

Image specific error rate: A biometric performance metric

Elham Tabassi

National Institute of Standards and Technology

Email: elham.tabassi@nist.gov

Abstract—Image-specific false match and false non-match error rates are defined by inheriting concepts from the biometric zoo. These metrics support failure mode analyses by allowing association of a covariate (e.g., dilation for iris recognition) with a matching error rate without having to consider the covariate of a comparison image. Image-specific error rates are also useful in detection of ground truth errors in test datasets. Images with higher image-specific error rates are more “difficult” to recognize, so these metrics can be used to assess the level of difficulty of test corpora or partition a corpus into sets with varying level of difficulty. Results on use of image-specific error rates for ground-truth error detection, covariate analysis and corpus partitioning is presented.

Keywords-biometric performance; DET, covariate analysis; failure analysis; image or corpus difficulty

I. INTRODUCTION

It is known that different users exhibit different levels of recognizability in biometric recognition systems. Some people are easy to recognize, while others can impersonate or be impersonated. The literature makes the analogy between the various Type I and Type II error rate heterogeneities and a biometric zoo.

The issue of performance variability among different users was first addressed by Campbell et al. [1]. Later, Doddington et al. [2] developed a statistical framework to identify four categories of speakers based on the recognition error of each speaker. Specifically, they introduced:

- ▷ Sheep - users who are recognized easily;
- ▷ Goats - users who are particularly difficult to recognize;
- ▷ Lambs - users who are particularly easy to imitate;
- ▷ Wolves - who are users that are particularly successful in imitating others.

Others [4], [7], [8] have investigated the existence of a biometric menagerie in face and fingerprint recognition systems. More recently Yager and Dunstone [9] introduced four new groups of animals.

Cappelli et al. [12] proposed a metric for intrinsic difficulty of individual fingerprints by averaging overall FNMR (or FMR) when each genuine comparison score (or impostor comparison score) of a particular fingerprint is set as threshold. We propose a different way of measuring image difficulty: we measure the difficulty of an image (or image specific error) at a particular (global) threshold as opposed to [12] which averages error rates computed at all possible thresholds (i.e. the set of genuine comparison

scores involving the image), and in computing an image specific error rate, we only consider comparison scores that the image was involved in, while [12] uses all comparison scores.

Recognizing the user-dependent performance variability, Poh et al. [6] ranked users based on the strength of their performance and stated that this information could be used to do multi-modal fusion on a per-user basis by selecting only a subset of the most discriminative biometric traits for each person, rather than using all available modalities.

This non-uniform performance is of interest to the designers of biometric recognition systems. The difficult-to-recognize users are responsible for the major share of biometric errors. Goats contribute to false non-match rate (FNMR) but this poor performance in genuine comparisons does not necessarily elevate false match rate (FMR). Goats are particularly problematic in access control systems where reliable, convenient, verification of users is the main interest (i.e. low FNMR is desirable). Wolves and lambs adversely affect the security of biometric systems by contributing to the FMR. Their biometric samples tend to match impostors, or be matched by impostors. Similarly, different images of the same subject could exhibit different levels of matchability. Image performance variation is often ascribed to the capture device (e.g., different physical imaging properties of sensors), the environment (e.g., low light) or the user (e.g., squinting), and the thrust of research is therefore to make recognition algorithms more tolerant of such variations. Stated another way, algorithms that bound or constrain FNMR and FMR are more reliable and secure. To reduce false non-matches and improve reliability, it is a common policy to allow multiple acquisitions of the same biometric at the time of authentication (e.g., to re-acquire after a moistening of the finger). Dealing with false match occurrences, however, is a more difficult problem. In operational verification systems, false matches are likely to be undetected; in identification systems they lead to spurious entries on candidate lists and these elevate workload.

The cause of the performance variations is interesting. Beveridge et al. [10] developed a framework for biometric covariate analysis and investigated factors affecting performance of face recognition algorithms in the Face Recognition Grand Challenge [11]. For other modalities (iris or finger), the extent to which combinations of image (or user) covariates and/or matching algorithm might cause this is,

to our knowledge, unreported. Another interesting problem is whether, regardless of the matching algorithm, there are wolves (or goats) at large (i.e. users or images that account for disproportional share of the overall FMR or FNMR). We intend to pursue both subjects in the future. For now, we focus on investigating a) how to quantify the level of difficulty of an image (and so a dataset) and b) the ability of matching algorithms to produce comparison scores that are robust to variation in image (or user) covariate. In other words, if an algorithm operating at a fixed threshold could maintain a relatively constant false match rate or false non-match rate regardless of image (or user) properties.

The rest of this paper is organized as follows: we introduce image-specific error rates in section II. Results and use cases for these metrics are presented in section III which precedes conclusion. The concept presented in this paper is modality independent, but results are presented for iris recognition technology.

II. IMAGE SPECIFIC ERROR RATES

To examine performance variation among different images, we define the following image-specific error rates:

- ▷ Image false match rate iFMR - the proportion of comparisons for which an image produces false matches (i.e. for dis-similarity scores, non-match comparisons at or below the operating threshold).
- ▷ Image false non-match rate iFNMR - the proportion of comparisons for which an image produces a false non-match (i.e. for dis-similarity scores, genuine comparisons above the operating threshold).

Specifically, if we define s_{kl}^{ij} to be the comparison score of the k -th image of subject i with the l -th image of subject j then the set of impostor scores of the k -th image of subject i is

$$\mathcal{I}(i, k) = \{ s_{kl}^{ij}, i \neq j, j = 1 \dots J, l = 1 \dots N_j \} \quad (1)$$

for comparison against all N_j images of all J persons in an enrolled set. The image false match rate is then defined as

$$\text{iFMR}(\tau, i, k) = \frac{\sum_{s \in \mathcal{I}(i, k)} 1 - H(s - \tau)}{\sum_{s \in \mathcal{I}(i, k)} 1} \quad (2)$$

where $H(s)$ is the step function defined here as

$$H(x) = \begin{cases} 0 & x \leq 0 \\ 1 & x > 0 \end{cases} \quad (3)$$

If the threshold is set to τ in the conventional manner (i.e. over some large cross comparison set) to give a global FMR of f , then the general case is that $\text{iFMR} \neq f$.

For the image false non-match rate, we use the set of non-self genuine scores of the k -th image of subject i

$$\mathcal{G}(i, k) = \{ s_{kl}^{ii}, l = 1 \dots N_i, k \neq l \} \quad (4)$$

to compute

$$\text{iFNMR}(\tau, i, k) = \frac{\sum_{s \in \mathcal{G}(i, k)} H(s - \tau)}{\sum_{s \in \mathcal{G}(i, k)} 1} \quad (5)$$

where $H(x)$ is again the step function of equation 3.

The threshold (τ) can be set to any value. Given the number of per image genuine and impostor scores that were available, threshold was set such that over all impostor comparisons $\text{FMR} = 0.001$ is achieved. Computation of iFNMR requires multiple (non-self) genuine comparison scores, which means multiple images of the same subject biometric is needed. Furthermore, it should be computed at thresholds that false match error actually occurs.

For any given image, its image-specific error rates will be expected to vary when computed using comparison scores of different algorithms or different thresholds. This variations are addressed here.

A. Dependence on comparison algorithm

An image with a high FMR or FNMR for one comparison algorithm, might result in low image error rates when a different comparison algorithm is applied. In other words, a difficult to process and recognize image for one algorithm, might be easy for another algorithm.

To examine the dependence of image errors on comparison algorithms, we computed image error for several comparison algorithms for 31,415 images of ICE2006 corpus [5]. Figure 1 shows image error rate for two different iris recognition algorithms. The dotted black lines mark the FMR and FNMR at the operating threshold of $\text{FMR} = 0.001$. The spread of points demonstrates image performance variability. The ideal case is when a matching algorithm produces constant false matches and false non-match (with a very small spread) for any image regardless of its underlying properties and quality. However, the relatively wide spread and heavy tails of the distributions in figure 1b suggest that it is not the case in the real world. It can be seen that the spread of the blue cloud (and the histogram of image errors) is different for the two algorithms. That means an image that presents wolf-like behavior to one algorithm (i.e. its iFMR is larger than overall FMR of the system at the operating threshold) might be a sheep when matched by a different comparison algorithm (i.e. not causing any false matches). This is a strong argument for fusion of matching algorithms.

B. Dependence on operating threshold

We studied variation in image specific error rates computed at two different operating thresholds that give overall FMR of 0.001 and 0.0001. Image errors computed at operating threshold of $\text{FMR} = 0.0001$ results in larger iFNMR and slightly smaller iFMR than image errors computed at operating threshold of 0.001. Changes in both iFMR and iFNMR were significant. It is desirable to measure image difficulty independent of operating threshold.

C. Biometric zoo

Images can be categorized according to their level of recognizability (or difficulty):

CLEAR ICE These are images for which iFMR is less than the nominal FMR, and iFNMR is less than the nominal FNMR. These images occupy the lower left quadrant of the plots of figure 1 and may well be considered *easy* to recognize.

BLACK ICE These images occupy the upper right quadrant. They are the most challenging images of the ICE2006 dataset since their image error rates are higher than the nominal error rates indicated by the dotted black line.

BLUE GOATS Images in the top left quadrant have $iFNMR > FNMR(\tau)$ and $iFMR \leq FMR(\tau)$. These are more frequently falsely rejected but do not attract false matches.

BLUE WOLVES Images residing in the bottom right quadrant have $iFMR > FMR(\tau)$ and $iFNMR \leq FNMR(\tau)$. These images are implicated in more false matches and are generally easy to match. Further we compute an *aggregate* iFMR as the arithmetic mean of image false match rates over several (in this case 19) comparison algorithms. Similarly the aggregate iFNMR is the arithmetic mean of image false non-match rates over the 19 different algorithms. Aggregate iFNMR and iFMR are primarily useful in assessing the difficulty of an image (and so the corpus if proper summarization is performed) because image error rates are computed across a diverse set of comparison algorithms. (Results not shown due to space limitation.)

III. USES OF IMAGE SPECIFIC ERROR RATES

This section presents results on uses of image specific error rates for image covariate analysis, assessing level of difficulty of image (or dataset) and ground-truth validation.

Failure mode analysis: Image specific error rates support failure mode analyses by allowing association of a covariate with a matching error rate without having to consider the covariate of a comparison image. As an example, use of image specific error rate in assessing the predictive power of iris image quality scores is presented below. An effective quality algorithm should assign the highest scores to CLEAR ICE images and the lowest to BLACK ICE images [3]. BLUE GOATS and BLUE WOLVES should have quality scores in between. Any other result is undesirable. Three different quality algorithms were used to assess quality of images in ICE2006 corpus. Table I shows correlation coefficients between aggregated image error and quality scores of these three algorithms. Two of them (algorithms X and Y) assign higher quality scores to images with lower image error rate, which is exhibited by a modest negative correlation. The weak correlation between the other algorithm’s quality scores (algorithm Z) and image-specific error rates indicates the lack of a strong relationship between quality algorithm Z and image-specific error rates. Algorithm Z assigned lower quality scores to BLACK ICE images, but only slightly higher

Table I
SPEARMAN CORRELATION COEFFICIENTS FOR QUALITY SCORES OF ALGORITHMS Z, Y AND X AND AGGREGATE IMAGE-SPECIFIC ERRORS. (IFMR AND IFNMR) COMPUTED AT OVERALL THRESHOLD OF IFMR =0.001. ALGORITHM Z QUALITY SCORES SHOW LITTLE CORRELATION WITH IMAGE ERROR, WHILE X AND Y SHOW SOME CORRELATION.

Correlation with quality	Z	Y	X
iFMR	-0.051	-0.403	-0.321
iFNMR	-0.104	-0.236	-0.334

quality scores to CLEAR ICE than BLUE GOATS or BLUE WOLVES.

A. Level of difficulty

Images with higher image-specific error rates are more “difficult” to recognize, so these metrics can be used to assess the level of difficulty of test corpora or partition a corpus into sets with varying level of difficulty. As described in section II, image-specific error rates are used to create four partitions of the ICE2006 dataset: CLEAR ICE, BLUE GOATS, BLUE WOLVES, and BLACK ICE. The latter consists of those images that have pathological error rates on all comparison algorithm. Figure 2 shows example images of each partition.

Level of test corpus difficulty is an important subject in biometric performance evaluations. It can be computed as a summary statistics of its image specific error rates. Results on various summarization techniques is intended to be presented in a future publication.

B. Ground-truth validation

Ground-truth errors is a problem in biometric performance evaluation. Manual examination of all test images is too expensive and often impossible. Images with high iFNMR (e.g. greater than 0.9) are possible ground-truth errors. Figure 2e and 2f shows images with iFNMR =1.0 which have no iris information. Not excluding these images will (erroneously) inflate false non-match rate.

IV. CONCLUSION AND FUTURE WORK

This paper advances image specific error rates as a metric for biometric performance evaluation. It can be used to assess comparison algorithm robustness to image quality variation. It is particularly useful in data mining to explore patterns in biometric images that cause recognition failure. Future work includes multivariate statistical analysis to relate iris image properties such as dilation, contrast, and focus to image specific error rates.

ACKNOWLEDGMENT

This work was supported by the Department of Homeland Security Science and Technology Directorate under project HSHQDC-09-X-00467, RSTS-09-00019. The author is grateful to the sponsor of this work. This publication only reflects the author’s view.

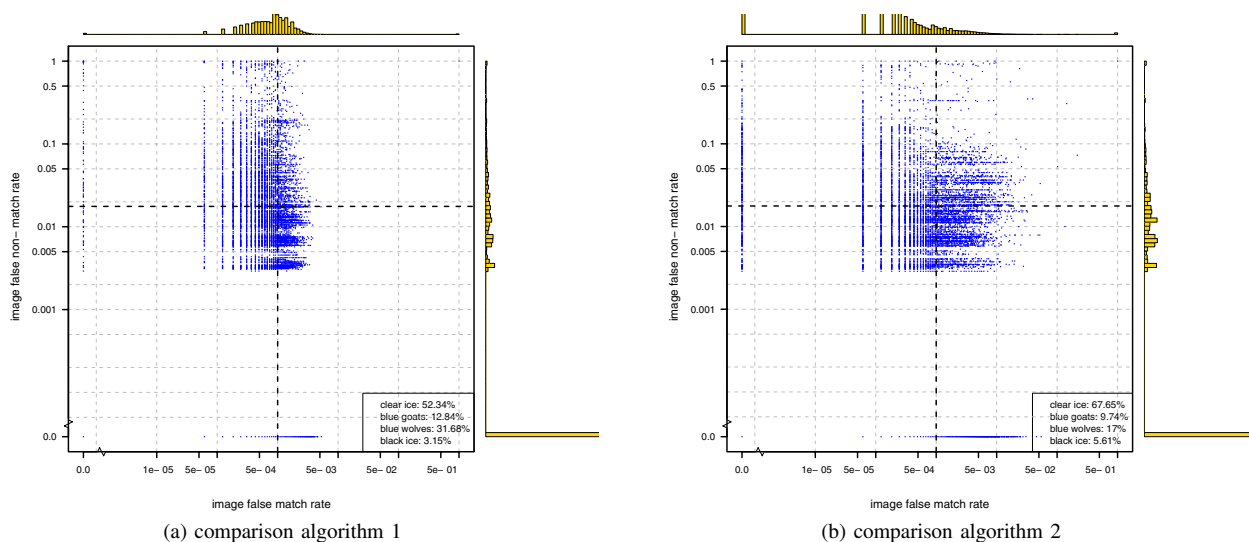


Figure 1. Image FNMR vs. image FMR for 31,415 images of the ICE2006 dataset for two iris recognition algorithms. Image errors are computed at the threshold that gives global FMR = 0.001 for each algorithm. The black dotted lines correspond to overall error rate of the system. Image false non-match (and image false match) probability density is plotted on the top (left side). The relative spread of the image errors suggests comparison algorithm 1 is more robust to image variation than comparison algorithm 2. The legend shows percentage of clear ice, blue goat, blue wolf and black ice images.

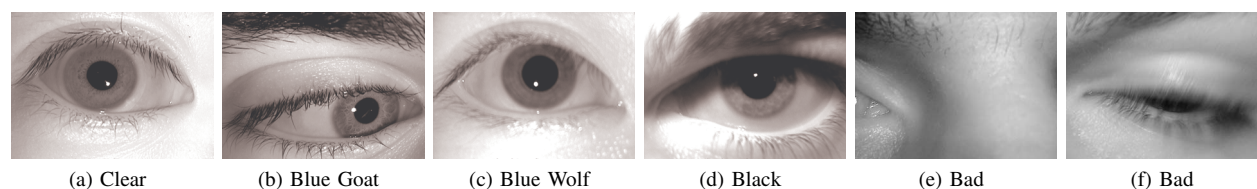


Figure 2. Example images in ICE2006 partition: (a)clear ice($iFMR = 0.0001$, $iFNMR = 0.0$), (b)blue goat ($iFMR = 0.0002$, $iFNMR = 0.746$), (c)blue wolf ($iFMR = 0.1$, $iFNMR = 0.005$), and (d)black ice ($iFMR = 0.024$, $iFNMR = 0.263$) images. (e) and (f) are images with no usable iris information and both have $iFNMR = 1.0$. Images with $iFNMR$ equals to 1.0 are almost always ground-truth bugs and shall be excluded from analyses and performance evaluations. Partition was computed at threshold of $FMR = 0.001$.

REFERENCES

- [1] J. P. Campbell. Speaker recognition: A tutorial. In *Proc. of the IEEE*, volume 85, 1997.
- [2] G. Doddington, W. Liggett, A. Martin, M. Przybocki, and D. Reynolds. Sheep, goats, lambs and wolves: A statistical analysis of speaker performance. In *Proc. Fifth Int'l Conf. Spoken Language Processing (ICSLP)*, pages 1351–1354, 1998.
- [3] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. PAMI*, 29(4):531–543, April 2007.
- [4] A. Hicklin, C. Watson, and B. Ulery. The myth of goats: How many people have fingerprints that are hard to match. Technical report, NIST, 2005. IR7271.
- [5] P. J. Phillips, W. T. Scruggs, A. J. O’Toole, P. J. Flynn, K. W. Bowyer, C. L. Schott, and M. Sharpe. Fvt 2006 and ice 2006 large-scale experimental results. *IEEE Trans. PAMI*, 99(1), 2009.
- [6] N. Poh, S. Bengio, and A. Ross. Revisiting doddington’s zoo: A systematic method to assess user-dependent variability. *Multimodal User Authentication (MMU)*, 13(1):234–778, 2003.
- [7] J. L. Wayman. Multifinger penetration rate and roc variability for automatic fingerprint identification systems, 1999. National Biometric Test Center.
- [8] M. Wittman, P. Davis, and P. J. Flynn. Empirical studies of the existence of the biometric menagerie in the frgc 2.0 color image corpus. In *Proceedings of CPVR ’06 Workshop*, page 33. IEEE Computer Society, 2006.
- [9] N. Yager and T. Dunstone. The biometric menagerie. *IEEE Trans. on PAMI*, 2009.
- [10] J. R. Beveridge, G. H. Givens, J. P. Phillips and B. A. Draper. Factors that Influence Algorithm Performance in the Face Recognition Grand Challenge. *Computer Vision and Image Understanding, Volume 113, Pages 750-762*, 2009.
- [11] J. P. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer and W. Worek. Preliminary Face Recognition Grand Challenge Results. In *Proc. Seventh Int’l Conf. Automatic Face and Gesture Recognition (FGR)*, pages 15–24, 2006.
- [12] R. Cappelli, D. Maio, D. Maltoni, J. L. Wayman, and A. K. Jain. Performance Evaluation of Fingerprint Verification Systems. In *IEEE Trans. PAMI*, pages 3–18, January 2006.