Issues Related to Face Recognition Accuracy Varying Based on Race and Skin Tone

K. S. Krishnapriya, Vítor Albiero, *Student Member, IEEE*, Kushal Vangara, Michael C. King, *Member, IEEE*, and Kevin W. Bowyer, *Fellow, IEEE*

***Abstract*—Face recognition technology has recently become controversial over concerns about possible bias due to accuracy varying based on race or skin tone. We explore three important aspects of face recognition technology related to this controversy. Using two different deep convolutional neural network face matchers, we show that for a fixed decision threshold, the African-American image cohort has a higher false match rate (FMR), and the Caucasian cohort has a higher false nonmatch rate. We present an analysis of the impostor distribution designed to test the premise that darker skin tone causes a higher FMR, and find no clear evidence to support this premise. Finally, we explore how using face recognition for one-to-many identification can have a very low false-negative identification rate and still present concerns related to the false-positive identification rate. Both the ArcFace and VGGFace2 matchers and the MORPH dataset used in our experiments are available to the research community so that others should be able to reproduce or reanalyze our results.**

***Index Terms*—Biometrics, face recognition, false match rate (FMR), false nonmatch rate (FNMR), gender, one-to-many identification, race, skin tone, social impact.**

I. Introduction

FACE recognition technology has been the subject of privacy concerns in the past [5], but recently has also become controversial over concerns about bias, or fairness. Waves of critical media coverage (e.g., [29] and [46]) have resulted from incidents, such as the ACLU's matching of members of the U.S. Congress to arrest photographs [43] and the Georgetown Law Center's "Perpetual Line-up" report [16]. In the wake of this publicity, first San Francisco and then other localities passed laws banning or regulating the use of face recognition technology by government agencies [33]. It seems clear that government use of face recognition technology can give rise to strong reactions that may or may not be based on an accurate understanding of the technology.

In this article, we explore three aspects of face recognition related to concerns about possible bias. First, we explore how face recognition accuracy varies between African-American and Caucasian image cohorts. Different from some past research analyzing older face recognition technology, our results show that the African-American image cohort has a slightly *better* receiver operating characteristic (ROC) curve. However, analysis of the impostor and genuine distributions underlying the ROC curves shows that the African-American cohort has a higher false match rate (FMR), and the Caucasian cohort has a higher false nonmatch rate (FNMR). Thus, accuracy does vary between demographic groups, with each cohort being advantaged on one error rate and disadvantaged on the other. Second, we present an experimental test of the premise that darker skin tone alone drives an increase in the FMR. To our knowledge, this is the first such experiment designed to separate the effects of skin tone from the effects of differences in face morphology. Our results indicate that a *similar* skin tone between a pair of images is correlated with increased likelihood of a false match. However, there is no clear evidence that darker skin tone by itself is correlated with an increased likelihood of a false match. Third, we show that while face recognition can support a one-to-many search with a very low false-negative identification rate (FNIR), it may be more difficult to control the number of false positives.

Manuscript received December 20, 2019; revised February 6, 2020; accepted February 8, 2020. Date of publication February 17, 2020; date of current version March 11, 2020. This article was recommended for publication by Associate Editor R. A. Calvo. *(Corresponding author: Kevin W. Bowyer.)*

K. S. Krishnapriya, Kushal Vangara, and Michael C. King are with the Computer Science Department, Florida Institute of Technology, Melbourne, FL 32901 USA (e-mail: kks2017@my.fit.edu; kvangara2015@my.fit.edu; michaelking@fit.edu).

Vítor Albiero and Kevin W. Bowyer are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: valbiero@nd.edu; kwb@nd.edu).

Digital Object Identifier 10.1109/TTS.2020.2974996

II. Literature Review

There is a small but growing body of literature dealing with differences in face recognition accuracy between demographic groups. This section reviews work that deals specifically with accuracy differences between African-American and Caucasian image cohorts.

Klare *et al.* [27] analyzed demographic accuracy differences using the Pinellas County Sheriff's Office (PCSO) image dataset and multiple commercial, off-the-shelf (COTS) and open-source matchers. The matchers are from before the wave of deep convolutional neural network (CNN) algorithms in face recognition. As pointed out in a recent National Institute of Standards and Technology (NIST) report [20], deep learning algorithms have been an "industrial revolution" for face recognition, and current algorithms may have different properties than past algorithms. Klare *et al.* [27] reported accuracy in terms of ROC curves and verification rate at a fixed FMR. They conclude that "The female, Black, and younger cohorts are more difficult to recognize for all matchers used in this article (commercial, nontrainable, and trainable)."

This article appears to be the primary technical basis cited by the Georgetown Law Center's "Perpetual Line-up" report [16]. The PCSO dataset used in this article is *not* available for analysis by others in the research community.

El Khiyari and Wechsler [14] reported results using a small subset of the MORPH dataset [41]. They select 362 subjects with 1448 images for an African-American cohort and the same for a Caucasian cohort. Details of the selection are not provided. Comparing a matcher based on the Oxford Visual Geometry Group's VGG network [39] and a COTS matcher, they find substantially better ROCs for the VGG-based matcher. Comparing accuracy across Caucasian and African-American cohorts, for the VGG matcher the ROC for the Caucasian cohort is much better than that for the African-American. For the COTS matcher, the two ROCs are much more similar. There is a principal components analysis (PCA) dimensionality reduction of the VGG feature vector, which is seemingly trained on the same data that accuracy is reported for, which may have resulted in optimistic results for the VGG-based matcher.

Grother [19] presented results from the NIST "FRVT Ongoing" study that document race and gender accuracy differences. ROC results show that the Caucasian ROC is better than the African-American ROC. However, specifically for FMR and FNMR, he reports that "African-Americans give slightly lower FNMR than Whites" and "African-Americans give much higher FMR than Whites." These results are based on a dataset of about 600,000 mugshot-style images. The dataset used in this article is *not* available to other researchers.

Cook *et al.* [10] analyzed demographic factors in a study of 11 unattended image acquisition stations. The "same-day" portion of their dataset contains 363 subjects. Each subject has one image acquired under supervised conditions, and 11 additional images, one from each of 11 different unattended acquisition stations. In addition to these same-day images, 326 of the 363 persons, plus another 199 persons, contributed an average of 3.5 images each to a historic gallery of 1848 images acquired in the previous four years by varied cameras. Exploiting the 18% gray background in the controlled enrollment image, a measure of relative skin reflectance was computed for the 363 persons. They report that darker skin tone is associated with longer image acquisition times and with lower similarity scores for genuine image pairs. They also report that the skin reflectance measure was a statistically better predictor than self-reported race labels (56% Black or African-American, and 35% White). This may be due in part to the binary nature of the self-reported race labels versus the continuous value for skin reflectance. The dataset used in this article is *not* available for analysis by other researchers.

Howard *et al.* [26] reported a unique FMR result—"we observed the highest FMR for older males that self-identified as White and the lowest for older males that self-identified as Black or African-American." They recognize that this result appears to conflict with some previous work, and suggest that the apparent conflict is due to an interaction between race and age. Breaking the FMR comparison down by age ranges, they find that for subjects < 40 years old, African-American males have higher FMR than Caucasian males, and for subjects $>= 40$ years old, Caucasian males have higher FMR than African-American males.

In previous work, we compared the accuracy of African-American and Caucasian image cohorts for two deep CNN matchers and two COTS matchers [28]. Results of the ROC accuracy comparison were inconsistent, with two matchers yielding a better ROC for African-Americans and two yielding a better ROC for Caucasians. However, accuracy comparison based on the FMR and FNMR curves was consistent across the four matchers. For a given decision threshold on similarity, the African-American cohort had a higher FMR and the Caucasian cohort had a higher FNMR. These results illustrate the importance of comparing accuracy at a deeper level than the ROC curve. The deep CNN matchers used in [28] have been surpassed in accuracy by the ArcFace matcher used in experiments in this article. The MORPH dataset used in this previous work and in this article is described in more detail in a later section.

Lu *et al.* [30] reported on face recognition accuracy broken out by Fitzpatrick skin tone [15]. Their analysis uses the IARPA Janus Benchmark (IJB) datasets, which do not include metadata for race but do include skin tone ratings obtained via Mechanical Turk [31]. They report that the image cohort for skin tone I (lightest) has the best ROC. However, they find that the image cohort for skin tone V has the worst ROC, with the skin tone VI (darkest) cohort having the second worst. Limitations of the IJB datasets for this analysis include that the numbers of subjects and images are small for some skin tones and uneven across skin tones, that the variation in other factors (gender balance, age balance, facial expression, off-angle pose, etc.) is not controlled across skin tones, and that some skin tones may include substantial numbers of persons of multiple races.

Wang *et al.* [47] evaluated four COTS algorithms and four CNN matchers, including a version of ArcFace trained on CASIA-Webface, and report that Caucasians have higher verification accuracy on a subset of the racial faces in the wild (RFW) dataset focused on difficult pairs. Their RFW dataset is constructed by selecting images from the MS-Celeb-1M dataset, using the nationality attribute of Freebase [8] to select Indians and Chinese, and using a commercial tool to predict Caucasian and African-American. Their results could differ from those in this article due to the use of a different training dataset, and/or the method of creating their RFW dataset. They propose a deep information maximization adaptation network designed to reduce bias in face recognition accuracy.

Gong *et al.* [17] introduced "DebFace" for "debiasing" face recognition by "disentangling" features related to demographics and identity. Accuracy is reported in terms of area under the ROC curve (AUC), a value that ranges from 0.5 (random guessing) to 1 (no errors). DebFace achieves an AUC of 0.731 for White and 0.691 for Black, compared to 0.737 and 0.678, respectively, for a baseline matcher. Thus, DebFace reduces the accuracy gap by decreasing the accuracy for White and increasing the accuracy for Black by a larger amount. Interestingly, accuracy was highest for the Indian

image cohort, at 0.809 for DebFace and 0.835 for the baseline. Verification accuracy for DebFace on the labeled faces in the wild (LFW) dataset is reported as 98.97%, as compared to 99.38% for the baseline matcher and 99.83% for the ArcFace matcher. These results illustrate two challenges for "de-biased" matchers: 1) does reducing bias necessarily mean lower accuracy for the initially higher accuracy group? 2) can a de-biased matcher improve on the state-of-the-art accuracy?

Grother *et al.* [22] presented extensive results in a recent report on demographic effects, spanning a number of algorithms, datasets of different quality, and multiple races. Results relevant to comparing accuracy for African-American and Caucasian cohorts are largely similar to our findings. False positives are generally higher for West and East Africans and lower for Eastern Europeans. False positives are also higher for women than for men. With mugshot quality images, false negatives are higher for Caucasians and lower for African-Americans. However, with lower quality images, "false negatives are generally higher in people born in Africa and the Caribbean, the effect being stronger in older individuals" [22]. They also find that false negatives are generally higher for women, but that there are exceptions.

Cavazos *et al.* [9] studied the accuracy comparison of face recognition algorithms across Caucasian and Asian datasets. They also discuss the "other race effect" for face recognition by humans and by algorithms. And, importantly, they discuss some general considerations for sound research in comparing face recognition accuracy. Nagpal *et al.* [35] explored the effects of training data demographics on face recognition accuracy. They show that, for the matcher and dataset used in their work, if the training data is entirely African-American, then accuracy is higher for African-Americans than Caucasians, and if the training data is entirely Caucasian, then accuracy is higher for Caucasians.

Buolamwini and Gebru [7] evaluated the accuracy of three commercial facial analytics tools that predict gender from a face image. They use a set of images collected from the web, for male and female parliamentarians from six countries in northern Europe and in Africa. Images were manually labeled with gender and Fitzpatrick skin tone, and the accuracy of each of the commercial gender prediction tools was computed. They conclude that each of the three gender prediction tools is: 1) more accurate for males than for females; 2) more accurate for lighter skin tones than for darker skin tones; and 3) least accurate for darker skin tone female faces.

Muthukumar *et al.* [34] followed [7] to further explore accuracy in gender classification. Based on an experiment using color space operations to alter apparent skin tone, they conclude that, "The main finding, perhaps surprisingly, is that skin type is not the driver." In their initial results, of the gender classification errors of darker-skin-tone females, 29 of 33 misclassified images had short hairstyle. However, an experiment using images cropped to remove the hairstyle cues shows the same pattern of relative accuracy, indicating that hairstyle is also not the driving factor. In an experiment to determine what features drive gender classification, they find that "the lips, eyes, and cheeks show up very prominently as a sufficient explanation for a female classification, whereas the nose and forehead areas support a male classification."

Das *et al.* [12] presented an approach for gender, age, and race classification that is designed with the goal of minimizing accuracy differences between demographic groups. They report results for their multitask CNN approach that ranked first in the bias estimation in face analytics (BEFA) 2018 workshop challenge. They also explore how the accuracy of the prediction of one attribute varies with other attributes.

There is work in psychology aimed at teasing apart the effects of skin color and face structure for human face recognition. Bar-Haim *et al.* [4] reported on an experiment with observers viewing four categories of face images—"...16 original African faces (African features/black skin), 16 original Caucasian faces (Caucasian features/white skin), 16 whitened African faces (African features/white skin), and 16 blackened Caucasian faces (Caucasian features/black skin)." They conclude that skin color is less important than face structure—"...despite the notion that skin color plays a major role in categorizing faces into own and other-race faces, its effect on face recognition is minor relative to differences across races in facial features." Brooks and Gwynn [6] reported similar conclusions. Manipulating skin tone of face images in a classification task, they find that face morphology rather than skin tone drives categorization of perceived race—"Our results show a clear dissociation between ratings of perceived race and perceived skin tone. Although surrounds of Black (White) faces cause a central face to appear lighter (darker), they produce no change in perceived racial typicality. This suggests that perceived race is not determined by skin tone but instead by morphological characteristics, which are unaffected by these surrounds."

This work from psychology deals with human abilities to analyze face images, rather than face recognition by algorithms. But it does suggest that the role of skin tone should be carefully evaluated as a possible cause of accuracy differences rather than simply assumed to be a dominant factor. This is especially true in light of the results by Muthukumar *et al.* [34] indicating the importance of face features/morphology in gender prediction.

Table I summarizes selected elements of the works described above. Note that the datasets used in [10], [17], [19], [22], [26], and [27] are *not* available for analysis by other researchers. Results presented in this article are obtained using the MORPH dataset, as used in [14] and [28], which is openly available. The IJB dataset [32] is also available to the research community, but, as discussed above, has limitations for this article. Note that the COTS face matcher(s) used in previous work are generally unnamed, due to license restrictions. Results in this article are obtained using the publicly available ArcFace and VGGFace2 matchers. Finally, note that there is not uniform agreement across the accuracy comparisons in Table I, but the most common result for the most recent matchers is that, for a fixed decision threshold, African-Americans experience a higher FMR and Caucasians experience a higher FNMR.

TABLE I
COMPARISONS OF AFRICAN-AMERICAN AND CAUCASIAN FACE RECOGNITION ACCURACY IN PREVIOUS WORKS

Reference	Accuracy Comparisons			Dataset			Matcher(s)
	ROC	FMR	FNMR	Subjects	Images	Available?	
Klare et al. [27], 2012	C better	NA	NA	16K	103K	No (PCSO)	LBP, Gabor, 4SF, COTS x3
El Khiyari [14], 2016	C better	NA	NA	724	2,896	MORPH	VGG, COTS x1
Grother [19], 2017	C better	A-A better	C better	?	300K	No	?
Cook et al. [10], 2019	NA	C better	NA	363+199	6K	No	COTS x1
Howard et al [26], 2019	NA	NA	C better	363+199	6K	No	COTS x1
Lu et al. [29], 2019	Fitzpatrick skin tone ROCs: 1 best, 5 worst			3.5K	33K	IJB-C	fusion of 5 CNN models
Krishnapriya et al. [28], 2019	C better	A-A better	C better	11K	45K	MORPH	VGG, COTS-A
Krishnapriya et al. [28], 2019	C better	A-A better	C better	11K	45K	MORPH	Resnet, COTS-B
Wang et al [47], 2019	NA	NA	C better	6K	20K	Yes	COTS x4, open CNN x4
Gong et al [17], 2019	C better	NA	NA	16K	103K	No (PCSO)	DebFace
Grother et al [22], 2019	NA	C better	A-A better	multiple large datasets		No	multiple
Buolamwini [7], 2019	face analytics for gender prediction, no face recognition results						
Muthukumar [34], 2018	face analytics for gender prediction, no face recognition results						
Das [12], 2018	face analytics for gender, race and age prediction, no face recognition results						

III. FACE MATCHERS AND IMAGE DATASET

ArcFace is a current state-of-the-art deep CNN matcher, and VGGFace2 is a slightly older matcher using a different network structure, loss function, training set, input image size, and feature vector size. MORPH is a widely used dataset that is particularly well suited for this article. ArcFace, VGGFace2, and MORPH are freely available to the research community. Thus, other researchers should be able to reproduce and extend the results presented here, if desired.

TABLE II
SUMMARY OF CURATED MORPH ALBUM 2

	Male	Female	Total
African-American	36,837 images 8,863 persons	5,782 images 1,500 persons	42,619 images 10,363 persons
Caucasian	8,006 images 2,133 persons	2,606 images 637 persons	10,612 images 2770 persons
Total	44,843 images 10,996 persons	8,388 images 2,137 persons	53,231 images 13,133 persons

A. ArcFace and VGGFace2

ArcFace [13] is based on the ResNet-100 architecture [25], trained with the MS1M V2 dataset (a cleaned version of the MS-Celeb-1M dataset [24]), using additive angular margin loss. We use the instance of ArcFace represented by a set of pretrained weights, available online [23], with no fine-tuning.

As input to ArcFace, faces were detected and aligned using MTCNN [48], and resized to the 112×112 used by ArcFace. The 512-D feature vector for a face is taken from the next-to-last layer of the network. Cosine similarity is used for the similarity between feature vectors from two face images.

ArcFace training optimizes the geodesic distance margin by utilizing the exact correspondence between arc and angle in a normalized hypersphere. ArcFace appears to provide stable performance without being combined with other loss functions, and also appears to facilitate relatively easy convergence on the training data.

VGGFace2 is representative of the state-of-the-art in CNN matchers prior to ArcFace. (The name "VGGFace2" has come to refer both to the second image dataset from the Oxford VGG and to a popular deep CNN matcher trained on the dataset.) It is based on the ResNet-50 network structure trained on the VGGFace2 dataset [8] with standard softmax loss. As with ArcFace, faces are detected and aligned using MTCNN. Faces are resized to 224×224 pixels for VGGFace2, and a 2048-D feature vector taken from the next-to-last layer. Again, cosine similarity is used for the similarity between feature vectors. The VGGFace2 model used in this article is publicly available [31].

B. MORPH Dataset

MORPH [41] is available from UNC-Wilmington [44]; the noncommercial release of MORPH is used in this article. Images contained within the MORPH dataset come from public records. All metadata associated with the images are also available from public records. MORPH has been extensively used in research in face aging; for example, a recent paper on predicting age from face images states [42]—"FG-NET and MORPH datasets with face images and (real) age labels are the most used datasets allowing for comparison of methods..." MORPH has recently also been used in research on demographic variation in face recognition accuracy [2], [14], [28], and is particularly useful for this because it has a large number of images of both African-Americans and Caucasians, acquired under similar, controlled conditions. MORPH images are frontal with controlled lighting and background; examples are shown in Fig. 1.

The numbers of subjects and images in the version of MORPH used in this article are summarized in Table II. For example, the African-American male cohort of MORPH is 8,863 subjects and 36,837 images. These numbers are far larger than the 562 (363+199) total subjects in the dataset used in [8] and [26], which in any case is not available

Fig. 1. Example images from the MORPH dataset. Images in the MORPH dataset are nominally frontal pose, neutral expression, controlled indoor lighting, and uniform background; minor variation in pose, expression, and use of eyeglasses can be observed. Face matching accuracy with this quality of images may be higher than for images under "in the wild" or "Web-scraped" conditions. The original images are from public records; however, black rectangles are added to the images in this figure in an effort to respect individual anonymity and privacy.

to other researchers. The 8,850 African-American male subjects are also greater than the 3,531 total subjects in the IJB-C dataset [32] used in [30]. The IJB-C dataset has the additional complications that: 1) although it has Mechanical Turk-supplied Fitzpatrick skin tone ratings, it does not have race labels and 2) it intentionally samples highly varied pose and occlusion, making it difficult to assume approximately equal image quality across demographic groups.

Overall, we conclude that MORPH is currently the best openly available dataset to explore issues of race-based differences in face recognition accuracy. Other datasets are either 1) not available to the research community; and/or 2) do not have race labels; and/or 3) are much smaller.

IV. Accuracy Differences Based on Race

The relative accuracy of competing algorithms, compared on the same dataset, is commonly presented in ROC curves. ROC curves have also been used to compare the accuracy of a particular face recognition algorithm across different demographic groups [14], [27], [28], [30]. However, there is a potential problem with ROC comparisons of accuracy for different demographic groups, meaning different datasets.

The horizontal axis of the ROC curve is the FMR. An algorithm typically achieves the same FMR for different demographic groups at different thresholds on the similarity score. But operational scenarios typically use a fixed similarity threshold for images of all persons. Thus, the ROC comparison may not correspond to how face recognition is used in an operational scenario. Researchers have recently begun to make this explicit by annotating the ROC curves [1], [14]. In this section, acknowledging its limitations, we first present the ROC comparison. Then, we present the comparison for the underlying FMR and FNMR curves, which make the dependence on the similarity threshold explicit, and so give a more explicit comparison.

A. ROC Curve Comparison

Fig. 2 compares the ROC curves for the African-American and Caucasian cohorts of the MORPH dataset, broken out by male and female, for the two matchers. For each of the demographic cohorts, at an FMR of 1-in-100,000, ArcFace achieves a TPR exceeding 99%. The MS-Celeb-1M V2 dataset used to train ArcFace has no overlap of subjects or images with MORPH, and so ArcFace's accuracy here, noticeably higher than that of VGGFace2, indicates a substantive improvement over earlier deep learning matchers.

Several important observations can be noted from Fig. 2. First, the ROC comparisons show better accuracy for African-Americans than for Caucasians. Also, for both the African-American cohort and the Caucasian cohort, the ROC for males shows higher accuracy than the ROC for females. This accuracy difference between males and females appears to be a general phenomenon, noted by a number of researchers, whose cause(s) are not yet understood [2]. The gender-based accuracy difference shown here is at least as large as the race-based accuracy difference.

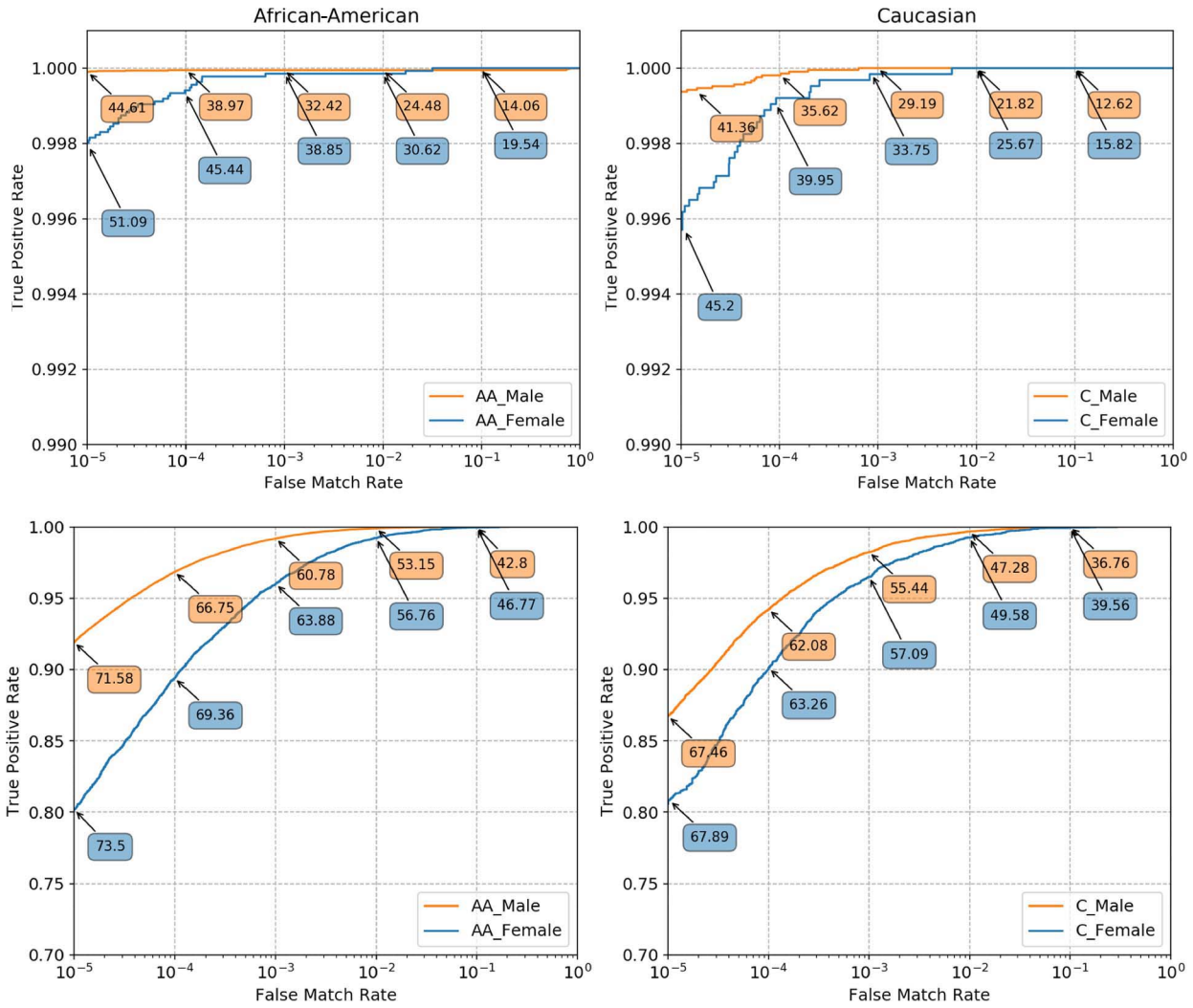

Fig. 2. Accuracy comparison by ROC curves for African-American and Caucasian, with male/female breakdown. ArcFace results above, VGGFace2 results below. Due to the higher accuracy of ArcFace, in order to better visualize differences, the vertical axis of its plots starts at 99%. On the horizontal axis, the FMR ranges from 100\% on the right down to 1-in-100,000 on the left. Annotations on the curves indicate the decision threshold value at which that cohort achieves the particular FMR.

Second, note that the similarity thresholds at which the various cohorts achieve a given FMR are different. For the ArcFace matcher, Caucasian males achieve a given FMR at the lowest similarity threshold, followed by African-American males, then Caucasian females, and then African-American females. This suggests that the relative accuracy between demographic groups in terms of FMR and FNMR is more nuanced than is revealed in the ROC curves. This is analyzed in more detail in the FMR and FNMR curves analyzed in the next section.

Finally, it is interesting to note that African-American subjects are greatly *under-represented* in the MS-Celeb-1M V2 dataset that was used to train ArcFace. Wang *et al.* [47] estimated the African-American fraction of MS-Celeb-1M at about 15%. Thus, the ROC results here *do not support* the common assumption that the training dataset must be demographically balanced in order for the accuracy achieved by the matcher to be comparable. Rather, the results show that it is possible for the ROC accuracy for the African-American cohort to be as high or higher than that of the Caucasian cohort even when African-Americans are substantially under-represented in the training data.

B. FMR/FNMR Comparison

Comparing the FMR and FNMR across the range of the similarity threshold gives a more detailed picture of relative accuracy than does comparing ROC curves. The FMR and FNMR for the African-American and Caucasian male cohorts, separated by male and female, for both deep CNN matchers, are shown in Fig. 3.

The (newer) ArcFace matcher has noticeably higher accuracy than the (older) VGGFace2 matcher, for either race and either gender. However, both matchers show the same qualitative trends across race and gender.

For both matchers, the FMR curves indicate that, at any similarity value used as a threshold, the FMR for African-Americans is higher (worse) than it is for Caucasians. The ArcFace threshold value that corresponds to a 1-in-10,000 FMR for Caucasian male (35.62) and Caucasian female

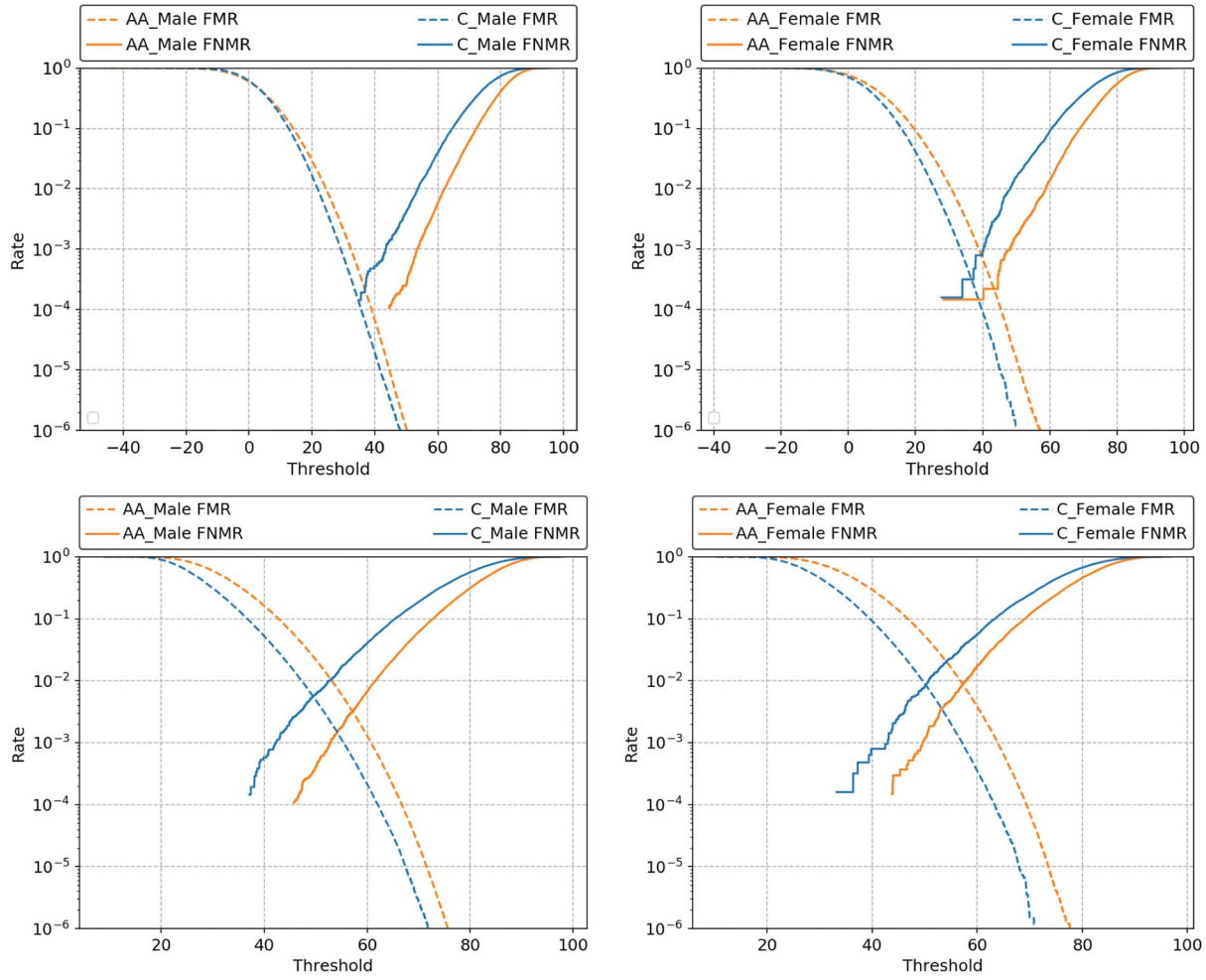

Fig. 3. FMR and FNMR accuracy comparison for African-American and Caucasian, male and female. ArcFace results above, VGGFace2 results below. The FMR curves run from an FMR of 100% (top) to 1-in-1M (bottom)., The FNMR curves run from 100% (top) to an FNMR of 1-in-10,000 (bottom). The FNMR curves cannot extend as far as the FMR curves because there are fewer genuine image pairs than impostor pairs. The horizontal axis is the useful range of the cosine similarity value. At any given threshold on similarity, the African-American cohort has a higher FMR and the Caucasian cohort has a higher FNMR. Also, for either African-Americans or Caucasians, the female cohort has a higher FMR and a higher FNMR than the male cohort.

(39.95) could approach an FMR closer to 1-in-1000 for African-American male (32.42) and African-American female (38.85). On the other hand, the FNMR curves make it clear that, over the range of decision threshold values, the FNMR for Caucasians is higher (worse) than that for African-Americans. The qualitative trends shown here for race-based and gender-based differences in FMR and FNMR are largely the same as those shown in [28] for four other (older) matchers.

C. Likely Presence of Monozygotic Twins in MORPH

Monozygotic ("identical") twin births occur at a rate of 3 to 4 per 1000 births worldwide [36]. Given that there are nearly 9000 African-American males in MORPH, and that images were acquired in limited geographic locations, it is possible that MORPH contains instances of identical twins. MORPH metadata does not include any specific indication of whether two persons are monozygotic twins. MORPH does include metadata for date of birth, and identical twins would typically have the same date of birth.

In examining the forty African-American male impostor image pairs with ArcFace similarity scores over 70, it appears that all of these result from image pairs belonging to five sets of likely monozygotic twins. The metadata for the highest-similarity Caucasian male impostor pairs does not show any pairs of subjects with the same date of birth. Thus, it is unlikely that there are any monozygotic twins in the Caucasian male image cohort. Given that the African-American male image cohort is larger that the Caucasian male cohort, and that the number of likely monozygotic twins in the African-American cohort is only five, it appears that different number of twins in the cohorts is due to random variation.

It is known that face-matching algorithms have difficulty distinguishing between identical twins [38] and that monozygotic twins can naturally cause instances of false matches that persist into higher similarity scores. Thus, if there are monozygotic twins in one cohort but not in another, then the high-similarity tails of the impostor distributions of the two cohorts must be interpreted carefully. While it appears that there are image pairs of monozygotic twins in the high-similarity tail of the African-American impostor distribution,

the number of such image pairs is too small to be a significant factor in the overall difference between the African-American and Caucasian impostor distributions.

D. Outlier Low-Similarity Genuine Image Pairs

Outlier low-similarity image pairs of the same person can occur for various reasons, and cause artifacts in the tails of the genuine distribution and the FNMR curve. Manual examination of the genuine pairs with lowest ArcFace similarity score reveals that they are at least partly caused by unusual conditions, such as bruises and/or bandages, or by off-angle pose and unusual facial expression. The presence of such image pairs can complicate the comparison of face recognition accuracy between cohorts at very low FNMR values.

V. Does Darker Skin Tone Increase False Matches?

The premise that face recognition accuracy is worse for darker skin tones has appeared in articles from major media sources. For example, a BBC article includes a figure with the caption "Face recognition tech is less accurate the darker your skin tone" [46]. Similarly, a New York Times article states that "... the darker the skin tone, the more errors arise..." [29].

Results in the previous section indicate that the African-American image cohort has a higher FMR than the Caucasian. However, as pointed out in [4] and [6], comparisons across African-American and Caucasian face images confound differences in skin tone together with differences in face morphology. In this section, we present the first experiment that we are aware of that is designed to isolate the effect of skin tone on face recognition accuracy, for a direct test of the premise that "face recognition is less accurate the darker your skin tone."

For this experiment, we analyze the African-American male impostor distribution. By focusing on a single race and gender, we avoid confounding factors of face morphology, and African-American male is the largest cohort in the dataset. We focus on the impostor distribution because the finding that the FMR for the African-American cohort is higher than for the Caucasian makes it plausible to consider if darker skin tone is a causal factor for increased FMR.

By definition, the impostor image pairs most likely to result in a false match are those with the highest similarity score. This suggests a comparison of the frequency of darker skin tone in the high-similarity tail of the impostor distribution with, for example, the center of the distribution. To represent the high-similarity tail of the impostor distribution, we sample the 500 image pairs that just exceed the 1-in-10,000 FMR threshold. To represent the center of the distribution, we sample 500 image pairs from just above a threshold corresponding to one half of the distribution. The conceptual structure of the experiment is shown in Fig. 4.

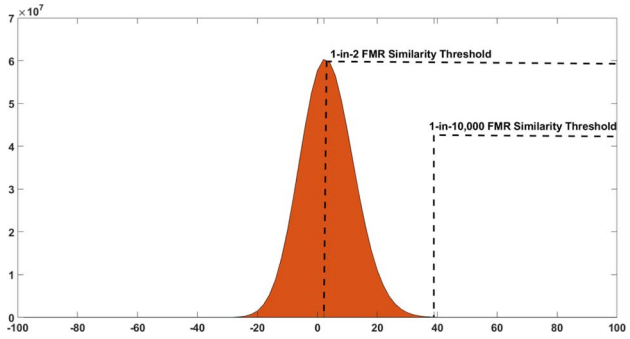

Fig. 4. Is darker skin tone correlated with increased FMR? The ArcFace impostor distribution for the African-American male cohort of MORPH is shown. The frequency of darker skin tone is compared for a "no false match region" in the middle of the distribution and a "high-likelihood false match region" in the high-similarity tail of the distribution.

A. Fitzpatrick Skin Tone Ratings

The Fitzpatrick skin typing system [15] was developed as a tool for classifying skin type in terms of reaction to ultraviolet radiation, and is widely used in dermatology; e.g., [3]. The Fitzpatrick scale runs from Type I, characterized as "Always burns—never tans" in reaction to sun exposure and "pale white" in terms of skin color, up to Type VI, characterized as "Never burns—tans profusely" and "dark brown or black" [45]. We note that Pichon *et al.* [40] argued that the "cultural sensitivity and cross-cultural utility" of the Fitzpatrick system should be improved, and also that Fitzpatrick himself indicated a need for "research and development of a more objective method that would replace this simple classification of skin types" [15]. However, the Fitzpatrick I–VI skin tone rating is the appropriate choice for this article due to its simplicity and widespread use, including prior use in the face recognition research community; e.g., metadata for face images in the IARPA IJB datasets [32], work by Buolamwini and Gebru [7], Lu *et al.* [30], and Muthukumar *et al.* [34].

We have 500 image pairs sampled from a "no-false-match region" in the center of the impostor distribution, and another 500 sampled from a "high-likelihood-false-match region" in the high-similarity tail of the distribution. There are up to 1000 distinct images, of up to 1000 different persons, for each region. In practice, the number of distinct images and persons can be less due to one image occurring as part of multiple image pairs, and/or one person having multiple images among the 500 pairs. For the ArcFace matcher, for the 500 image pairs from the center of the distribution, we have 915 distinct persons and 982 distinct images, and for the 500 image pairs from the high-similarity tail of the distribution, we have 872 distinct persons and 967 distinct images. The numbers are similar for the VGGFace2 matcher.

Each of three viewers independently assigned a skin tone rating for each image, without knowing which region of the impostor distribution an image came from, and without knowing other viewers' ratings. If two or three of the viewers agreed on the skin tone rating, that was used as the rating for the image. If the three viewers gave different ratings, then the middle rating of the three was used.

For ArcFace, for the 982 distinct images from the center of the distribution, all three reviewers gave the same rating to 352 images (36%), two reviewers agreed for 586 of the images (60%), and the reviewers gave three different ratings

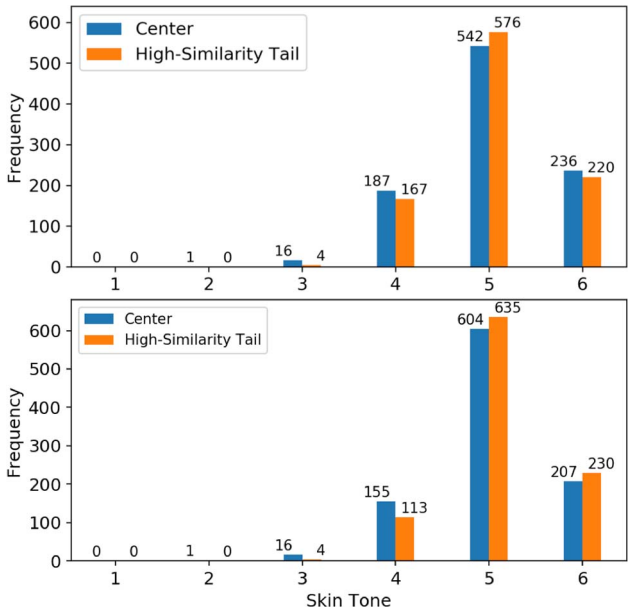

Fig. 5. Comparison of skin tone ratings for images from center and high-similarity tail of impostor distribution. ArcFace above, VGGFace2 below.

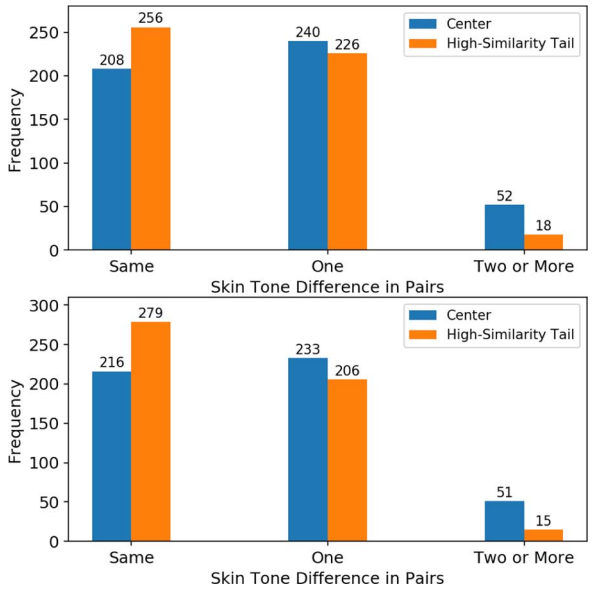

Fig. 6. Distribution of difference in skin tone rating for image pairs from high-similarity tail and center of distribution. ArcFace above, VGGFace2 below. For both matchers, image pairs from the high-similarity tail of the distribution have more similar skin tone.

for 44 images (4%). For the 967 images from the high-similarity tail of the distribution, all three reviewers agreed on 302 images (31%), two agreed for 602 of the images (62%), and the reviewers gave three different ratings for 63 images (7%). For VGGFace2, for the images from the center of the distribution, three reviewers all agreed on 332/983 (34%), two agreed on 606/983 (62%), and the three disagreed on 45/983 (5%). For VGGFace2, for the images from the high-similarity tail of the distribution, three reviewers all agreed on 237/982 (24%), two agreed on 686/982 (70%), and the three disagreed on 59/982 (6%). Overall, across the four samplings (two samples from each of two matchers), two or three viewers agreed on the skin tone rating for 93% to 96% of images. This level of consistency was judged as good for the purposes of our experiment. However, this analysis also suggests that using a single viewer to assign skin tone ratings could be problematic.

B. Comparison of Per-Image Skin Tone Distribution

Fig. 5 shows the distribution of skin tone ratings for images sampled from the center and the high-similarity tail of the impostor distribution for ArcFace and for VGGFace2. For both matchers and both samples, the most frequent skin tone rating is V, followed by IV, then VI, and then III. The number of skin tone VI images decreases slightly for ArcFace between the center and the high-similarity tail, and increases slightly for VGGFace2. Overall, the distribution of single-image ratings does not reveal any major shift toward darker skin tone for images from the high-similarity tail of the distributions.

C. Comparison of Image-Pair Skin Tone Distribution

A false match is a result obtained for an image pair, rather than a single image. This motivates examining the frequency of image pairs where both images are rated as darker skin tone. Image pairs from the high-similarity tail of the distribution naturally have greater similarity than pairs from the center of the distribution. We are interested in whether pairs that are more similar and also rated as having darker skin tone occur more frequently in the high-similarity tail.

We first examine whether image pairs with similar skin tone are more frequent in the high-similarity sample. We consider image pairs in terms of the skin tone ratings being the same, being one skin tone apart, or more than one skin tone apart. The distribution of image pairs in these three categories is shown in Fig. 6. For both matchers, in going from the center of the distribution to the high-similarity tail, the frequency of same-skin-tone pairs increases markedly, the frequency of one-skin-tone-difference pairs decreases, and the frequency of more-than-one-skin-tone-difference decreases markedly. This shows that same-skin-tone pairs occur more frequently in the high-similarity tail of the distribution, but does not necessarily support a conclusion that darker-skin-tone pairs occur more frequently.

To get at the particular question of whether same-skin-tone pairs with darker skin tone occur more frequently, we compare the distribution of skin tone IV, V, and VI pairs across the two samples. If darker skin tone is a driving cause of false matches, then we could expect skin tone VI pairs to increase disproportionately in the high-similarity sample. The results in Fig. 7 show that, for ArcFace, skin-tone-IV pairs increase from 22 to 31 (41%), skin-tone-V pairs increase from 157 to 185 (18%), and skin-tone-VI pairs increase from 29 to 40 (38%). Thus, while the number of skin-tone-VI pairs does go up by 38%, the number of skin tone IV pairs goes up slightly more, by 41%. For VGGFace2, skin-tone-IV pairs increase from 8 to 13 (63%), skin-tone-V pairs increase from 184 to 220 (20%), and skin-tone-VI pairs increase from 24 to 46 (92%). Thus, while skin-tone-VI pairs increase in the high-similarity sample, skin-tone-IV pairs also increase. In fact, for ArcFace, the more

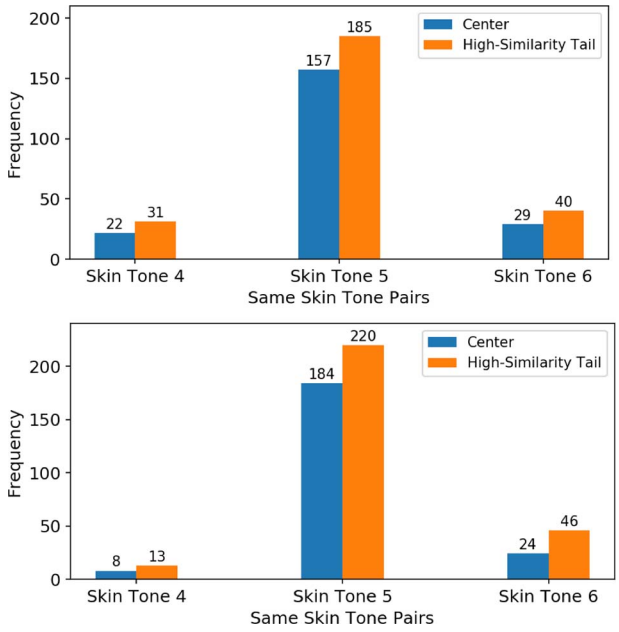

Fig. 7. Comparison of distribution of skin tone IV, V, and VI pairs at the center versus high-similarity tail.

accurate of the two matchers, skin-tone-IV pairs increase by a larger percent. For VGGFace2, skin-tone-VI pairs increase by a larger percent than skin-tone-IV pairs, but just a few image pairs difference would change this. Overall, there is no clear evidence to support that darker skin tone, in and of itself, is a driving factor in causing false match results. The issue merits further study, because of its importance, but these results suggest that factors other than simply skin tone should be considered in looking for the cause of the accuracy difference between African-Americans and Caucasians.

VI. One-to-Many Identification Searches

Results presented in previous sections are in the context of identity verification. In verification, two images are compared to determine whether they are of the same person. Modern face recognition technology has become powerful enough that it can also achieve impressive accuracy in one-to-many identification. For example, the recent NIST report [20]. In one-to-many identification, a probe image of a person whose identity is to be determined is matched against many enrolled images with known identities in order to determine if one is a match. In most real-world application scenarios, the person in the probe image may or may not have an image in the set of enrolled images. This is termed an "open-set search." In general, an open-set, one-to-many search may return zero, one or more possible matching identities.

Two types of errors occur in one-to-many identification. A false-negative identification is when there is an enrolled image corresponding to the person in the probe image, but that image is not returned as the result of the search. A false-positive identification is when either the person in the probe image has no enrolled image, and the search returns an image of a different person or the person in the probe image has an enrolled image, but the search returns an image of a different person.

The recent NIST report [21] makes an important observation about false positives in one-to-many search—"Note that when a fixed number of candidates are returned, the false-positive identification rate of the automated face recognition engine will be 100%, because a probe image of anyone not enrolled will still return candidates." To control the number of false positives, a threshold can be used for the minimum similarity required to report a match. Alternatively, in some scenarios, a manual review of the identification results can be used as a means to filter out false positives.

To illustrate the accuracy of face recognition for one-to-many search, we split the African-American male image cohort into a set of probe images and a set of enrolled images. The set of probe images consists of the most recent image of each person who has two or more images in MORPH. There are 8544 probe images. The set of enrolled images consists of all the remaining 28,294 images. Each probe is matched against all the enrolled images to find the one with the highest similarity value, the "rank-one match." In this experiment, with the ArcFace matcher, 8541 of the 8544 probe images, or 99.96%, result in a rank-one match that is the correct identity. The lowest similarity value for across the 8541 correct rank-one matches is 42.5. This high level of rank-one identification accuracy is in keeping with the results found in the recent NIST report, using different matchers and datasets [20].

The 99.96% rank-one identification accuracy for persons who are enrolled is impressive, but we must also consider what happens when the person in the probe image does not have an enrolled image. A variation of the above experiment can be used to estimate the rate of false-positive identifications for probes that are not enrolled. For each probe image, we find the rank-one match for the set of enrolled images excluding those of the person in the probe image. In this case, the number of enrolled images varies slightly between probe images. Of the 8544 probe images in this experiment, each of which has no true match in the enrollment when it is used as a probe, 3172 have a rank-one match above the similarity threshold of 42.5. This means that, for this particular dataset, the 99.96% correct rank-one identification accuracy of persons who *are* enrolled is achieved at a false-positive identification rate of 37% (3172/8544) for persons who *are not* enrolled. Raising the threshold required to report a match will reduce the false-positive identification rate, with the tradeoff of some increase in the FNIR. For this dataset, raising the threshold to 60 lowers the false-positive identification rate to 0.16%, or about 1 in 500, while the rank-one identification accuracy is still 98.7%, for an FNIR of under 2%. This illustrates that in general it should be possible to introduce a useful threshold for identification search results.

However, tuning a threshold for an identification search can be a complicated process for multiple reasons. The false-positive identification rate naturally grows as the size of the enrollment dataset grows, and so to maintain a given false-positive identification rate, the threshold must be updated as the size of the enrollment dataset grows. Also, the false-positive identification rate can depend in subtle ways on the

quality of the probe image, and automatic threshold adjustments to keep the same false-positive identification rate across a range of image quality may not be feasible. For these reasons and others, in an operational scenario related to public safety or national security, the results of an automated face identification search are often put through a manual review before a candidate match is considered "correct" or "real."

The Police Commissioner of New York City has described their use of face identification search in an operational scenario [37]. A probe image (or video) is sent to the Facial Identification Section of the Detective Bureau, and "a database consisting solely of arrest photographs is then searched as the sole source of potential candidates." The results of the automated face identification are then reviewed by an examiner who winnows the results to zero or one possible identification—"the facial identification team will provide only a single such lead to the case detective." This process of automated face identification embedded within a manual evaluation by an examiner results in about three quarters of searches being returned with no match—"In 2018, detectives made 7024 requests to the Facial Identification Section, and in 1851 cases possible matches were returned" [37].

This process of manual examination of the results from an automated facial identification search appears to build in real safeguards to control the number of false-positive identifications. However, such a process requires resources and experienced personnel that may not be available in all locations or feasible in all scenarios. For example, it is unrealistic to imagine such a process embedded in the analysis of video streaming in real-time from body-worn cameras.

VII. Conclusion and Discussion

Media coverage promoting the premise that "face recognition is not accurate" typically does not sufficiently characterize the quality of the data used nor reference any current, independent study of face recognition accuracy; e.g., the most recent NIST report [20]. The data in Fig. 2 shows that, at a similarity threshold that results in an FMR of 1-in-100,000, the ArcFace verification rate exceeds 99%. Similar results, for different matchers and datasets, can be found in many other recent research publications. Given this, it seems that there is no objective basis for a general assertion that "face recognition is not accurate" when good quality, frontal images such as those in this article are used in the analysis.

Concerns about accuracy seem to be expressed as a kind of shorthand for concern that accuracy varies between demographic groups in some way that is biased or unfair. As shown in Figs. 3 and 4, current face recognition technology does have different accuracy for different demographics. In comparing accuracy for African-Americans and Caucasians, for a scenario in which a fixed decision threshold is used for all persons, African-Americans have a higher FMR and Caucasians have a higher FNMR. In comparing accuracy for males and females, females have both a higher FMR and a higher FNMR. Assessing the accuracy for each demographic cohort at an operationally relevant FMR (e.g., 1-in-10,000) it is observed that African-American males have the highest verification rate, followed by Caucasian males, African-American females, and Caucasian females. Achieving this in an operational scenario would require a different similarity threshold setting for each demographic, which currently seems intractable.

The premise that "face recognition is less accurate for darker skin tone" appears to have come about as an oversimplification of previous research results on the ROC curve or FMR for African-Americans and Caucasians. Our results from an experiment designed to isolate the effect of skin tone on FMR do not support a general conclusion that darker skin tone, in and of itself, causes an increased FMR. However, this is the first such experiment that we are aware of, and more work is needed to better explore this issue. Replication by other researchers on other datasets would be valuable.

Some uncertainty and variation in the rating of images on the Fitzpatrick scale seems natural because a continuous skin tone range is being discretized into six categories. Our results suggest that a single person rating an image a single time may produce results that are too noisy, and so one future line of research could consider the method of image presentation and the number of viewers that is needed to maximize the consistency of the ratings. It would be generally useful if an accepted procedure existed that could be followed by different research groups in order to obtain consistent/comparable ratings.

The message that "face recognition is nearly 100% accurate in finding an enrolled person in a one-to-many identification search" is true. However, this message is incomplete in two important respects. The simpler of the two is that a focus on the accuracy in finding a match for a person who is enrolled must be balanced with a focus on the false matches that occur for persons who are not enrolled. The more complex of the two involves understanding how face recognition accuracy is transformed by the manner in which the technology is used to accomplish the usecase objective while ensuring the best possible experience for the user/public. Consider the context of false-positive identifications resulting from one-to-many searches with an image of a person who is not enrolled. The number of false-positive identifications experienced by the public depends on the number of times such searches are performed. If such searches are performed, for example, many more times for African-Americans than for Caucasians, then the African-American public will experience a higher false-positive identification rate than the Caucasian public. A manual review of automated face identification search results by an expert examiner [37] will not necessarily solve the problem of a bias in the initiation of identification searches. A similar concern exists for the enrollment dataset. If many more persons are enrolled for one demographic than for another, then more high-similarity false-positive identifications will naturally result.

The research reported in this article should be seen as initial results toward defining the size and scope of the problem of demographic variations in face recognition accuracy. Much work remains to be done and no doubt many researchers will make important contributions. Additional datasets are needed.

Datasets used in face recognition research are typically what other fields of research might characterize as "secondary" datasets [11]. That is, they are datasets not collected for the particular research purpose, and possibly have unknown confounding factors. Datasets may vary in their distribution of numerous covariates that impact accuracy: age of persons, time lapse between images, occlusion, facial expression, camera angle, image quality, and others. Grother *et al.* [22] suggested that the results of some accuracy comparisons are dependent on the image quality. Additional experiments are needed to replicate, disprove or better explain many points before the problem can be fully understood and effective solutions can be proposed. Experiments are needed to determine why accuracy comparisons turn out as they do. For example, is skin tone a driving factor in increasing error rates? And if skin tone really is not the driving factor in a comparison, then what is? Experiments are also needed to determine exactly how the demographic balance of the training data drives the accuracy. While it seems clear and intuitive that representation in the training data is important, some results [2] suggest that the demographic balance of the training data does not translate simply and directly into the demographic balance of accuracy. More research is also needed to investigate whether and how a demographic-aware training process might result in effectively the same impostor and genuine distributions across demographics [17], [47]. Research in demographic effects on face recognition accuracy is a growth area of significant public importance.

ACKNOWLEDGMENT

K. S. Krishnapriya led the analysis in Section IV, V. Albiero led the analysis in Section V, K. Vangara led the analysis in Section VI, K. W. Bowyer led the writing, and M. C. King and K. W. Bowyer directed the overall research.

REFERENCES

[1] V. Albiero, K. W. Bowyer, K. Vangara, and M. C. King, "Does face recognition accuracy get better with age? Deep face matchers say no," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2020.

[2] V. Albiero *et al.*, "Analysis of gender inequality in face recognition accuracy," in *Proc. WACV Workshop Demographic Variation Perform. Biometric Syst.*, 2020. [Online]. Available: https://arxiv.org/pdf/2002.00065.pdf

[3] O. Arosarena, "Options and challenges for facial rejuvenation in patients with higher Fitzpatrick skin phototypes," *JAMA Facial Plastic Surg.*, vol. 17, no. 5, pp. 358–359, 2015.

[4] Y. Bar-Haim, T. Saidel, and G. Yovel, "The role of skin colour in face recognition," *Perception*, vol. 38, no. 1, pp. 145–148, 2009.

[5] K. W. Bowyer, "Face recognition technology: Security versus privacy," *IEEE Technol. Soc. Mag.*, vol. 23, no. 1, pp. 9–19, 2004. [Online]. Available: https://ieeexplore.ieee.org/document/1273467

[6] K. Brooks and O. Gwinn, "No role for lightness in the perception of black and white? Simultaneous contrast affects perceived skin tone, but not perceived race," *Perception*, vol. 39, no. 8, pp. 1142–1145, 2010.

[7] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Mach. Learn. Conf. Fairness Accountability Transparency*, 2018, pp. 1–15.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. Face Gesture Recognit.*, 2018, pp. 67–74.

[9] J. G. Cavazos, P. J. Phillips, C. D. Castillo, and A. J. O'Toole. *Accuracy Comparison Across Face Recognition Algorithms: Where Are We on Measuring Race Bias?* Accessed: Feb. 25, 2020. [Online]. Available: https://arxiv.org/abs/1912.07398

[10] C. M. Cook, J. J. Howard, Y. B. Sirotin, J. L. Tipton, and A. R. Vemury, "Demographic effects in facial recognition and their dependence on image acquisition: An evaluation of eleven commercial systems," *IEEE Trans. Biometr. Behav. Identity Sci.*, vol. 40, no. 1, pp. 32–41, Jan. 2019.

[11] E. Cox, B. C. Martin, T. Van Staa, E. Garbe, U. Siebert, and M. L. Johnson, "Good research practices for comparative effectiveness research: Approaches to mitigate bias and confounding in the design of nonrandomized studies of treatment effects using secondary data sources," *Value Health*, vol. 12, no. 8, pp. 1053–1061, 2009.

[12] A. Das, A. Dantcheva, and F. Bremond, "Mitigating bias in gender, age and ethnicity classification: A multi-task convolution neural network approach," in *Proc. Eur. Conf. Comput. Vis. Workshops*, Sep. 2018, pp. 573–585.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4690–4699.

[14] H. El Khiyari and H. Wechsler, "Face verification subject to varying (age, ethnicity, and gender) demographics using deep learning," *J. Biometr. Biostat.*, vol. 7, no. 4, pp. 1–5, 2016.

[15] T. B. Fitzpatrick, "The validity and practicality of sun-reactive skin types I through VI," *Archives Dermatol.*, vol. 124, no. 6, pp. 869–871, 1988.

[16] C. Garvie, A. Bedoya, and J. Frankle. *The Perpetual Line-Up*. Accessed: Feb. 25, 2020. [Online]. Available: https://www.perpetuallineup.org

[17] S. Gong, X. Liu, and A. Jain. *DebFace: De-Biasing Face Recognition*. Accessed: Feb. 25, 2020. [Online]. Available: https://arxiv.org/abs/1911.08080

[18] Google. *Freebase API*. Accessed: Feb. 25, 2020. [Online]. Available: https://developers.google.com/freebase

[19] P. Grother, "Bias in face recognition: What does that even mean? and is it serious?" in *Proc. Biometr. Congr.*, 2017.

[20] P. Grother, M. Ngan, and K. Hanaoka. *NISTIR 8238: On-Going Face Recognition Vendor Test (FRVT) Part 2: Identification*. Accessed: Feb. 25, 2020. [Online]. Available: https://doi.org/10.6028/NIST.IR.8238

[21] P. Grother, M. Ngan, and K. Hanaoka. *NISTIR 8271: Face Recognition Vendor Test (FRVT) Part 2: Identification*. Accessed: Feb. 25, 2020. [Online]. Available: https://doi.org/10.6028/NIST.IR.8271

[22] P. Grother, M. Ngan, and K. Hanaoka. *NISTIR 8280: Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects*. Accessed: Feb. 25, 2020. [Online]. Available: https://doi.org/10.6028/NIST.IR.8280

[23] J. Guo. *InsightFace: 2D and 3D Face Analysis Project*. Accessed: Jun. 2019. [Online]. Available: https://github.com/deepinsight/insightface

[24] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 87–102.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: arXiv:1512.03385.

[26] J. J. Howard, Y. B. Sirotin, and A. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *Proc. Biometr. Theory Appl. Syst. (BTAS)*, Sep. 2019.

[27] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain, "Face recognition performance: Role of demographic information," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 6, pp. 1789–1801, Dec. 2012.

[28] K. S. Krishnapriya, K. Vangara, M. C. King, V. Albiero, and K. W. Bowyer, "Characterizing the variability in face recognition accuracy relative to race," in *Proc. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2019.

[29] S. Lohr, *Facial Recognition Is Accurate, If You're A White Guy*, New York Times, New York, NY, USA, Feb. 2018. [Online]. Available: https://www.nytimes.com/2018/02/09/technology/facial-recognitionrace-artificial-intelligence.html

[30] B. Lu, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An experimental evaluation of covariates effects on unconstrained face verification," *IEEE Trans. Biometr. Behav. Identity Sci.*, vol. 1, no. 1, pp. 42–55, Jan. 2019.

[31] R. C. Malli. *VggFace Implementation With Keras Framework*. Accessed: Jun. 2019. [Online]. Available: https://github.com/rcmalli/keras-vggface

[32] B. Maze *et al.*, "IARPA janus benchmark—C: Face dataset and protocol," in *Proc. Int. Conf. Biometr.*, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8411217

[33] R. Metz, *Beyond San Francisco, More Cities Are Saying No to Facial Recognition*, CNN, Atlanta, GA, USA, Jul. 2019. [Online]. Available: https://www.cnn.com/2019/07/17/tech/cities-ban-facial-recognition/index.html

[34] V. Muthukumar *et al.* (2018). *Understanding Unequal Gender Classification Accuracy From Face Images*. [Online]. Available: https://arxiv.org/abs/1812.00099
[35] S. Nagpal, M. Singh, R. Singh, and M. Vatsa. *Deep Learning for Face Recognition: Pride or Prejudiced?* Accessed: Feb. 25, 2020. [Online]. Available: https://arxiv.org/abs/1904.01219
[36] National Institutes of Health. *Is the Probability of Having Twins Determined by Genetics?* Accessed: Feb. 25, 2020. [Online]. Available: https://ghr.nlm.nih.gov/primer/traits/twins
[37] J. O'Neill, *How Facial Recognition Makes You Safer*, New York Times, New York, NY, USA, Jun. 2019.
[38] J. R. Paone *et al.*, "Double trouble: Differentiating identical twins by face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 2, pp. 285–295, Feb. 2014.
[39] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 471–487. [Online]. Available: https://www.robots.ox.ac.uk/˜vgg/software/vgg_face/
[40] L. C. Pichon, H. Landrine, Y. Hao, J. A. Mayer, and K. D. Hoerster, "Measuring skin cancer risk in African Americans: Is the Fitzpatrick skin type classification culturally sensitive?" *Ethnicity Disease*, vol. 20, no. 2, pp. 174–179, 2010.
[41] K. Ricanek and T. Tesafaye, "MORPH: A longitudinal image database of normal adult age-progression," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 2006, pp. 341–345.
[42] R. Rothe, R. Timofte, and L. V. Gool, "Deep expectation of real and apparent age from a single image without facial landmarks," *Int. J. Comput. Vis.*, vol. 126, pp. 144–157, Apr. 2018.
[43] J. Snow. *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*. Accessed: Feb. 25, 2020. [Online]. Available: https://www.aclu.org/blog/privacytechnology/surveillance-technologies/amazons-face-recognition-falselymatched-28
[44] *University of North Carolina—Wilmington, Morph Academic Dataset*. Accessed: Feb. 25, 2020. [Online]. Available: https://uncw.edu/oic/tech/morph.html
[45] U.S. Food and Drug Administration. *Your Skin*. Accessed: Feb. 25, 2020. [Online]. Available: https://www.fda.gov/radiation-emitting-products/tanning/your-skin
[46] M. Wall, *Biased and Wrong? Facial Recognition Tech in the Dock*, BBC News, London, U.K., Jul. 2019. Accessed: Feb. 25, 2020. [Online]. Available: https://www.bbc.com/news/business-48842750
[47] M. Wang, W. Deng, J. Hu, X. Tao, and Y. Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, 2019.
[48] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Process. Lett.*, vol. 23, no. 10, pp. 1499–1503, Oct. 2016.

K. S. Krishnapriya received the B.Tech. degree in computer science and engineering from the SCMS School of Engineering and Technology, Kochi, India, in 2014, and the M.Tech. degree in computer science and engineering from the Rajiv Gandhi Institute of Technology, Mumbai, India, in 2016. She is currently pursuing the Ph.D. degree in computer science with the Florida Institute of Technology, Melbourne, FL, USA.

She worked as an Assistant Professor with the Sahrdaya College of Engineering and Technology, Kodakara, India, for one and a half years. Her research interests include biometrics, pattern recognition, and machine learning.

Vítor Albiero (Student Member, IEEE) received the B.S. degree in computer science from the University of West Santa Catarina, Joaçaba, Brazil, in 2015, and the M.S. degree in computer science from the Federal University of Paraná, Curitiba, Brazil, in 2018. He is currently pursuing the Ph.D. degree in computer science and engineering with the University of Notre Dame, Notre Dame, IN, USA.

His research focus is in understanding the demographic aspects of accuracy of deep-learning-based face recognition.

Kushal Vangara received the B.Tech. degree in electronics and communications engineering from Jawaharlal Nehru Technological University, Hyderabad, India, in 2014, and the M.S. degree in information assurance and cybersecurity from the Florida Institute of Technology, Melbourne, FL, USA, in 2018, where he is currently pursuing the Ph.D. degree in computer science.

His research interests include biometrics, computer vision, data science, and deep learning.

Michael C. King (Member, IEEE) received the Ph.D. degree in electrical engineering from North Carolina Agricultural and Technical State University, Greensboro, NC, USA.

He joined the Harris Institute for Assured Information, Florida Institute of Technology's, Melbourne, FL, USA, as a Research Scientist in 2015 and holds a joint appointment as an Associate Professor of computer engineering and sciences. Prior to joining academia, he served for more than ten years as a Scientific Research/Program Management Professional with the United States Intelligence Community. While in government, he created, directed, and managed research portfolios covering a broad range of topics related to biometrics and identity to include: advanced exploitation algorithm development, advanced sensors and acquisition systems, and computational imaging. He crafted and led the Intelligence Advanced Research Projects Activity's Biometric Exploitation Science and Technology Program to transition technology deliverables successfully to several Government organizations.

Kevin W. Bowyer (Fellow, IEEE) received the Ph.D. degree in computer science from Duke University, Durham, NC, USA.

He is the Schubmehl-Prein Family Professor of computer science and engineering with the University of Notre Dame, Notre Dame, IN, USA, and also serves as the Director of the International Summer Engineering Programs, Notre Dame College of Engineering, Notre Dame.

Prof. Bowyer received the Technical Achievement Award from the IEEE Computer Society, with the citation "for pioneering contributions to the science and engineering of biometrics." He served as the Editor-in-Chief for the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE. He is currently serving as the Editor-in-Chief for the IEEE TRANSACTIONS ON BIOMETRICS, BEHAVIOR AND IDENTITY SCIENCE. In 2019, he was elected as a fellow of the American Association for the Advancement of Science and IAPR.