Analysis of Facial Images across Age Progression by Humans

Jingting Zeng, Haibin Ling, Longin Jan Latecki, Shanon Fitzhugh, and Guodong Guo

1 Department of Computer and Information Sciences, Temple University, Philadelphia, PA 19122, USA
2 Department of Psychology, Temple University, Philadelphia, PA 19122, USA
3 CSEE Department, West Virginia University, Morgantown, WV 26506, USA

Correspondence should be addressed to Haibin Ling, hbling@temple.edu

Received 25 July 2011; Accepted 25 August 2011

Academic Editors: O. Ghita and R.-H. Park

Copyright © 2012 Jingting Zeng et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The appearance of human faces can undergo large variations over aging progress. Analysis of facial images taken over age progression recently attracts increasing attentions in computer-vision community. Human abilities for such analysis are, however, less studied. In this paper, we conduct a thorough study of human ability on two tasks, face verification and age estimation, for facial images taken at different ages. Detailed and rigorous experimental analysis is provided, which helps understanding roles of different factors including age group, age gap, race, and gender. In addition, our study also leads to an interesting observation: for age estimation, photos from adults are more challenging than that from young people. We expect the study to provide a reference for machine-based solutions.

1. Introduction

Human faces are important in revealing the personal characteristic and understanding visual data. The facial research has been studied over several decades in computer vision community [1, 2]. Analysis facial images across age progression recently attracts increasing research attention [3] because of its important real-life applications. For example, facial appearance predictor of missing people and ID photo automatic update system are playing important roles in simulating face aging of human beings. Age estimation can also be applied to age-restricted vending machine [4]. Most recent studies (see Section 2) of age-related facial image analysis mainly focus on three tasks: face verification, age estimation, and age effect simulation. In comparison, it remains unclear how humans perform on these tasks.

In this paper, we study human ability on face verification and age estimation for face photos taken at across age progression. Such studies are important in that it not only provides a reference for future machine-based solutions, but also provides insight on how different factors (e.g., age gaps, gender, etc.) affect facial analysis algorithms. There are previous works on human performance for face recognition and age estimation; however, most of them are either focusing on nonage related issues such as lighting [5] or limited by the scale of image datasets (e.g., [6]). Taking advantage of the recent available MORPH dataset [7], which to the best of our knowledge is the largest publicly available face aging dataset, we are able to conduct thorough human studies on facial analysis tasks.

For face verification, the task is to let a human subject decide whether two photos come from the same person (at different ages). In addition to report the general performance on our human subjects’ performance, we also analyze the effects of difference factors, including age group, age gap, race, and gender. In addition, we also compare human performance with previous reported baseline algorithm. For age estimation, similarly, we report and analyze human performance for general cases as well as for different factors.

The rest of the paper is organized as follows. Section 2 shows the related works on different databases. Section 3 describes the details of human experiments of face-recognition and age-estimation problems. Then, in Section 4,
the results are compared to the existing results of human experiments and computer algorithms. The conclusion is given in Section 5.

2. Related Works

Face recognition across age progression and age estimation have been studied widely in recent years. Large numbers of algorithms have been implemented based on different databases [9]. One of the earliest works in Lanitis et al. [3] uses a statistical model to capture the variation of facial shapes over age progression. Then, the model is used for age estimation and face recognition on a database containing 500 face images of 60 subjects. In [10], Ramanathan and Chellappa use the probabilistic eigenspace framework for face recognition. Ling et al. [11] propose gradient orientation pyramids operators derived from multiple resolutions and then uses SVM to perform face recognition experiments. These two algorithms are conducted on a private passport database. A recent work in Biswas et al. [12] studies feature drifting on face images at different ages and applies it to face-verification tasks. Other studies using age transformation for recognition include [13–16].

For age-estimation problem, Fu and Huang [17] construct a low-dimensional manifold from a set of age-separated face images to estimate the ages of faces. Manifold learning approach adopted in Guo et al. [18] is to estimate the age from the low-dimensional representation of faces. Hybrid features are recently included to further improve the estimation accuracy [19]. Other related researches in age estimation can be found in [20–23].

A major issue in the research of age-related facial image analysis is the database. For a long time, the FGNet face aging database [8], which is collected by Lanitis and colleagues, is the only publicly available database dedicated to face aging study (some other public databases, e.g., the FERET database [24], also contain limited number of facial photos of the same person taken at different ages, and they are designed for more general facial analysis tasks other than aging). The dataset contains about 1000 facial photos from 82 subjects taken at different ages. Since its introduction, the FGNet database has been widely used in face aging analysis [25–27]. Recently, Ricanek and Tesařy introduce the MORPH Database [7], which contains across age photos from a large amount of subjects (see Section 3). Albert and Ricanek [28] implemented a baseline facial verification on the MORPH database using the eigenface algorithm. In our experiment, the MORPH database is used, because it involves more subjects as well and larger age ranges than other public and private databases.

The most related works to ours are previous studies of human ability for face recognition (with age progression) and age-estimation. In [25], 30 subjects participated with 100 pairs of face images randomly selected from the FGNet database. For each test pair, subjects were requested to tell if the two images are from the same person and if the images belong to the same age group. The human performance is compared to Support Vector Machine classifiers using Mahalanobis distance. Experiments demonstrated that the SVM classifiers could perform better than the human performances. Geng et al. [6] collected the human performance of an age-estimation experiment, where 29 subjects were asked to estimate the ages of images from the FGNet database. Then, they provided a method by learning a representative subspace to model the aging pattern. Experiments showed promising performances to the results of human experiments. Compared to previous work, our study is more thorough in several aspects: (1) a much larger database is used, (2) we conducted experiments on both face-verification and age-estimation tasks, and (3) rigorous statistical analysis of the experimental results is provided.

3. Human Experiments Implementation

3.1. Data and Subjects. To conduct a thorough study, we use facial images from the recently available MORPH face aging dataset [7]. The dataset contains two albums; we use the first one (i.e., MORPH Album 1), since a baseline face verification result is provided on it [28]. MORPH Album 1 contains 1690 facial photos of 515 individuals with ages in the range of 15–68 years (see Figure 1 for typical examples). All images are in gray scale. We have removed 92 blurred and noisy photos (see, e.g., Figure 2). In addition to ages, the metadata in the MORPH dataset also provides detailed information about the subject (e.g., gender and race) and the photographic conditions (e.g., pose, lighting conditions, and image quality).

One interesting difference between the MORPH dataset and the FGNet dataset is that photos from MORPH are mainly from adult persons, while most photos from FGNet are dominated by children. In this way, by comparing with previous work on FGNet, our study on MORPH provides a comparison of the performances on photos from children versus those from adults.

For experimental subjects, thirty one participants were recruited from Temple University. All the subjects were naive
3.2. Face-Verification Experiment. In the face-verification experiment, each participant attends 300 trials, and therefore, the total number of the trials is $300 \times 31 = 9300$. In each trial, a pair of face images is randomly selected from the database and then presented to the participant. The participant is requested to decide if the two photos come from the same person. Among the 300 trials, about 30% are from the same persons. The user interface of the experiment is shown in Figure 3(a). When a pair of photos is shown, the participant is required to click either the “Same” button, if they think the two photos are of the same person, or the “Diff” button, if they think otherwise. The choice and reaction time of each trial is recorded.

3.3. Age-Estimation Experiment. In the age-estimation experiment, similar to the verification experiment, each participant attends 300 trials, and there are 9300 trials in total. In each trial, a participant is requested to estimate the age of a photo randomly selected from the whole database. The user interface of the experiment is presented in Figure 3(b). Given a photo, the participant is request to “choose” among buttons named “1” to “80”, corresponding to age 1 to 80, respectively. Note that the actual age range of the MORPH dataset is 15–68. We purposely allow a participant to chose from a larger range to avoid bias.

4. Experimental Results

4.1. Face Verification across Age Progression

4.1.1. Face-Verification Performance. To measure the face-verification performance, we report the average face-verification accuracy over all the participants. The accuracy represents the participants’ correct selections—“Same” for photos from the same person or “Diff” for those from different persons. In addition to the performance on the whole dataset, we also report the performance on the subgroups of the dataset to study how gender and race affects human accuracy. These subgroups include African American versus European American and male versus female. The results are summarized in Table 1. Overall, from the table, we see that neither gender nor race has significant effects on the verification task. Several examples that are incorrectly verified during the test are given in Figure 4 (false positive) and Figure 5 (false negative).

A similar human study was reported by Lanitis [25] on the FGNet database, where an overall accuracy of 66.9% is achieved for images in similar conditions as ours (i.e., cropped grayscale images). Given the fact that FGNet contains much more child photos than does MORPH, this observation shows that verification of a person using child and adult photos is more challenging than only adult photos.

| All Data | African American | European American | Male | Female |
|----------|------------------|-------------------|------|--------|
| 0.788    | 0.790            | 0.781             | 0.788| 0.787  |

Table 1: Face-verification accuracies of human experiments.
Figure 4: Some typical photos that lead to false-positive input (participants chose "same" when the pairs actually are not the same).

Figure 5: Some typical photos that lead to false-negative input (participants chose "Diff" when the pairs actually are the same).

### Table 2: Face-verification accuracy results of human and a baseline algorithm [28].

| Age Group | Gap 0–5 | Gap 6–10 | Gap 11–15 | Gap 16+ |
|-----------|---------|----------|-----------|---------|
| <18       | 0.775   | 0.744    | 0.751     | 0.789   |
| 18–29     | 0.778   | 0.740    | 0.753     | 0.782   |
| 30–39     | 0.821   | 0.803    | 0.830     | 0.821   |
| 40+       | 0.750   | 0.800    | *         | *       |

Note: (1) "|" means "accuracy of human experiment | accuracy of PCA algorithm"; (2) ** indicates there are not enough data for evaluation. The same notations apply for other tables as well.

We also compare the result on MORPH using baseline machine solutions reported in [28]. In their experiment, a standard eigenface algorithm was implemented and achieved an overall accuracy of 38.1%, which is much lower than the human performance (78.8%).

To further study roles of different factors in the verification performance, we follow [28] to divide the data into several groups according to different criteria, including ages, age gaps, races, and genders. Specifically, we divide the photos into four age groups: younger than 18 (<18), from 18 to 29 (18–29), from 30 to 39 (30–39), and older than age 40 (≥40). The age gap is defined as the absolute age difference between two photos from the same person. Age gaps between the first photos and the second photos in our experiment are also divided into four gap groups: less than 6 years (named gap 0–5), from 6 to 10 years (named gap 6–10), from 11 to 15 years (named gap 11–15), and more than 16 years (named gap 16+). The verification accuracies on each groups in our experiment are reported in Tables 2–6 to show the effect of different ages and age gaps on the accuracies. In particular, Table 2 shows the verification accuracies for each joint age and age gap group. Tables 3–6 show these accuracies for African American photos, European American photos, American male photos, and American female photos, respectively. In addition, the results from [28] are included for comparison wherever available.

In general, our experiment shows that elder persons are better verified than younger persons. For example, the accuracy of age group 18–29 is higher than the group <18 but lower than the group 30–39. This observation is consistent with previous studies in human and machine face recognition [1].

By comparing Table 3 versus Table 4 and Table 5 versus Table 6, we observe that the races and genders do not affect the human ability of face verification. This is consistent with the conclusion drawn from Table 1.

### 4.1.2. Statistical Analysis.

We now analyze the result statistically. Specifically, we investigate the relationship between the user-related information and the photo-related profile, such as the race, gender, and the age gap of pairs of photos. The user-related information includes the correctness of the participant response and the reaction time of the response. Multiway ANOVA test is adopted for this purpose. Because the accuracy of the participant response is the most important, it is used as the dependent variable. Independent factors include the mutual effect of the gender and race of the photos, the participant response time (in seconds), the age gap, and the gap group of photo pairs. The ANOVA result is shown in Table 7. It shows that $P < 0.0001$ in the model,
For this, we use one-way ANOVA. In this model, the overall effect of independent variables on dependent variables. Here, the response time has significant effect on the correctness. The mutual effect of race and gender is also playing a major factor in human performance, and the age factor of the photo is important too.

We are also interested in whether the response time is affected by the mutual effect of the race and the gender. For this, we use one-way ANOVA. In this model, the overall F-value is 12.51, \( P = 0.0004 \). The partial F-value of the dependent variable (the mutual effect) is 12.51, \( P = 0.0004 \). Thus, both the model and the dependent variable are statistically significant, which indicates that race and gender do affect the response time.

### 4.2. Age-Estimation Experiment

#### 4.2.1. Age-Estimation Performance

To measure the age-estimation performance, we use the same criterion, \textit{mean absolute error} (MAE), used in previous studies [3, 6, 9]. Formally, MAE is defined as

\[
\text{MAE} = \frac{1}{n} \sum_{k=1}^{n} | \text{Estimated Age}_k - \text{True Age}_k | ,
\]

where Estimated Age\(_k\) and True Age\(_k\) are the estimated age and the true age, respectively, of the photo used in the \( k \)-th trial and \( n \) is the number of trials. The experiment result is shown in Table 8, where it shows that the average MAE is 8.58 years. Some example photos with large estimation errors are shown in Figure 6.

For comparison, the table also lists the MAE on the FGNet database [6], in which 30 participants attended a similar human experiments. In [6], the experiments are conducted on two types of images: one with cropped grayscale images and the other with uncropped color images. The cropped grayscale images are similar to images in the MORPH dataset. Table 8 shows that the MAE on the MORPH dataset (our experiment) is larger than those on the FGNet dataset. Given the fact that FGNet contains photos that are in average much younger than those in MORPH, we conclude that ages of adults are harder to be estimated than do ages of children.

To further investigate the gender and race factors in the age-estimation task, we summarize the MAE for each subgroup in Table 9. It shows that ages of European Americans are, in average, harder to be estimated than ages of African Americans. Moreover, it also shows that it is harder to estimate ages of male photos than of female photos. The significance of the conclusion is statistically verified in the following subsection.

#### 4.2.2. Statistical Analysis

We use statistical model to analyze the age-estimation result. In the multiway ANOVA analysis model, the dependent variable is the absolute difference between a participant’s answer and the actual age of the trial photo. The gender, race, and the response time in seconds are independent variables. The result is reported in Table 10. As shown in the table, the gender and race
separately have significant effect on the dependent variable though the participant response time is not related. This confirmed our observation that genders and faces do affect the performance of age estimation.

4.3. Summary of Observations. We summarize the observations in the above experiments, in addition to the performance scores, as below.

(i) For humans, face verification for young people is more difficult than for senior people.

(ii) Human ability for age estimation is affected by the race and gender of the person in the photo to be estimated. In average, for age estimation, photos of European Americans are harder than those of African Americans; photos of male subjects are harder than photos of female subjects.

(iii) For humans, it is relative easier to estimate ages of children than to estimate ages of adults. This could be due to the fact that facial profiles usually stay stable after age 18 [29].

5. Conclusion

In this paper, we describe our study of human ability of facial image analysis for photos across age progression. We implemented two experiments including face-verification task and age-estimation task. In addition to report the overall performance, we also investigate how different factors affect human performance. These factors include the age of person in a photo, the age gaps at which two photos are generated, the gender, and the race. We expect our study to provide a reference for future studies of related topics.

Acknowledgments

The authors thank Thomas F. Shipley for helpful suggestions on the analysis of data, Karl Ricanek and A. Midori Albert for providing the MORPH dataset, and anonymous reviewers for insightful comments.

References

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, “Face recognition: a literature survey,” ACM Computing Surveys, vol. 35, no. 4, pp. 399–458, 2003.

[2] N. Ramanathan, R. Chellappa, S. Biswas et al., “Age progression in human faces: a survey,” Journal of Visual Languages and Computing, vol. 15, pp. 3349–3361, 2009.

[3] A. Laniit, C. J. Taylor, and T. F. Cootes, “Toward automatic simulation of aging effects on face images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 4, pp. 442–455, 2002.

[4] BBC, “Japanese smokers to face age test,” 2008, http://news.bbc.co.uk/2/hi/asia-pacific/7395910.stm.

[5] A. J. O’Toole, P. Phillips, F. Jiang, I. Ayyad, N. Pénard, and H. Abdi, “Face recognition algorithms surpass humans matching faces over changes in illumination,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 9, pp. 1642–1646, 2007.

[6] X. Geng, Z. H. Zhou, Y. Zhang, G. Li, and H. Dai, “Learning from facial aging patterns for automatic age estimation,” in Proceedings of the 14th Annual ACM International Conference on Multimedia, (MM ’06), pp. 307–316, Santa Barbara, Calif., USA, October 2006.

[7] K. Ricanek and T. Tesafaye, “MORPH: a longitudinal image database of normal adult age-progression,” in Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition, (FGR ’06), pp. 341–345, Southampton, UK, April 2006.

[8] “FGNet aging database,” http://www.fgnet.rsunit.com/.

[9] A. Laniit, C. Draganova, and C. Christodoulou, “Comparing different classifiers for automatic age estimation,” IEEE Transactions on Systems, Man, and Cybernetics, Part B, vol. 34, no. 1, pp. 621–628, 2004.

[10] N. Ramanathan and R. Chellappa, “Face verification across age progression,” IEEE Transactions on Image Processing, vol. 15, no. 11, pp. 3349–3361, 2006.

[11] H. Ling, S. Soatto, N. Ramanathan, and D. W. Jacobs, “Face verification across age progression using discriminative methods,” IEEE Transactions on Information Forensics and Security, vol. 5, no. 1, Article ID 5353681, pp. 82–91, 2010.

[12] S. Biswas, G. Aggarwal, N. Ramanathan, and R. Chellappa, “A Non-generative approach for face recognition across aging,” in Proceedings of the 2nd IEEE International Conference on Biometrics: Theory, Applications and Systems, (BTAS ’08), pp. 1–6, Washington, DC, USA, October 2008.

[13] R. Singh, M. Vatsa, A. Noore, and S. K. Singh, “Age transformation for improving face recognition performance,” in Proceedings of the 2nd International Conference on Pattern Recognition and Machine Intelligence, (PReMI ’07), pp. 576–583, Kolkata, India, 2007.

[14] J. Wang, Y. Shang, G. Su, and X. Lin, “Age simulation for face recognition,” in Proceedings of the 18th International Conference on Pattern Recognition, (ICPR ’06), pp. 913–916, Hong Kong, China, August 2006.

[15] U. Park, Y. Tong, and A. K. Jain, “Face recognition with temporal invariance: a 3D aging model,” in Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition, (FG’08), pp. 1–7, Amsterdam, The Netherlands, September 2008.

[16] E. Patterson, A. Sethuram, M. Albert, K. Ricanek, and M. King, “Aspects of age variation in facial morphology affecting biometrics,” in Proceedings of the 1st IEEE International Conference on Biometrics: Theory, Applications, and Systems, (BTAS ’07), Crystal City, Va, USA, September 2007.

[17] Y. Fu and T. S. Huang, “Human age estimation with regression on discriminative aging manifold,” IEEE Transactions on Multimedia, vol. 10, no. 4, Article ID 4523958, pp. 578–584, 2008.

[18] G. Guo, Y. Fu, C. R. Dyer, and T. S. Huang, “Image-based human age estimation by manifold learning and locally adjusted robust regression,” IEEE Transactions on Image Processing, vol. 17, no. 7, pp. 1178–1188, 2008.

[19] G. Guo, G. Mu, Y. Fu, C. Dyer, and T. Huang, “A study on automatic age estimation using a large database,” in Proceedings of the 12th International Conference on Computer Vision, (ICCV ’09), pp. 1986–1991, Kyoto, Japan, October 2009.
[20] S. K. Zhou, B. Georgescu, X. S. Zhou, and D. Comaniciu, “Image based regression using boosting method,” in Proceedings of the 10th IEEE International Conference on Computer Vision, (ICCV ’05), vol. 1, pp. 541–548, Beijing, China, October 2005.

[21] S. Yan, H. Wang, X. Tang, and T. S. Huang, “Learning auto-structured regressor from uncertain nonnegative labels,” in Proceedings of the 11th IEEE International Conference on Computer Vision, (ICCV’07), Rio de Janeiro, Brazil, October 2007.

[22] Y. H. Kwon and N. Da Vitoria Lobo, “Age classification from facial images,” Computer Vision and Image Understanding, vol. 74, no. 1, pp. 1–21, 1999.

[23] A. Montillo and H. Ling, “Age regression from faces using random forests,” in Proceedings of the IEEE International Conference on Image Processing, (ICIP ’09), pp. 2465–2468, Cairo, Egypt, November 2009.

[24] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 10, pp. 1090–1104, 2000.

[25] A. Lanitis, “Evaluating the performance of face-aging algorithms,” in Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition, (FG ’08), pp. 1–6, Amsterdam, The Netherlands, September 2008.

[26] N. Ramanathan and R. Chellappa, “Modeling age progression in young faces,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR ’06), pp. 387–394, New York, NY, USA, June 2006.

[27] J. Suo, F. Min, S. Zhu, S. Shan, and X. Chen, “A multi-resolution dynamic model for face aging simulation,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR ’07), Minneapolis, Minn, USA, June 2007.

[28] A. M. Albert and K. Ricanek, “The MORPH database: investigating the effects of adult craniofacial aging on automated face-recognition technology,” Forensic Science Communications, vol. 10, p. 2, 2008.

[29] J. B. Pittenger and R. E. Shaw, “Aging faces as visco-elastic events: implications for a theory of nonrigid shape perception,” Journal of Experimental Psychology, vol. 1, no. 4, pp. 374–382, 1975.
Submit your manuscripts at
http://www.hindawi.com