Face Recognition Performance: Role of Demographic Information

Brendan F. Klare, *Member, IEEE,* Mark J. Burge, *Senior Member, IEEE,* Joshua C. Klontz, Richard W. Vorder Bruegge, and Anil K. Jain, *Fellow, IEEE*

***Abstract*—This paper studies the influence of demographics on the performance of face recognition algorithms. The recognition accuracies of six different face recognition algorithms (three commercial, two non-trainable, and one trainable) are computed on a large scale gallery that is partitioned so that each partition consists entirely of specific demographic cohorts. Eight total cohorts are isolated based on gender (male and female), race/ethnicity (Black, White, and Hispanic), and age group (18 to 30, 30 to 50, and 50 to 70 years old). Experimental results demonstrate that both commercial and the non-trainable algorithms consistently have lower matching accuracies on the same cohorts (females, Blacks, and age group 18 to 30). Additional experiments investigate the impact of the demographic distribution in the training set on the performance of a trainable face recognition algorithm. We show that the matching accuracy for race/ethnicity and age cohorts can be improved by training exclusively on that specific cohort. Operationally, this leads to a scenario, called dynamic face matcher selection, where multiple face recognition algorithms (each trained on different demographic cohorts), are available for a biometric system operator to select based on the demographic information extracted from a probe image. This procedure should lead to improved face recognition accuracy in many intelligence and law enforcement face recognition scenarios.**

***Index Terms*—face recognition, demographics, race/ethnicity, gender, age, training, dynamic face matcher selection**

I. Introduction

Sources of errors in automated face recognition algorithms are generally attributed to the well studied variations in pose, illumination, and expression [1], collectively known as PIE. Other factors such as image quality (e.g., resolution, compression, blur), time lapse (facial aging), and occlusion also contribute to face recognition errors [2]. Previous studies have also shown within a specific demographic group (e.g., race/ethnicity, gender, age) that certain cohorts are more susceptible to errors in the face matching process [3], [4]. However, there has yet to be a comprehensive study that investigates whether or not we can train face recognition algorithms to exploit knowledge regarding the demographic cohort of a probe subject.

This study presents a large scale analysis of face recognition performance on three different demographics (see Figure 1): (i) race/ethnicity, (ii) gender, and (iii) age. For each of these demographics, we study the performance of six face recognition algorithms belonging to three different types of systems: (i) three commercial off the shelf (COTS) face recognition systems (FRS), (ii) face recognition algorithms that do not utilize training data, and (iii) a trainable face recognition algorithm. While the COTS FRS algorithms leverage training data, we are not able to re-train these algorithms; instead they are black box systems that output a measure of similarity between a pair of face images. The non-trainable algorithms use common feature representations to characterize face images, and similarities are measured within these feature spaces. The trainable face recognition algorithm used in this study also outputs a measure of similarity between a pair of face images. However, different versions of this algorithm can be generated by training it with different sets of face images, where the sets have been separated based on demographics. Both the trainable algorithms, and (presumably) the COTS FRS, initially use some variant of the non-trainable representations.

The study of COTS FRS performance on each of the demographics considered is intended to augment previous experiments [3], [4] on whether these algorithms, as used in government and other applications, exhibit biases. Such biases would cause the performance of commercial algorithms to vary across demographic cohorts. In evaluating three different COTS FRS, we confirmed that not only do these algorithms perform worse on certain demographic cohorts, they consistently perform worse on the same cohorts (females, Blacks, and younger subjects).

Even though biases of COTS FRS on various cohorts were observed in this study, these algorithms are black boxes that offer little insight into to why such errors manifest on specific demographic cohorts. To understand this, we also study the performance of non-commercial trainable and non-trainable face recognition algorithms, and whether statistical learning methods can leverage this phenomenon.

By studying non-trainable face recognition algorithms, we gain an understanding of whether or not the errors are inherent to the specific demographics. This is because non-trainable algorithms operate by measuring the (dis)similarity of face images based on a specific feature representation that, ideally, encodes the structure and shape of the face. This similarity is measured independent of any knowledge of how face images vary for the same subject and between different subjects. Thus, cases in which the non-trainable algorithms have the same relative performance within a demographic group as the COTS FRS indicates that the errors are likely due to one of the

B.F. Klare, M.J. Burge, and J. Klontz are with The MITRE Corporation, McLean, VA, U.S.A.

R.W. Vorder Bruegge is with the Science and Technology Branch, Federal Bureau of Investigation, Quantico, VA, U.S.A.

A.K. Jain and B.F. Klare are with the Dept. of Computer Science and Engineering, Michigan State University, East Lansing, MI, U.S.A.

A.K. Jain is also with the Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Korea

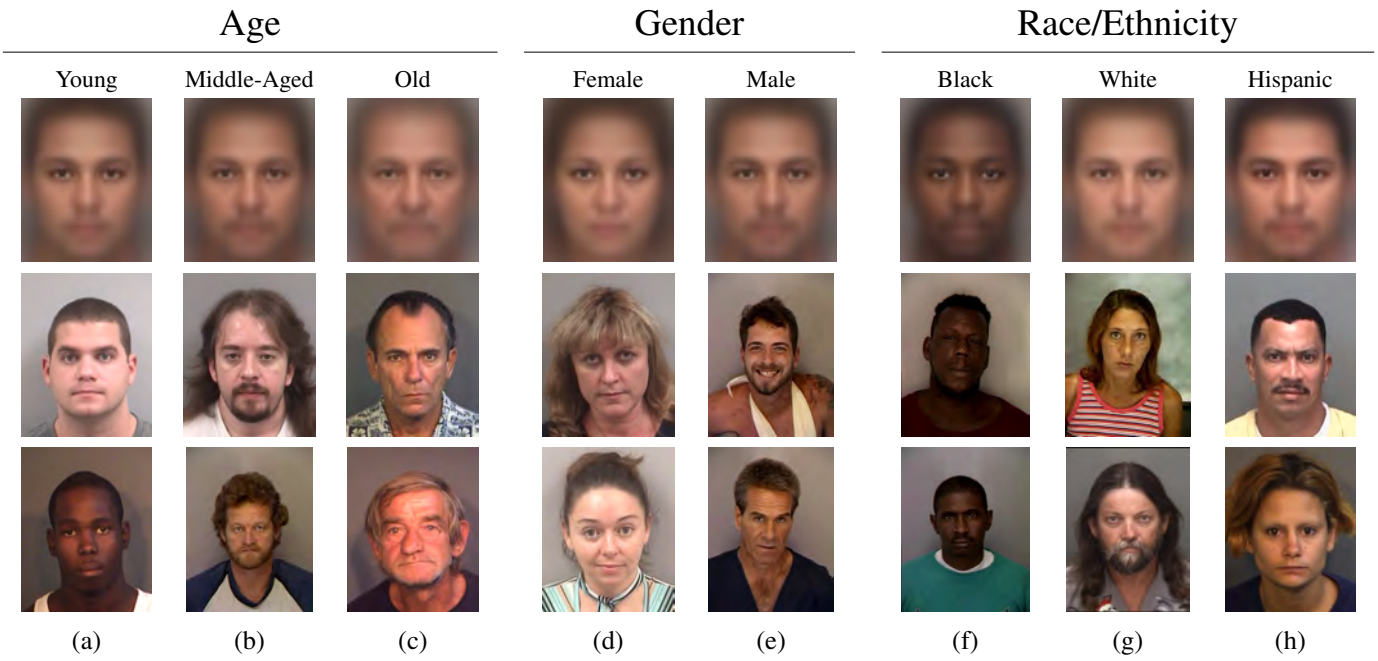


Fig. 1. Examples of the different demographics studied. (a-c) Age demographic. (d-e) Gender demographic. (f-h) Race/ethnicity demographic. Within each demographic, the following cohorts were isolated: (a) ages 18 to 30, (b) ages 30 to 50, (c) ages 50 to 70, (d) female gender, (e) male gender, (f) Black race, (g) White race, and (h) Hispanic ethnicity. The first row shows the "mean face" for each cohort. A "mean face" is the average pixel value computed from all the aligned face images in a cohort. The second and third rows show different sample images within the cohorts.

cohorts being inherently more difficult to recognize.

Relative differences in performance between the non-trainable algorithms and the COTS FRS indicate that the lower performance of COTS FRS on a particular cohort may be due to imbalanced training of the COTS algorithm. We explore this hypothesis by training the Spectrally Sampled Structural Subspace Features (4SF) face recognition algorithm [5] (i.e., the trainable face recognition algorithm used in this study) on image sets that consist exclusively of a particular cohort (e.g., White only). The learned subspaces in 4SF are applied to test sets from different cohorts to understand how unbalanced training with respect to a particular demographic impacts face recognition accuracy.

The 4SF trained subspaces also help answer the following question: to what extent can statistical learning improve accuracy on a demographic cohort? For example, it will be shown that females are more difficult to recognize than males. We will investigate how much training on only females, for example, can improve face recognition accuracy when matching females. Such improvements suggest the use of multiple discriminative subspaces (or face recognition algorithms), with each trained exclusively on different cohorts. The results of these experiments indicate we can improve face recognition performance on the race/ethnicity cohort by using an algorithm trained exclusively on different demographic cohorts. This finding leads to the notion of *dynamic face matcher selection*, where demographic information may be submitted in conjunction with a probe image in order to select the face matcher trained on the same cohort. This framework, illustrated in Figure 2, should lead to improved face recognition accuracies.

The remainder of this paper is organized as follows. In Section II we discuss previous studies on demographic introduced biases in face recognition algorithms and the design of face recognition algorithms. Section III discusses the data corpus that was utilized in this study. Section IV identifies the different face recognition algorithms that were used in this study (commercial systems, trainable and non-trainable algorithms). Section V describes the matching experiments conducted on each demographic. Section VI provides analysis of the results in each experiment and summarizes the contributions of this paper.

II. Prior Studies and Related Work

Over the last twenty years the National Institute of Standards and Technology (NIST) has run a series of evaluations to quantify the performance of automated face recognition algorithms. Under certain imaging constraints these tests have measured a relative improvement of over two orders of magnitude in performance over the last two decades [4]. Despite these improvements, there are still many factors known to degrade face recognition performance (e.g., PIE, image quality, aging). In order to maximize the potential benefit of face recognition in forensics and law enforcement applications, we need to improve the ability of face recognition to sort through facial images more accurately and in a manner that will allow us to perform more specialized or targeted searches. Facial searches leveraging demographics represents one such avenue for performance improvement.

While there is no standard approach to automated face recognition, most face recognition algorithms follow a similar pipeline [6]: face detection, alignment, appearance normalization, feature representation (e.g., local binary patterns [7], Gabor features [8]), feature extraction [9], [10]), and matching [11]. Feature extraction generally relies on an offline training

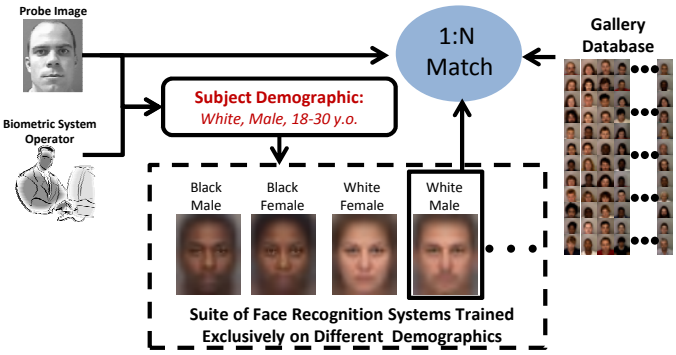


Fig. 2. Dynamic face matcher selection. The findings in this study suggest that many face recognition scenarios may benefit from multiple face recognition systems that are trained exclusively on different demographic cohorts. Demographic information extracted from a probe image may be used to select the appropriate matcher, and improve face recognition accuracy.

stage that utilizes exemplar data to learn improved feature combinations (such as feature subspaces). For example, variants of the linear discriminant analysis (LDA) algorithm [9], [10] use training data to compute between-class and within-class scatter matrices. Subspace projections are then computed to maximize the separability of subjects based on these scatter matrices.

This study examines the impact of training on face recognition performance. Without leveraging training data, face recognition algorithms are not able to discern between noisy facial features and facial features which offer consistent cues to a subject's identity. As such, automated face recognition algorithms are ultimately based on statistical models of the variance between individual faces. These algorithms seek to minimize the measured distance between facial images of the same subject, while maximizing the distance between the subject's images and those of the rest of the population. However, the feature combinations discovered are functions of the data used to train the recognition system. If the training set is not representative of the data a face recognition algorithm will be operating on, then the performance of the resulting system may deteriorate. For example, the most distinguishing features for Black subjects may differ from White subjects. As such, if a system was predominantly trained on White faces, and later operated on Black faces, the learned representation may discard information useful for discerning Black faces.

The observation that the performance of face recognition algorithms could suffer if the training data is not representative of the test data is not new. One of the earliest studies reporting this phenomenon is not in the automated face recognition literature, but instead in the context of human face recognition. Coined the "other-race effect", humans have consistently demonstrated a decreased ability to recognize subjects from races different from their own [12], [13]. While there is no generally agreed upon explanation for this phenomenon, many researchers believe the decreased performance on other races is explained by the "contact" hypothesis, which postulates that the lower performance on other races is due to a decreased exposure [14]. While the validity of the contact hypothesis has been disputed [15], the presence of the "other-race effect" has not.

From the perspective of automated face recognition, Phillips et als findings in the 2002 government sponsored NIST Face Recognition Vendor Test (FRVT) is believed to be the first finding that face recognition algorithms have different recognition accuracies depending on a subject's demographic cohort [3]. Among other findings, this study demonstrated for commercial face recognition algorithms on a dataset containing roughly 120,000 images that (i) female subjects were more difficult to recognize than male subjects, and (ii) younger subjects were generally more difficult to recognize than older subjects.

More recently, Grother et als measured the performance of seven commercial face recognition algorithms and three university face recognition algorithms in the 2010 NIST Multi-Biometric Evaluation [4]. The experiments conducted also concluded that females were more difficult to recognize than males. This study also measured the recognition accuracy of different races and ages.

Previous studies have investigated what impact the distribution of a training set has on recognition accuracy. Furl et als [16] and O'Toole et als [17] conducted studies to investigate the impact of cross training and matching on White and Asian races. Similar training biases were investigated by Klare and Jain [18], who showed that aging-invariant face recognition algorithms suffer from decreased performance in non-aging scenarios.

The study in [17] was motivated by a rather surprising result in the 2006 NIST Face Recognition Vendor Test (FRVT) [19]. In this test, the various commercial and academic face recognition algorithms tested exhibited a common characteristic: algorithms which originated in East Asia performed better on Asian subjects than did algorithms from the West. The reverse was true for White subjects: algorithms developed in the western hemisphere performed better. O'Toole et als suggested that this discrepancy was due to the different racial distribution in the training sets for the Western and Asian algorithms.

The impact of these training sets on face recognition algorithms cannot be overemphasized; face recognition algorithms do not generally rely upon explicit physiological models of the human face for determining match or non-match. Instead, the measure of similarity between face images is based on statistical learning, generally in the feature extraction stage [10], [20] or during the matching stage [11].

In this work, we expand on previous studies to better demonstrate and understand the impact of a training set on the performance of face recognition algorithms. While previous studies [16], [17] only isolated the race variate, and only considered two races (i.e., Asian and White), this study explores both the inherent biases and training biases across gender, race (three different races/ethnicities) and age. To our knowledge, no studies have investigated the impact of gender or subject age for training face recognition algorithms.

III. Face Database

This study was enabled by a collection of over one million mug shot face images from the Pinellas County Sheriff's

TABLE I
NUMBER OF SUBJECTS USED FOR TRAINING AND TESTING FOR EACH DEMOGRAPHIC CATEGORY. TWO IMAGES PER SUBJECT WERE USED. TRAINING AND TEST SETS WERE DISJOINT. A TOTAL OF 102,942 FACE IMAGES WERE USED IN THIS STUDY.

Demographic	Cohort	# Training	# Testing
Gender	Female	7995	7996
	Male	7996	7998
Race	Black	7993	7992
	White	7997	8000
	Hispanic	1384	1425
Age	18 to 30	7998	7999
	30 to 50	7995	7997
	50 to 70	2801	2853

Office[1] (examples of these images can be found in Figure 1). Accompanying these images are complete subject demographics. The demographics provide the race/ethnicity, gender, and age of the subject in each image, as well as a subject ID number.

Given this large corpus of face images, we were able to use the metadata provided to control the three demographics studied: race/ethnicity, gender, and age. For gender, we partitioned image sets into cohorts of (i) male only, and (ii) female only. For age, we partitioned the sets into three cohorts: (i) young (18 to 30 years old), (ii) middle-age (30 to 50 years old), and (iii) old (50 to 70 years old). There were very few individuals in this database with age less than 18 and older than 70. For race/ethnicity[2], we partitioned the sets into cohorts of (i) White, (ii) Black, and (iii) Hispanic[3]. A summary of these cohorts and the number of subjects available for each cohort can be found in Table I. Asian, Indian, and Unknown race/ethnicities were not considered because an insufficient number of samples were available.

For each of the eight cohorts (i.e., male, female, young, middle-aged, old, White, Black, and Hispanic), we created independent training and test sets of face images. Each set contains a maximum of 8,000 subjects, with two images (one probe and one gallery) for each subject. Table I lists the number of subjects included for each set. Cohorts far less than 8,000 subjects (i.e., Hispanic and older) reflect a lack of data available to us. Cases with cohorts containing only slightly fewer than 8,000 subjects are the result of removing a few images that could not be successfully enrolled in the COTS FRS.

The dataset of mug shot images did not contain a large enough number of Asian subjects to measure that particular race/ethnicity cohort. However, studies by Furl et al. [16] and O'Toole et al. [17] investigated the impact of the Whites and East Asians. As previously discussed, these studies concluded that algorithms developed in the Western Hemisphere did better on White subjects and Asian algorithms did better on Asian subjects.

[1]The mug shot data used in this study was acquired in the public domain through Florida's "Sunshine" laws. Subjects shown in this manuscript may or may not have been convicted of a criminal charge, and thus should be presumed innocent of any wrongdoing.

[2]Racial identifiers (i.e. White, Black, and Hispanic) follow the FBI's National Crime Information Center code manual.

[3]Hispanic is not technically a race, but instead an ethnic category.

IV. FACE RECOGNITION ALGORITHMS

In this section we will discuss each of the six face recognition algorithms used in this study. We have organized these algorithms into commercial algorithms (Sec. IV-A), non-trainable algorithms (Sec. IV-B), and trainable algorithms (Sec. IV-C).

A. Commercial Face Recognition Algorithms

Three commercial face recognition algorithms were evaluated in this study: (i) Cognitec's FaceVACS v8.2, (ii) PittPatt v5.2.2, and (iii) Neurotechnology's MegaMatcher v3.1. The results in this study obfuscate the names of the three commercial matchers.

These commercial algorithms are three of the ten algorithms evaluated in the NIST sponsored Multi-Biometrics Evaluation (MBE) [4]. As such, these algorithms are representative of the state of the art performance in face recognition technology.

B. Non-Trainable Face Recognition Algorithms

Two non-trainable face recognition algorithms were used in this study: (i) local binary patterns (LBP), and (ii) Gabor features. Both of these methods operate by representing the face with Level 2 facial features (LBP and Gabor), where Level 2 facial features are features that encode the structure and shape of the face, and are critical to face recognition algorithms [21].

These non-trainable algorithms perform an initial geometric normalization step (also referred to as alignment) by using the automatically detected eye coordinates (eyes were detected using FaceVACS SDK) to scale, rotate, and crop a face image. After this step, the face image has a height and width of 128 pixels. Both algorithms are custom implementations by the authors.

1) Local Binary Patterns: A seminal method in face recognition is the use of local binary patterns [7] (LBP) to represent the face [22]. Local Binary Patterns are Level 2 features that represent small patches across the face with histograms of binary patterns that encode the structure and texture of the face.

Local binary patterns describe each pixel using a p-bit binary number. Each bit is determined by sampling p pixel values at uniformly spaced locations along a circle of radius r, centered at the pixel being described. For each sampling location, the corresponding bit receives the value 1 if it is greater than or equal to the center pixel, and 0 otherwise.

A special case of LBP, called the uniform LBP [7], is generally used in face recognition. Uniform LBP assigns any non-uniform binary number to the same value, where uniformity is defined by whether more than u transitions between the values 0 and 1 occur in the binary number. In

the case of $p = 8$ and $u = 2$, the uniform LBP has 58 uniform binary numbers, and the 59th value is reserved for the remaining $256 - 58 = 198$ non-uniform binary numbers. Thus, each pixel will take on a value ranging from 1 to 59. Two different radii are used ($r = 1$ and $r = 2$), resulting in two different local binary pattern representations that are subsequently concatenated together (called Multi-scale Local Binary Patterns, or MLBP).

In the context of face recognition, LBP values are first computed at each pixel in the (normalized) face image as previously described. The image is tessellated into patches with a height and width of 12 pixels. For each patch i, a histogram of the LBP values $S_i' \in \mathbb{Z}^{d_s}$ is computed (where $d_s = 59$). This feature vector is then normalized to the feature vector $S_i \in \mathbb{R}^{d_s}$ by $S_i = \frac{S_i'}{\sum_i^{d_s} S_i'}$. Finally, we concatenate the N vectors into a single vector x of dimensionality $d_s \cdot N$.

In our implementation, the illumination filter proposed by Tan and Triggs [23] is used prior to computing the LBP codes in order to suppress non-uniform illumination variations. This filter resulted in improved recognition performance.

2) Gabor Features: Gabor features are one of the first Level 2 facial features [21] to have been used with wide success in representing facial images [8], [20], [24]. One reason Gabor features are popular for representing both facial and natural images is their similarity with human neurological receptor fields [25], [26].

A Gabor image representation is computed by convolving a set of Gabor filters with an image (in this case, a face image). The Gabor filters are defined as

$$G(x, y, \theta, \eta, \gamma, f) = \frac{f^2}{\pi\gamma\eta} e^{-\left(\frac{f^2}{\gamma^2}x'^2 + \frac{f^2}{\gamma^2}y'^2\right)} e^{\left(j2\pi f x'\right)} \quad (1)$$

$$x' = x\cos\theta + y sin\theta \quad (2)$$

$$y' = -x sin\theta + y\cos\theta \quad (3)$$

where f sets the filter scale (or frequency), θ is the filter orientation along the major axis, γ controls the filter sharpness along the major axis, and η controls the sharpness along the minor axis. Typically, combinations across the following values for the scale f and orientation θ are used: $f = \{0, 1, \ldots, 4\}$ and $\theta = \{\pi/8, \pi/4, 3\pi/8, \ldots, \pi\}$. This creates a set (or bank) of filters with different scales and orientations. Given the bank of Gabor filters, the input image is convolved with each filter, which results in a Gabor image for each filter. The combination of these scale and orientation values results in 40 different Gabor filters, which in turn results in 40 Gabor images (for example).

In this paper, the recognition experiments using a Gabor image representation operate by: (i) performing illumination correction using the method proposed by Tan and Triggs [23], (ii) computing the phase response of the Gabor images with $f = \{1, 2\}$, and $\theta = 0, \pi/4, \pi/2, 3\pi/4$, (iii) tessellating the Gabor image(s) into patches of size 12x12, (iv) quantizing the phase response (which ranges from 0 to 2π) into 24 values and computing the histogram within each patch, and (v) concatenating the histogram vectors into a single feature vector. Given two (aligned) face images, the distance between their corresponding Gabor feature vectors is used to measure the dissimilarity between the two face images.

C. Trainable Face Recognition Algorithm

The trainable algorithm used in this study is the Spectrally Sampled Structural Subspace Features algorithm [5], which is abbreviated as 4SF@. This algorithm uses multiple discriminative subspaces to perform face recognition. After geometric normalization of a face image using the automatically detected eye coordinates (eyes were detected using FaceVACS SDK), illumination correction is performed using the illumination correction filter presented by Tan and Triggs [23]. Face images are then represented using histograms of local binary patterns at densely sampled face patches [22] (to this point, 4SF is the same as the non-trainable LBP algorithm described in Sec. IV-B1). For each face patch, principal component analysis (PCA) is performed so that 98.0% of the variance is retained. Given a training set of subjects, multiple stages of weighted random sampling is performed, where the spectral densities (i.e., the eigenvalues) from each face patch are used for weighting. The randomly sampled subspaces are based on Ho's original method [27], however the proposed approach is unique in that the sampling is weighted based on the spectral densities. For each stage of random sampling, LDA [10] is performed on the randomly sampled components. The LDA subspaces are learned using subjects randomly sampled from the training set (i.e., bagging [28]). Finally, distance-based recognition is performed by projecting the LBP representation of face images into the per-patch PCA subspaces, and then into each of the learned LDA subspaces. The sum of the Euclidean distance in each subspace is the dissimilarity between two face images. The 4SF algorithm is summarized in Figure 3.

As shown in the experiments conducted in this study, the 4SF algorithm performs on par with several commercial face recognition algorithms. Because 4SF is initially the same approach as the non-trainable LBP matcher, the improvement in recognition accuracies (in this study) between the non-trainable LBP matcher and the 4SF algorithm clearly demonstrates the ability of 4SF to leverage training data. Thus, a high matching accuracy and the ability to leverage training data make 4SF an ideal face recognition algorithm to study the effects of training data on face recognition performance. The 4SF algorithm was developed in house.

V. Experimental Results

For each demographic (gender, race/ethnicity, and age), three separate matching experiments are conducted. The results of these experiments are presented per demographic. Figure 4 delineates the results for all the experiments on the gender demographic. Figure 5 delineates the results for all experiments on the race/ethnicity demographic. Finally, Figure 6 delineates the results for all experiments on the age demographic. The true accept rate at a fixed false accept rate of 0.1% for all the plots in Figures 4 to 6 are summarized in Table II.

The first experiment conducted on each demographic measures the relative performance within the demographic cohort

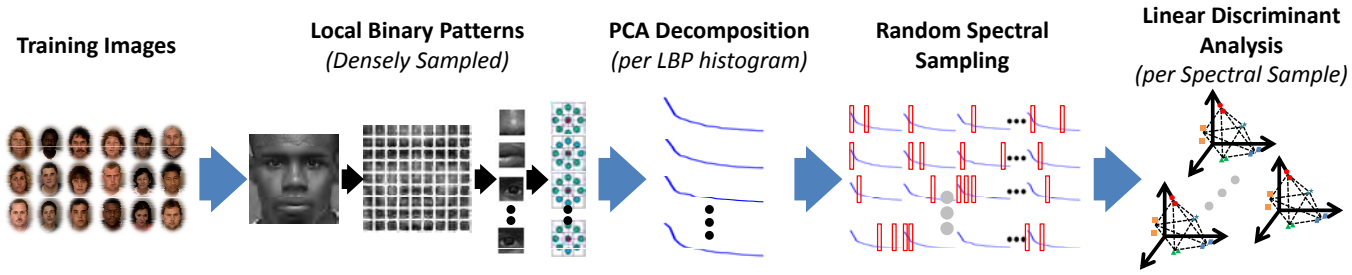


Fig. 3. Overview of the Spectrally Sampled Structural Subspace Features (4SF) algorithm. This custom algorithm is representative of state of the art methods in face recognition. By changing the demographic distribution of the training sets input into the 4SF algorithm, we are able to analyze the impact the training distribution has on various demographic cohorts.

for each COTS FRS@. That is, for a particular commercial matcher (e.g., COTS-A), we compare it's matching accuracy on each cohort within that demographic. For example, on the gender demographic, this experiment will measure the difference in recognition accuracy for commercial matchers on males versus females. The results from this set of experiments can be found in Figures 4 (a-c) for the gender demographic, Figures 5 (a-c) for the race/ethnicity demographic, and Figures 6 (a-c) for the age demographic.

The second experiment conducted on each demographic cohort measures the relative performance within the cohort for non-trainable face recognition algorithms. Because the non-trainable algorithms do not leverage statistical variability in faces, they are not susceptible to training biases. Instead, they reflect the inherent (or a priori) difficulty in recognizing cohorts of subjects within a specific demographic group. The results from this set of experiments can be found in Figures 4 (d-e) for the gender demographic, Figures 5 (d-e) for the race/ethnicity demographic, and Figures 6 (d-e) for the age demographic.

The final experiment investigates the influence of the training set on recognition performance. Within each demographic cohort, we train several versions of the 4SF algorithm (one for each cohort). These differently trained versions of the 4SF algorithm are then applied to separate testing sets from each cohort within the particular demographic. This enables us to understand within the gender demographic (for example), how much training exclusively on females (i) improves performance on females, and (ii) decreases performance on males. In addition to training 4SF exclusively on each cohort, we also use a version of 4SF trained on an equal representation of specific demographic cohorts (referred to as "Trained on All"). For example, in the gender demographic, this would mean that for "All", 4SF was trained on 4,000 male subjects and 4,000 female subjects. The results from this set of experiments can be found in Figures 4 (f-h) for the gender demographic, Figures 5 (f-i) for the race/ethnicity demographic, and Figures 6 (f-i) for the age demographic.

VI. Analysis

In this section we provide an analysis of the findings of the experiments described in Section V. A strength of this study is the large face dataset leveraged; accuracies measured on each cohort (except Hispanic and Old cohorts) are from roughly 8,000 subjects.

A. Gender

Each of the three commercial face recognition algorithms performed significantly worse on the female cohort than the male cohort (see Figures 4 (a-c)). Additionally, both non-trainable algorithms (LBP and Gabor) performed significantly worse on females (see Figures 4 (d-e)).

The agreement in relative accuracies of the COTS FRS and the non-trainable LBP method on the gender demographic suggests that the female cohort is more difficult to recognize using frontal face images than the male cohort. That is, if the results in the COTS algorithms were due to imbalanced training sets (i.e., training on more males than females), then the LBP matcher should have yielded similar matching accuracies on males and females. Instead, the non-trained LBP and Gabor matchers performed worse on the female cohort. When training on males and females equally (Figure 4(h)), the 4SF algorithm also did significantly worse on the female cohort. Together, *these results strongly suggest that the female cohort is inherently more difficult to recognize*.

The results of the 4SF algorithm on the female cohort (Figure 4 (f)) offer additional evidence about the nature of the discrepancy. The performance of training on only females is not higher than the performance of training on a mix of males and females (labeled "All"). Further, the difference in performance when training on only males versus training on only females is much lower than the difference in performance between males and females on the non-trainable algorithm. In other words, the difficulty in recognizing females seems to be due to a higher ratio of inter-class variance to intra-class variance in the initial face image representations.

Different factors may explain why females appear more difficult to recognize than males. One explanation may be that because females often use cosmetics (i.e., makeup), and males generally do not, there is a higher within-class variance in females. This hypothesis is supported by the match score distributions for males and females (see Figure 7). A greater difference in the true match distributions is noticed when compared to the false match distributions. The increased dissimilarities between images of the same female subjects demonstrate intra-class variability. Again, a cause of this may be due to the application of cosmetics.

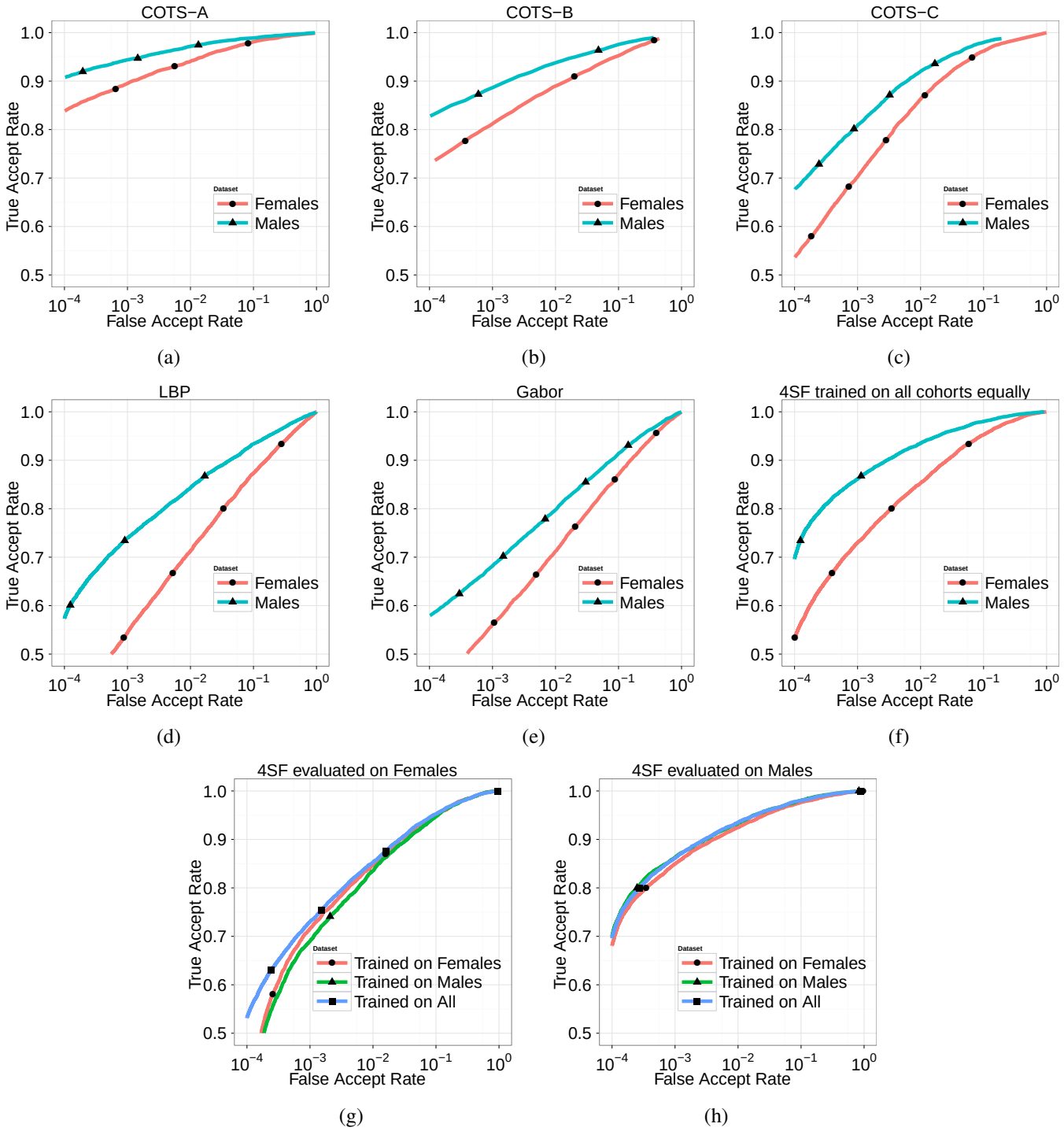


Fig. 4. Performance of the six face recognition systems on datasets seperated by cohorts within the gender demographic. (a) COTS-A, (b) COTS-B, (c) COTS-C, (d) Local binary patterns (non-trainable), (e) Gabor (non-trainable), (f) 4SF trained on equal number of samples from each gender, (g) 4SF algorirthm (trainable) on the Females cohort, (h) 4SF algorirthm (trainable) on the Males cohort.

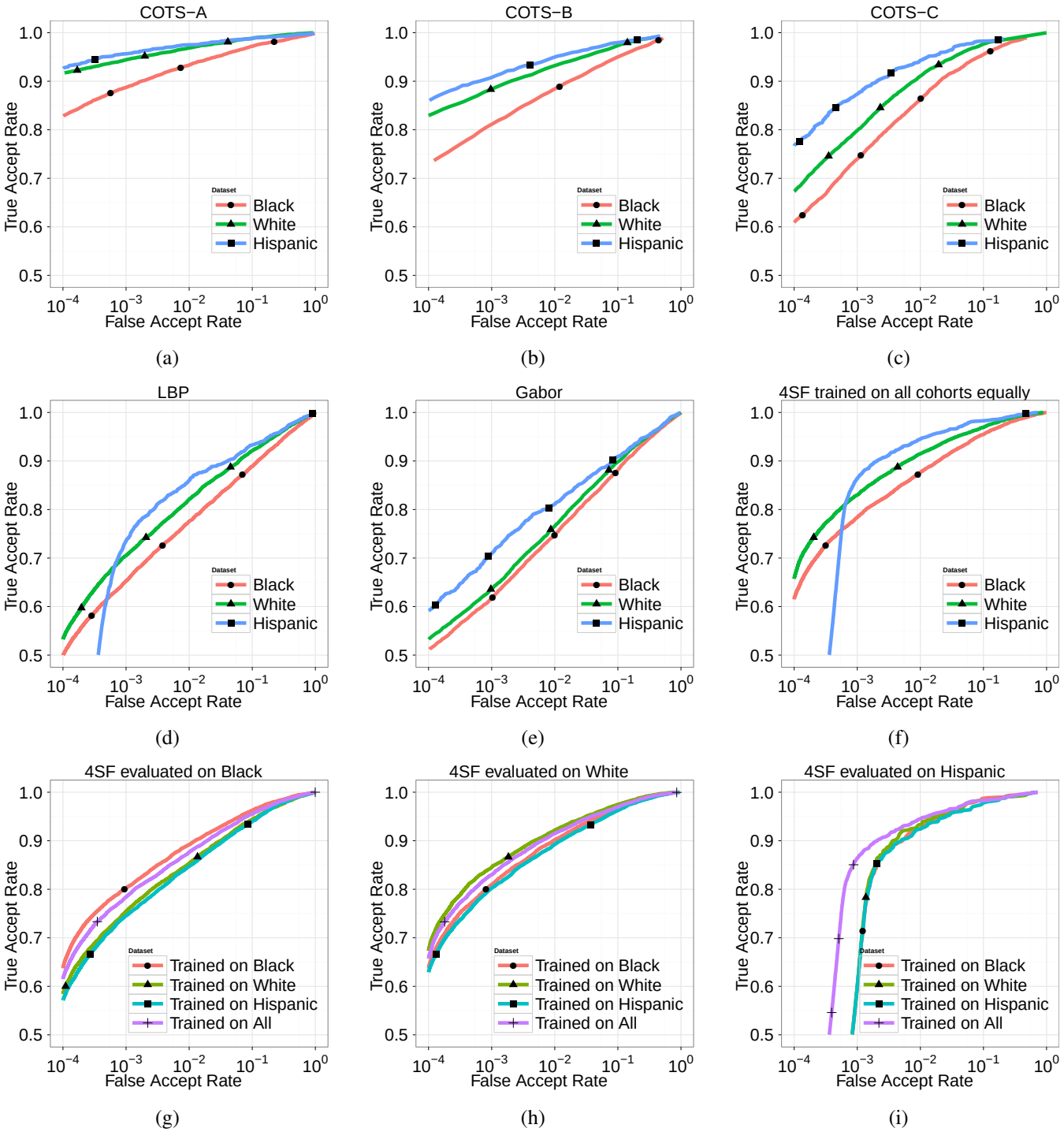


Fig. 5. Performance of the six face recognition systems on datasets seperated by cohorts within the race demographic. (a) COTS-A, (b) COTS-B, (c) COTS-C, (d) Local binary patterns (non-trainable), (e) Gabor (non-trainable), (f) 4SF trained on equal number of samples from each race, (g) 4SF algorirthm (trainable) on the Black cohort, (h) 4SF algorirthm (trainable) on the White cohort, (i) 4SF algorirthm (trainable) on the Hispanic cohort.

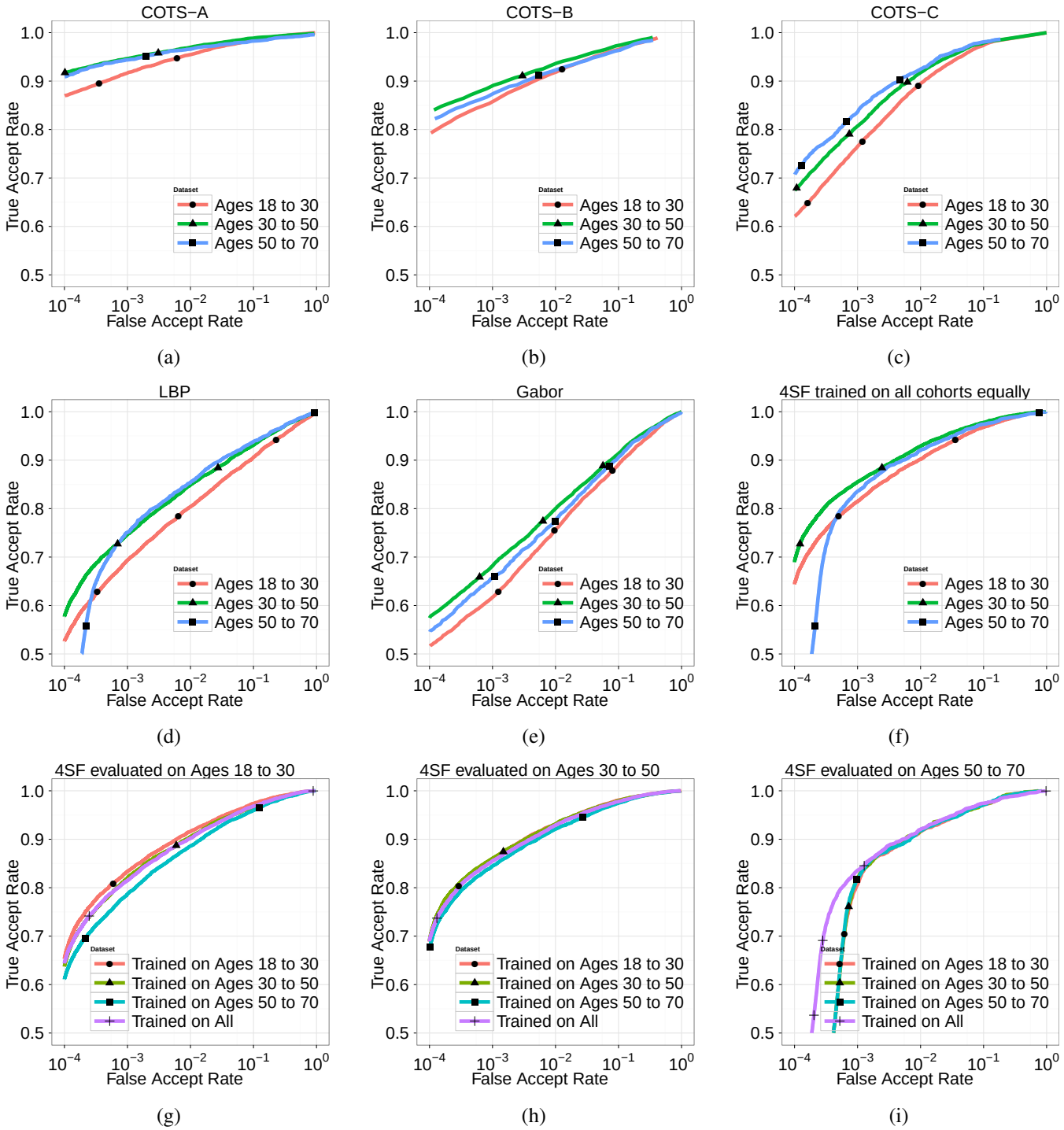


Fig. 6. Performance of the six face recognition systems on datasets seperated by cohorts within the age demographic. (a) COTS-A, (b) COTS-B, (c) COTS-C, (d) Local binary patterns (non-trainable), (e) Gabor (non-trainable), (f) 4SF trained on equal number of samples from each age, (g) 4SF algorirthm (trainable) on the Ages 18 to 30 cohort, (h) 4SF algorirthm (trainable) on the Ages 30 to 50 cohort, (i) 4SF algorirthm (trainable) on the Ages 50 to 70 cohort.

TABLE II
LISTED ARE THE TRUE ACCEPT RATES AT A FIXED FALSE ACCEPT RATE OF 0.1% FOR EACH MATCHER AND DEMOGRAPHIC DATASET.

	Females	Males
COTS-A	89.5	94.4
COTS-B	81.6	89.3
COTS-C	70.3	80.9
LBP	54.4	74.0
Gabor	56.0	68.2
4SF trained on All	73.0	86.2
4SF trained on Females	71.5	85.0
4SF trained on Males	69.0	86.3

	Black	White	Hispanic
COTS-A	88.7	94.4	95.7
COTS-B	81.3	89.0	90.7
COTS-C	74.0	79.8	87.3
LBP	65.3	70.5	73.5
Gabor	61.6	63.7	70.9
4SF trained on All	78.4	83.0	86.3
4SF trained on Black	80.2	81.0	59.8
4SF trained on White	75.4	84.5	59.9
4SF trained on Hispanic	74.5	80.2	60.1

	18 to 30 y.o.	30 to 50 y.o.	50 to 70 y.o.
COTS-A	91.7	94.6	94.4
COTS-B	86.1	89.1	87.5
COTS-C	76.5	80.7	83.6
LBP	69.4	74.7	75.1
Gabor	61.7	68.2	65.7
4SF trained on All	81.5	85.6	83.6
4SF trained on 18 to 30 y.o.	83.3	85.9	80.7
4SF trained on 30 to 50 y.o.	82.1	86.0	82.2
4SF trained on 50 to 70 y.o.	78.7	84.5	82.0

B. Race

When examining the race/ethnicity cohort, all three commercial face recognition algorithms achieved the lowest matching accuracy on the Black cohort (see Figures 5(a-c)). The two non-trained algorithms had similar results (Figures 5 (d-e)).

When matching against only Black subjects (Figure 5 (f)), 4SF has higher accuracy when trained exclusively on Black subjects (about a 5% improvement over the system trained on Whites and Hispanics only). Similarly, when evaluating 4SF on only White subjects (Figure 5 (g)), the system trained on only the White cohort had the highest accuracy. However, when comparing the 4SF algorithm trained equally on all race/ethnicity cohorts (Figure 5 (i)), we see that the performance on the Black cohort is still lower than on the White cohort. Thus, even with balanced training, the Black cohort still is more difficult to recognize.

The key finding in the training results shown in Figures 5 (f-i) is the ability to improve recognition accuracy by training exclusively on subjects of the same race/ethnicity. Compared to balanced training (i.e., training on "All"), the performance of 4SF when trained on the same race/ethnicity it is recognizing is higher. Thus, by merely changing the distribution of the training set, we can improve the recognition rate by nearly 2% on the Black cohort and 1.5% on the White cohort (see Table II).

The inability to effectively train on the Hispanic cohort is likely due to the insufficient number of training samples available for this cohort. However, the biogeographic ancestry of the Hispanic ethnicity is generally attributed to a three-way admixture of Native American, European, and West Black populations [29]. Even with an increased number of training samples, we believe this mixture of races would limit the ability to improve recognition accuracy through race/ethnicity specific training.

C. Age Demographic

All three commercial algorithms had the lowest matching accuracy on subjects grouped in the ages 18 to 30 (see Figures 6 (a-c)). The COTS-A matcher performed nearly the same on the 30 to 50 year old cohort as the 50 to 70 year old cohort. However, COTS-B had slightly higher accuracy on 30 to 50 age group than 50 to 70 age group, while COTS-C performed slightly better on 50 to 70 than 30 to 50 age groups.

The non-trainable algorithms (Figures 6 (d-e)) also performed the worst on the 18 to 30 age cohort.

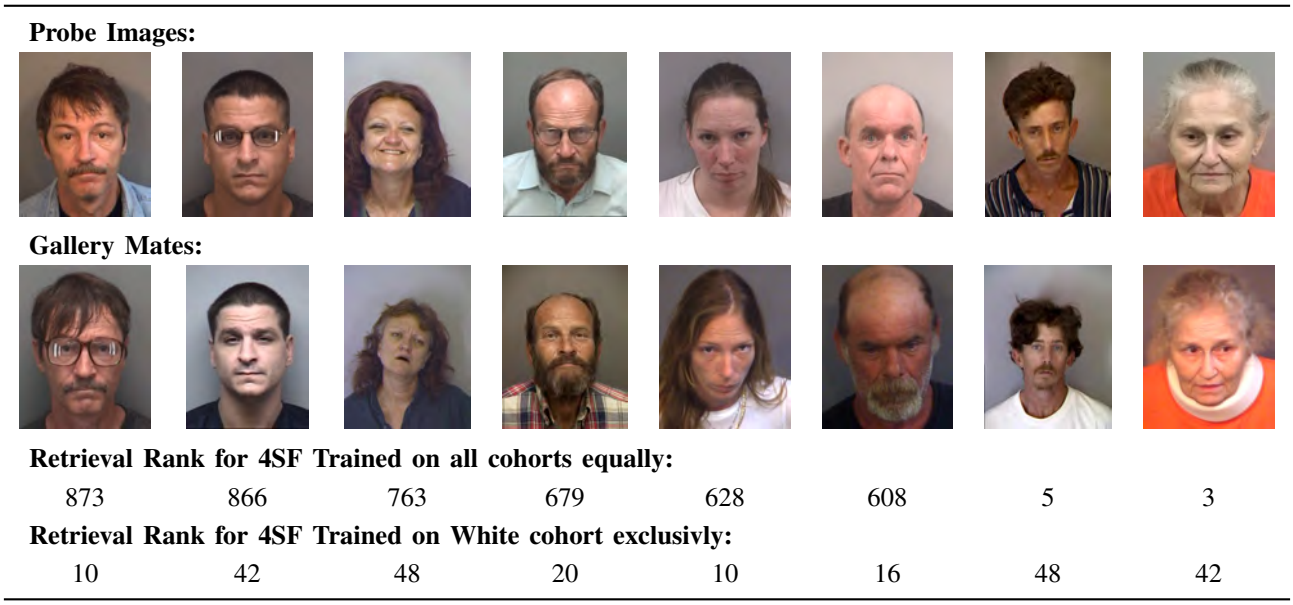
When evaluating 4SF on only the 18 to 30 year old cohort (Figure 6 (f)) and the 30 to 50 year old cohort (Figure 6 (g)), the highest performance was achieved when training on the same cohort. Table II helps elaborate on the exact accuracies. Similar to race, we were able to improve recognition accuracy by merely changing the distribution of the training set.

When comparing the 4SF system that is trained with equal number of subjects from all age cohorts, the performance on the 18 to 30 year old cohort is the lowest. This is consistent with the accuracies of the commercial face recognition algorithms.

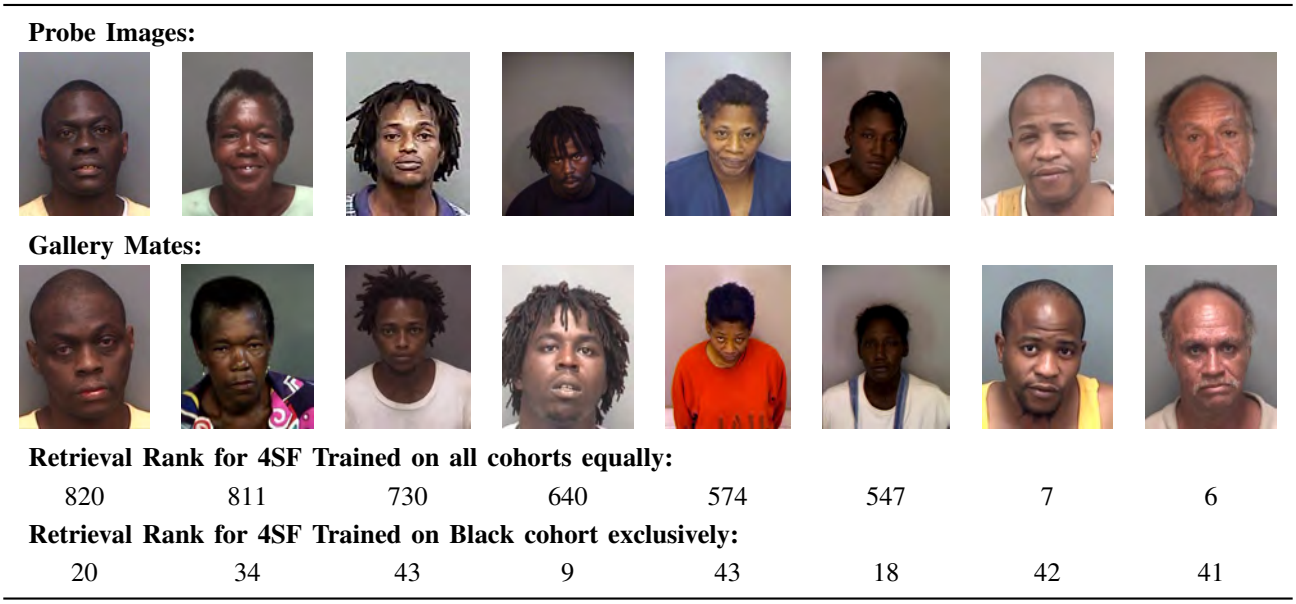
The less effective results from training on the 50 to 70 year old cohort is likely due to an small number of training subjects. This is consistent with the training results on the Hispanic cohort, which also had a small number of training subjects.

D. Impact of Training

The demographic distribution of the training set generally had a clear impact on the performance of different demographic groups. Particularly in the case of race/ethnicity, we see that training on a set of subjects from the same demographic cohort as being matched offers an increase in the True Accept Rate (TAR). This finding is particularly important because in most operational scenarios, particularly those dealing with forensics and law enforcement, the use of face recognition is not being done in a fully automated, "lights out" mode. Instead, an operator is usually interacting with a face recognition system, performing a one-to-one verification task, or exploring the gallery to group together candidates in clusters for further exploitation. Each of these scenarios can benefit from the use of demographic-enhanced matching algorithms, as described below.



(a)



(b)

Fig. 8. Shown are examples where dynamic face matcher selection improved the retrieval accuracy. The final two columns show the less frequent case where such a technique reduced the retrieval accuracy. Retrieval ranks are out of 8,000 gallery subjects for the White cohort (a), and 7,992 for the Black cohort (b). Leveraging demographic information (such as race/ethnicity in this example) allows a face recognition system to perform the matching using statistical models that are tuned to the differences within the specific cohort.

a) Scenario 1 - 1:N Search: In many large face recognition database searches, the objective is to have the true match candidates ranked high enough to be found by the analyst performing the candidate adjudication. While it will not always be the case, under many conditions, the analyst will be able to categorize the demographics of the probe image based on age, gender, and/or race/ethnicity. In such a situation, if the analyst has the option to select a different matching algorithm that has been trained for that specific demographic group, then improved matching results should be expected. An schematic of this is shown in Figure 2. This individual could be searched using an algorithm trained on male, Whites, and aged 18 to 30. If a true match is not found using that algorithm, then a more generic algorithm might be used as a follow up to further search the gallery. Note that this scenario does not require that the gallery images be pre-classified based on specific demographic information. Instead, the algorithm should simply generate higher match scores for subjects that share the characteristics of that demographic cohort. We call this method of face search *dynamic face matcher selection*. In cases where the demographic is unclear (e.g., a mixed race/ethnicity subject), the matcher trained on all cohorts

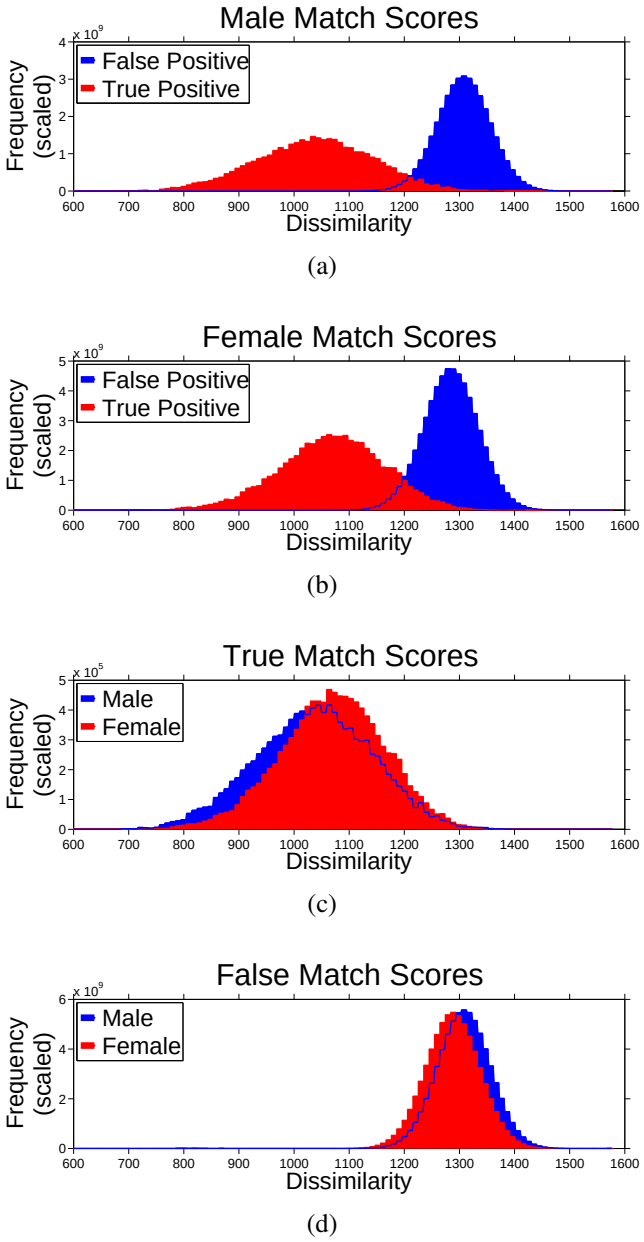


Fig. 7. Match score distributions for the male and female genders using the 4SF system trained with an equal number of male and female subjects. The increased distances (dissimilarities) for the true match comparisons in the female cohort suggest increased within-class variance in the female cohort. All histograms are aligned on the same horizontal axis.

equally can be used. Examples of improved retrieval instances through applying this technique can be found in Figure 8.

b) Scenario 2 - 1:1 Verification: It is often the case that investigators will identify a possible match to a known subject and will request an analyst to perform a 1:1 verification of the match. This also happens as a result of a 1:N search, once a potential match to a probe is identified. In either case, the analyst must reach a determination of match or no-match. In fully automated systems, this decision is based on a numerical similarity threshold. In some environments, the analyst is prevented from seeing the similarity score out of concern that his judgment will be biased. But in others, the analyst is permitted to incorporate this into his analysis. In either case, it is anticipated that an algorithm trained on a specific demographic group will return higher match scores for true matches than one that was more generic. As a result, the analyst is more likely to get a hit and the 1:1 matching results process will be improved.

c) Scenario 3 - Verification at Border Crossings: The results presented here provide support for further testing of additional demographic groups, potentially including specific country or geographic-region of origin. Assuming such demographics proved effective at improving match scores, then use of dynamic face matcher selection could be extended to immigration or border checks on entering subjects to verify that their passport or other documents accurately reflects their country of origin.

d) Scenario 4 - Face Clustering: Another analyst-driven application involves the exploitation of large sets of uncontrolled face imagery. Images encountered in intelligence or investigative applications often include large sets of videos or arbitrary photographs taken with no intention of enrolling them in a face recognition environment. Such image sets offer a great potential for development of intelligence leads by locating multiple pictures of specific individuals and giving analysts an opportunity to link subjects who may be found within the same photographs. Clustering methods are now being used on these datasets to group faces that appear to represent the same subject. Implementations of such clustering methods today usually rely upon a single algorithm to perform the grouping and an analyst must perform the quality control step to determine if a particular cluster contains only a single individual. By combining multiple demographic-based algorithms into a sequential analysis, it may be possible to improve the clustering of large sets of face images and thereby reduce the time required for the analyst to perform the adjudication of individual clusters.

VII. Conclusions

This paper examined face recognition performance on different demographic cohorts on a large operational database of 102,942 face images. Three demographics were analyzed: gender (male and female), race/ethnicity (White, Black, and Hispanic), and age (18 to 30 years old, 30 to 50 years old, and 50 to 70 years old).

For each demographic cohort, the performances of three commercial face recognition algorithms were measured. The performances of all three commercial algorithms were consistent in that they all exhibited lower recognition accuracies on the following cohorts: females, Blacks, and younger subjects (18 to 30 years old).

Additional experiments were conducted to measure the performance of non-trainable face recognition algorithms (local binary pattern-based and Gabor-based), and a trainable subspace method (the Spectrally Sampled Structural Subspace Features (4SF) algorithm). These experiments offered additional evidence to form hypotheses about the observed discrepancies between certain demographic cohorts.

Some of the keys findings in this study are:

- The female, Black, and younger cohorts are more difficult to recognize for all matchers used in this study (commercial, non-trainable, and trainable).
- Face recognition performance on race/ethnicity and age cohorts generally improve when training exclusively on that same cohort.
- The above finding suggests the use of *dynamic face matcher selection*, where multiple face recognition systems, trained on different demographic cohorts, are available as a suite of systems for operators to select based on the demographic information of a given query image (see Figure 2).
- In scenarios where dynamic matcher selection is not possible, training face recognition systems on datasets that are well distributed across all demographics is critical to reduce face matcher vulnerabilities on specific demographic cohorts.

Finally, as with any empirical finding, additional ways to exploit the findings of this research are likely to be found. Of particular interest is the observation that women appear to be more difficult to identify through facial recognition than men. If we can determine the cause of this difference, it may be possible to use that information to improve the overall matching performance.

The experiments conducted in this paper should have a significant impact on the design of face recognition algorithms. Similar to the large body of research on algorithms that improve face recognition performance in the presence of other variates known to compromise recognition accuracy (e.g., pose, illumination, and aging), the results in this study should motivate the design of algorithms that specifically target different demographic cohorts within the race/ethnicity, gender and age demographics. By focusing on improving the recognition accuracy on such confounding cohorts (i.e., females, Blacks, and younger subjects), researchers should be able to further reduce the error rates of state of the art face recognition algorithms and reduce the vulnerabilities of such systems used in operational environments.

ACKNOWLEDGEMENTS

We would like to thank Scott McCallum and the Pinellas County Sheriff's Office for assisting with the data. This study would not have been possible without their invaluable support. Feedback provided by Nick Orlans was instrumental in the completion this paper. We appreciate Patrick Grother's insightful comments on this study. This research was supported by the Office of the Director of National Intelligence (ODNI). Richard Vorder Bruegge's research is partially supported by the Director of National Intelligence (DNI) Science and Technology (S&T) Fellows program.

REFERENCES

[1] P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An introduction to the good, the bad, & the ugly face recognition challenge problem," in *Proc. of Automatic Face Gesture Recognition*, 2011.

[2] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval: Applications in forensics," *IEEE Multimedia (to appear)*, 2011.

[3] P. J. Phillips, P. J. Grother, R. J. Micheals, D. Blackburn, E. Tabassi, and J. M. Bone, "Face recognition vendor test 2002: evaluation report," *National Institute of Standards and Technology, NISTIR*, vol. 6965, 2003.

[4] P. J. Grother, G. W. Quinn, and P. J. Phillips, "MBE 2010: Report on the evaluation of 2D still-image face recognition algorithms," *National Institute of Standards and Technology, NISTIR*, vol. 7709, 2010.

[5] B. Klare, "Spectrally sampled structural subspace features (4SF)," in *Michigan State University Technical Report, MSU-CSE-11-16*, 2011.

[6] S. Z. Li and A. K. Jain, Eds., *Handbook of Face Recognition*, 2nd ed. Springer, 2011.

[7] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[8] L. Wiskott, J. Fellous, N. Kuiger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 775–779, 1997.

[9] X. Wang and X. Tang, "Random sampling for subspace face recognition," *Int. Journal of Computer Vision*, vol. 70, no. 1, pp. 91–104, 2006.

[10] P. Belhumeur, J. Hespanda, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, 1997.

[11] B. Moghaddam, T. Jebara, and A. Pentland, "Bayesian face recognition," *Pattern Recognition*, vol. 33, no. 11, pp. 1771–1782, 2000.

[12] R. K. Bothwell, J. Brigham, and R. Malpass, "Cross-racial identification," *Personality & Social Psychology Bulletin*, vol. 15, pp. 19–25, 1989.

[13] P. N. Shapiro and S. D. Penrod, "Meta-analysis of face identification studies," *Psychological Bulletin*, vol. 100, pp. 139–156, 1986.

[14] P. Chiroro and T. Valentine, "An investigation of the contact hypothesis of the own-race bias in face recognition," *Quarterly Journal of Experimental Psychology, Human Experimental Psychology*, vol. 48A, pp. 879–894, 1995.

[15] W. Ng and R. C. Lindsay, "Cross-race facial recognition: Failure of the contact hypothesis," *Journal of Cross-Cultural Psychology*, vol. 25, pp. 217–232, 1994.

[16] N. Furl, P. J. Phillips, and A. J. O'Toole, "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis," *Cognitive Science*, vol. 26, no. 6, pp. 797–815, 2002.

[17] A. O'Toole, P. J. Phillips, A. Narvekar, F. Jiang, and J. Ayyad, "Face recognition algorithms and the other-race effect," *Journal of Vision*, vol. 8, no. 6, 2008.

[18] B. Klare and A. K. Jain, "Face recognition across time lapse: On learning feature subspaces," in *Int. Joint Conference on Biometrics*, 2011.

[19] P. Phillips, W. Scruggs, A. O'Toole, P. Flynn, K. Bowyer, C. Schott, and M. Sharpe, "FRVT 2006 and ICE 2006 large-scale experimental results," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 831–846, 2010.

[20] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, 2002.

[21] B. Klare and A. Jain, "On a taxonomy of facial features," in *Proc. of IEEE Conference on Biometrics: Theory, Applications and Systems*, 2010.

[22] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.

[23] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1635–1650, 2010.

[24] L. Shen and L. Bai, "A review on Gabor wavelets for face recognition," *Pattern Analysis & Applications*, vol. 9, pp. 273–292, 2006.

[25] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature Neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[26] E. Meyers and L. Wolf, "Using biologically inspired features for face processing," *Int. Journal of Computer Vision*, vol. 76, no. 1, pp. 93–104, 2008.

[27] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 8, pp. 832–844, 1998.

[28] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[29] X. Mao, A. W. Bigham, R. Mei, G. Gutierrez, K. M. Weiss, T. D. Brutsaert, F. Leon-Velarde, L. G. Moore, E. Vargas, P. M. McKeigue, M. D. Shriver, and E. J. Parra, "A genomewide admixture mapping

panel for hispanic/latino populations," *The American Journal of Human Genetics*, vol. 80, no. 6, pp. 1171–1178, 2007.

Brendan F. Klare is a doctoral candidate in the Pattern Reconition and Image Processing Lab at Michigan State University. He received the B.S. and M.S. degrees in Computer Science and Engineering from the University of South Florida in 2007 and 2008. From 2001 to 2005 he served as an airborne ranger infantryman in the 75th Ranger Regiment. Brendan has authored several papers on the topic of face recognition, and he recieved the Honeywell Best Student Paper Award at the 2010 IEEE Conference on Biometrics: Theory, Applications and Systems (BTAS). His other research interests include pattern recognition and computer vision.

Mark J. Burge is a scientist with The MITRE Corporation, McLean, Virgina. Dr. Burge's research interests include image processing, pattern recognition, and biometric authentication. He served as a Program Director at the National Science Foundation, a research scientist at the Swiss Federal Institute of Science (ETH) in Zurich, Switzerland, and the TUWien in Vienna, Austria. While a tenured computer science professor, he co-authored the three volume set, "Principles of Digital Image Processing" which has been translated into both German and Chinese. He holds patents in multispectral iris recognition and is the coeditor of Springer Verlag's forthcoming "Handbook of Iris Recognition".

Joshua C. Klontz is a scientist with The MITRE Corporation, McLean, Virgina. He recieved a B.S. in Computer Science from Harvey Mudd College in 2010. His research interests include pattern recognition and cross-platform C++ software development. Josh also has an academic interest in epidemiology and is recently published in the Journal of Food Protection.

Richard W. Vorder Bruegge is a Senior Level Photographic Technologist for the Federal Bureau of Investigation. In this role, he is responsible for overseeing FBI science and technology developments in the imaging sciences. He serves as the FBIs subject matter expert for face recognition and is the current chair of the Facial Identification Scientific Working Group. He has multiple publications on forensic image analysis and biometrics and he was co-editor of Computer-Aided Forensic Facial Identification (2010). He has testified as an expert witness over sixty times in criminal cases in the United States and abroad. Dr. Vorder Bruegge is a fellow of the American Academy of Forensic Sciences and was named a Director of National Intelligence Science and Technology Fellow in January 2010.

Anil K. Jain is a university distinguished professor in the Department of Computer Science and Engineering at Michigan State University, East Lansing. His research interests include pattern recognition and biometric authentication. He served as the editor-in-chief of the IEEE Trans. on Pattern Analysis and Machine Intelligence (1991-1994). The holder of six patents in the area of fingerprints, he is the author of a number of books, including Handbook of Fingerprint Recognition (2009), Handbook of Biometrics (2007), Handbook of Multibiometrics (2006), Handbook of Face Recognition (2005), Biometrics: Personal Identification in Networked Society (1999), and Algorithms for Clustering Data (1988). He served as a member of the Defense Science Board and The National Academies committees on Whither Biometrics and Improvised Explosive Devices. Dr. Jain received the 1996 IEEE Trans. on Neural Networks Outstanding Paper Award and the Pattern Recognition Society best paper awards in 1987, 1991, and 2005. He is a fellow of the AAAS, ACM, IAPR, and SPIE@. He has received Fulbright, Guggenheim, Alexander von Humboldt, IEEE Computer Society Technical Achievement, IEEE Wallace McDowell, ICDM Research Contributions, and IAPR King-Sun Fu awards. ISI has designated him a highly cited researcher. According to Citeseer, his book Algorithms for Clustering Data (Englewood Cliffs, NJ: Prentice-Hall, 1988) is ranked #93 in most cited articles in computer science.