

Face Recognition Performance Under Aging

Debyan Deb
Michigan State University
East Lansing, MI, USA
debdebay@msu.edu

Lacey Best-Rowden
Michigan State University
East Lansing, MI, USA
bestrowl@msu.edu

Anil K. Jain
Michigan State University
East Lansing, MI, USA
jain@cse.msu.edu

Abstract

With the integration of face recognition technology into important identity applications, it is imperative that the effects of facial aging on face recognition performance are thoroughly understood. As face recognition systems evolve and improve, they should be periodically re-evaluated on large-scale longitudinal face datasets. In our study, we evaluate the performance of two state-of-the-art commercial off the shelf (COTS) face recognition systems on two large-scale longitudinal datasets of mugshots of repeat offenders. The largest of these two datasets has 147,784 images of 18,007 subjects with an average of 8 images per subject over an average time span of 8.5 years. We fit multi-level statistical models to genuine comparison scores (similarity between images of the same face) from the two COTS face matchers. This allows us to analyze the degradation in recognition performance due to elapsed time between a probe (query) and its enrollment (gallery) image. We account for face image quality to obtain a better estimate of trends due to aging, and analyze whether longitudinal trends in genuine scores differ by subject gender and race. Based on the results of our statistical model, we infer that the state-of-the-art COTS matchers can verify 99% of the subjects at a false accept rate (FAR) of 0.01% for up to 10.5 and 8.5 years of elapsed time. Beyond this time lapse of 8.5 years, there is a significant loss in face recognition accuracy. This study extends and confirms the findings of earlier longitudinal studies on face recognition.

1. Introduction

It is now well established that accuracies of face recognition systems are adversely affected by factors including facial pose, illumination, expression and aging, collectively known as PIE-A. While image acquisition conditions and subject cooperation can accommodate for PIE variations in controlled capture scenarios, facial aging factors are intrinsic and cannot be controlled. A face undergoes various temporal changes across time, including wrinkles, weight,

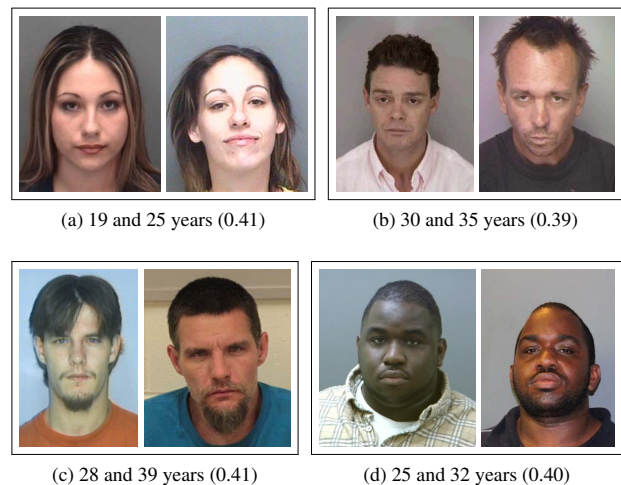


Figure 1: Examples of low-scoring genuine face image pairs of four subjects from (a), (b) PCSO and (c), (d) MSP longitudinal mugshot datasets. Ages at image acquisitions are given along with similarity scores from COTS-A for each pair. The thresholds for COTS-A at 0.1% and 0.01% false accept rate (FAR) are 0.42 and 0.49, respectively, for both PCSO and MSP datasets. Each of the genuine pairs would be falsely rejected at 0.1% and 0.01% FARs.

facial hair, etc. Figure 1 shows that large time lapse between two genuine face images can result in false reject errors. Therefore, it is critical to systematically evaluate face recognition technology on longitudinal face datasets to determine state-of-the-art accuracy with respect to time lapse between a probe and its enrollment image in the gallery.

A considerable amount of research has been performed and reported on age-invariant face recognition [6],[7], synthetic aging [8],[9], automatic age estimation from face images [10], aging analysis [11], and appearance prediction across ages [12]. In this study, we focus on quantifying the impact of aging on the performance of face recognition systems. Studies within this realm traditionally have followed a methodology of dividing a given population into discrete age groups and then reporting recognition performance (e.g. TAR or EER) for each age group independently. Two major

conclusions have been drawn based on this cross-sectional analysis: (i) Face recognition performance decreases as the elapsed time between two images of the same subject increases [1], [3], [4], and (ii) faces of younger individuals are more difficult to recognize than faces of older individuals [4],[5]. Studies, where arbitrary age group assignment are considered, make it difficult to compare results from different studies. Additionally, the primary concern with facial aging is a decrease in *genuine* similarity scores over time lapse, which causes an increase in false non-match rates. Reporting summary performance measures for subsets of a facial aging dataset does not provide any insight into how the genuine similarity scores of individuals are changing over time. Previous facial aging studies [19] primarily used FG-NET [17] and MORPH [18] face datasets, which are both limited in terms of the number of subjects and number of images per subject. FG-NET contains only 1,002 images from 82 subjects and MORPH contains only 317 subjects that have at least 5 images over at least 5 years time lapse. Due to their small size, both in terms of number of subjects and number of images per subject over time, FG-NET and MORPH are not suitable for statistical modeling for longitudinal study. For this study, we obtained two large-scale longitudinal face datasets, denoted as PCSO and MSP (see Table 1) and report our inference based on these datasets. While these two datasets, due to privacy issues, are not available in the public domain, there is no way to carry out a meaningful longitudinal study without such large-scale data that is only available from government agencies.

Multilevel statistical models have been adopted for large-scale longitudinal study of iris [13], fingerprint [16], and face [2] recognition. Grother *et al.* analyzed a large-scale longitudinal dataset of 7,876 subjects to quantify the iris aging effect on recognition over time in a study called IREX VI [13]. They concluded that an increase in Hamming distance due to elapsed time between enrolled and query images has no significant effect on iris recognition failures. Some limitations of the IREX VI study were identified [14], [15]. Yoon and Jain analyzed a longitudinal fingerprint dataset of 15,597 subjects and found a decreasing trend in genuine match scores; however, the recognition accuracy, at operational FARs between 0.01% and 0.00001%, remained stable up to 12 years (the maximum time span in the dataset) [16]. Best-Rowden and Jain performed a subject-specific analysis using two longitudinal face datasets: PCSO dataset containing 147,784 images of 18,007 subjects and LEO dataset¹ containing 31,852 images of 5,636 subjects [2]. Best-Rowden and Jain concluded: (i) while genuine scores declined significantly over time, scores for 99% of the population remained above the thresh-

¹Face images from this dataset are not available to us and hence, we do not use this dataset in our study. Without the face images, we could not, for example, extract the image quality covariate.

old at FAR of 0.01% (0.1%) until 6.5 (8.5) and 5.5 (8.0) years of elapsed time for PCSO and LEO datasets, respectively, and (ii) subject demographics had marginal effects on the rates of change in genuine scores over time. Because these results are tied to the particular COTS face matchers used in [2], it is imperative to periodically repeat the longitudinal study utilizing new and improved face matchers as they evolve, as well as additional large-scale longitudinal datasets.

This paper repeats and expands on the longitudinal face study in [2]. The contributions of this paper are as follows:

- Evaluate the performance of two state-of-the-art COTS face matchers (COTS-A and COTS-B)² on two longitudinal mugshot datasets (PCSO and MSP)³ from two different law enforcement sources. COTS-A is a new version from the same vendor of the COTS-A face matcher used in [2].
- PCSO and MSP datasets used in this study are the largest longitudinal face datasets studied to date. See Table 1 for details on PCSO and MSP datasets.
- Analyze the rate of change in genuine scores due to the elapsed time between enrollment and probe images, as well as covariates such as gender, race, and the quality of the gallery and probe images.

The key differences between this study and [2] are:

- A newer version of the COTS-A face matcher is utilized with significantly improved face recognition performance.
- Evaluation of COTS-A on both PCSO and MSP datasets, and evaluation of COTS-A and COTS-B on MSP. In contrast, [2] evaluated two matchers, each on only a single dataset.
- While [2] used inter-pupillary distance (IPD) and face frontalness as quality measures, we utilize an overall face quality measure, namely, Rank-based Quality Score (RQS) proposed by Chen *et al.* [20].

The organization of the paper is as follows. Section 2 describes the two large-scale longitudinal datasets (PCSO and MSP) used in this study. In Section 3, we revisit the multilevel statistical models from [2]. Section 4 makes inferences from fitting the models on genuine scores from COTS-A and COTS-B. Section 5 summarizes and concludes our paper.

²COTS-A is one of the top-3 performers in the NIST FRVT 2014. COTS-B algorithm is based on the deep convolutional networks. These are state-of-art face matchers which are essential for such a longitudinal study.

³COTS-B was not evaluated on PCSO dataset because, according to the vendor, it was trained on PCSO.

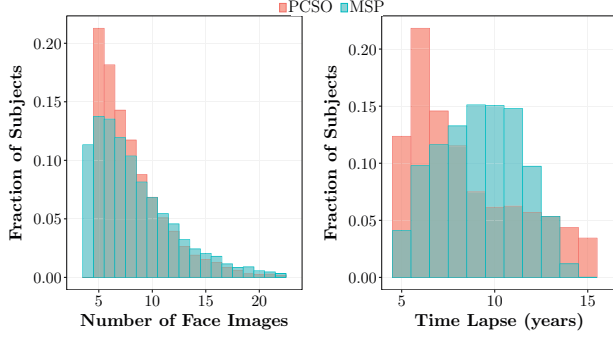


Figure 2: Statistics of the two longitudinal face datasets used in this study (PCSO and MSP). Histograms denote the number of face images per subject, and show the time lapse between the enrollment and the latest probe image of a subject. PCSO dataset contains 147,784 face images of 18,007 subjects and MSP dataset contains 82,450 images of 9,572 subjects.

2. Longitudinal Face Datasets

The two longitudinal face datasets used in this study are obtained from two different law enforcement sources: (i) Pinellas County Sheriff’s Office (PCSO) and (ii) Michigan State Police (MSP). The PCSO longitudinal dataset contains 147,784 mugshots of 18,007 recidivists spanning from the years 1994 to 2010. The MSP longitudinal dataset contains 82,450 mugshots of 9,572 recidivists spanning from the years 2002 to 2015. Statistics for these two datasets are shown in Figure 2.

The criteria used to select the subjects and images are the following. (i) Each subject has sufficient number of image acquisitions (at least 5 for PCSO and at least 4 for MSP) over a minimum of 5 year time span, (ii) each consecutive image acquisition of a subject is separated by at least one month, and (iii) the youngest subject is at least 18 years old. Let $I_{i,j}$ represent the j th mugshot acquisition of subject i . Then $I_i = \{I_{i,1}, I_{i,2}, I_{i,3}, \dots, I_{i,N_i}\}$ is the set of N_i total images for subject i . We order $I_{i,j}$ by the time of acquisition, so $T_{i,1} < T_{i,2} < T_{i,3} < \dots < T_{i,N_i}$, where $T_{i,j}$ is the date of acquisition of $I_{i,j}$. Let $Age_{i,j}$ denote subject i ’s age during the j th face acquisition. In summary,

- $N_i \geq 5$ and $N_i \geq 4$ for PCSO and MSP, respectively.
- $(T_{i,N_i} - T_{i,1} \geq 5)$ years for both PCSO and MSP.
- $(T_{i,(j+1)} - T_{i,j} \geq 1)$ month for both PCSO and MSP.
- $Age_{i,1} \geq 18$ years for both PCSO and MSP.

3. Multilevel Statistical Model

The large-scale longitudinal datasets described in Section 2 contain repeated observations (face acquisitions and, consequently, genuine scores) for each subject over

Table 1: Statistics of PCSO and MSP datasets.

	PCSO	MSP
Number of Images	147,784	82,450
Number of Subjects	18,007	9,572
Average no. of images/subject	8	9
Max no. of images/subject	60	48
Average time lapse (years)	8.5	9.0
Max time span (years)	16	14
Average age of subjects (years)	31	33
Youngest age of subject (years)	18	18
Oldest age of subject (years)	83	78
Male/Female ratio (%)	83/17	88/12
Black/White ratio (%)	61/39	52/48
Average IPD (pixels)	113	126

time. Additionally, the datasets are both time-unstructured ($T_{i,k} \neq T_{j,k}$) and unbalanced ($N_i \neq N_j$). To study such hierarchical data, multilevel (or mixed-effects) statistical models have been widely used to evaluate the correlation of within-subject response variables across time in many important fields of research including epidemiology, sociology, psychology, etc.

The models used in this work contain two levels, similar to those used in [2]. The Level-1 model describes the changes in genuine scores, $Y_{i,j}$, for each subject over time (within-subject variation), whereas the Level-2 model describes how these changes differ across subjects (between-subject variation). Genuine scores are standardized using z-score normalization so that coefficients obtained from the models are interpreted as the change in standard deviations of the genuine distribution per year (e.g. for elapsed time covariate). The genuine score distributions for COTS-A on PCSO and MSP and COTS-B on MSP datasets are shown in Figure 3. We model changes in genuine scores over time, $Y_{i,j}$, as a linear function of various covariates, $X_{i,j}$,

$$Y_{i,j} = \pi_{0i} + \pi_{1i}X_{i,j} + \varepsilon_{i,j}, \quad (1)$$

where π_{0i} and π_{1i} are subject i ’s intercept and slope, respectively. Equation 1 is the Level-1 model which corresponds to within-subject changes in face comparison scores over time. The Level-1 residual variation, $\varepsilon_{i,j}$, represents the variance in each individual’s comparison scores from his/her linear longitudinal trend. The slope and intercept parameters, π_{0i} and π_{1i} , are modeled as a combination of fixed and random effects. Fixed effects, γ_{00} and γ_{10} , are the overall means of the population intercepts and slopes, whereas random effects, b_{0i} and b_{1i} , are subject i ’s deviation from the population means. Therefore, subject i ’s slope and intercept parameters can be written as,

Table 2: Multilevel models with different covariates

Model	Level-1 Model	Level-2 Model	Covariates
Model B_T	$Y_{ij} = \pi_{0i} + \pi_{1i}\Delta T_{i,j} + \varepsilon_{i,j}$	$\pi_{0i} = \gamma_{00} + b_{0i}$, $\pi_{1i} = \gamma_{10} + b_{1i}$	Time lapse
Model C_{GR}	$Y_{ij} = \pi_{0i} + \pi_{1i}\Delta T_{i,j} + \varepsilon_{i,j}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}Gend_i + \gamma_{02}Race_i + b_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{11}Gend_i + \gamma_{12}Race_i + b_{1i}$	Time lapse, gender, and race
Model Q_T	$Y_{ij} = \pi_{0i} + \pi_{1i}\Delta T_{i,j} + \pi_{2i}Qual_{i,j \neq 1}$ $+ \pi_{3i}\Delta T_{i,j}Qual_{i,j \neq 1} + \varepsilon_{i,j}$	$\pi_{0i} = \gamma_{00} + \gamma_{01}Qual_{i,1} + b_{0i}$ $\pi_{1i} = \gamma_{10} + \gamma_{11}Qual_{i,1} + b_{1i}$ $\pi_{2i} = \gamma_{20} + \gamma_{21}Qual_{i,1}$ $\pi_{3i} = \gamma_{30} + \gamma_{31}Qual_{i,1}$	Time lapse, quality

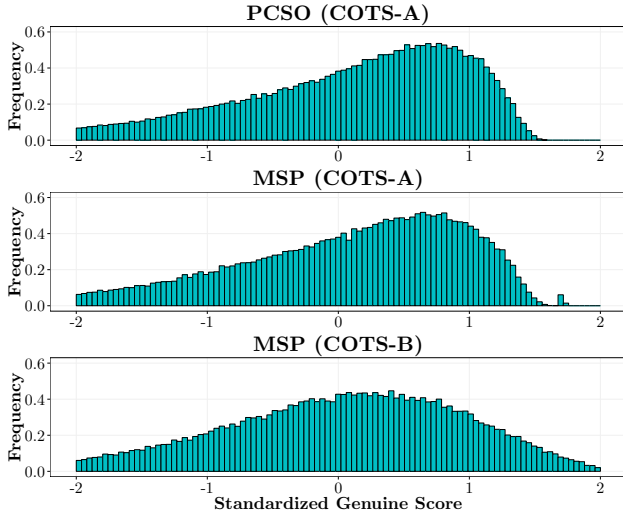


Figure 3: Genuine score distributions for (top to bottom) COTS-A on PCSO, COTS-A on MSP, and COTS-B on MSP datasets.

$$\pi_{0i} = \gamma_{00} + b_{0i}, \quad (2a)$$

$$\pi_{1i} = \gamma_{10} + b_{1i}. \quad (2b)$$

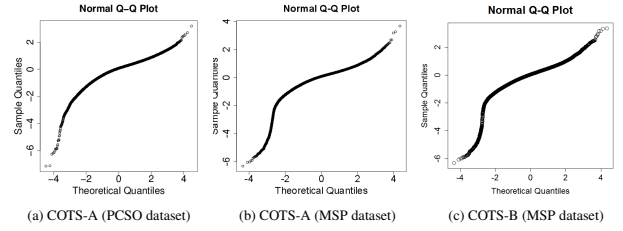
Equations 2a and 2b constitute the Level-2 model which models changes in face comparison scores across subjects. Equations 1 and 2 can be combined to get,

$$Y_{ij} = [\gamma_{00} + b_{0i}] + [\gamma_{10} + b_{1i}]X_{i,j} + \varepsilon_{i,j}. \quad (3)$$

If the random effects and residual variation are equal to their expected means of zero, then Equation 3 reduces to the population-mean trend $Y_{ij} = \gamma_{00} + \gamma_{10}X_{i,j}$.

The different covariates incorporated in the models are as follows:

- $\Delta T_{i,j}$: time lapse between the j th image acquisition and the enrollment face image of subject i
- $Age_{i,1}$: age at enrollment of subject i


 Figure 4: Normal probability plots for level-1 residuals ε_{ij} from Model B_T fit to COTS-A on (a) PCSO and (b) MSP and COTS-B on (c) MSP genuine scores.

- $Gend_i$: gender of subject i (0 for female, 1 for male)
- $Race_i$: race of subject i (0 for black, 1 for white)
- $Qual_{i,j}$: face quality metric for j th mugshot image of subject i . Face quality values used here are obtained from the methodology proposed by Chen *et al.* [20]. $Qual_{i,1}$ is the face quality value for enrollment image and $Qual_{i,j \neq 1}$ is the quality value for the j th probe image of subject i .

Note that $\Delta T_{i,j}$ and $Qual_{i,j \neq 1}$ are time-varying covariates and affect within-subject variation in genuine scores (Level-1). On the other hand, $Age_{i,1}$, $Gend_i$, $Race_i$, and $Qual_{i,1}$ are time-invariant covariates and affect between-subject variation in genuine scores (Level-2). Table 2 describes all models and covariates used in this study.

4. Experimental Results

Genuine scores from COTS-A and COTS-B were obtained by comparing each subject's enrollment image (youngest acquisition) to his/her subsequent face images. Hence, for subject i , there are $(N_i - 1)$ genuine comparisons. For PCSO dataset, there are a total of 130,878 genuine and 11.1 billion impostor comparison scores, whereas, for MSP dataset, there are 82,150 genuine and 4.1 billion impostor comparison scores. Increasingly complex models (Table 2) are successively fit to evaluate the variation in genuine scores over time and the impact of additional covariates. Models were fit using the LME4 package in R

Table 3: Bootstrap parameter estimates of fixed-effects and variance components for COTS-A on PCSO and MSP datasets and COTS-B on MSP dataset. Estimates where the 95% bootstrap confidence intervals contained zero are highlighted in bold.

	PCSO (COTS-A)			MSP (COTS-A)			MSP (COTS-B)		
	B_T	C_{GR}	Q_T	B_T	C_{GR}	Q_T	B_T	C_{GR}	Q_T
γ_{00}	0.7233	0.3958	0.6889	0.5858	0.4914	0.5059	0.5599	0.6847	0.4647
γ_{10}	-0.1429	-0.1170	-0.1399	-0.1076	-0.1006	-0.0945	-0.1036	(-0.0849)	-0.0899
γ_{01}		0.4286	0.0595		0.1665	0.1173		0.0424	0.1762
γ_{11}		-0.0087	0.0067		(0.0078)	(-0.0029)		(-0.0084)	-0.0058
γ_{02}		-0.0433	0.0858		-0.0929	0.1900		-0.3128	0.2408
γ_{12}		-0.0308	0.0069		-0.0291	0.0104		-0.0251	0.0056
γ_{20}			0.0605			0.0833			0.1380
γ_{21}			(-0.0002)			(-0.0001)			-0.0064
σ_ϵ^2	0.2951	0.2955	0.2587	0.5300	0.5230	0.3878	0.5218	0.5065	0.4034
σ_0^2	0.2465	0.2202	0.2322	0.4338	0.4232	0.3878	0.5878	0.5547	0.5258
σ_1^2	0.0036	0.0033	0.0032	0.0081	0.0077	0.0067	0.0080	0.0072	0.0068
σ_{01}	-0.0020	-0.0017	-0.0026	-0.0367	-0.0356	-0.0305	-0.0470	-0.0447	-0.0425
AIC	254659	252696	246185	178710	177435	162928	177703	174684	163758
BIC	254717	252794	246371	178765	177527	163103	177758	174776	163932
Deviance	254647	252676	246147	178699	177415	162890	177691	174664	163720

using maximum likelihood estimation. Thresholds at different FAR values are calculated from the full impostor distributions in order to evaluate how the longitudinal trends in genuine scores affect the recognition accuracies of COTS-A and COTS-B.

Inferences from multilevel models are based on the assumption that the residual errors are normally distributed. Figure 4 shows the normal probability plots (Q-Q plots) of the residuals, $\epsilon_{i,j}$, from fitting Model B_T to genuine scores. For both datasets, linearity is violated, indicating that the validity of normality assumption does not hold. When parametric model assumptions are violated, non-parametric bootstrap can be performed to obtain confidence intervals for the parameter estimates [16]. Therefore, non-parametric bootstrapping is conducted with 1,000 bootstrap sets, obtained by sampling 18,007 and 9,572 subjects with replacement from PCSO and MSP, respectively. The multilevel models are then fit to each bootstrap set, and the mean parameter estimates over all 1,000 bootstraps are reported. Table 3 gives parameter estimates and variances obtained from the models after bootstrapping; 95% bootstrap confidence intervals have been omitted due to space, but parameters for which confidence intervals contained zero are indicated in bold. These parameters are statistically zero and the null hypothesis of the parameter equal to 0 cannot be rejected at significance level of 0.05.

4.1. Time Lapse

Model B_T contains a single covariate, namely the time lapse between a subject’s enrollment image and probe images ($\Delta T_{i,j}$). The population-mean trend for Model B_T , given by γ_{00} and γ_{10} , estimates that COTS-A genuine scores decrease by $\gamma_{10} = 0.1429$ and 0.1076 standard deviations per year for PCSO and MSP datasets, respectively. Similar to COTS-A on MSP, Model B_T estimates

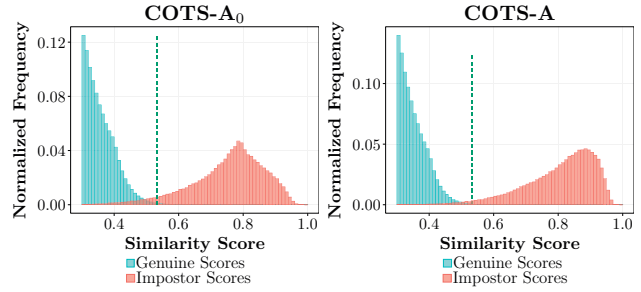


Figure 5: Genuine score distributions for (a) COTS-A₀ and (b) COTS-A face matchers. Thresholds at 0.01% FAR for COTS-A₀ and COTS-A face matchers are 0.53 and 0.49, respectively. The thresholds are shown with dashed lines.

Table 4: Bootstrap parameter estimates of fixed-effects and variance components from Model B_T for COTS-A₀ and COTS-A on PCSO dataset

		COTS-A ₀ [2]	COTS-A
intercept	γ_{00}	0.6734	0.7233
slope	γ_{01}	-0.1364	-0.1429
	σ_ϵ^2	0.3912	0.2951
	σ_0^2	0.3243	0.2465
	σ_1^2	0.0028	0.0036
	σ_{01}	-0.0039	-0.0020

† COTS-A₀ is the older version of COTS-A face matcher used by Best-Rowden and Jain [2].

that COTS-B genuine scores from MSP decrease by 0.1036 standard deviations per year. In other words, this implies that COTS-A genuine scores decrease by one full standard deviation of the PCSO (MSP) score distribution after $1/\gamma_{10} = 7.0$ (9.3) years of elapsed time. Again, similar to COTS-A on MSP, COTS-B genuine scores decrease by one full standard deviation of the MSP score distribution after $1/\gamma_{10} = 9.7$ years of elapsed time.

Table 4 compares the longitudinal performance on PCSO of the COTS-A face matcher with the previous version of

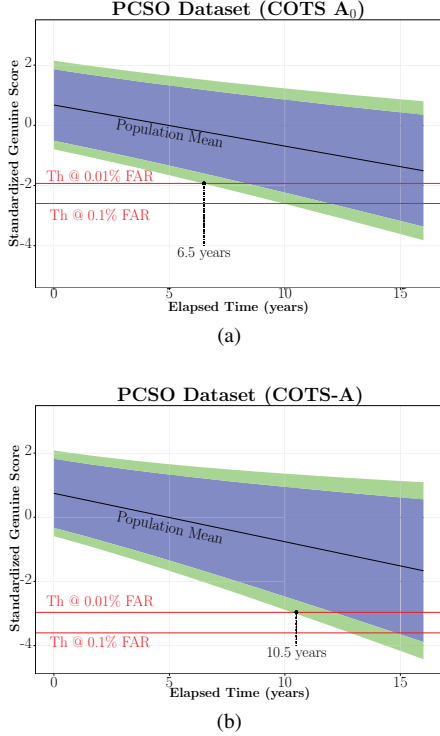


Figure 6: Results from Model B_T on (a) COTS-A₀ and (b) COTS-A match scores on PCSO dataset. The blue and green bands plot regions of 95% and 99% confidence for subject-specific variations around the population-mean trend. Hence, Model B_T estimates that 95% and 99% of the subject trends fall within the blue and green bands. Thresholds at 0.01% and 0.1% FAR for COTS-A₀ and COTS-A are shown as red lines.

COTS-A, denoted as COTS-A₀, used in [2]. The estimated slopes from Model B_T for COTS-A₀ and COTS-A indicate that genuine scores decrease by one standard deviation of their respective genuine distributions after 7.3 and 6.9 years of elapsed time, respectively. These two estimates are fairly close for the two versions of COTS-A, but suggest that COTS-A₀ may be slightly more robust to aging.

Following [2], using estimated variation in slope and intercept parameters (σ_0^2 , σ_1^2 , and σ_{01}), we plot regions that contain the longitudinal trends for 95% and 99% of the population. The regions are then used to determine when genuine scores for 95% and 99% of the population begin to drop below thresholds for FARs of 0.01% and 0.1%. In other words, we estimate the elapsed time in years which is tolerated by the COTS matchers before the decrease in genuine scores begins to cause errors at different FARs. Best-Rowden and Jain reported that genuine scores for 99% of the population remained above the threshold at 0.01% FAR until 6.5 years for COTS-A₀ on the PCSO dataset, whereas, from Figure 6, we estimate this time lapse to be 10.5 years for COTS-A. Figure 5 shows that the score distribution for COTS-A face matcher has a better separation between impostor and genuine score distributions, compared to COTS-

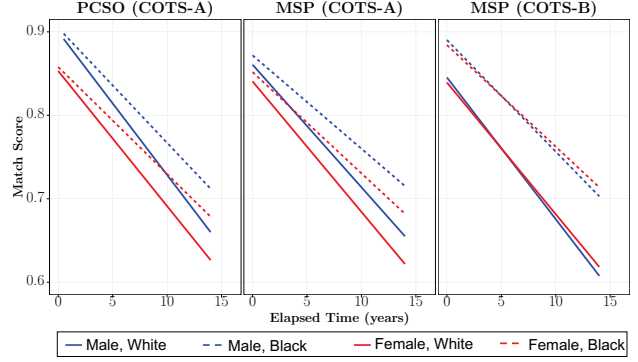


Figure 7: Population-mean trends in COTS-A and COTS-B genuine scores on PCSO dataset and MSP dataset for the four demographic groups in the datasets. Trends were obtained from Model C_{GR} .

A₀, due to lower thresholds at different FARs and a shift in the entire genuine distribution, which can accommodate more for decreasing genuine scores. This may explain the improved longitudinal performance of COTS-A compared to COTS-A₀.

From Figure 6, we find a significant amount of variability in subject-specific longitudinal trends in COTS-A genuine scores on PCSO dataset over time. We observed that large subject-specific variability exists for COTS-A and COTS-B on MSP dataset as well. Therefore, we consider other covariates such as gender and race, and face image quality to further explain this variability and to obtain a better estimate of longitudinal trends due to face aging.

4.2. Gender and Race

This section investigates whether variability in subject-specific longitudinal trends in genuine scores over time can be explained by subject demographics of gender and race. Population-mean trends for gender and race, estimated by Model C_{GR} are shown in Figure 7. Two primary conclusions can be drawn about the effects of demographics on trends in genuine scores over time. Differences due to demographics are (i) consistent for COTS-A on both datasets, but (ii) the demographic effects are matcher-dependent when COTS-A and COTS-B are both evaluated on the MSP dataset.

For COTS-A on PCSO and MSP, average genuine scores are significantly lower for females than for males, but rates of change (slopes) are not statistically different between males and females. We observe the opposite effect for COTS-A with respect to race; average genuine scores are not statistically different between white and black subjects, but rates of change (slopes) are significantly steeper for white subjects than for black subjects. These results are consistent with the effects of demographics on the COTS-A₀ version reported in [2].

Comparing COTS-A and COTS-B on the MSP dataset, we observe that trends between males and females are almost identical for COTS-B. However, COTS-B average genuine scores for black subjects are significantly higher and change at a slower rate than the genuine scores of white subjects. Hence, for the MSP dataset, COTS-B is most affected by race, while COTS-A is most affected by gender. Because COTS-A performed consistently on both PCSO and MSP, the different performance on demographic groups of COTS-A and COTS-B can likely be attributed to the distribution of demographic groups in the training sets used for COTS-A and COTS-B.

4.3. Face Quality

Figure 8 shows mugshot images of subjects from PCSO and MSP datasets whose longitudinal trends estimated by Model B_T lie outside of the 95% confidence band. A majority of these subjects have poor quality face images which can result in low genuine scores. Hence, in this section, we attempt to obtain better estimates of longitudinal performance by accounting for the varying quality of the face images. The raw quality values from Chen *et al.* [20] range from 0 to 100; for our study, we standardize the quality values to have a mean of 0 and standard deviation of 1 for ease of interpretation of the models.

From Table 3, it can be inferred that the model with the best goodness-of-fit for both PCSO and MSP datasets is Model Q_T . Using the equations for Model Q_T in Table 2, the composite form of Model Q_T is written as,

$$\begin{aligned}
 Y_{ij} = & \gamma_{00} + \gamma_{01}Qual_{i,1} + b_{0i} + \\
 & (\gamma_{10} + \gamma_{11}Qual_{i,1} + b_{1i}) \Delta T_{i,j} + \\
 & (\gamma_{20} + \gamma_{21}Qual_{i,1} + b_{2i}) Qual_{i,j \neq 1} + \\
 & (\gamma_{30} + \gamma_{31}Qual_{i,1} + b_{3i}) \Delta T_{i,j} Qual_{i,j \neq 1} + \varepsilon_{i,j}
 \end{aligned} \quad (4)$$

Because we standardize the quality values to have a mean of 0, if we assume average quality of enrollment and probe images, Equation 4 reduces to,

$$Y_{ij} = [\gamma_{00} + b_{0i}] + [\gamma_{10} + b_{1i}] \Delta T_{i,j} + \varepsilon_{i,j}, \quad (5)$$

which is the same as Equation 3 for time lapse. We then investigate the change in COTS-A and COTS-B genuine scores over time, assuming average mugshot quality, by plotting 95% and 99% confidence bands around the population-mean trends in Figure 9. The elapsed times when confidence bands cross thresholds at different FARs are also given in Table 5. Comparing the longitudinal performance estimated by Models B_T and Q_T in Table 5, COTS-B face matcher is impacted the most by accounting for varying face image quality. The Pearson coefficients between the quality values for probe images and genuine scores are 0.35 for COTS-B on MSP, and 0.16 and 0.04 for














Enrollment Image	Query Images age in years (quality of image)			
 34 (57.09)	 37 (60.72)	 38 (54.57)	 40 (52.40)	 47 (52.79)
 33 (72.26)	 34 (40.22)	 37 (39.16)	 39 (42.91)	 43 (43.17)
(a)				
 44 (65.39)	 55 (50.28)	 56 (56.64)	 57 (49.79)	 58 (23.07)
 49 (83.61)	 55 (28.76)	 56 (43.89)	 58 (41.87)	 59 (56.94)
(b)				

Figure 8: Examples of subjects, in (a) PCSO dataset and (b) MSP dataset, whose longitudinal trends estimated by Model B_T lie outside of the 95% confidence band. Age at image acquisition along with quality value of the image (in parentheses) are given. The mean (standard deviation) of the quality distributions for PCSO and MSP datasets are 73.89 (10.29) and 76.38 (10.69), respectively.

COTS-A on PCSO and MSP datasets, respectively. This suggests that COTS-B genuine scores are more correlated with quality values of probe images which may explain why the longitudinal performance estimated by Model B_T is lower than that estimated by Model Q_T for COTS-B. Another explanation may be that COTS-B is more sensitive to low quality images than COTS-A, and hence the performance due to aging significantly changes when low genuine scores caused by face image quality are account for in Model Q_T .

In summary, assuming average quality mugshot images, the genuine scores of 99% of the population remain above the threshold at 0.01% FAR for an elapsed time of 10.5 years for COTS-A on PCSO dataset. Genuine scores of 99% of the population remain above the threshold at 0.01% FAR for an elapsed time of 10.5 (8.5) years for COTS-A (COTS-B) on MSP dataset.

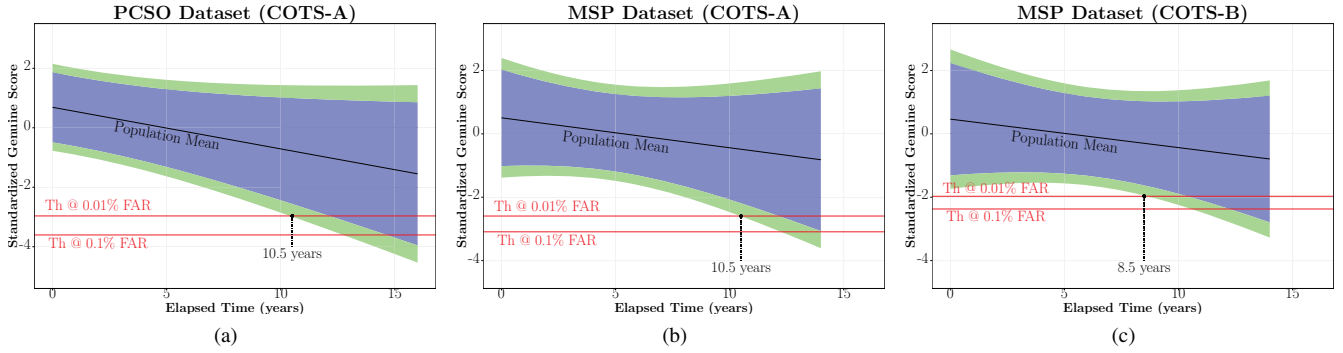


Figure 9: Results from Model Q_T on COTS-A match scores on (a) PCSO and (b) MSP datasets and (c) COTS-B matches scores on MSP dataset. The blue and green bands plot regions of 95% and 99% confidence for subject-specific variations around the population-mean trend. Hence, Model Q_T estimates that 95% and 99% of the subject trends fall within the blue and green bands. Thresholds at 0.01% and 0.1% FAR for COTS-A and COTS-B are shown as red lines.

Table 5: Results from Models B_T and Q_T for COTS-A genuine scores on PCSO and MSP datasets and COTS-B genuine scores on MSP dataset. Values represent time lapse in years tolerated by the matchers before genuine scores for 95/99% of the population drop below thresholds at 0.01% and 0.1% FAR.

(a) Model B_T

	95% Confidence		99% Confidence	
	0.01% FAR	0.1% FAR	0.01% FAR	0.1% FAR
PCSO (COTS-A)	12.0	15.0	10.5	13.0
MSP (COTS-A)	12.0	14.0	9.5	12.5
MSP (COTS-B)	9.0	12.5	5.5	9.5

(b) Model Q_T

	95% Confidence		99% Confidence	
	0.01% FAR	0.1% FAR	0.01% FAR	0.1% FAR
PCSO (COTS-A)	12.0	14.5	10.5	12.5
MSP (COTS-A)	12.0	14.0	10.5	12.5
MSP (COTS-B)	10.5	12.0	8.5	10.5

5. Conclusions

We have presented a longitudinal study of face recognition, using two operational longitudinal face datasets of mugshot images, PCSO (147,784 mugshots of 18,007 subjects, 8 images per subject on average over an average time lapse of 8 years) and MSP (82,450 images of 9,572 subjects, 9 images per subject on average over an average time lapse of 9 years). Each subject has at least 5 and 4 images for PCSO and MSP, respectively, acquired over a time lapse of at least 5 years. Multilevel statistical models were used to analyze variations in genuine scores due to covariates such as time lapse, gender, race, and face quality. Face similarity scores were obtained from state-of-the-art face matcher, COTS-A, for both PCSO and MSP dataset and another state-of-the-art face matcher, deep-network based COTS-B, on MSP dataset. The findings of this paper as follows:

- Differences due to demographics of gender and race are (i) consistent for COTS-A evaluated on both datasets, but (ii) matcher-dependent when COTS-A and COTS-B are both evaluated on the MSP dataset.
- Accounting for varying face image quality significantly impacted the estimated longitudinal performance for the weaker matcher COTS-B, but did not affect the estimated longitudinal performance of COTS-A.
- Assuming average quality of mugshot images, we estimate the longitudinal performance of the state-of-the-art COTS face matchers to be the following: (i) Genuine scores of 99% of the population remain above the threshold at 0.01% FAR for an elapsed time of 10.5 years for COTS-A on PCSO dataset. (ii) Genuine scores of 99% of the population remain above the threshold at 0.01% FAR for an elapsed time of 10.5 (8.5) years for COTS-A (COTS-B) on MSP dataset. These results are summarized in Table 5b also for 0.1% FAR and 95% of the population.

Future work will include: (i) In this study, we evaluated face recognition performance over time only in verification scenarios. This needs to be repeated for face identification performance over time. (ii) Analyzing rates of change of comparison scores for original face images versus rates of change of comparison scores for age-progressed or age-simulated face images. A longitudinal study such as ours needs to be conducted periodically to assess current state-of-the-art in age-invariant face recognition.

References

- [1] B. Klare and A. K. Jain. Face recognition across time lapse: On learning feature subspaces. In *Proc. IJCB*, 2011. 2
- [2] L. Best-Rowden and A. K. Jain. Longitudinal Study of Automatic Face Recognition. To appear in the *IEEE Trans. Pattern Analysis & Machine Intelligence*, 2017. 2, 3, 5, 6
- [3] C. Otto, H. Han, and A. K. Jain. How does aging affect facial components? In *ECCV WIAF Workshop*, 2012 2
- [4] 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 2
- [5] P. Grother and M. Ngan. FRVT: Performance of face identification algorithms. *NIST Interagency Report 8009*, May 2014. 2
- [6] U. Park, Y. Tong, and A. K. Jain. Age-invariant face recognition. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 32(5):947 -954, 2010. 1
- [7] Z. Li, U. Park, and A. K. Jain. A discriminative model for age invariant face recognition. *IEEE Trans. Information Forensics and Security*, 6(3):1028-1037, 2011. 1
- [8] J. Suo, S.-C. Zhu, S. Shan, and X. Chen. A compositional and dynamic model for face aging. *IEEE Trans. Pattern Analysis & Machine Intelligence*, 32(3):385-401, 2010. 1
- [9] H. Yang, D. Huang, Y. Wang, H. Wang, and Y. Tang. Face aging effect simulation using hidden factor analysis joint sparse representation. *IEEE Transactions on Image Processing*, 25(6): 2493-2507, 2016. 1
- [10] G. Guo and G. Mu. Human age estimation: What is the influence across race and gender? In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, 2010. 1
- [11] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, and M. King. Aspects of age variation in facial morphology affecting biometrics. In *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2007. 1
- [12] N. Ramanathan and R. Chellappa. Modeling Age Progression in Young Faces. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2006 1
- [13] 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. 2
- [14] K. W. Bowyer and E. Ortiz. Critical examination of the IREX VI results. *IET Biometrics*, vol. 4, pp. 192-199, 2015. 2
- [15] P. Grother, J. R. Matey, and G. W. Quinn. IREX VI: Mixed-effects Longitudinal Models for Iris Aging: Response to Bowyer and Ortiz. *IET Biometrics*, Nov. 2015. 2
- [16] S. Yoon and A. K. Jain. Longitudinal Study of Fingerprint Recognition. *Proc. National Academy of Sciences (PNAS)*, Vol. 112, No. 28, pp. 8555-8560, July 2015. 2, 5
- [17] A. Lanitis, C. J. Taylor, and T. F. Cootes. Toward automatic simulation of aging effects on face images. In *IEEE Trans. on PAMI*, vol. 24, no. 4, Apr. 2002. 2
- [18] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *FGR 2006: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, vol. 2006, pp. 341-345. 2
- [19] 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. 2
- [20] J. Chen, Y. Deng, G. Bai, G. Su. Face image quality assessment based on learning to rank. *IEEE Signal Processing Letters*, 22(1): 90-94, 2015. 2, 4, 7