Demographic Effects on Estimates of Automatic Face Recognition Performance

Alice J. O'Toole[a], P. Jonathon Phillips[b], Xiaobo An[a], Joseph Dunlop[a]

[a] *The University of Texas at Dallas*
Richardson, TX 75080, USA
[b] *National Institute of Standards and Technology*
Gaithersburg, MD 20899, USA

Abstract

The intended applications of automatic face recognition systems include venues that vary widely in demographic diversity. Formal evaluations of algorithms do not commonly consider the effects of population diversity on performance. We document the effects of racial and gender demographics on estimates of the accuracy of algorithms that match identity in pairs of face images. In particular, we focus on the effects of the "background" population distribution of non-matched identities against which identity matches are compared. The algorithm we tested was created by fusing three of the top performers from a recent US Government competition. First, we demonstrate the variability of algorithm performance estimates when the population of non-matched identities was demographically "yoked" by race and/or gender (i.e., "yoking" constrains non-matched pairs to be of the same race or gender). We also report differences in the match threshold required to obtain a false alarm rate of .001 when demographic controls on the non-matched identity pairs varied. In a second experiment, we explored the effect on algorithm performance of progressively increasing population diversity. We found systematic, but non-general, effects when the balance between majority and minority populations of non-matched identities shifted. Third, we show that identity match accuracy differs substantially when the non-match identity population varied by race. Finally, we demonstrate the impact on performance when the non-match distribution consists of faces chosen to resemble a target face. The results from all experiments indicate the importance of the demographic composition and modeling of the background population in predicting the accuracy of face recognition algorithms.

Preprint submitted to Image and Vision Computing *December 7, 2011*

Keywords: face recognition, algorithm evaluation, demographics

1. Introduction

The appearance of a face is determined by its gender, race/ethnicity, age, and identity. Any given image of a face depends also on the viewing angle, illumination, and resolution of the sensor. The goal of most face recognition algorithms is to identify someone as a unique individual. Often this test requires the algorithm to match the identity of faces between images that may vary in the quality or nature of the viewing conditions. The diversity of faces in the real world means that face recognition algorithms must operate across a backdrop of appearance variability that is *not related* to an individual's unique identity. Thus, face recognition algorithms intended for real-world applications must perform predictably over changes in the demographic composition of the intended application populations. The most likely application sites for algorithms include airports, border crossings, and crowded city sites such as train and metro stations. These locations are characterized by ethnically diverse populations that may vary by the time of year (e.g., tourist season) or even by the time of day (e.g., flights from the Far East arrive in the morning and European flights in the afternoon).

The performance of state-of-the-art automatic face recognition algorithms has been tested extensively over the last two decades in a series of U.S. Government-sponsored tests (e.g., [1]). Measures of algorithm performance in these tests provide the best publicly available information for making decisions about the suitability and readiness of algorithms for security and surveillance applications of importance. Traditionally, these tests have emphasized measuring performance over photometric conditions such as illumination and image resolution [2, 3, 4, 5]. They have also concentrated primarily on the quality of the "match" (i.e., similarity) between images of the same individuals, considering this to be the critical factor determining algorithm performance. When the degree of similarity between matched identities is high, as is the case when photometric factors are controlled, the algorithm is expected to perform well.

Much less consideration has been given to the distribution against which matched identities are compared. To digress briefly, all measures of the performance of face recognition algorithms rely both on the distribution of data for the population of identity matches (i.e., pairs of images of the same

person) and on the distribution of data for non-matched identities (i.e., pairs of images of different people). By definition, the identity match population contains image pairs of the same race and gender. The non-matched identity population, however, may be structured in several ways. To date, in most evaluations, this distribution consists of pairs of faces that have different identities and which may, or may not, be of different races or genders.

More formally, identity match decisions (same person or different people?) are generally based on a computed similarity score between two face images. If this score exceeds a threshold similarity score, the faces are judged as an "identity match" (sometimes referred to as an "identity verification"). Otherwise, the images are judged as non-matched identities. False alarm errors occur when the computed similarity score between a pair of face images of different individuals exceeds the match threshold. In most applications, the similarity cutoff threshold is set to achieve a low false alarm rate (commonly on the order of 0.001).

As noted, in many formal evaluations of algorithms, the distribution of similarity scores for the non-matched face images includes face pairs that may differ in gender, race/ethnicity, and age [1]. The inclusion of these categorically mismatched image pairs may lead to an over-estimation of an algorithm's ability to discriminate the *identity* of face pairs. In other words, when non-match face pairs differ on categorical variables such as race and gender, some part of the performance of the algorithms may be due to the easier problem of discriminating faces based on race or gender, rather than to the more challenging problem of recognizing individual face identities. From a theoretical point of view, any demographic factor that decreases the average similarity of the non-matched face pairs will increase the estimated performance of an algorithm.

In this study, we focus on the problem of how the demographic composition of non-match populations affects estimates of algorithm accuracy. Although demographic variables have been shown to affect both human ([6, 7, 8, 9]) and machine ([10, 11]) accuracy recognizing faces, the effects of these variables in the non-matched identity distributions have not been studied previously. The first goal of this study was to document the effects of yoking non-match pairs according to the categorical variables of race and gender, individually, and together. By "yoking" we mean controlling the demographic variables within a non-match pair, such that both faces in the pair are of the same gender, same race, or same gender and race. We also examine the implications of demographic control of the non-match pairs for

the choice of a threshold for match/non-match decisions. The second goal was to examine algorithm accuracy with systematic variations in the proportion of a "second" ethnic group in the non-match distribution. Third, we measured algorithm accuracy when the identity matches were of a particular race and identity mismatches were from another race. This was compared to the case when the match and non-match distributions were of the same race. Finally, we look at the challenging, but plausible security application scenario, that occurs when a deliberate attempt is made to impersonate another person. Specifically, we evaluate algorithm performance estimates when the non-match distribution consists of selected imposters, chosen to be similar to a target.

2. Algorithm Fusion and Test Protocol

In this section, we overview the algorithm and test protocol common to all experiments. The source of data for these experiments was the FRVT 2006 [1], a U.S. Government sponsored test of face recognition algorithms conducted by the National Institute of Standards and Technology (NIST) (Details of this test and its results can be found elsewhere, [1]). We used face recognition algorithms from the FRVT 2006 international test because they are among the best algorithms available for testing and may be typical of algorithms currently in use in real world applications. Due to the fact that many of the submitted algorithms were proprietary, the FRVT 2006 was conducted using executable versions of each algorithm's code. Thus, although it would be desirable to know specifics of the how the algorithms work (i.e., training regimes, "experience" with faces from different demographic backgrounds, etc.) this was not possible. Despite this shortcoming, the present experiments may instead provide a realistic look at how the performance of algorithms that are currently "in the field" varies as a function of the demographic assumptions made in the measurement process.

Here, we focused on the *uncontrolled-uncontrolled* identity match test in the FRVT 2006 (*Stimulus Set 1*). In that test, algorithms matched identity in pairs of images taken under uncontrolled illumination conditions. To demonstrate the general nature of the problem, we replicated a subset of our results with a second data set (*Stimulus Set 2*) that differed substantially in demographic characteristics from Stimulus Set 1. Specifically, Stimulus Set 1 consisted of mostly young adults between the ages of 18 and 29, with a substantial female majority. Stimulus Set 2, by contrast was distributed

Figure 1: Example same-identity image pair for the uncontrolled-uncontrolled identity matching task.

Table 1: Summary demographic composition of the Stimulus Set 1.

Ethnicity	Total	Female	Male
Caucasian	716	318	398
Asian	264	121	143
Hispanic	26	10	16
Southern Asian	23	2	21
Middle Eastern	6	1	5
African-American	12	1	11
Unknown	41	16	21
Total	1088	469	619

more evenly across adult age groups and was more gender balanced. The majority race for both datasets was Caucasian, but there was also a substantial minority population in each. For Stimulus Set 1, the minority population was East Asian, whereas for Stimulus Set 2 the minority population was Hispanic. Combined, the broad differences in the demographic structure of the two datasets provided a strong test of the generality of the effects we report.

2.1. Stimulus Set 1

The Notre Dame multi-biometric data set [1] was the primary source of face stimuli for the test. This dataset consists of 9,307 images of 570 individuals. The images were photographed with a 6 Megapixel Nikon D70 camera and were taken under uncontrolled illumination conditions, either outside or in a corridor or hallway. An example image pair appears in Figure 1 and the relevant demographic breakdown of faces appears in Table 1. (Recall that multiple images of individuals are represented in these lists).

Table 2: Summary demographic information for sex, race, and age in Stimulus Sets 1 and 2. Values within a demographic category are by percent (numbers are rounded). If the number of subjects in a given category is less than 2.5 percent, then the cell is left blank. Note: the race categories are Caucasian (C), East Asian (EA) and Hispanic (H).

	Sex		Race			Age				
Dataset	F	M	C	EA	H	18-29	30-39	40-49	50-59	60+
Notre Dame	62	38	76	13		92	7			
Sandia	55	45	64		21	15	11	23	35	18

2.2. Stimulus Set 2

The Sandia data set was collected at the Sandia National Laboratory and consists of high-resolution frontal face images taken under both controlled and uncontrolled illumination [1]. The images were taken with a 4 Megapixel Canon PowerShot G2. The average face size for the controlled images was 350 pixels between the centers of the eyes and 110 pixels for the uncontrolled images. For comparison with Stimulus Set 1, the demographic breakdown of faces appears in Table 2, with additional detail given to the age of subjects, which is more variable in this data set.

2.3. Algorithm Test Procedure

In the FRVT 2006 evaluation, each algorithm computed a similarity score for all possible pairs of *target* and *query* images. This yielded a matrix of similarity scores where element $s_{i,j}$ represents the similarity between the i^{th} target and the j^{th} query image. The goal of the algorithms was to distinguish matched-identity image pairs and non-matched identity image pairs using the similarity scores. The performance of the algorithms was evaluated at NIST using receiver operator characteristic (ROC) curves that plot the proportion of false alarms (false accepts) against the proportion of hits (identity verifications). Because face recognition algorithms are required to operate at low false alarm rates, a second measure was also computed – the verification or hit rate at the 0.001 false alarm rate. In this paper, we use both types of measures.

2.4. Algorithm Fusion Data

The experiments were conducted using data extracted from a fusion of three of the top performing algorithms in the FRVT 2006. The fusion of the

similarity matrices operated by first estimating the median and the median absolute deviation (MAD) from 1 in 1023 similarity scores ($median_k$ and MAD_k reference the median and MAD for the k^{th} algorithm). The fused similarity scores were computed as the sum of each algorithm's similarity scores after the median has been subtracted and then divided by the MAD. Thus, if s_k is a similarity score for the k^{th} algorithm, and s_f is a fusion similarity score, then $s_f = \sum_k (s_k - \text{median}_k) \ / \ \text{MAD}_k$.

The fused data were partitioned subsequently into three performance strata, representing face image pairs that were matched at high, moderate, and low levels of performance. This performance stratification has been referred to elsewhere as the "Good, Bad, and Ugly (GBU) Challenge Problem" [12]. In this study, we used face pairs from the "bad" and "ugly" performance strata, which we refer to here as the *moderately difficult* and the *difficult* face pair conditions. The GBU challenge problem highlights the broad range of algorithm performance when frontal image matches are made. An important constraint imposed across these strata was that the pairs of *identities* used for both the matched and non-matched images were held constant. Thus, only the *images* of these identities differed across the three strata. As such, the performance variations among the three levels were due to photometric quality issues (e.g., illumination) or to other extraneous variations in facial expression rather than to recognizability differences of the individual identities.

For any given matched identity pairing (i.e., images of the same person), there were multiple similarity scores available in all three performance strata. We used data from the moderately difficult and difficult face pairs because they had substantial error rates. In each condition, the fused algorithm data consisted of a matrix of similarity scores between all possible pairs of $1,088$ target face images and $1,088$ query images[1]. This included $3,306$ matched identity pairs and $1,180,438$ non-matched identity pairs. Table 1 shows the ethnic and gender composition of the identities represented in the $1,088$ identities of the target and query image lists used to create the similarity matrix. (Recall that multiple images of individuals are represented in these lists). As Table 1 makes clear, Caucasians and Asians are the best represented races in the dataset. We use only these two races in the experiments we report

[1]The results in this paper use an earlier version of the GBU than released previously [12]. The difference between the two versions is the removal of one identity.

in Sections 4 and 5 of this paper.

3. Demographic Pairing in Non-Match Identity Distributions

The goal of these experiments is to document changes in performance estimates for face recognition algorithms as a function of the demographic characteristics of the non-matched face pairs. In the first section, we show that performance estimates vary substantially when the non-match population consists of face pairs that are yoked by demographic groups (gender, race, gender and race). In the second section, we examined the effects of demographic controls on the appropriate choice of a threshold cutoff point for assigning an identity match decision to the face pairs.

3.1. Partitioning the Performance Estimation Process by Demographic Bins

We computed ROC curves in four ways that vary in the demographic controls applied to the non-matched identity distribution. Figure 2 shows the fusion algorithm's performance for the *moderately difficult* (left) and *difficult* (right) face pairs. In the *No Demographic Matching* (No DM) condition, we used all available pairs of mis-matched identities to compute the ROC curve. Thus, non-match pairs could be of the same or different gender and/or race. In the *Gender-matched Demographic* (DM Gender Only) condition, we used only same-sex non-match pairs to compute the ROC curve, although these pairs could be of a different race. In the *Race-matched Demographic* (DM Race Only) condition, we used only same-race non-match pairs, although these pairs could be of a different gender. In the *Gender and Race matched Demographic* (DM Gender-Race) condition, the non-matched pairs were of the same-race and same-gender.

The graphs show a substantial decrease in performance as the non-match distribution becomes more demographically controlled. The difference in performance is further seen in Table 3, which shows the verification rates at the .001 false alarm rate, as the demographic controls change. These rates change markedly between the condition where there is no demographic control and the condition where both gender and race are controlled. Notably, the verification rate for the difficult face pairs nearly doubles when the demographic controls are removed completely. For the moderately difficult pairs, the verification rate differs by 10 percent between the no control and full control conditions.

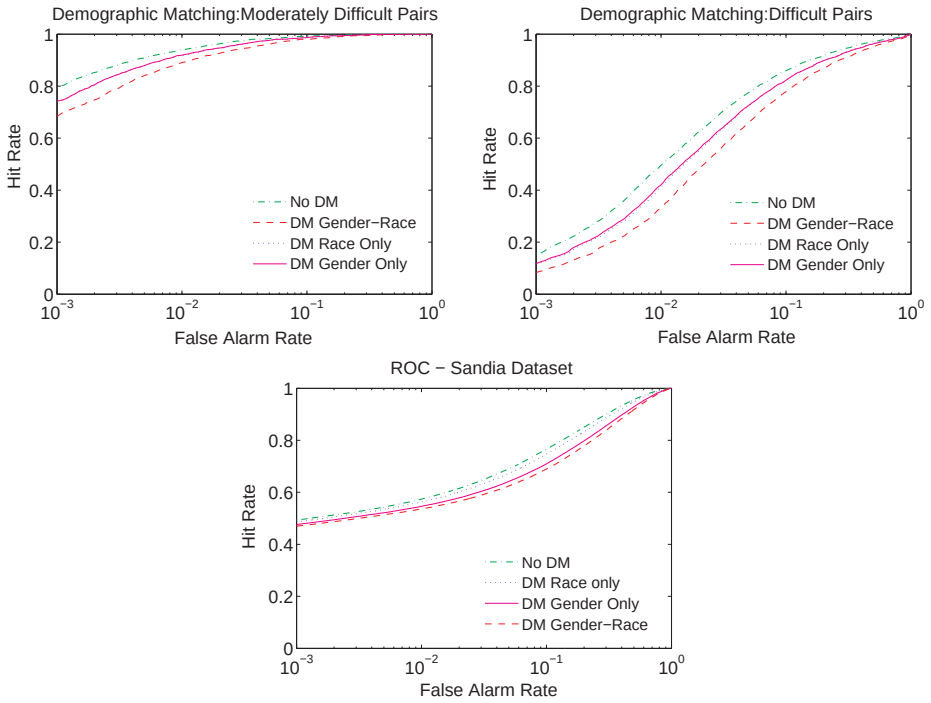

Figure 2: ROC curves for the moderately difficult (left) and difficult (right) face pairs with different kinds of demographic control on the non-matched identity pairs. Performance is best when the pairs are uncontrolled (No DM) and worst when the pairs are same-gender and same-race pairs (DM Gender-Race). The gender-only and race-only controls yield performance in between the no-control and full control conditions. Analogous ROC curves for the Sandia stimulus set appear in the second row and support these findings using a database with different demographic characteristics.

Analogous ROC curves for the Sandia data set also appear in Figure 2. Similar to the results for Stimulus Set 1, the more demographically controlled conditions showed decreased performance estimates. This finding bolsters our conclusions about the effects of demographic control using an independent data set with a markedly different demographic structure.

3.2. Estimating the Decision Threshold Based on Demographic Composition of Non-Matched Identities

In application-based scenarios, face recognition algorithms use a threshold similarity score to determine if two face images are the same identity or are different identities. Threshold similarity scores are usually chosen to optimize

Table 3: Verification rate at the .001 False Alarm Rate with Variable Types of demographic control

Demographic Control	Moderately Difficult	Difficult
No DM	.7925	.1515
DM Gender Only	.7402	.1165
DM Race Only	.7408	.1137
DM Gender and Race	.6851	.0823

an operational criterion such as a false alarm rate. Typically, thresholds for these automatic face recognition systems are generally set to maintain a false alarm rate of 0.001. Here, we look at the implications of variable demographic controls in the choice of a similarity threshold cutoff for identity match decisions.

In Figure 3, we plot the false alarm rate as a function of the threshold similarity score, using the four demographic control procedures described previously (No DM, DM Gender Only, DM Race Only, DM Gender-Race). (Note that lower values indicate higher similarities.) The left graph shows the results for the moderately difficult pairs and the right graph shows the results for the difficult pairs. The graph below shows the shift for the Sandia data set. In all cases, the threshold that produces a false alarm rate of .001 is shifted for the different demographic control conditions. Moreover, the full functions of thresholds are similarity shifted. This indicates that to operate at a particular false alarm rate, threshold values must be set taking into consideration the nature of demographic controls in the non-match population. The results indicate that as the demographic controls tighten, the similarity cutoff threshold shifts to higher similarity scores.

4. Simulations on Mixed Demographics

In real world applications, the representativeness of different demographic categories varies in different population contexts, from nearly 100 percent of a single majority race to various degrees of inclusion of (an)other race(s). In this section, we systematically explore the effects of progressive increases in population diversity on algorithm performance. Again, we focus on diversity in the non-matched identity distribution. We begin by measuring algorithm performance when only one race of faces is included in the non-match identity distribution (Caucasian). Next, we gradually increase the

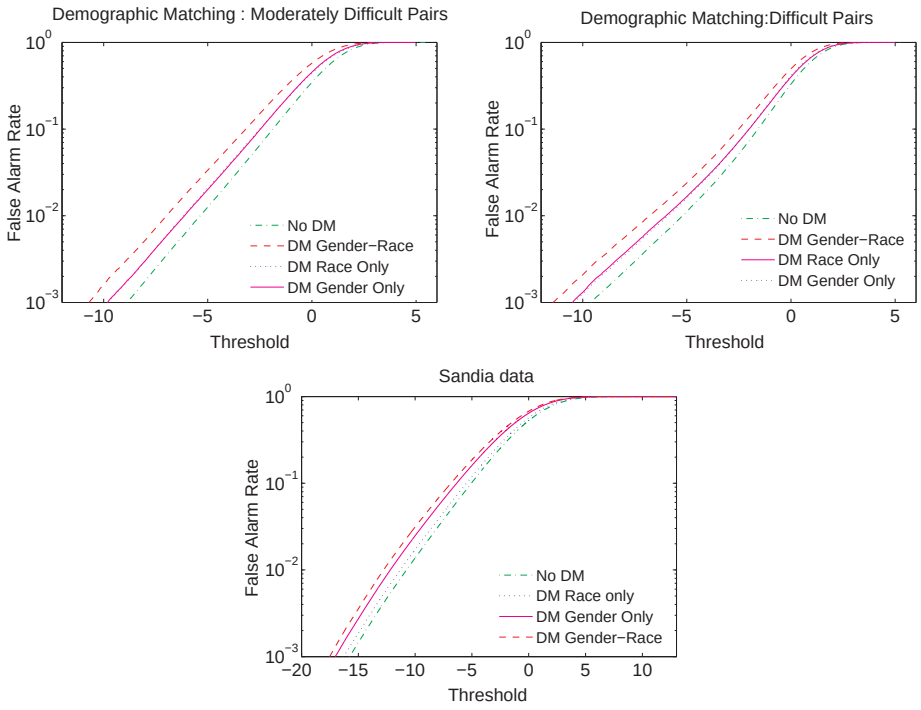

Figure 3: False alarm rate is plotted as a function of threshold similarity score, using the four demographic control procedures. (Note that lower values indicate higher similarities.) The left graph shows the results for the moderately difficult pairs and the right graph shows the results for the difficult pairs. The graph below shows analogous data from the Sandia stimulus set. The threshold shifts toward values in the target distribution (i.e., higher similarity scores) as the demographic controls tighten.

numbers of faces of a second race (Asian) in this distribution and reassess performance. We refer to this case as the *Caucasian-to-Asian* condition. In a second simulation, we reversed the process. In this case, we start with only Asians in the non-matched identity distribution, and progressively include more Caucasians. We refer to this case as the *Asian-to-Caucasian* condition. In the Caucasian-to-Asian condition, the matched identity distribution contained only Caucasian face pairs. In the Asian-to-Caucasian condition, the matched identity distribution contained only Asian face pairs. (In Section 5, we explore other combinations of the identity match distributions.)

Because of the imbalance in the number of Caucasian and Asian faces available in the database, these two conditions investigate different scenarios of population shifts. The *Caucasian-to-Asian* condition simulates a population shift that begins with a single race and ends with a population that is characterized by a strong majority race and moderately sized minority population (slightly over 10 percent). The *Asian-to-Caucasian* condition simulates a population shift that begins with a single race (Asian) and progressively shifts to equal inclusion of the two races (Asian and Caucasian). As additional Caucasian pairs are included, the population continues to shift toward a majority of Caucasians and a minority Asian population.

4.1. Methods

The stimulus set for this experiment included faces of the two races that were best represented in the database: Asians and Caucasians. (See Table 1 for a listing of numbers of people in these races). Note that "Asians" in this study refer to people from the Far East (e.g., China, Japan, Korea). Recall that the original similarity matrix consisted of 1088×1088 entries containing the similarity scores for all possible pairs of the images. The data for this experiment consisted of a subset of those similarity scores. Specifically, they were the set of similarity scores from same-gender and same-race Caucasian and Asian face pairs.

Identity Match Data. The identity match data for both conditions did not vary across the simulations. For the Caucasian-to-Asian condition, the similarity scores for all possible pairs of Caucasian identity matches ($n = 2,110$) were used to create the identity match distribution. For the Asian-to-Caucasian condition, the similarity scores of all possible pairs of Asian identity matches ($n = 866$) were used to create the identity match distribution. These numbers represent approximately three images of each person

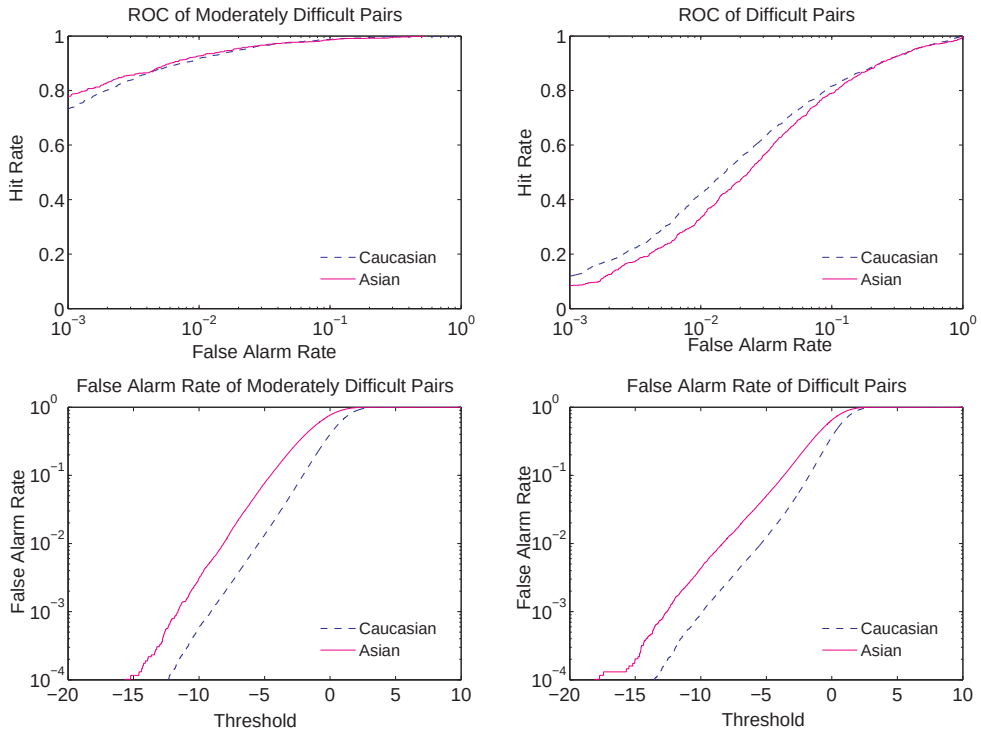

Figure 4: ROC curves that show the performance of the fusion algorithm on the Caucasian and Asian face pairs (top panels). There is a marginal advantage for Asian faces in the moderately difficult pair condition and an inversion of this pattern for the difficult pairs. Consistent with previous experiments, the bottom panels of the figure show shifts in the similarity threshold functions that produce a constant false alarm rate.

in the database, allowing for approximately three pairings of a person with him/herself.

Identity Non-Match Distribution. For the Caucasian-to-Asian condition, the simulation began with the non-match distribution consisting of the similarity scores from all possible gender-matched Caucasian pairs of different identities ($n = 257,418$). For the Asian-to-Caucasian case, the simulation began with the non-match distribution consisting of the similarity scores from all possible gender-matched Asian pairs of different identities ($n = 34,224$).

In both cases, face pairs from the other race, were added into the non-match distribution, $1,000$ at a time, and the verification rate at the 0.001 false alarm rate was computed.

4.2. Results

For completeness and clarity, before reporting the results of the demographic mixing in the non-match distribution, we first present the baseline performance of the algorithm for the Asian versus Caucasian face pairs in Figure 4. The top two panels indicate that algorithm performance for the moderately difficult face pairs was marginally better for the Asian faces (left panel). When the face pairs were difficult, the performance was consistently better for the Caucasian faces (right panel). We do not offer a strong interpretation of this finding, because it is arbitrarily dependent on the three algorithms chosen for the fusion. Two of these algorithms were from Western countries and the third algorithm was from a country in East Asia. In the bottom two panels of Figure 4, we see a substantial shift in the similarity threshold that produces a constant false alarm rate when the face race differs. Again this indicates the importance of demographics in the choice of a threshold value selected with the goal of achieving a constant false alarm rate.

The results from the demographic mixing of the non-match distribution appear in Figure 5. The top lines in the figure show the performance for the moderately difficult pairs and the bottom lines show performance for the difficult pairs. The blue lines show verification rate at 0.001 false accept rate as Asian face pairs are added into the non-match distribution for Caucasian face pairs. The red lines show this rate as Caucasian face pairs are added into the non-match distribution for Asian face pairs. As noted, the Caucasian to Asian condition approximates a case where a minority race gradually becomes better represented in the population until it reaches its maximum, in this case at 13.3 percent of the population.

Figure 5 shows that the effect of this gradual increase in representation of the minority race is a steady decline in the verification rate at the .001 false accept rate. The Asian-to-Caucasian condition approximates a case where a minority race gradually becomes better represented in the population until it reaches its equilibrium with the first race (50 percent diversity). Ultimately, Caucasians become the majority race (red lines on the graph). Two results are apparent from this simulation. First, in contrast to the Caucasian-to-Asian simulation, which shows a theoretically comparable situation up to 13 percent diversity, verification rate *increases* in this range. This suggests that the effect on verification rate in these situations is not a general result, but rather, is likely to depend both on the percentage diversity and on the algorithm itself, which may process faces of different races in different ways.

The second result is the gradual continued increase in verification rate as Caucasians become the majority race in the non-match distribution. Over the full range of diversity percentages, the verification rate at the .001 false alarm rate varies by as much as 0.12.

5. Demographically "Reversed" Identity Match and Mismatch Distributions

It is easy to imagine an application scenario in which the background distribution of non-match faces contains faces of one race, but the target population contains faces of another race. This might happen when the algorithm is developed in one geographic venue but is deployed in another venue. In that case, estimates of the similarity of match and non-match faces would be based on different races of faces. If the homogeneity of the two populations differs, one would expect variable estimates of algorithm performance depending upon which race is used to create the match versus mis-match distributions.

In this experiment, we explored the case where the population of face matches to be detected are of a different race than the population against which these matches will be compared. Specifically, we compared verification rate at the .001 false alarm rate with the following four combinations of race in the identity match and identity non-match distributions: a.) Caucasians in the match distribution and Caucasians in the non-match distribution (Caucasian-Caucasian); b.) Caucasians in the match distribution and Asians in the non-match distribution (Caucasian-Asian); c.) Asians in the match distribution and Asians in the non-match distribution (Asian-Asian); and d.) Asians in the match distribution and Caucasians in the non-match distribution (Asian-Caucasian).

Table 4 lists the verification rates at the .001 false accept rate for the four conditions. There was a sharp drop in verification rate from the Caucasian-Caucasian condition to the Caucasian-Asian case. This was true for both the moderately difficult and difficult face pairs. Surprisingly, the inverse pattern was found in comparing the Asian-Asian condition to the Asian-Caucasian case. In this latter case, the verification rate *increased* when the race of the non-matched identity distribution differed from the race of the matched identity.

An explanation of this might be found in the mean of the similarity distributions for the match and non-matched identity distributions or in the form

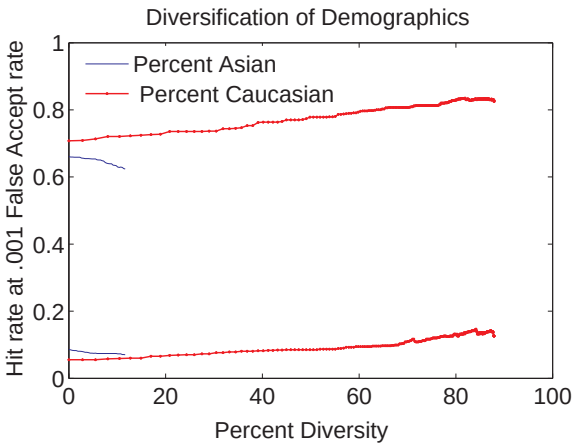

Figure 5: Verification rate at the 0.001 false accept rate is plotted as a function of the diversification of the the non-match distribution. When the non-match matrix begins with 100 percent Caucasians (blue), verification rate declines as the representation of the Asian minority race increases to its maximum of 13.3 percent. When the non-match matrix begins with 100 percent Asians in the matrix (red), the verification rate increases as the representation of the Caucasians increases, equals Asians, and reverses.

Table 4: Verification rate at the .001 False Alarm Rate with Reversed Demographics.

Match	Non-Match	Moderately Difficult	Difficult
Caucasian	Caucasian	.66	.08
Caucasian	Asian	.49	.03
Asian	Asian	.70	.06
Asian	Caucasian	.85	.16

of the distributions themselves. A careful inspection of our data suggested both factors at work. Differences in the separation of the various match and non-match distribution means explained some, but not all, of the verification rate data in Table 4. An explanation of the remaining differences must come, therefore, from subtle aspects of the shapes of the distributions themselves.

6. Imposter Distributions

In the experiments reported up to this point in the paper, we have considered what happens to estimates of algorithm performance as a function of changes to the demographic constraints of the background population of non-matched faces. Consequently, our results are due to the effects of these constraints on the structure of the similarity scores in the non-match distribution. In this final experiment, we modeled the scenario of deliberate attempts to impersonate other people. Analogous to the demographic manipulations, this too is likely to have profound effects of the structure of the similarity distributions. In these situations, we assume that a person who intends to impersonate someone else, chooses someone similar. In other words, if an individual steals an identity card, they are likely to search for a person with a similar face. To test the effects of this scenario on estimates of the performance of face recognition algorithms, we modeled person-specific non-match face pairs.

This analysis was carried out as follows. From a fixed population, one wants to find the best candidate to impersonate a specific subject, ℓ. For subject ℓ we have two images, $t_{\ell(1)}$ and $t_{\ell(2)}$. Subject ℓ is enrolled in a system using image $t_{\ell(1)}$. For this system, decisions are made by algorithm $\mathcal{A}$. We model the case where we know the person we want to impersonate, but do not have a copy of the image enrolled in the system. Instead, we have another image, $t_{\ell(2)}$, of subject ℓ. The next step selects an image to impersonate subject ℓ and compare with the enrolled image $t_{\ell(1)}$. The image that will impersonate the subject is selected from a query set $\mathcal{Q}_\ell$, where $\mathcal{Q}_\ell$ does not contain images of subject ℓ. The image selected $q(t_{\ell(2)})$, is the image that has the highest similarity score between $t_{\ell(2)}$ and the images in $\mathcal{Q}_\ell$. All similarity scores are computed by algorithm $\mathcal{A}$.

Given a target set of images $\mathcal{T}$ and a query set of images $\mathcal{Q}$, the set of non-match face pairs is computed by the following method. For subjects with only one image in the target set $\mathcal{T}$, there are no non-match face pairs. For subjects with two images, two non-match face pairs are generated. The

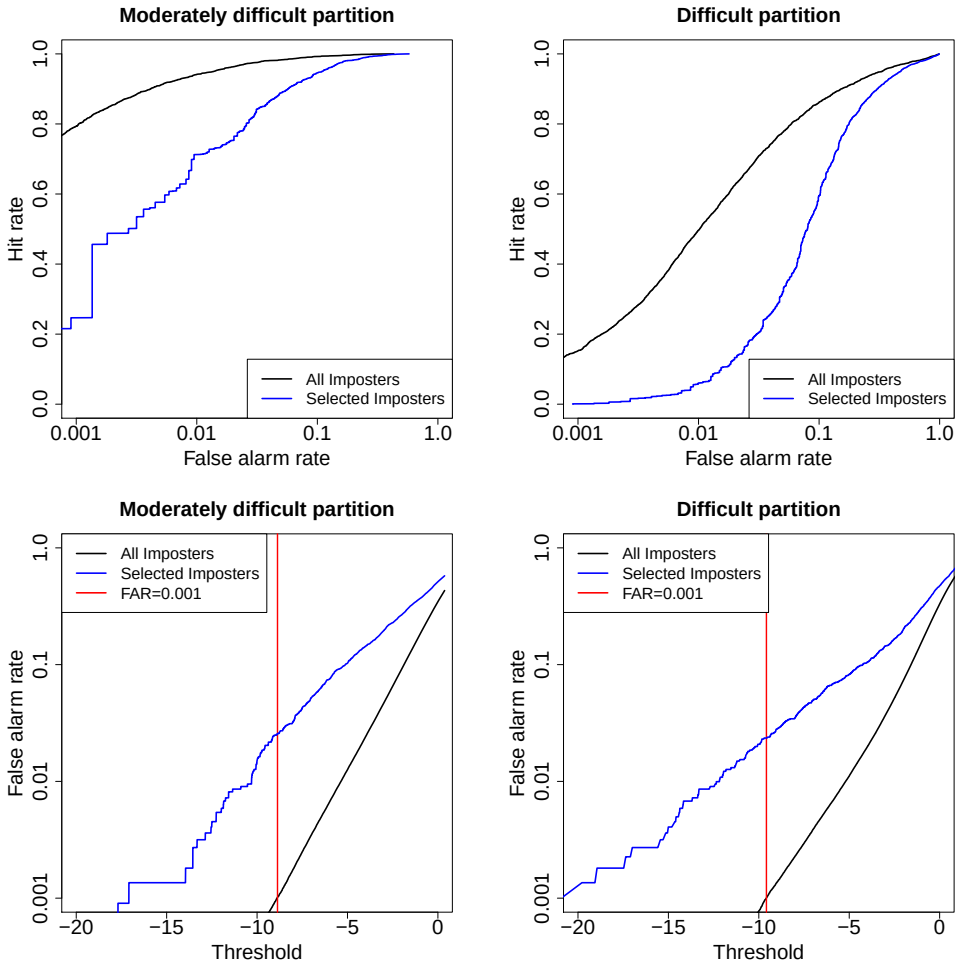

Figure 6: ROC curves in the top panel show the performance of the fusion algorithm for selected versus all imposters. As expected, performance estimates are substantially lower with the selected imposter distributions. The lower panel shows the shift in the similarity score cutoff to achieve particular false alarm rates.

first non-match face pair is generated by having the first image simulate the enrolled image ($t_{\ell(1)}$ in the previous paragraphs); the second image serves the roll of image $t_{\ell(2)}$ in the model. The second non-match face pair is generated by having the second image serving as the enrolled image and the first image serving as the image $t_{\ell(2)}$.

If there are three or more images of a subject in the target set $\mathcal{T}$, we repeat the process in the previous paragraph for all possible pairs of images of a subject. If there are n images of a subject, then there will be $n(n-1)$ non-match face pairs.

The results of this analysis appear in Figure 6. The ROC curves in the top panel show the performance of the fusion algorithm for selected versus all imposters. As expected, the performance estimates are substantially lower with the selected imposter distributions. The lower panel shows the shift in the similarity score cutoff to achieve particular false alarm rates. These results confirm and extend the results with demographic matching to the more challenging case of "imposters" chosen based on their similarity to a target.

7. Conclusions

In summary, all measures of the performance of face recognition algorithms rely both on the distribution of data for identity matches and on the distribution of data for mismatched identities. Traditionally, attempts to improve the performance of face recognition algorithms have emphasized methods that increase the degree of match between images of the same person (e.g., by bridging differences in illumination). Less consideration has been given to the effects of the composition of the non-match identity distributions in producing stable estimates of algorithm performance. These estimates are important for predicting how the algorithms will perform in real world environments. In this study, we show that differences in the treatment of demographic diversity in the non-match distribution can radically alter the estimates of algorithm performance.

The results of this study point to the following factors as determinants of performance. First, the demographic pairing of non-matched identity items can affect both the overall level of performance estimated and the choice of thresholds for match/non-match decisions. If no demographic pairing constraints are imposed, and if the database is diverse, the ability of algorithms to recognize unique identities will be over-estimated. This is because some

part of the performance will be based on face categorization (e.g., by gender or race) rather than identity discrimination. Second, we show systematic, but not general, effects on the verification rate at the .001 false alarm rate when a non-match identity distribution increases in racial diversity. Third, again using the measure of verification rate at the .001 false alarm rate, we demonstrate that comparisons of match and non-match distributions based on the same and different race of faces can lead to substantial differences in performance expectations. Finally, real world performance in some cases must anticipate the worst case scenario for which an imposter is chosen deliberately to resemble a particular target. Here, a realistic expectation of accuracy is well below the more general scenarios used to estimate algorithm performance.

The second and third findings are particularly troubling because they suggest that it may be difficult to reliably predict algorithm performance without good estimates of the way the match and non-match identity distributions are structured demographically. This poses a new and pressing challenge for this literature to find a method for tuning algorithm performance to the constantly changing demographic environments in which these systems must operate reliably. Again, returning to the context of an airport or tourist attraction, the choice of an appropriate non-match distribution and threshold may have to be re-assessed periodically and adjusted as needed. The present study offers the first quantitative evidence on the importance of considering the demographic "background" in setting performance expectations for face recognition systems in the field.

Acknowledgments

The authors wish to thank Technical Support Working Group (TSWG)/ DOD and the Federal Bureau of Investigation (FBI) for their support of this work. The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology or the The University of Texas at Dallas. The authors thank Jay Scallan for his assistance in preparing the GBU challenge problem and Allyson Rice for comments on a previous version of this manuscript.

[1] P. J. Phillips, W. T. Scruggs, A. J. O'Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, M. Sharpe, FRVT 2006 and ICE 2006 large-scale results, IEEE Trans. PAMI 32 (2010) 831–846.

[2] R. Gross, S. Baker, I. Matthews, T. Kanade, Face recognition across pose and illumination, in: S. Z. Li, A. K. Jain (Eds.), Handbook of Face Recognition, Springer, 2005, pp. 193–216.

[3] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, W. Worek, Overview of the face recognition grand challenge, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 947–954.

[4] P. J. Phillips, H. Moon, S. Rizvi, P. Rauss, The FERET evaluation methodology for face-recognition algorithms, IEEE Trans. PAMI 22 (2000) 1090–1104.

[5] P. J. Phillips, P. J. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, J. M. Bone, Face Recognition Vendor Test 2002: Evaluation Report, Technical Report NISTIR 6965, National Institute of Standards and Technology, 2003. Http://www.frvt.org.

[6] R. S. Malpass, J. Kravitz, Recognition for faces of own and other race faces, Journal of Personality and Social Psychology 13 (1969) 330–334.

[7] R. K. Bothwell, J. C. Brigham, R. S. Malpass, Cross-racial identification, Personality & Social Psychology Bulletin 15 (1989) 19–25.

[8] C. A. Meissner, J. C. Brigham, Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review, Psychology, Public Policy, and Law 7 (2001) 3–35.

[9] P. N. Shapiro, S. D. Penrod, Meta-analysis of face identification studies, Psychological Bulletin 100 (1986) 139–156.

[10] N. Furl, P. J. Phillips, A. J. O'Toole, Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis, Cognitve Science 26 (2002) 797–815.

[11] P. J. Phillips, F. Jiang, A. Narvekar, A. J. O'Toole, An other-race effect for face recognition algorithms, ACM Trans. Applied Perception 8 (2010).

[12] P. J. Phillips, J. R. Beveridge, B. A. Draper, G. Givens, A. J. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, S. Weimer, An Introduction to the Good, the Bad, and the Ugly Face Recognition Challenge Problem, in: Proceedings, Ninth International Conference on Automatic Face and Gesture Recognition.