Automatic Configuration for a Biometrics-Based Physical Access Control System

Michael Beattie, B.V.K. Vijaya Kumar, Simon Lucey, and Ozan K. Tonguz

Carnegie Mellon University; Pittsburgh, PA 15213-3890, USA

Abstract. Selecting appropriate thresholds and fusion rules for a system involving multiple biometric verifiers requires knowledge of the match score statistics for each verifier. While this statistical information can often be measured from training data, that data may not be representative of the environment into which each verifier is deployed. To compensate for missing statistics, we present a technique for estimating the error rates of each verifier using decisions made after a system has been deployed. While this post-deployment data lacks class labels, it is guaranteed to be representative. Extracted error rates can be used to select appropriate fusion rules and search for thresholds that meet operational requirements.

1 Introduction

Physical access control represents an important application for biometric verification that is already available as an optional component in high end building security systems. We consider the next step in biometrics-based access control: multiple cooperating biometric verifiers to protect multiple security zones [1]. Decisions from these verifiers can be combined to balance security and convenience as appropriate for the resource being protected.

Configuration represents a major challenge for this type of system. Each biometric verifier must be configured with an appropriate threshold, and rules for combining decisions must be determined. Good thresholds ensure that each single verifier produces useful decisions. For example, if a threshold is too high then a verifier will generate mostly reject decisions—regardless of whether the claimant is authentic or an imposter. Standard techniques for combining decisions attempt to weight individual decisions by their relative accuracy [2, 3]. In both cases, knowledge of the False Accept Rate (FAR) and False Reject Rate (FRR) can be used to select appropriate configuration parameters.

In many applications, configuration of both thresholds and fusion rules can be completed using training data. Matching algorithms are be applied to a database of biometric samples and the resulting match scores can be used to select appropriate thresholds and measure error rates. In other applications, representative training data may not be readily available. Configuration using mismatched training data may lead to poor performance in the deployed system. We expect this to be the case in physical access control where biometric samples can differ significantly from building to building and even within a single building.

S.Z. Li et al. (Eds.): IWBRS 2005, LNCS 3781, pp. 241–248, 2005.

Further intensifying this problem, physical access control is often a single component of a larger building security system that can involve a variety of components from several different vendors. The parties building and installing these systems may not have the expertise to adequately configure biometric devices. In these cases, we expect systems to be left using default thresholds—a practice that we will show to be dangerous.

To facilitate a solution to this problem, we propose a strategy for estimating the FAR and FRR using verification decisions created during normal system operation. These decisions have the potential to provide accurate error rate information because they are formed using samples captured in the exact environment of interest. On the other hand, a lack of class labels (i.e. imposter or authentic) makes direct calculation of error rates impossible.

To overcome this challenge, we have derived an estimation technique based on the Expectation-Maximization (EM) Algorithm [4]. Using this this technique, it is possible to estimate error rates for each individual verifier based on sequences of decisions constructed as a claimant moves through a building. Our evaluation against synthetic verifier sequences with a wide variety of underlying error rates implies that this estimation technique can produce highly accurate estimates of individual verifier error rates.

Knowledge of verifier error rates provide an essential step toward enabling automatic system configuration. Comparing FAR and FRR to security and convenience requirements can indicate whether a local threshold should be adjusted. Additionally, error rates interpreted as estimators for class-conditional error probabilities can be used to determine how the sequence of decisions should be combined.

We proceed in Sect. 2 by further motivating the need for automatic configuration in biometrics-based physical access control systems. Sect. 3 presents the details of our error rate estimation, and Sect. 4 continues with the results of an empirical evaluation of that technique using a large number of synthetic data sets. Finally, we conclude our report in Sect. 5.

2 Motivation

Variation among biometric samples collected in different environments is the primary motivation for employing error rate estimation to configure a deployed set of biometric verifiers. Such variation can result from differing environmental conditions such as lighting and humidity. Variation due to more subtle conditions such as sensor positioning or individual user behavior is also possible. All of these variations will lead to different match score behavior in different buildings or even different sections of the same building. As a result, we cannot expect to find a single verification threshold to work in all environments. Furthermore, if we use the same threshold in two different environments, we may observe dramatically different error rates in each.

To demonstrate an extreme example of this phenomenon we used frontal face images from the Version 1 data set of the Face Recognition Grand Chal-

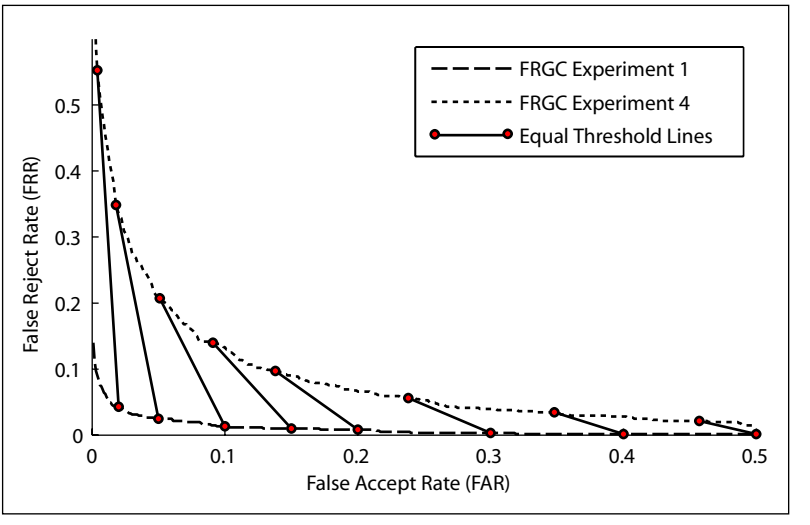

Fig. 1. Difference in performance between FRGC Experiments 1 and 4.

lenge (FRGC) [5]. This particular data set includes samples from two different collection scenarios: controlled and uncontrolled. Samples from the controlled scenario are captured in a studio setting with frontal lighting and a solid background. Samples from the uncontrolled setting are captured in a hallway and are plagued by shadows and irregular backgrounds. The FRGC protocol defines two experiments of interest for our evaluation. Experiment 1 compares controlled templates with controlled samples while Experiment 4 compares controlled templates with uncontrolled samples. To evaluate the difference in verification performance between these two experiments, we constructed match scores for each using a variant of Linear Discriminant Analysis (LDA) described in [6].

Experiments 1 and 4 can be interpreted as a single biometric verification product deployed to two different environments within the same building. In Fig. 2, we present two Receiver Operating Characteristics (ROCs): one for Experiment 1 and another for Experiment 4. Also visible in this figure are a series of equal-threshold lines that connect points on the two curves that correspond to the same threshold. It is clear from this figure that a particular threshold can lead to vastly different error rates in different environments. This figure also demonstrates that the change in performance from one experiment to the other is due largely to changes in FRR. This implies that the distribution of authentic scores is changing while the distribution of imposter scores remains relatively stable. We are unsure if this phenomenon is peculiar to this particular experiment or if it would be true for a wide range of algorithms and data sets.

Changes in accuracy for a fixed threshold suggest that the same biometric verifier deployed in two different locations may require different thresholds to meet the same operating requirements. It is also possible that in some environ-

ments, the verifier may not be able to meet the requirements at all. To motivate the need for automating the process of threshold selection, consider the installation of a biometrics-based physical access control system. Building owners purchase biometric verifiers from a vendor specializing in such devices (often via an intermediate system integrator). Because different thresholds are required for different environments, the problem of choosing a threshold is pushed from the biometric vendor down to the building administrators or system integrators. While biometric vendors are likely to have access to standard test databases, this is not the case for administrators or integrators. Their core competence is likely to fall outside the realm of biometrics, and they should not be required to maintain training data to assist in the configuration of biometric verifiers.

The solution we propose is analogous to a thermostat in a heating system. An identical furnace can be placed in two different buildings, and that furnace may need to be running more often in one building than the other. Building administrators are not required to calculate precisely how often the furnace must be running in each building. Instead, they determine a target temperature and the thermostat turns the furnace on and off as often as necessary to maintain that temperature. This control loop is made possible by the thermostat's ability to measure the current temperature. In much the same way, estimated verifier error rates enable an access controller to adjust verification thresholds and combination rules according to each verifier's operating environment.

3 Error Rate Estimation

The main challenge in estimating error rates from post-deployment verification decisions is a lack of class labels. We cannot be certain whether a particular decision was created from an authentic claimant or an imposter. As a result, we are unable to declare whether or not the decision is in error. There are a number of possible strategies to overcome this problem based on assumptions about the building environment. Most of these potential strategies are ad hoc in nature and rely on assumptions that may not be accurate. For example, we could estimate FRR by asking authentic claimants to report when they are rejected. This assumes that each time an authentic claimant is rejected, he or she will report that event exactly once. This is unlikely to be the case. Another simple strategy for estimating FRR is to assume that the vast majority of claimants will be authentic and simply count every reject decision as a false reject for the purpose of computing error rates.[1] In this case, an unexpectedly large number of impersonation attempts can heavily bias the estimate.

We propose a more structured approach to estimating error rates based on the Expectation-Maximization (EM) Algorithm [4]. The data available for this procedure are sequences of decisions from a single claimant with the true class (authentic or imposter) of that claimant interpreted as a hidden variable. Each sequence is created as a single claimant moves past multiple verifiers (e.g.,

[1] Of course, this does not imply that all claimants will be accepted as authentic; this assumption would be used strictly for error rate estimation.

through multiple doors). A per-subject access token connects a sequence of access attempts at different verifiers. We collect these sequences into vectors and label the j^{th} decision sequence $\boldsymbol{u}_j$ and a the i^{th} decision from that sequence as u_j^i.

Each decision u_j^i is an observation of the random variable u^i representing a decision from verifier i. We assume that u^i takes on a value of 1 to denote an accept decision (i.e. accept the claimant as authentic) and 0 to denote a reject decision. Similarly, the authentic and imposter classes are denoted by ω_1 and ω_0 respectively. Within this framework, we use the class-conditional decision probabilities in (1) and (2) as proxies for FRR and FAR. The EM Algorithm aims to find values for these probabilities that maximize the likelihood of the provided decision sequences.

$$\text{False Reject Probability: } p(\, u^i = 0 \mid \omega_1 \,) \tag{1}$$

$$\text{False Accept Probability: } p(\, u^i = 1 \mid \omega_0 \,) \tag{2}$$

To complete the estimation procedure, EM is initialized with a pair of initial probabilities for each verifier. These initial values may be provided by the biometric vendor or they may be nothing more than guesses. For example, the results in this paper all used initial error probabilities of 0.05. Given these initial values, the EM algorithm iteratively refines the estimated probabilities using (4) where $1\,(\,\cdot\,)$ is the indicator function. As usual, the posterior probability can be calculated with Bayes formula and an assumption of conditional independence as in (4).

$$p(\, \omega_k \mid \boldsymbol{u}_j \,) = \frac{p(\, \omega_k \,) \prod_{l=1}^{N} p(\, u_j^l \mid \omega_k \,)}{p(\, \boldsymbol{u}_j \,)} \tag{3}$$

$$p(\, u_i = k \mid \omega_k \,) \leftarrow \frac{\sum_{j=1}^{T} 1\,(\, u_j^i = k \,)\, p(\, \omega_k \mid \boldsymbol{u}_j \,)}{\sum_{j=1}^{T} p(\, \omega_k \mid \boldsymbol{u}_j \,)} \tag{4}$$

After constructing updated error rate estimates for each verifier, the process can be repeated. In each iteration, the class-conditional decision probabilities in (4) are replaced by the estimates from the previous iteration and their complements. After a small number of iterations (5, in our evaluation), we can usually converge to an accurate pair of error rate estimates for each verifier.

In (4) we have assumed that each available decision sequence was generated by the same series of verifiers. This need not be the case. It is possible to extend this methodology to allow for decisions from different sequences of verifiers. The only requirement is that each $\boldsymbol{u}_j$ contains at least one decision from verifier i. Using this strategy, we can envision a process running on the access controller that collects all decision sequences presented over a specific interval and use them to estimate error probabilities for each verifier in the building. By repeating this procedure at scheduled intervals, it is possible to monitor the accuracy of all verifiers in the building and adjust the thresholds and fusion rule accordingly.

One of the difficulties with this procedure is determining the *a priori* class probabilities in (4). We expect that for most buildings, the number of authentic claimants will be much larger than the number of imposters; however, the exact probabilities are unknown. A small number of imposters also presents a potential difficulty in using estimated FAR as a proxy for the probability of false accept. If only a few imposters attempt to gain access, then even perfect error rate measurements would be poor indicators of false accept probabilities. For the stated purpose of configuration, error probabilities are our primary interest.

Our solution to this problem is to generate additional decision sequences using pseudo-imposters. The notion of pseudo-imposters is common in the speech community and compares each sample to more than one template [8, 9]. In addition to comparing each collected sample to the template of the claimed identity, the system compares the same sample to an additional template for a different identity. If the number of actual imposters is small relative to the number of authentic claimants, this results in a nearly an equal number of authentic and imposter decisions. We have found that this preprocessing step can improve estimation accuracy significantly. As we will show in Sect. 4, submitting a balanced set of decision sequences to the estimation algorithm leads to accurate error rate estimates for a wide range of verifier configurations.

4 Evaluation

In a sense, the estimation technique that we have presented relies on the combined decisions of an ensemble of verifiers to determine whether a local decision is in error. The technique will work well when the ensemble is accurate, with estimation accuracy degrading as the ensemble error rate increases. For this reason, the evaluation of our error rate estimation technique should be parameterized by the accuracy of the ensemble of verifiers. In the following, we present an empirical evaluation of estimation accuracy for a wide range of verifier ensembles and present the results with respect to ensemble error probabilities.

Our evaluation applies the estimation technique from Sect. 3 to a number of synthetic data sets and measures the resulting estimation error. To evaluate estimation error for a single ensemble of verifiers, we first assign a pair of underlying error probabilities to each verifier in the sequence. Then we generate T binary decisions for each verifier according to the chosen error probabilities. Counting the number of actual errors produces a pair of error rates for each verifier: the measured FAR and FRR. We can then use raw decision data to construct decision sequences and submit these sequences to our estimation algorithm. Estimation error is calculated as the difference between estimated error rates and measured error rates. Generating a new set of decision data and repeating this procedure for a single set of underlying error probabilities allows us to measure the statistics of the estimator.

Fig. 4 presents the average magnitude error in estimating verifier error rates for a wide range of ensembles. For brevity, we limit our presentation to error rate estimation for the third verifier in each sequence. Plots for other positions would

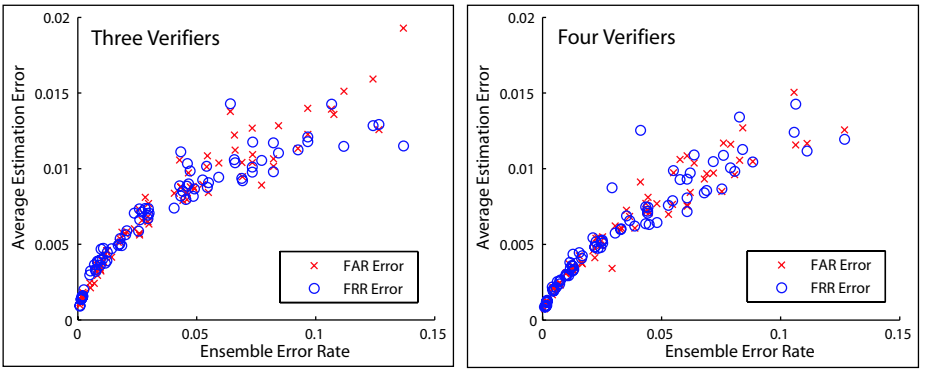

Fig. 2. Average magnitude of estimation error at the third verifier.

be nearly identical. Each point in the figure corresponds to a single set of error rates, and the error magnitude is averaged over 1000 data sets each containing $T = 1000$ decision sequences. The data contains an equal number of imposter and authentic decision sequences under the assumption that this balancing has been completed using pseudo-imposter decisions as described in Sect. 3. Ensemble error rates are predicted using underlying verifier error probabilities according to the minimum probability of error criterion.

For both the three and four verifier plots in Fig. 4, estimation errors less than 0.01 are possible for ensemble error rates less than 0.05. For ensemble error rates as high as 0.10, estimation error is still small. Although not presented in the figures, we have found that for ensemble error rates between 0.10 and 0.40 estimation error follows a linear trend with a moderate slope of 0.10. We expect that most reasonable ensembles will have combined error rates less than 0.15, but it is reassuring to see that even for large ensemble error rates the strategy maintains only linear growth in estimation error.

5 Conclusion

In the preceding, we have argued that representative training data is typically not available for evaluating the accuracy of a biometric verifier in a particular environment. This is particularly relevant in the context of physical access control, where verifiers may be deployed to many different environments. Typically accuracy information would be used to configure thresholds and decision fusion rules. In environments where a single configuration is insufficient, we believe that the selection of these configuration parameters should be completed automatically. To support this automatic configuration, we have presented a strategy for estimating the FAR and FRR of each verifier in an ensemble when class labels are unavailable.

These error rates can be directly applied to the selection of a fusion rule for combining decisions from multiple verifiers. We also expect that knowledge of these error rates can be used to search for appropriate thresholds for each local verifier. By comparing estimated error rates to specified requirements, the system can infer when thresholds require adjustment and the necessary direction of change. We do not provide further detail for adjusting thresholds in this manner, but we expect future work in this area to investigate this topic more thoroughly.

We expect that this work is applicable in a variety of applications beyond physical access control. An ability to estimate verifier error rates is an important tool for such automated configuration. Without knowledge of how well the system works under current configuration parameters, there is no way of knowing when parameters should be changed. Estimating verifier error rates provides feedback to remedy this situation.

Acknowledgments

This work was supported by the NIST Building and Fire Research Laboratory.

References

1. Beattie, M., Kumar, B.V.K., Lucey, S., Tonguz, O.: Combining verification decisions in a multi-vendor environment. In: Audio- and Video-based Biometric Person Authentication (AVBPA). (2005)
2. Kittler, J., Hater, M., Duin, R.: On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) **20** (1998) 226–239
3. Ross, A., Jain, A.: Information fusion in biometrics. Pattern Recognition Letters **24** (2003) 2115–2125
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society **39** (1977) 1–38
5. Phillips, P., Flynn, P.J., Scruggs, T., K.W. Bowyer, J.C., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2005)
6. Chen, L., Liao, H., Lin, J., Ko, M., Yu, G.: A new LDA-based face recognition system which can solve the small sample size problem. Pattern Recognition **33** (2000) 1713–1726
7. Belhumeur, P., Hespanha, J., Kriegman, D.: Eigenfaces vs. Fisherfaces : Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **19** (1997) 711–720
8. Mak, M., Zhang, W., He, M.: Determination of a priori decision threshold for phrase-prompted speaker verification. In: International Workshop on Multimedia Data Storage, Retrieval, Integration and Applications. (2000) 96–103
9. Pierot, J.B., Lindberg, J., Koowaaij, J., Hutter, H.P., Genoud, D., Blomberg, M., Bimbot, F.: A comparison of a priori threshold setting procedures for speaker verification in the cave project. In: International Conference on Acoustics, Speech, and Signal Processing (ICASSP). Volume 1. (1998) 125–128