Calculation of a Composite DET Curve

Andy Adler[1] and Michael E. Schuckers[2]

[1] School of Information Technology and Engineering,
University of Ottawa, Ontario, Canada
adler@site.uOttawa.ca
[2] Mathematics, Computer Science and Statistics Department,
St. Lawrence University, Canton, NY, USA and
Center for Identification Technology Research (CITeR)
West Virginia University, Morgantown, WV, USA
schuckers@stlawu.edu **

Abstract. The verification performance of biometric systems is normally evaluated using the receiver operating characteristic (ROC) or detection error trade-off (DET) curve. We propose two new ideas for statistical evaluation of biometric systems based on these data. The first is a new way to normalize match score distributions. A normalized match score, $\hat{t}$, is calculated as a function of the angle from a representation of (*FMR*, *FNMR*) values in polar coordinates from some center. This has the advantage that it does not produce counterintuitive results for systems with unusual DET performance. Secondly, building on this normalization we develop a methodology to calculate an average DET curve. Each biometric system is represented in terms of $\hat{t}$ to allow genuine and impostor distributions to be combined, and an average DET is then calulated from these new distributions. We then show that this method is equivalent to direct averaging of DET data along each angle from the center. This procedure is then applied to data from a study of human matchers of facial images.

1 Introduction

One common way to represent the performance of a biometric classification algorithm is the detection error tradeoff (DET) curve. A sample population containing matching (*genuine*) and non-matching (*impostor*) image pairs is presented to the biometric algorithm and the match score, t, calculated to estimate the genuine ($g(t)$) and impostor ($f(t)$) match score distributions. From these distributions, the DET is typically plotted as the false match rate (*FMR*) on the *x-axis* against the false non-match rate (*FNMR*) on the *y-axis*, by varying a threshold τ, and calculating $FMR(\tau) = \int_{\tau}^{\infty} f(x)dx$ and $FNMR(\tau) = \int_{-\infty}^{\tau} g(y)dy$. The DET summarizes the verification performance of the biometric algorithm on the sample population on which it is calculated. Technology evaluations, such as the

** This work is supported by NSERC Canada (Dr. Adler), and by NSF grant CNS-0325640 (Dr. Schuckers) which is cooperatively funded by the National Science Foundation and the United States Department of Homeland Security

FRVT and FpVTE tests [14][15] use DET curves — or a variant, the Receiver Operating Characteristic (ROC) — to describe their results.

Given its ubiquity, it is perhaps somewhat surprising that few statistical methods have been proposed for analysis and interpretation of DET data in biometric classification. On the other hand, there is a large body of research in the statistical literature, e.g. Zhou et al. [19], and a growing body of work in the machine learning/artificial intelligence literature, e.g. Hernández-Orallo et al. [10]. ROC analysis is used in a wide variety of classification settings including radiography, human perception, and industrial quality control. Zhou et al. ([19]) provide a excellent overview of this work. One limitation of inferential tools for ROC's is the common assumption of Guassian distributions for $g(t)$ and $f(t)$, e.g. Green and Swets [6]. The methodology we propose here does not depend on any distributional assumptions. Another focal area for this research has been the area under the curve or AUC, e.g. Hanley and McNeil [9]. However, biometric authentication has emphasized the equal error rate (EER) as an overall summary of system performance rather than the AUC.

Although most of the literature analyses the ROC, we focus on DET curves since they are more commonly used in biometric identification systems. Here we are motivated to develop methods for a composite DET curve given classification pairs from multiple sources $\mathit{FMR}(\tau)$, $\mathit{FNMR}(\tau)$ in which the original genuine and impostor distributions are either lost, or the match score values, t, are calculated in different spaces. Four types of DET or ROC averaging have been proposed. Bradley [2] suggests using an average based upon the i^{th} ordered threshold in DET space. However, this method leads to difficulties when the number of thresholds tested varies greatly from curve to curve. Vertical averaging (along the *FMR*) has been suggested by Provost et al. [17], but this method is only appropriate if one of the error rates is more important for some *a priori* reason. When the data to be averaged have very different error rates this method can produce very non-intuitive results, such as if one system reaches $\mathit{FNMR} = 1.0$ at non-zero *FMR*. Fawcett [5] proposes averaging at the thresholds; however, this method fails when the systems use different match score scales. Finally, Karduan et al. [12] proposed averaging the log-odds transformation of one error rate given the other. In this paper we propose a new method for averaging based on the radial sweep methodology of Macskassy and Provost [13]. This approaches, described below, transforms each curve from the (*FMR*, *FNMR*) space to polar coordinates.

In this paper we were specifically motivated by how to average the separate DET curves of human volunteers who were asked to perform face recognition [1], by evaluating the whether pairs of images were of the same individual. There are few other reports of comparisons of human face recognition performance to that of automatic systems. Burton and collaborators [3][8] compared PCA based and graph-matching algorithms against human ratings of similarity and distinctiveness, and human memory performance. These studies were focussed on the extent to which automatic algorithms explain features of human performance, rather than as a comparison of recognition performance levels. These studies did not pursue advanced statistical techniques to synthesize an average masure of human performance. As is typical with data collected from subjective evaluations, assessed values cannot be directly compared between participants. However, in order to compare human face recognition performance levels to each other and to those of automatic software, we wanted a way to calculate the composite human face recognition performance. Because a DET is inherently a two dimensional curve it is difficult to average the curves in a way that properly maintains the importance of both dimensions. In order to address this problem, we develop a technique to calculate an average DET based on regeneration of normalized match scores and distributions. We then show that this is equivalent to a geometrical averaging directly on the DET curves.

The rest of this paper is organized in the following manner. Our method for a composite DET is described in section 2. We then apply this method to data from a group of human subjects (section 3). Finally, in Section 4 and we discuss the applicability of this technique for analysis and interpretation of biometric system verification results.

2 Methods

We use the following notation. A collection of J biometric score distributions are available; each is measured in terms of its own match score $t_i, i = 1 \ldots n_j$. There are no conditions on the match scores other than they be scalar, and increase with match likelihood. The genuine and impostor distributions are represented as $f_j(t_i)$ and $g_j(t_i)$, respectively for $j = 1 \ldots J$. Based on these distributions, the false match rate (FMR_j) and false non-match rate ($FNMR_j$) for biometric system j may be calculated as

$$FMR_j(\tau) = \int_{\tau-}^{\infty} f_j(t)dt = 1 - \int_{-\infty}^{\tau+} f_j(t)dt \tag{1}$$

$$FNMR_j(\tau) = \int_{-\infty}^{\tau-} g_j(t)dt \tag{2}$$

by varying the threshold τ. Clearly, real biometric match score data are not continuous, in which case sums must be used instead of integrals. In this case, it is important that the calculation of either *FMR* or *FNMR* but not both, include the distribution value at τ; we include it in the *FMR*. Implicitly this assumes that the decision process is to accept if the match score is greater than or equal to the threshold, τ. This calculation is illustrated in Fig. 1.

2.1 Normalized match scores via polar coordinates

In order to perform further analysis on multiple DET curves, it is necessary to calculate a normalized match score common to all curves. In this section, we describe an approach, based on representing the curve in polar coordinates, as illustrated in Fig. 1.

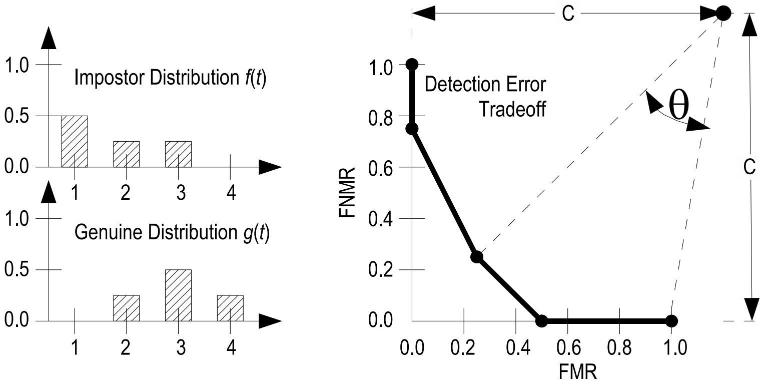

Fig. 1. Calculation of *FMR* and *FNMR* from sample distributions and regeneration of match score t using polar coordinates. Given the discrete *genuine* and *impostor* distributions shown on the left, the DET curve on the right is calculated. From a center at (c, c) an angle θ is calculated to each $FMR, FNMR$ point. A normalized match score t is then calculated from θ. In this example, the distributions are discrete, and the DET curve uses a linear interpolation between points.

We have $FMR, FNMR$ coordinate pairs $(x_{ij}, y_{ij}), i = 1, \ldots, n_j; j = 1, \ldots, J$ for a series of J DET curves. By the monotonicity of the DET curves, we know that $x_{1j} \leq x_{2j} \leq \ldots \leq x_{n_j j}$ and $y_{1j} \geq y_{2j} \geq \ldots \geq y_{n_j j}$.

We also assume that no other information is available that would assist us in knowing how the knots in the splines are selected. These points are, as is made clear below, a function of some threshold, τ. Equivalently, we are assuming that no information is available concerning the threshold values. (For example, it would be possible to assume that the thresholds are equally spaced and to derive approximate genuine and imposter distributions following such an assumption.)

Thus, from the DET curve, we calculate an angle

$$\theta_{ij} = tan^{-1}\left(\frac{c - x_{ij}}{c - y_{ij}}\right). \tag{3}$$

We define an angle with respect to the bottom-right of the DET, since at $\tau = -\infty$, $FMR = 1$ and $FNMR = 0$. The DET curve moves left and upward with increasing τ. The limits for θ are

$$\theta_{min} = tan^{-1}\left(\frac{c-1}{c}\right) \tag{4}$$

$$\theta_{max} = tan^{-1}\left(\frac{c}{c-1}\right) \tag{5}$$

Since we wish to calculate a normalized match score $\hat{t}$ in the range $0, \ldots, 1$ from θ, we define

$$\hat{t} = \frac{\theta - \theta_{min}}{\theta_{max} - \theta_{min}} \tag{6}$$

2.2 Distributions from DET curves

In this section, we use the polar-coordinate representation, to reconstruct candidate genuine, $\hat{g}(\hat{t})$, and impostor, $\hat{f}(\hat{t})$ distributions. Based on the equations 1 and 2, we calculate for each DET curve j.

$$f_j(\hat{t}) = -\frac{dFMR_j}{d\hat{t}} \tag{7}$$

$$g_j(\hat{t}) = \frac{dFNMR_j}{d\hat{t}}. \tag{8}$$

Fig. 2 illustrates the calculations. Since *FMR* and *FNMR* data are not continuous, but are sampled from the DET, the distributions must be defined in terms of discrete approximations to the derivative. One consequence of the discrete derivative is that $\hat{g}$ and $\hat{f}$ are noisy, but this does not matter for this application.

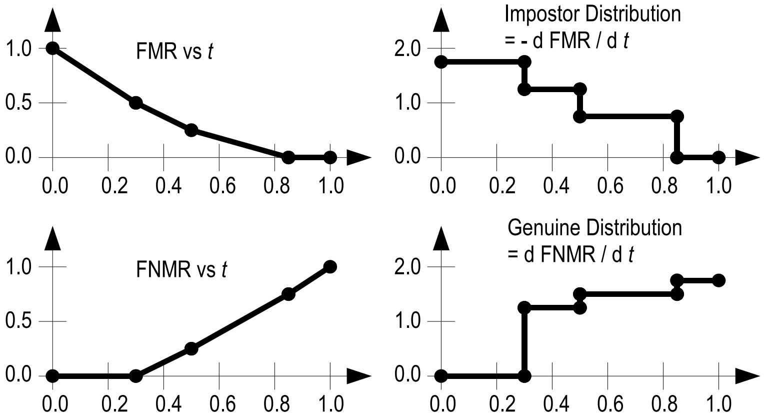

Fig. 2. Reconstructed genuine, $\hat{g}(\hat{t})$, and impostor, $\hat{f}(\hat{t})$, distributions: From the DET curve of Fig. 1 the *FMR* (upper left) and *FNMR* (lower left) are calculated as a function of the normalized match score $\hat{t}$. From these curves, the *impostor* (upper right) and *genuine* (lower right) distributions are calculated as $-\frac{d}{d\hat{t}}FMR$ and $\frac{d}{d\hat{t}}FNMR$, respectively.

Using this calculation, we now have a collection of distributions $\hat{g}_j, \hat{f}_j$ for $j = 1 \ldots J$, which are all based on the same match scores, $\hat{t}$'s. It is thus possible to combine the distributions, weighted by the number of samples in each (if known). The number of samples in each genuine and impostor distribution are represented as $n_{g,j}$ and $n_{f,j}$, respectively. If the number of samples is unknown, all n values are assumed to be equal. The combined distributions $\bar{f}$ and $\bar{g}$ are

$$\bar{f} = \frac{1}{N_f} \sum_{j=1}^{J} n_{f,j} \hat{f}_j \tag{9}$$

$$\bar{g} = \frac{1}{N_g} \sum_{j=1}^{J} n_{g,j} \hat{g}_j \tag{10}$$

where $N_f = \sum n_{f,j}$ and $N_g = \sum n_{g,j}$.

However, this expression may be shown to be equivalent to a direct averaging of the DET curves in $(\mathit{FMR}, \mathit{FNMR})$ space, as follows:

$$\begin{aligned}
\hat{\mathit{FNMR}}(\hat{t}) &= \int_{-\infty}^{\tau-} \bar{g}(t) dt &(11)\\
&= \int_{-\infty}^{\tau-} \frac{1}{N_g} \sum_{j=1}^{J} \frac{1}{dt} d\mathit{FNMR}_j(t) dt &(12)\\
&= \int_{-\infty}^{\tau-} \frac{1}{N_g} \sum_{j=1}^{J} n_{g,j} \frac{1}{dt} d\mathit{FNMR}_j(t) d\hat{t} &(13)\\
&= \frac{1}{N_g} \sum_{i=1}^{J} n_{g,j} (\mathit{FNMR}_j(\hat{t}) - \mathit{FNMR}_j(-\infty)) &(14)\\
&= \sum_{j=1}^{J} \frac{n_{g,j}}{N_g} \mathit{FNMR}_j(\hat{t}) &(15)
\end{aligned}$$

Similarly,

$$\hat{\mathit{FMR}}(\tau) = \sum_{j=1}^{J} \frac{n_{f,j}}{N_f} \mathit{FMR}_j(\hat{t}) \tag{16}$$

Thus, the average DET at each angle θ can be calculated by a (possibly weighted) average the distance of each curve from (c, c).

3 Results

This paper uses data from a comparison of human and automatic face recognition performance [1]. This study investigated the ability of interested and motivated non-specialist volunteers to perform face identification tasks matched against performance by several commercial face recognition software packages. Images were obtained from the NIST mugshot database [16]. Pairs of frontal pose face images were randomly created from this database. Two-thirds of the pairs were impostors (images of different persons), and one third were genuines (different images of the same person). No special effort was made to select images of the same gender or ethnicity for the impostor pairs.

Twenty one people (16 male, 5 female) participated in the experiments. They were predominantly Caucasian and in the age range 20–40. Participants were asked to log onto a web site, where an application server would present pairs of face images, and the participant was asked whether they were from the same person. Participants were not given any information about the distribution of genuines and impostors, or any feedback about their success. Participants were presented the following options: *same*, *probably same*, *not sure*, *probably different*, or *different*. Each option was converted to a match score value (such that *different*= 1 and *same*= 5).

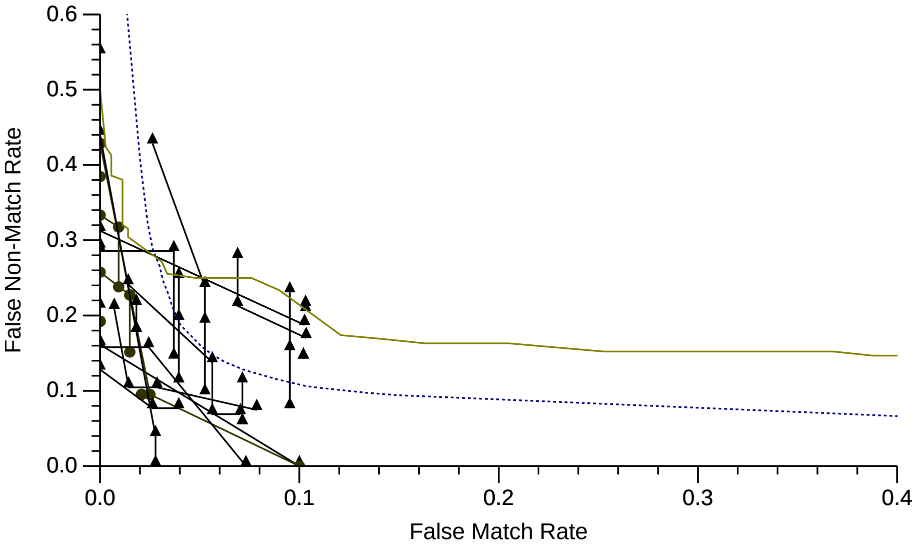

Fig. 3. Calculation of an average DET curve for human face recognizers. Individual human DET curves are shown by symbols (circle=female, triangle=male). The average curve (dotted line) is calculated using the method of this paper. For comparison, the highest performing software available to us in 2003 is also shown (solid line).

4 Discussion

In this paper we have presented a new methodology for combining and averaging DET or ROC curves. This approach was motivated by the need to create a composite DET curve for human evaluators of human faces. This methodology was developed independently of [13]; however, it uses the same basic technique of radially sweeping across the DET curve to create a normalized match score. This permits the creation of normalized distributions for *FMR* and *FNMR* that are a composite of individual DET curves. This normalization is a significant advance in and of itself and adds to a growing body of methods for this purpose

[11]. We have used this normalization to to average at normalized radial match scores.

Several issues arise from radial sweeping of DET curves. The first is where to locate the center of the sweeping. Because we would like the averaging to not depend on which error rate is on which axis, we limited possible center points to (c, c) for some constant c. It is immediately clear that choosing a center along the *FMR* = *FNMR* line results in an average curve that is independent of the selection of axes. We considered three possible values for c, $0, 1$ and ∞. Choosing $c = 0$, often resulted in composite or average curves that were counter-intuitive because of the acute angles near the axes. This is especially important for biometric systems which are often placed in settings where low *FMR*'s are required. There was little difference between the curves when $c = 1$ and $c = \infty$. However, we prefer $c = 1$ because the radial angles match the typical curvature of a DET curve and, hence, are more likely to be perpendicular to such curves. The choice of $c = \infty$ results in averaging across parallel 45° lines.

Another issue is the choice of how to "average" the curves. Here we have effectively taken an arithmetic average of the curves. Other choices are possible including a weighted average, to account for database size or importance by varying the weights to be given to each DET. An alternative would be to use a radial median at each angle. This would results in a spline that is not as smooth as the radial mean DET but which may be more robust to "outlying" DET curves.

The question of inferential methods based on the radial mean DET is one that is important for future study. Here we are interested in creating confidence intervals for an individual curve (as in [13]) as well as being create a confidence interval for the difference of two DET curves. Similarly we would like to create tests for significant differences between two or more DET curves. It might also be of interest to test a single observed DET against a hypothetical DET curve. This last case may take the form of a Kolmogorov-Smirnov type test.

References

1. Adler, A., Maclean, J.: "Performance comparison of human and automatic face recognition" *Biometrics Consortium Conference 2004* Sep. 20-22, Washington, DC, USA (2004)
2. Bradley, A. P.: "The use of the area under the ROC curve in the evaluation of machine learning algorithms." *Pattern Recognition* **7**, 1145-1159 (1997).
3. Burton, A. M., Miller, P., Bruce, V., Hancock, P. J. B., Henderson, Z.: "Human and automatic face recognition: a comparison across image formats" *Vision Research*, 41:3185-3195, 2001.
4. Drummond, C., Holte, R. C.: "What ROC Curves Can't Do (and Cost Curves Can)" In *Proc. 1st Workshop ROC Analysis in AI:ROCAI*, 19-26, (2004).
5. Fawcett, T.: *ROC graphs: Notes and practical considerations for data mining researchers*, Technical Report HPL-2003-4. HP Labs. (2003).
6. Green, D. M., Swets, J. A.: *Signal Detection Theory and Psychophysics.* John Wiley & Sons, New York, (1966).

7. Golfarelli, M., Maio, D., Maltoni D.: "On the Error-Reject Trade-Off in Biometric Verification Systems" *IEEE Trans. Pattern Anal. Machine Intel.* **19** 786–796 (1997)
8. Hancock, P. J. B., Bruce, V., Burton, M. A.: "A comparison of two computer-based face identification systems with human perceptions of faces" *Vision Research* 38:2277-2288 (1998).
9. Hanley, J. A., McNeil, B. J.: "The meaning and use of the area under a receiver operating characteristic (ROC) curve" *Radiology* **143** 29–36 (1982).
10. Hernández-Orallo, J., Ferri, C., Lachiche, N. Flach, P.A.,ed.: *ROC Analysis in Artificial Intelligence, 1st Int. Workshop, ROCAI-2004*, Valencia, Spain (2004).
11. Jain, A.K, Nandakumar, K.: Ross, A.: "Score Normalization in Multimodal Biometric Systems", *Pattern Recognition*, (in press, 2005).
12. Karduan, J., Karduan, O.: "Comparative diagnostic performance of three radiological procedures for the detection of lumbar disk herniation" *Methods Inform. Med.* **29** 12-22 (1990).
13. Macskassy, S., Provost, F.: "Confidence Bands for ROC Curves: Methods and an Empirical Study." In *Proc. 1st Workshop ROC Analysis in AI:ROCAI*, 61-70, (2004).
14. NIST: *Face Recognition Vendor test 2002* `http://frvt.org/frvt2002`
15. NIST: *Fingerprint Vendor Technology Evaluation (FpVTE) 2003* `http://fpvte.nist.gov/`
16. NIST: *NIST Special Database 18: Mugshot Identification Database (MID)* `http://www.nist.gov/srd/nistsd18.htm`
17. Provost, F. J., Fawcett, T., Kohavi, R.: "The case against accuracy estimation for comparing induction algorithms" In *Proc. 15th Int. Conf. Machine Learning*, 445-453 (1998).
18. Rukhin, A., Grother, P., Phillips, P.J., Newton, E.: "Dependence characteristics of face recognition algorithms" *Proc. Int. Conf Pattern Recog.* **16** 36–39 (2002)
19. Zhou, X.-H., McClish, D. K., Obuchowski, N. A.: *Statistical Methods in Diagnostic Medicine* John W. Wiley & Sons, (2002).