the LEO_LS and PCSO_LS databases. To ensure consistent labels within each subject's record, we determine the gender, race, and date of birth of a subject as the majority vote from the set of face images for that subject. A cursory examination of the PCSO_LS database revealed 134 subject records that contained multiple identities (see Fig. 3). These subjects were removed from our study. While the LEO_LS and PCSO_LS databases contain relatively constrained face images, some confounding factors are still present (*e.g.* sunglasses and facial injury shown in Fig. 4). We have retained such images in this study.

B. *Face Comparison Scores*

Face comparison scores (similarities) were obtained from two different commercial-off-the-shelf (COTS) face matchers, both of which were among the top performers in the FRVT 2013 [1]. We will denote the two matchers used for the PCSO_LS and LEO_LS databases as COTS-A and COTS-B, respectively.[5] Genuine comparison scores s_{ij} between the

[5]Scores for the LEO_LS database were provided by the Image group, National Institute of Standards and Technology (NIST).

enrollment and jth face images of subject i, were standardized so $y_{ij} = (s_{ij} - \mu)/\sigma$, where μ and σ are the mean and standard deviation, respectively, of all the genuine scores from all subjects. This standardized response y_{ij} is in terms of standard deviations from the mean of the genuine distribution, which allows interpretation of coefficients from regression models as quantifying the change in genuine scores as β standard deviations per year, for example.

Responses for all mixed-effects models in this study are *genuine* comparison scores. To evaluate face recognition performance, changes in genuine scores should be considered in context with respect to an impostor distribution. Hence, for both the LEO_LS and PCSO_LS databases, we compute all possible impostor scores, and then calculate thresholds at different fixed false accept rates (FARs). The threshold at, say, 0.01% FAR is used to determine when genuine scores drop below the threshold, hence, causing false rejection/negative errors.

As mentioned in Section III-A, the LEO_LS database consists of 31,852 images of 5,636 subjects, and the PCSO_LS database consists of 147,784 images of 18,007 subjects. Under the scenario where each subject's set of images are compared to his/her enrollment image, this results in a total of 26,216 and 129,773 genuine scores and 546,940,788 and 11,102,014,369 impostor scores for the LEO_LS and PCSO_LS databases, respectively.

IV. Methods

Mixed-effects models (also known as random-effects, multilevel, and hierarchical models) are widely used in various scientific disciplines for studying data that is hierarchically structured, including longitudinal data of repeated observations over time [18], [19]. In our case, face images (and comparison scores) are grouped by subject because we have repeated observations of each individual in our study. When data is structured in such a manner, responses from the same cluster/group/individual are correlated with each other and across time (for longitudinal data); hence, variation in the response (here, face comparison scores) occurs at different levels of the data hierarchy. Mixed-effects models enable analysis of these different sources of variation.

Ideally, longitudinal data collection would observe all individuals in the study following the exact same schedule over the entire duration of interest. However, longitudinal data is typically not this nice; either it is difficult (and expensive) to collect or it must be analyzed retrospectively. Even if we were to attemp to *design* a longitudinal data collection procedure, the number of cohorts, overlap between cohorts, and the frequency of measurements are all important design decisions to consider, and no "best" design exists [20]. Instead, longitudinal data is most often *time-unstructured* and *unbalanced*, meaning individuals in the study population are observed at different schedules and have different numbers of observations. For the mugshot databases, this translates to different rates of recidivism for each subject. Figure 2 shows that subjects in the LEO_LS and PCSO_LS databases have anywhere from 4 to more than 20 mugshots, and Fig. 6 shows that the age spans of the subjects are highly unstructured.

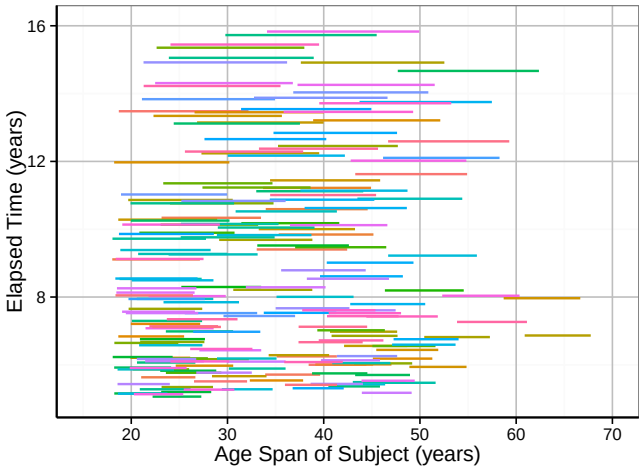


Fig. 6: Age distribution of a random subset of 200 subjects from the PCSO_LS database. Each line denotes the age span of a subject (*i.e.* age of the youngest image to the age of the oldest image of that subject), separated along the y-axis by the elapsed time for each subject (*i.e.* the length of the age span).

A. Statistical Analysis

Given n_i face images of subject i, let AGE_{ij} denote the absolute age of the ith individual in their jth face image, where $AGE_{ij} < AGE_{ik}$ for $j = 0, \ldots, n_i - 2$ and $k = j + 1, \ldots, n_i - 1$ (*i.e.* the n_i images are ordered by increasing age). To begin with, assume that the youngest image (first acquisition) of each subject is enrolled in the gallery, and let $AGE_{ie} = AGE_{i0}$ denote the age of individual i at enrollment where $AGE_{ie} < AGE_{ij}$ for $j = 1, \ldots, n_i - 1$. We can compute $m_i = n_i - 1$ genuine comparison scores by comparing every other image to the enrollment image. Hence, in this scenario, y_{ij} $(j = 1, \ldots, m_i)$ is the comparison score between the jth face image of individual i and his/her enrollment image. AGE_{ij} is the age of the jth query/probe image of subject i, so the elapsed time between enrollment and query image is $\triangle T_{ij} = AGE_{ij} - AGE_{ie}$.

When studying age-related effects on automatic face recognition performance, there are two different, albeit closely related, time-varying covariates which are of primary interest: (i) the *elapsed time* between and (ii) the *absolute ages* of the two face image acquisitions in comparison.

1) Function of Elapsed Time: The simplest notion of face recognition performance over time is a function of the elapsed time between a subject's enrollment and query face images, $f(\triangle T_{ij})$. A linear mixed-effects model with two levels (to account for subject-specific trends) and a single covariate for elapsed time can be formulated as follows. At level-1, the comparison score y_{ij} between the enrollment and jth query image of subject i can be modeled as a linear function of $\triangle T_{ij}$:

$$y_{ij} = \varphi_{0i} + \varphi_{1i} \triangle T_{ij} + \varepsilon_{ij}, \tag{1}$$

TABLE III: Mixed-Effects Model Formulations

Model	Level-1 Model	Level-2 Model: Intercept	Level-2 Model: Slope
A	$y_{ij} = \varphi_{0i} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + b_{0i}$	
BT	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
CT	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
CT	$y_{ij} = \varphi_{0i} + \varphi_{1i}AGE_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
D	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{02}AGE_{ie}^2 + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}AGE_{ie} + b_{1i}$

where the ith individual's intercept, φ_{0i}, and slope, φ_{1i}, are

$$\begin{aligned} \varphi_{0i} &= \beta_{00} + b_{0i}, \\ \varphi_{1i} &= \beta_{10} + b_{1i}. \end{aligned} \tag{2}$$

The *level-1* equation in (1) models *within-subject* longitudinal change in y_{ij} where a subject's scores can vary around his/her linear trend by ε_{ij}. The *level-2* model in (2) accounts for *between-subject* variation in comparison scores because each subject's intercept and slope parameters, φ_{0i} and φ_{1i}, respectively, are modeled as a combination of fixed and random effects. The *fixed effects*, β_{00} and β_{10}, are the grand means of the population intercepts and slopes, respectively, and define the overall *population-mean trend*, while the *random effects*, b_{0i} and b_{1i}, are subject-specific deviations from the population-mean parameters. Since each subject can have his/her own intercept and slope parameters, mixed-effects models are flexible in handling/allowing for biometric zoo [21], [22] effects (some subjects generally have higher or lower scores and subject scores change at different rates over time).

The random structure of the above two-level model includes the level-1 residuals, $\{\varepsilon_{ij}\}$, as well as the random effects, b_{0i} and b_{1i}, which can be thought of as level-2 residuals. The assumptions of these error terms are:

$$\varepsilon_{ij} \sim N(0,\sigma_\varepsilon^2) \tag{3}$$

and

$$\begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right), \tag{4}$$

where $N(.,.)$ denotes a Gaussian distribution.

Substituting the level-2 equations for subject-specific intercepts and slopes into the level-1 model in (1), the composite form of the two-level mixed-effects model is:

$$y_{ij} = \big[\beta_{00} + b_{0i}\big] + \big[\beta_{10} + b_{1i}\big]\triangle T_{ij} + \varepsilon_{ij}. \tag{5}$$

Here, the model terms inside the two brackets in (5) correspond to all coefficients for the intercept and slope terms, respectively.

When the error terms are equal to their assumed means of zero, (6) reduces to the population-mean trend of $y_{ij} = \beta_{00} + \beta_{10}\triangle T_{ij}$. The grand mean intercept β_{00} quantifies the expected *marginal* mean comparison score when $\triangle T_{ij} = 0$. Note that this intercept is not particularly meaningful, as our data does not contain any same-day comparisons. However, interpretation of β_{00} does give us some notion of differences in subject's initial statuses, or comparison scores at baseline. The primary coefficient we are interested in is β_{10} which quantifies the expected change in mean comparison score per one-year increase in elapsed time since enrollment. Because this model, as well as all others considered in this paper, include random terms for both intercepts and slopes (b_{0i} and b_{1i}), we can also analyze the variation in the population's parameters (*i.e.* differences in the trends of individuals in the population).

2) Function of Elapsed Time and Age at Enrollment: If rates of change in comparison scores are steeper or flatter throughout an individual's lifetime, then face recognition performance may be a function of elapsed time, as well as absolute age. If we add the age of the enrollment image[6] to the mixed-effects model in (5), the model becomes:

$$y_{ij} = \big[\beta_{00} + \beta_{01}AGE_{ie} + b_{0i}\big] + \big[\beta_{10} + b_{1i}\big]\triangle T_{ij} + \varepsilon_{ij}. \tag{6}$$

Because AGE_{ie} is a constant for each subject, it is a *time-invariant*, or fixed, effect, and the above composite model actually has a two-level specification with the same level-1 model in (1). Hence, AGE_{ie} cannot improve the model fit at level-1 (within-subject); it can only influence the level-2 subject-specific variations. The population-mean trend for (6) is:

$$\begin{aligned} y_{ij} &= \beta_{00} + \beta_{01}AGE_{ie} + \beta_{10}\triangle T_{ij} \\ &= \beta_{00} + \beta_{01}AGE_{ie} + \beta_{10}(AGE_{ij} - AGE_{ie}). \end{aligned} \tag{7}$$

By definition, $\triangle T_{ij}$ is a *centered* version of AGE_{ij}, where the centering term (AGE_{ie}) is subject-specific. Hence, the model for aging as a function of elapsed time and age at enrollment, $f(\triangle T_{ij}, AGE_{ie})$, is mathematically equivalent to a model for aging as a function of the age of the query image and age at enrollment, $f(AGE_{ij}, AGE_{ie})$:

$$y_{ij} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{10}AGE_{ij}. \tag{8}$$

The two models in (7) and (8) will result in the same estimate for longitudinal change, β_{10}. What distinguishes them is the interpretation of the coefficient β_{01} quantifying the effect of AGE_{ie}. Note the relationship between the two models: $\beta_{01}^{(8)} = \beta_{01}^{(7)} - \beta_{10}^{(7)}$. Hence, $\beta_{01}^{(8)}$ is the "contextual" effect that models the *difference* between the within- and between-subject effects of time [23].[7] The significance of subject age at enrollment in Model (8) is tested with the null hypothesis of $H_0 : \beta_{01} = 0$, whereas *restricted inference* is needed to test

[6]Comparing all images of an individual to a fixed enrollment image means that AGE_{ij} and $\triangle T_{ij}$ are perfectly correlated at level-1 (within-subject) of the model. Hence, we cannot include both of these covariates; the effect of age must be added as a level-2 covariate.

[7]The equality $\beta_{01}^{(8)} = \beta_{01}^{(7)} - \beta_{10}^{(7)}$ holds for mixed-effects models with random intercepts, and is approximately true for models with both random intercepts and random slopes.

significance in Model (7). In particular, the null hypothesis must instead be $H_0 : \beta_{01} = \beta_{10}$.

The relationship between these two models is similar to common approaches for decoupling the longitudinal and cross-sectional effects of a time-varying covariate. A time-varying covariate at level-1 (*e.g.* age or elapsed time) exhibits variability *within*, but also *between* individuals; models which assume that the within- and between-individual effects are equal do not properly estimate either of these effects [19], [24], [25], [23]. The estimated regression coefficient (*i.e.* β_{10} in (2) and (5)) which quantifies the expected change in the response for a unit increase in a covariate x_{ij}, is actually a weighted combination of the true longitudinal (within-subject) change and the cross-sectional (between-subject) effect, where the weights are related to the proportion of between-subject variation in the covariate, relative to the within-subject variation [19]. Typically, the time-varying covariate is "centered" on subject-specific means. Adjusting for "cluster-level" mean covariate levels can remove potential confounding bias in estimation of the effect of the individual-level covariate and the outcome/response [25].

B. Model Comparison and Evaluation

All models in our analysis are fit with full maximum likelihood (ML) estimation via iterative generalized least-squares (GLS). Goodness-of-fit measures based on log-likelihood statistics can be used to compare models with different covariates and degrees of complexity. Deviance quantifies how much worse the current model is compared to the (hypothetical) saturated model that includes all possible covariates to perfectly fit the data. Because the log-likelihood (LL) of the saturated model is 0,

$$\text{Deviance} = -2[LL_{\text{current}} - LL_{\text{saturated}}] = -2LL_{\text{current}}. \quad (9)$$

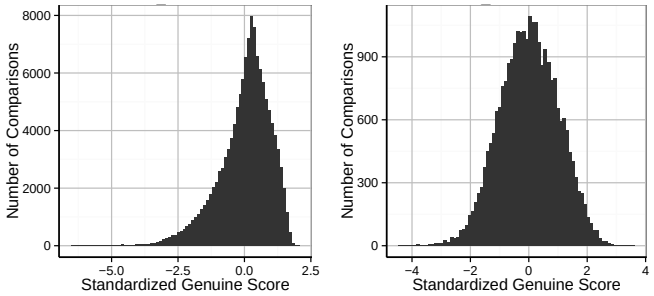
Deviance can be used to calculate χ^2 statistics for comparing nested models (*i.e.* the simpler model is a reduced form of the more complex model, where some coefficients in the complex model are equal to 0) that are fit to the same data. To compare non-nested models, Akaike Information Criterion (AIC) penalizes the log-likelihood by the number of parameters[8] in the model, and Bayesian Information Criterion (BIC) additionally penalizes large sample sizes. Note that for all three goodness-of-fit measures, smaller values indicate better fit.[9] Additionally, to evaluate different models, normal probability plots of the level-1 residuals and level-2 random effects (for both intercepts and slopes) can be used to determine whether model assumptions (3) and (4) are valid; the error terms follow Gaussian distributions if the normal probability plots are linear.

V. Results

The goal of statistical modeling is to find a model that includes substantive predictors and excludes unnecessary ones (parsimony). In this paper, we are interested in determination of significant factors that explain the variation in genuine comparison scores of face images, particularly over time. A common approach to arrive at a "final model" is to fit increasingly complex models to successively evaluate the impact of adding different covariates [18]. We first focus on analysis of the PCSO_LS database, starting with the simple models discussed in Section IV and progressing to more complex models including covariates for face quality and subject sex and race. We then present results for the LEO_LS database.

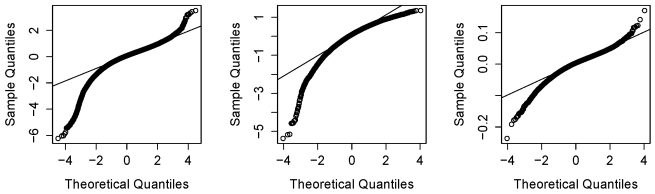
[8]For full ML estimation, the number of parameters includes both the fixed effects and the variance components.

[9]For AIC and BIC, the magnitude of the reduction in fit is difficult to interpret.



(a) PCSO_LS (b) LEO_LS

Fig. 7: Distributions of standardized genuine comparison scores from (a) PCSO_LS and (b) LEO_LS longitudinal face databases; scores are from (a) COTS-A and (b) COTS-B.



(a) (b) (c)

Fig. 8: Normal probability plots of (a) level-1 residuals, ε_{ij}, and level-2 random effects for (b) intercepts, b_{0i}, and (c) slopes, b_{1i} from Model BT on the PCSO_LS database. Departure from normality at the tails of the distributions is likely due to low quality face images or errors in subject IDs.

A. Model Assumptions

While mixed-effects models are capable of handling non-Gaussian response distributions (*e.g.* COTS-A genuine scores for the PCSO_LS database in Fig. 7a), the error terms must follow Gaussian distribution. Figure 8a shows normal probability plots of the level-1 residuals, ε_{ij}, from fitting Model BT to genuine scores from the PCSO_LS database. Since significant departure from linearity is observed at the tails, we cannot verify that the model assumptions hold; normal probability plots of random effects, b_{0i} and b_{1i}, also depart from linearity (Figs. 8b, 8c). This behavior was observed for other models as well, precluding the use of standard errors for formal hypothesis tests of parameters [26].

In situations where parametric model assumptions are violated, it is common to resort to non-parametric bootstrap to establish confidence intervals for the parameter estimates. Hence, for the PCSO_LS database, we conduct a non-parametric bootstrap by *case resampling* [26]; 1,000 bootstrap replicates

TABLE IV: Bootstrap results for mixed-effects models on the PCSO_LS database (COTS-A genuine scores)

		Model A	Model BT	Model CT	Model CA	Model D
FIXED EFFECTS (95% CONFIDENCE INTERVALS):						
(INTERCEPT)	β_{00}	0.02736	0.67343	0.72258	0.71267	0.51576
		$(0.01711, 0.03755)$	$(0.66238, 0.68490)$	$(0.69048, 0.75563)$	$(0.68134, 0.74335)$	$(0.40727, 0.62394)$
TIME	β_{10}		-0.13640	-0.13639	-0.13772	-0.13718
			$(-0.13792, -0.13494)$	$(-0.13792, -0.13493)$	$(-0.13922, -0.13630)$	$(-0.14257, -0.13155)$
AGE GROUP	β_{01}			-0.00160	0.13676	0.01199
				$(-0.00265, -0.00055)$	$(0.13490, 0.13854)$	$(0.00472, 0.01891)$
AGE GROUP × TIME	β_{11}					$0.00002^{\#}$
						$(-0.00016, 0.00020)$
AGE GROUP^2	β_{02}					-0.00020
						$(-0.00032, -0.00009)$
VARIANCE COMPONENTS:[a]						
Level-1 Residual	σ^2_ε	0.60762	0.39115	0.39116	0.40781	0.39117
Random Intercepts	σ^2_0	0.38409	0.32433	0.32393	0.78544	0.32311
Random Slopes	σ^2_1		0.00283	0.00283	0.00060	0.00283
Covariance	σ_{01}		-0.00386	-0.00385	-0.01725	-0.00382
GOODNESS-OF-FIT:[b]						
	AIC	333433	287016	287006	288148	286985
	BIC	333462	287074	287075	288217	287073
	Deviance	333427	287004	286992	288134	286967

[a] Confidence intervals for variance components have been omitted due to space limitations.

[b] Goodness-of-fit values are the mean values of the 1,000 bootstrap samples.

are generated by sampling 18,007 subjects with replacement. Multilevel models are fit to each bootstrap replicate, and the mean parameter estimates over all 1,000 bootstraps are reported. Tests for fixed effects parameters can be conducted by examining the bootstrap confidence intervals.[10] Table IV gives the bootstrap parameter estimates (with 95% confidence intervals), variance components, and goodness-of-fit for the multilevel models in Table III.

B. Unconditional Means Model

The simplest mixed-effects model is the unconditional means model, which partitions the total variation in comparison scores by subject. Denoted Model A in Table III, and with composite form of

$$y_{ij} = \beta_{00} + b_{0i} + \varepsilon_{ij}, \tag{10}$$

b_{0i} is the *subject-specific mean* and β_{00} is the *grand mean*. Similar to analysis of variance (ANOVA), Model A provides initial estimates of the within-subject variance σ^2_ε (*i.e.* deviations around each subject's own mean comparison score) and the between-subject variance σ^2_0 (*i.e.* deviations of subject-specific means around the grand mean). The intraclass correlation coefficient (ICC) quantifies the proportion of between-subject variation in the response, $\rho = \sigma^2_0/(\sigma^2_0 + \sigma^2_\varepsilon)$. Variance components for Model A shown in Table IV indicate that between-subject differences account for 38.8% ($\rho = 0.3873$) of the total variation in genuine scores from the PCSO_LS database. Baseline goodness-of-fit measures are also shown in Table IV.

Further comparisons of models depend on whether the successive model has included a time-invariant (*e.g.* gender, race) or time-varying (*e.g.* face quality, age) covariate to the baseline model. For both cases, pseudo-R^2 statistics can be calculated to measure the proportional reduction in variance attributable to additional covariates. For addition of time-invariant predictors, we can examine the level-2 variance components (*e.g.* σ^2_0 for Model A); Note that time-invariant predictors cause no change in the level-1 residual variance σ^2_ε. For time-varying predictors, we can examine proportional reductions in the level-1 residual variance.

C. Unconditional Growth Model: Elapsed Time

The next model to consider in longitudinal analysis is the unconditional growth model that includes the time-related covariate. In our case, we add elapsed time, $\triangle T_{ij}$, as well as random effects for slopes, b_{1i} to Model A, resulting in Model BT. Table IV shows that Model BT estimates that PCSO_LS genuine scores decrease by 0.1364 standard deviations per one-year increase in elapsed time (see solid black line in Fig. 9). Comparing the level-1 residual variation of Models A and BT, elapsed time explains 35.6% of the variation in a given subject's genuine scores around his/her own average genuine score.[11]

Longitudinal change estimated by Model BT implies that the *population-mean trend* will drop below the thresholds for 0.001% and 0.01% FAR after 13.5 and 19.1 years elapsed time, respectively. However, this only provides insight into performance on the average (*i.e.* typical) subject in the population. Figure 9 plots a region of two standard deviations of subject-specific deviations around the population-mean trend and shows that subjects can have individual trends that are quite different from the population-mean. Since two standard deviations is approximately 95% of a distribution, this figure indicates that genuine scores for 95% of the population will remain above the threshold at 0.01% FAR for up to 5 years elapsed time. From 5 to 10 years elapsed time, an additional 14% (1σ) of the population will begin to drop below the 0.01% FAR threshold.

[10] The null hypothesis of the parameter value equal to 0 can be rejected at significance of 0.05 if the 95% confidence interval does not contain 0.

[11] Using pseudo-$R^2 = (\sigma^2_\varepsilon(A) - \sigma^2_\varepsilon(BT))/\sigma^2_\varepsilon(A)$.

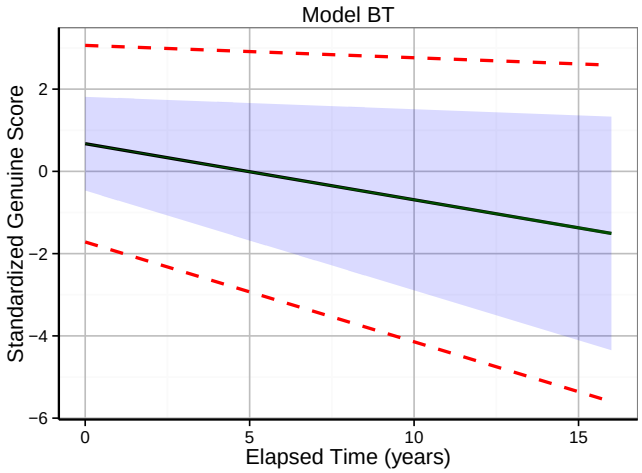


Fig. 9: Model BT on COTS-A genuine scores from the PCSO_LS database. The bootstrap-estimated population-mean trend is shown in black (confidence intervals are too small to be visible). The blue band plots two standard deviations of the subject-specific intercepts and slopes around the population-mean trend; dashed red lines additionally add σ_ε to the subject deviations. Hence, approximately 95% of the subjects lie within the blue band, but scores around their trends can further extend to the red dashed lines. Note that the thresholds (based on the full impostor distribution of PCSO_LS) at 0.001%, 0.01%, and 0.1% FAR for COTS-A are -1.16, -1.93, and -2.60, respectively.

D. Elapsed Time and Subject Age Group

We next investigate whether the population-mean trends in genuine scores over time depend on a subject's absolute age (*i.e.* whether variation in subject-specific trends observed in Model BT can be explained by differences in subject age). Firstly, whether considering face recognition performance as a function of age at enrollment and elapsed time, $f(AGE_{ie}, \triangle T_{ij})$, or age at enrollment and age of query image, $f(AGE_{ie}, AGE_{ij}$, both Models CT and CA result in similar estimates of rates of change due to elapsed time, as well as a signifcant effect of age at enrollment. However, note the difference between the magnitude of the estimated β_{01} coefficients in Table IV for Models CT and CA. Recall that β_{01} in Model CT is the true between-subject effect of AGE_{ie} because $\triangle T_{ij}$ at level-1 uses AGE_{ie} as a centering constant for each subject, whereas β_{01} in Model CA actually estimates the *difference* between the longitudinal change and the effect of age at enrollment. Hence, from the relatively small magnitude of β_{01} in Model CT, we can conclude that the effect of age at enrollment is statistically significant, but a 100-year increase in age at enrollment is required for the change in genuine scores to be of the same order of magnitude as a one-year increase in elapsed time.

To further test the complexity of the effects of age at enrollment, we add additional terms associated with AGE_{ie} to Model CT, resulting in Model D (recall Table III). The hypotheses of interest are:

1) Older subjects are easier to recognize than younger subjects, and
2) Younger subjects age at faster rates than older subjects.

These two hypotheses manifest in younger subjects having higher genuine scores, on average, and steeper negative rates of change. Note that the significance of the AGE_{ie} term in Model CT actually suggests a *negative* linear relationship between age at enrollment and genuine scores.

Table IV shows that the interaction term $AGE_{ie} \times \triangle T_{ij}$ in Model C is not significant because the 95% confidence interval for β_{11} contains zero; hence, we cannot conclude that subject enrollment age has a linear effect on rates of change in COTS-A genuine scores for the PCSO_LS database. The statistically significant β_{02} coefficient indicates a quadratic relationship between subject enrollment age and intercepts, and goodness-of-fit measures are lower compared to Model BT. However, further comparing to Model BT, level-2 variation in random effects for intercepts (σ_0^2) is only reduced by 0.4% after including AGE_{ie} terms. The differences between scores for different ages at enrollment are marginal compared to the change in scores due to elapsed time; the change in score between a a 20 year-old and a 30 or 50 year-old (at enrollment) is equivalent to only 7 and 5 *months* of elapsed time (longitudinal change), respectively.

The relatively small, arguably marginal, effects of subject age at enrollment could be explained as follows: (i) This particular matcher, COTS-A, is not largely sensitive to subject age. (ii) The PCSO_LS database is limited in that the age range is primarily 18–60 years old (473 subjects older than 60). Additional data for both younger and older subjects may result in a larger effect of subject age. (iii) The relationship between elapsed time and subject age is more complex than our models allow for. For example, Model C does not account for whether subjects with enrollment age of 20 years have query images spanning 21–26 or 31–36 years old. Figure 10 shows that this type of scenario is present in the highly time-unstructured PCSO_LS database. (iv) There are significant omitted covariates, particularly at level-1 (within-subject).

E. Face Quality

Adding level-2 covariates (*i.e.* time-invariant values for each subject, such as AGE_{ie}) cannot improve the fit of the model at level-1 (within-subject). Table IV shows that the level-1 residual variation σ_ε^2 (*i.e.* deviation of scores around each subject's own linear trend) is quite large when time is the only level-1 covariate for all models considered thus far. One standard deviation of level-1 residual variation estimated by Model BT (and similarly Models CT and D) is equivalent to 4.6 years of elapsed time (calculated as $\sqrt{\sigma_\varepsilon^2}/\beta_{10} = \sqrt{0.39117}/-0.13718$). This is visually shown by the dotted red lines in Fig. 9.

Level-1 residual variation can only be reduced by time-varying covariates; in this section we investigate whether face image quality measures can be used to improve the model fit. The quality measures considered are interpupillary distance (IPD) and a "frontal" score, both of which are output by COTS-A. While higher frontalness indicates better quality, the

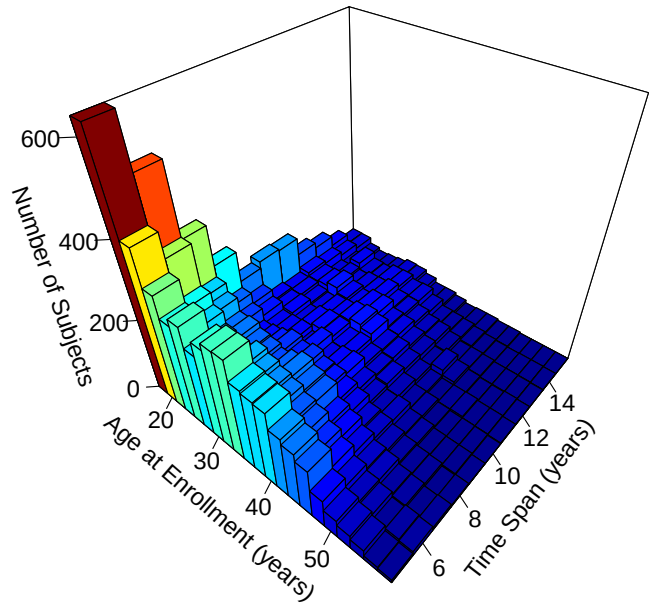


Fig. 10: Histogram of number of subjects for time span (age of oldest image minus age of youngest image) versus subject age at enrollment (age of youngest image) for the 18,007 subjects in the PCSO_LS database.

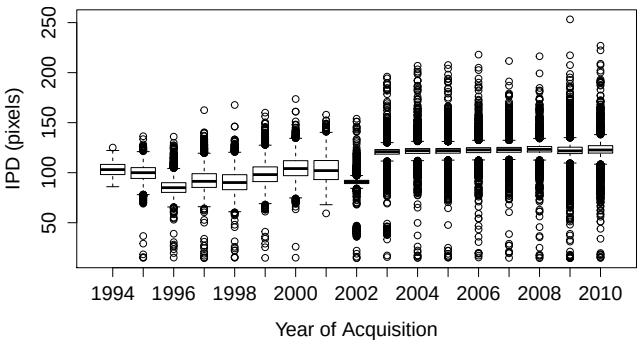


Fig. 11: A boxplot of interpupillary distances (IPDs) versus year of acquisition shows that mean IPDs systematically changed over time for the PCSO_LS database, likely due to booking stations adhering to face imaging standards in more recent years.

range of the frontal score has little meaning, since its computation is proprietary. We standardize the frontalness score so we can interpret model parameters as standard deviations from the mean of the frontalness scores from all images in PCSO_LS.

First, we fit models with a quality measure, either IPD or frontalness, as the only covariate (including separate measures for both enrollment and query images). In summary, we found that neither of the quality measures alone explain variation in genuine scores as well as Model BT with only elapsed time as covariate (details are omitted due to space limitations).

Second, we add the quality measures to Model BT such that the new level-1 (within-subject) model is:

$$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varphi_{2i}Q_{ij} + \varphi_{3i}Q_{ij}\triangle T_{ij} + \varepsilon_{ij}, \quad (11)$$

TABLE V: Bootstrap results for mixed-effects models with elapsed time and face quality covariates for the PCSO_LS database (COTS-A genuine scores)

	Model QF	Model QI	Model QFI
LEVEL-1 RESIDUAL VARIANCE:			
σ_ε^2	0.33022	0.35385	0.32175
GOODNESS-OF-FIT:			
AIC	275108	281296	273643
BIC	275283	281471	273848
Deviance	275072	281260	273601

[a]Confidence intervals for variance components have been omitted due to space limitations.

[b]Goodness-of-fit are the mean of the 1,000 bootstrap samples.

TABLE VI: Elapsed times in years for when population-mean trends in genuine scores drop below the decision thresholds at 0.001% and 0.01% FAR for different measures related to face quality (frontalness and IPD) of the enrollment image Q_{ie} and the query image Q_{ij}

	Q_{ie}	Q_{ij}	0.001% FAR	0.01% FAR
Frontal	-1σ	-1σ	10.9	15.6
	μ	μ	13.0	18.4
	1σ	1σ	16.8	23.0
IPD	100 pixels	100 pixels	13.8	19.4
	100 pixels	120 pixels	14.0	20.0
	120 pixels	120 pixels	13.0	18.4

and the level-2 (between-subject) model is:

$$\begin{aligned} \varphi_{0i} &= \beta_{00} + \beta_{01}Q_{ie} + b_{0i}, \\ \varphi_{1i} &= \beta_{10} + \beta_{11}Q_{ie} + b_{1i}, \\ \varphi_{2i} &= \beta_{20} + \beta_{21}Q_{ie} + b_{2i}, \\ \varphi_{3i} &= \beta_{30}, \end{aligned} \quad (12)$$

where Q_{ie} and Q_{ij} denote the quality measure of the enrollment and jth query images of subject i, respectively.

Table V gives estimated level-1 residual variation and goodness-of-fit for models with frontalness, IPD, and both frontalness and IPD (Model QF, QI and QFI, respectively). Model QF has a better overall fit than Model QI. Table VI gives the elapsed times for when population-mean scores cross thresholds at 0.001% and 0.01% FAR for different values of frontalness and IPD. Note how changing frontalness has a greater impact on when population-mean genuine scores cross the thresholds than changes in IPD. Model QFI with both measures of quality further reduces the level-1 residual variation and goodness-of-fit values.

The values of 100 and 120 pixels for IPD in Table VI were chosen because we observed systematic changes in IPDs over time (see Figure 11); in particular, mean IPD varies around 100 pixels from 1994–2002 but increases to a consistent ∼120 pixels after the year 2003. This observation, along with correspondence with Pinellas County Sheriff's Office, suggests that booking agencies began to adopt imaging standards during the years 2001–2003. To investigate whether this aspect of the data confounds the estimation of longitudinal effects (face

TABLE VII: Mixed-effects model results for the LEO_LS database

		Model A	Model BT	Model CT	Model CA	Model D
FIXED EFFECTS (STANDARD ERRORS):						
(INTERCEPT)	β_{00}	0.00368	0.53953	0.54676	0.56094	0.08944
		(0.00981)	(0.01273)	(0.03249)	(0.03413)	(0.10566)
TIME	β_{10}		−0.16994	−0.16994	−0.17028	−0.19801
			(0.00231)	(0.00231)	(0.00211)	(0.00764)
AGE GROUP	β_{01}			−0.00025	0.16971	0.03464
				(0.00105)	(0.00234)	(0.00682)
AGE GROUP × TIME	β_{11}					0.00098
						(0.00025)
AGE GROUP^2	β_{02}					−0.00060
						(0.00010)
VARIANCE COMPONENTS:						
Level-1 Residual	σ^2_ε	0.59854	0.42755	0.42755	0.45174	0.42745
Intercepts	σ^2_0	0.40091	0.55433	0.55417	1.10361	0.55164
Slopes	σ^2_1		0.00585	0.00584	0.00063	0.00578
Covariance	σ_{01}		−0.03173	−0.03171	−0.02165	−0.03161
GOODNESS-OF-FIT:						
	AIC	68705	62647	62649	62811	62606
	BIC	68730	62697	62707	62868	62679
	Deviance	68699	62635	62635	62797	62588

images in later years may be of higher quality), we also tested for a difference in slope prior to 2003 versus after 2003 by using a piecewise linear formulation for the mixed-effects model (with a breakpoint at 2003). We found that slope after 2003 was significantly flatter (less negative).

Additional face quality factors known to cause changes in face recognition performance are illumination, expression, and occlusions. However, there are no widely accepted methods for quantifying such variations in face images and doing so is beyond the scope of this paper.

F. LEO_LS Database

Table VII gives results for the models in Table III fit to COTS-B genuine scores from the LEO_LS database; fixed-effects parameter estimates are given with standard errors (bootstrapping was not needed for LEO_LS models because the error terms followed Gaussian distributions). Firstly, Model A estimates that 40% of the total variation in genuine scores is due to between-subject differences in subject-specific average scores. The longitudinal change in genuine scores estimated by both Model BT and Model CT indicates that a one year increase in elapsed time decreases genuine scores by $\beta_{10} = -0.16994$ standard deviations. Model CT also estimates that a one year between-subject increase in age at enrollment decreases genuine scores by $\beta_{01} = -0.00025$ standard deviations. Although this between-subject effect of age is orders of magnitude smaller than the longitudinal effect, is is still significantly different from β_{10}. The significance of age at enrollment can also be seen in Model CA where the null hypothesis of $\beta_{01} = 0$ is rejected at $p = 0$ significance level. Although both models show reductions in goodness-of-fit compared to Model A, the goodness-of-fit values (as well as level-1 residual variation σ^2_ε) for Model CT are lower than for Model CA, indicating that elapsed time as the level-1 covariate better explains the variation in comparison scores than the age of the query image as the level-1 covariate.

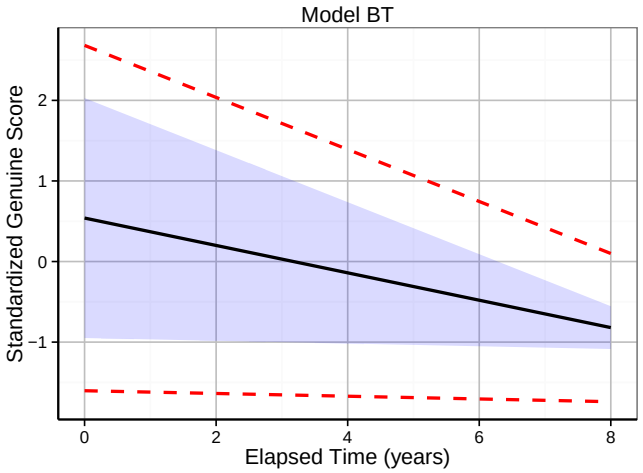


Fig. 12: Model BT on COTS-B genuine scores from the LEO_LS database. The population-mean trend is shown in black. The blue band plots two standard deviations of the subject-specific intercepts and slopes around the population-mean trend; dashed red lines additionally add σ_ε to the subject deviations. Hence, approximately 95% of the subjects lie within the blue band, but scores around their trends can further extend to the red dashed lines. Note that the thresholds (based on the full impostor distribution of LEO_LS) at 0.001%, 0.01%, and 0.1% FAR for COTS-B are -2.25, -2.54, and -2.84, respectively (all of which are below the lower limit of the y-axis).

The significant interaction term $AGE_{ie} \times \triangle T_{ij}$ in Model C indicates that longitudinal change in comparison scores tends to vary with subject's age at enrollment; a 10 year increase in subject's age results in a longitudinal slope that is $\beta_{11} = -0.00980$ standard deviations steeper. Population-mean rates of change range from -0.17841 to -0.14901 standard deviations per year for subjects with age at enrollment of 20 and 50 years, respectively (calculated as $\beta_{10}+\beta_{11}AGE_{ie}$). The significant β_{02} coefficient indicates a quadratic relationship between subject age group and intercepts ($p < 0.001$).

To test for effects of subject sex and race, we add sex and race to Model D as both fixed effects and interactions with elapsed time. Results indicate that intercepts (*i.e.* genuine scores at hypothetical $\triangle T_{ij} = 0$) are 0.05651 and 0.42375 standard deviations higher for black and male subjects, respectively (so, black-male subjects have intercepts that are 0.48026 standard deviations higher than white-female subjects). Slopes are not statistically different for black and white subjects, but the population-mean slope for males is 0.02106 *steeper* (*i.e.* more negative) than for females. These population-mean trends are shown in Fig. 13 for different ages at enrollment; while male genuine scores decrease at slightly faster rates than female scores, males are clearly easier to recognize with higher genuine scores overall. Fig. 13 also shows that the differences between subject race are minor compared to differences between males and females.

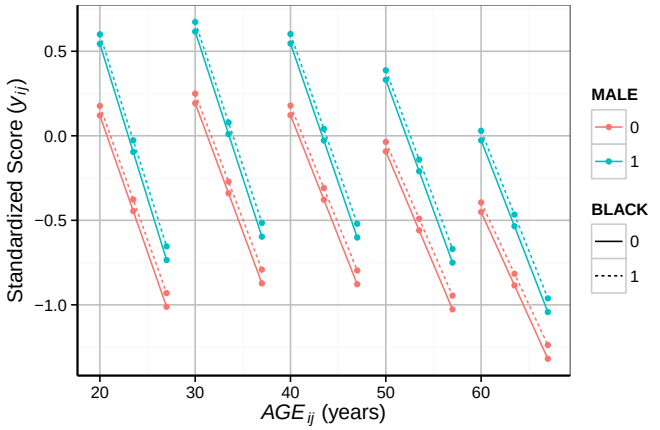


Fig. 13: Model C with additional subject sex and race covariates for COTS-B genuine scores from the LEO_LS database. Population-mean trends are plotted by subject demographics of sex and race, in addition to five different ages at enrollment (20 to 60 years). Each trend line represents seven years of elapsed time from the age at enrollment. For example, the solid blue line beginning at $AGE_{ij} = 20$ years represents the population-mean trend for white males enrolled at age 20. Note that the thresholds (based on the full impostor distribution of LEO_LS) at 0.001%, 0.01%, and 0.1% FAR for COTS-B are -2.25, -2.54, and -2.84, respectively (all of which are below the lower limit of the y-axis).

VI. Discussion

We presented a longitudinal study of automatic face recognition, utilizing two large operational databases of mugshots, PCSO_LS (147,784 images of 18,007 subjects) and LEO_LS (31,852 images of 5,636 subjects), where each subject has at least four face images acquired over at least a five-year time span. Mixed-effects statistical models were used to analyze variation in genuine scores due to elapsed time, age, sex, and race, as well as subject-specific differences in scores (*i.e.* biometric zoo effects). Face comparison scores were obtained from a state-of-the-art COTS matcher for both the PCSO_LS and LEO_LS databases. Based on our analysis of these two databases, we make the following observations:

1) Population-mean trends indicate that genuine scores significantly decrease with increasing elapsed time, to no surprise. However, the population-mean trends, if sustained, estimate that average genuine scores do not fall below thresholds at 0.01% FAR until after 15 years elapsed time for both databases.
2) Analysis of subject-specific variations in aging trends suggests that genuine scores for 95% of the population will still remain above thresholds at 0.01% FAR up to approximately 5 years elapsed time.
   a) But going from 5 years elapsed time to 10 years elapsed time, more subjects start to quickly fall below the 0.01% FAR threshold.
   b) A rough approximation is that when matching a query image after 10 years elapsed time since enrollment, 20% of the subjects would experience false rejection errors.
3) We observed a downward quadratic trend in average genuine scores with respect to age at enrollment (with maximum scores at approximately 30 years-old). Of the ages studied in our analysis, older subjects (55–60) were more difficult to recognize (with the lowest average genuine scores). However, rates of change in genuine scores over time get flatter (less negative) with increasing age at enrollment. This suggests that younger individuals age faster than older individuals.
   a) While the effects of age at enrollment considered here were statistically significant, they resulted in only marginal reductions in the variation in subject-specific trends. Hence, again, while the population-mean trends for all subject age groups remain above thresholds at low FARs, the subject-specific deviations still indicate that false rejections due to aging are an issue.
   b) Caveat: Our models make no distinction between, for example, two subjects with age at enrollment of 20 years-old but with query images of ages 21-27 versus 26-33. These two subjects are considered to have the same effect of age.
4) Subject sex has a larger effect on genuine scores and rates of change than subject race. Both COTS matchers compute significantly higher genuine scores for males than for females. The two COTS matchers/databases did not agree on the direction of the effect of subject race (black/white), but the magnitude of the effect is consistently minor compared to the effect of subject sex.
   a) Rates of change in genuine scores (slopes) tend to depend more on subject sex than race (effect of race on slopes was not significant for COTS-B on the LEO_LS database); male scores decrease faster over time than female scores.
5) While the model fit improved for more complex models incorporating simple measures of face quality (for the PCSO_LS database), the models are still limited for *prediction* purposes.
   a) The smallest level-1 residual variation for the most complex model considered for PCSO_LS was $\sigma_\varepsilon^2 = 0.33$, so one standard deviation around a given subject's trend means that genuine scores can change by ± 0.57 standard deviations of the full genuine score distribution. This is equivalent to the change in scores due to approximately 3.5 years of elapsed time (if $\beta_{10} = -0.165$). Stated otherwise, at a given value of elapsed time, short-term variations such as illumination, expression, etc. can cause genuine scores to change by the same amount as ± 3.5 years of aging.
   b) Comparatively, short-term variations in iris and fingerprint genuine scores was orders of magnitudes smaller than or equivalent to, respectively, the decrease in scores across 10 and 16 years elapsed time (the full periods of time of the iris and fingerprint databases used in longitudinal studies [8] and [9]). This suggests that elapsed time plays a minor factor in iris and

fingerprint scores compared to other factors like pupil dilation and fingerprint image quality.

In the absence of a reliable face quality measure, mixed-effects models quickly become very complicated for analysis when various factors are considered (such as IPD, a measure of frontalness, etc.), especially considering that both the quality of the enrollment and query face images should be considered. Longitudinal analysis, in general, is a very difficult problem, and to the best of our knowledge, no proper statistical analysis has yet been conducted for studying face recognition performance over long periods of time. In this paper, we attempted to analyze the covariates of interest that were available to us (elapsed time, age, sex, race, etc.), but there are additional covariates that cannot be accounted for because we do not have the information (*e.g.* camera characteristics, IPD for the LEO_LS database, expression variations, etc.). Additionally, observations detailed in this paper are both matcher and database dependent. However, the longitudinal study on automatic face recognition presented here utilizes two of the *largest*, *deepest*, and *longest* (in terms of number of subjects, number of images per subject, and time spans of subject images, respectively) face image databases studied to date, and the COTS matchers are representative of current state-of-the-art.

Future work will include: (i) Evaluation of face *identification* (both closed-set and open-set) performance over time. Statements about recognition accuracy in this paper apply to face *verification* scenarios (*i.e.* one-to-one comparisons) operating with a global threshold. (ii) Models for all pairwise comparisons to include $\triangle T_{ijk}$ and AGE_{ijk} as two time-varying level-1 covariates, and a covariance structure that can account for the pairwise relationships. (iii) Investigation into a single face quality measure for mugshot type face images.

ACKNOWLEDGMENT

The authors would like to thank Patrick Grother and Mei Ngan at the National Institute of Standards and Technology (NIST) for collaboration in providing covariates and COTS-B comparison scores for the LEO_LS database.

REFERENCES

[1] P. Grother and M. Ngan, "FRVT: Performance of face identification algorithms," NIST Interagency Report 8009, May 2014.
[2] D. Wang, C. Otto, and A. K. Jain, "Face search at scale: 80 million gallery," http://arxiv.org/abs/1507.07242, Jul. 2015.
[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *Proc. CVPR*, 2014.
[4] B. Klare and A. K. Jain, "Face recognition across time lapse: On learning feature subspaces," in *Proc. IJCB*, 2011.
[5] C. Otto, H. Han, and A. Jain, "How does aging affect facial components?" in *ECCV WIAF Workshop*, 2012.
[6] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, "Face verification across age progression using discriminative methods," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 1, pp. 82–91, Mar. 2010.
[7] M. Bereta, P. Karczmarek, W. Pedrycz, and M. Reformat, "Local descriptors in application to the aging problem in face recognition," *Pattern Recognition*, vol. 46, no. 10, pp. 2634–2646, Oct. 2013.
[8] P. Grother, J. R. Matey, E. Tabassi, G. W. Quinn, and M. Chumakov, "IREX VI: Temporal stability of iris recognition accuracy," NIST Interagency Report 7948, Jul. 2013.
[9] S. Yoon and A. K. Jain, "Longitudinal study of fingerprint recognition," *Proc. National Academy of Sciences*, vol. 112, no. 28, pp. 8555–8560, Jul. 2015.
[10] L. Best-Rowden and A. K. Jain, "A longitudinal study of automatic face recognition," in *Proc. International Conference on Biometrics*, 2015.
[11] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. FGR*, 2006.
[12] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. on PAMI*, vol. 24, no. 4, Apr. 2002.
[13] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips, "A meta-analysis of face recognition covariates," in *Proc. BTAS*, 2009.
[14] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proc. ICCV*, 2013.
[15] F. Juefei-Xu, K. Luu, M. Savvides, T. D. Bui, and C. Y. Suen, "Investigating age invariant face recognition based on periocular biometrics," in *Proc. IJCB*, 2011.
[16] M. Erbilek and M. Fairhurst, "A methodological framework for investigating age factors on the performance of biometric systems," in *Proc. Multimedia and Security*, 2012.
[17] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper, "Factors that influence algorithm performance in the face recognition grand challenge," *CVIU*, vol. 113, pp. 750–762, 2009.
[18] J. D. Singer and J. B. Willett, Eds., *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford Univ. Press, Inc., 2003.
[19] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied Longitudinal Analysis*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2011.
[20] S. Galbraith, J. Bowden, and A. Mander, "Accelerated longitudinal designs: An overview of modelling, power, costs and handling missing data," *Statistical Methods in Medical Research*, vol. 0, no. 0, pp. 1–25, Aug. 2014.
[21] 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 *Proc. ICSLP*, 1998.
[22] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Trans. on PAMI*, vol. 32, no. 2, pp. 220–230, Feb. 2010.
[23] A. Bell and K. Jones, "Explaining fixed effects: Random effects modeling of time-series cross-sectional and panel data," *Political Science Research and Methods*, vol. 3, no. 1, pp. 133–153, Jan. 2015.
[24] J. M. Neuhaus and J. D. Kalbfleisch, "Between- and within-cluster covariate effects in the analysis of clustered data," *Biometrics*, vol. 54, no. 2, pp. 638–645, Jun. 1998.
[25] M. D. Begg and M. K. Parides, "Separation of individual-level and cluster-level covariate effects in regression analysis of correlated data," *Statistics in Medicine*, vol. 22, no. 16, pp. 2591–2602, Aug. 2003.
[26] R. van der Leeden, F. M. Busing, and E. Meijer, "Bootstrap methods for two-level models," in *Multilevel Conf.*, 1997.