Rethinking the effective assessment of biometric systems

Yingzi Du and Chein-I Chang

The receiver operating characteristic curve is widely used to evaluate recognition systems, but it has too many limitations to serve as a sole criterion.

Biometrics is a means of identifying and verifying people based on physiological characteristics. Popular forms of the technology seek to recognize fingerprint, face, or the iris of the eye. Compared with traditional methods, biometrics is more convenient for users, less vulnerable to fraud, and more secure. It is becoming an important ally of e-commerce, law enforcement, and homeland security, to name just a few applications. The value of a biometric system depends on its capacity to correctly accept or reject the identity of an individual. Incorrectly identifying someone or incorrectly rejecting them are problems. The effectiveness of a biometric system can be represented graphically by the so-called receiver operating characteristic (ROC) curve. The ROC plots the effectiveness of a system with reference to correct identifications and misidentifications.

Figure 1. (a) A receiver operating characteristic (ROC) curve. (b) The alternative ROC curve of (a). FAR: False alarm rate. FRR: False rejection rate.

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>n</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification results</td>
<td>TN</td>
<td>FN</td>
</tr>
<tr>
<td>p</td>
<td>FP</td>
<td>TP</td>
</tr>
</tbody>
</table>

We can illustrate the utility of biometrics by considering a so-called two-class problem. Let’s say a person is being identified by a biometric system. The possible outputs are either positive, \( p \) (verified as the person in the system database) or negative, \( n \) (identified as someone who is not in the database). There are four possible outcomes (see Table 1). If the output is \( p \) and this person is the person in the database, the result is a true positive (TP); however, if this person is not really the person in the database, the outcome is a false positive (FP). Conversely, a true negative (TN) results when the output and the actual identity are \( n \), and a false negative (FN) when the result is \( n \) but the actual identity is \( p \).

The detection power (also called the true positive rate) is the fraction of all positives that are correctly classified as positive. The false rejection rate (FRR) is the fraction of all positives that are incorrectly classified negative. And the sum of the detection power and the false rejection rate is unity: in other words, assessment of all positives can be classified as one or the other. The false alarm rate (FAR) is the proportion of all negatives that are incorrectly classified positive. Figure 1(a) shows an ROC curve based the relationship between detection power and the FAR. Figure 1(b) presents an alternative way of plotting the ROC curve of Figure 1(a) based on the relationship between the FAR and FRR.

Despite decades of application in fields that range from medicine to psychology to machine learning, the ROC curve has a number of features that limit its effectiveness.

The first limitation is that cost is not reflected in the ROC curve. The design of a biometric system required for a highly secured environment will be very different from one used for personal computer log-ins. In the high-security scenario, even one falsely accepted terrorist or criminal can cause substantial damage to the facility. For that reason, the FAR is a high priority. In contrast, for personal home computer log-ins, convenience is an important consideration, and the FRR counts more.

Continued on next page
Figure 2. An example of two crossed ROC curves with the same equal error rate (EER).

Unfortunately, the ROC curve cannot reflect the cost of classification, that is, errors. The equal error rate (EER) in an ROC curve—the point where the FAR equals the FRR—can be misleading. Figure 2 shows two crossed ROC curves with identical EERs. For situations where security is paramount, the FAR is weighted more and system 1 (the solid curve) would be preferable. For consumer electronics applications, on the other hand, system 2 (the dashed curve) has the advantage.

The second limitation is that the ROC curve gives no indication of optimal threshold. It is not sensitive to the bias of a system to misclassify one way or the other and, more important, the ROC curve cannot predict the optimum threshold for a system or the threshold’s accuracy.

Thirdly, the ROC curve ignores the amount of data. Database size is a critical parameter affecting biometric accuracy. For the same system, the FRR and FAR will increase along with the size of the database. Yet ROC curves say nothing about how large a data set is. It is impossible to compare ROC curves of biometric systems tested on different information repositories.

Finally, variable data affect the ROC curve’s predictive power. The condition of the data can affect the performance of a biometric system. Low quality dramatically decreases accuracy, yet ROC curves do not reflect the state of the data used in recognition. Consequently, when quality is variable, it is difficult to predict a biometric system’s performance based purely on its ROC curve.

There are additional concerns to. Other factors can affect the accuracy and performance of biometric systems that ROC curves cannot measure. Examples include recognition time, testing and evaluation protocol, template size (the amount of computer memory taken up by the biometric data), failure-to-enroll rate (the proportion of the population of end-users who fail to complete enrollment), comfort, convenience, and acceptability.

Possible solutions
We propose a 3D combinational accuracy curve, shown in Figure 3, as one way of obtaining a balanced assessment of FAR, FRR, threshold \( T \), and cost. Six 2D curves can be derived from the 3D combinational accuracy curve: the conventional 2D ROC curve; the 2D curve of \((FRR, T)\); the 2D curve of \((FAR, T)\); the 2D curve of \((FRR, cost)\); the 2D curve of \((FAR, cost)\); and the 2D curve of \((T, cost)\). A 3D combinational performance curve can be derived from the 3D combinational accuracy curve, which weighs security, convenience, \( T \), and cost. Overall, these curves provide more comprehensive information about system accuracy and performance than the ROC curve alone.

In addition, different systems should be tested and evaluated using identical database(s) and testing and evaluation protocols. The National Institute of Standards and Technology has taken a lead in evaluating and testing biometric systems and algorithms, such as the Iris Evaluation Challenge, the Face Recognition Grand Challenge, and the Face Recognition Vendor Test.

Appropriate metrics for data quality should be included when evaluating system performance and accuracy. For example, the feature information-based method objectively assesses the quality of an iris image, which helps in comparing system accuracy when using different data sets. Because performance in-

Continued on next page
volves a number of factors, measures should be adopted to facilitate comparisons between systems. Examples include, but are not limited to, the 3D combinational accuracy and performance curves, quality assessment, ease of use, and user acceptance.

Author Information

Yingzi Du
Department of Electrical and Computer Engineering
Indiana University-Purdue University Indianapolis
Indianapolis, IN
http://www.ece.iupui.edu/ece/yidu/

Yingzi Du is an assistant professor with the Department of Electrical and Computer Engineering. Her research interests include biometrics, image processing, and pattern recognition. She is a member of SPIE, IEEE, Phi Kappa Phi, and Tau Beta Pi. She received an Office of Naval Research Young Investigator Program award in 2007.

Chein-I Chang
Department of Computer Science and Electrical Engineering
University of Maryland, Baltimore County
Baltimore, MD

Chein-I Chang, now a professor, received his PhD in electrical engineering from the University of Maryland, College Park. He has authored a book titled Hyperspectral Imaging, and has published 90 journal articles. He is a SPIE Fellow and associate editor of IEEE Transactions on Geoscience and Remote Sensing.

References