To appear: Proceedings of the IEEE Conference on Biometrics: Theory, Applications, and Systems, 2012.

Face Recognition: Impostor-based Measures of Uniqueness and Quality

Brendan F. Klare
Noblis
Falls Church, VA, U.S.A.
brendan.klare@noblis.org

Anil K. Jain
Michigan State University
East Lansing, MI, U.S.A.
jain@cse.msu.edu

Abstract

We present a framework, called uniqueness-based non-match estimates (UNE), which demonstrates the ability to improve face recognition performance of any face matcher. The first aspect of the framework is a novel metric for measuring the uniqueness of a given individual, called the impostor-based uniqueness measure (IUM). The UNE the maps face match scores from any any face matcher into non-match probability estimates that are conditionally dependent on the probe image's IUM. Using this framework we demonstrate: (i) an improved generalization of matching thresholds (and, subsequently, improved matching accuracy), (ii) a score normalization technique that improves the interoperability for users of different face matchers, and (iii) the predictive ability of IUM towards face recognition accuracy. Studies are conducted on an operational dataset with 16,000 subjects using three different face matchers (two commercial, one proprietary) to demonstrate the effectiveness of the proposed framework.

1. Introduction

Certain faces have intrinsic attributes that impact their ability to be successfully matched by an automatic face recognition system. Specifically, the facial appearance of some subjects exhibits high levels of similarity to other subjects in some population. For example, Figure 1 shows six groups of subjects who all look very similar with other members of the group. Different from the pose, expression, illumination, and aging variates that are known to compromise performance, such intrinsic factors cannot be compensated for through image processing and face modeling techniques. Instead, these subjects pose challenge to the face matching process because they do not conform to global thresholds designed to produce accurate genuine matches at some fixed false accept rate.

In this work we present a *matcher-independent* framework for improving face recognition performance. This is achieved by measuring the perceived uniqueness of the subject in a probe image, and mapping the match score to a non-match probability estimate [11] which is conditionally dependent on the uniqueness measure. Called Uniqueness-based Non-match Estimates (UNE), this proposed framework offers the following benefits and contributions. (i) An improved generalization of decision thresholds used in the face matching process. In turn, recognition performance is increased across the three different matchers studied. (ii) UNE results in a common range and distribution of match scores across all face matchers. Consistency between match scores from different matchers improves the interoperability of face recognition with users, (such as biometric system analysts). (iii) Finally, we introduce the matcher independent Impostor-based Uniqueness Measure (IUM), which measures the uniqueness of subject's face given a single image of their face.

The paper is organized as follows. In Section 2 we discuss past studies and their relation to this work. In Section 3 we discuss the different face recognition algorithms used in this study (two commercial, one proprietary). In Section 4 we propose a novel quality metric that requires only black box access to any arbitrary face recognition system. In Section 5 we detail the procedure for converting face match scores into empirical non-match probabilities. In Section 6 we present and analyze various experimental results which demonstrate the value of uniqueness-based non-match estimates. Finally, we conclude this paper in Section 7.

2. Related Work

Doddington's Biometric Zoo [4] categorizes people based on how their biometric traits interact with the rest of the population. The majority of the population are considered "sheep", who are generally easy to distinguish given a good quality sample. A small percentage of the population fall into the category of "goats", "lambs", and "wolves". Goats are subjects that have traits which are difficult to match. Lambs are subjects whose traits exhibit high levels of similarity to other subjects (all the subjects shown within each group in Figure 1 would be considered lambs). Wolves are those able to best mimic other subjects' traits, and hence

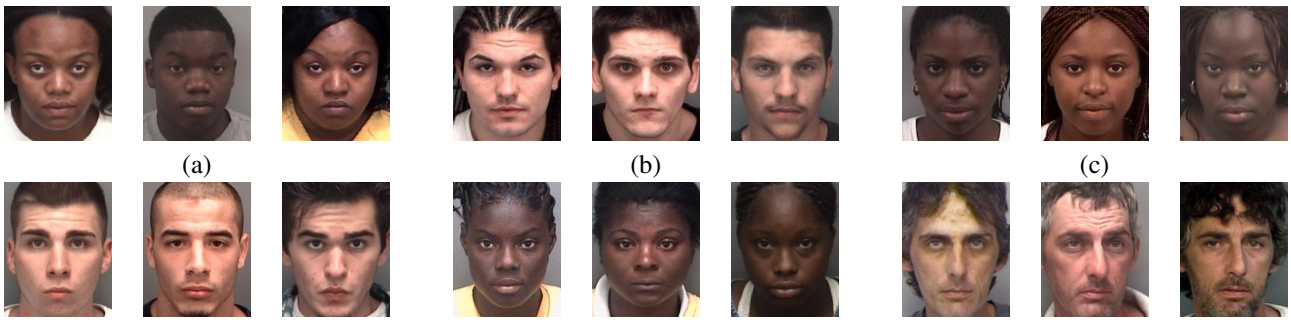


Figure 1. Six different groups of subjects consisting of members with high facial similarity to the other members of the group. Such subjects are often referred to as "lambs" [4], and recognizing them across a large population is a difficult task when using a common threshold on match scores. In this work we detect such subjects using a matcher-independent measure, called the Impostor-based Uniqueness Measure (UIM). Match scores are mapped to non-match probability estimates based on the IUM score, resulting in improved recognition accuracy.

present risks for spoofing a biometric system. This work focuses on the lambs in the face biometric.

Improving face matching in the presence of lambs is better suited than goats because lambs are a function of their impostor match score distributions [4]. By contrast, goats are defined based on their true match score distribution. While measuring the true match score distribution requires prior knowledge regarding a subject's identity and multiple biometric samples, impostor match scores require no prior information regarding the subject's identity and can generally be determined from a single image.

Because the accuracy of a face recognition system will be dependent on whether or not a subject exhibits sheep-like properties, there is strong motivation to measure such properties in a biometric sample (in our case, a face image). In addition to Doddington et al.'s original measures for each animal in the biometric zoo [4], Poh and Kittler offered measures of "sheepishness" (such as their F-ratio), and demonstrated the ability of these measures to predict matching accuracy [13]. Poh and Kittler later introduced the biometric menagerie index (BMI) to further characterize a subject's membership in Doddington's zoo [14]. We refer to such a measure of where a subject falls in Doddington's zoo (e.g., F-ratio, BMI, IUM), by "uniqueness measure". Similar to the approach presented in this paper, Poh et al. performed group specific score normalization based on the genuine and impostor match scores of a subject [15]. Ross et al. used such uniqueness measures in a multi-modal biometric system to apply adaptive fusion based on the intrinsic weakness of a subject's biometric characteristics [16]. Other studies have been conducted to measure the generalization of zoos across different fingers [7] and face matchers [17]. The presence of zoos on datasets with several hundred subjects has been previously studied [19, 18].

These earlier studies [13, 14, 16, 15] that measure are limited in that the uniqueness measures for a specific subject were based on generative methods that require multiple samples per subject. Thus, these approaches perform user-specific training. In most applications of face recognition one cannot expect multiple face images per subject, limiting the role of these methods. Further, these approaches are designed for 1:1 verification. In identification tasks, there is no such identity claim and the per-user model would result in a different uniqueness measure for each subject in the gallery that is compared to a probe image. For large face databases in the order of millions of subjects, such an approach is again not feasible.

By contrast, the impostor-based uniqueness measure offered in this paper (see Section 4) utilizes only a single image of a subject, and does not require any claimed identity. Thus, while it is not directly comparable to previous techniques, it serves as a compliment to those approaches for scenarios in which it cannot be assumed that multiple images per subject are available in the database.

A second aspect of our work is a mapping of a match score to a uniqueness-based non-match probability estimate (see Section 5). The approach is motivated by Choi et al.'s use of evidential non-match probabilities in fingerprint recognition [11]. Using training data in the form of imposter and genuine match scores from a biometric matcher, they computed the probability that a given match score would result from a comparison of two different subjects. While Choi et al. applied this technique to facilitate the admissibility of fingerprint match scores as legal evidence, we instead show how this approach can be used to alter decision thresholds in the face recognition process. Choi et al.'s also reported that fingerprints of different qualities resulted in different evidential non-match probability distributions. We also extend this notion of quality to the uniqueness (or intrinsic quality) of a biometric (face) sample, as motivated by our discussions of the biometric zoo. The approach was similarly applied by Poh et al. using multiple samples per subject to categorize their uniqueness [15].

Our motivation to alter decision thresholds based on con-

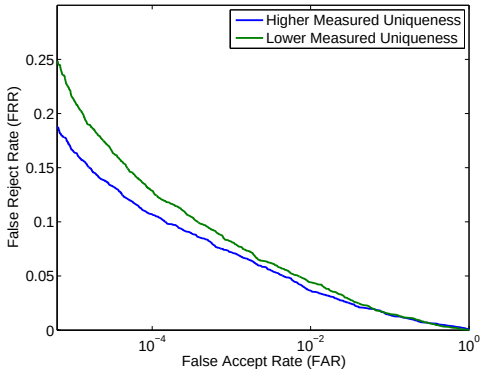


Figure 2. ROC plot comparing the face recognition accuracy when using probe images with "higher" uniqueness and probe images with "lower" uniqueness, using the proposed impostor-based uniqueness metric (IUM). Match scores were generated from 8,000 subjects using FaceVACS. This work reduces the FRR by conditionally mapping a match score to an estimated non-match probability based on the uniqueness of a face image.

textual information is also based on several findings that biometric systems exhibit different accuracies depending on the general quality of a biometric sample being matched. It is well documented that the accuracy of face recognition systems is a generally monotonic function of the quality of the face image samples [10, 12, 8, 3, 9]. In general, quality is encompassing of the pose, expression, illumination, compression, time lapse, and uniqueness of a subject.

In addition to face recognition systems being sensitive to the quality of the samples being matched, Grother and Tabassi [6] also demonstrated that the accuracy of a biometric system can generally be predicted by the biometric quality metric. A key contribution of our work here is the demonstration that such decision thresholds can be normalized by converting face match scores to the empirical non-match probability estimates which are conditionally dependent on the facial uniqueness (see Figure 5 for an overview of this approach).

3. Face Recognition Algorithms and Databases

This study evaluates the benefits of proposed uniqueness-based non-match estimates using three different face recognition algorithms. The first two algorithms are commerical off the shelf (COTS) matchers: Cognitec's FaceVACS [1] and Neurotechnology's VeriLook [2]. The third face recognition algorithm is our proprietary algorithm, called Anonymous Proprietary Algorithm[1] These matchers are all used as black box systems which are capable of outputting a measure of similarity between two facial photographs.

This study uses face images from 16,000 subjects which come from an operational database maintained by the Pinellas County Sheriff's Office.

Two images per subject were used in this study: one for a gallery seed, and the other for a probe/query. We used the first 8,000 subjects in the dataset for training purposes, and the remaining 8,000 subjects for testing. Both the training and testing set were controlled to contain 2,000 white males, 2,000 black males, 2,000 white females, and 2,000 black females. No subject used in training was used for testing. The training dataset was used in two capacities: (i) to compute the impostor-based uniqueness measure (see Section 4), and (ii) to learn the non-match probability distributions (see Section 5).When computing the IUM for the images in the training set, their mated images were removed (i.e. only the impostor comparisons were considered).

4. Uniqueness Measure

A subject whose face is non-unique, or is a "lamb", will generally exhibit high levels of similarity to many other subjects in a large population (by definition). Based on this basic assertion, a subject's uniqueness is measurable through how similar he is to a population of impostors. We assume that (i) only a single image for the subject is available, and (ii) the identity of the subject is not claimed. Given these constraints, we define the impostor-based uniqueness measure as the mean match score to a set of impostor subjects.

Let $S(i, j, m)$ denote the match score between face images from subjects i and j using matcher m (e.g., $m =$ FaceVACS). Given a set of n subjects $J = \{j_1, j_2, \ldots, j_n\}$, which represent impostor face images, the impostor-based uniqueness metric u for subject i against the set J and matcher m is defined as

$$u'(i, J, m) = \frac{1}{n} \sum_{k=1}^{n} s(i, j_k, m) \quad (1)$$

$$u(i, J, m) = \frac{\max(u'_J) - u'(i, J, m)}{\max(u'_J) - \min(u'_J)} \quad (2)$$

where $\min(u'_J)$ and $\max(u'_J)$ are the minimum and maximum u' values computed within the set J (respectivelly). If $u(i, \cdot, \cdot)$ is high, we can infer that subject i's face is generally unique. Conversely, a low value for u would infer the subject contains a more typical looking face.

We found that the IUM is quite consistent regardless of the impostor subjects used. This is shown in Figure 4, where the IUM value for all 8,000 probe images in the test set were computed using both the training set (labeled Set 1) and the 7,999 impostors in the test set (labeled Set 2). When the IUM was computed for Set 1, the number of impostors was steadily increased from 100 subjects to 8,000 subjects. In each case, plots also list the linear correlation between the IUM measure computed from the two different sets. Even when the IUM was computed using only 100 impostor images from Set 1, the correlation with the IUM computed

[1] Details of the algorithm have been omitted to maintain anonymity.

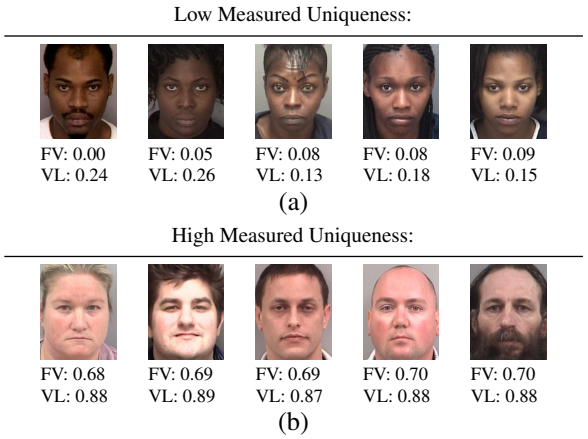


Figure 3. (a) Face images with low computed uniqueness using the proposed IUM. (b) Faces images with high computed uniqueness. Faces in (b) will generally match with higher accuracy than those in (a), and should therefor be thresholded differently. Black subjects were found to have consistently lower IUM's than white subjects in our study (indicating a potential bias in the studied face recognition systems). FV:FaceVACS; VL: VeriLook.

using 8,000 subjects from Set 2 was 0.92. If we increased the impostors in Set 1 to 1,000, subjects the correlation was 0.99. We see from Figure 4 that the proposed IUM value becomes stable with a database of 1,000 subjects.

Considering operational use of such a measure, storing 1,000 templates for computing IUM is not a hurdle. As an example, in FaceVACS a face template is roughly 2KB in size, so storing 1,000 templates would require only 2MB of space in the enrollment module. In terms of computation time, modern face matchers are generally expected to operate at speeds of at least 1,000 face comparisons per second. Thus, computing this measure should take no more than one second during the offline enrollment.

Finally, IUM offers a generally accurate predication of recognition accuracy. For example, using the 8,000 test subjects, Figure 2 shows the ROC plots from the 4,000 subjects with the highest measured uniqueness and the 4,000 subjects with the lowest measured uniqueness (using the IUM computed against the training set). A clear difference in recognition accuracy is noted between subjects with high IUM and subjects with low IUM, particularly in the critical region of ROC corresponding to low false accept rates (FAR). For example, the more unique subjects matched with nearly 5% higher accuracy than the less unique subjects at low false accept rates. Example of face images with the lowest and highest IUM's can be seen Figure 3.

5. Evidential Non-Match Probabilities

Similar to the demonstration by Grother and Tabassi [6] of how the accuracy of a biometric system can generally be predicted by the systems quality metric, the results in Figure 2 show a correlation between recognition performance and inferred uniqueness. This suggests a global decision threshold on match scores is not generally desirable. Instead, the decision threshold should be altered based on the image quality or any other contextual information that correlates with match scores.

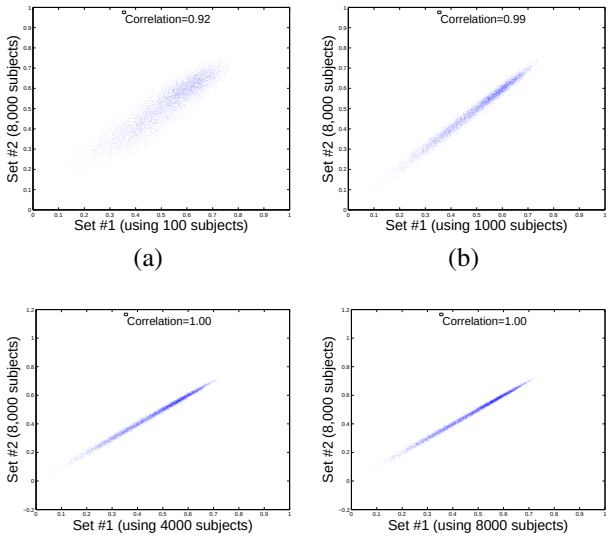


Figure 4. Generalization of the impostor-based uniqueness measure (IUM). The scatters plots contain corresponding IUM for 8,000 images computed using impostors from two non-overlapping datasets. The high correlation indicates this measure of uniqueness generalizes well when using different sets of impostors. The number of impostors used to compute the IUM from the first set (x-axis) are: (i) 100, (ii) 1,000, (iii) 4,000, and (iv) 8,000.

Recently, Choi et al. [11] proposed normalizing match scores into evidential non-match probabilities. Specifically, given a match score s, the non-match probability (NMP) is defined as

$$NMP = P(I|s) = 1 - P(G|s) \tag{3}$$

$$P(G|s) = \frac{P(s|G)P(G)}{P(s|I)P(I) + P(s|G)P(G)} \tag{4}$$

where I is the event of an impostor comparison, and G is the even of a genuine comparison. Given a sufficiently large training set of match scores, genuine and impostor score densities $P(s|G)$ and $P(s|I)$ can be estimated. Following Choi et al., we estimate these distributions using kernel density estimation (KDE) [5]. For KDE, we used a Gaussian kernel with a bandwidth set as the standard deviation of the match scores. We tacitly assume that the prior probabilities for impostor and genuine subjects $P(I)$ and $P(G)$ to be equal as we simply use the computed non-match probability estimates for score normalization.

Using this approach, match scores output by face match-

ers from any range, such as $(-\infty, \infty)$ or $(0, 255)$, will now be converted to the range $[0, 1]$. Further, having learned the match score distributions from the kernel density estimates, a NMP match score from any face matcher generally has the same interpretation. For example, an NMP score of 0.95 from both FaceVACS and VeriLook should have roughly the same likelihood of being the same subject. Section 6 will show how such a normalization of match scores can improve users who interact with multiple face recognition systems.

5.1. Uniqueness-based NMP

The aforementioned non-match probabilities convert match scores to probability estimates that share a common range of [0,1]. It is important to point out that such mappings are monotonic with respect to the input match scores, and they will not alter the ROC performance.

In order to choose an appropriate decision threshold, we leverage the impostor-based uniqueness measure defined in Section 4. If we separate subjects into n_U different levels of uniqueness based on their IUM, then we can learn a separate NMP mapping for each uniqueness level. Let $U \in \mathbb{R}^n$ be the vector of IUM values u for the n training subjects for a given matcher. Recall we can compute the IUM using the impostors within the training set. For a given uniqueness level $l \in l_1, l_2, \ldots, l_{n_U}$, we compute the non-match probability NMP_l as

$$NMP_l = P(I|s, l) = 1 - \frac{P(s|G, l)}{P(s|I, l) + P(s|G, l)} \quad (5)$$

where the likelihoods $P(s|G, l)$ are $P(s|I, l)$ are conditionally dependent on the uniqueness level l as well as the match score s. We estimate these likelihoods in the same manner as before: kernel density estimation. However, these distributions are estimated using only probe subjects whose IUM falls within the level l. The gallery images from subjects whose probe image is not in this level are still retained as impostors when generating the density estimates.

We partition subjects into three levels of uniqueness: low, average, and high. High subjects are those whose IUM is more than one standard deviation above the mean IUM. Average uniqueness subjects are those who are within one standard deviation from the mean. Finally, low uniqueness subjects are those whose IUM is more than one standard deviation below the mean. The mean and standard deviation statistics are computed using the training set.

Figure 6 shows the NMP mappings that are learned from each of the three uniqueness levels using the FaceVACS match scores from the training set; a clear separation is observed among the three learned mappings. The difference between these mappings is intuitive: subjects with low uniqueness are required to have a high match score in order

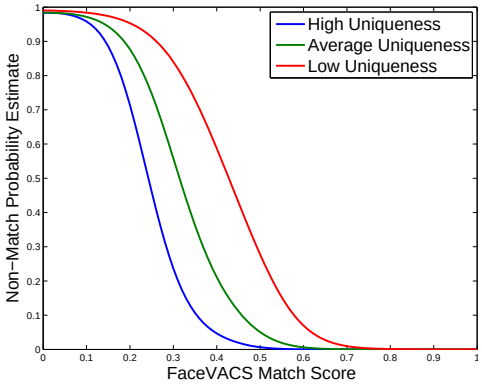


Figure 6. Mapping of FaceVACS match scores to uniqueness-based non-match estimates. Each mapping is learned from subjects with different levels of uniqueness, as defined by the proposed IUM. The learned mappings resulted faciliates a single threshold for allowing less unique subjects to have higher FaceVACS match scores than their more unique counterparts.

to generate the same non-match probability as a subject with high uniqueness. For example, a FaceVACS match score of 0.4 will result in a non-match probability estimate of 0.05 for a subject who falls in the high uniqueness group. In other words, it is estimated there is a 95% chance that the two images being compared are from the same subject. By contrast, the same FaceVACS match score of 0.4 will result in a non-match probability estimate of 0.24 for a subject who falls in the low uniqueness group. That is, we now estimate that there is only a 76% chance the two images being compared are from the same subject. Thus, when a subject is more unique, an initial match score of 0.4 (for example) will be treated with more confidence. This is intuitive: the subject does not generally match highly to impostor subjects. By contrast, the subject who is not unique will treat this same match score more pessimistically as it is more likely to to be from an impostor subject.

The final framework, called Uniqueness-based Non-match Estimates (UNE), is shown in Figure 5. For a given probe image, the measured uniqueness will be used to map the match scores to non-match probability estimates using the mapping from the corresponding uniqueness group. In practice, we use $1 - NMP$ to convert the NMP from a dissimilarity measure to a similarity measure.

6. Experiments

We first examine how the proposed uniqueness-based non-match probability estimates affect face recognition performance. We use the 8,000 subjects in the training set to learn the NMP mappings and populate the impostors for the IUM. Figure 8 shows the ROC plots when matching the probe set using (i) the initial match scores from each matcher and (ii) the proposed uniqueness-based non-match estimates (UNE). It is noticed that each matcher exhibits

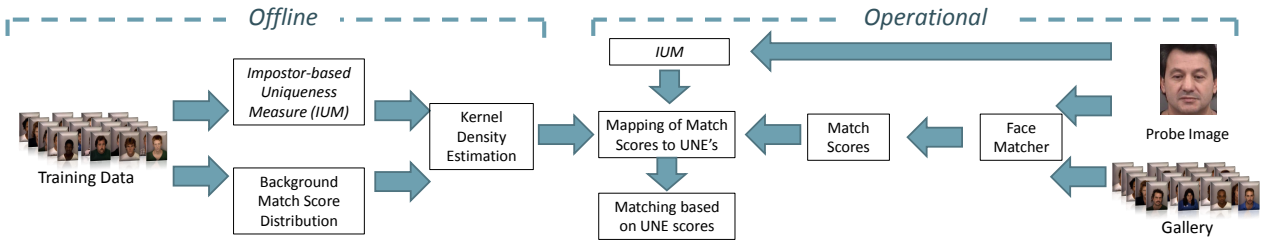


Figure 5. The proposed uniqueness-based non-match estimate (UNE) framework for learning and applying empirical non-match probability estimates. Given a background distribution of match scores (computed from a training set of images), empirical non-match probabilities are computed using kernel density estimation which are conditionally dependent on the measured uniqueness (IUM) of the face image.

Figure 7. The probe images (first column) from each subject correctly matched their mate (second column) using FaceVACS at a threshold that yields a false accept rate of 0.1%, . However, the impostor images in columns 3-6 where incorrectly classified as matches to the corresponding probe image. When the FaceVACS match scores were mapped to QNE scores, all subjects still correctly matched their mate, but no longer matched the impostors shown. The implicit change in threshold for each subject based on their measured uniqueness allowed for such improvements.

a decrease in recognition error with the proposed method. Thus, without any specific knowledge of a matcher (e.g. face representation, feature extraction, matching strategy), we are able to improve the face recognition accuracy.

Figure 7 shows examples of subjects who matched correctly to their genuine mate as well as to the shown impostor subjects using the FaceVACS match scores when operating at a false accept rate of 0.1%. After converting these match scores to UNE scores, the impostors for each subject no longer exceeded the matching threshold and were properly classified as non-matches. These examples show the value of UNE scores when applied to watch list or open set identification $(1{:}\,N+1)$ scenarios. In these cases it is ideal to set a stringent threshold that does not generate many false matches (as these are costly to manually eliminate), yet still maintains a low false reject rate. As observed from the ROC plots in Figure 8, the largest accuracy improvement obtained using UNE was obtained when operating at low false accept rates.

Figure 9 shows the histogram distributions of true and false match scores with and without UNE. The histograms are computed from 4,000 true match and 6.37×10^7 false match comparisons for each of the three matchers in this study (FaceVACS, VeriLook, and 4SF). When using UNE instead of the raw match scores, the distributions exhibit less overlap in the middle of the range of the distributions. Further, the extremes of each distributions are more peaked when using UNE. The change in distributions from raw match scores to UNE scores results in match score distributions which are better suited for a threshold. That is, with the false and true match distributions attenuating faster with UNE than without, improved match decisions are made within the range of thresholds containing these regions of decreased overlap. In turn, this results in the decreased FRR shown in Figure 8.

A final contribution of this work is a demonstration of how non-match probability estimates conditioned on uniqueness can be used to increase user operability of face recognition systems in retrieval systems. Because the measure of similarity output by face recognition systems differs between vendors and algorithms, match scores are typically not shown to system analysts when retrieving match candidates. However, using uniqueness-based non-match estimates: (i) the mean match scores of correct (Rank-1) retrievals from different face recognition algorithms are

Without UNE:

	FaceVACS	VeriLook	4SF
Rank-1*	0.95 ± 0.11	132.50 ± 90.02	0.94 ± 0.10
True Scores	0.89 ± 0.22	115.77 ± 91.88	0.82 ± 0.23
Impostor Scores	0.06 ± 0.07	2.37 ± 4.57	0.07 ± 0.11

(a)

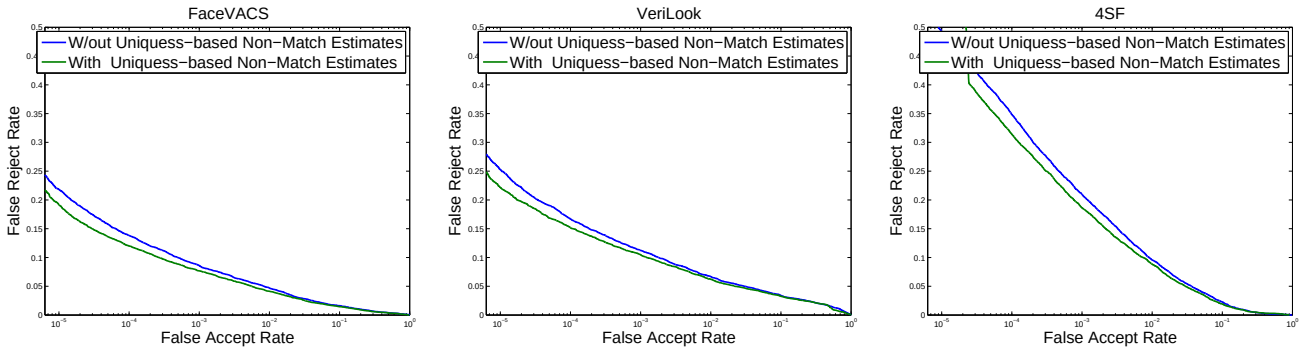
With UNE:

	FaceVACS	VeriLook	4SF
Rank-1*	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
True Scores	0.96 ± 0.17	0.94 ± 0.21	0.93 ± 0.19
Impostor Scores	0.04 ± 0.09	0.05 ± 0.11	0.07 ± 0.12

(b)

** Rank-1 is the true match scores for subjects which were correctly retrieved at Rank-1.*

Table 1. The mean and standard deviation statistics of match scores with and without the propose UNE score normalization. The statistics demonstrate the improved intra-matcher consistency (i.e., the stability of match scores within a single matcher) and inter-matcher consistency (i.e., the consistency of match scores from different matchers) when match scores are converted to uniqueness-based non-match estimates. High inter-matcher consistency offers strong benefits in terms of human inter-operability with face recognition systems.



(a) (b) (c)

Figure 8. ROC plots showing the recognition performance both with and without the proposed uniqueness-based non-match estimates. The decrease in the false reject rate (FRR) demonstrates an improved generalization of the decision thresholds after applying the proposed score normalization technique. (a) FaceVACS. (b) VeriLook. (c) 4SF.

made to be nearly the same regardless of the given matchers score distribution, and (ii) the variance of correct Rank-1 retrieval match scores can be significantly reduced. These two achievements greatly increase a system analyst's ability interpret match scores across different queries and face recognition algorithms.

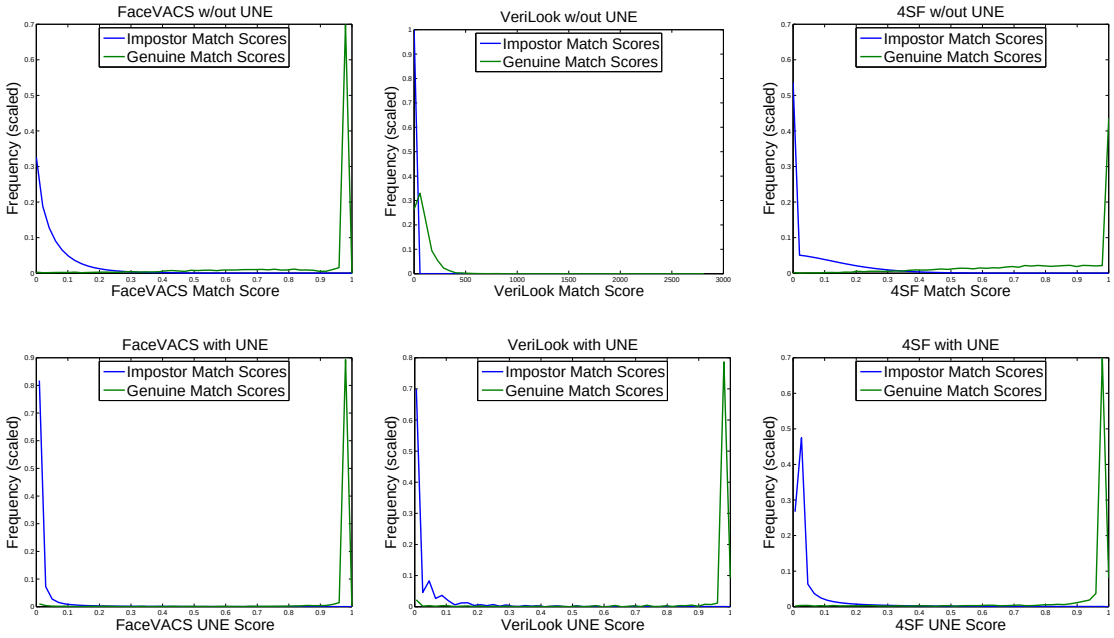
For example, Table 1 lists the mean and standard deviation of match scores on the test set of 8,000 subjects with and without the proposed UNE score normalization. The first row lists the average match score of a correct Rank-1 retrieval. Without QNE, we see that FaceVACS outputs an average score of 0.95 whereas VeriLook outputs an average score of 132.5. This substantial difference demonstrates the diffuses users of multiple face recognition systems face when interpreting match scores. By contrast, FaceVACS and VeriLook both have an average match score of 1.0 when using UNE scores. Thus, in addition to improving face recognition accuracy, UNE's also improve the consistency across heterogeneous face matchers.

7. Conclusions

Face recognition systems are operated as blackbox systems that, given two face images or templates, output a measure of similarity between the two biometric samples. In most forensic and operational scenarios these similarity measures are used to retrieve a list of candidate matches (1: N matching), determine whether two specific match companions are the same subject (1: 1 matching), or determine if an image matches to a watch list (open set identification, 1: $N+1$).

In mapping the three studied matchers' match scores to uniqueness-based non-match estimates (UNE), we demonstrated how each of these matching scenarios can be improved. For retrieval applications, the ability of UNE to map different score distributions to a common range and shape offers a consintent interface for a human analyst and the retrieval results provided. For 1: 1 comparisons, the UNE match scores offer improved generalized of the threshold used for determining a match. Similarly, in open set identification the improved threshold generalization allows for less false positive hits while mainting a given accuracy.

While the results provided demonstrated the efficacy of the proposed UNE framework, we see several avenues to expand the score normalization technique. For instance, in addition to uniqueness, factors such as image quality and subject demographics can influence face recognition accuracy. Thus, future studies will determine whether emperical non-match probability estimates that are conditionally dependend on image quality, race, or gender (in addition the IUM) can further improve the recognition accuracy.



(a) (b) (c)

Figure 9. Histograms of match score distributions with (top row) and without (bottom row) uniqueness-based non-matches estimates (UNE). (a) FaceVACS. (b) VeriLook. (c) 4SF. When match scores are converted to UNE's, (i) the tails of the impostor and genuine scoore distributions attenuate faster (resulting in less overlap), and (ii) the UNE distributions for each matcher are highly similar to one another.

References

[1] FaceVACS Software Developer Kit, Cognitec Systems GmbH, http://www.cognitec-systems.de.

[2] VeriLook SDK, Neurotechnology, http://www.neurotechnology.com/verilook.html.

[3] N. R. Council. *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press, 2009.

[4] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance in the NIST 1998 speaker recognition evaluation. In *Proc. of Int. Conference on Spoken Language Processing*, 1998.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. Wiley-Interscience, 2000.

[6] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):531 –543, 2007.

[7] A. Hicklin, C. Watson, and B. Ulery. The myth of goats: How many people have fingerprints that are hard to match. In *Technical Report NISTIR 7271, NIST*, 2005.

[8] A. K. Jain, B. Klare, and U. Park. Face matching and retrieval: Applications in forensics. *IEEE Multimedia*, 19(1):20–28, 2012.

[9] S. Li, X. Lu, X. Hou, X. Peng, and Q. Cheng. Learning multiview face subspaces and facial pose estimation using independent component analysis. *IEEE Trans. on Image Processing*, 14(6):705 –712, 2005.

[10] S. Z. Li and A. K. Jain, editors. *Handbook of Face Recognition*. Springer, 2 edition, 2011.

[11] A. Nagar, H. Choi, and A. K. Jain. Evidential value of latent fingerprint match. *IEEE Trans. on Information Forensics and Security (to appear)*, 2012.

[12] 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.

[13] N. Poh and J. Kittler. A methodology for separating sheep from goats for controlled enrollment and multimodal fusion. In *Biometrics Symposium*, pages 17 –22, 2008.

[14] N. Poh and J. Kittler. A biometric menagerie index for characterising template/ model-specic variation. In *Proc. Int. Conference on Biometrics*, 2009.

[15] N. Poh, J. Kittler, A. Rattani, and M. Tistarelli. Group-specific score normalization for biometric systems. In *Proc. of Computer Vision and Pattern Recognition Workshop*, 2010.

[16] A. Ross, A. Rattani, and M. Tistarelli. Exploiting the Doddington Zoo effect in biometric fusion. In *Proc. on Biometrics: Theory, Applications, and Systems*, 2009.

[17] M. Teli, J. Beveridge, P. Phillips, G. Givens, D. Bolme, and B. Draper. Biometric zoos: Theory and experimental evidence. In *Int. Joint Conference on Biometrics*, 2011.

[18] M. Wittman, P. Davis, and P. Flynn. Empirical studies of the existence of the biometric menagerie in the FRGC 2.0 color image corpus. In *Proc. of Computer Vision and Pattern Recognition Workshop*, 2006.

[19] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(2):220 –230, 2010.