To appear in the *Proc. IEEE ICB, 2015*

# A Longitudinal Study of Automatic Face Recognition

Lacey Best-Rowden and Anil K. Jain
Dept. of Computer Science and Engineering
Michigan State University, East Lansing, MI, U.S.A.
{bestrow1,jain}@msu.edu

## 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 changes drastically over time. This heterogeneity (inter-subject variability) suggests the need for a subject-specific aging analysis. In this paper, we conduct such an analysis using a longitudinal database of 147,784 operational mug shots of 18,007 repeat criminal offenders, where each subject has at least five face images acquired over a minimum of five years. By fitting multilevel statistical models to genuine similarity scores from two commercial-off-the-shelf (COTS) matchers, 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 of the scores from the more accurate COTS matcher shows that despite decreasing genuine scores over time, the average subject can still be correctly verified at a false accept rate (FAR) of 0.01% across all 16 years of elapsed time in our database. We also investigate (i) the effects of several other covariates (gender, race, face quality), and (ii) the probability of true acceptance over time.*

## 1. Introduction

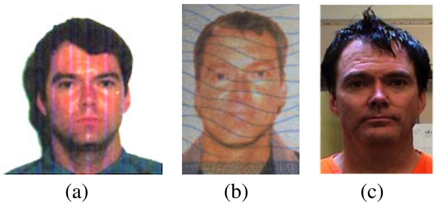


Figure 1. In July of 2014, automatic face recognition matched Neil Stammer's (a) FBI's most wanted photo (mugshot from 1999) to (b) his fraudulent passport photo from 2011. He was then (c) captured after almost 15 years as a fugitive.

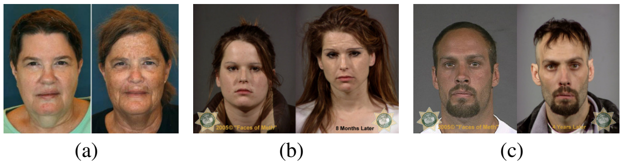
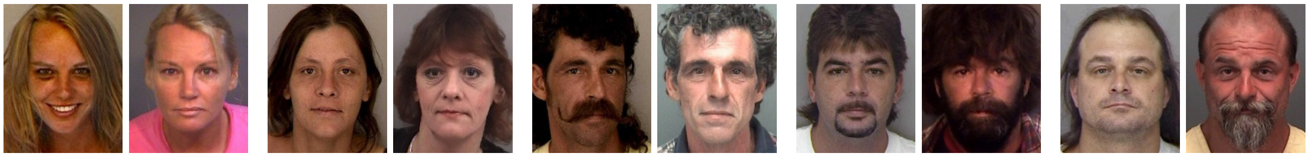


Figure 2. Aging influences facial appearance: (a) Identical twins at age 61 appear to age differently due to smoking and sun exposure. Drastic changes in facial appearance of two individuals after (b) eight months and (c) four years of methamphetamine use.

Studies on the persistence of face recognition performance across large time lapse will be extremely valuable to law enforcement, homeland security, as well as other agencies that use face images for de-duplication or person identification. One notable example of the strength of COTS face matchers is the case of Neil Stammer, a fugitive who who was first arrested in 1999. In January of 2014, facial recognition software returned a match between Stammer's photo on the FBI's most wanted list (his mug shot from 1999, Fig. 1(a)) and the photo from a 2011 passport with a different name (Fig. 1(b)); Stammer had been living in Nepal under a stolen identity.[1] While Stammer's capture is a successful example of face recognition across large time lapse (12 years), we need to determine whether this capability holds across a large population of subjects.

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. In addition to anatomical factors, environmental and/or lifestyle factors also have a significant impact on facial appearance over time. Smoking, sun exposure, and stress levels have been claimed to be plausible explanations for different aging patterns in the faces of

[1] http://www.fbi.gov/news/stories/2014/august/long-time-fugitive-neil-stammer-captured



(a) Ages 30.5 and 39.6 (0.4231)

(b) Ages 32.2 and 40.3 (0.4325)

(c) Ages 39.2 and 48.7 (0.4666)

(d) Ages 29.5 and 38.3 (0.4984)

(e) Ages 39.2 and 48.6 (0.5002)

Figure 3. Genuine face image pairs separated by eight to ten years in the PCSO_LS database. COTS-A similarity scores for each pair are shown in parentheses. The thresholds for COTS-A 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.

identical twins (Fig. 2(a)).[2] Figures 2(b) and (c), show how the use of methamphetamine can drastically alter a person's face over just a short period of time.[3] Due to the cumulative effects of both biological and environmental factors, facial aging affects each individual differently.

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. Figure 3 shows that elapsed time of eight to ten years between two face images can cause a false non-match.

Many prior studies have claimed that face recognition performance decreases as the time interval[4] between two acquisitions of a person's face image increases [8]. However, these studies did not (i) conduct any formal tests of hypotheses, (ii) make any distinction between population trends and subject-specific variations in these trends, and (iii) quantify how much of the change in accuracy is due to other confounding factors such as face quality (*e.g.* facial pose, ambient illumination), gender, age, and race.

In this paper, we apply multilevel statistical models to a large-scale longitudinal database of 147,784 face images of 18,007 subjects. To study the effect of elapsed time, we use two state-of-the-art commercial off-the-shelf (COTS) face matchers to obtain genuine and impostor comparison scores.[5] The contributions of this study are: (i) Provides the largest (to date) statistical analysis of longitudinal effects on face recognition. (ii) Use of statistical models to determine temporal trends in genuine scores and probability of true acceptance with respect to different covariates. (iii) The first study to quantify the variance in *subject-specific* temporal trends in genuine scores. (iv) Conclusion that false rejection rates at 0.01% FAR of one of the COTS matchers remain below 2% up to approximately 10 years time interval for the longitudinal face database used here.

## 2. Related Work

The facial aging process has received a considerable amount of attention with respect to automatic methods for age estimation, age simulation/progression, and age-invariant face recognition (see [9] for a survey). In this work, we are primarily concerned with how facial aging, namely elapsed time, affects the performance of face recognition systems. All of the previous studies on this topic follow a similar approach: (i) partition the database (face pairs) 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. Based on this procedure, the following conclusions have been reported [8]: (i) Face recognition performance decreases as the time elapsed between two images of the same person increases. For example, Klare and Jain report a decrease in TAR @ 1% FAR of 7.7% and 14.3% from 0-1 to 5-10 years time lapse for two COTS matchers on a database of mug shot images [6]. (ii) Faces of younger individuals are more difficult to recognize than faces of older individuals. See, for example, the NIST FRVT 2014 evaluation [5] on seven age groups (baby, kid, pre-teen, teen, young, parents, older) in a visa database of 19,972 subjects.

In these cohort-based approaches, which age group or time lapse partitions are evaluated is often arbitrary and varies from one study to another, thereby, making comparisons between studies difficult [2]. Furthermore, cohort-based analysis with summary statistics does not investigate whether age-related performance trends are due to adverse effects on the genuine distribution, the impostor distribution, or both.

Multilevel (hierarchical or mixed-effects) statistical models have been used for determining important factors to explain the performance of face recognition systems. Beveridge *et al.* [1] apply generalized linear mixed models to verification decisions made by three algorithms from the FRGC Exp. 4. In addition to eight levels of FAR as a covariate, they analyze gender, race, image focus, eye distances, age, and elapsed time, but the maximum elapsed time is less than one year, and the study only involves 351 subjects.

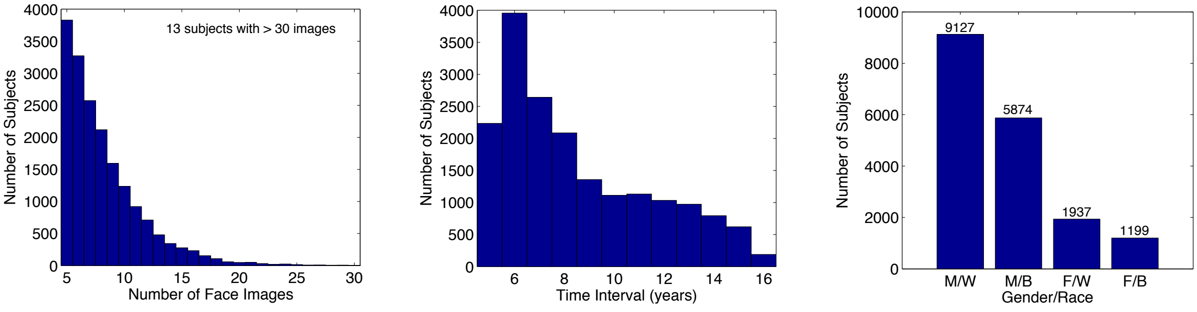
The use of multilevel statistical models for large-scale

[2] http://www.nbcnews.com/id/33385839/ns/health-skin_and_beauty/t/twin-study-reveals-secrets-looking-younger/

[3] http://www.facesofmeth.us/main.htm

[4] The terms time interval, time lapse, elapsed time, and age gap are used interchangeably in this paper.

[5] COTS-A was ranked among the top three performers in the FRVT 2014 face recognition evaluation [5], and COTS-B is PittPatt v5.2.2.



(a) (b) (c)

Figure 4. Statistics of the longitudinal database of face images (mug shots) used in this study: (a) number of face images per subject, (b) elapsed time between the first and last image of each subject, and (c) gender (male or female) and race (white or black) distribution. In total, there are 147,784 images of 18,007 subjects in the database (PCSO_LS).

longitudinal studies has been advocated by [4] for iris and [12] for fingerprint. Grother *et al*. utilized 715,612 Hamming distances from 7,876 subjects enrolled in an operational border crossing system over 4 to 9 years; they estimated that the increase in genuine Hamming distances due to time lapse has no effect on iris recognition failures over a subject's lifetime [4]. Grother *et al*. further identified that the aging effects on iris scores from a prior study [3] on a smaller database were in fact due to changes in pupil dilation. For fingerprint, Yoon and Jain report a decreasing trend in genuine similarity scores from a longitudinal fingerprint database of 15,597 subjects; however, they further determine that fingerprint image quality better explains the variance in match scores, and that the decreasing trend in genuine scores did not indicate a decrease in recognition accuracy over time [12]. The longitudinal study on face recognition in this paper follows the methodologies outlined in [4] and [12].

## 3. Longitudinal Face Database

While the FG-NET [7] and MORPH[6] databases have contributed to studies on face aging, they are not suitable for longitudinal study. FG-NET has only 82 subjects in total with rather large variations in pose, expression, and image quality, and half of the face images are younger than 13 years old. While MORPH is a much larger database, there are still only 317 subjects with at least 5 images acquired over at least 5 years elapsed time. For these reasons, we compiled a new longitudinal database of face images, denoted PCSO_LS.

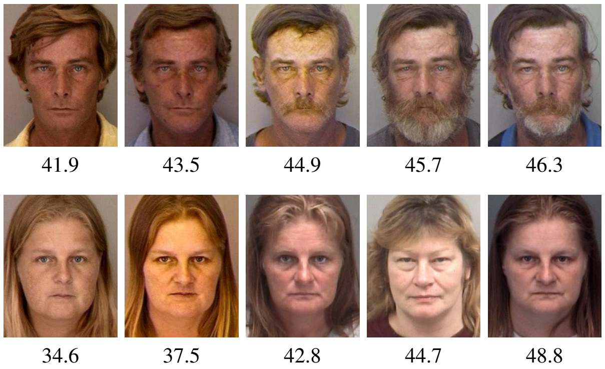


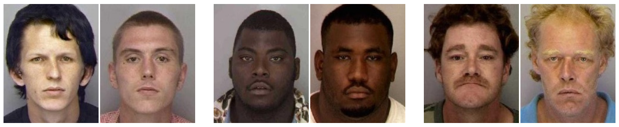
Figure 5. Face images and corresponding ages (in years) of two subjects in the PCSO_LS database.

The PCSO_LS database consists of 147,784 operational mug shots of 18,007 repeat criminal offenders booked by the Pinellas County Sheriff's Office (PCSO) from 1994 to 2010.[7] This subset of images was selected from a larger database using the following criteria. Each subject has at least 5 face images that were collected over at least a 5 year time span, where each pair of consecutive images is age-separated by at least one month. The database statistics are shown in Fig. 4. PCSO_LS has an average of eight (maximum of 60) images per subject that were acquired over an average of 8.5 years (maximum of 16). The average age of the youngest image of a subject is 31 years old; all ages range from 18 to 83 years old. Examples of the booking records for two subjects are shown in Fig. 5. Each booking record (*i.e.* face image) also includes ancillary information (*e.g.* gender, race, date of birth, date of arrest). We only include white and black race subjects in this study because there are too few subjects of other races. Figure 4(c) shows the number of subjects in each demographic group.

Human labeling errors of demographic attributes, as well as subject ID, are typical of large-scale legacy databases. Identifying all such errors in PCSO_LS is not feasible due to the large size of the database. 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 all of their images. A cursory examination of the PCSO_LS

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

[7] Those interested in obtaining this data can contact PCSO: http://www.pcsoweb.com



(a) (b) (c)

Figure 6. Three examples of labeling errors in the PCSO_LS database. All pairs show different subjects who are labeled with the same subject ID.

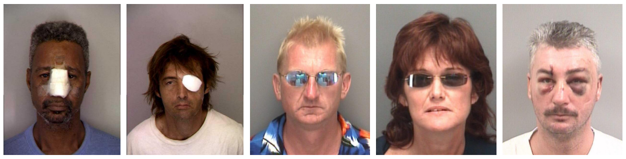


Figure 7. Examples of facial occlusions in the PCSO_LS database.

database revealed 134 subject records that contained multiple identities (see Fig. 6). These subjects were removed from our study. PCSO_LS is relatively constrained which facilitates longitudinal study, but some confounding issues are still present (*e.g.* sunglasses and facial injury shown in Fig. 7). We have retained these images in this study.

## 4. Multilevel Statistical Model

A longitudinal database of face images consists of repeated observations on subjects over time. Face comparison scores generated from such a database can be grouped by subject. To address this hierarchical structure of data, multilevel statistical models have been widely used to properly handle the correlation between intra-subject scores across time.[8] Multilevel models are also appropriate for two properties typical of longitudinal data: time-unstructured and unbalanced (*i.e.* image acquisition schedules and number of face images vary for each subject). The multilevel model in this work consists of two levels:

- Level-1 Model (intra-subject variability)

$$y_{ijk} = \varphi_{0i} + \varphi_{1i} x_{ijk} + \varepsilon_{ijk}, \quad \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2) \quad (1)$$

- Level-2 Model (inter-subject variability)

$$\begin{gathered} \varphi_{0i} = \beta_{00} + b_{0i}, \quad \varphi_{1i} = \beta_{10} + b_{1i}, \\ \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} = N\left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{10} & \sigma_1^2 \end{bmatrix} \right). \end{gathered} \quad (2)$$

The level-1 model describes intra-subject variation in a response variable $y_{ijk}$ as a linear function of a covariate $x_{ijk}$, while level-2 explains inter-subject variation by modeling each subject's true parameters, $\varphi_{0i}$ and $\varphi_{1i}$, as a combination of fixed and random effects. Fixed effects, $\beta_{00}$ and $\beta_{10}$, are the grand means of the population slopes and intercepts and define the *population-mean trend*; random effects, $b_{0i}$ and $b_{1i}$, are each subject's deviations from the population-mean parameters.

In this study, the primary response variable ($y_{ijk}$) is $\tilde{s}_{i,jk} = \frac{s_{i,jk} - \mu}{\sigma}$, where $s_{i,jk}$ is the genuine score between the $j$-th and $k$-th face images of subject $i$ output by one of the two COTS face matchers, and $\mu$ and $\sigma$ are the mean and standard deviation of $\{s_{i,jk}\}$, respectively. We investigate the effects of the following covariates ($x_{ijk}$) to explain variations in genuine scores:

- $\triangle T_{i,jk}$: Time interval between the $j$-th and $k$-th face acquisitions of subject $i$.
- $Q^{\text{ipd}}_{i,jk}$: One minus the ratio of the smallest to largest inter-pupillary distances of the $j$-th and $k$-th face images of subject $i$. Each COTS algorithm outputs estimated eye locations.
- $Q^{\text{yaw}}_{i,jk}$: The absolute value of the difference between the two yaw values of the $j$-th and $k$-th face images of subject $i$. Each COTS algorithm outputs an estimate of the face yaw angle.
- $M_i$: A binary indicator of gender of subject $i$ (1 for male, 0 for female).
- $W_i$: A binary indicator of race of subject $i$ (1 for white, 0 for black).

Table 1 details the multilevel models used in our study that include the above covariates. Note that time-varying covariates ($\triangle T_{i,jk}$, $Q^{*}_{i,jk}$) affect the level-1, while time-invariant covariates ($M_i$, $W_i$) affect the level-2 model specification.

## 5. Experimental Results

Face comparison scores were obtained by two COTS matchers. For genuine scores, we include all pairwise comparisons between a subject's images (*i.e.* if a subject has $n_i$ images, there are $\binom{n_i}{2}$ genuine scores). For the PCSO_LS database, the total number of genuine and impostor scores are 639,531and 11.1 billion, respectively. We compute all impostor scores and calculate thresholds for different FARs to evaluate longitudinal effects on accuracies of the two COTS matchers. Results are presented following a common approach to statistical modeling: we fit increasingly complex models (Table 1) to evaluate the impacts of additional covariates [10]. All models were fit with the LME4 package (v1.1-7)[9] for R (v3.1.1) using full maximum likelihood estimation.

### 5.1. Model Assumptions

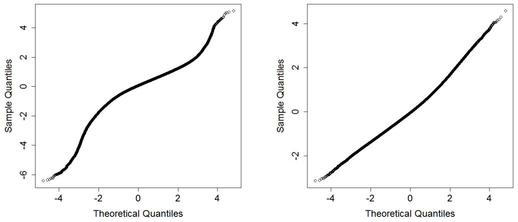
Multilevel models assume normality of the residuals (Eq. (1)) and random effects (Eq. (2)). Figure 8 shows normal probability plots of the residuals, $\varepsilon_{i,jk}$, from fitting

[8] This is in contrast to cross-sectional studies which compare partitions of age groups or time intervals at a single point in time.

[9] http://cran.r-project.org/package=lme4

Table 1. Multilevel models with different covariates used in this study

| Model | Level-1 Model | Level-2 Model | Covariates |
|---|---|---|---|
| Model A | $y_{ijk} = \varphi_{0i} + \varepsilon_{ijk}$ | $\varphi_{0i} = \beta_{00} + b_{0i}$ | None (null model) |
| Model $\mathrm{B_T}$ | $y_{ijk} = \varphi_{0i} + \varphi_{1i} \triangle T_{ijk} + \varepsilon_{ijk}$ | $\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$ | Time interval |
| Model $\mathrm{B_Q}$ | $y_{ijk} = \varphi_{0i} + \varphi_{1i} Q^{\text{ipd}}_{ijk} + \varphi_{2i} Q^{\text{yaw}}_{ijk} + \varepsilon_{ijk}$ | $\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$,<br>$\varphi_{2i} = \beta_{20} + b_{2i}$ | Ratio of eye distances and facial pose |
| Model $\mathrm{C_{GR}}$ | $y_{ijk} = \varphi_{0i} + \varphi_{1i} \triangle T_{ijk} + \varepsilon_{ijk}$ | $\varphi_{0i} = \beta_{00} + \beta_{01} M_i + \beta_{02} W_i + b_{0i}$,<br>$\varphi_{1i} = \beta_{10} + \beta_{11} M_i + \beta_{12} W_i + b_{1i}$ | Time interval, gender, and race |
| Model D | $y_{ijk} = \varphi_{0i} + \varphi_{1i} \triangle T_{ijk} + \varphi_{2i} Q^{\text{ipd}}_{ijk} + \varphi_{3i} Q^{\text{yaw}}_{ijk} + \varepsilon_{ijk}$ | $\varphi_{0i} = \beta_{00} + b_{0i}$, $\varphi_{1i} = \beta_{10} + b_{1i}$,<br>$\varphi_{2i} = \beta_{20} + b_{2i}$, $\varphi_{3i} = \beta_{30} + b_{3i}$ | Time interval, ratio of eye distances, and facial pose |



(a) COTS-A (b) COTS-B

Figure 8. Normal probability plots of level-1 residuals $\varepsilon_{ijk}$ from Model $\mathrm{B_T}$ fit to (a) COTS-A and (b) COTS-B genuine scores.

Model $\mathrm{B_T}$ to genuine scores. Departure from linearity is observed at the tails, particularly for COTS-A, so we can not verify that the model assumptions hold; normal probability plots of random effects $b_{0i}$ and $b_{1i}$ also departed from linearity. This behavior was observed for other models as well, precluding formal hypothesis tests parameters [11].

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, we conduct a non-parametric bootstrap by *case resampling* [11]; 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; single parameter tests for fixed-effects and variance components can be conducted by examining the bootstrap confidence intervals. Table 3 gives the bootstrap parameter estimates, variance components, and 95% confidence intervals for the multilevel models in Table 1. Due to space limitations, results are only shown for COTS-A.

Table 2. Fitting results for unconditional means model (Model A)

| | COTS-A | COTS-B |
|---|---|---|
| $\beta_{00}$ | 0.0609<br>(0.0534, 0.0682) | 0.0265<br>(0.0182, 0.0346) |
| $\sigma^2_\varepsilon$ | 0.6894<br>(0.6741, 0.7065) | 0.7235<br>(0.7167, 0.7299) |
| $\sigma^2_0$ | 0.2418<br>(0.2338, 0.2501) | 0.2850<br>(0.2780, 0.2915) |

## 5.2. Unconditional Means Model

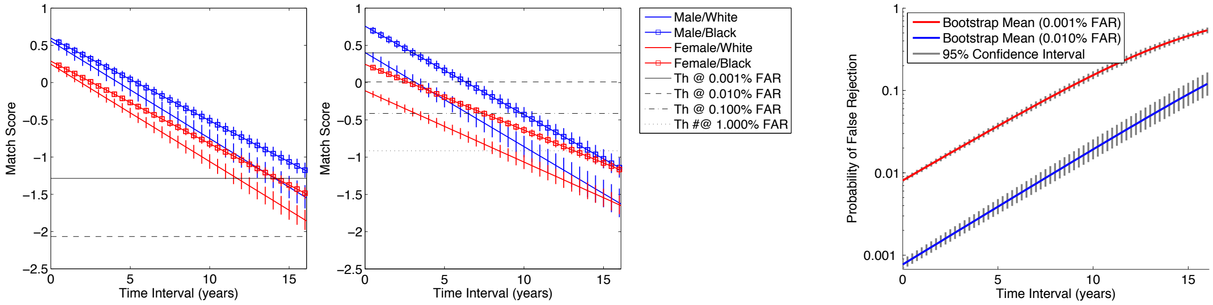
The unconditional means model, Model A, is a multilevel model with subject ID as a random effect but no other covariates. Model A partitions the total variation in genuine scores by subject; each subject's estimated trajectory is the mean of his/her genuine scores (*i.e.* a flat line). Similar to analysis of variance, $b_{0i}$ is the *subject-specific mean* and $\beta_{00}$ is the *grand mean*. The purpose for fitting the unconditional means model is to obtain initial estimates of the random effects which will serve as a baseline for subsequent models. Table 2 gives the estimated intra-subject variance $\sigma^2_\varepsilon$ (*i.e.* deviations around each subject's own mean) and inter-subject variance $\sigma^2_0$ (*i.e.* deviations of subject means around the population mean). Estimates of the intra-subject correlation coefficient, $\rho = \sigma^2_0/(\sigma^2_0 + \sigma^2_\varepsilon)$, for COTS-A and COTS-B are 0.2597 and 0.2826 which indicate that approximately a quarter of the total variation in genuine scores is due to differences between subjects.

## 5.3. Unconditional Growth Model

The next model of interest is the unconditional growth model, Model $\mathrm{B_T}$, which groups match scores by subject and includes time interval $\triangle T_{i,jk}$ as covariate. The level-1 residuals $\varepsilon_{ijk}$ now quantify the variance of each subject's genuine scores around his/her linear trajectory ($\varphi_{0i} + \varphi_{1i} \triangle T_{ijk}$), rather than around his/her subject-specific mean as in Model A. The random effects in Model $\mathrm{B_T}$ allow the estimated slope and intercept to vary for each subject.

Our first observation is that the population-mean trend for Model $\mathrm{B_T}$, given by fixed-effects $\beta_{00}$ and $\beta_{01}$ in Table 3, estimates that COTS-A (COTS-B) genuine scores decrease by 0.1253 (0.1178) standard deviations[10] per year; the null hypothesis of $\beta_{10} = 0$ is rejected at significance level of 0.05 since the 95% bootstrap confidence intervals do not contain zero for both COTS face matchers. Comparing Model A with Model $\mathrm{B_T}$ using a pseudo-$\mathrm{R}^2$ statistic, $\{\sigma^2_\varepsilon(\mathrm{A}) - \sigma^2_\varepsilon(\mathrm{B_T})\}/\sigma^2_\varepsilon(\mathrm{A})$, we can conclude that about 24%

[10] Interpretation of the aging rate w.r.t. standard deviations is made possible by the standardization of genuine scores in Sec. 4.



(a) COTS-A (b) COTS-B (c) COTS-A

Figure 9. Population-mean trends in genuine scores estimated by Model $\text{C}_{\text{GR}}$ for (a) COTS-A and (b) COTS-B. Thresholds at different FARs are overlaid on the plots to indicate implications on recognition accuracy. (c) Population-mean trends in binary verification decisions estimated by Model $\text{B}_{\text{T}}$ for COTS-A show that false rejection rates remain below 2% up to 10 years time interval at 0.01% FAR.

Table 3. Bootstrap estimates of fixed-effects parameters with 95% confidence intervals and variance components for COTS-A

| | | Model $\text{B}_{\text{T}}$ | Model $\text{C}_{\text{GR}}$ | Model $\text{B}_{\text{Q}}$ | Model D |
|---|---|---|---|---|---|
| Fixed Effects | $\beta_{00}$ | 0.5184<br>(0.5103, 0.5263) | 0.2871<br>(0.2644, 0.3085) | 0.4443<br>(0.4361, 0.4521) | 0.6696<br>(0.6619, 0.6776) |
| | $\beta_{10}$ | -0.1253<br>(-0.1265, -0.1240) | -0.1108<br>(-0.1143, -0.1072) | -2.1639<br>(-2.2016, -2.1272) | -0.1161<br>(-0.1174, -0.1147) |
| | $\beta_{20}$ | | | -0.0680<br>(-0.0701, -0.0658) | -0.6256<br>(-0.6648, -0.5891) |
| | $\beta_{30}$ | | | | -0.0607<br>(-0.0625, -0.0588) |
| | $\beta_{01}$ | | 0.3093<br>(0.2880, 0.3318) | | |
| | $\beta_{11}$ | | -0.0030*<br>(-0.0064, 0.0005) | | |
| | $\beta_{02}$ | | -0.0400<br>(-0.0548, -0.0248) | | |
| | $\beta_{12}$ | | -0.0200<br>(-0.0224, -0.0173) | | |
| Variance | $\sigma_\varepsilon^2$ | 0.5237 | 0.5238 | 0.5432 | 0.4424 |
| | $\sigma_0^2$ | 0.2294 | 0.2152 | 0.2195 | 0.2077 |
| | $\sigma_1^2$ | 0.0038 | 0.0037 | 4.1159 | 0.0041 |
| | $\sigma_2^2$ | | | 0.0063 | 2.9934 |
| | $\sigma_3^2$ | | | | 0.0053 |
| | $\sigma_{01}$ | -0.0061 | -0.0062 | ** | ** |

*The null hypothesis that the parameter is equal to zero cannot be rejected because the 95% confidence interval contains zero.

**Covariance components have been omitted due to space limitations.

and 19% of the intra-subject variation can be explained by a linear relationship between genuine scores and time interval for COTS-A and COTS-B, respectively.

### 5.4. Other Covariates

Goodness-of-fit measures are used to compare different models. Deviance can only be used to compare the fit of *nested* models (*e.g*. model pairs ($\text{B}_{\text{T}}$, $\text{C}_{\text{GR}}$) or ($\text{B}_{\text{T}}$, D)). While AIC and BIC allow comparison of non-nested models (*e.g*. models $\text{B}_{\text{T}}$ and $\text{B}_{\text{Q}}$), the magnitude has little meaning [10]. Table 4 shows reductions in goodness-of-fit for each additional covariate (*i.e*. every covariate considered explains some of the variation in genuine scores). Pseudo-$\text{R}^2$ statistics are also used to measure the proportional reduction in level-2 variance ($\sigma_*^2$) and level-1 residual variance ($\sigma_\varepsilon^2$) attributable to inclusion of time-invariant and time-variant covariates, respectively.

Table 4. Goodness-of-fit for multilevel models in Table 1

(a) COTS-A

| Model | AIC | BIC | Deviance |
|---|---|---|---|
| A | 1,665,285 | 1,665,320 | 1,665,279 |
| $\text{B}_{\text{T}}$ | 1,508,569 | 1,508,638 | 1,508,557 |
| $\text{C}_{\text{GR}}$ | 1,507,242 | 1,507,355 | 1,507,222 |
| $\text{B}_{\text{Q}}$ | 1,540,801 | 1,540,915 | 1,540,781 |
| D | 1,425,416 | 1,425,586 | 1,425,386 |

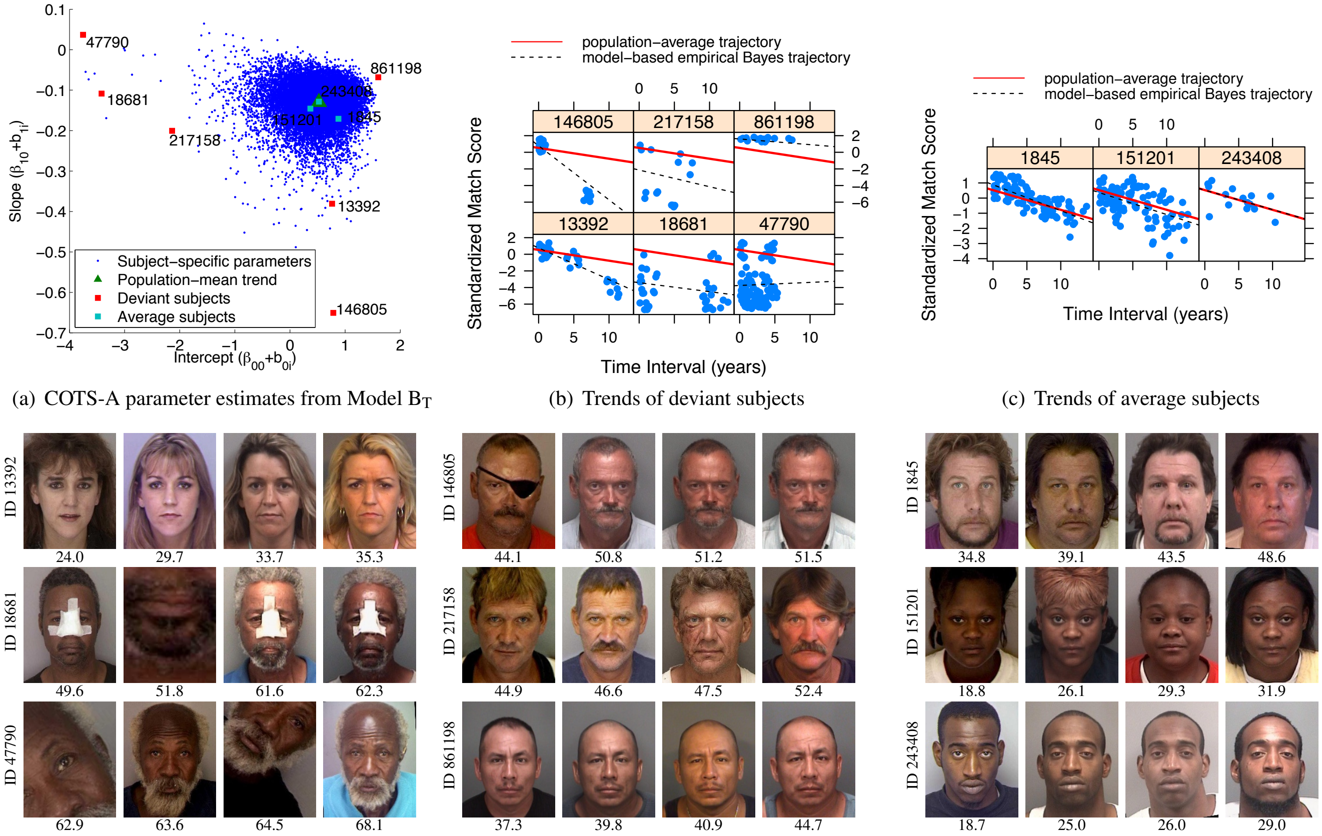
(b) COTS-B

| Model | AIC | BIC | Deviance |
|---|---|---|---|
| A | 1,651,355 | 1,651,389 | 1,651,349 |
| $\text{B}_{\text{T}}$ | 1,531,939 | 1,532,008 | 1,531,927 |
| $\text{C}_{\text{GR}}$ | 1,527,868 | 1,527,982 | 1,527,848 |
| $\text{B}_{\text{Q}}$ | 1,560,491 | 1,560,605 | 1,560,471 |
| D | 1,477,954 | 1,478,124 | 1,477,924 |

**Gender and Race:** Comparing Model $\text{B}_{\text{T}}$ with Model $\text{C}_{\text{GR}}$, pseudo-$\text{R}^2$ statistics measure that 6.2% (17.1%) of the variation in intercept and 2.6% (5.7%) of the variation in slope parameters is explained by gender and race covariates for COTS-A (COTS-B). Population-mean trends for each demographic plotted in Figs. 9(a) and (b), along with the 95% bootstrap confidence intervals, suggest that males are easier to recognize than females and black subjects are easier to recognize than white subjects [1, 8]. The 1,928 white females have the lowest and the 5,808 black males have the highest average intercepts. Note that gender has no effect on differences in slope for COTS-A (null hypothesis of $\beta_{11} = 0$ cannot be rejected); the effect of gender on differences in slope is significant for COTS-B.

Verification thresholds at different FARs are overlaid on the plots in Figs. 9(a) and (b) to show how these trends relate to verification accuracy; COTS-A correctly verifies the average individual, regardless of demographic, across 16 years time interval at 0.01% FAR, while this only holds up to 6 years time interval for COTS-B.

**Face Quality:** Overall, the model with the lowest (*i.e*. best) goodness-of-fit was Model D which includes covariates of $\triangle T_{ijk}$, $Q_{ijk}^{\text{yaw}}$, $Q_{ijk}^{\text{ipd}}$. Compared with Model $\text{B}_{\text{T}}$, inclusion of the face quality covariates reduces the level-1 residual variation by 15.5%. The intercept of Model D refers to the score when all covariates are zero. Because of how we defined the quality covariates, the intercept corresponds to the highest quality images (pairs) when $\triangle T_{ijk} =$



(a) COTS-A parameter estimates from Model $B_T$

(b) Trends of deviant subjects

(c) Trends of average subjects

(d) Face images of deviant subjects

(e) Face images of average subjects

Figure 10. Parameter estimates for Model $B_T$ fit to COTS-A genuine scores; (a) subject-specific and population-mean parameter estimates are marked with blue dots and green triangle, respectively, and six examples of deviant subjects are shown as red squares. Subject-specific growth plots of the (b) six deviant subjects and (c) three average subjects. Population-mean and subject-specific trajectories, estimated from Model $B_T$, are overlaid in solid red and dotted black lines, respectively. Subject IDs are shown above each plot and correspond to face images of the (d) six deviant subjects and (e) three average subjects, which were all aligned and cropped using COTS-A eye locations.

0; hence, $\beta_{00}$ increases compared to Model $B_T$. Note also that $\beta_{10}$ is the expected change in the score w.r.t. time interval when face quality covariates are held constant; $\beta_{10}$ for Model D has decreased compared to Model $B_T$.

## 5.5. Subject-Specific Trends

In addition to estimated population-mean trends, we are also interested in the overall behavior of the population. The distribution of the subject-specific parameters around the population-mean parameters $(\beta_{00}, \beta_{10})$ is given for Model $B_T$ in Fig. 9(a) for COTS-A. The parameter estimates for the majority of subjects are symmetrically distributed around the population-mean trend; the genuine scores of almost all of the 18,007 subjects in PCSO_LS decrease with elapsed time. Figures 10(c) and (e) show three examples of subjects who exhibit the population-mean trend. However, Figs. 10(b) and (d), provide six examples of subjects whose estimated parameters deviate significantly from the population-mean trend. As demonstrated by the face images of subject IDs 13392 and 861198 in Fig. 10(e), deviations can be due to aging more or less than average. However, we found that most of the significant deviations are due to occlusions, incorrect eye locations, facial injury, etc. (*i.e.* confounding issues that are difficult to model.) Note that in preliminary results, examining deviant subjects in this manner helped to identify some of the subject ID labeling errors mentioned in Sec. 3.

## 5.6. Probability of True Acceptance

We also fit Model $B_T$ to binary verification decisions on COTS-A genuine scores at different FARs to study probability of true acceptance over time. Thresholds were calculated from the distribution of 11.1 billion impostor scores. Binary response variables can be viewed as Bernoulli trials with probability of success (*i.e.* true acceptance), $\pi_{ijk}$. The multilevel model is then specified as:

$$\text{Level 1: } g(\pi_{ijk}) = \varphi_{0i} + \varphi_{1i}\triangle T_{ijk} + \varepsilon_{ijk} \tag{3}$$

$$\text{Level 2: } \varphi_{0i} = \beta_{00} + b_{0i}, \quad \varphi_{1i} = \beta_{10} + b_{1i} \tag{4}$$

where $g(\pi_{ijk}) = log(\pi_{ijk}/(1 - \pi_{ijk}))$.

Figure 9(c) shows the bootstrap estimated population-mean trends of probability of false rejection ($1 -$ true acceptance) with respect to time interval and corresponding 95% bootstrap confidence intervals. At 0.01% FAR, the probability of false rejection for COTS-A remains below 2% until approximately 10 years; at 16 years, it is estimated to increase to 10%. However, at 0.001% FAR, the probability of false rejection begins to increase almost immediately, dropping to about 20% for 10 years.

## 6. Summary and Conclusions

We presented a longitudinal study on face recognition, utilizing a database (PCSO_LS) of 147,784 face images of 18,007 subjects. Each subject has at least 5 images which were acquired over at least a 5 year time interval. Multilevel statistical models were used to estimate population-mean trends in genuine scores, particularly with respect to increasing elapsed time between two face images. Based on the results of our longitudinal analysis, our findings are summarized as follows:

1. Significantly decreasing trends in genuine scores over time were observed for the two state-of-the-art COTS face matchers, with COTS-A showing overall better performance than COTS-B.
2. Despite decreasing genuine scores, population-mean trends of COTS-A genuine scores from all demographic groups remained above the threshold at 0.01% FAR for time intervals up to 16 years (the maximum in the PCSO_LS database);
3. The probability of false rejection by COTS-A was stable (at less than 2%) across time intervals up to approximately 10 years for 0.01% FAR.
4. Some of the variation in subject-specific estimated parameters was explained by gender, race, and face quality covariates. Face quality measures did not explain the variation in genuine scores as well as time interval alone, but improved the model fit when included with time interval.

We stress that most prior studies evaluated aging effects on performance of face matchers at 1% FAR (*e.g.* [6]). Our results indicate that it is important to utilize state-of-the-art matchers and large operational databases. This study in no way claims to have reached a "final model." Additional covariates and interactions are likely needed; statistically significant non-zero intra-subject and inter-subject variance exists in our models that could be further reduced.

Future work will include: (i) Studying the stability of the impostor distribution over time; recognition errors can also manifest in increased impostor similarity scores. (ii) Mitigate the correlation that exists between all pair-wise genuine comparisons. (iii) Investigate non-linear models and include absolute age and other quality measures as covariates. (iv) While large-scale longitudinal databases of face images are not easy to obtain, it would be desirable to repeat this study on a database with different demographic makeup, particularly a database of civilians rather than criminals.

## Acknowledgments

This research was supported by a grant from the NSF Center for Identification Technology Research (CITeR). The authors thank Soweon Yoon at Michigan State University and Patrick Grother at the National Institute of Standards and Technology (NIST) for their valuable feedback and suggestions.

## References

[1] 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*, 113:750–762, 2009.

[2] M. Erbilek and M. Fairhurst. A methodological framework for investigating age factors on the performance of biometric systems. In *Proc. Multimedia and Security*, 2012.

[3] S. Fenker and K. W. Bowyer. Experimental evidence of a template aging effect in iris biometrics. In *WACV*, 2012.

[4] 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.

[5] P. Grother and M. Ngan. FRVT: Performance of face identification algorithms. NIST Interagency Report 8009, May 2014.

[6] B. Klare and A. K. Jain. Face recognition across time lapse: On learning feature subspaces. In *Proc. IJCB*, 2011.

[7] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. *IEEE Trans. on PAMI*, 24(4), Apr. 2002.

[8] 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.

[9] N. Ramanathan, R. Chellappa, and S. Biswas. Computational methods for modeling facial aging: A survey. *Journal of Visual Languages and Computing*, 20:131–144, 2009.

[10] J. D. Singer and J. B. Willett, editors. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York: Oxford Univ. Press, Inc., 2003.

[11] R. van der Leeden, F. M. Busing, and E. Meijer. Bootstrap methods for two-level models. In *Multilevel Conf.*, 1997.

[12] S. Yoon and A. K. Jain. Longitudinal study of fingerprint recognition. Tech. Report MSU-CSE-14-3, Michigan State Univ., East Lansing, MI, USA, Jun. 2014.