Appeared in 26th International Conference on Pattern Recognition (ICPR 2022), Fairness in Biometrics Workshop, Montreal, Quebec, August 2022.

Disparate Impact in Facial Recognition Stems from the Broad Homogeneity Effect: A Case Study and Method to Resolve

John J. Howard*[1], Eli J. Laird*†[1], and Yevgeniy B. Sirotin*[1]

The Identity and Data Sciences Lab at The Maryland Test Facility, Maryland, USA
{elaird, jhoward, ysirotin}@idslabs.org

Abstract. Automated face recognition algorithms generate encodings of face images that are compared to other encodings to compute a similarity score between the two originating face images. These face encodings, also known as feature vectors, contain representations of various facial features. Some of these facial features, but not all, have been shown to resemble each other across different subjects that happen to share a demographic group assignment, such as having the same race or gender. Recent work has shown that these demographically dependent features can increase similarity scores between different individuals who belong to the same demographic group compared to similarity scores for different individuals in different groups. When one feature vector is compared to many other feature vectors, as in identifications, this effect, referred to as "demographic clustering", can lead to un-equal false positive identification error rates for different demographic groups. In this study, we propose a method of mitigating this clustering effect from face recognition algorithms to reduce these un-equal error outcomes. Our method presumes that feature space patterns shared within demographic groups can be removed while preserving other distinct features of individuals. In this paper, we prove that this is possible, in principle, by applying linear dimensionality techniques to the feature space of two ArcFace face recognition algorithms. We show this method increases four distinct "fairness" measures while preserving useful true match rates.

Keywords: Face Recognition · Demographic Differentials · Disparate Impact · Fairness.

1 Introduction

In the 2010s, face recognition algorithms significantly improved in accuracy due to advances in deep learning methods in computer vision. Specifically, the introduction of deep convolutional neural networks (DCNNs) to the face recognition task achieved near human performance for the first time in 2014, with an accuracy of 97.35% on the Labeled Faces in the Wild (LFW) dataset [18, 26]. The following year, a modified DCNN architecture achieved "better-than-human" performance on the same task (accuracy of 99.63%) [25]. By 2020, government tests

[*] All authors contributed equally to this research. Authors listed alphabetically.
[*†] Corresponding author.

of face recognition algorithm performance documented false positive outcomes occurring on 3 out of every 1000 searches, and false negative rates nearing 1 in 1000, when searching galleries of up to of 12 million individuals [10].

These impressive error rates on large galleries may lead advocates of face recognition technology to claim that face recognition is a solved problem. While there are many means to dispute such claims (ageing, gallery size, pose, etc.) [10], one aspect of face recognition regularly receives far less attention than these well-studied problems. Face recognition algorithms have been shown to routinely judge that different individuals, who happen to share gender, race, age, and country of origin designations are more similar than individuals who don't share these categories [17] [11]. This group similarity effect has been given different names, first "broad homogeneity" in [15], then "demographically matched individuals" in [11], and "imposter pairs across homogeneous and heterogeneous categories" in [8] (we will use the first term in this manuscript).

Regardless of nomenclature, broad homogeneity was shown in Annex 5 of [11] to exist in *all* of the 138 facial recognition algorithms submitted as part of this global face recognition evaluation in 2019. Furthermore, there seems to be an acceptance, both in the research and commercial communities, that face recognition algorithms *should* behave in this manner, despite these effects being unique to face recognition and decidedly not present in other common biometric modalities, such as fingerprint and iris recognition. Of additional concern are the mathematics first highlighted in [17] and later in [5] that show, in the presence of broad homogeneity effects and imbalanced facial recognition identification galleries, a strong tendency for un-equal identification error rates across demographic groups.

For these reasons, we contend broad homogeneity effects to be an undesirable, but unfortunately, currently universal, characteristic of face recognition algorithms. Additionally, methods to reduce this effect are presently under-researched. This manuscript presents one such method that removes demographically clustered components of the facial biometric feature vector (also known as the face template) that cause broad homogeneity effects. We demonstrate the utility of this approach on two disjoint datasets of test subjects who self reported their gender and race affiliations. We further show that, after applying this technique, each of four currently proposed facial recognition fairness metrics shows an improvement.

2 Background

2.1 Broad versus Specific Homogeneity Fairness Criteria

A face verification operation involves a one-to-one comparison of two face images. Images are first converted to face feature vectors within a p-dimensional feature space. Two face feature vectors can then be mathematically compared to compute a similarity score that represents the similarity between the two originating face images. If the resulting score is greater than some threshold τ, the algorithm is indicating that the original face images are from the same person. A false match error occurs when an algorithm produces a score greater than τ for two face images that, in fact, were from different people (also known as a non-mated pair). False match rate (FMR) is the frequency with which a false match error occurs in all possible non-mated image pairs within a given evaluation dataset.

When considering the "fairness" of facial recognition systems in regards to FMR, [17] introduced two separate criteria to consider. Briefly, the first, termed "specific homogeneity fairness criteria", stipulated that FMR measured within demographic group should be equal for each group, but that FMR measured between different groups could still take a different (presumably lower) value. For example, the false match rate when Black Males are compared to other Black Males would equal the false match rate when White Females were compared to other White Females ($\text{FMR}_{(\text{WM},\text{WM})} == \text{FMR}_{(\text{BM},\text{BM}))}$ but the false match rate between Black Males and White Males may be lower than the false match rate between Black Males ($\text{FMR}_{(\text{WM},\text{BM})} < \text{FMR}_{(\text{BM},\text{BM})}$). That a face recognition algorithm would operate in this way is intuitive to humans because human facial recognition processes behave in this way as well.

However, [17] also demonstrated (along with [5]) that, should specific homogeneity fairness criteria be the goal, disparities in face recognition identification (one-to-N) error rates could still persist, particularly in the presence of demographically imbalanced identification galleries. For this reason [17] advocates for a second criteria for assessing facial recognition systems in regards to FMR, the "broad homogeneity fairness criteria". This criteria states that the false match rate for cross demographic groups should equal the FMR of within demographic groups. Using our previous example, in this model $\text{FMR}_{(\text{WM},\text{WM})} == \text{FMR}_{(\text{BM},\text{BM})} == \text{FMR}_{(\text{WM},\text{BM})}$. This face recognition algorithm would operate in a way that is un-intuitive to humans, as it would confuse White Males for Black Males equally often as it would confuse White Males for other White Males. However, a face recognition algorithm that operated in this fashion may be able to achieve metrically more fair identification outcomes. Graphic descriptions of the specific and broad homogeneity fairness criteria are shown in Figure 1A and B, respectively.

2.2 Achieving Broad Homogeneity

As stated, demographic clustering effects were found to exist in every face identification algorithm tested in [11]. With such an ubiquitous effect, one might be inclined to think it a natural characteristic of face recognition in general. However, [17] showed that only a small portion of the information content available in a human face appears to be consonant across different people within gender and race categories. On five separate leading commercial face recognition algorithms, [17] found that just 10% of the variation in non-mated similarity score could be attributed to race and gender clustering. This suggests that if a face recognition algorithm ignored these clustering components, it may be able to achieve broad homogeneity while still maintaining useful levels of performance.

One limitation of [17] was that it measured these grouping effects only in similarity scores between individuals, i.e. the "score space". The researchers achieved

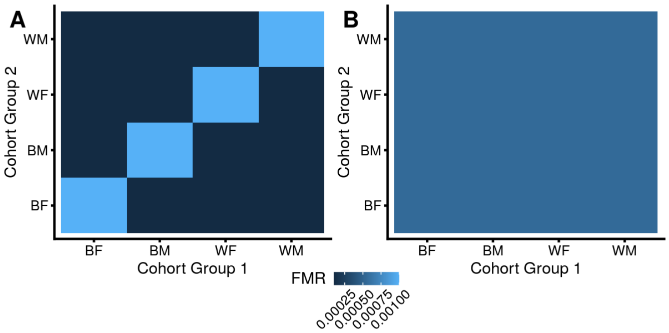


Fig. 1. Examples of cohort matrices that demonstrate fair false match rates (FMR) across demographic groups according to A) the specific homogeneity fairness condition and B) the broad homogeneity fairness condition. Example cohorts are Black Female (BF), Black Male (BM), White Female(WF), and White Male (WM), although the concept applies to categorical demographic groups generally.

.

this by performing eigenvalue decomposition on a matrix of cross subject similarity scores, as produced by each commercial algorithm. This technique is useful for measuring the magnitude of demographic clustering across algorithms. It could also potentially be useful for removing demographic clustering on a *static* population of identities, such as an access control scenario where every individual attempting to interact with a system is known and the population is relatively stable. However, removing demographic clustering in score space is not practical for algorithms meant to operate on *dynamic* populations, such as a system where new enrolles are frequent or where out-of-gallery or non-mate comparisons are frequent. This is because the specific identities in the database form the basis of score space and correction relies on removing patterns across these identities, not across face features. Furthermore, this correction necessitates establishing demographic group membership of each identity in the sample.

To adapt to the dynamic setting, one must develop a transformation that when computed on one set of subjects can be successfully applied to a disjoint set, i.e. a generalized transformation. Here, we will show that such a transformation can be computed using the p-dimensional feature vectors of face recognition algorithms. We will also show that transformations derived from this space can be applied to the embeddings generated from identities not in the original set and that when error rates and comparison scores generated from the original and transformed features are evaluated, fairness measures consistently improve.

3 Methods

3.1 Dataset

Three sets of images/subjects are used in this research. The first, referred to as ‘S1’, is a demographically balanced set. S1 contains one image per subject across 600 unique subjects (exactly 150 per demographic group). The second set, referred to as 'S2', is a disjoint set of subjects. No subject in S1 is present in S2. The S2 dataset contains one image per subject across 192 unique subjects (approximately 50 per demographic group, but in some cases less). The third set, referred to as 'S3' is a set of mated image pairs to subjects in S1. The purpose of S3 is to validate that our transforms (see Section 3.3) do not corrupt face templates to the point that matching transformed templates to other mated pairs is no longer possible. S3 is not intended to validate transforms that reduce demographic clustering. S3 contains between 1-6 images per subject across 466 unique subjects. Not all subjects in S1 had a corresponding mated image. All datasets were collected by a trained biometric collection operator, minimizing any issues related to image acquisition quality. All samples were collected at biometric scenario evaluations that took place from 2018-2021 [1] [13].

The disjoint property of S1 and S2 is a purposeful and important characteristic. Simply because two subjects identify into the same demographic group, White Male for example, does not signify that they necessarily share similar facial features. S1 and S2 may therefore have legitimately different patterns with respect to face features, despite having the exact same demographic groups.

Table 1. Number of subjects and samples for datasets used in this research.

Dataset	Subjects (Samples)			
	Black Female	**Black Male**	**White Female**	**White Male**
S1	150 (150)	150 (150)	150 (150)	150 (150)
S2	50 (50)	50 (50)	49 (49)	43 (43)
S3	106 (300)	117 (339)	126 (321)	117 (278)

3.2 ArcFace Face Recognition Algorithm

In 2019, Deng et al. [4] proposed and open-sourced a new face recognition loss function named ArcFace that reached the state-of-the-art verification accuracy of 99.83% on the LFW dataset. The ArcFace algorithm belongs to a family of 'margin-based' loss functions that apply margins to their logits to encourage class separability. ArcFace's predecessors, such as SphereFace [21] and CosFace [27], introduced this concept of penalizing class centers in the angular space using margins. ArcFace expanded on these techniques by introducing an additive angular margin loss that improves the compactness of intra-class samples and the separation of inter-class samples in the face embedding space.

A face recognition model trained with ArcFace loss was chosen for this study for three reasons. First, the techniques we outline here require "white-box" algorithms, where the feature space can be interpreted. In many commercial face recognition algorithms, this is not possible. Second, a model trained with ArcFace has been shown to be one of the highest-ranking, open-source, face recognition algorithms in 1-to-1 comparisons according to NIST's 2021 Face Recongition Vendor Test (FRVT) [10]. Third, the developers of ArcFace open-sourced several pre-trained models [20].

In this work we leverage two pre-trained models obtained from [20]. The first is a ResNet-100 [14], trained on a refined version of the MS-Celeb-1M dataset [12] (referred to here as “ArcFace-MS1MV2”). This is the model that was evaluated in FRVT. We also utilize a second model that is an iResNet-100 [6] trained on the Glint360k dataset [2] (referred to here as “ArcFace-Glint360k”). This model is an improvement over the initial ArcFace-MS1MV2 model submitted to FRVT, both in terms of the training dataset and the architecture used. The Glint360k dataset is much larger and more demographically diverse than MS-Celeb-1M, which when used to train the iResNet architecture led to better performance reported across demographic groups according to [20].

3.3 Identifying And Removing Demographic Clustering in Feature Vectors

The S1 and S2 face samples described in Table 1 were processed using both the ArcFace-MS1MV2 and ArcFace-Glint360k algorithms producing a set of 1584 feature vectors. No failure to process errors occurred. ArcFace feature vectors are 512-dimensional. However, in general the techniques described here apply to any arbitrarily length feature vector $v \in \mathbb{R}^{1 \times p}$.

Identifying Demographic Clustering We first use the $n = 600$ samples in dataset S1 to identify feature vectors that exhibit demographic clustering using the following approach. First, we normalize the feature vectors such that $\hat{v} = \frac{\bar{v}}{\|\bar{v}\|}$. We then construct a normalized matrix of feature vectors for n subjects $\hat{V}$, where $\hat{V} \in \mathbb{R}^{n \times p}$. This matrix can be decomposed into its subject and feature specific components using singular value decomposition (SVD). The singular value decomposition of feature matrix $\hat{V}$ is defined by $\hat{V} = U \Sigma W^T$, where $U \in \mathbb{R}^{n \times n}$, $\Sigma \in \mathbb{R}^{n \times p}$, and $W^T \in \mathbb{R}^{p \times p}$.

Given the matrix of subject specific components, U, and the demographic labels for each feature vector (see Table 1), we identify which components cluster by demographic group by calculating the clustering index [17] shown in Equation (1). C_k describes the percent of variance in subject-specific component k that is explained by race and gender features. C_k is calculated by taking the ratio of *within group* variance for a demographic group D $(\sum_D \sum_{i \in D} (u_i - \bar{u}_D)^2)$ and dividing by the *overall* variance in the subject-specific component space $(\sum_i (u_i - \bar{u})^2)$. In this model, if component k had subjects spread over the full space in equal proportion to the variance of that space, both numerator and denominator would be equal and $C_k = 0$. However, if subjects in a group D are spread over less than the full component space, i.e. *cluster* in that space, the numerator becomes less than the denominator and C_k rises.

$$C_k = 1 - \frac{\sum_D \sum_{i \in D} (u_i - \bar{u}_D)^2}{\sum_i (u_i - \bar{u})^2}, \quad k, i \in \{1, ..., n\} \tag{1}$$

Empirically, every C_k is bound to have a non-zero value due to noise in the feature space, therefore we must identify which components have a statistically significant clustering indices. To evaluate the significance of each clustering index we generate a null distribution C_{null} by randomly shuffling each subject’s demographic labels and calculating C_k; this is repeated 1000 times to generate the distribution of C_k values. We then define statistically significant features to be those with C_k values greater than the 99^{th} percentile of the C_{null} distribution.

We note that the use of the linear SVD method limits us to only removing clustering based on linear relationships within groups. Removing non-linear clustering will require the use of non-linear decomposition techniques and is the focus of future research.

Removing Demographic Clustering Once identified, features that exhibit significant demographic clustering in the encoded subject space (U) can be removed from the encoded feature space (W^T). The result of this reduced matrix is $\hat{W}$, where $\hat{W} \in \mathbb{R}^{p \times m}$, and $m = p - r$. The r components are identified from $R = \{U_i...U_r\}$. Using the reduced matrix $\hat{W}$, we can reconstruct a modified feature vector $\dot{v}$ for any arbitrary v by applying the transformation $\dot{v} = v\hat{W}\hat{W}^T$. If the components of R were appropriately selected, the reconstructed feature vector $\dot{v}$ should have reduced demographic clustering and thus reduced overall specific homogeneity effects.

This technique specifically is an extension of the method proposed in [17], in which a similar transformation is performed in the score space using the eigenvalue decomposition of the similarity matrix S. More broadly, this is a modification on a widely used pattern for dimensionality reduction using SVD [24]. Our novel contributions are the application of this approach to biometric feature vectors and the selection mechanism using the clustering index C_k.

3.4 Biometric Fairness Metrics

To evaluate the efficacy of our proposed method, we apply four biometric fairness metrics to quantify the method's ability to reduce demographic bias. The four fairness metrics include the Net Clustering metric from [17], the Gini Aggregation Rate for Biometric Equitability (GARBE) from [16], the Fairness Discrepancy Rate (FDR) from [23], and the NIST Inequity Ratio from [9].

The Net Clustering metric [17], defined in Equation 2, measures the proportion of total variance in the feature vectors explained by demographic clustering, where C_k is the clustering index defined in Section 3.3, σ_k^2 is the variance of the k^{th} feature, and σ_{net}^2 is the total variance in the feature vectors. For the Net Clustering metric, a value closer to zero indicates a more "fair" algorithm.

$$C_{net} = \frac{1}{\sigma_{net}^2} \sum_k \sigma_k^2 C_k \tag{2}$$

The GARBE metric [16] is a fairness measure inspired by the Gini coefficient, a historical measure of dispersion often used in measuring wealth inequality [7]. In [16], the Gini coefficient is applied to biometric error rates, specifically the false match rate (FMR) and false non-match rate (FNMR) across demographic group D, as shown in Equations (3) to (5). As an extension of the Gini coefficient, GARBE combines measures of FMR and FNMR dispersion using a weighing factor α as shown in Equation (5). Similarly to Net Clustering, a value closer to zero indicates a more fair algorithm according to the GARBE metric.

$$G_x = \left(\frac{n}{n-1}\right)\left(\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \mid x_i - x_j \mid}{2n^2\bar{x}}\right) \forall d_i, d_j \in D \tag{3}$$

$$A(\tau) = G_{FMR_\tau};\ B(\tau) = G_{FNMR_\tau} \tag{4}$$

$$\mathit{GARBE(\tau)} = \alpha A(\tau) + (1-\alpha)B(\tau) \tag{5}$$

The Fairness Discrepancy Rate (FDR) [23] is made up of two values: the first, shown in Equation (6), measures the maximum difference in FMR values between demographic groups D for a given threshold τ and the second, shown in Equation (7), measures the maximum difference in FNMR between demographic groups at the same threshold τ. As in Equation (5), Equations (6) and (7) are mixed with the hyper-parameter α and subtracted from 1 to form the Fairness Discrepancy Rate shown in Equation (8). Unlike other fairness metrics, the FDR metric increases as “fairness” increases, meaning a value closer to 1 is more desirable.

$$A(\tau) = \max(|FMR_{d_i}(\tau) - FMR_{d_j}(\tau)|)\ \ \forall d_i, d_j \in D \tag{6}$$

$$B(\tau) = \max(|FNMR_{d_i}(\tau) - FNMR_{d_j}(\tau)|)\ \ \forall d_i, d_j \in D \tag{7}$$

$$\mathit{FDR}(\tau) = 1 - (\alpha A(\tau) + (1-\alpha)B(\tau)) \tag{8}$$

The NIST Inequity Ratio takes the maximum difference in FMR and FNMR values into account as a ratio as shown in Equations (9) and (10). This approach then proposes multiplicative and exponential scaling by risk ratios α and β as opposed to additive scaling as shown in Equation (11).

$$A(\tau) = \frac{\max(FMR_{d_i}(\tau))}{\min(FMR_{d_j}(\tau))}\ \ \forall d_i, d_j \in D \tag{9}$$

$$B(\tau) = \frac{\max(FNMR_{d_i}(\tau))}{\min(FNMR_{d_j}(\tau))}\ \ \forall d_i, d_j \in D \tag{10}$$

$$\mathit{INEQ}(\tau) = A(\tau)^{\alpha}B(\tau)^{\beta} \tag{11}$$

We note a special case of the Inequity Ratio is to only calculate this ratio when $i == j\ \ \forall d_i, d_j \in D$, essentially calculating the ratio across the diagonal of a cohort matrix (see Figure 1). We refer to this measure as $\mathit{INEQ}(\tau)^{\star}$.

4 Results

Before applying the transform to the feature vectors, we performed comparisons and calculated false match rates for both the S1 and S2 datasets, resulting in 360,000 and 36,864 comparisons respectively. We found that threshold values of 0.647 and 0.635 produced a false match rate of $1e^{-3}$ globally across the S1 dataset for comparisons performed by ArcFace-MS1MV2 and ArcFace-Glint360k, respectively. We use a global false match rate of $1e^{-3}$ to represent the use case of access control or small gallery matching at border exit and entry sites [22]. For the disjoint S2 set, threshold values of 0.657 and 0.635 produced the same false match rates, again respectively. Cross comparisons between untransformed S1 and S3 sets produced a mated comparison set of 1,238. Of these, one similarity score was below the respective S1 dataset thresholds for the ArcFace-MS1MV2 and ArcFace-Glint360k algorithms ($FNMR = 8.1e^{-4}$).

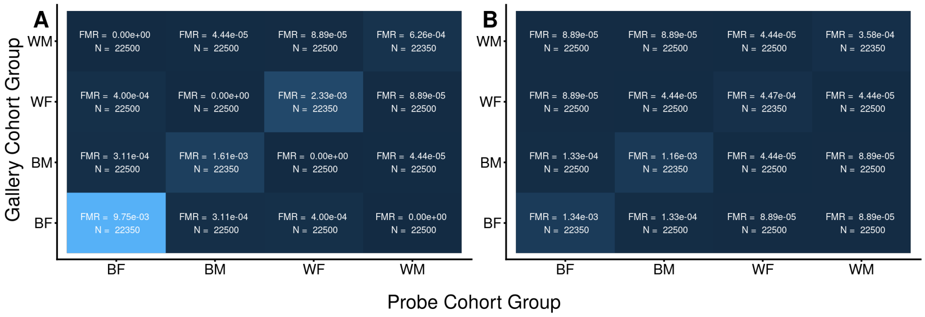
Two experiments were then performed to evaluate the proposed transform's ability to remove clustering. In Experiment 1, the de-clustering transform is calculated from the S1 dataset and evaluated on the S1 dataset. In Experiment 2, the transform calculated in Experiment 1 is applied to the S2 dataset to test the ability of de-clustering on a disjoint set. In these experiments, when evaluating fairness measure outcomes for metrics with hyper-parameters, we set α to 1 and β to 0. This focuses the measure on variations in false match rate, which are material to broad homogeneity effects.

4.1 Experiment 1 - De-clustering Learned and Applied to the Same Dataset

The first experiment's purpose is to show that the transform described in Section 3.3 is capable of removing demographic clustering effects and increasing fairness from the same dataset the transform is derived from. This is a first-order check that this technique may be useful more broadly. Before applying the de-clustering transformation, at a population FMR of $1e^{-3}$ the Black Female cohort-specific FMR was the largest at $9.75e^{-3}$ and the FMR for the White Male cohort was the smallest at $6.26e^{-4}$. Note the Black Female FMR is roughly 10x larger than the population FMR and the White Male FMR is roughly half of the population FMR. The ratio between the max and min within cohort FMRs, is a factor of roughly 15 ($INEQ(\tau)^{\star} = 9.75e^{-3}/6.26e^{-4} = 15.58$).

Once the transform is applied to the ArcFace-MS1MV2 templates, this dispersion is noticeably reduced, with the highest FMR still belonging to the Black Female cohort but now at a rate of $1.34e^{-3}$ (1.34x larger than the population FMR) and the lowest FMR still belonging to the White Male cohort but now at a rate of $3.58e^{-4}$ (35% of the population FMR). The full FMR spectrum is shown in Figure 2. Accordingly, every biometric fairness measure introduced in Section 3.4 moved in a "more fair" direction after the transform (see Table 2).

A similar effect was observed when applying the de-clustering transformation to the ArcFace-Glint360k templates. Despite being trained on a larger, more diverse dataset, templates generated with this ArcFace model still had a noticeable spread in FMR. At a population FMR of $1e^{-3}$ the FMR of the Black



A

	BF	BM	WF	WM

WM	FMR = 0.00e+00
	N = 22500
		FMR = 4.44e-05
		N = 22500
			FMR = 8.89e-05
			N = 22500
				FMR = 6.26e-04
				N = 22350

WF	FMR = 4.00e-04
	N = 22500
		FMR = 0.00e+00
		N = 22500
			FMR = 2.33e-03
			N = 22350
				FMR = 8.89e-05
				N = 22500

BM	FMR = 3.11e-04
	N = 22500
		FMR = 1.61e-03
		N = 22350
			FMR = 0.00e+00
			N = 22500
				FMR = 4.44e-05
				N = 22500

BF	FMR = 9.75e-03
	N = 22350
		FMR = 3.11e-04
		N = 22500
			FMR = 4.00e-04
			N = 22500
				FMR = 0.00e+00
				N = 22500

B

	BF	BM	WF	WM

WM	FMR = 8.89e-05
	N = 22500
		FMR = 8.89e-05
		N = 22500
			FMR = 4.44e-05
			N = 22500
				FMR = 3.58e-04
				N = 22350

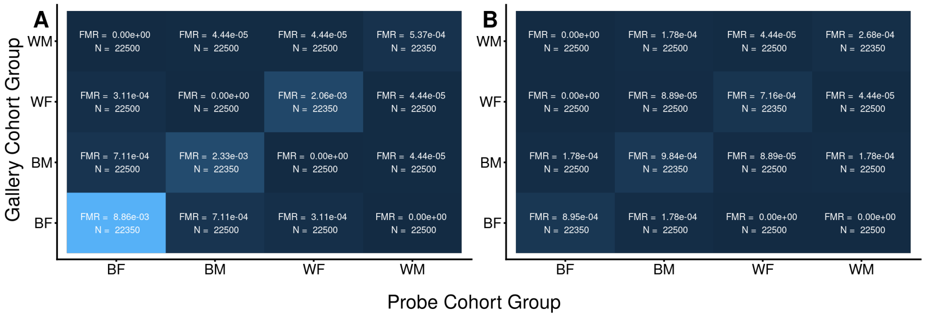
WF	FMR = 8.89e-05
	N = 22500
		FMR = 4.44e-05
		N = 22500
			FMR = 4.47e-04
			N = 22350
				FMR = 4.44e-05
				N = 22500

BM	FMR = 1.33e-04
	N = 22500
		FMR = 1.16e-03
		N = 22350
			FMR = 4.44e-05
			N = 22500
				FMR = 8.89e-05
				N = 22500

BF	FMR = 1.34e-03
	N = 22350
		FMR = 1.33e-04
		N = 22500
			FMR = 8.89e-05
			N = 22500
				FMR = 8.89e-05
				N = 22500

Fig. 2. A) False match rates across demographic groups *before* removing demographic clustering from ArcFace-MS1MV2 templates on S1 dataset. B) False match rates across demographic groups *after* removing demographic clustering from ArcFace-MS1MV2 templates on S1 dataset.

Female cohort was again the highest at $8.66e^{-3}$, 8.6x higher than the disaggregated measure. The FMR for White Males was again the lowest at $5.37e^{-4}$ or roughly half of the disaggregated measure. The ratio between these disaggregated FMRs is on a similar scale to the ratio observed in the MS1MV2 templates $(INEQ(\tau)^{\star} = 8.66e^{-3}/5.37e^{-4} \ = 16.23)$.



A

	BF	BM	WF	WM

WM	FMR = 0.00e+00
	N = 22500
		FMR = 4.44e-05
		N = 22500
			FMR = 4.44e-05
			N = 22500
				FMR = 5.37e-04
				N = 22350

WF	FMR = 3.11e-04
	N = 22500
		FMR = 0.00e+00
		N = 22500
			FMR = 2.06e-03
			N = 22350
				FMR = 4.44e-05
				N = 22500

BM	FMR = 7.11e-04
	N = 22500
		FMR = 2.33e-03
		N = 22350
			FMR = 0.00e+00
			N = 22500
				FMR = 4.44e-05
				N = 22500

BF	FMR = 8.86e-03
	N = 22350
		FMR = 7.11e-04
		N = 22500
			FMR = 3.11e-04
			N = 22500
				FMR = 0.00e+00
				N = 22500

B

	BF	BM	WF	WM

WM	FMR = 0.00e+00
	N = 22500
		FMR = 1.78e-04
		N = 22500
			FMR = 4.44e-05
			N = 22500
				FMR = 2.68e-04
				N = 22350

WF	FMR = 0.00e+00
	N = 22500
		FMR = 8.89e-05
		N = 22500
			FMR = 7.16e-04
			N = 22350
				FMR = 4.44e-05
				N = 22500

BM	FMR = 1.78e-04
	N = 22500
		FMR = 9.84e-04
		N = 22350
			FMR = 8.89e-05
			N = 22500
				FMR = 1.78e-04
				N = 22500

BF	FMR = 8.95e-04
	N = 22350
		FMR = 1.78e-04
		N = 22500
			FMR = 0.00e+00
			N = 22500
				FMR = 0.00e+00
				N = 22500

Fig. 3. A) False match rates across demographic groups *before* removing demographic clustering from ArcFace-Glint360k templates on S1 dataset. B) False match rates across demographic groups *after* removing demographic clustering from ArcFace-Glint360k templates on S1 dataset.

After applying the transform described in Section 3.3, all within cohort FMR's were within an order of magnitude of each other. The highest FMR now belonged to the Black Male cohort at $9.84e^{-4}$ and the lowest FMR still belonged to White Males at $2.68e^{-4}$. Importantly, this spread is only a factor of roughly 4x. Accordingly, all biometric fairness measures outlined in Section 3.4 improved (see Table 2).

4.2 Experiment 2 - De-clustering Learned on One Dataset and Applied to a Disjoint Dataset

While encouraging, the results in Section 4.1 are of limited utility if they do not generalize beyond the subjects used to learn the de-clustering transform. Ideally, a transform $\hat{W}$ (See Section 3.3) would apply to other subjects with similar demographics. To test this capability, a second experiment using a $\hat{W}$ learned from S1 was applied to the dataset S2. Recall, there is no subject overlap from S1 to S2, although there is demographic overlap (see Section 3.1). Prior to transform, at a population FMR threshold of $1e^{-3}$, the FMR for the Black Female cohort was the highest at $8.98e^{-3}$ and the FMR for White Males was lowest at 0, using the ArcFace-MS1MV2 templates. The FMR for the second lowest cohort, White Females, was $8.5e^{-4}$, making the best calculable spread ratio roughly an order of magnitude ($INEQ(\tau)^{\star} = 8.98e^{-3}/8.5e^{-4} \;\; = 10.56$).

When the transformation, derived from the S1 dataset, is applied to the disjoint S2 dataset, we again see a decrease in error-rate disparity for FMR as shown in Figure 4. While the correction is smaller in magnitude than when it was when learned and applied within S1 (as expected), the decrease shows that the transformation can be generalized to an extent to feature vectors derived from unseen faces. After the de-clustering transform, the highest FMR was still for black females at $7.35e^{-3}$. The lowest, non-zero FMR is for White Males, at $1.11e^{-3}$. This factor of $INEQ(\tau)^{\star} \approx 7$ is an improvement on the spread observed before the transform. All other biometric fairness measures outlined in Section 3.4 also improved (see Table 2).

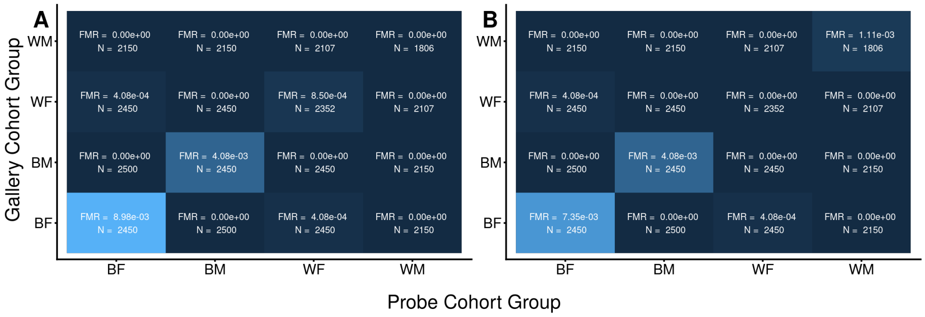


Fig. 4. A) False match rates across demographic groups *before* removing demographic clustering from ArcFace-MS1MV2 templates on S2 dataset. B) False match rates across demographic groups *after* removing demographic clustering from ArcFace-MS1MV2 templates on S2 dataset.

A similar effect was observed when applying the de-clustering transformation, derived from the S1 dataset, to the ArcFace-Glint360k templates in S2. Originally in the S2 dataset, at a population FMR of $1e^{-3}$, the Black Female cohort experienced a FMR of nearly 10x that rate ($1.06e^{-2}$). The lowest FMR was experienced by the White Male group at 0 and the second lowest by the White Female group at $8.5e^{-4}$. The spread ratio between the highest and lowest calculable within cohort FMRs is thus ($\mathit{INEQ}(\tau)^{\star} = 1.06e^{-2}/8.5e^{-4} \ = 12.5$). After the de-clustering transform is applied all within cohort FMR's were within an order of magnitude of each other. The highest FMR was still for the Black Female cohort at $4.08e^{-3}$ and the lowest non-zero FMR was for White Males at $1.11e^{-3}$, leading to a $\mathit{INEQ}(\tau)^{\star}$ of 4x. All other biometric fairness measures outlined in Section 3.4 also improved (see Table 2).

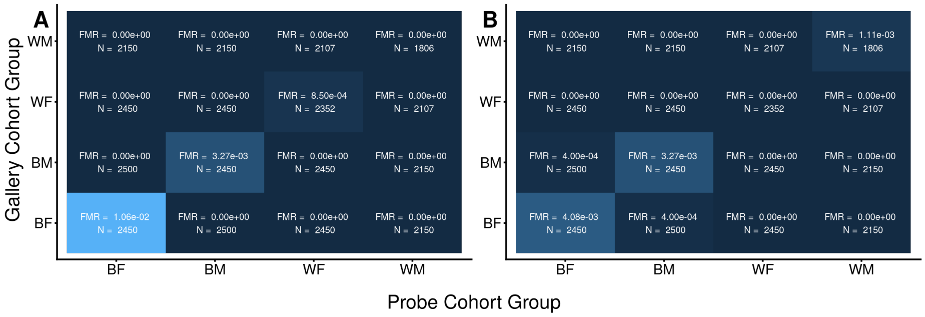


Fig. 5. A) False match rates across demographic groups *before* removing demographic clustering from ArcFace-Glint360k templates on S2 dataset. B) False match rates across demographic groups *after* removing demographic clustering from ArcFace-Glint360k templates on S2 dataset.

The de-clustering transform derived from the S1 dataset, on both algorithms, was also applied to feature vectors in the S3 dataset and mated similarity scores between S1 and transformed S3 were calculated. Of 1,238 mated comparisons, one had a similarity score below the S1 non-transformed threshold, giving an $\mathit{FNMR} = 8.1e^{-4}$. Upon further inspection this one false non-match was for a subject who's clothing had a distractor face, meaning the true FNMR for both the transformed and untransformed templates was likely 0. FNMRs at these levels confirm the feature vector transform documented here both improves fairness (see Table 2) while preserving useful true match rates.

5 Discussion and Conclusions

5.1 Summary

In this research, we've shown that the clustering index metric can be used to measure demographic clustering in the space of face recognition feature vectors. We

Table 2. Fairness metric values calculated on the training (S1) and test (S2) datasets, using the ArcFace-MS1MV2 and ArcFace-Glint360k algorithm, before and after demographic clustering correction. Optimal fairness measures for each experiment are shown in **bold**. Note that fairness measures *universally* move in a "more fair" direction (increasing for FDR, decreasing for Net Clustering, GARBE, and NIST INEQ) once the demographic clustering correction method from Section 3.3 is applied.

Algorithm	**Fairness Metric**	**Experiment 1**		**Experiment 2**	
		S1 Original	**S1 Transformed**	**S2 Original**	**S2 Transformed**
ArcFace-MS1MV2	Net Clustering	0.0163	**0.00549**	0.0252	**0.0207**
	GARBE	0.8540	**0.65000**	0.922	**0.909**
	FDR	0.9900	**0.99900**	0.991	**0.993**
	INEQ	219.00	**30.2000**	22.00	**18.00**
	INEQ*	15.58	**3.74**	10.56	**6.62**
ArcFace-Glint360k	Net Clustering	0.0150	**0.00497**	0.0250	**0.0197**
	GARBE	0.8350	**0.67100**	0.955	**0.881**
	FDR	0.9910	**0.99900**	0.990	**0.996**
	INEQ	199.00	**22.1000**	12.5	**10.20**
	INEQ*	16.23	**3.67**	12.47	**3.68**

then show how we can use this knowledge to form a matrix transformation that removes feature vector components that exhibit demographic clustering from a disjoint test set of feature vectors. Applying this transformation decreases the disparity in false match rates across demographic groups. As evidence of this, we show increases in four published "fairness" metrics. We replicate these findings across two, separately trained biometric algorithms, ArcFace-MS1MV2 and ArcFace-Glint360k. We believe this is evidence of this approaches generalizability and utility.

5.2 Impact On Human & Algorithm Identification Workflows - Why Does this Matter?

When performing face identifications in practice, it is common to use a face recognition algorithm to generate similarity scores between a probe image and a gallery of images. The results are then ranked by decreasing similarity score and down-selected to include only the top n possible matches, referred to here as a "rank-n candidate list". This candidate list is then passed on to a human adjudicator whose task is to choose the image from the list that matches the probe subject.

Broad homogeneity effects in the identification context mean that the candidate list will consist largely of subjects belonging to the same demographic group as the probe subject. This consequently makes the identification task for the human more difficult, which can result in errant outcomes.

To explore the broad homogeneity effect in an identification operation, we performed identifications for the 466 subjects in the S3 dataset against the 600 subjects in the S1 dataset and ranked them by their similarity scores. This process was performed for both the original and transformed ArcFace-MS1MV2 face templates. In Figure 6 we show two, Rank-3 candidate lists resulting from the identification of two subjects; one list generated using the original face templates, Figure 6A-B, and the other generated using the transformed templates, Figure 6C-D. To simulate what the human adjudicator would see in the identification process we embed a mated image amongst the Rank-3 non-mated images in a random position.

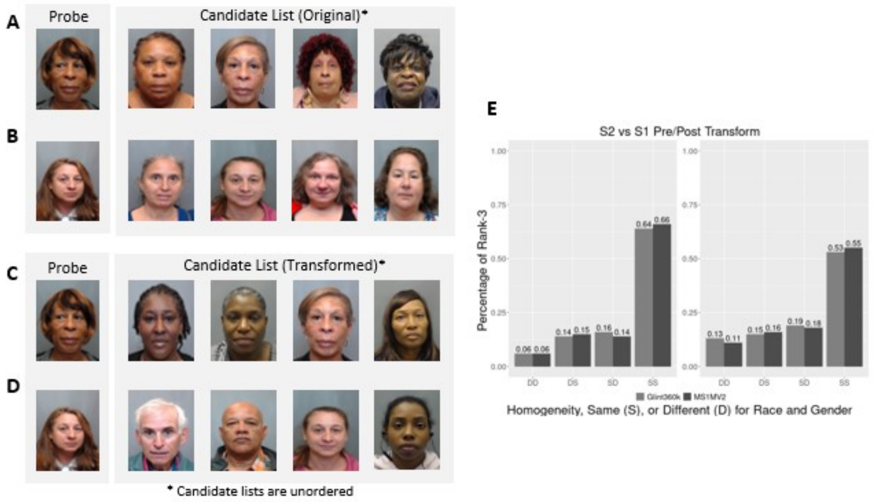


Fig. 6. **(A-B)** Rank-3 Candidate lists, with mated images inserted at random, for two subjects in S3 dataset compared against images in S1 dataset for non-transformed ArcFace-MS1MV2 face templates. **(C-D)** Rank-3 Candidate lists, with mated images inserted at random, for two subjects in S3 dataset compared against images in S1 dataset for transformed ArcFace-MS1MV2 face templates. **(E)** Percentage of homogeneity in Rank-3 identification results.

For both of the non-transformed candidate lists in Figure 6A-B, we note that all subjects included in the list are of the same demographic group. After applying the transform, the candidate list for one probe changed from a demographically homogeneous list in Figure 6A, to demographically diverse list in Figure 6D. We also note that the demographic homogeneity for the other candidate list, Figure 6C, did not change after applying the transformation, highlighting that future work in developing more sophisticated transformations is needed. Despite the remaining homogeneity of some candidate lists in this experiment, we do note an 11% decrease in the percentage subjects belonging to the same demographic group within the Rank-3 results, as seen in Figure 6E.

5.3 Further Research

This research demonstrates that broad homogeneity effects can be reduced by removing components of the face feature vectors that show demographic clustering. However, due to the limited size of the datasets used here, we suggest further analysis is needed to confirm the effectiveness of the proposed method on larger, open-source identification galleries comparable to those used in practice. We also suggest analysis of the proposed method on face recognition models trained with other loss functions, such as CurricularFace [19] or ElasticFace [3], as well as the evaluation of the approach when α and β parameters for the fairness metrics vary.

In addition to the use of larger identification galleries and other loss functions, we intend to experiment with integrating the proposed methodologies into the deep neural network training procedures. This avenue of research involves the development of loss functions designed to limit the effects of demographic clustering during the training of a face recognition algorithm. We hope that the applications of this research increases focus in the biometrics community on the development of more equitable systems.

Acknowlegments

This research was funded by the Department of Homeland Security, Science and Technology Directorate (DHS S&T) on contract number W911NF-13-D-0006-0003. The views presented do not represent those of DHS, the U.S. Government, or the author's employers. The biometric dataset utilized in this work is not publicly available at this time.

Paper contributions: All authors conceived the work; Eli J. Laird and John J. Howard performed the statistical analysis and wrote the paper; Yevgeniy B. Sirotin devised mathematical method to mitigate clustering, advised on statistical analysis, and edited the paper. Jerry L. Tipton (IDSLabs) and Arun R. Vemury (DHS S&T) also conceived the work.

References

1. 2021 Rally - Maryland Test Facility. https://mdtf.org/Rally2021
2. An, X., Zhu, X., Gao, Y., Xiao, Y., Zhao, Y., Feng, Z., Wu, L., Qin, B., Zhang, M., Zhang, D., Fu, Y.: Partial fc: Training 10 million identities on a single machine. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 1445–1449 (2021). https://doi.org/10.1109/ICCVW54120.2021.00166
3. Boutros, F., Damer, N., Kirchbuchner, F., Kuijper, A.: Elasticface: Elastic margin loss for deep face recognition. CoRR **abs/2109.09416** (2021), https://arxiv.org/abs/2109.09416
4. Deng, J., Guo, J., Niannan, X., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: CVPR (2019)

5. Drozdowski, P., Rathgeb, C., Busch, C.: The watchlist imbalance effect in biometric face identification: Comparing theoretical estimates and empiric measurements. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3757–3765 (2021)
6. Duta, I.C., Liu, L., Zhu, F., Shao, L.: Improved residual networks for image and video recognition. arXiv preprint arXiv:2004.04989 (2020)
7. Gini, C.: Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E (1912)
8. Gong, S., Liu, X., Jain, A.K.: Jointly De-Biasing Face Recognition and Demographic Attribute Estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 330–347. Springer International Publishing, Cham (2020)
9. Grother, P.: Face recognition vendor test (frvt) part 8: Summarizing demographic differentials (2022)
10. Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test (FRVT) part 2: Identification (2018)
11. Grother, P., Ngan, M., Hanaoka, K.: Face recognition vendor test (FRVT) part 3: Demographic effects (2019)
12. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 87–102. Springer International Publishing, Cham (2016)
13. Hasselgren, J.A., Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: A scenario evaluation of high-throughput face biometric systems: Select results from the 2019 Department of Homeland Security Biometric Technology Rally. The Maryland Test Facility (2020)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)
15. Howard, J.J., Sirotin, Y.B., Vemury, A.R.: The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance. In: 2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS). pp. 1–8 (2019). https://doi.org/10.1109/BTAS46853.2019.9186002
16. Howard, J.J., Laird, E.J., Sirotin, Y.B., Rubin, R.E., Tipton, J.L., Vemury, A.R.: Evaluating proposed fairness models for face recognition algorithms (2022). https://doi.org/10.48550/ARXIV.2203.05051, https://arxiv.org/abs/2203.05051
17. Howard, J.J., Sirotin, Y.B., Tipton, J.L., Vemury, A.R.: Quantifying the extent to which race and gender features determine identity in commercial face recognition algorithms (2020)
18. Huang, G., Mattar, M.A., Berg, T.L., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
19. Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., Huang, F.: Curricularface: Adaptive curriculum learning loss for deep face recognition. CoRR **abs/2004.00288** (2020), https://arxiv.org/abs/2004.00288
20. insightface: State-of-the-art 2d and 3d face analysis project, https://github.com/deepinsight/insightface/tree/master/model_zoo
21. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 212–220 (2017)

22. Manaher, C.: Privacy impact assessment for the traveler verification service (2018), https://www.dhs.gov/publication/dhscbppia-056-traveler-verification-service
23. Pereira, T.d.F., Marcel, S.: Fairness in biometrics: a figure of merit to assess biometric verification systems. IEEE Transactions on Biometrics, Behavior, and Identity Science pp. 1–1 (2021). https://doi.org/10.1109/TBIOM.2021.3102862
24. Rajaraman, A., Ullman, J.D.: Mining of massive datasets. Cambridge University Press (2011)
25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **07-12-June**, 815–823 (2015). https://doi.org/10.1109/CVPR.2015.7298682
26. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
27. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018)