Jointly De-biasing Face Recognition and Demographic Attribute Estimation (Supplementary Material)

Sixue Gong Xiaoming Liu Anil K. Jain
{gongsixu, liuxm, jain}@msu.edu

Michigan State University

In this supplementary material we include: (1) Section 1: the statistics of datasets used in the experiments, (2) Section 2: implementation details and performance of the three demographic models trained to label MS-Celeb-1M, (3) Section 3: distributions of the scores of the imposter pairs across homogeneous versus heterogeneous, (4) Section 4: performance comparisons of cross-age face recognition.

1 Datasets

Table 1 reports the statistics of training and testing datasets involved in the experiments, including the total number of face images, the total number of subjects (identities), and whether the dataset contains the annotation of gender, age, race, or identity (ID).

2 Demographic Estimation

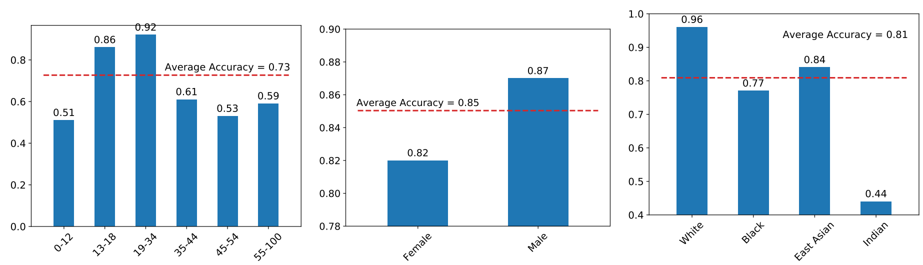
We train three demographic estimation models to annotate age, gender, and race information of the face images in MS-Celeb-1M for training DebFace. For all three models, we randomly sample equal number of images from each class and set the batch size to 300. The training process finishes at $35K^{th}$ iteration. All hyper-parameters are chosen by testing on a separate validation set. Below gives the details of model learning and estimation performance of each demographic.

Gender: We combine IMDB, UTKFace, AgeDB, AFAD, and AAF datasets for learning the gender estimation model. Similar to age, 90% of the images in the combined datasets are used for training, and the remaining 10% are used for validation. Table 2 reports the total number of female and male face images in the training and testing set. More images belong to male faces in both training and testing set. Figure 1b shows the gender estimation performance on the validation set. The performance on male images is slightly better than that on female images.

Table 1: Statistics of training and testing datasets used in the paper.

Dataset	# of Images	# of Subjects	Contains the label of			
			Gender	Age	Race	ID
CACD [2]	163,446	2,000	No	Yes	No	Yes
IMDB [12]	460,723	20,284	Yes	Yes	No	Yes
UTKFace [15]	24,106	-	Yes	Yes	Yes	No
AgeDB [10]	16,488	567	Yes	Yes	No	Yes
AFAD [11]	165,515	-	Yes	Yes	Yes[a]	No
AAF [3]	13,322	13,322	Yes	Yes	No	Yes
FG-NET [1]	1,002	82	No	Yes	No	Yes
RFW [14]	665,807	-	No	No	Yes	Partial
IMFDB-CVIT [13]	34,512	100	Yes	Age Groups	Yes*	Yes
Asian-DeepGlint [1]	2,830,146	93,979	No	No	Yes[a]	Yes
MS-Celeb-1M [5]	5,822,653	85,742	No	No	No	Yes
PCSO [4]	1,447,607	5,749	Yes	Yes	Yes	Yes
LFW [7]	13,233	5,749	No	No	No	Yes
IJB-A [8]	25,813	500	Yes	Yes	Skin Tone	Yes
IJB-C [9]	31,334	3,531	Yes	Yes	Skin Tone	Yes

[a] East Asian
* Indian



(a) Age (b) Gender (c) Race

Fig. 1: Demographic Attribute Classification Accuracy on each group. The red dashed line refers to the average accuracy on all images in the testing set.

Table 2: Gender distribution of the datasets for gender estimation.

Dataset	# of Images	
	Male	Female
Training	321,590	229,000
Testing	15,715	10,835

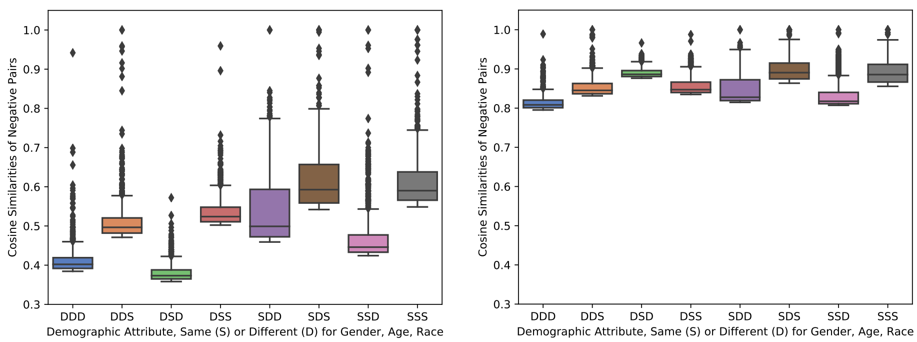
Table 3: Race distribution of the datasets for race estimation.

Dataset	# of Images			
	White	Black	East Asian	Indian
Training	468,139	150,585	162,075	78,260
Testing	9,469	4,115	3,336	3,748

Race: We combine AFAD, RFW, IMFDB-CVIT, and PCSO datasets for training the race estimation model. UTKFace is used as validation set. Table 3 reports the total number of images in each race category of the training and

Table 4: Age distribution of the datasets for age estimation

Dataset	# of Images in the Age Group					
	0-12	13-18	19-34	35-44	45-54	55-100
Training	9,539	29,135	353,901	171,328	93,506	59,599
Testing	1,085	2,681	13,848	8,414	5,479	4,690



(a) BaseFace (b) DebFace

Fig. 2: BaseFace and DebFace distributions of the similarity scores of the imposter pairs across homogeneous versus heterogeneous gender, age, and race categories.

testing set. Similar to age and gender, the performance of race estimation is highly correlated to the race distribution in the training set. Most of the images are within the White group, while the Indian group has the least number of images. Therefore, the performance on White faces is much higher than that on Indian faces.

Age: We combine CACD, IMDB, UTKFace, AgeDB, AFAD, and AAF datasets for learning the age estimation model. 90% of the images in the combined datasets are used for training, and the remaining 10% are used for validation. Table 4 reports the total number of images in each age group of the training and testing set, respectively. Figure 1a shows the age estimation performance on the validation set. The majority of the images come from the age 19 to 34 group. Therefore, the age estimation performs the best on this group. The performance on the young children and middle to old age group is significantly worse than the majority group.

It is clear that all the demographic models present biased performance with respect to different cohorts. These demographic models are used to label the MS-Celeb-1M for training DebFace. Thus, in addition to the bias from the dataset itself, we also add label bias to it. Since DebFace employs supervised feature disentanglement, we only strive to reduce the data bias instead of the label bias.

Table 5: Evaluation Results (%) of Cross-Age Face Recognition

Method	Datasets	
	FG-NET	CACD-VS
BaseFace	90.55	98.48
DebFace	93.3	99.45

3 Distributions of Scores

We follow the work of [6] that investigates the effect of demographic homogeneity and heterogeneity on face recognition. We first randomly select images from CACD, AgeDB, CVIT, and Asian-DeepGlint datasets, and extract the corresponding feature vectors by using the models of BaseFace and DebFace, respectively. Given their demographic attributes, we put those images into separate groups depending on whether their gender, age, and race are the same or different. For each group, a fixed false alarm rate (the percentage of the face pairs from the same subjects being falsely verified as from different subjects) is set to 1%. Among the falsely verified pairs, we plot the top 10^{th} percentile scores of the negative face pairs (a pair of face images that are from different subjects) given their demographic attributes. As shown in Fig. 2a and Fig. 2b, we observe that the similarities of DebFace are higher than those of BaseFace. One of the possible reasons is that the demographic information is disentangled from the identity features of DebFace, increasing the overall pair-wise similarities between faces of different identities. In terms of de-biasing, DebFace also reflects smaller differences of the score distribution with respect to the homogeneity and heterogeneity of demographics.

4 Cross-age Face Recognition

We also conduct experiments on two cross-age face recognition datasets, i.e., FG-NET [2] and CACD-VS [2], to evaluate the age-invariant identity features learned by DebFace. The CACD-VS consists of 4,000 genuine pairs and 4,000 imposter pairs for cross-age face verification. On FG-NET, the evaluation protocol is the leave-one-out cross-age face identification. Table 5 reports the performance of BaseFace and DebFace on these two datasets. Compared to BaseFace, the proposed DebFace improves both the verification accuracy on CACD-VS and the rank-1 identification accuracy on FG-NET.

[2] `https://yanweifu.github.io/FG_NET_data`

References

1. http://trillionpairs.deepglint.com/overview
2. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: ECCV (2014)
3. Cheng, J., Li, Y., Wang, J., Yu, L., Wang, S.: Exploiting effective facial patches for robust gender recognition. Tsinghua Science and Technology **24**(3), 333–345 (2019)
4. Deb, D., Best-Rowden, L., Jain, A.K.: Face recognition performance under aging. In: CVPRW (2017)
5. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. Springer (2016)
6. Howard, J., Sirotin, Y., Vemury, A.: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: IEEE BTAS (2019)
7. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
8. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR (2015)
9. Maze, B., Adams, J., Duncan, J.A., Kalka, N., Miller, T., Otto, C., Jain, A.K., Niggel, W.T., Anderson, J., Cheney, J., et al.: Iarpa janus benchmark-c: Face dataset and protocol. In: 2018 ICB (2018)
10. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: CVPRW (2017)
11. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: CVPR (2016)
12. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. IJCV (2018)
13. Setty, S., Husain, M., Beham, P., Gudavalli, J., Kandasamy, M., Vaddi, R., Hemadri, V., Karure, J.C., Raju, R., Rajan, Kumar, V., Jawahar, C.V.: Indian Movie Face Database: A Benchmark for Face Recognition Under Wide Variations. In: NCVPRIPG (2013)
14. Wang, M., Deng, W., Hu, J., Tao, X., Huang, Y.: Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In: ICCV (2019)
15. Zhang, Z., Song, Y., Qi, H.: Age progression/regression by conditional adversarial autoencoder. In: CVPR. IEEE (2017)