Impact of Blur and Resolution on Demographic Disparities in 1-to-Many Facial Identification

Aman Bhatta[1]   Gabriella Pangelinan[2]   Michael C. King[2]   Kevin W. Bowyer[1]

University of Notre Dame[1]
Florida Institute of Technology[2]

{abhatta,kwb}@nd.edu, gpangelinan@my.fit.edu, michaelking@fit.edu

Abstract

Most studies to date that have examined demographic variations in face recognition accuracy have analyzed 1-to-1 matching accuracy, using images that could be described as "government ID quality". This paper analyzes the accuracy of 1-to-many facial identification across demographic groups, and in the presence of blur and reduced resolution in the probe image as might occur in "surveillance camera quality" images. Cumulative match characteristic curves (CMC) are not appropriate for comparing propensity for rank-one recognition errors across demographics, and so we use three metrics for our analysis: (1) the well-known d' metric between mated and non-mated score distributions, and introduced in this work, (2) absolute score difference between thresholds in the high-similarity tail of the non-mated and the low-similarity tail of the mated distribution, and (3) distribution of (mated – non-mated rank-one scores) across the set of probe images. We find that demographic variation in 1-to-many accuracy does not entirely follow what has been observed in 1-to-1 matching accuracy. Also, different from 1-to-1 accuracy, demographic comparison of 1-to-many accuracy can be affected by different numbers of identities and images across demographics. More importantly, we show that increased blur in the probe image, or reduced resolution of the face in the probe image, can significantly increase the false positive identification rate. And we show that the demographic variation in these high blur or low resolution conditions is much larger for male / female than for African-American / Caucasian. The point that 1-to-many accuracy can potentially collapse in the context of processing "surveillance camera quality" probe images against a "government ID quality" gallery is an important one.
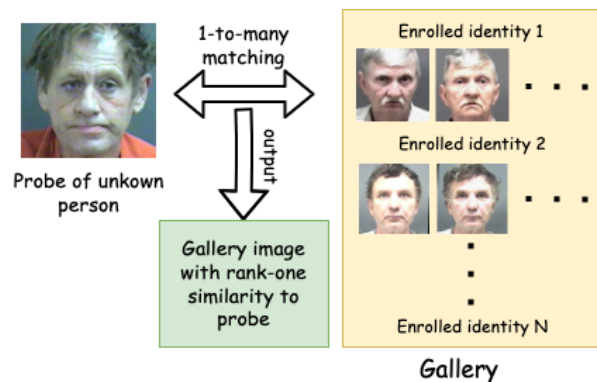
Figure 1. Does 1-to-many rank-one match error rate vary across demographics? In 1-to-many identification, an image of a person with unknown identity (the probe) is matched against a list of persons with known identity (the gallery) to find a candidate identity.

1. Introduction

Many studies to characterize demographic disparities in the accuracy of face recognition systems employed techniques designed to evaluate 1-to-1 match accuracy in the context of verification or authentication tasks, with only a few focusing on open-set 1-to-many identification scenarios. By all accounts, 1-to-many facial identification has been shown to be highly accurate in the case where the probe and gallery images are "all government ID quality" and the person in the probe image has one or more enrolled images [20, 24].

In this work, we report experimental results investigating (a) demographic accuracy variation in 1-to-many face matching, and (b) impact of varying probe image quality on accuracy of 1-to-many face matching.

In 1-to-many matching, an image of an unknown person (a "probe") is matched against enrolled images of many known persons (the "gallery") in order to find a candidate identity for the unknown person. In this paper, we focus on the use of 1-to-many matching to return the gallery im-

age with the greatest similarity to the probe, the "rank-one match." More generally, 1-to-many matching can also be used to return the top N matches, and / or to return matches above some threshold of similarity. Such use-cases typically assume a human examiner is involved to decide among the candidate results, and experiments incorporating review of algorithm-selected candidate images by a human examiner are beyond the scope of this paper.

Contributions of this work include the following:

- Using ArcFace, MagFace and AdaFace, we compare 1-to-1 verification accuracy and 1-to-many identification accuracy across female / male and African-American / Caucasian. We find that the ranking of demographics for likelihood of rank-one false positive identification is not the same as the ranking for false match rate in 1-to-1 verification. This is due to 1-to-1 matching being based on a fixed threshold on similarity, whereas rank-one false positive identification is based on the difference between a probe's rank-one mated and non-mated scores.

- We introduce two metrics to supplement the well known d' between mated and non-mated score distributions to estimate the propensity for 1-to-many matching to produce false positive identifications. The introduced metrics are: (1) absolute score difference between thresholds in the high-similarity tail of the non-mated and the low-similarity tail of the mated distribution, and (2) distribution of (mated – non-mated rank-one scores) across the set of probe images. These metrics allow an assessment of how close a distribution of scores is to generating false positive identifications even if no specific false positive identification is observed. The relative ranking of demographics is the same for the three metrics, but our analysis suggests that using d' alone could miss subtle differences.

- We create increasingly blurred versions of the original probe images in order to assess how blur in the probe image impacts false positive identification rate. We also create lower-resolution versions of the original probe images in order to assess how lower resolution in the probe image impacts false positive identification. For both increasing blur and lower resolution, there is a point at which one-to–many matching "breaks" and the false positive identification rate increases dramatically. This suggests that 1-to-many search accuracy cannot be objectively evaluated without the context of (at least) the level of blur in the image and the resolution of the face.

2. Related Work

For a broad overview of issues related to demographic bias in biometrics, see the survey by Drozdowski et al. [16]. We briefly touch on related works in two areas: (1) demographic differences in face recognition accuracy and (2) 1-to-many matching for face recognition.

The earliest observation that accuracy varies across demographic groups may be the 2002 Face Recognition Vendor Test [26], which found lower accuracy for females than for males. The well-known study by Klare et al. [23] analyzed results from multiple algorithms and found that, "The female, Black, and younger cohorts are more difficult to recognize for all matchers used in this study (commercial, nontrainable, and trainable)". It is interesting that [23] found that *training on demographic-balanced training data did not balance accuracy across demographics*. These papers [23, 26] were before the rise of deep-CNN-based face matching. One recent study involving deep CNN matchers opposes the result that younger cohorts are more difficult to recognize [6]. Another recent study agrees that the demographic balance of the training data does not lead to balanced accuracy across demographics [9], meanwhile the study by Wu et al. [32] proposes a bias-aware cross-demographic validation methodology for bias factor investigation. Furthermore, a recent study demonstrates demographic disparities in accuracy at the facial detail attribute level too [31]. The consensus across recent studies is that females, compared to males, have a worse impostor distribution and a worse genuine distribution and that African-American males have a worse impostor distribution compared to Caucasian males, but also a better genuine distribution [8, 10, 12, 24, 29, 30]. All of the above studies are in the context of 1-to-1 matching.

Grother and colleagues have released NIST reports on 1-to-many facial identification [20] and on demographic effects [19, 21]. The report on 1-to-many identification [20] analyzes results from a number of matchers and large image sets. They note the high 1-to-many rank-one accuracy of current matchers, the fact that accuracy decreases linearly with the number of enrolled identities, and that enrolling multiple images for an identity increases accuracy. The report on demographic effects [21] notes results similar to those shown later in this paper for African-American / Caucasian, and for female / male. They also discuss the specialized forms of 1-to-many matching that are designed to produce a list of candidates for review by a human analyst. The most recent NIST report [19] summarizes demographic comparisons across groups from various locations around the globe, and notes that demographic differences in false positive identification rate (FPIR) are larger than in false negative identification rate (FNIR), and that universal generalizations are difficult. They also note that some commerical matchers do not implement 1-to-many matching in
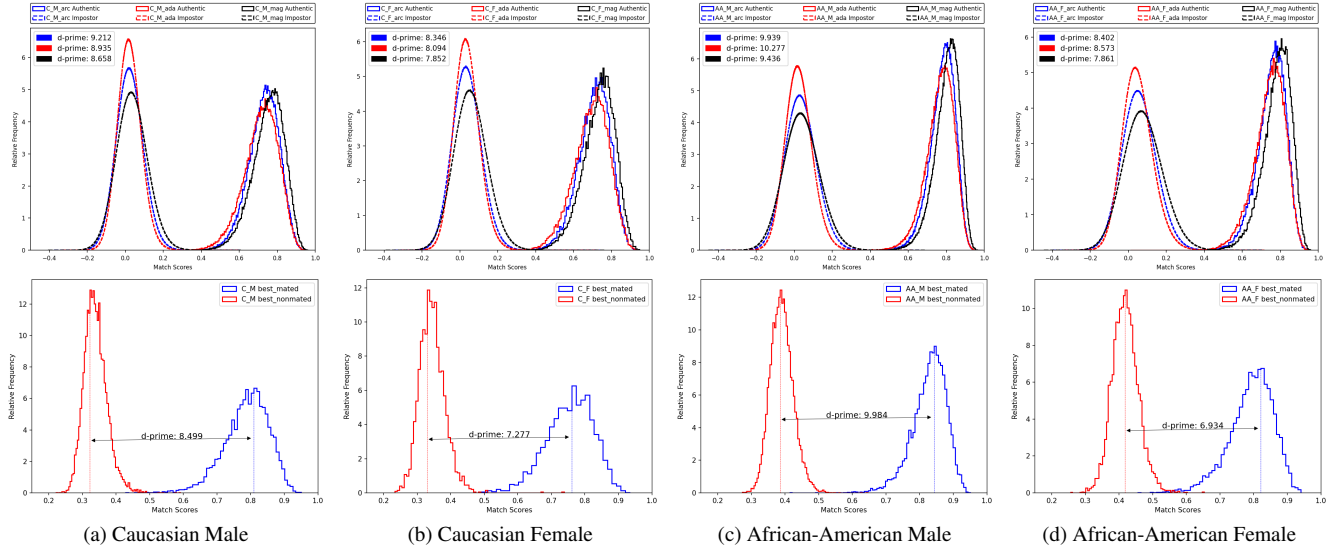
Figure 2. Baseline 1-to-1 (verification) and 1-to-many (identification) distributions. Top row shows 1-to-1 impostor and genuine distributions; d-prime is greatest for African-American Male and lowest for Caucasian Female. Bottom row shows 1-to-many mated and non-mated distributions for ArcFace (results for AdaFace and MagFace in Table 1) ; d-prime is greatest (lowest false positive identification rate) for African-American Male and lowest (highest false positive identification rate) for African-American Female. In the legend of figures in the top row, the labels "arc," "ada," and "mag" correspond to the results obtained using the ArcFace, AdaFace, and MagFace loss functions, respectively.

the straightforward manner [19], and so comparisons involving commercial matchers can be difficult. Krishnapriya et al [24] also look at 1-to-many matching accuracy, using the MORPH dataset, and also report that when the person in the probe image does have an enrolled image, the 1-to-many search is highly likely to return the correct identity. Both [21] and [24] emphasize that a 1-to-many search when the person in the probe image does not have an enrolled image can by definition only return a false-positive identification (FPI). Recent work by Drozdowski et al. [15] examines the "watchlist imbalance effect" that occurs in 1-to-many matching when different demographics represent different fractions of the watchlist. They compare theoretical and empirical estimates of FPIR across demographics for balanced and unbalanced watchlists. They also demonstrate that equitable 1-to-1 matching doesn't necessarily ensure equitable 1-to-many matching.

This paper goes beyond the works above in the following respects. While the NIST reports use datasets of different quality images - mugshots, webcam and in-the-wild - they do not systematically examine quality dimensions associated with surveillance-camera images. We perform systematic experiments examining the degree of blur and the degree of reduced resolution in the probe image, and how these factors impact 1-to-many accuracy. Blur and resolution are two important elements in surveillance-camera quality images. Our experiments use image datasets and matchers that in principle are available to the research community (some government datasets are not [19, 21]), enabling reproducibility by other researchers.

3. Dataset and Matcher

MORPH [4, 27] may be the best-known dataset in face aging research and is also widely used in studying demographic accuracy variation [5, 6, 8, 15, 17, 24, 29]. Drozdowski et al. [15] note that MORPH is particularly appropriate for demographic studies "... due to its large size, relatively constrained image acquisition conditions, and the presence of ground-truth labels (from public records) for sex, race, and age of the subjects". We use the same version of MORPH used in [10], which has 35,276 images of 8,835 Caucasian males, 10,941 images of 2,798 Caucasian females, 56,245 images of 8,839 African-American males, and 24,855 images of 5,928 African-American females. The average number of images per identity is about 4 for Caucasian male and Caucasian female, 4.2 for African-American female and 6.4 for African-American male.

We use three open-source matchers: 1) combined margin model based on ArcFace loss [14] trained on Glint-360K(R100) [11] with weights at [2], 2) MagFace [25] trained on MS1MV2(R100) with weights at [3], and c) AdaFace [22] trained on MS1MV2(R100) with weights at [1]. The input to each network is an aligned face resized to 112x112, and the output is a 512-d feature vector that is matched using cosine similarity. Faces are detected and aligned using img2pose [7].

Group	ArcFace			AdaFace			MagFace		
	1-to-1	1-to-many original	1-to-many balanced	1-to-1	1-to-many original	1-to-many balanced	1-to-1	1-to-many original	1-to-many balanced
AA M	9.94	9.98	9.55	10.27	9.54	8.98	9.44	9.23	9.10
C M	9.21	8.50	8.43	8.93	8.12	8.07	8.66	7.99	7.91
C F	8.35	7.28	6.84	8.09	6.84	6.40	7.85	6.54	6.17
AA F	8.40	6.93	6.82	8.57	6.44	6.41	7.86	6.07	5.95

Table 1. d' Comparisons Across Demographics. d' for 1-to-1 impostor and genuine, 1-to-many non-mated and mated with original unbalanced dataset, and 1-to-many non-mated and mated with demographics balanced on number of identities and images.

4. 1-to-1 Matching Accuracy Across Demographics

The top row of Figure 2 shows the 1-to-1 impostor and genuine distributions. The impostor contains similarity scores for all pairs of images representing different identities. The genuine contains similarity scores for all pairs of images representing the same identity. We also give the d' for separation of the impostor and genuine distributions [13]:

$$d' = \frac{|\mu_G - \mu_I|}{\sqrt{\frac{\sigma_G^2 + \sigma_I^2}{2}}} \qquad (1)$$

where μ_G and σ_G are the mean and standard deviation for the genuine distribution, and μ_I and σ_I for the impostor distribution. Larger d' means greater separation between impostor and genuine distributions, implying greater matching accuracy. (Using d' assumes that the distributions are reasonably approximated as Gaussian.)

Figure 2 shows that African-American male has the largest d', followed by Caucasian male, African-American female, and Caucasian female. However, the difference in d' for African-American female and Caucasian female is small. This pattern of d' implies that if 1-to-1 matching was performed separately for each demographic, the lowest false non-match rate (FNMR) at a fixed false match rate (FMR) could be achieved for African-American male, followed by Caucasian male, African-American female, and Caucasian female. However, in the typical operational scenario, the same threshold on similarity score is used to classify all image pairs, regardless of demographics. Using a fixed threshold results in the different demographics having different FMRs, with Caucasian male having the lowest FMR, followed by African-American male, Caucasian female and African-American female.

5. 1-to-Many Accuracy Across Demographics in "government ID" quality setting

For our analyses of 1-to-many matching, each identity's most recent image is designated as its probe image and the older images as its enrolled images for the gallery. A few MORPH identities have just one image, and so have a probe but no enrolled images. Given the high rank-one

matching accuracy of modern matchers [20, 21, 24], the size of MORPH, and the quality of images in MORPH, experiments will not generate large numbers of rank-one false positive identifications. However, we can make useful comparisons across demographics in at least three ways: (1) the d' separation of rank-one non-mated and mated score distributions, (2) separation in the FPIR-centric tails of the non-mated and mated distributions, and (3) the distribution of (mated – non-mated) score differences. Because of the straightforward analogy to 1-to-1 impostor and genuine distributions, we begin with rank-one non-mated and mated score distributions.

The 1-to-many mated distribution contains one score for each identity that has one or more enrolled images. That score is the maximum similarity score of the probe to any enrolled image of the same identity. The 1-to-many non-mated distribution also contains one score for each probe, and that score is the maximum similarity of the probe to any enrolled image of a different identity.

The 1-to-many non-mated and mated distributions are shown in the bottom row of Figure 2. The 1-to-many non-mated and mated distributions are centered at higher similarity than the 1-to-1 impostor and genuine distributions. This is because the 1-to-many score for a probe is the maximum similarity score across a set of images, whereas in 1-to-1 impostor and genuine distributions include similarity scores for all image pairs.

The d' for the 1-to-1 matching distributions and 1-to-many matching distributions in Figure 2 are summarized in Table 1, columns "1-to-1" and "1-to-many original". As with the 1-to-1 distributions, African-American male has the largest d' for the 1-to-many distributions, followed by Caucasian male. However, the order of the d' values flips for Caucasian female and African-American female, so that African-American female has the lowest d' in the 1-to-many matching results. This is an example of how demographic differences in 1-to-many matching do not necessarily repeat differences observed in 1-to-1 matching.

Results in Table 1 suggest that, other factors equal, African-American males should have the lowest false positive identification rate (FPIR). However, there are caveats. One is that the number of enrolled identities and the number of images per identity vary across demographics. Another
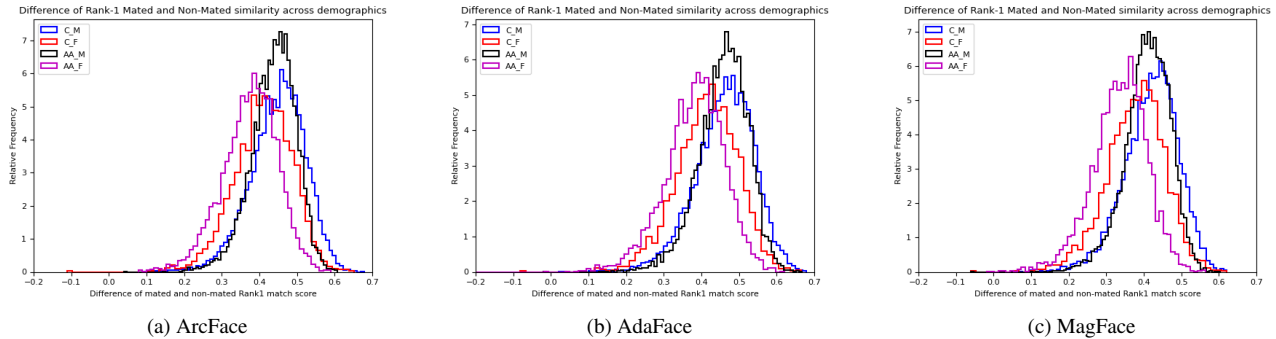
Figure 3. (mated - non-mated) distributions by demographic for three face matchers. The X-axis represents the "Relative Frequency", while the Y-axis represents the "Difference of mated and non-mated Rank1 match score".

is that d' may not adequately reflect variations in skew in the tails of the distributions. We address these two concerns in later sections. Another concern is that operational scenarios are typically "open-set" in the sense that there is no knowledge of whether the identity in the probe image has any image in the gallery. Every 1-to-many search using a probe that has no image in the gallery can only result in a FPI, as pointed out in [20, 24]. For this reason, a general conclusion about the FPIR across demographics in an operational scenario requires knowledge of the rate of open-set searches across demographics.

5.1. Comparison With Equal Identities and Images

The non-mated and mated distributions can be affected by the number of identities enrolled and the number of images per identity [18, 20]. Therefore, a comparison focused on difference due to demographic alone should have the same number of identities and images across the demographics. In our dataset, Caucasian female has the least number of identities, 2,797. So we randomly select 2,797 identities from each of the other demographics. The most recent image of each identity is again used as the probe. To equalize the number of enrolled images, the next most recent image is selected as the one enrolled image per identity. This results in each of the demographics having 2,797 probe images and 2,797 gallery images. Further, to check that the age difference between probe and enrolled image for an identity is approximately balanced across demographics, we checked the distribution of time between mated images.

The d' values for the non-mated and mated distributions based on the balanced image sets are given in Table 1 ("1-to-many balanced" column). The d' order across demographics is the same as for the original unbalanced datasets. However, the differences between the demographics are smaller for the balanced dataset.

Group	ArcFace		AdaFace		MagFace	
	Δ original	Δ balanced	Δ original	Δ balanced	Δ original	Δ balanced
AA M	0.073	0.060	0.056	0.048	0.034	0.029
C M	0.010	0.059	-0.027	0.021	-0.023	-0.048
C F	-0.005	0.018	-0.038	-0.068	-0.060	-0.123
AA F	-0.059	-0.044	-0.096	-0.058	-0.101	-0.083

Table 2. Non-parametric "Recognition Power" Across Demographics. Δ is the distance between the 1-in-1000 threshold in the high-similarity tail of the non-mated distribution and the low-similarity tail of the mated. Balanced indicates that the number of identities and number of enrolled images per identity are the same across demographic groups.

5.2. Metric for Tails of Non-mated and Mated

As mentioned earlier, using d' assumes the distributions are reasonably approximated as Gaussian. If the high-similarity tail of the non-mated and/or the low-similarity tail of the mated are skewed differently enough across demographics, comparison based on d' could give a misleading impression of relative FPIR. To address this, we create a metric that reflects the separation between the tails of the distributions without assuming the distributions are Gaussian. We compute the distance between the similarity values representing the lower one-thousandth of the mated distribution and the upper one-thousandth of the non-mated distribution. A higher value of this metric indicates greater separation in the relevant tails of the distributions.

Values of this metric are shown in Table 2, computed both for the original dataset with varying numbers of identities and images across demographics, and for the balanced dataset. Note that the order of the demographics is the same for unbalanced and balanced datasets. However, the difference between African-American male and Caucasian male is minimal for the balanced dataset. This again illustrates the need to make a more detailed analysis than d' alone.

(a) $\sigma = $ none   (b) $\sigma = 1$ (c) $\sigma = 2$ (d) $\sigma = 3$ (e) $\sigma = 4$ (f) $\sigma = 5$

Figure 4. Sample probe images with increasing Gaussian blur levels.

(a) Caucasian Male   (b) Caucasian Female   (c) African-American Male   (d) African-American Female
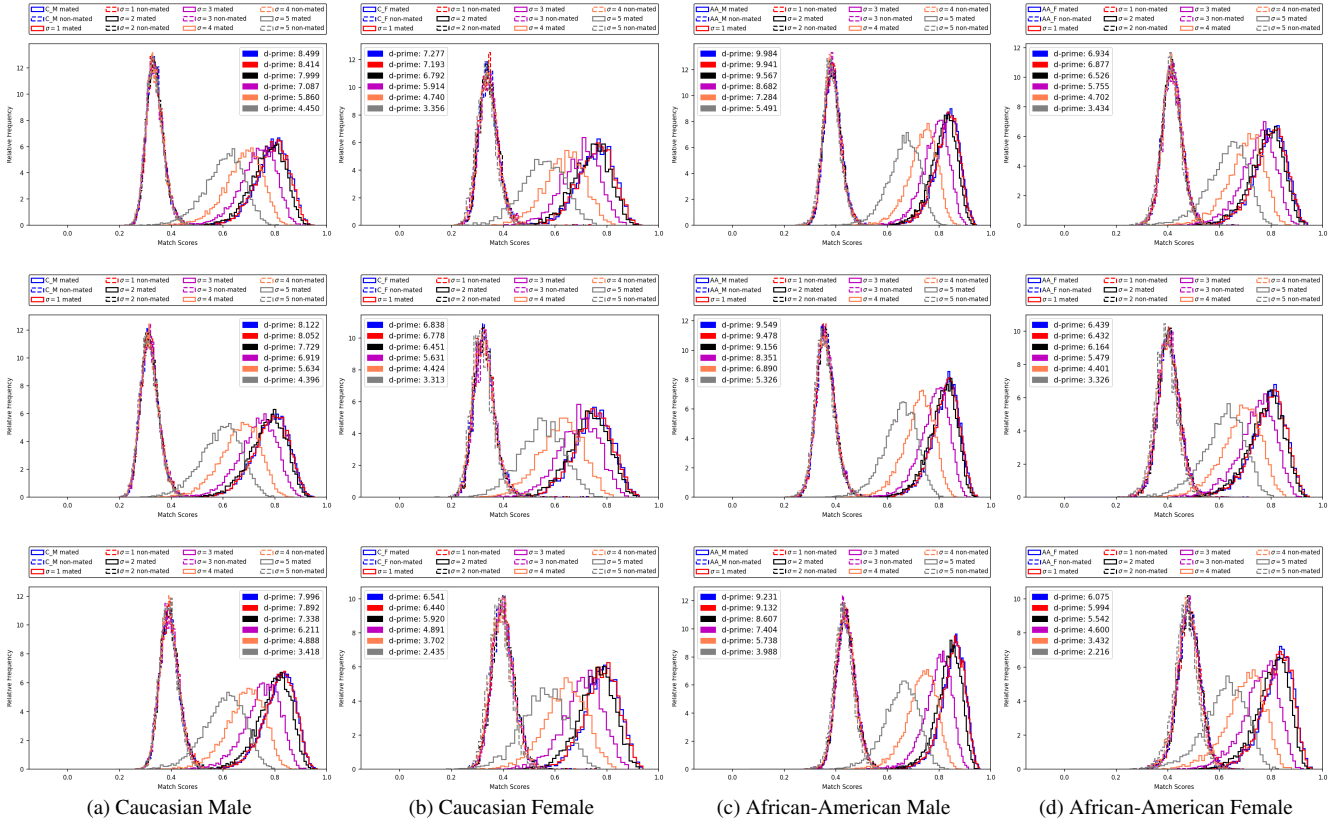
Figure 5. Impact of probe image blur on one-to-many matching accuracy. The top row shows the results for ArcFace, the middle row for AdaFace, and the bottom row for MagFace. The impact of blurred probe images remains consistent across the matchers. In all plots, the X-axis represents "Relative Frequency," while the Y-axis represents "Match Scores."

5.3. Distribution of (Mated - Non-mated) Difference

For a given probe image, a negative (mated - non-mated) score difference represents an FPI, and larger positive differences represent further distance from possible FPI. In this sense, the distribution of (mated - non-mated) score differences is a more direct indication of potential for FPI. The distributions of (mated - non-mated) differences for the balanced datasets are shown in Figure 3. Distributions for the male demographics are relatively similar. Distributions for the female demographics are also relatively similar. The larger demographic difference is between male / female, and the difference between African-American / Caucasian

of the same gender is smaller.

6. 1-to-Many Accuracy Across Demographics in "surveillance" setting

Results in earlier sections are based on both the probe image and the enrolled images both being "government ID quality". However, one important scenario for the use of 1-to-many facial identification involves a "surveillance camera" probe image matched against government ID-quality enrolled images. Probe images derived or extracted from surveillance media may be lower quality, with respect to image blur, resolution of the face, and other factors. In this
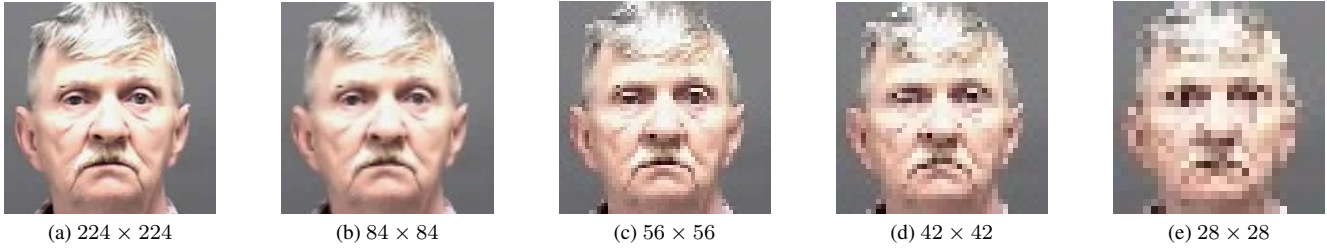
(a) 224 × 224 (b) 84 × 84 (c) 56 × 56 (d) 42 × 42 (e) 28 × 28

Figure 6. Sample probe images with increasing downscaling levels.

(a) African-American Male (b) African-American Female (c) African-American Female (d) African-American Female
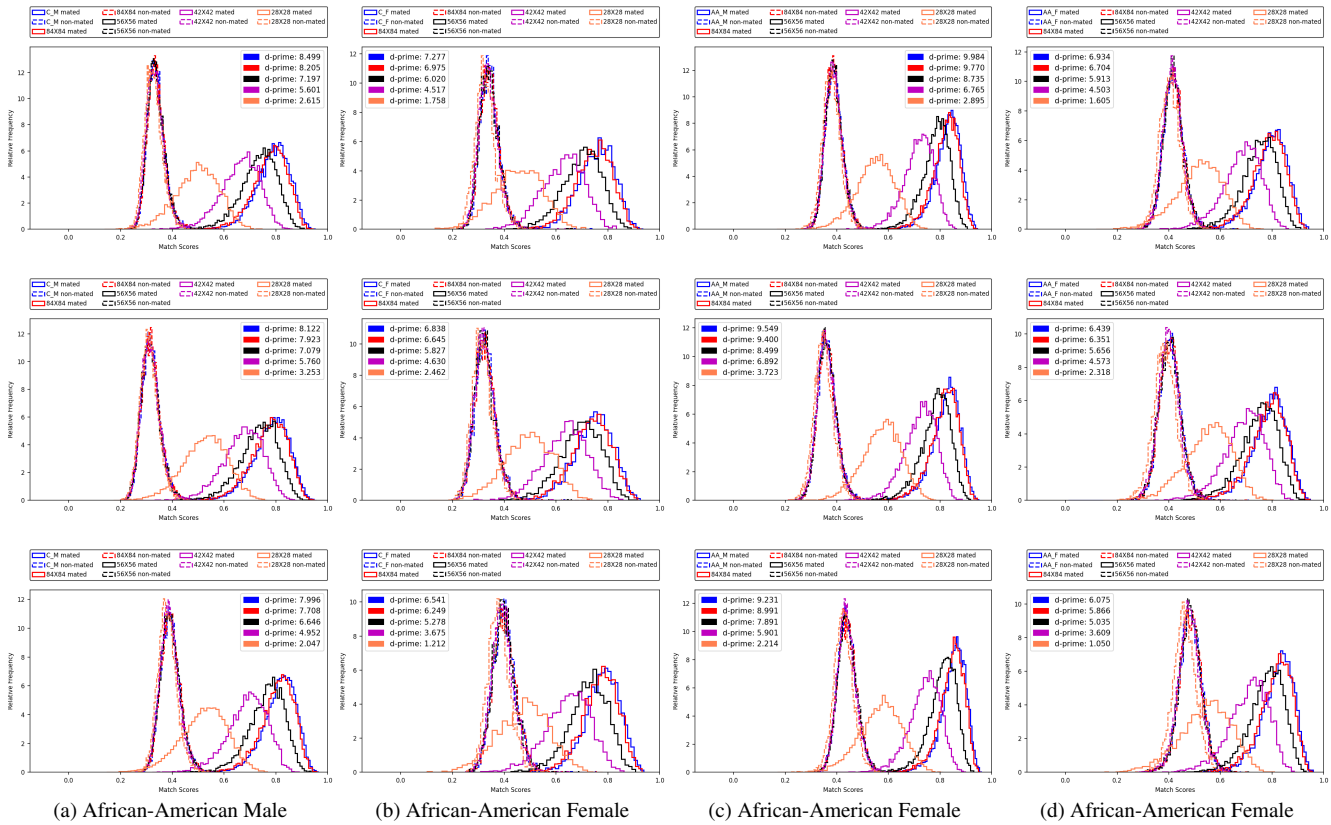
Figure 7. Impact of lower-resolution probe on one-to-many matching accuracy. The top row shows the results for ArcFace, the middle row for AdaFace, and the bottom row for MagFace. The impact of downsampled probe images remains consistent across the matchers. In all plots, the X-axis represents "Relative Frequency," while the Y-axis represents "Match Scores."

section we explore how *blur in the probe* image impacts 1-to-many accuracy, and subsequently, we explore the impact of the *probe having lower resolution*.

6.1. Impact of Blur in Probe Image

For this experiment, we used Gaussian smoothing of $\sigma = 1, 2, 3, 4, 5$ to create a sequence of increasingly blurred versions of the probe images (the same Gaussian blur settings as in [28]). Sample probe images are shown in Figure 4. We repeated the 1-to-many matching experiment done with the original probe images with each of the increasing blurred versions of the probe images. The resulting 1-to-many mated and nonmated distributions are shown in Fig-

ure 5. Increasing blur in the probe image has no large effect on the nonmated distribution, but causes the mated distribution to move to lower similarity. As an example, Figure 8a shows that, with ArcFace as the chosen face matcher, at a blur level of $\sigma = 5$, both the Caucasian female and African-American female demographics display false positive identification rates of over 1% for probe images. These findings imply that estimating the level of blur in probe images may be a crucial quality control measure for any 1-to-many search result.

The results consistently align across all three matchers utilized in this study — ArcFace, AdaFace, and Mag-Face. This finding is particularly intriguing given that the
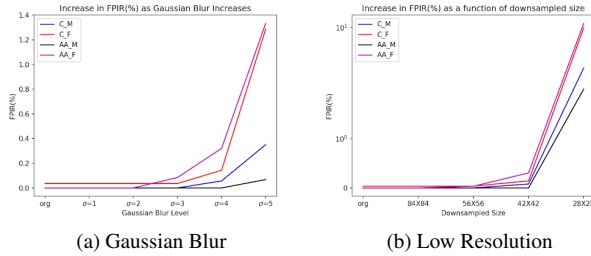
(a) Gaussian Blur   (b) Low Resolution

Figure 8. FPIR(%) for blurred (left) and lower resolution (right) probe images. The plots shown here are for ArcFace. In both plots, the Y-axis represents FPIR(%), while the X-axis represents the gaussian blur level for the figure on the left and image resolution for the figure on the right.

AdaFace and MagFace loss functions are intentionally designed to consider image quality during training, showing promise in 1-to-1 matching scenarios. However, it's worth noting that these promising results in 1-to-1 matching do not seamlessly extend to 1-to-many, as illustrated in Figure 5.

6.2. Impact of Lower-Resolution Probe Image

Faces cropped from frames of surveillance video may have lower spatial resolution than government ID face images. To explore the impact of having a lower-resolution probe face image, we used bicubic interpolation to create 84×84, 56×56, 42×42 and 28×28 versions of the original 224×224 probe images. Sample images of this sequence are shown in Figure 6.

The different resolution versions of the probe images are all converted to 112×112 for input to the CNN. The 1-to-many matching experiment is repeated for each resolution of the probe images, and the results are shown in Figure 7. The nonmated distribution is relatively unaffected by the resolution of the probe image before it is input to the network. But lower resolution in the probe image causes the mated distribution to shift to lower similarity. As an example, Figure 8b shows that, with ArcFace as the chosen face matcher, for 28×28 versions of probe images, the female demographics exhibit a false positive identification rate of more than 10%. It is important to acknowledge that the experimental setup used in this study provided the face embedding network with more prior information than what is commonly encountered in real-world scenarios. Specifically, the images utilized were pre-aligned and labeled, with the underlying assumption that the "face is certainly in the image." In practice, face detection algorithms may discard images that are too small or cannot be aligned, unless a human operator annotates the face image. Furthermore, in alignment with our observations from the blur experiment, as presented in Section 6.1, we find a consistent pattern in the results across all three matchers for this experiment as well.

7. Conclusions and Discussion

We focus on accuracy of 1-to-many matching in terms of the false positive identification rate for the rank-one match of a probe image against a set of gallery images that contains one or more other images of the same identity as the probe ("closed-set" matching). We do not consider schemes in which the 1-to-many search returns all matches above a similarity threshold or the top N matches. For a series of results based on both the probe image and the enrolled images being government ID-quality, we use three different approaches to assess the propensity for false positive identification in 1-to-many search. We show that the relative ranking across demographics for 1-to-many matching is different than the ranking for 1-to-1 matching. This is because 1-to-1 matching uses a fixed similarity threshold across demographics, while rank-one 1-to-many matching results in a false positive identification when the rank-one nonmated match is higher similarity than the rank-one mated match. We also point out that comparison of 1-to-many accuracy across demographics should be done with equal number of identities and images/identity across demographics. And we show that using only the d' for separation of mated and non-mated distributions may not be sufficient to appropriately compare accuracy across demographics.

We show that increasing blur in the probe image causes the mated distribution to shift to lower similarity and thereby increases the FPIR. Similarly, we show that lower spatial resolution in the probe image causes the mated distribution to shift to lower similarity and thereby increase the FPIR. The high accuracy of modern CNN-based matchers, when both probe and enrolled images are of government ID quality [21, 24], can degrade dramatically if the probe image exhibits substantial blur or low resolution on the face, as is often the case in real-world "surveillance" settings.

Image blur level and spatial resolution of the cropped face are quality factors that are readily estimated. Our results suggest that these factors should be considered when assessing the usefulness of a probe for 1-to-many search. Our analysis here, employing three matchers (ArcFace, AdaFace, and MagFace) and utilizing MORPH as a test dataset, is an example of the type of analysis that could be done for a particular operational scenario's matcher and face dataset in order to establish quality thresholds for probe images used in facial identification.

References

[1] Adaface: Quality adaptive margin for face recognition. https://github.com/mk-minchul/AdaFace. 3

[2] Insightface: 2d and 3d face analysis project. https://github.com/deepinsight/insightface/tree/master/model_zoo. 3

[3] Magface: A universal representation for face recognition and quality assessment. https://github.com/IrvingMeng/MagFace. 3

[4] Morph dataset. https://uncw.edu/oic/tech/morph.html. 3

[5] Vítor Albiero and Kevin W Bowyer. Is face recognition sexist? no, gendered hairstyles and biology are. In *Proc. British Mach. Vision Conf.*, 2020. 3

[6] Vítor Albiero, Kevin W. Bowyer, Kushal Vangara, and Michael C. King. Does face recognition accuracy get better with age? Deep face matchers say no. In *Winter Conf. on App. of Comput. Vision*, 2020. 2, 3

[7] Vítor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner. img2pose: Face alignment and detection via 6dof, face pose estimation. In *Compt. Vis. & Patt. Recog.* IEEE, 2021. 3

[8] Vitor Albiero, Krishnapriya KS, Kushal Vangara, Kai Zhang, Michael C King, and Kevin W Bowyer. Analysis of gender inequality in face recognition accuracy. In *Compt. Vis. & Patt. Recog. Workshops*, 2020. 2, 3

[9] Vítor Albiero, Kai Zhang, and Kevin W Bowyer. How does gender balance in training data affect face recognition accuracy? In *Int. Joint. Conf. on Biometrics*, 2020. 2

[10] Vítor Albiero, Kai Zhang, Michael C King, and Kevin W Bowyer. Gendered differences in face recognition accuracy explained by hairstyles, makeup, and facial morphology. In *Trans. on Inform. Forensics and Security*, 2021. 2, 3

[11] Xiang An, Xuhan Zhu, Yuan Gao, Yang Xiao, Yongle Zhao, Ziyong Feng, Lan Wu, Bin Qin, Ming Zhang, Debing Zhang, et al. Partial fc: Training 10 million identities on a single machine. In *Int. Conf. Comput. Vision*, pages 1445–1449. IEEE, 2021. 3

[12] Aman Bhatta, Vítor Albiero, Kevin W Bowyer, and Michael C King. The gender gap in face recognition accuracy is a hairy problem. In *Winter Conf. on App. of Comput. Vision Workshop*, 2023. 2

[13] John Daugman. How iris recognition works. In *The essential guide to image processing*, pages 715–739. Elsevier, 2009. 4

[14] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Compt. Vis. & Patt. Recog.* IEEE, 2019. 3

[15] Pawel Drozdowski, Christian Rathgeb, and Christoph Busch. The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In *Int. Conf. Comput. Vision Workshops*, pages 3750–3758, 2021. 3

[16] Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. Demographic bias in biometrics: A survey on an emerging challenge. In *Trans. Technology and Society*, 2020. 2

[17] Markos Georgopoulos, James Oldfield, Mihalis A. Nicolaou, Yannis Panagakis, and Maja Pantic. itigating demographic bias in facial datasets with style-based multi-attribute transfer. In *Int. J. Comput. Vision*, volume 129, page 2288–2307, 2021. 3

[18] Patrick Grother and P Jonathon Phillips. Models of large population recognition performance. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004. 5

[19] Patrick J. Grother. Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials. Technical Report 8429, NIST, July 2022. 2, 3

[20] Patrick J. Grother, Mei Ngan, and Kayee Hanaoka. Face recognition vendor test (FRVT) part 2: Identification. Technical Report 8271, NIST, Sep. 2019. 1, 2, 4, 5

[21] Patrick J. Grother, Mei Ngan, and Kayee Hanaoka. Ongoing face recognition vendor test (FRVT) part 3: Demographic effects. Technical Report 8280, NIST, Dec. 2019. 2, 3, 4, 8

[22] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Compt. Vis. & Patt. Recog.*, pages 18750–18759. IEEE, 2022. 3

[23] Brendan F. Klare, Mark J. Burge, Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain. Face recognition performance: Role of demographic information. In *Trans. on Inform. Forensics and Security*, volume 7, pages 1789–1801, 2012. 2

[24] K.S. Krishnapriya, V. Albiero, K. Vangara, M.C. King, and K.W. Bowyer. Issues related to face recognition accuracy varying based on race and skin tone. In *Trans. Technology and Society*, 2020. 1, 2, 3, 4, 5, 8

[25] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Compt. Vis. & Patt. Recog.*, pages 14225–14234. IEEE, 2021. 3

[26] P. Jonathon Phillips, Patrick J. Grother, Ross J. Michaels, Duane M. Blackburn, Elham Tabassi, and J.M. Bone. Face Recognition Vendor Test 2002: Evaluation Report. Technical Report 6965, NIST, 2003. 2

[27] Karl Ricanek and Tamirat Tesafaye. MORPH: A longitudinal image database of normal adult age-progression. In *Automatic Face and Gesture Recognition*, 2006. 3

[28] Wes Robbins and Terrance E Boult. On the effect of atmospheric turbulence in the feature space of deep face recognition. In *Compt. Vis. & Patt. Recog. Workshops*, pages 1618–1626, 2022. 7

[29] Kushal Vangara, Michael C King, Vitor Albiero, Kevin Bowyer, et al. Characterizing the variability in face recognition accuracy relative to race. In *Compt. Vis. & Patt. Recog. Workshops*, 2019. 2, 3

[30] Haiyu Wu, Vítor Albiero, KS Krishnapriya, Michael C King, and Kevin W Bowyer. Face recognition accuracy across demographics: Shining a light into the problem. In *Compt. Vis. & Patt. Recog. Workshops*, 2023. 2

[31] Haiyu Wu, Grace Bezold, Aman Bhatta, and Kevin W Bowyer. Logical consistency and greater descriptive power for facial hair attribute learning. In *Compt. Vis. & Patt. Recog.*, pages 8588–8597, 2023. 2

[32] Haiyu Wu and Kevin W Bowyer. A real balanced dataset for understanding bias? factors that impact accuracy, not numbers of identities and images. In *Proc. British Mach. Vision Conf.*, 2023. 2