

How Does Aging Affect Facial Components?

Charles Otto, Hu Han, and Anil Jain

Michigan State University
{ottochar,hhan,jain}@cse.msu.edu

Abstract. There is growing interest in achieving age invariant face recognition due to its wide applications in law enforcement. The challenge lies in that face aging is quite a complicated process, which involves both intrinsic and extrinsic factors. Face aging also influences individual facial components (such as the mouth, eyes, and nose) differently. We propose a component based method for age invariant face recognition. Facial components are automatically localized based on landmarks detected using an Active Shape Model. Multi-scale local binary pattern and scale-invariant feature transform features are then extracted from each component, followed by random subspace linear discriminant analysis for classification. With a component based representation, we study how aging influences individual facial components on two large aging databases (MORPH Album2 and PCSO). Per component performance analysis shows that the nose is the most stable component during face aging. Age invariant recognition exploiting demographics shows that face aging has more influence on females than males. Overall, recognition performance on the two databases shows that the proposed component based approach is more robust to large time lapses than FaceVACS, a leading commercial face matcher.

Keywords: face recognition, aging, facial components, demographics.

1 Introduction

Automatic face recognition has attracted much attention due to its widespread potential applications in homeland security and law enforcement [1]. Major challenges in designing a robust face recognition system include variations in lighting, expression, head pose, and age. A number of approaches have been proposed to achieve lighting and/or pose invariant face recognition. Among these methods, novel appearance synthesis and discriminative feature extraction are two of the popular approaches [2–5].

Unlike other sources of variation (lighting, pose, and expression) which can be controlled during face image acquisition, face aging is an unavoidable natural process during the lifespan of a person. There has been growing interest in achieving age invariant face recognition to meet the requirements of several applications in law enforcement and forensic investigation (e.g. de-duplication of driver's licenses, identifying missing children). In these applications, the age

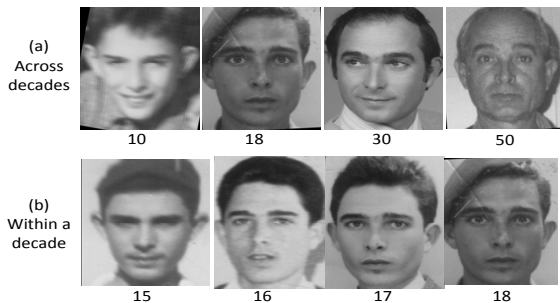


Fig. 1. Face aging of a subject in the FG-NET [15] database. (a) Face aging across several decades. (b) Face aging within a decade.

difference between probe and gallery face images from the same subject becomes one of the main challenges for face verification.

Age variation in face recognition can be handled in a manner similar to illumination and pose variations by using novel appearance synthesis or discriminative feature extraction [6]. For example, synthesis methods can eliminate age gaps between face images by synthesizing facial appearances via learning the face aging process [7–9]. One disadvantage of this approach is that the synthesized face image is just a pseudo-photo due to its inferred content, and the low quality of that inferred content can limit recognition performance. Discriminative feature extraction methods seek to extract facial features that are robust to age variations [10–13]. In these methods [7, 8, 10], faces are usually represented holistically. However, intrinsic and extrinsic factors have varying influence on different facial components [14]. This suggests that a component based face representation should provide better understanding of the face aging process, and facilitate an analysis of the role of individual facial components in age invariant face recognition.

While there are large variations in facial appearance across decades (see Figure 1 (a)), facial appearance within a decade may not always be clearly visible (see Figure 1 (b)). This observation suggests learning an age gap specific model for robust face recognition [10, 16]. However, other demographic factors, e.g. gender and race, which can also be inferred from faces, have not been adequately exploited in age invariant face recognition. In this paper, we present a study on the influence of aging on different facial components and demographic groups.

2 Data Sets

We use two face data sets in our study: the public domain MORPH album 2 data set [17], and a dataset comprised of mug shots collected in the state of Florida by the Pinellas County Sheriff’s office (referred to here as the PCSO data set). The PCSO dataset was acquired in the public domain through Florida’s “Sunshine” laws. Both data sets provide ground truth information on the race and gender

Group	PCSO			MORPH	
	0-1	1-5	5-10	0-1	1-5
White-Male	2,000	2,000	2,000	900	450
White-Female	2,000	2,000	2,000	900	450
Black-Male	2,000	2,000	2,000	900	450
Black-Female	2,000	2,000	2,000	900	450
Hispanic-Male	2,000	2,000	557	900	450
Total	10,000	10,000	8,557	4,500	2,250

(a)

(b)

Fig. 2. Numbers of subjects sampled from different demographic categories and age gaps (a), and (b) example images from MORPH (left) and PCSO (right) databases

of subjects, and contain significant age gaps between multiple face images of a given subject.

In order to study the influence of face aging on different facial components and demographic groups, we constructed subsets of the MORPH and PCSO databases containing specific age lapses between the probe and gallery sets. Subjects were further restricted by demographics to build data sets containing relatively balanced numbers of samples from several race and gender categories: white male, white female, black male, black female, and Hispanic male. The number of samples available for other demographic groups fell off sharply, leaving too few subjects to perform reasonable training and evaluation for our study.

We used data sets exhibiting 0-1, 1-5, and 5-10 year age gaps, and randomly sampled subjects from each demographic category. The numbers of subjects sampled from each demographic category are shown in Figure 2. MORPH Album2 contained no subjects exhibiting an age gap larger than 5 years, so we could only build 0-1 and 1-5 year age gap subsets for MORPH. For each subject, we sampled two images with the desired age gap. Half of the subjects in each subset are used for training and the rest are used for evaluation.

3 Face Representation

Our representation is a combination of random subspace linear discriminant analysis (RS-LDA) based classifiers [18] trained on two types of local descriptors, and is generally analogous to the multi-feature discriminative analysis (MFDA) scheme proposed in [13]. In contrast to MFDA, we apply RS-LDA to local features extracted from explicitly detected facial components (following the component localization scheme described in [19]), rather than horizontal slices of the face. Following RS-LDA for each component and feature type, we perform PCA and then train multiple LDA classifiers on subspaces sampled from the PCA dimensions. For computational efficiency we train a single LDA classifier for each sampled PCA subspace rather than using bagging to train multiple LDA classifiers per PCA subspace, as done in MFDA.

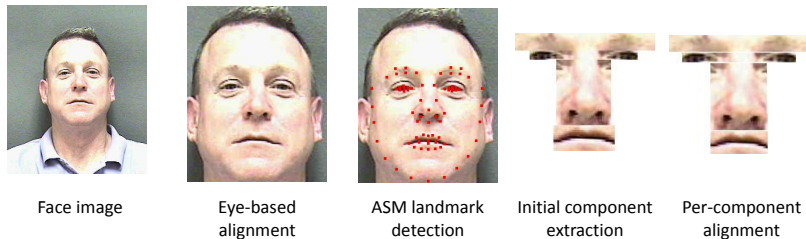


Fig. 3. Component extraction stages

3.1 Component Localization

Figure 3 shows the stages of our component localization process, which is essentially the same as the one described in [19]. We start by aligning all face images based on their eye locations, which we detect using the FaceVACS SDK [24]. After the initial alignment, we detect 76 facial landmarks automatically using Stasm [20], an open source Active Shape Model (ASM) [21] implementation. Stasm is designed to work on upright frontal faces, and while we are working with frontal face images, they do exhibit minor pose variations. The goal of our initial alignment step is to eliminate those pose variations, thereby improving Stasm’s landmark detection accuracy.

We localize components based on subsets of landmarks corresponding to the eyes, nose, mouth, and eyebrows. Due to the initial normalization step, the eyes are typically well aligned, but the remaining three components (nose, mouth, and eyebrow) are not. We, therefore, perform a Procrustes alignment for each component on the corresponding subset of landmarks across all images. This per-component normalization results in fixed-size sub-images for each component.

Figure 4 shows component sub-images averaged across all subjects in the PCSO 0-1 year age gap subset. The blue and green ellipses in Figure 4 illustrate differences between the average components across different races and genders, respectively. The sharpness of the average component images for different demographic categories demonstrates the effectiveness of per-component alignment.

3.2 Feature Extraction

For each facial component, we densely sample both MLBP [22] and SIFT [23] descriptors, using a patch size of 16x16 with 8 pixels overlap between neighboring patches. For a given feature type (MLBP or SIFT), features extracted from all patches are concatenated. Thus, a face image is represented by 8 different feature vectors (4 components \times 2 feature descriptors).

3.3 Training

We use random subspace linear discriminant analysis (RS-LDA) [18] to compensate for the small number of samples per subject, similarly to MFDA [13]. We

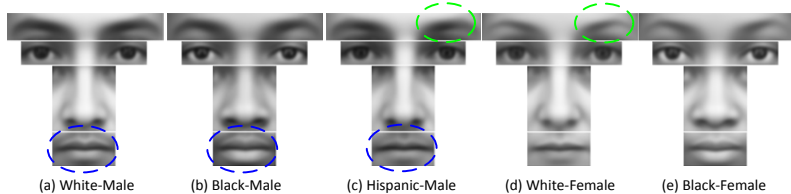


Fig. 4. Average component images for the subjects in the PCSO database for different demographic groups: (a) White-Male; (b) Black-Male; (c) Hispanic-Male; (d) White-Female, and (e) Black-Female. The blue and green ellipses highlight differences between the average components across different races and genders, respectively

carry out training separately for each component, and for both feature types we first apply PCA on the concatenated feature vectors in the training set for dimensionality reduction. Then we build 10 different subspaces, each with 300 dimensions by sampling from the PCA dimensions. For each subspace, the first 200 dimensions are fixed as the eigenvectors that correspond to the 200 highest eigenvalues, and the remaining 100 dimensions are randomly selected from the top 1,000 PCA eigenvectors. Finally, we perform LDA on each random subspace, leaving us with a total of 80 LDA subspaces (4 components \times 2 feature descriptors \times 10 random subspaces). In our experiments, training subjects are disjoint from testing subjects.

3.4 Matching

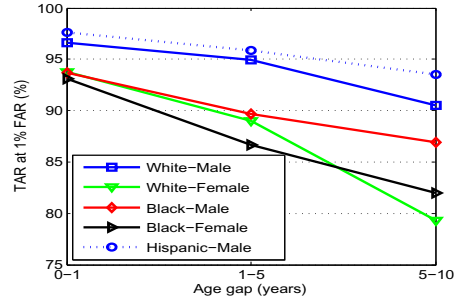
For the face images in probe and gallery sets, the extracted features are projected onto the learnt LDA subspaces. The similarity between two corresponding components is measured using the cosine similarity. To arrive at a final similarity between two faces, we first combine the scores from all LDA subspaces from a given component and descriptor type, giving us different 8 scores (one for each descriptor and component combination). We take the combined LDA subspace scores and combine both scores for each component, leaving us with 4 scores (one for each facial component). These per component scores are then combined to give a final similarity measure between two faces. We use min-max normalization followed by the sum rule for each stage of score fusion. The minimum and maximum values used in the normalization steps are the minimum and maximum scores seen when comparing all face images in the training set.

4 Results and Analysis

We evaluate our component based algorithm and FaceVACS by performing all vs. all matching, then calculating receiver operating characteristic (ROC) curves. We report either true accepts rates (TAR) attained at a fixed false accept rate (FAR), or full ROC curves. Results for specific demographic groups are based on all vs. all matching between all testing subjects within those groups.

Table 1. Per component TAR (%) at 1% FAR across time lapses on PCSO

Age gaps	Eyes	Nose	Mouth	Eyebrows
0-1 year	79.78	92.48	79.10	83.86
1-5 year	76.06	89.68	68.54	73.66
5-10 year	65.83	81.70	55.41	59.76

**Fig. 5.** Per demographic performance across three age gaps on PCSO

4.1 Per Component Performance Analysis

Table 1 shows true accept rates at 1% FAR to summarize the verification performance of individual facial components across different time lapses on the PCSO data. These results show that as the time lapse increases, the accuracy of each component decreases. The nose is the most stable component across all time lapses, which is consistent with our intuition about face aging. The influence of aging on the other components is not uniform across different time lapses. On the 0-1 year age gap set, the eyes are the second worst performing component, with a TAR less than 1% higher than the mouth's, but on the 1-5 and 5-10 year lapse data sets the eyes gain in performance relative to the other components.

4.2 Per Demographic Performance Analysis

Tables 2, 3, and 4 contain face verification results on subsets of the PCSO data sets containing only specific demographic groups. Both our method and FaceVACS generally have higher accuracy on males than females, and higher accuracy on whites and Hispanics than blacks. These results are consistent with the results reported in [24], and with most algorithms evaluated in [25]. There are a couple of exceptions to these trends, on the 0-1 year age gap FaceVACS is slightly more accurate on black females than black males, while on the 5-10 year age gap our method has higher accuracy for black females than white females. For most groups our accuracy is fairly close to FaceVACS, but the component based method performs significantly worse on white females across all time gaps. Figure 5 demonstrates the loss in verification accuracy of the component based method for each demographic group as the time lapse increases. All demographic groups lose accuracy as the age gap increases; however, males lose less accuracy than females.

Observing the performance of individual components across demographics, we see some variations along gender lines. On the 0-1 year gap set, the eyes outperform the mouth component for females, while for males the eyes are either on

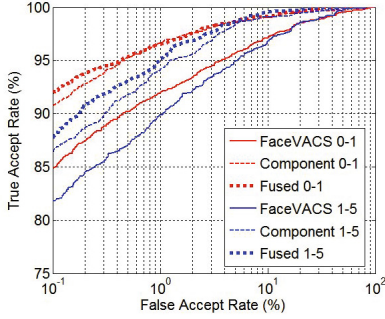


Fig. 6. MORPH results on 0-1 and 1-5 year age gap data sets

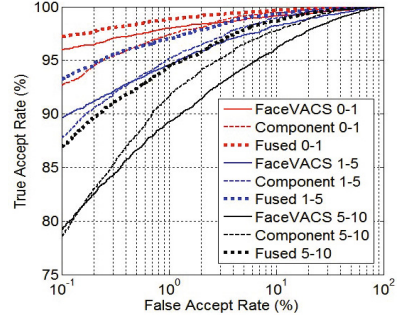


Fig. 7. PCSO results on 0-1, 1-5, and 5-10 year age gap data sets

par or better than the mouth. On the 1-5 year gap set there are again differences in the relative accuracy of components across gender lines; for females, the eyes are more accurate than the eyebrows, for males the opposite. On the 5-10 year gap set, for black and Hispanic males the eyebrows still perform better than the eyes, while for white males the eyes perform just 1% better than the eyebrows. In contrast, for females the eyes are 8-10% more accurate than the eyebrows. Across all time lapses studied, the eyes perform better for females relative to the mouth and eyebrows than they do for males.

4.3 Overall Performance

We evaluated FaceVACS and our component based system on the MORPH and PCSO data sets, and also performed a simple sum of scores fusion (with min-max normalization) on our method and FaceVACS. The ROC curves of both face recognition systems, and their fusion are shown in Figures 6 and 7.

Our component based system performs significantly better than FaceVACS on both MORPH subsets (0-1 and 1-5 year age gaps); in fact the component based system shows better performance on the 1-5 year age gap set than FaceVACS does on the 0-1 year set. On the 0-1 year age gap set the fusion of our method and FaceVACS showed minor improvement over the component based method alone on lower FAR operating points, while on the 1-5 year age gap set the fused system was consistently better than the component based method.

Table 2. Per demographic TAR (%) at 1% FAR, PCSO 0-1 year age gap set

PCSO 0-1 year	Eyes	Nose	Mouth	Eyebrows	Fused Components	FaceVACS
White-Male	74.30	89.20	74.80	82.10	96.60	98.60
White-Female	71.10	85.40	61.10	73.70	93.80	97.60
Black-Male	67.80	85.30	66.30	78.80	93.70	95.20
Black-Female	67.40	85.70	59.70	74.30	93.10	95.90
Hispanic-Male	79.60	94.80	85.90	87.30	97.60	98.70

Table 3. Per demographic TAR (%) at 1% FAR, PCSO 1-5 year age gap set

PCSO 1-5 year	Eyes	Nose	Mouth	Eyebrows	Fused Components	FaceVACS
White-Male	73.10	87.80	61.70	75.90	94.90	96.40
White-Female	64.50	79.40	53.10	57.40	89.00	93.20
Black-Male	66.00	81.20	52.90	69.50	89.70	90.30
Black-Female	62.60	80.30	55.40	56.80	86.70	87.30
Hispanic-Male	73.10	90.70	69.10	77.80	95.90	97.00

Table 4. Per demographic TAR (%) at 1% FAR, PCSO 5-10 year age gap set

PCSO 5-10 year	Eyes	Nose	Mouth	Eyebrows	Fused Components	FaceVACS
White-Male	61.20	77.20	46.80	60.00	90.50	91.90
White-Female	52.50	64.50	36.30	41.70	79.30	85.10
Black-Male	57.00	75.30	44.00	60.90	86.90	85.10
Black-Female	51.80	72.60	44.00	43.70	82.00	80.60
Hispanic-Male	62.01	86.02	56.63	63.44	93.55	92.83

Figure 8 shows some example false-reject pairs from the MORPH 0-1 year set. The left 3 pairs were falsely rejected by FaceVACS, but correctly classified by the component method. The right 2 pairs were incorrectly rejected by the component method, but accepted by FaceVACS at the 1% FAR operating point. The FaceVACS errors show its sensitivity to pose and hair variations, while the errors of component based method show its sensitivity to changes in expression (this may be the case since most facial regions we focus on are easily affected by expression changes).

The results on the PCSO subsets are mixed. FaceVACS outperforms the component based system on the 0-1 year age gap set; however, as the age gap increases the component based system performs relatively better. The component based system outperforms FaceVACS at the 1% FAR operating point on both the 1-5 and 5-10 year age gap sets, although FaceVACS has greater accuracy at lower FAR operating points. The fusion of FaceVACS and our method improved the performance on all age gap sets, with notable improvement on the larger age gaps. These results indicate that the component based method can provide complementary information to FaceVACS.

The performance of our system on complete data sets is better than its performance on most per demographic subsets because the complete data sets include cross-group matches, which are easier to reject than impostor within group matches on average. For example, on the 0-1 year age gap PCSO data set our system had 97.38% TAR at 1% FAR (better than all individual demographic groups except Hispanic males). On this data set, the mean within group impostor match score was 0.2416, while the mean cross group impostor score was 0.1273 on a score range of $[0, 1]$.

Both our method and FaceVACS perform better on PCSO than MORPH. We attribute this to the generally lower image quality in the MORPH Album2 data set. The PCSO images have higher resolution (486×624) than MORPH (400



Fig. 8. Example false reject matches from MORPH 0-1 year age gap. (a) Pairs falsely rejected by FaceVACS but accepted by the component method at 1% FAR. (b) Pairs falsely rejected by the component method but accepted by FaceVACS at 1% FAR

$\times 480$ or 200×240). Additionally, as shown in Figure 2 some MORPH images are slightly blurred, while in others the face is washed out by strong lighting. Our method's significantly better performance on MORPH than FaceVACS can be attributed to a model trained specifically on that data set compensating for the lower quality images.

5 Conclusions and Future Work

We have investigated the influence of aging on different facial components and demographic categories using a component based face representation and matching algorithm. Our experiments on the MORPH and PCSO databases show that the nose is the most stable component across face aging, and that aging has more influence on females than males. Comparisons with a leading commercial matcher (FaceVACS) show that the proposed approach is more robust to face recognition across large time lapses, while still achieving at least comparable performance to FaceVACS even across less than 1 year time lapses. Experiments also show that a score level fusion between our method and FaceVACS can improve the overall accuracy. Our future work will address how to improve face recognition accuracy by automatically estimating demographic information from face images.

References

1. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition, 2nd edn. Springer (2011)
2. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. PAMI 25, 1063–1074 (2003)
3. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Trans. IP 19, 1635–1650 (2010)
4. Han, H., Shan, S., Qing, L., Chen, X., Gao, W.: Lighting Aware Preprocessing for Face Recognition across Varying Illumination. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 308–321. Springer, Heidelberg (2010)

5. Zhang, X., Gao, Y.: Face recognition across pose: A review. *Pattern Recognition* 42, 2876–2896 (2009)
6. Ramanathan, N., Chellappa, R., Biswas, S.: Computational methods for modeling facial aging: A survey. *Journal of Visual Languages and Computing* 20, 131–144 (2009)
7. Geng, X., Zhou, Z., Smith-Miles, K.: Automatic age estimation based on facial aging patterns. *IEEE Trans. PAMI* 29, 2234–2240 (2007)
8. Park, U., Tong, Y., Jain, A.K.: Age-invariant face recognition. *IEEE Trans. PAMI* 32, 947–954 (2010)
9. Suo, J., Zhu, S.C., Shan, S., Chen, X.: A compositional and dynamic model for face aging. *IEEE Trans. PAMI* 32, 385–401 (2010)
10. Ramanathan, N., Chellappa, R.: Face verification across age progression. *IEEE Trans. IP* 15, 3349–3361 (2006)
11. Mahalingam, G., Kambhamettu, C.: Age invariant face recognition using graph matching. In: *Proc. BTAS*, pp. 1–7 (2010)
12. Ling, H., Soatto, S., Ramanathan, N., Jacobs, D.: Face verification across age progression using discriminative methods. *IEEE Trans. IFS* 5, 82–91 (2010)
13. Li, Z., Park, U., Jain, A.K.: A discriminative model for age invariant face recognition. *IEEE Trans. IFS* 6, 1028–1037 (2011)
14. AgingSkinNet: Causes of aging skin (2010), <http://www.skincarephysicians.com/agingskin-net/basicfacts.html>
15. Face and Gesture Recognition Research Network: FG-NET AGING DATABASE (2010), <http://www.fgnet.rsunit.com>
16. Klare, B., Jain, A.K.: Face recognition across time lapse: On learning feature subspaces. In: *Proc. IJCB*, pp. 1–8 (2011)
17. Ricanek, K.J., Tesafaye, T.: Morph: A longitudinal image database of normal adult age-progression. In: *Proc. AFGR*, pp. 341–345 (2006)
18. Wang, X., Tang, X.: Random sampling lda for face recognition. In: *IEEE CVPR*, vol. 2, pp. 259–265 (2004)
19. Klare, B., Paulino, A.A., Jain, A.K.: Analysis of facial features in identical twins. In: *Proc. IJCB*, pp. 1–8 (2011)
20. Milborrow, S., Nicolls, F.: Locating Facial Features with an Extended Active Shape Model. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part IV*. LNCS, vol. 5305, pp. 504–513. Springer, Heidelberg (2008), <http://www.milbo.users.sonic.net/stasm>
21. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. *CVIU* 61, 38–59 (1995)
22. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI* 24, 971–987 (2002)
23. Lowe, D.G.: Object recognition from local scale-invariant features. In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 1150–1157 (1999)
24. Klare, B., Burge, M., Klontz, J., Bruegge, R.W.V., Jain, A.K.: Face recognition performance: Role of demographic information (2012), http://www.mitre.org/work/tech_papers/2012/11_4962/
25. Grother, P.J., Quinn, G.W., Phillips, P.J.: MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms. NIST Report (2010)