# Does Face Recognition Accuracy Get Better With Age? Deep Face Matchers Say No

Vítor Albiero, Kevin W. Bowyer
University of Notre Dame
Notre Dame, Indiana
{valbiero, kwb}@nd.edu

Kushal Vangara, Michael C. King
Florida Institute of Technology
Melbourne, Florida
kvangara2015@my.fit.edu, michaelking@fit.edu

## Abstract

*Previous studies generally agree that face recognition accuracy is higher for older persons than for younger persons. But most previous studies were before the wave of deep learning matchers, and most considered accuracy only in terms of the verification rate for genuine pairs. This paper investigates accuracy for age groups 16-29, 30-49 and 50-70, using three modern deep CNN matchers, and considers differences in the impostor and genuine distributions as well as verification rates and ROC curves. We find that accuracy is lower for older persons and higher for younger persons. In contrast, a pre deep learning matcher on the same dataset shows the traditional result of higher accuracy for older persons, although its overall accuracy is much lower than that of the deep learning matchers. Comparing the impostor and genuine distributions, we conclude that impostor scores have a larger effect than genuine scores in causing lower accuracy for the older age group. We also investigate the effects of training data across the age groups. Our results show that fine-tuning the deep CNN models on additional images of older persons actually lowers accuracy for the older age group. Also, we fine-tune and train from scratch two models using age-balanced training datasets, and these results also show lower accuracy for older age group. These results argue that the lower accuracy for the older age group is not due to imbalance in the original training data.*

## 1. Introduction

Differences in face recognition accuracy across demographic groups have attracted a lot of attention in recent years. In this paper, our goal is to develop a better understanding of how face recognition accuracy varies across young, middle and older age groups.

Our experiments use a large, publicly available dataset [1], that was originally assembled for studying face aging [24]. Importantly, *this is the largest generally available dataset that has meta-data for subject age with each image*. Web-scraped datasets typically do not have such meta-data.

Contributions of this work include: (1) finding a consistent qualitative effect across race, (2) investigation of how genuine and impostor distributions contribute to accuracy across age groups, (3) investigation of the effect of elapsed time between genuine pairs, (4) insights on how the effect of age difference between persons in an impostor pair depends on the age groups, (5) investigation of the effects of balancing the training data across age groups, and (6) use of publicly available matchers and dataset, supporting reproducible experiments.

## 2. Related Work

The 2002 Face Recognition Vendor Test [22] looked at how identification rate varies with age for three commercial matchers, and concluded that older persons are recognized more accurately than younger. Beveridge et al. [5], using the Face Recognition Grand Challenge dataset [23], found that verification rate increases as subject age increases. In a 2009 meta-analysis of previous works, Lui et al. [19] found that 20 out of 22 age-related results agree that the accuracy is higher for older persons. In a 2018 review, Abdurrahim et al. [2] found general agreement in the literature that older persons are recognized more accurately.

Klare et al. [15] evaluated six matchers on the age groups 18-30, 30-50 and 50-70, using the Pinellas County Sheriff's Office (PCSO) dataset. The three commercial matchers showed lower verification rates for 18 to 30 age group, indicating that younger people are harder to verify. Also, despite not agreeing on which age group has the highest accuracy, the two non-trainable matchers (Local Binary Patterns and Gabor filters) show an overall lower accuracy for the young age group. Best-Rowden and Jain [3] reported results for two commercial matchers on a subset of the PCSO dataset, separated to study the effects of elapsed time between genuine pairs. They conclude that the genuine scores significantly decrease over time for both matchers. An extended version of their work [4] added an additional dataset and more matchers.

In the FRVT Ongoing effort at NIST, Grother [9] reported higher false match rates (FMRs) for older persons than for younger. With a fixed threshold to give an overall FMR of 1-in-10,000, a group of age 20-somethings had a FMR of 1-in-100, but a group with 70-somethings had a FMR of 5-in-100. On the other hand, they reported higher false non-match rates (FNMRs) for younger subjects, 0.05 for 20-somethings, and 0.02 for 70-somethings.

Lu et al. [18] looked at results from four deep learning matchers for the IARPA JANUS Benchmark B (IJB-B) dataset, which is a collection of web-scraped images. The authors found that accuracy increases with age up until age 50, which agrees with previous results in the literature. They also found that after age 50, accuracy starts to drop. However, the age annotations for IJB-B images were derived using Amazon Mechanical Turk, and we are not aware of how accurate they may be.

Cook et al. [7] analyzed demographic factors in the 2018 Biometric Technology Rally [13]. They used images captured from eleven different image acquisition systems, and matched them against same-day enrollment images and against historical images using a commercial matcher. Their age experiment was conducted using two age groups: 20 to 40; and 40 to 85. While the same-day matching shows a similar result for both age ranges, they report higher accuracy for older persons in the historical matching.

Cao et al. [6] analyzed results of matching young (less than 34 years) and mature (34 years or more) persons, using a deep CNN matcher. They report that matching young-to-mature faces is harder than young-to-young or mature-to-mature. They also report that mature-to-mature matching results in higher similarity scores than young-to-young matching.

Previous works generally focus on how match scores for genuine pairs vary across age groups [5, 7, 6], and do not explicitly consider both the impostor and genuine distributions. Some previous works use datasets that are not available [15, 9, 7]. Further, most previous work does not experiment with race or gender subsets, which is important as an imbalance in other demographics across age groups could be a confounding factor for accuracy variations.

We present the most extensive analysis to date of how face recognition accuracy varies across age groups. In addition to ROC curves, we analyze genuine/impostor score distributions, and how they contribute to accuracy differences. We analyze how elapsed time between genuine pairs and age difference between impostor pairs affects accuracy. We also investigate how training data affects accuracy, with models fine-tuned across age groups, age balanced datasets, and trained from scratch on age balanced datasets.

| | **Whole Dataset** | | **AA Male** | | **C Male** | |
|---|---|---|---|---|---|---|
| **Group** | **Subjects** | **Images** | **Subjects** | **Images** | **Subjects** | **Images** |
| **Young** | 5,778 | 21,665 | 4,085 | 16,799 | 833 | 2,690 |
| **Middle** | 6,532 | 25,604 | 4,235 | 16,891 | 1,140 | 4,140 |
| **Old** | 1,074 | 3,622 | 726 | 2,387 | 221 | 837 |

Table 1: Division of age groups for whole dataset, African American (AA) males and Caucasian (C) males.

## 3. Dataset

We identified the largest dataset that is available to the research community and that has meta-data for subject age at the time each image was acquired [24]. Many well-known datasets do not have the age of the subject recorded with each image. This is the case for the IJB datasets [27, 21], MegaFace [14], and more generally, for any web-scraped dataset. Some datasets used in previous studies are not available to others. FG-NET [17] has images with annotated ages, however the number of images with a subject older than 50 years is only 23. In contrast, the MORPH [24] dataset used in this work has recorded age meta-data and a relatively large number of subjects and images across the age groups, as it was collected to study face aging.

The MORPH dataset [24] was assembled from public records and has been widely used in face aging research. We curated a subset of MORPH Album 2, which, after removal of a small number of images that do not contain a face or that are repeated images, has 53,231 images of 13,119 subjects. For each image, the meta-data includes a date of birth (DOB) and date of acquisition. This makes it possible to split the dataset into images of subjects for different age ranges. The age ranges used are 16 to 29, 30 to 49, and 50 to 70, termed “young”, “middle” and “old”, respectively.

In initial studies, we found that 1,708 of the 13,119 subjects had inconsistent DOBs across their images. As extreme cases, one subject had six different DOBs across their images, and another had a 32-year difference across the DOBs for their images. However, in the large majority of cases, the inconsistent DOBs could be corrected in a straightforward manner. If the DOB was consistent for 75% or more of a subject’s images, then the inconsistent DOBs for the subject’s other images were corrected to the consistent DOB. This resolved DOB inconsistency for 1,364 subjects. The DOB for the remaining 344 was deemed too inconsistent to correct, and these subjects were dropped. Thus the total number of subjects in the dataset for our analysis is 12,775. The image names and meta-data corrections used in this paper will be made available online.

The dataset with curated DOB was divided into the three age ranges. Note that a particular subject may have images in both the “young” and “middle” range, or the “middle” and “old” range. However, if a subject has only one image for a given age range, the subject and image were dropped

(a) 22 and 25 years (b) 41 and 44 years (c) 58 and 61 years

Figure 1: Sample authentic pairs with 3 years difference in young, middle and old groups.

from that age range. The numbers of subjects and images for the division of the whole dataset into young, middle and old age ranges is given in Table 1.

Other research has shown that face recognition accuracy is different for African-American and Caucasian, and for male and female [15, 5, 7, 16]. The young, middle and old ranges of our dataset may have varying demographic mix. To create subsets with consistent demographic mix across age ranges, we split the dataset into four subsets: African-American male, African-American female, Caucasian male and Caucasian female. However, the female old age range had just 93 subjects and 286 images for African-American, and 34 subjects and 112 images for Caucasian. This is too few subjects and images for reliable analysis, and so the subgroup analysis was done only for the two male subsets. Sample authentic pairs are shown in Figure 1. Results of our analysis are presented for the whole dataset, with its varying demographics across age ranges, and for the African-American male and the Caucasian male subsets.

## 4. Deep Learning Face Matchers

We use three recent deep CNN face matchers. These matchers are chosen to represent training with different loss functions and different training sets: (1) VGGFace2 (ResNet-50) [12], trained on VGGFace2 dataset [6] with standard softmax loss; (2) FaceNet [25], trained on MS-Celeb-1M dataset [11] with triplet loss; and (3) ArcFace [8] (ResNet-100), trained on MS-Celeb-1M V2 dataset with additive angular margin loss. Each network was used with pre-trained weights that are publicly available [20, 26, 10].

As input to the matchers, the faces were detected and aligned using MTCNN [28], and resized to 224x224 pixels (VGGFace2), 160x160 (FaceNet), or 112x112 (ArcFace). For feature extraction, the last but one layer was used, which corresponds to a 2048-d feature vector for VGGFace2, 128-d for FaceNet, and 512-d for ArcFace. After extraction, the features were matched using cosine similarity.

## 5. Experimental Results

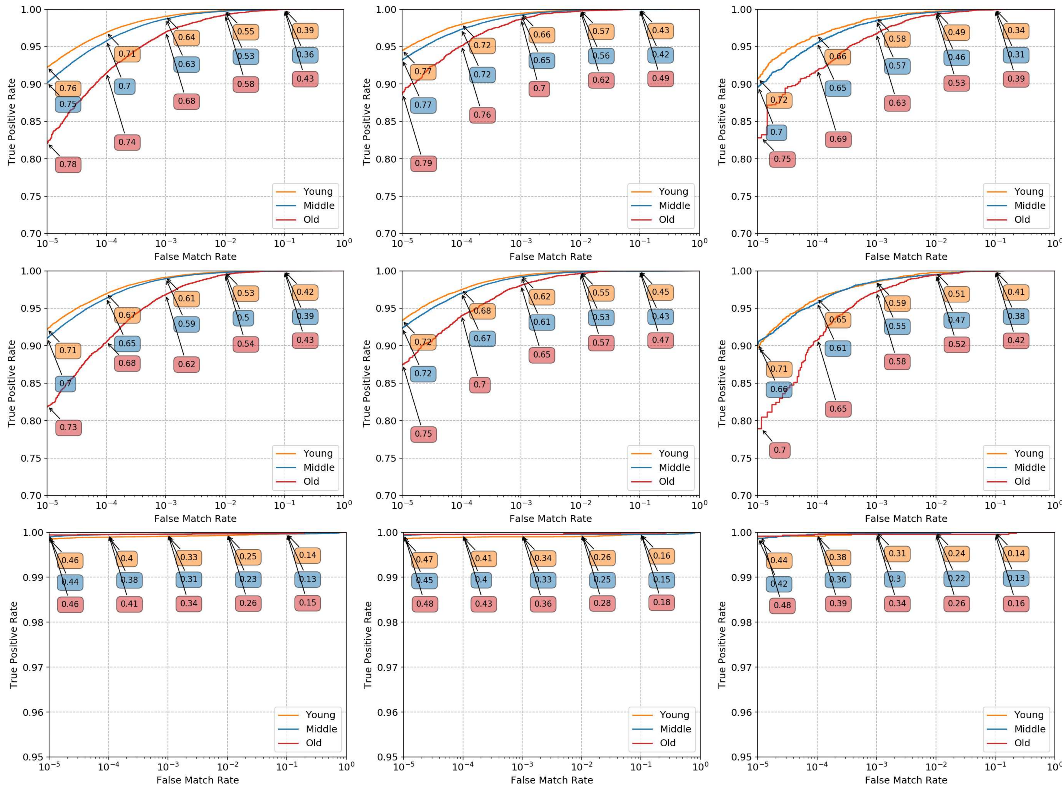
This section first presents ROC results for the matchers, for the whole dataset and for the African-American male and Caucasian male cohorts. The relative accuracy shown by the ROC curves across the age ranges is different from what most past studies have found. To better understand the effects of the two types of errors that are summarized in the ROC curve, we next present the impostor and genuine distributions. We further investigate how elapsed time between impostor and genuine pairs affects their matching scores across the age groups. Moreover, to understand why our results differ from those of some previous studies, we present results from a pre deep CNN face matcher as used in a well-known previous work [15]. Finally, we analyze how the training data affects the performance of the deep models, by fine-tuning models on separate age groups, age balanced subsets and training from scratch on age balanced subsets.

### 5.1. ROC Curve Comparison

ROC curves for the different age ranges are compared in Figure 2. The general pattern of the ROC results is consistent, whether considering the whole dataset with its varying mixture of gender and race across age ranges, or considering either of the two same-race-same-gender subsets. It is important to note that ROC curves are more appropriately used to compare accuracy of different algorithms on the same dataset. In this case, we are comparing the accuracy of the same algorithm for different datasets (different age ranges). The ROC format hides the fact that the same FMR for different age ranges is obtained at different decision thresholds. For this reason, the decision threshold value is marked on the plots for sample values of FMR.

We see that except for VGGFace2 on Caucasian males, for all ROC curves, the old age range has higher thresholds to achieve the same FMR. Following the old age range, the young age range has the second highest thresholds. The thresholds give an idea of how much the overlap of the impostor and genuine distributions is shifted towards higher similarity scores, which can be correlated to a worse impostor distribution.

For six out of nine ROC curves, the young age range has the best ROC, and the old age range has the worst. For FaceNet and VGGFace2, the gap between middle and old is noticeably larger than the gap between young and middle. With the ArcFace matcher, the older group has a slightly better ROC curve than the other two age groups, but achieved using a higher threshold. In the whole dataset, the ArcFace true positive rate (TPR) with a false match rate (FMR) of $10^{-4}$ is 99.89%, 99.95% and 99.95% for the young, middle and old age groups, respectively. Overall,




(a) Whole dataset (b) African American males (c) Caucasian males

Figure 2: ROC curves for FaceNet (top), VGGFace2 (middle), and ArcFace (bottom). Annotated values correspond to thresholds used for the correspondent FMR. ArcFace is displayed at a different scale for better visualization.

the pattern of ROC results disagrees with previous studies that found that older persons are easier to recognize than younger persons. For this reason, the next section presents the impostor and genuine distributions that underlie the ROC curves.

## 5.2. Impostor and Genuine Distributions

The impostor and genuine distributions that underlie the ROC curves in Figure 2 are shown in Figure 3. We see that in each of the nine plots, the impostor distribution for the old group is the worst, with a noticeable shift toward the genuine distributions. The impostor distribution for the middle age range is the best, with a slight shift toward lower similarity scores. And the impostor distribution for the young age range is between the other two.

The young age range generally has the best genuine distribution, showing a higher peak of high-similarity scores than the other two age ranges. There is not a noticeable difference between the genuine distributions for the middle and old age ranges.

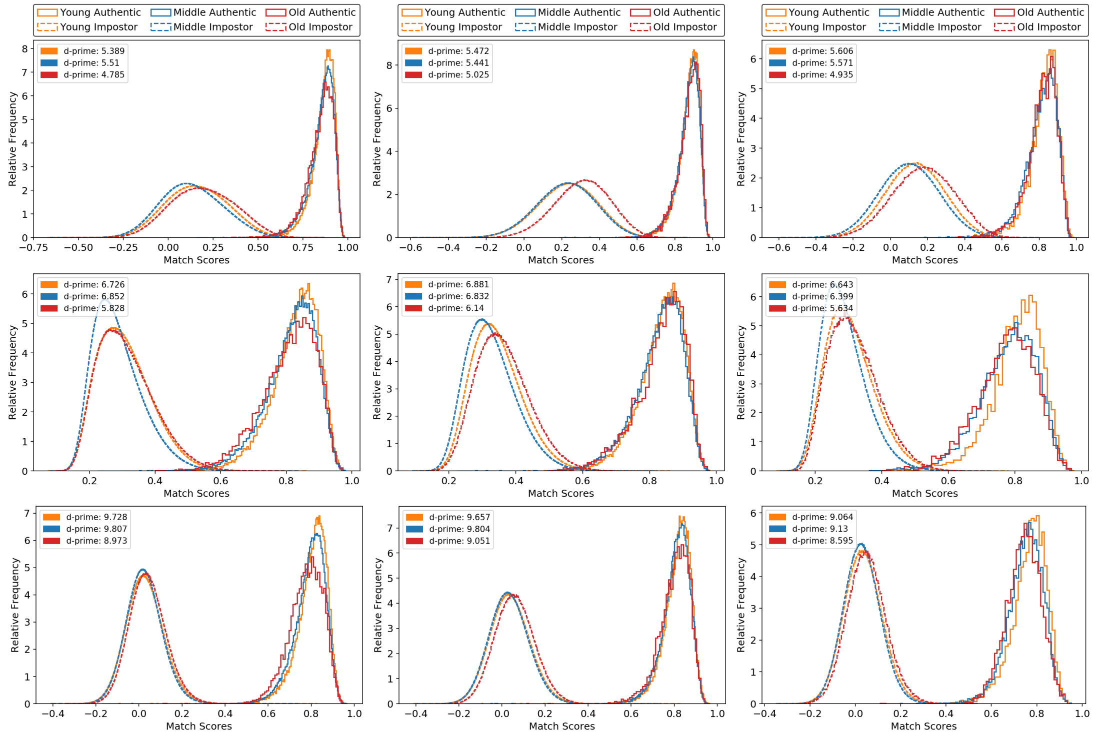
From analyzing the impostor and genuine distributions, we can infer that the main factor driving the poorer ROC curves for the old age range is its poorer impostor distribution. We can also infer that the main factor driving the generally better ROC curves for the young age range is its (slightly) better genuine distribution.

The impostor and genuine distributions are not perfectly Gaussian. However, the d-prime statistic may still indicate the relative separation of the two distributions. As shown in Figure 3, in general, the d-prime values for younger and middle age ranges are more similar, and the d-prime for the old age range is noticeably worse.

In summary, the impostor distribution is the main cause for the old age range having a worse ROC curve than the other age ranges. In contrast, the young age range's better ROC curves are explained mostly by its genuine distribution; its impostor distribution is worse than the middle age range.

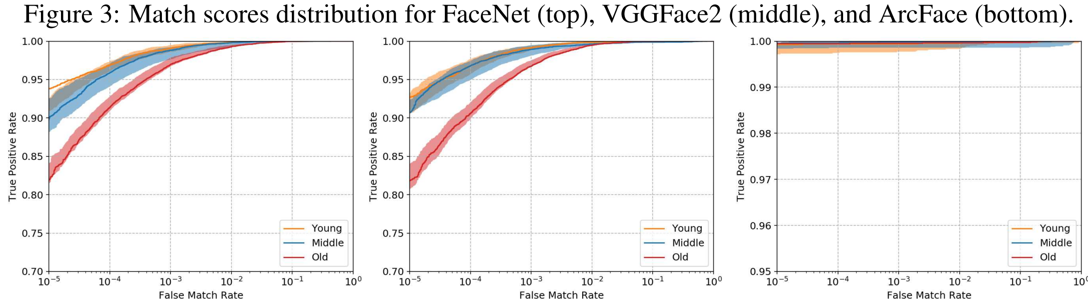
## 5.3. Bootstrap Confidence Analysis

To check whether our results might be an accident of the particular set of subjects and images, we randomly selected 80% (860) of the subjects in the old group 100 times. And to check whether our results might be partly due to the different number of subjects and images in the age ranges, we randomly selected 860 subjects in the young and middle age groups 100 times, so that they have the same number of subjects as the old age group. We then computed 100 ROC




(a) Whole dataset (b) African American males (c) Caucasian males

Figure 3: Match scores distribution for FaceNet (top), VGGFace2 (middle), and ArcFace (bottom).




(a) FaceNet (b) VGGFace2 (c) ArcFace

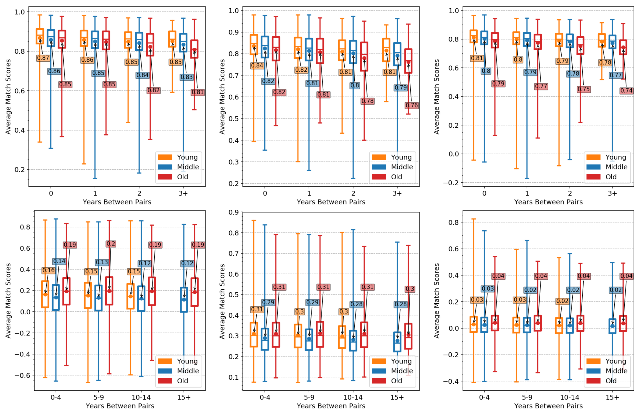
Figure 4: ROC curves with 90% confidence interval using 860 subjects randomly selected 100 times. ArcFace is displayed at a different scale for better visualization.

curves for each age range and ordered them by the area under the curve (AUC) between a FMR of $10^{-5}$ and $10^{-3}$. Figure 4 shows the median ROC curve with a 90% confidence interval using the ROC curves with 5-th and 95-th AUCs.

It is clear that generally the old group has a much worse ROC than the young and middle groups, even though all three age ranges have the same number of subjects. Also, the ROC curves from Figure 2 lay within the confidence intervals, indicating that the previous results are not the result of different subject/image quantity, or of an unusual distribution of subjects. Finally, the middle age group shows a wider confidence interval, which overlaps with the young age confidence interval.

## 5.4. Influence of Elapsed Time

Figure 5 shows the average match score for impostor pairs and for genuine pairs, as a function of age difference between the person(s) in the two images. For genuine pairs, this is age difference (time lapse) between images of the same person. For impostor pairs, this is age difference between different persons. The concept is likely most familiar for genuine matches. The average match score between two images of the same person generally decreases with increase elapsed time. A similar result holds for impostor matches. Images of two different persons are on average more similar if the persons are the same age, and increased difference in ages generally decreases the average similarity. The results are presented only for the whole dataset, as




(a) FaceNet (b) VGGFace2 (c) ArcFace

Figure 5: Match scores with increasing elapsed time between authentic (top) and impostors (bottom) pairs for whole dataset.

it has more images to break into separate bins of elapsed time.

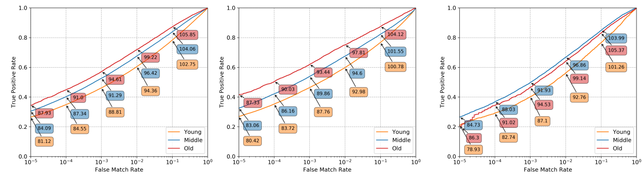
For the genuine score plots, we see that the younger age range generally has the highest similarity scores and the older age range generally has the lowest similarity scores, for all time lapse bins and for all three matchers. In effect, that data says that, on average, for the same length of time lapse, two images of the same older person look less similar than two images of the same younger person. Also, we see a general trend of decreasing similarity score with increased elapsed time, which was expected.

For the impostor score plots, the older age range generally has the worst (highest) average similarity scores and the middle age range generally has the best (lowest) average similarity scores. For the young and middle age ranges, there is a small decrease in scores as a function of age increase between impostor pairs. In the other hand, the old age range shows steady scores as the age difference between impostor pairs increase.

Overall, we see no evidence that a difference in the distribution of time lapse between images across the different age ranges is a driving factor in the observed accuracy differences between the age ranges.

### 5.5. Comparison with Pre Deep CNN Algorithm

The best-known previous works comparing accuracy across different age ranges date to before the wave of deep CNN matchers; e.g. [15]. Also, previous works use different datasets from the one that we use. Therefore, our results being different from previous results could potentially be due to (a) newer deep learning matchers having different properties than older matchers, and / or (b) our dataset being different somehow from those used in previous work. The PCSO dataset analyzed in [15] is no longer available. However, we re-implemented the Local Binary




(a) Whole dataset (b) AA males (c) C males

Figure 6: ROC curves for LBP matcher for whole dataset, African American (AA) males, and Caucasian (C) males. Annotated values correspond to thresholds used for the correspondent FMR.

Patterns (LBP) matcher as described in [15] and used it to analyze our dataset.

Figure 6 shows ROC curves for the LBP matcher. We can see that the trend across age ranges for the LBP matcher is the opposite of the trend for the deep CNN matchers. For the LBP matcher on our dataset, the young age range consistently has the worst performance, and the old age range has the best performance in two of the subsets. Thus the results of the LBP matcher on our dataset agree qualitatively with the results of the LBP matcher on the PCSO dataset as reported in [15]. This indicates that our results differ from those of past studies due to modern deep CNN matchers operating differently from older matchers and not because of some difference in the datasets.

### 5.6. Training Data Analysis

Using an CNN age predictor [10], we predicted the subject age in the images in the three datasets that were used in training the three CNN matchers. The age predictor mean absolute error (MAE) on the whole MORPH dataset is 8.55, and is 4.1 on the validation set. Because the age predictor has MAE around 4 on the validation test, we classified faces predicted as 34 or less years as young, 54 or less as middle age, and 55 or more as old. Table 2 clearly shows that all three datasets are imbalanced, with the middle age range having the most representation and the older age range having the least. This shows that the better accuracy observed for the young age range is not directly caused by the fraction of young faces in the training data, as the middle age range represents a larger fraction of the training data. It also demonstrates that the much better and more similar accuracy of ArcFace across age groups is not due to the training data being more balanced than for the other matchers, but mostly because of its better loss function.

To further investigate the effects of the subjects' ages in the training data, we fine-tuned the two higher-accuracy matchers using subjects in a specific age range. The age range with the smaller amount of images and subjects is the old range, which was selected as the starting point for each training group preparation. First, for each age range, we removed all subjects that have less than 50 images, which

| | Number of Images | | |
|---|---|---|---|
| **Dataset** | **Young** | **Middle** | **Old** |
| **VGGFace2** | 1.01M (32.5%) | 1.74M (55.9%) | 0.36M (11.6%) |
| **MS1M** | 2.97M (35.2%) | 4.49M (53.1%) | 0.99M (11.7%) |
| **MS1M V2** | 1.66M (28.6%) | 3.26M (56.1%) | 0.89M (15.3%) |

Table 2: Number of images and ratio of young, middle and old age groups in the training datasets.

yields 2,009 subjects with 308,640 images for VGGFace2, and 5,033 subjects with 348,742 images for MS1M V2. Then, we randomly selected the same number of subjects and images for the young and middle age groups, to make training subsets balanced on number of subjects and images.

Using each age group separately, we fine-tuned the VGGFace2 matcher using data from the MS1M V2 dataset, and the ArcFace matcher using data from the VGGFace2 dataset, so that we do not re-use images from a matcher's original training data as fine-tuning data. Table 3 (upper half) shows the TPR at a FMR of $10^{-4}$ for each fine-tuned model. We observe that the performance dropped for all the fine-tuned models compared to the original results (shown in middle), which may be expected as there are many fewer images and subjects, thus less generalization capacity. In general, no matter the age range used for fine-tuning, the old age group has the worst accuracy. Contrary to what might be expected, the best results for each age group are not correlated to fine-tuning on the same age range, e.g. the young age range had better accuracy when the matcher was fine-tuned either on a middle age group (VGGFace2) or an old age group (ArcFace). For the VGGFace2 matcher, in general, the best accuracy for the three age groups were with the fine-tuning on the middle age group, with d-prime values in the whole dataset of 5.273, 5.518 and 5.082 for young, middle, and old age groups, respectively. ArcFace achieved better accuracy for all age groups when fine-tuned on an old age group, with d-prime values in the whole dataset of 7.919, 7.908, 7.247 for young, middle, and old age groups, respectively. Moreover, in five out of six results, the difference between the best and worst group was lower when the matcher was fine-tuned using an old age group. When fine-tuning with a young age group, the difference between the best and worst age group is much higher for both matchers, which indicates that middle to old faces are better for a more uniform and higher accuracy between age groups.

It is possible that the worse accuracy for the old age range across the fine-tunings is due to some bias in the initial training. To explore that, we combined the three age subsets together, creating age balanced datasets. Then, we fine-tuned and trained from scratch the VGGFace2 and ArcFace matchers on the age balanced datasets.

Table 3 (lower half) shows results for the matchers fine-tuned and trained from scratch on the age balanced datasets. Only one out of six results shows closer performance between the worst and best group for the fine-tuning results. For all other five results, a more similar performance was achieved training from scratch. With the VGGFace2 matcher, the accuracy with the fine-tuning is much higher than when trained from scratch, but the performance is still lower for the old group, and the d-prime values for the whole dataset are 4.563, 4.661, and 4.315 for young, middle, and old age groups, respectively. On the other hand, ArcFace achieved better performance when training from scratch on the MS1M V2 dataset, with a higher accuracy for the old group compared to the young, but still with a lower d-prime value for the old group, which is 6.876, compared to 7.06 for the young group, and 7.367 for middle group. Moreover, ArcFace training and fine-tuning on VGGFace2 dataset also show worse results for the old group compared to the other two groups. Finally, the results show that even when the dataset is age balanced, in general, the old group has lower performance than young or middle.

## 6. Conclusions and Discussion

**Younger people are easier, older people are harder.** Previous works have found that people in the old age range are easier to recognize (have higher accuracy) than people in the young age range. However, we find the opposite. For two of the three modern deep CNN face matchers used in our work, the old age range has a noticeably worse ROC curve than the young age range, and the other matcher, which has much higher accuracy for all age ranges, shows similar accuracy because all age groups are basically "at ceiling". The bootstrap results also confirm that this pattern is not due to number of subjects or particular age group distributions of the dataset.

The genuine distributions for the three age ranges seem more similar than their impostor distributions. Therefore it seems that the impostor distribution for the old age range is driving its worse ROC. Also, the d-prime values show how much worse the separation between the old group distributions is compared to young and middle.

**The new deep CNN face matchers are different.** The different pattern of accuracy across age ranges in our results could, in principle, be due to a difference in testing datasets or to modern face matchers having different properties. Cao et al. [6] describe an experiment involving two age groups, "young" (less than 34 years) and "mature" (34 years or more), in which they report that "mature" is recognized more accurately than "young" with the VGGFace2 matcher. This result appears to align better with the traditional expectation for young/old accuracy, than with our result. However, this result is based just on 100 subjects. Also, it uses 2 "templates" of 5 images each, for each age range for each subject, for a total of just 2,000 images. Examining the 20 images in the templates for the first subject number of the 100 subjects, we found that there are du-

| Matcher@Trainining subset | Whole Dataset Young | Middle | Old | Diff. | African American Males Young | Middle | Old | Diff. | Caucasian Males Young | Middle | Old | Diff. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| VGGFace2@MSM1V2_Young | 92.32 | 87.3 | 81.39 | 10.93 | 94.43 | 91.05 | 87.06 | 7.37 | 94.33 | 90.96 | 84.18 | 10.15 |
| VGGFace2@MSM1V2_Middle | **93.24** | **90.44** | **85.64** | **7.6** | **95.41** | **93.2** | **90.16** | 5.25 | **94.82** | **93.44** | 85.76 | 9.06 |
| VGGFace2@MSM1V2_Old | 88.54 | 86.67 | 80.75 | 7.79 | 91.02 | 90.41 | 88.45 | **2.57** | 92.93 | 91.75 | **89.17** | **3.76** |
| ArcFace@VGGFace2_Young | 98.39 | 98.47 | 95.08 | 3.39 | 99.12 | 99.14 | 97.32 | 1.82 | 98.69 | 99.31 | 97.41 | 1.9 |
| ArcFace@VGGFace2_Middle | 98.19 | 98.33 | 95.77 | 2.56 | 98.93 | 98.93 | 97.41 | 1.52 | 98.41 | 98.93 | 97.97 | 0.96 |
| ArcFace@VGGFace2_Old | **99.37** | **99.37** | **99.27** | **0.1** | **99.47** | **99.51** | **99.35** | **0.16** | **99.55** | **99.94** | **99.81** | **0.39** |
| *VGGFace2 Baseline* | *96.94* | *96.94* | *90.51* | *6.43* | *97.49* | *97.02* | *94.04* | *3.45* | *96.3* | *95.71* | *90.81* | *5.49* |
| *ArcFace Baseline* | *99.89* | *99.95* | *99.95* | *0.06* | *99.89* | *99.95* | *99.95* | *0.06* | *99.91* | *99.95* | *99.96* | *0.05* |
| VGGFace2@MS1MV2_All_ft | **95.84** | **93.23** | **89** | 6.84 | **96.83** | **95.15** | **93.51** | 3.32 | **95.92** | **94.08** | **90.69** | 5.23 |
| VGGFace2@MS1MV2_All | 77.01 | 75.53 | 69.05 | 7.96 | 77.35 | 76.66 | 74.94 | 2.41 | 76.82 | 74.04 | 73.03 | **3.79** |
| VGGFace2@VGGFace2_All | 71.21 | 70.99 | 67.85 | **3.36** | 72.15 | 73.18 | 74.19 | **2.04** | 76.28 | 75.28 | 70.3 | 5.98 |
| ArcFace@VGGFace2_All_ft | 98.27 | 98.56 | 97.16 | 1.4 | 98.74 | 99 | 98.29 | 0.45 | 98.26 | 99.06 | 98.51 | **0.8** |
| ArcFace@VGGFace2_All | 95.82 | 96.72 | 95.02 | 1.7 | 96.55 | 97.69 | 96.44 | 1.25 | 95.66 | 97.36 | 96.82 | 1.7 |
| ArcFace@MS1MV2_All | **98.85** | **99.21** | **99.27** | **0.42** | **99.05** | **99.39** | **99.12** | **0.34** | **98.2** | **99.14** | **99.68** | 1.48 |

Table 3: True positive rates (%) with a false match rate of $10^{-4}$ and difference (best − worst) for models fine-tuned with different age group subsets (top half), and fine-tuned (ft) or trained from scratch on all age groups combined (bottom half). The baseline results are shown in the middle of the table.

plicate images, incorrect identity labels, incorrect age category labels, as well as images with extreme pose variation. (These 20 images are shown in the Supplemental Materials.) Given these issues with the images and meta-data, combined with the small number of subjects and images, the young/mature result found by Cao et al. [6] simply may not be reliable.

Lu et al. [18] report results across five age ranges (that were predicted using crowd sourcing) with deep CNN face matchers. Their results show increasing accuracy until the age of 50, then a drop in accuracy, which agrees with our results on the old age group, but disagrees on the young age group. However, as the dataset used is unconstrained, there are many factors that can be affecting one age group more than other, e.g., pose, illumination, and facial expression. So, it is possible that the lower performance of subjects in the middle 20s is not related to their ages, but to external factors.

As far as we know, MORPH [24] is the only currently available dataset with recorded age meta-data, enough subjects across ages, and consistent quality, so, we cannot reproduce the experiments on another dataset. However, we re-implemented a pre deep CNN face matcher (LBP) as described in [15] and ran it on our dataset and obtained results similar to those in [15]. This indicates that the pattern of results in our work is due to newer deep CNN matchers, rather than a property of the dataset.

Deep CNN face matching technology seems to have changed the default expectation for differences in accuracy across age ranges. The datasets that current deep CNN matchers have been trained on do not have a large fraction of older subjects. For instance, all the datasets used to train the matchers tested in this work, have less than 16% of the images with a subject older than 50 years. It would be possible that models trained on a dataset with a more balanced age distribution would have a similar performance across age groups. However, the highest-accuracy matcher, ArcFace, achieves more similar results across age groups, and has a very imbalanced training dataset. To further investigate the training data effect, we fine-tuned two matchers on different age groups, and the results are generally worse for the old age group, no matter the age subset fine-tuned on. Moreover, fine-tuning or training from scratch on an age balanced dataset does not achieve the same performance across ages, showing in general worse results for the older age group and better results for the younger or middle groups. Therefore, we conclude that balanced training data is not the simple answer to achieve same performance across ages; rather, better loss functions are desirable.

**Time-lapse changes impostor as well as genuine.** The "template aging" effect is well known - increasing time lapse between images in an genuine pair results in a lower similarity score. We show that an increased difference in the age between two persons in an impostor pair also results in a lower similarity score. This is perhaps intuitive – images of two different persons of the same age are likely to look more alike than images of two different persons with large age difference. This result appears consistent with results presented by Grother [9]. However, our result for the old age range is atypical on this point; increased age difference between persons in an impostor pair has a less predictable effect, staying about the same or even increasing slightly in one case. We do not yet have any confident speculation for the cause of this effect.

## References


[1] Morph dataset. `https://www.faceaginggroup.com/?page_id=1414`.

[2] S. H. Abdurrahim, S. A. Samad, and A. B. Huddin. Review on the effects of age, gender, and race demographics on automatic face recognition. *Visual Computer*, 34(11):1617–1630, 2018.

[3] L. Best-Rowden and A. K. Jain. A longitudinal study of automatic face recognition. In *International Conference on Biometrics*, 2015.

[4] L. Best-Rowden and A. K. Jain. Longitudinal study of automatic face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):148–162, 2018.

[5] J. R. Beveridge, G. H. Givens, P. J. Phillips, and B. A. Draper. Factors that influence algorithm performance in the Face Recognition Grand Challenge. *Computer Vision and Image Understanding*, 113(6):750–762, 2009.

[6] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Face and Gesture Recognition*, 2018.

[7] C. M. Cook, J. J. Howard, Y. B. Sirotin, and J. L. Tipton. Demographic Effects in Facial Recognition and their Dependence on Image Acquisition: An Evaluation of Eleven Commercial System. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1), 2019.

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] P. Grother. Bias in face recognition: What does that even mean? and is it serious? *Biometrics Congress*, 2017.

[10] J. Guo. Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface, last accessed on June 2019.

[11] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European Conference on Computer Vision*, 2016.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

[13] J. J. Howard, A. A. Blanchard, Y. B. Sirotin, J. A. Hasselgren, and A. Vemury. An investigation of high-throughput biometric systems: Results of the 2018 department of homeland security biometric technology rally. In *IEEE International Conference on Biometrics: Theory Applications and Systems*, 2018.

[14] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.

[16] K. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K. Bowyer. Characterizing the variability in face recognition accuracy relative to race. In *Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2019.

[17] A. Lanitis and T. Cootes. Fg-net aging data base. *Cyprus College*, 2002.

[18] B. Lu, J. Chen, C. D. Castillo, and R. Chellappa. An experimental evaluation of covariates effects on unconstrained face verification. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 40(1), 2019.

[19] Y. M. Lui, D. Bolme, B. A. Draper, J. R. Beveridge, G. Givens, and P. J. Phillips. A meta-analysis of face recognition covariates. In *IEEE 3rd International Conference on Biometrics: Theory, Applications and Systems*, 2009.

[20] R. C. Malli. Vggface implementation with keras framework. https://github.com/rcmalli/keras-vggface, last accessed on June 2019.

[21] B. Maze, J. Adams, J. A. Duncan, N. Kalka, T. Miller, C. Otto, A. K. Jain, W. T. Niggel, J. Anderson, J. Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *International Conference on Biometrics*, 2018.

[22] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone. Face Recognition Vendor Test 2002: Evaluation Report. *NIST IR 6965*, 2003.

[23] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Computer Vision and Pattern Recognition (CVPR)*, 2005.

[24] K. Ricanek and T. Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, 2006.

[25] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] H. Taniai. Facenet implementation with keras framework. https://github.com/nyoki-mtl/keras-facenet, last accessed on June 2019.

[27] C. Whitelam, E. Taborsky, A. Blanton, B. Maze, J. Adams, T. Miller, N. Kalka, A. K. Jain, J. A. Duncan, K. Allen, J. Cheney, and P. Grother. IARPA Janus Benchmark-B Face Dataset. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2017-July:592–600, 2017.

[28] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 2016.