Longitudinal Study of Automatic Face Recognition

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

Abstract—The two underlying premises of automatic face recognition are uniqueness and permanence. This paper investigates the permanence property by addressing the following: Does face recognition ability of state-of-the-art systems degrade with elapsed time between enrolled and query face images? If so, what is the rate of decline w.r.t. the elapsed time? While previous studies have reported degradations in accuracy, no formal statistical analysis of large-scale longitudinal data has been conducted. We conduct such an analysis on two mugshot databases, which are the largest facial aging databases studied to date in terms of number of subjects, images per subject, and elapsed times. Mixed-effects regression models are applied to genuine similarity scores from state-of-the-art COTS face matchers to quantify the population-mean rate of change in genuine scores over time, subject-specific variability, and the influence of age, sex, race, and face image quality. Longitudinal analysis shows that despite decreasing genuine scores, 99% of subjects can still be recognized at 0.01% FAR up to approximately 6 years elapsed time, and that age, sex, and race only marginally influence these trends. The methodology presented here should be periodically repeated to determine age-invariant properties of face recognition as state-of-the-art evolves to better address facial aging.

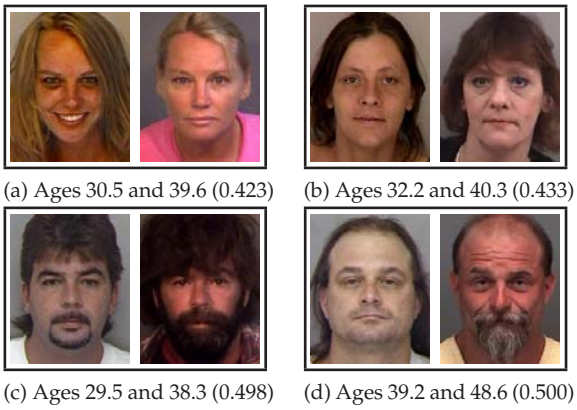
Index Terms—face recognition, facial aging, longitudinal study, mixed-effects models, multilevel models, random effects.

1 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 accuracies as high as 99% true accept rate (TAR) at a false accept rate (FAR) of 0.1% 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" (*e.g.,* images posted on the web) to account for PIE, occlusion, and partial face images. Comparatively, aging variations (*i.e.,* large time lapse between pairs of images being compared) have received considerably less attention in the face recognition community.

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 because assumptions of independent observations require that there be only one measurement per individual in the study. Past and future measurements are either not considered or are summarized into a single measurement which loses information; trends of individuals over time are not analyzed. Hypotheses about facial aging are, instead, *longitudinal* by nature and require multiple measurements of the same individuals over time to reveal trends in comparison scores with respect to facial aging.



(a) Ages 30.5 and 39.6 (0.423) (b) Ages 32.2 and 40.3 (0.433)

(c) Ages 29.5 and 38.3 (0.498) (d) Ages 39.2 and 48.6 (0.500)

Fig. 1: Face image pairs of four 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) are shown in parentheses (score range is [0.0, 1.0]). The thresholds at 0.01% and 0.1% FAR are 0.533 and 0.454, respectively. Hence, all of these genuine pairs would be falsely rejected at 0.01% FAR, while the two female subjects, (a) and (b), would also be rejected at 0.1% FAR.

To what extent facial aging affects the performance of automatic face recognition systems is of more than academic 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 five years for minors

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

and ten years for adults, while U.S. driver's licenses typically require renewal every five years. Additionally, to our knowledge, ensuring that a new (more recent) photo has been submitted for renewal is not verified, especially for renewals by mail or online. 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. Fig. 1 shows that elapsed times of eight to ten years between two face images can cause false non-match errors. 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 accuracy.

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 addressing the following question: *How robust are state-of-the-art automatic face recognition systems to facial aging?* In this paper, we conduct a longitudinal analysis of the performance of state-of-the-art COTS face matchers on two longitudinal face image databases consisting of repeat criminal offenders (mugshots) from two different law enforcement agencies (see Table 2). The COTS matchers used here are among the top-ranked performers in the FRVT 2013 face recognition evaluation [1]. The contributions of this paper can be summarized as follows:

1) *Longitudinal analysis of two of the largest longitudinal databases studied to date.* LEO_LS contains 31,852 images of 5,636 subjects, and PCSO_LS contains 147,784 images of 18,007 subjects, where the average time span between a subject's multiple image acquisitions is 6.1 and 8.5 years, respectively. Such large-scale databases allow for evaluation of performance at low FAR values (*e.g.,* 0.01% and 0.1%). Previous studies (*e.g.,* [4], [5]) evaluated at 1% FAR and higher.
2) *Determine the age-invariant properties of current state-of-the-art face matchers.* Rates of change over time in genuine comparison scores are analyzed using mixed-effects regression models, which are appropriate for longitudinal data. In doing so, we quantify (i) the population-mean rate of change in genuine scores over time and (ii) the variability in subject-specific longitudinal trends (*i.e.,* how closely individuals in the population follow the population-mean trend). We also investigate the influence of age at enrollment, sex, race, and face image quality.
3) *Methodology and analysis tools for advancing the development and evaluation of age-invariant face recognition algorithms.* The analysis conducted in this paper can be applied to any matcher and any database. Periodic reevaluation will be necessary as face recognition technology evolves to better address facial aging.[2]

Our previous longitudinal analysis of automatic face recognition was first published in [10]. The present work extends and refines our previous study in significant ways. The primary differences are as follows. (i) We study longitudinal effects of both *aging* (elapsed time) and *age* (biological age); [10] only studied elapsed time. (ii) Genuine scores are computed to represent a scenario where the youngest image of each subject is enrolled in a gallery (a subject with n_i total images has $n_i - 1$ scores, whereas [10] computed all $\binom{n_i}{2}$ genuine scores). Comparing query images to an enrollment image (a fixed point in time) simplifies the complex correlation structure that is present for all pairwise comparisons. (iii) We analyze an additional longitudinal face database (namely, LEO_LS) from a different law enforcement agency than the PCSO_LS database used in [10], and a different COTS matcher is used to obtain genuine scores for LEO_LS. Still, longitudinal analysis shows similar results for both databases and matchers.

The remainder of this paper is organized as follows. Section 2 highlights related work on facial aging as it pertains to automatic face recognition. Section 3 details the two longitudinal face databases used in this study. Section 4 explains the methodology used for longitudinal analysis. Section 5 gives results for both the PCSO_LS and LEO_LS face databases. Section 6 summarizes our observations about the current longitudinal capabilities of automatic face recognition.

2 Related Work

Almost all of the published studies that investigate 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 conclusions from the differences in performance across the partitions. Such an approach has led to the following general conjectures [11]: (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 1 for a summary of these studies.[3]

Partitioning of data (images or subjects) based on age group or time lapse is often arbitrary and varies from one study to another. Erbilek and Fairhurst show that different age group partitionings result in different performance trends for both iris and signature modalities [14]. 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.* [18] apply generalized linear mixed models to verification decisions (accept or reject) 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

1. A biometric system operating with a global threshold uses the same decision threshold for all subjects across all comparisons.

2. To facilitate longitudinal study on other face datasets and matchers, the code of our longitudinal analysis will be made publicly available at http://biometrics.cse.msu.edu/.

3. Studies that address developing age-invariant face recognition algorithms (*e.g.,* [12], [13]) are beyond the scope of this paper.

TABLE 1: Table of related work on the effects of facial aging on face recognition performance.

Study	Database	Age or Elapsed Time Partitions	Summary of Findings
Ling *et al.* [6]	Passports (private)	4–11 years elapsed time	Degradation in EER saturates after 4 years elapsed time.
	FG-NET	0–8, 8–18, and 18+ years old	Verification accuracies increase with increasing age group.
Klare and Jain [4]	PCSO (200,000 mugshots, 64,000 subjects)	0-1, 1-5, 5-10, 10+ years elapsed time	TARs at 1% FAR are 96.3%, 94.3%, 88.6%, and 80.5% for the listed elapsed time partitions. Training/testing on different aging partitions decreases performance in some non-aging scenarios.
Otto *et al.* [5]	MORPH-II	0–1, 1–5 years elapsed time	TARs at 1% FAR are 97% and 95% for the listed elapsed time partitions. The nose is the most stable facial component over time.
Bereta *et al.* [7]	FG-NET	0–5, 6–10, 11–15, 16–20, 21–30, and 30+ years elapsed time; 23–30, 31–40, 41–50, and 50+ years old	Identification accuracies of local descriptors (*e.g.*, variants of LBP) when combined with Gabor wavelet magnitudes become relatively consistent across absolute ages and age gap groups, but accuracies are still fairly low for a small gallery.
NIST FRVT [1]	Visa images (19,972 subjects)	baby, kid, pre-teen, teen, young, parents, older	Error rates (for open-set identification) are higher for younger age groups when the same threshold is used for all age groups.

EER = equal error rate; TAR = true accept rate; FAR = false accept rate

TABLE 2: Facial Aging Databases

Database	Num. Subjects	Total Num. Imgs	Num. Imgs per Subject	Age Range (years)
FG-NET [15]	82	1,002	6–18 (avg. 12)	0–69 (avg. 16)
MORPH-II [16]	13,000	55,134	2–53 (avg. 4)	16–77 (avg. 42)
MORPH-II commercial [16][a]	20,569	78,207	1–76 (avg. 4)	15–77 (avg. 33)
CACD [17]	2,000	163,446	n.a. (avg. 81)	16–62 (n.a.)
LEO_LS[b]	5,636	31,852	4–20 (avg. 6)	12–69 (avg. 31)
PCSO_LS[b]	18,007	147,784	5–60 (avg. 8)	18–83 (avg. 35)

[a] This largest version of MORPH-II only has 317 subjects with at least 5 images acquired over at least 5 years.
[b] The longitudinal face image databases used in this study (details in Sec. 3).

time between face images of the same subject is less than one year, and (ii) it only involves 351 subjects. Poh *et al.* [19] utilized regression models to estimate subject-specific biometric (face and speech) performance trends over time, but the database used only contains 150 subjects and the elapsed times are less than two years. 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.

The two main databases used for research on facial aging, including automatic age estimation, age progression, and age-invariant face recognition, are FG-NET [15] and MORPH [16]. Panis *et al.* [20] provide a recent overview of research that has utilized the FG-NET database. While the public release of these databases greatly encouraged progress in these areas, the databases 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 5 years).[4] The Cross-Age Celebrity Dataset (CACD) [17] was recently released, containing 163,446 images of 2,000 celebrities across 10 years. However, because the images were downloaded from the web (via Google search), the unconstrained quality makes it difficult to statistically model the effects of facial aging. Variations in pose, illumination, expression, etc., may largely influence the trends in similarity scores. Such covariates are difficult to quantify in order to "tease out" these effects from the longitudinal effects, so standardized imaging (near-frontal, neutral expression, uniform illumination) is preferable for the longitudinal study conducted in this paper. Relatively constrained images, such as mugshots, help to ensure that other effects, such as PIE variations, are captured in the noise term in the statistical models. For the above reasons, our longitudinal analysis utilizes two new longitudinal face databases, detailed in Section 3.

3 Longitudinal Face Databases

Operational face image datasets maintained by government and law enforcement agencies can contain longitudinal records of individuals of magnitudes that are infeasible to collect in laboratory settings (*e.g.*, elapsed times over 10+ years). 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 repeat criminal offenders. The sources of face images in our longitudinal analysis are mugshot bookings. While we acknowledge that lifestyle factors (*e.g.*, drug[5] and alcohol use, trauma, etc.) may increase aging rates for some individuals in this population (adult repeat criminal offenders), these accelerated agers are expected to be outliers in the statistical models in our analysis; the overall trends should be relatively robust to this factor. Additionally, we were not able to access any other longitudinal face data. We did attempt to use longitudinal face images from the State Department visa databases. However, we discovered that roughly 5% of genuine face

4. Images in FG-NET are relatively unconstrained (scanned from personal photo collections), while the MORPH databases are mugshots, similar to LEO_LS and PCSO_LS used in this paper but with different database properties (see Table 2).

5. See Yadav *et al.* [21] for work specifically on the effects of drug abuse on face recognition performance.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

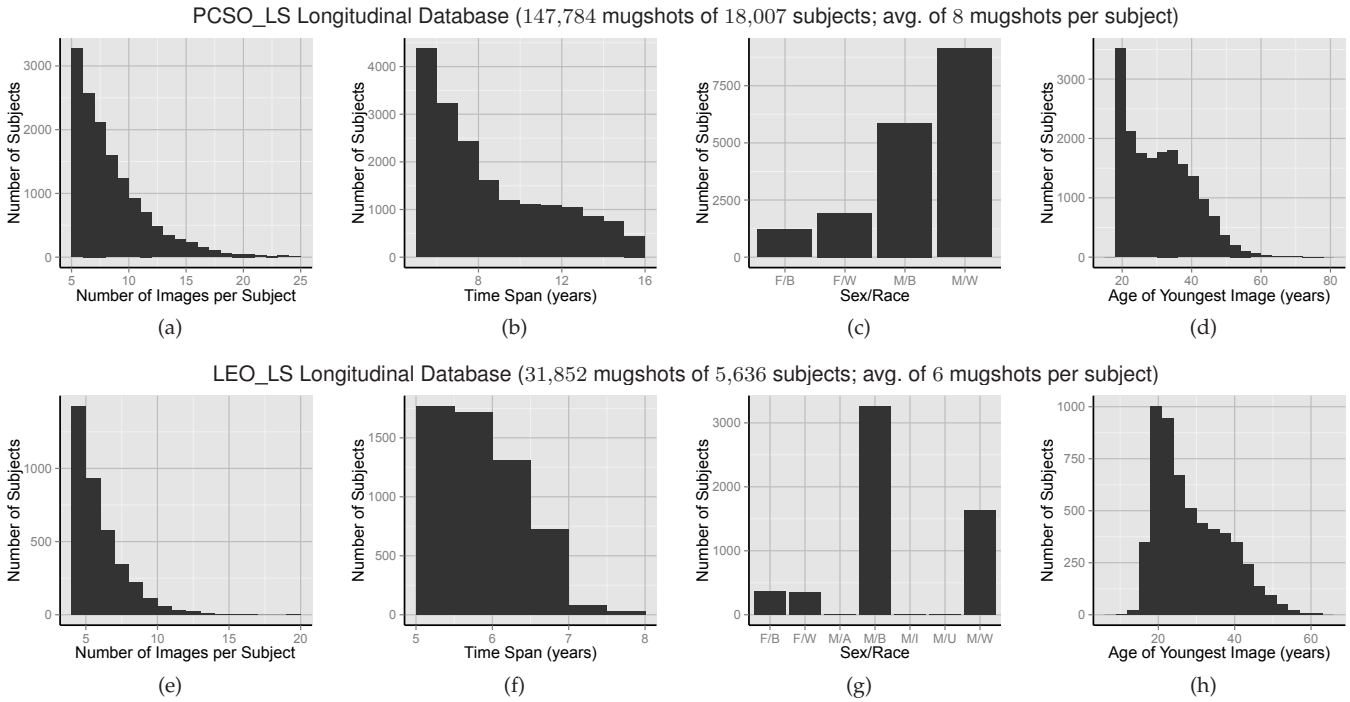


Fig. 2: Statistics of the two longitudinal face image databases (PCSO_LS and LEO_LS) used in this study. (a) and (e) Number of face images per subject, (b) and (f) the time span of each subject (*i.e.*, the number of years between a subject's youngest and oldest face image acquisitions), (c) and (g) demographic distributions of sex (male, female) and race (white, black, Asian, Indian, unknown), and (d) and (h) the age of the youngest image of each subject (in years).

images were duplicate photo submissions (*e.g.*, an individual reuses the same photo for a visa renewal application), so the corresponding inaccurate age information rendered it unsuitable for longitudinal study.

The two databases used in this longitudinal study (LS), denoted LEO_LS and PCSO_LS, are subsets of subjects and images from two larger mugshot databases initially consisting of 3.7 and 1.5 million images, respectively. The following criteria were used to compile the subsets: (i) Each subject has at least 4 (LEO_LS) or 5 (PCSO_LS) face images that were (ii) acquired over at least a 5 year time span, and (iii) each pair of consecutive images is time-separated by at least one month. Database statistics are shown in Fig. 2.

The facial variations in the PCSO_S and LEO_LS databases are well-controlled because the mugshots adhere to standards similar to those detailed in the ANSI/NIST-ITL 2011 face image standards.[6] The standards specify that mugshots should be captured at frontal pose, with neutral expression, uniform illumination, and a background set to 18% gray, for examples. Because these databases are both from operational sources, some confounding factors are still present, such as minor pose and expression variations (see Fig. 6). We also observed rare occurrences of facial occlusions or injury, as shown in Fig. 4, but have retained such images in this study.

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. Since human labeling errors pertaining to demographic attributes and subject ID can be inadvertently introduced in large-scale legacy databases, we determine the sex, race, and date of birth of a subject as the majority vote from each subject's records to ensure consistent labels within each subject. Identifying all such errors was not feasible due to the large size of these databases, but a cursory examination of the PCSO_LS database revealed 134 subject records that contained multiple identities (Fig. 3). These subject records were removed from our study.

6. https://www.nist.gov/itl/iad/image-group/ansinist-itl-standard-history

(a) (b) (c)

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

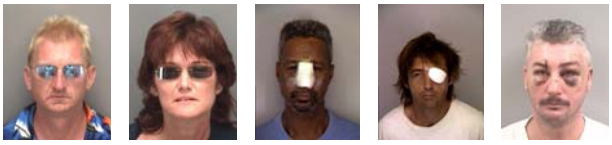


Fig. 4: Examples of facial occlusions (sunglasses, bandages, and bruises) in the PCSO_LS face database.

3.1 LEO_LS Face Database

The LEO_LS database contains 31,852 images of 5,636 subjects from an operational dataset of law enforcement images.

TABLE 3: Overall true accept rates (TARs) at fixed false accept rates (FARs) for various face matchers on the PCSO_LS and LEO_LS databases.

		0.01% FAR	0.1% FAR	1% FAR
PCSO_LS	**COTS-A**	**94.98**	**97.83**	**99.14**
	PittPatt	41.54	58.65	78.30
LEO_LS	**COTS-B**	**99.35**	**99.66**	**99.84**
	COTS-2	90.62	94.96	97.92
	COTS-3	78.97	86.87	93.49
	COTS-4	96.68	98.47	99.31

Each subject has an average of 6 images over an average time span of 5.8 years (maximum of 8 years). Demographic makeup of the LEO_LS database includes 2,009 white and 3,627 black subjects where 4,922 subjects are males and 714 are females. Subjects in LEO_LS are primarily adults, but there are 656 images of 369 subjects that are younger than 18 years-old; these may be juvenile[7] arrests or they could be data entry errors. Due to privacy considerations, we only have access to the comparison scores (both genuine and impostor), so we cannot show face images from this database.

3.2 PCSO_LS 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. Each subject has an average of 8 images over an average time span of 8.5 years (maximum of 16 years). Demographic makeup of the PCSO_LS database includes 11,002 white and 7,004 black subjects where 14,882 subjects are males and 3,124 are females. Example face images from PCSO_LS are shown in Fig. 6. Each booking record in PCSO_LS contains both the date of birth and the date of arrest (actual dates were unavailable for LEO_LS, only the ages were provided to us).

3.3 Face Comparison Scores

Face comparison scores (similarities) were obtained from various commercial face matchers with the aim of evaluating current state-of-the-art longitudinal performance. Two matchers were applied to the PCSO_LS database, and comparison scores were obtained from four different matchers for the LEO_LS database.[8] As shown in Table 3, COTS-A and COTS-B were the overall most accurate matchers. Due to space limitations, longitudinal results are only reported for COTS-A and COTS-B throughout the remainder of the paper. COTS-A and COTS-B were both among the top-3 performers in the FRVT 2013 [1].

The original mugshot images were input to each COTS matcher, and a total of 26,216 and 129,773 genuine scores were computed for the LEO_LS and PCSO_LS databases, respectively, under the scenario where each subject's set of face images are compared to his/her enrollment image. Genuine comparison scores, s_{ij}, between the 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 of 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 mixed-effects regression models as quantifying the change in genuine scores as β standard deviations per year. Fig. 7 shows the distributions of COTS-A and COTS-B standardized genuine scores.

The response variable for all mixed-effects models in this study are standardized *genuine* comparison scores. However, to evaluate face recognition performance, trends in genuine scores should be considered in context with an impostor distribution. For both the LEO_LS and PCSO_LS databases, we computed all possible impostor scores (5.5 million and 11.1 billion, respectively) to calculate thresholds at different fixed FAR values. The threshold at 0.01% FAR, for example, is used to determine when genuine scores drop below the threshold, causing false rejection errors.

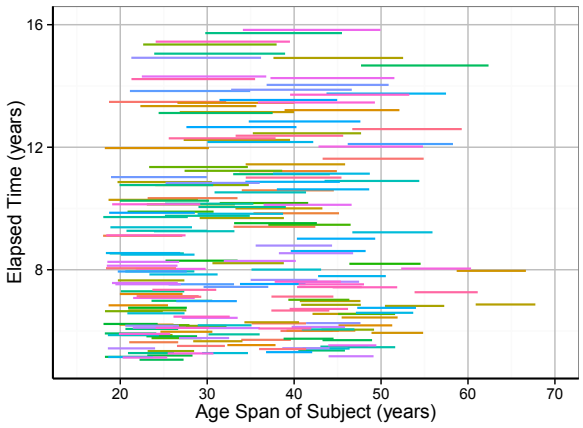


Fig. 5: Age distribution of a random sample of 200 subjects from the PCSO_LS database. Each line denotes the age span of a subject (*i.e.,* age of youngest image to age of the oldest image), separated along the y-axis by the elapsed time for each subject (*i.e.,* the length of the age span).

4 Mixed-Effects Models

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 [22], [23]. In our case, face images 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). Mixed-effects models enable analysis of variation in the response (here, standardized face comparison scores) that occurs at different levels of the data hierarchy.

7. In the United States, a juvenile is typically under the age of 17.

8. Comparison scores and ancillary information (sex, race, age) for the LEO_LS face image database were provided by the Image Group, National Institute of Standards and Technology (NIST), http://www.nist.gov/itl/iad/ig/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

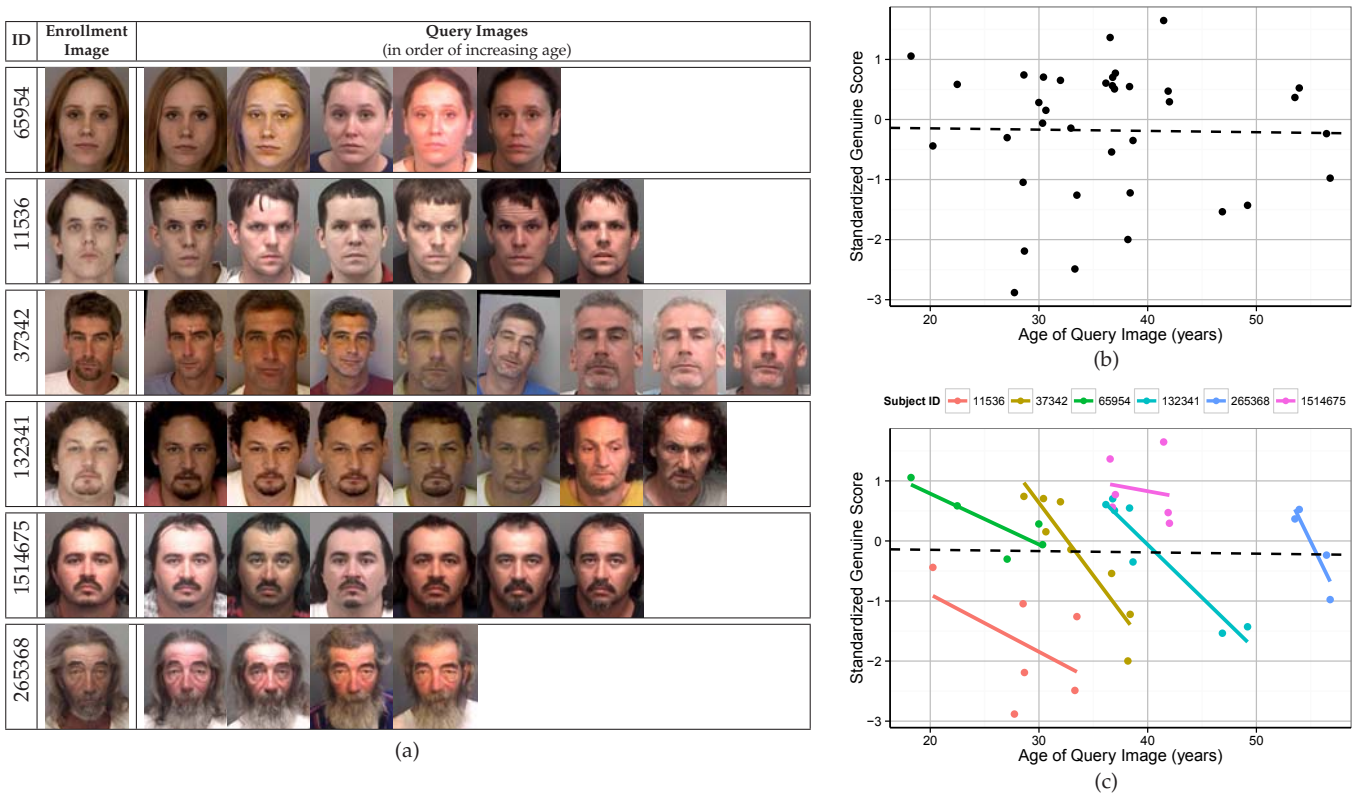


Fig. 6: An example of cross-sectional vs. longitudinal analysis. (a) Face images of six example subjects from the PCSO_LS database. The enrollment face image (leftmost column) is the youngest image of each subject, and all query images are in order of increasing age. In this study, genuine similarity scores are computed by comparing the query images of each subject to his/her enrollment image. In (b), a cross-sectional approach (ordinary least squares (OLS) linear regression) is applied, which incorrectly assumes that all the scores are independent. In (c), OLS is instead applied six times, separately to each subject's set of scores. The slope estimated by cross-sectional analysis (black dotted line) is much flatter than the slopes of subject-specific trends in (solid colored lines in (c)). The longitudinal analysis in this paper utilizes mixed-effects models, which provide "shrunken" OLS estimates for each subject, where the OLS trends shrink towards a population-mean trend [22], [23], further accounting for the correlation that exists between scores from the same subject.

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 nicely structured because it is difficult (and expensive) to collect, or it must be analyzed retrospectively, as is the case with the mugshot databases used in this study. 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. Fig. 2 shows that subjects in the LEO_LS and PCSO_LS databases have anywhere from 4 to more than 20 mugshots, and Fig. 5 shows that the age spans of the subjects are highly unstructured.

Mixed-effects models can handle imbalanced and time-unstructured data and are preferable over other approaches because they model both the mean response (fixed effects define the population-mean trend), as well as the covariance structure (random effects allow deviations of individuals from the population-mean). In longitudinal data, this covariance structure has a complicated form which stems from the fact that error terms are *not* independent (as is assumed in standard linear regression). The remainder of this section provides details of the models and covariates of interest.

4.1 Model Formulations

Given n_i face images of subject i, let AGE_{ij} denote the absolute age of the ith individual for the 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 image acquisitions

TABLE 4: 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	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
CA	$y_{ij} = \varphi_{0i} + \varphi_{1i}AGE_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + b_{1i}$
D	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{02}AGE_{ie}^2 + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}AGE_{ie} + b_{1i}$
E	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{02}AGE_{ie}^2 + \beta_{03}M_i + \beta_{04}B_i + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}AGE_{ie} + \beta_{12}M_i + \beta_{13}B_i + b_{1i}$
Q	$y_{ij} = \varphi_{0i} + \varphi_{1i}\triangle T_{ij} + \varphi_{2i}Q_{ij} + \varphi_{3i}Q_{ij}\triangle T_{ij} + \varepsilon_{ij}$	$\varphi_{0i} = \beta_{00} + \beta_{01}Q_{ie} + b_{0i}$	$\varphi_{1i} = \beta_{10} + \beta_{11}Q_{ie} + b_{1i}$, $\varphi_{2i} = \beta_{20} + \beta_{21}Q_{ie} + b_{2i}$, $\varphi_{3i} = \beta_{30}$

$\triangle T_{ij}$: elapsed time (years) between the enrollment and jth face image of subject i;
AGE_{ie}: age (years) of subject i in her enrollment face image; AGE_{ij}: age (years) of subject i in her jth face image;
M_i: binary indicator of subject sex ($M_i = 1$ if male, 0 if female); B_i: binary indicator of subject race ($B_i = 1$ if black, 0 if white)
Q_{ie}: quality (*e.g.,* frontalness or interpupillary distance) of the enrollment image of subject i;
Q_{ij}: quality (*e.g.,* frontalness or interpupillary distance) of the jth query image of subject i

and (ii) the *absolute ages* of the subject in the two face images being compared. Below, we discuss mixed-effects models which include these and other covariates.

4.1.1 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 $\triangle T_{ij}$:

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

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} (level-1 residual variation). 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 effects [24], [25] (some subjects generally have higher or lower scores). Fig. 6 shows six example subjects from the PCSO_LS database at different ages, with their subject-specific trends in genuine scores over time shown in Fig. 6c.

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 distributional assumptions of these two 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.

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 comparison scores at a projected baseline of zero years elapsed time. 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 parameters (*i.e.,* differences in the trends of individuals in the population).

4.1.2 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 also be a function of absolute age. If we add the age of the enrollment image to (5):

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

Because AGE_{ie} is a *fixed* effect for each subject (*time-invariant*), 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.[9] The population-mean trend for (6) is:

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

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, $f(AGE_{ij}, AGE_{ie})$:

$$\mathbf{E}(y_{ij}) = \beta_{00} + \beta_{01}AGE_{ie} + \beta_{10}AGE_{ij}. \quad (8)$$

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 aging [26].[10] The significance of subject age at enrollment in (8) is tested with the null hypothesis of $H_0 : \beta_{01} = 0$, whereas *restricted* inference is needed to test significance in (7) because the null hypothesis must instead be $H_0 : \beta_{01} = \beta_{10}$.

The relationship between these two models (CT and CA) 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 [23], [26], [27], [28]. Typically, the time-varying covariate is "centered" on subject-specific means, so as to remove between-subject variation at level-1 of the model.

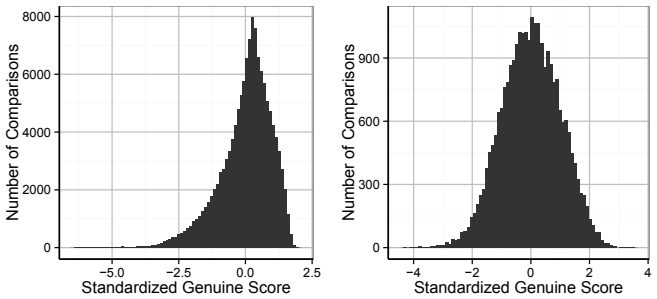
4.2 Model Comparison and Evaluation

The goal of statistical modeling is to find a model that includes substantive predictors and excludes unnecessary ones (parsimony). A common approach is to fit increasingly complex models to successively evaluate the impact of adding different covariates [22]. Models can be compared using goodness-of-fit measures based on log-likelihood statistics: deviance, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). 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 zero,

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

9. Comparing all images of a given subject to her 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.

10. 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.



(a) PCSO_LS (b) LEO_LS

Fig. 7: Distributions of standardized genuine comparison scores from the two longitudinal face databases used in this study: (a) COTS-A on PCSO_LS and (b) COTS-B on LEO_LS. There are a total of 129,773 and 26,216 genuine scores in (a) and (b), respectively.

Deviance can be used to compare nested models (*i.e.,* the more complex model can be reduced to the simpler model by placing constraints on its parameters) that are fit to the same data. To compare non-nested models, AIC and BIC penalize the log-likelihood based on the complexity of the models[11] and the sample size. Smaller values indicate better fit for all three goodness-of-fit measures.[12]

Further comparisons of models depend on whether the successive model has included a time-invariant (*e.g.,* sex, race) or time-varying (*e.g.,* face image quality) covariate to the baseline model. For both cases, pseudo-R^2 statistics can be used to measure the proportional reduction in level-2 variance (σ_0^2, σ_1^2) and level-1 residual variance (σ_ε^2) attributable to inclusion of time-invariant and time-variant covariates, respectively.

5 Results

We first focus on analysis of the PCSO_LS database, starting with simpler models (*i.e.,* Model A and BT) and progressing to more complex models including covariates for subject sex/race and face image quality. We then present results for the LEO_LS database. Recall that models are discussed in Section 4 and equations are provided in Table 4. All models in our analysis are fit with full maximum likelihood (ML) estimation via iterative generalized least-squares (GLS) using the **lme4** package (v1.1-9) [29] for R (v3.2.2).

5.1 Model Assumptions

While mixed-effects models are capable of handling non-Gaussian response distributions (*e.g.,* COTS-A genuine scores in Fig. 7a), the error terms must follow Gaussian distribution. Fig. 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

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

12. For AIC and BIC, the magnitude of the reduction in model fit is difficult to interpret.

TABLE 5: Bootstrap results for mixed-effects models on the PCSO_LS database and COTS-A genuine scores.

		Model A	Model BT	Model CT	Model D
FIXED EFFECTS (95% CONFIDENCE INTERVALS):					
INTERCEPT	β_{00}	0.0274 (0.0171, 0.0376)	0.6734 (0.6624, 0.6849)	0.7226 (0.6905, 0.7556)	0.5158 (0.4073, 0.6239)
TIME	β_{10}		−0.1364 (−0.1379, −0.1349)	−0.1364 (−0.1379, −0.1349)	−0.1372 (−0.1426, −0.1316)
AGE GROUP	β_{01}			−0.0016 (−0.0027, −0.0006)	0.0120 (0.0047, 0.0189)
AGE GROUP × TIME	β_{11}				0.0000# (−0.0002, 0.0002)
AGE GROUP^2	β_{02}				−0.0002 (−0.0003, −0.0001)
VARIANCE COMPONENTS:[a]					
Level-1 Residual	σ^2_ε	0.6076	0.3912	0.3912	0.3912
Random Intercepts	σ^2_0	0.3841	0.3243	0.3239	0.3231
Random Slopes	σ^2_1		0.0028	0.0028	0.0028
Covariance	σ_{01}		−0.0039	−0.0039	−0.0038
GOODNESS-OF-FIT:[b]					
	AIC	333433	287016	287006	286985
	BIC	333462	287074	287075	287073
	Deviance	333427	287004	286992	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.

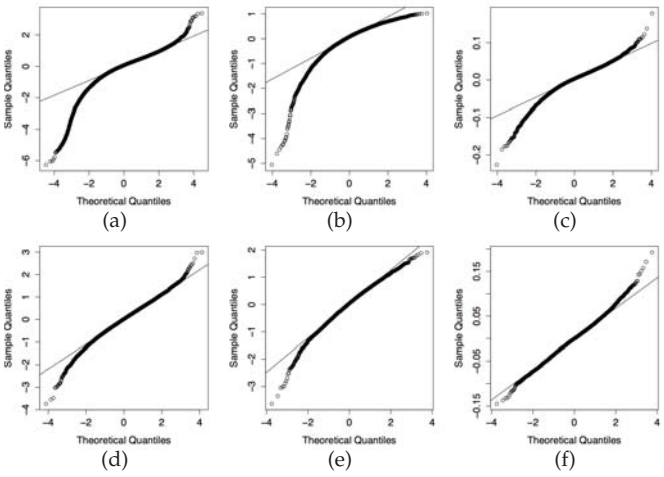


Fig. 8: Normal probability plots of ((a) and (d)) level-1 residuals, ε_{ij}, and level-2 random effects for ((b) and (e)) intercepts, b_{0i}, and ((c) and (f)) slopes, b_{1i}, from Model BT on the PCSO_LS and LEO_LS databases (top and bottom rows, respectively). Departure from normality at the tails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

(Figs. 8b, 8c). This behavior was observed for other models as well, precluding the use of standard errors for formal hypothesis tests of parameters [30].

When parametric model assumptions are violated, it is common to resort to non-parametric bootstrap to establish confidence intervals for the parameter estimates, as followed in Yoon and Jain [9]. Hence, for the PCSO_LS database, we conduct a non-parametric bootstrap by *case resampling* [30]; 1,000 bootstrap replicates 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.[13] Table 5 gives the bootstrap parameter estimates (with 95% confidence intervals), variance components, and goodness-of-fit for the models in Table 4.

5.2 Unconditional Means Model (Model A)

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 4, and with composite form of $y_{ij} = \beta_{00} + b_{0i} + \varepsilon_{ij}$, 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 5 indicate that between-subject differences in genuine scores (*i.e.,* biometric zoo) account for 38.7% ($\rho = 0.3873$) of the total variation in genuine scores from the PCSO_LS database. Baseline goodness-of-fit measures are also shown in Table 5.

5.3 Unconditional Growth Model (Model BT)

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 5 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

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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

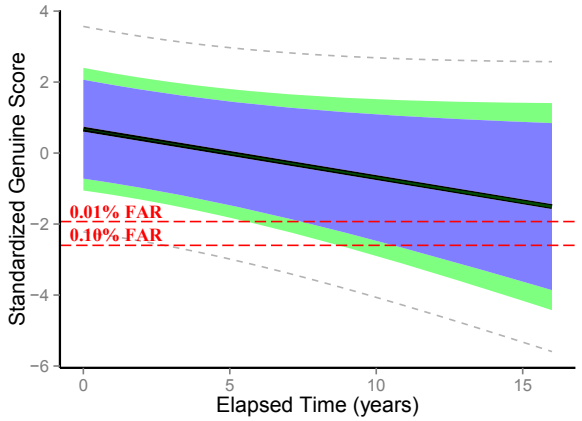


Fig. 9: Results from Model BT on COTS-A genuine scores from the PCSO_LS database. The bootstrap-estimated population-mean trend is shown in black (bootstrap confidence intervals are too small to be visible). The blue and green bands plot regions of 95% and 99% confidence, respectively, for subject-specific variations around the population-mean trend. Grey dotted lines additionally add one standard deviation of estimated residual variation, σ_ε. Hence, Model BT estimates that 95% and 99% of the subject trends fall within the blue and green bands, but scores can vary around their trends, extending to the grey dotted lines. Thresholds at 0.01% and 0.1% FAR for COTS-A are shown as dashed red lines.

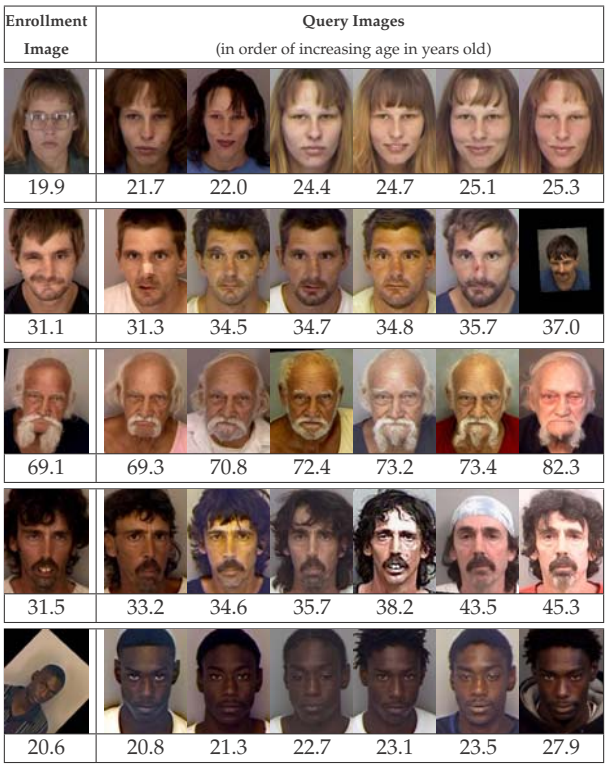


Fig. 10: Example outlier subjects, *i.e.*, subjects whose subject-specific trends, estimated by Model BT, significantly deviate from the spread of the population in the PCSO_LS database. All images were aligned using COTS-A eye locations.

variation in a given subject's genuine scores around his/her own average genuine score.[14]

Longitudinal change estimated by Model BT implies that the *population-mean trend* will drop below the thresholds for 0.01% and 0.1% FAR after 19.1 and 24.0 years elapsed time, respectively, but this only provides insight into performance on subjects in the population with average (or higher) genuine scores over time. A reliable face recognition system must be able to recognize much more than just 50% of the population it encounters, so we are also interested in the spread of the population around the population-mean trend. Do all subjects closely follow the population-mean trend, or is there large variability between subjects? Do biometric zoo effects extend to rates of change over time?

Using the estimated variance components for slopes and intercepts (σ_0^2, σ_1^2, and σ_{01}), we compute a 2D confidence ellipse (random effects are assumed to be 2D Gaussian distributed) to define a region that contains, for example, 95% of the estimated subject-specific parameters. In order to translate from the 2D space of intercepts and slopes to obtain a confidence region for genuine scores versus elapsed time, we sample 100 combinations of intercept and slope parameters along the contour of the confidence ellipse, compute the predicted genuine scores for each of the 100 trends, and define the confidence region as between the minimum and maximum predicted scores for different values of elapsed time. Results are shown in Fig. 9.

From the confidence bands of subject variations in Fig. 9, we infer that genuine scores for 99% of the population will remain above the threshold at 0.01% FAR for up to approximately 5.5 years elapsed time, which reduces to 95% of the population after 7 years (*i.e.*, false reject errors would occur, on average, for 5% of subjects after 7 years since enrollment). Similarly, at a higher FAR of 0.1%, 99% of subjects can be recognized up to 8.5 years elapsed time, which reduces to 95% after 10.5 years. Fig. 10 shows face images from six example outlier subjects whose estimated trends lie outside the 99% region of confidence due to extreme intercepts and/or slopes; subjects significantly deviate from the population spread due to alignment errors, face quality issues (illumination, facial occlusion), and changes to facial hair, for example.

5.4 Age at Enrollment (Models CT and D)

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). The significance of the AGE_{ie} term in Model CT suggests a *negative* linear relationship between age at enrollment and genuine scores, but the magnitude of β_{01} is relatively small.

To further test the complexity of the effects of age at enrollment, we add additional terms associated with AGE_{ie}, resulting in Model D (see Table 4). 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 lower genuine scores, on average, and steeper negative rates of change. Table 5 shows that

14. Using pseudo-$R^2 = (\sigma_\varepsilon^2(A) - \sigma_\varepsilon^2(BT))/\sigma_\varepsilon^2(A)$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

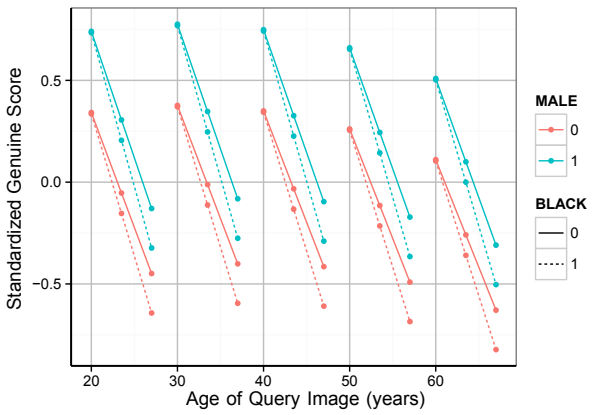


Fig. 11: Model E fit to COTS-A genuine scores from the PCSO_LS database. Population-mean trends are plotted by subject demographics of sex and race and for five different ages at enrollment (20 to 60 years). Each trend line represents seven years of elapsed time after enrollment. For example, the solid blue line beginning at $AGE_{ij} = 20$ years represents the average decrease in genuine scores for white males enrolled at age 20 with query images until age 27.

the interaction term $AGE_{ie} \times \triangle T_{ij}$ in Model D is not significantly different from zero 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. 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 20 year-old and a 30 or 50 year-old (at enrollment) is equivalent to only 7 and 5 *months* of elapsed time (within-subject longitudinal change), respectively.

5.5 Sex and Race (Model E)

Model E in Table 4 is used to test the effects of subject sex and race. First, we observed that Model E results in better model fit than Model D (deviance for Model E is 285,712 compared to 286,967 for Model D). The main effect of subject sex is statistically non-zero at significance level of 0.05, but the main effect of subject race is not (the 95% bootstrap confidence interval contains 0). Male genuine scores at baseline ($\triangle T_{ij} = 0$ years) are 0.3987 standard deviations higher than female scores. Significant interactions with elapsed time indicate that rates of change in genuine scores depend on both sex and race; population-mean slopes are -0.0113 and -0.0267 standard deviations steeper for males and black subjects, respectively. Population-mean trends separated by subject demographics are shown in Fig. 11 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

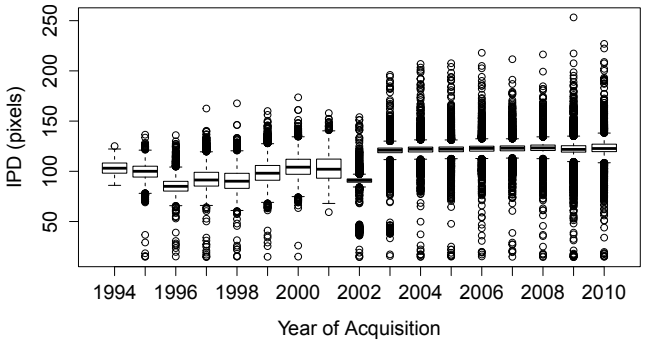


Fig. 12: A boxplot of interpupillary distances (IPDs) versus year of acquisition shows that mean IPDs systematically changed over time for the PCSO_LS database, likely due to booking stations adhering to face imaging standards only in more recent years.

TABLE 6: Bootstrap results for mixed-effects models with elapsed time and face quality covariates for the PCSO_LS database and COTS-A genuine scores.

	Model QF	Model QI	Model QFI
σ_ε^2	0.3302	0.3539	0.3218
AIC	275108	281296	273643
BIC	275283	281471	273848
Deviance	275072	281260	273601

overall. Fig. 11 also shows that the differences between subject race are minor compared to differences between males and females.

5.6 Face Image Quality (Model Q)

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 5 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 $\sqrt{\sigma_\varepsilon^2}/\beta_{10} = \sqrt{0.3912}/-0.1372$). This is visually shown by the dotted grey lines in Fig. 9.

Level-1 residual variation can only be reduced by level-1 time-varying covariates (*i.e.,* image-specific); 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 range of the frontal score has little meaning, since its computation is proprietary. We standardize (z-score) 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.

After finding 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), we then added the quality measures to Model BT, resulting in Model Q in Table 4. Table 6 gives

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 7: 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

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 7 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 both the level-1 residual variation and values of goodness-of-fit values.

The values of 100 and 120 pixels for IPD in Table 7 were chosen because we observed systematic changes in IPDs over time (see Fig. 12); in particular, mean IPD varies around 100 pixels from 1994–2002 but increases to a consistent ∼120 pixels starting in 2003. This observation, along with correspondence with Pinellas County Sheriff's Office, suggests that booking agencies began to adhere to imaging standards around this time. To investigate whether this aspect of the data confounds the estimation of longitudinal effects (face 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.

5.7 LEO_LS Database

Table 8 gives results for the models in Table 4 fit to COTS-B genuine scores from the LEO_LS database. Fixed-effects parameter estimates are given with standard errors; bootstrapping was not conducted for LEO_LS models because the error terms better follow Gaussian distributions (see Fig. 8). Model results are summarized as follows.

Model A estimates that 40% of the total variation in genuine scores is due to between-subject differences. 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 $\beta_{10} = -0.1699$ standard deviations. From the confidence bands of subject variations in Fig. 13 (estimated by Model BT), we infer that genuine scores for 99% of the population will remain above

TABLE 8: Mixed-effects model results for the LEO_LS database and COTS-B genuine scores.

		Model A	Model BT	Model CT	Model D
Fixed Effects (standard errors):					
(Intercept)	β_{00}	0.0037	0.5395	0.5468	0.0894
		(0.0098)	(0.0127)	(0.0325)	(0.1057)
Time	β_{10}		−0.1699	−0.1699	−0.1980
			(0.0023)	(0.0023)	(0.0076)
Age Group	β_{01}			−0.0003	0.0346
				(0.0011)	(0.0068)
Age Group × Time	β_{11}				0.0010
					(0.0003)
Age Group^2	β_{02}				−0.0006
					(0.0001)
Variance Components:					
Level-1 Residual	σ^2_ε	0.5985	0.4276	0.4276	0.4275
Intercepts	σ^2_0	0.4009	0.5543	0.5542	0.5516
Slopes	σ^2_1		0.0059	0.0058	0.0058
Covariance	σ_{01}		−0.0317	−0.0317	−0.0316
Goodness-of-Fit:					
	AIC	68705	62647	62649	62606
	BIC	68730	62697	62707	62679
	Deviance	68699	62635	62635	62588

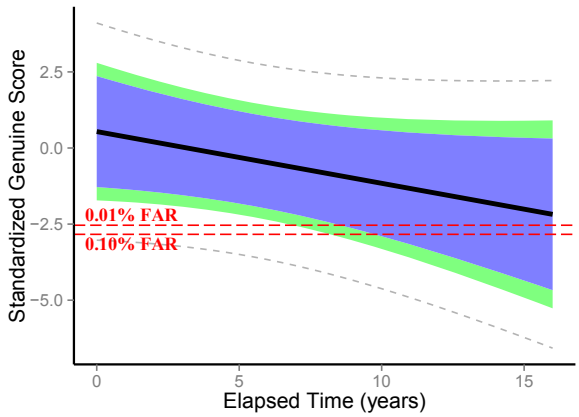


Fig. 13: Results from Model BT on COTS-B genuine scores from the LEO_LS database. The population-mean trend is shown in black. The blue and green bands plot regions of 95% and 99% confidence, respectively, for subject-specific variations around the population-mean trend. Grey dotted lines additionally add one standard deviation of estimated residual variation, σ_ε. Hence, Model BT estimates that 95% and 99% of the subject trends fall within the blue and green bands, but scores can vary around their trends, extending to the grey dotted lines. Thresholds at 0.01% and 0.1% FAR for COTS-B are shown as dashed red lines.

the threshold at 0.01% FAR for up to approximately 6.5 years elapsed time, which reduces to 95% of the population after 8.5 years (*i.e.*, false reject errors would occur, on average, for 5% of subjects after 8.5 years since enrollment). Similarly, at a higher FAR of 0.1%, 99% of subjects can be recognized up to 8.0 years, which reduces to 95% after 9.5 years elapsed time.

Although the between-subject effect of age at enrollment (β_{01}) is significantly different from β_{10} in Model CT, the effect is not significantly different from zero, indicating that there is no linear relationship between subject enrollment age and average genuine scores. However, additional terms

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2017.2652466, IEEE Transactions on Pattern Analysis and Machine Intelligence

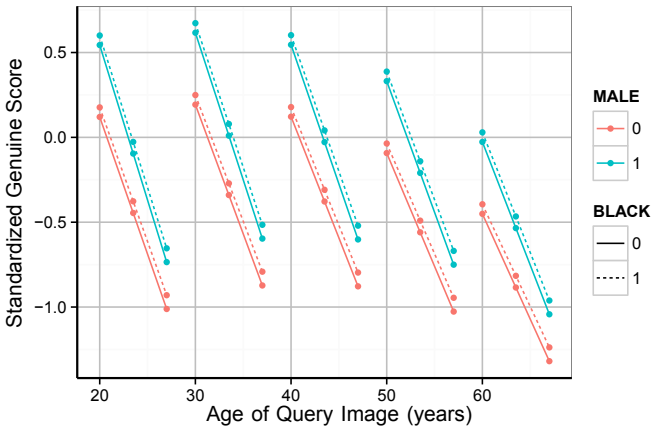


Fig. 14: Model E for COTS-B genuine scores from the LEO_LS database. Population-mean trends are plotted by subject demographics of sex and race, in addition to five different ages at enrollment (20 to 60 years). Each trend line represents seven years of elapsed time since enrollment. For example, the solid blue line beginning at $AGE_{ij} = 20$ years represents the average decrease in genuine scores for white males enrolled at age 20 with query images until age 27.

involving AGE_{ie} result in significant effects of enrollment age in Model D. The significant β_{02} coefficient indicates a downward quadratic relationship between age at enrollment and average genuine scores (similar to COTS-A on PCSO_LS). Furthermore, the significant interaction term $AGE_{ie} \times \triangle T_{ij}$ indicates that longitudinal change in scores tends to vary with subject's age at enrollment; a 10-year increase in subject age results in a longitudinal slope that is $\beta_{11} = -0.0098$ standard deviations steeper. Population-mean rates of change range from -0.1784 to -0.1490 standard deviations per year for subjects with age at enrollment of 20 to 50 years (calculated as $\beta_{10} + \beta_{11} AGE_{ie}$). Recall that age at enrollment had no effect on rates of change for COTS-A on PCSO_LS.

Model E results indicate that intercepts are 0.0565 and 0.4238 standard deviations higher for black and male subjects, respectively (so, black-male subjects have intercepts that are 0.4803 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 *steeper* (*i.e.*, more negative) than for females. These population-mean trends are shown in Fig. 14 for different ages at enrollment. Fig. 14 also shows that the differences between subject race are minor compared to differences between males and females, as was also the case for COTS-A on the PCSO_LS database.

6 Conclusions

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, avg. 8 images per subject over avg. 8.5 years) and LEO_LS ($31{,}852$ images of $5{,}636$ subjects, avg. 6 images per subject over avg. 5.8 years), where each subject has at least four face images acquired over at least a five-year time span. Linear mixed-effects regression 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 similarity scores were obtained from state-of-the-art COTS matchers for both the PCSO_LS and LEO_LS databases. Based on our analysis, we make the following observations (statements apply to both databases and matchers):

✦ Population-mean trends indicate that genuine scores significantly decrease with increasing elapsed time between enrollment (gallery) and query (probe) images, as expected. However, population-mean trends (average genuine scores) do not fall below thresholds at 0.01% FAR until after 15 years elapsed time. This suggests that in a practical application, an average individual's genuine scores decrease at a rate that will not affect the recognition accuracy at 0.01% FAR until more than 15 years since enrollment.

✦ Significant subject-specific variability around the population-mean trends is observed; genuine scores for some subjects decline at much faster rates than the population-mean. Analysis of the estimated variance in subject-specific parameters (intercepts and slopes) allowed for estimation of subject-based accuracies (*i.e.*, how many subjects are estimated to be falsely rejected, rather than standard image-based accuracy calculations). For example, the models estimate that genuine scores for 99% of the population will remain above the threshold at 0.01% FAR until 6.5 years elapsed time for PCSO_LS and 5.5 years for LEO_LS. Other calculations (e.g. 95% of the population) are also within approximately one year for both databases.

✦ Subject-specific variance in rates of change (*i.e.*, linear slopes) is only marginally attributable to subject age at enrollment, sex, and race. Subject sex was the most significant factor for between-subject differences in genuine scores, with males having significantly higher genuine scores than females. The magnitude of the difference suggests that false reject errors may occur approximately two years earlier for females than for males (assuming that a global threshold is used operationally).

✦ 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. The within-subject variability (*i.e.*, level-1 residual variance) is still quite large. All models considered in this study indicate that one standard deviation in genuine scores due to short-term variations (*e.g.*, illumination, hairstyle, etc.) is approximately equivalent to the change in genuine scores due to ± 4 years of elapsed time (for these particular databases and matchers).

Longitudinal analysis, in general, is an important, yet very difficult, problem. To the best of our knowledge, no proper statistical analysis has yet been conducted for studying face recognition performance on a large population over periods of time longer than five years. In this paper, we attempted to analyze the covariates of interest that were available to us (elapsed time, age, sex, race, some measures of quality), 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.). Despite this, 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. Given that the performance of face recognition systems continues to improve, longitudinal analysis should be conducted periodically to reevaluate robustness to facial aging (and other covariates).

Future work includes: (i) Evaluation of face *identification* (both closed-set and open-set) performance over time. Observations about recognition accuracy in this paper apply to *verification* scenarios (*i.e.*, one-to-one comparisons) operating with a global threshold. (ii) Development of a single face quality measure for mugshot type face images. (iii) Longitudinal analysis on different face cropping (particularly, pre-cropped images to exclude most of the hair region) to investigate the impact of changing hairstyle over time.

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 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. Computer Vision Pattern Recognition (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 Trans. 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] 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.
[12] D. Gong, Z. Li, D. Lin, J. Liu, and X. Tang, "Hidden factor analysis for age invariant face recognition," in *Proc. ICCV*, 2013.
[13] 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.
[14] M. Erbilek and M. Fairhurst, "A methodological framework for investigating age factors on the performance of biometric systems," in *Proc. Multimedia and Security*, 2012.
[15] A. Lanitis, C. J. Taylor, and T. F. Cootes, "Toward automatic simulation of aging effects on face images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, Apr. 2002.
[16] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. FGR*, 2006.
[17] B.-C. Chen, C.-S. Chen, and W. H. Hsu, "Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset," *IEEE Trans. on Multimedia*, vol. 17, no. 6, pp. 804–815, Apr. 2015.
[18] 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.
[19] N. Poh, J. Kittler, C.-H. Chan, and M. Pandit, "Algorithm to estimate biometric performance change over time," *IET Biometrics*, vol. 4, no. 4, pp. 236–245, Dec. 2015.
[20] G. Panis, A. Lanitis, N. Tsapatsoulis, and T. F. Cootes, "An overview of research on facial aging using the FG-NET aging database," *IET Biometrics*, May 2015.
[21] D. Yadav, N. Kohli, P. Pandey, R. Singh, M. Vatsa, and A. Noore, "Effect of illicit drug abuse on face recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2016.
[22] J. D. Singer and J. B. Willett, Eds., *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford Univ. Press, Inc., 2003.
[23] G. M. Fitzmaurice, N. M. Laird, and J. H. Ware, *Applied Longitudinal Analysis*, 2nd ed. Hoboken, New Jersey: John Wiley & Sons, Inc., 2011.
[24] 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.
[25] N. Yager and T. Dunstone, "The biometric menagerie," *IEEE Trans. on PAMI*, vol. 32, no. 2, pp. 220–230, Feb. 2010.
[26] 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.
[27] 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.
[28] 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.
[29] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.
[30] R. van der Leeden, F. M. Busing, and E. Meijer, "Bootstrap methods for two-level models," in *Multilevel Conf.*, 1997.

Lacey Best-Rowden received her B.S. degree in computer science and mathematics from Alma College, Alma, Michigan, in 2010. She is currently working towards the PhD degree in the Department of Computer Science and Engineering at Michigan State University, East Lansing, Michigan. Her research interests include pattern recognition, computer vision, and image processing with applications in biometrics. She is a student member of the IEEE.

Anil K. Jain is a University distinguished professor in the Department of Computer Science and Engineering at Michigan State University. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence (1991-1994). He served as a member of the United States Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. He has received Fulbright, Guggenheim, Alexander von Humboldt, and IAPR King Sun Fu awards. He was elected to the National Academy of Engineering in 2016.