Maximizing area under ROC curve for biometric scores fusion

Kar-Ann Toh*, Jaihie Kim, Sangyoun Lee

Biometrics Engineering Research Center, School of Electrical & Electronic Engineering, Yonsei University, 134 Shinchon-dong, Seodaemun-gu, Seoul 120-749, Republic of Korea

**Article Info**

**Abstract**

The receiver operating characteristics (ROC) curve has been extensively used for performance evaluation in multimodal biometrics fusion. However, the processes of fusion classifier design and the final ROC performance evaluation are usually conducted separately. This has been inevitable because the ROC, when taken from the error counting point of view, does not have a well-posed structure linking to the fusion classifier of interest. In this work, we propose to optimize the ROC performance directly according to the fusion classifier design. The area under the ROC curve (AUC) will be used as the optimization objective. When a fusion classifier has linear parameters, computation of the AUC provides a good representation of the ROC performance. Due to the piecewise cumulative structure of the AUC, a smooth approximate formulation is proposed. This enables a direct optimization of the AUC with respect to the classifier parameters. When a fusion classifier has linear parameters, computation of the solution to optimize a quadratic AUC approximation is surprisingly simple and yet effective. Our empirical experiments on biometrics fusion show strong evidences regarding the potential of the proposed method.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

Biometric is perhaps among the most promising security means to prevail conventional verification methods. This is because conventional token-based techniques such as keys, cards or passwords lie on the fact that they give authority to the tokens, but not to the users themselves. Anyone who has the keys, cards or passwords is given the authority to gain access to the token secured properties. However, there remain many problems to be resolved before biometric can gain possibly pervasive applications. For instance, due to inherent limitations as well as external sensing factors, no single biometric method can warrant a hundred percent authentication accuracy as well as universality of usage by itself. Noting that these problems may be alleviated by combining multiple biometric methods, the importance of multimodal biometrics thus need not be over emphasized.

Multiple modalities of biometrics can be combined either before matching or after matching. This combining process is also commonly called fusion. For fusion before matching, two levels namely, the sensor level and the feature level can be identified. For fusion after matching, three levels namely, the abstract level, the rank level and the match score level can be identified (see e.g. Ref. [1]). Existing fusion methods can generally be divided into two types: non-training-based methods and training-based methods. In non-training-based methods, it is often assumed that the outputs of individual biometrics are the probabilities that the input pattern belongs to a certain class (see e.g. Refs. [1,2]). The training-based methods do not require this assumption and they can operate directly on the match scores generated by biometric verification modules (see e.g. Refs. [3,4]). Our work here belongs to the training-based method working at match score level.

Since the outcome of biometric verification consists of only two labels, i.e. the query identity is recognized to be either a genuine-user or an imposter, the verification process can thus be treated as a two-category classification problem. This classification treatment holds well for multimodal biometrics fusion because similar decision labels are anticipated.

In the case of performance evaluation for multimodal biometric systems, the receiver operating characteristics (ROC) curve has been extensively adopted due to its good overview interpretation (see e.g. Refs. [5,6]). However, the processes of fusion design optimization and the final ROC performance evaluation are usually conducted separately. This has been inevitable because the ROC, when taken from the error counting point of view, does not have a well-posed structure linking to the fusion classifier of interest such that it can be directly computed.

In this work, we propose to optimize the ROC performance directly according to the fusion classifier design. The AUC (see e.g. Refs. [7-9]) will be used as the optimization objective since it provides a good overview representation of the ROC performance. Due to the piecewise cumulative structure of the AUC computation, a smooth approximate formulation is proposed. This enables a direct
optimization of the AUC with respect to the classifier parameters. When a fusion classifier has linear parameters, computation of the solution to optimize a quadratic AUC approximation can even be performed in closed-form. To our surprise, the empirical performance of this formulation shows promising potential. While noting that the proposed approach can be directly applied to other pattern analysis problems, main contributions of this paper are enumerated as follows: (i) proposal of an approximate AUC formulation as optimization objective for multimodal biometrics fusion, (ii) proposal of a deterministic fusion classifier solution for optimal AUC approximation, and (iii) provision of empirical evidences to support the proposal.

The paper is organized as follows: the next section provides some backgrounds regarding learning and prediction. Section 3 introduces the concept of AUC and existing formulations to compute this area. Section 4 focuses on maximizing the AUC. Essentially, a quadratic approximation is proposed to facilitate a convex loss functional. Based on a classifier which is linear in its parameters, the AUC is optimized directly. Some computational and application considerations are also discussed in the same section. Section 5 provides a brief description of biometrics used in experiments while Section 6 outlines the experimental setup. Section 7 presents empirical evidences for fusion of biometrics using both inhouse data and publicly available fusion data. Finally, some concluding remarks are provided in Section 8.

2. Background on learning and prediction

2.1. Linear parametric models

Linear parametric models have been widely used due to their tractability in optimization and related analysis. The embedding of nonlinearities such as kernels and other basis functions into linear regression models has even widened their scope of applications. The importance of linear parametric models that embed nonlinearities is thus obvious and we shall limit our scope to these linear models in this paper.

A good example of linear parametric model is the multivariate polynomial (MP) regression which has been shown to possess the capability of describing any arbitrary complex nonlinear input–output relationships attributed to the theoretical ground of Weierstrass’s approximation theory (see e.g. Ref. [10]). However, the number of independent adjustable parameters in MP would grow like \( \sum \binom{d}{l} \) for a \( r \)-order model with input dimension \( d \). This limitation has been recently addressed by Refs. [4,12] from reducing the number of polynomial expansion terms for classification applications. We shall adopt this reduced MP model in our experiments even though we know that the proposed formulation can easily be adapted for other types of linear models with embedded basis functions.

2.2. Learning and prediction

Consider a \( d \)-dimensional input \( \mathbf{x} \) and a \( r \)-th order polynomial operating on \( \mathbf{x} \) which gives rise to \( d \) polynomial expansion terms. A linear parametric model in this context can be written as

\[
g(\mathbf{x}, \mathbf{\alpha}) = \sum_{k=0}^{d-1} \alpha_k p_k(\mathbf{x}) = \mathbf{p}(\mathbf{x})^T \mathbf{\alpha},
\]

where \( p_k(\mathbf{x}) \) corresponds to the \( k \)th polynomial expansion term of the row vector \( \mathbf{p}(\mathbf{x}) = [p_0(\mathbf{x}), p_1(\mathbf{x}), \ldots, p_{d-1}(\mathbf{x})]^T \) and \( \mathbf{\alpha} = [\alpha_0, \alpha_1, \ldots, \alpha_{d-1}]^T \) is the column parameter vector.

When each input \( \mathbf{x} \in \mathbb{R}^d \) has a known target label \( y \in \mathbb{R} \), and given \( m \) number of such example learning data pairs \( (\mathbf{x}_i, y_i), i = 1, 2, \ldots, m \), the learning problem can be supervised. In biometric verification problems, these target labels \( (y_i, i = 1, 2, \ldots, m) \) are known as genuine-user and imposter. For multimodal biometrics fusion, the input vectors \( \mathbf{x}_i, i = 1, 2, \ldots, m \) come from the scores of \( l \) number of biometric modalities to be fused.

Learning of the target labels \( \{y_i\} \) as a system of \( m \) non-linear functions \( \mathbf{y} = \{y_i\} \) can be accomplished by minimizing a least squares error (LSE) criterion. To stabilize the solution for estimation, a weight decay regularization can be performed [12]. The criterion function to be minimized is thus

\[
J = \frac{1}{2} \sum_{i=1}^{m} [y_i - p(x_i)]^2 + \frac{b}{2} ||x||^2 = \frac{1}{2} ||y - \mathbf{P} \mathbf{x}||^2 + \frac{b}{2} ||x||^2.
\]

where \( b \) controls the weighting of regularization and \( \mathbf{P} \) packs the expanded training samples in matrix form:

\[
\mathbf{P} = \begin{bmatrix}
    p(x_1) \\
p(x_2) \\
\vdots \\
p(x_m)
\end{bmatrix}.
\]

The estimated training output is given by \( \hat{\mathbf{y}} = \mathbf{P} \mathbf{z} \) where the solution for \( \mathbf{z} \) which minimizes \( J \) is

\[
\text{LSE} : \mathbf{z} = (\mathbf{P}^T \mathbf{P} + bI)^{-1} \mathbf{P}^T \mathbf{y},
\]

with \( b \) being chosen to be a small value for stability and not introducing much bias [12]. \( \mathbf{I} \) is an identity matrix with similar dimension to \( \mathbf{P}^T \mathbf{P} \). When \( \mathbf{P} \) adopted the reduced polynomial model from [12], the LSE solution in Eq. (4) is also termed RM solution.

For unseen test data \( \mathbf{x}_t \), another polynomial matrix \( \mathbf{P} \) can be generated using \( \mathbf{p}(\mathbf{x}) \). Prediction of the class label \( \mathbf{y}_t \) can then be performed using the above learned \( \mathbf{z} \), i.e. \( \mathbf{y}_t = \mathbf{P} \mathbf{z} \). With these backgrounds in place, we are ready to discuss the ROC and related issues in the sequel.

3. Area under ROC curve (AUC)

Consider the basic case of binary classification. Suppose we have \( m \) learning examples \( \{\mathbf{x}_i\}_{i=1}^{m} \in \mathbb{R}^d \) (a \( d \)-dimensional feature vector) and their corresponding class labels \( y_i \in \{0,1\} \), let \( g : \mathbb{R}^d \mapsto \mathbb{R} \) be the hypothesis function (biometric fusion classifier) mapping these pattern features onto a scalar measure for decision inference. Suppose the classifier \( g(\mathbf{x}) \) produces a continuous output, then the output must be thresholded in order to label each example as positive class (genuine-user) or negative class (imposter). Given a decision threshold \( \tau \), the class label associated to a new example \( \mathbf{x}_n \) can be written as

\[
\text{cls}(g(\mathbf{x}_n)) = \begin{cases} 1 & \text{genuine-user} \quad \text{if } g(\mathbf{x}_n) \geq \tau, \\ 0 & \text{imposter} \quad \text{if } g(\mathbf{x}_n) < \tau. \end{cases}
\]

For each operational setting of \( \tau \), a true positive rate and a false positive rate can be obtained. By varying the decision threshold \( \tau \) over a range, say from 0 to 1 in a normalized case, the ROC curve is obtained by plotting the true positive rates over the false positive rates. The more bowed the curve is towards the upper left corner, the better the classifier’s ability to discriminate between the two pattern classes (see e.g. Refs. [7–9,13]). Hence, the AUC provides a measure related to classifier discrimination performance.

In continuous case, the AUC would need computation of an integral. However, for discrete case, a cumulative count can be adopted for AUC computation. Consider a random observation pair, one from each pattern class, the AUC can be expressed as the probability that the score of the positive class observation is greater than that of the negative class observation [14,15]. In other words, the AUC is the fraction of positive–negative pairs that are ranked correctly. Denote the variables \( (\mathbf{x}, m) \) that correspond to positive and negative examples by respective superscripts + and −, it is thus not difficult to see
that the AUC (see e.g. Refs. [15–19]) for the given \( m \) training examples can be expressed as

\[
A_c(X^+,X^-) = \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} g(X^+_i) g(X^-_j),
\]

where the term \( g(X^+_i) g(X^-_j) \) corresponds to a ‘1’ whenever the elements \( g(X^+_i) > g(X^-_j) \) (i.e., \( i = 1, 2, \ldots, m^+ \), \( j = 1, 2, \ldots, m^- \), and ‘0’ otherwise. Here we note that (6) is also known as Wilcoxon–Mann–Whitney statistic.

Define a unit step function \( u(\xi) \equiv 1_{\xi<0} \) with \( \xi_{ij} = g(X^+_i) - g(X^-_j) \), (6) can be re-written as

\[
A_c(X^+,X^-) = \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} u(\xi_{ij}).
\]

(7)

Suppose the classifier \( g \) consists of some \( d \) number of adjustable parameters \( x = \{x_0,x_1,\ldots,x_{d-1}\}^T \) operating on the feature vector \( X \) (scores from \( l \) number of single biometrics), e.g. \( g(x,X) \) in Eq. (1), then the goal to optimize the fusion classifier’s discriminative performance can be treated as to maximize the AUC:

\[
\text{AUC} : \arg \max x \text{ } 1_{\frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} u(\xi_{ij})}.
\]

(8)

where \( \xi_{ij} = g(x,X^+_i) - g(x,X^-_j) \), \( i=1,2,\ldots,m^+ \), \( j=1,2,\ldots,m^- \).

To solve the above maximization problem, it is usually re-stated in its dual form, i.e. to minimize the area above ROC curve (AAC):

\[
\text{AAC} : \arg \min x \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} u(\xi_{ij}),
\]

(9)

where \( \xi_{ij} = g(x,X^+_i) - g(x,X^-_j) \), \( j=1,2,\ldots,m^- \), \( i=1,2,\ldots,m^+ \).

Here we note that when a classifier achieves perfect accuracy, the AUC is equal to 1 (AAC = 0 since AAC = 1 − AUC). For random classifier, the associated AUC is 0.5 (AAC=0.5). With this formulation in place, we are ready to minimize the AAC directly according to the classifier parameters in the sequel.

4. Minimizing AAC

To solve the problem given by Eq. (9), an approximation to the non-differentiable step function \( u \) is often adopted. A natural choice to approximate the above step function is the sigmoid function [15,19,20] (see Fig. 1) where the minimization problem becomes

\[
\arg \min x \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} \sigma(\xi_{ij}),
\]

(10)

where

\[
\sigma(\xi_{ij}) = \frac{1}{1 + e^{-\gamma\xi_{ij}}}, \; \gamma > 0.
\]

(11)

There are two problems associated with this approximation. The first problem is that the formulation is nonlinear with respect to the learning parameters. Although an iterative search can be employed for local solutions, different initializations may end up with different local solutions, hence incurring laborious trial and error efforts to select an appropriate setting. The second problem is that the objective function could be ill-conditioned due to the much local plateaus resulting from summing the flat regions of the sigmoid. A lot of search effort may be spent upon making little progress at locally flat regions.

4.1. A deterministic solution

Existing statistical modelling and kernel approximation methods [21,22] require either assumption on data distribution or determination of additional smoothing function. Also, different from the nonlinear or heuristic search [13,15,19,23], quadratic programming [17,18,24,25], and boosting [26] based methods, we seek in this work a possible deterministic training-based solution without the need of prior knowledge on data distribution. This is achieved via matching the link–loss functional pair from the training formulation to facilitate a convex solution [27,28]. When a linear link function is adopted, a quadratic loss functional would match this linear link, and this leads to desired convexity for the link–loss pair. For instance, the polynomial link terms (denoted by a row vector \( p(x_X) \)) would give rise to a linear link function \( g(x,X) = p(x_X) \cdot x \) where \( x \) is the column parameter vector. However, some considerations regarding an appropriate approximation to the step function \( u(\cdot) \) would be necessary.

When we have all inputs normalized within [0,1], the step functional can be approximated by centering a quadratic functional at the origin. To cater for inputs which get beyond this range, an offset \( \eta \) can be introduced such that only the right arm of the quadratic functional is activated for the approximation (see Figs. 1, 2 and Remark 1). With this idea in mind, the following quadratic AAC approximation is proposed [20]:

\[
\text{AAC}_Q : \arg \min x \left\{ \frac{b}{2} \|x\|_2^2 + \frac{1}{2m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} \phi(\xi_{ij}) \right\},
\]

(12)

where

\[
\phi(\xi_{ij}) = (\xi_{ij} + \eta)^2 = |g(x,X^+_i) - g(x,X^-_j)| + \eta^2 = |p(x,X^+_i) - p(x,X^-_j)| \cdot x + \eta^2.
\]

(13)

for \( j=1,2,\ldots,m^- \), \( i=1,2,\ldots,m^+ \). Here we note that a weight decay has been included in the above optimization objective. This is to provide stabilization in case of matrix inversion.

The optimality condition for \( \text{AAC}_Q \) requires that

\[
\frac{\partial A(x,X^+,X^-)}{\partial x} = 0,
\]

(14)

which implies that

\[
bz + \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} (p(x,X^+_i) - p(x,X^-_j))^T [(p(x,X^+_i) - p(x,X^-_j)]x + \eta = 0
\]

\[
\Rightarrow \left[ b + \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} (p(x,X^+_i) - p(x,X^-_j))^T (p(X^+_i) - p(X^-_j)) \right] x 
\]

\[
+ \frac{\eta}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} (p(x,X^+_i) - p(x,X^-_j))^T = 0.
\]

(15)

Abbreviating the row polynomial vectors by \( \boldsymbol{p}_i = p(x,X^+_i) \in \mathbb{R}^d \) and \( \boldsymbol{p}_j = p(x,X^-_j) \in \mathbb{R}^d \), the solution for \( x \) which minimizes \( \text{AAC}_Q \) can be written as

\[
\begin{align*}
x & = \left[ b + \frac{1}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} (\boldsymbol{p}_i - \boldsymbol{p}_j)^T (\boldsymbol{p}_i - \boldsymbol{p}_j) \right]^{-1} \\
& \times \left[ -\frac{\eta}{m^+m^-} \sum_{i=1}^{m^+} \sum_{j=1}^{m^-} (\boldsymbol{p}_i - \boldsymbol{p}_j)^T \right].
\end{align*}
\]

(16)

where \( I \) is an identity matrix of \( d \times d \) size.
Remark 1. It is noted here that there have been two levels of approximations adopted, namely level-1: polynomial approximation for a nonlinear fusion classifier \( g(x, X) = P X a \), and level-2: quadratic approximation for the unit step function \( u(z_{ji}) \) where \( z_{ji} = g(x, X_j^+) - g(x, X_i^+) \). At level-1, the polynomial \( p(x) \) can be replaced by a row vector of fixed sigmoids \( r(x) \) or other activation functions such that \( g(x, X) = r(x) a \) [29–31]. However, this sigmoidal approximation does not apply to level-2 for \( u(z_{ji}) \) since it will result in a nonlinear formulation with respect to the optimizing parameter \( x \).

Also, notice that the learning solution in Eq. (16) appears to have similar structure to Eq. (4) but with a normalized double summation of differences between the regressor vectors \( P_i \) and \( P_j \) corresponding to each class label. This is differentiated from the LSE in Eq. (4) which lumped the two class specific regressor vectors into a single matrix \( P \).

4.2. An illustrative example

Similar to the LSE or RM method in Eq. (4), the solution \( x \) in Eq. (16) for minimal AAC classification decision can be obtained in closed-form which is also least squares optimal but in the AAC sense. This solution can be used to compute the output for unseen test data \( X_t \) using the polynomial matrix \( P_t \) as in the LSE case, i.e. \( \hat{y}_t = P_t x \). The
closed-form and deterministic solution is differentiated from those nonlinear or heuristic search, quadratic programming and boosting-based algorithms mentioned.

Fig. 2(a) shows a 1D learning example using Eq. (16) and a third order polynomial model at $b = 0.001$. This learning example is considered fundamental and yet non-trivial because it has overlapping class distributions. In this plot, the trained model outputs are shown for different offset $\eta$ settings. Here, it is seen that a high offset value results in a high output range, and this phenomenon applies to negative offset values but in a reverse sense (similar trends can be observed for $10^{-100} \leq b \leq 1$ from current numerical implementation). The distributions of these outputs show relevant mapping of the training data which distinguishes between the positive and negative classes. Fig. 2(b) shows the range of $\phi$ values from each of those $m \times m^{-1}$ summations in Eq. (12). This plot empirically verifies the validity of activating a single arm from quadratic approximation.

4.3. Biometric scores fusion

In biometric authentication, the user presents his/her biometric trait (e.g., face, fingerprint) to a camera or sensor, the system then extracts and generates a query template to compare with those enrolled templates in database. The score generated from a comparison between two biometric templates from the same identity is called the genuine-user score. The score generated from a comparison between two biometric templates from different identities is called the imposter score.

To train a multi-biometric system, a gallery of biometric samples (typically 1–10 samples) from each modality of each user needs to be collected. The training set is next obtained from intra-identity (typically 1–10 samples) from each modality of each user. Training scores, a polynomial matrix $P$ can be formed as in Eq. (3). This polynomial $P$ is then used together with the corresponding target vector $y = [y_1, y_2, ..., y_m]^T$ for LSE/RM parameter estimation using Eq. (4). For unseen test data, another polynomial $P_t$ will be generated using only the test data. The estimated parameter $\mathbf{x}$ from Eq. (4) can then be applied to generate the test output $\hat{y}_t = P_t \mathbf{x}$ for authentication decision, i.e. by applying $\text{cls}(\hat{y}_t)$ from Eq. (5) to make a final decision regarding whether the fused output is considered as a genuine-user or an imposter.

The process of parameter estimation in AAC$_Q$ training is similar to above except that the genuine-user and imposter scores are separated as $\mathbf{x}^+$ and $\mathbf{x}^-$ (notice that $\mathbf{x} = [\mathbf{x}^+, \mathbf{x}^-]$). These scores are then used to generate the polynomials $P(\mathbf{x}^+)$ and $P(\mathbf{x}^-)$ corresponding to the $i$th sample from $\mathbf{x}^+$ ($i = 1, ..., m^+)$ and $i$th sample from $\mathbf{x}^-$ ($j = 1, ..., m^-)$, respectively. Eq. (16) is then used for parameter estimation. The process test is similar to above description for LSE/RM application after having the polynomial $P_t$ generated from test data. We shall focus on verification accuracy in this work.

4.4. Some computational and application considerations

(i) Computational complexity: Consider a $d$-dimensional polynomial expansion. Then, the inverse term of Eq. (16) requires $m^+ \times m^- \times d^2$ operations for the covariance calculation followed by an $O(d^3)$ operation of the inverse assuming a Gaussian elimination process, giving a dominant computational complexity of $O(d^3 m^+ m^-)$. Assuming a non-empty set for each pattern category, then $m^+ \times m^-$ is bounded by $1 \times (m - 1) = (m - 1)$ and $m^+ \times m^- = m^2/4$, leading the computational complexity bounded between $O(d^3(m - 1))$ and $O(d^3 m^2/4)$. This shows an almost always higher computational complexity for AAC$_Q$ as compared to LSE estimation which is dominated by $O(d^3 m)$ for $m \geq d$. We shall show empirically that, for this price of a higher computational complexity, a stable estimation can be obtained from the larger number of summations in covariance computation for AAC$_Q$ as compared to that using LSE estimation.

(ii) Polynomials: As mentioned, since the number of terms of the full MPs grows exponentially with respect to feature dimension as well as model order, we adopt here the reduced multivariate (RM) polynomial proposed by Toh [4,12] where the number of parameters increases linearly with respect to feature dimension and model order:

$$g_{RM}(x) = a_0 + \sum_{i=1}^{d} \sum_{j=1}^{l} x_{i,j}P_{ij} + \sum_{i=1}^{d} \sum_{j=1}^{l} x_{i,j} \left( \sum_{k=1}^{l} x_{i,j} \right)^k + \sum_{i=1}^{d} \left( \sum_{j=1}^{l} x_{i,j} \right) ,$$

with $x_{ij}, i = 1,...,l$ being the polynomial inputs, $a_0, a_{ij}, a_{ijk},...$ being the weighting coefficients to be estimated, and $i, j, \sum_{i=1}^{d} \sum_{j=1}^{l} x_{i,j}$ correspond to input-dimension, order of system, respectively. Here we highlight the linear parametric formulation of the model by rewriting $g_{RM}(x) = \mathbf{p}_{RM} \mathbf{x}$ where $\mathbf{p}_{RM}$ denotes a row vector containing those polynomial expansion terms of the inputs and $\mathbf{x}$ denotes the column parameter vector to be estimated. As mentioned, this $\mathbf{p}_{RM}$ will be used in Eqs. (3) and (4) for LSE training, and in Eq. (16) for AAC$_Q$ training.

(iii) Performance measures: LAUC and conventional error rates. A frequently used evaluation technique for ROC performance is by graphical means. The higher the position of the ROC curve for a particular biometric with respect to other curves, the better the performance of this biometric. This technique is particularly useful for an overview of performance over a certain operating region.

However, in practical situations, the ROC curves may cross each other and difficulty may arise when comparing the performances of two biometrics. Under such a situation, a single numerical value of the AUC or the equal error rate (EER) would be helpful to quantify the performances of the two biometrics in terms of an overall or a particular performance. From the system tuning point of view, the use of AUC can be advantageous in the sense that there is no need to locate an optimal operating point and this redresses much the dilemma between over- and under-setting of the operating threshold. The numerical area value under or above ROC is thus useful in these aspects.

As the absolute AUC (or AAC) value shows little difference between two biometrics, particularly when both biometrics have high performances where the two results differ in the vicinity of small decimal places due to normalization according to the huge number of match scores used in generating the ROC, we shall use the negative base 10 logarithm of AUC values $[\log_{10}(AUC)]$ as overall performance measure (alternatively $[\log_{10}(AAC)]$ can also be used, but attention needs to be paid for those class separate cases with AAC = 0). The lower the value of $[\log_{10}(AUC)]$, the better the performance of a biometric is relative to others. Also, this numerical value shall roughly indicate the decimal order of accuracy, i.e. the number of decimal places for the square measure (area) of error rates (FAR × FRR). For instance, classifiers with AUC = 0.9, 0.99, 0.999, 0.9999, respectively, have values of $[\log_{10}(AUC)] = 4.58 \times 10^{-2}, 4.36 \times 10^{-4}, 4.35 \times 10^{-4}, 4.34 \times 10^{-5}$. Here we note that with particular $[\log_{10}(AAC)] = 4.34 \times 10^{-5}$ from $\text{FRR}_{\text{zeroFRR}}$ at $10^{-4}$ and $\text{FRR}_{\text{zeroFRR}}$ at 0.2, a classifier can either falsely accept approximately
Top row: some representative visual face samples for an identity from the BERc database which includes different lighting (left, front, and right illuminated), expression (smiling, normal, surprised) and pose (left 15°, front, right 15°) conditions; bottom row: infra-red face samples for the same identity taken under similar lighting, expression and pose conditions.

Fingerprint image samples: wet fingers in (a), (b) and normal fingers in (c), (d).

(a) A hand image sample, (b) extracted hand-geometry.

EER is also considered in our performance evaluation. The EER has its reasons for being widely used: (i) it is a single index measure and thus simple and direct in terms of interpretation as compared to the ROC, (ii) it is a compact term indicating both the false acceptance rate (FAR) and the false rejection rate (FRR) at the same time, and more importantly (iii) it is based on a projected optimal operating point (of total error rate) where the FAR curve meets the FRR curve.

5. Face and hand biometrics

To provide benchmarks over commonly adopted performance measure from the biometric literature (see e.g. Refs [32,33]), the

1 These figures are obtained by assuming a linear ROC line joining FAR\textsubscript{zeroFRR} and FRR\textsubscript{zeroFAR} giving rise to a triangular area of AAC = FAR\textsubscript{zeroFRR} × FRR\textsubscript{zeroFAR}/2, and based on the relationship AUC = 1 - AAC.

114 imposters (from the experimented 114 000 population) or falsely reject approximately 192 genuine-users (from the experimented 960 population). This shows the significance of decimal order particularly for high security applications. For convenience, we shall call this \(-\log_{10}(AUC)\) value as LAUC in brief.
reasons. Under a continuing effort to fuse several modalities in a natural manner in terms of usage, we consider the following face and hand biometrics for scores fusion.

5.1. Face-based biometrics

(i) Visual face verification: Face is the most common biometric used by humans. We inherently use this biometric to recognize people in our daily interactions. Face recognition is thus an important area in biometrics for it can also be covert as well as non-intrusive.

The main approaches to viewer centered 2D face recognition includes holistic, analytic and hybrid methods [33]. The holistic approach uses subspace techniques to reduce the image dimension and then compare image similarity using this subspace. A very widely used technique for this subspace reduction is the principal component analysis (PCA). The analytic approach uses geometrical features such as distances between face objects like eyes, nose, and mouth for similarity measure. The hybrid approach combines various means including the holistic and analytic approaches.

The visual face images used in this study were captured under various illumination and pose conditions using a Bumblebee CCD camera produced by Point Grey Research Inc. (see Ref. [34] for details). The resolution of the image used was 320 × 240 pixels. The top row of Fig. 3 shows some visual image samples for an identity under various illumination and pose conditions. In this work, we adopted the holistic approach using PCA. To compare similarity between two face images, the Euclidean distance was used for the first 100 eigenvalues.

(ii) Infra-red face verification: Due to the relatively high instrumental cost, the infra-red face is less studied as compared to the visual face. The infra-red face images used in this study were captured using a ThermoVision S65 produced by FLIR Systems Inc. As in the visual face case, the images were captured under varying illumination and pose conditions with the resolution of the image being fixed at 320 × 240 pixels. The bottom row of Fig. 3 shows some infra-red image samples for the same identity under various conditions. Similar to the visual face, we adopted the holistic approach using PCA for the infra-red face. To compare similarity between two face images, the Euclidean distance was used for the first 100 eigenvalues [34].

5.2. Hand-based biometrics

(i) Fingerprint verification: Beside the face biometric, fingerprint is perhaps the next most widely used biometric attributed to its identity and forensic applications. In general, an automatic
Fingerprint identification or verification (see e.g. Refs. [35,36]) system consists of three main processing stages namely, image acquisition, feature extraction and matching. In image acquisition, query and template database images are acquired through various input devices. Development over the years has seen through use of devices that mechanically scan the ink-based fingerprints into the computer system, to invention of devices which directly capture the fingerprints using sophisticated solid state sensors. With fingerprint images which could be distorted or contaminated by noise, the automated system seeks to extract characteristic features which are discriminating for different fingers and yet invariant with respect to image orientation for same fingers. The final stage of fingerprint identification is to search and verify matching image pairs.

The fingerprint images were collected using Veridicom’s iTouch sensor with a resolution of $300 \times 300$ pixels. Our representation for the fingerprint template consists of a global structure and a local structure. The global structure consists of positional and directional information of ridge endings and ridge bifurcations. The local structure consists of relative information of each detected minutia with other neighboring minutiae. Fingerprint verification is then performed by comparing the minutia information between two templates [37]. Fig. 4 shows four samples of fingerprint images with detected minutiae and area of interest segmentation. The interested readers are referred to Refs. [38,37] for details of minutia detection and matching.

(ii) Hand-geometry verification: The hand-geometry is considered to achieve medium security as compared to fingerprint technology. However, it has several advantages over the use of fingerprints namely, low computational cost, lack of relation to criminal records, and does not have imaging problems due to hand’s wetness. These features can be exploited to complement the high accuracy feature of fingerprint systems in this fusion development for a robust and yet highly secure system. The problem of having a finger being too wet or too dry for fingerprint image capture can thus be resolved without compromising the high security requirement.

A light box was used to flush the background of the captured images such that a sharp edge could be obtained for the hand-geometry. Segmentation on this back lit image became simple as the contrast was high. A pair of hand shape contours can be compared using the contour string-match method [39] or based on the so-called ‘handcrafted features’ [40]. Typical handcrafted features include the length and the width of each finger, aspect ratio of the palm or fingers, thickness of the hand, finger perimeter and areas, and so on [40].

The images for hand-geometry were collected using the light box set-up in VGA resolution [46]. In our current application, we use the width and length information. First, the hand contour is analyzed and dominant points are located [41]. These points are further identified as finger tips and valleys based on the convex or concave curvature of the contour. The principal axis of each finger is then found using a set of equal separated grid points starting from the respective finger tips. The widths are measured perpendicular to the axes at the grid points. The features used were similar to those in Ref. [40] except that a fixed interval was used for the width measurements with a total of 15-30 width features being collected for each hand image depending on the finger length. The length is found using the finger tip and its neighboring valleys information. These features of each finger from both the query image and the template image are compared separately. Their absolute matching differences are summed up and normalized as the matching score. Fig. 5 shows a sample captured hand image and its extracted hand-geometry including width/length features in our application.
6. Experimental setup

6.1. Data sets

In the following experiments, each data set corresponding to visual face verification, infra-red face verification, fingerprint verification, and hand-geometry verification consists of 96 identities, wherein each identity contains 10 samples. For training and test purposes, each of these biometric data sets are partitioned into two equal sets consisting of $\mathcal{S}_{\text{train}}$ and $\mathcal{S}_{\text{test}}$, each with $96 \times 5$ samples. The genuine-user and the imposter match scores are generated from these two sets by intra-identity and inter-identity matching among the image samples for each biometric. A total of $960$ ($96 \times 5 \times 5$) genuine-users and $114,000$ ($96 \times 95 \times 5 \times 5$) imposters (match scores) are generated for each training set and test set for each biometric. Since all four biometrics have the same number of

Fig. 8. Fusion(VS + IR): (a) average LAUC values versus model order $r$ and RBF width $\gamma$. A low LAUC value indicates good performance, (b) a zoom-in view.
genuine-user and imposter samples, an arbitrary one-to-one identity correspondence was assumed among the four biometric data sets. This is a reasonable assumption since our focus here is output scores fusion and not on correlation among different modalities for each identity.

The match scores for all biometrics are normalized to within the interval [0, 1], all having a higher match score for a genuine-user than that of an imposter, before performing data fusion. Fig. 6(a)–(h) show the normalized matching performances for the training and test sets, respectively, for individual fingerprint, hand-geometry, visual face and infra-red face verifications before multimodal fusion.

In order to have an overview of the relative performances of these individual biometric implementations, we plot all the ROC curves in the same plot in Fig. 7. Here, we see that the visual face system has the lowest ROC performance due to large variation of illumination and pose conditions. The infra-red face system appears to have good tolerance to these environmental factors relative to the visual face system. As for the fingerprint and hand-geometry systems, each of them shows its superiority in performance at different operating regions. We shall observe the impacts of fusing the weakest visual face system and the higher performed infra-red face system, and of fusing two comparable biometrics in the forthcoming experiments.
6.2. Fusion experiments setup

Using the data from the above biometrics and 33 publicly available fusion data sets from Refs. [42,43], we perform the following four sets of experiments:

(i) Fusion of visual and infra-red face verification scores: Here, we combine the weakest visual face biometric with a relatively stronger one, the infra-red face. This is a practical consideration because these face biometrics can be captured simultaneously. Our purpose here is to observe whether the stronger biometric would be affected by the much weaker one (which is quite close to random decision as shown in Fig. 7). We shall call this experiment as Fusion(FP + IR) in brief.

(ii) Fusion of fingerprint and hand-geometry verification scores: Here, we consider to combine two strong biometrics. The fingerprint and the hand-geometry are selected for fusion since their images can be captured simultaneously. We shall observe whether significant verification performance enhancement can be achieved from this kind of fusion. We call this experiment as Fusion(FP + HD) in brief.

(iii) To benchmark the fusion performance as well as to observe possible impacts on fusion due to use of distinct sets of identities and data acquisition times in training and testing, an evaluation is performed on a public domain database, NIST-BSSR1 [42], using the lowest performed face and fingerprint data sets: face-G and fing-li-V. This data set contains 517 identities, with 517 genuine match scores and 517 x 516 imposter match scores which are acquired over a period of nearly four years. We call this experiment NIST-BSSR1 in brief.

(iv) To further benchmark the performance, an additional set of experiments using 32 fusion data sets from publicly available database [43] is performed. Within these data sets, both inter-modal (data sets {1–15, 24–29}) and intra-modal (data sets {16–23, 30–32}) fusions have been considered. The experimental protocol has been adopted accordingly to that in Ref. [43] for the proposed AACQ such that the results can be directly compared in future.

6.3. Performance evaluation

Using those training and test data sets from the face and hand biometrics together with their randomly permuted versions, we perform 10 runs of twofold cross-validation experiments for statistical evidence. The average test results from the twofolds which are accumulated over 10 runs will be reported for the first three experiments. As for the last set of experiments using 32 fusion data sets, the experimental protocol of hold-out test from Ref. [43] will be followed such that the results can be easily compared in future.

To gain some ideas regarding the empirical performance, we shall compare the proposed learning method AACQ (16) with the original LSE learning (4), both using the RM model (17). For both learning methods (AACQ and LSE), the RM polynomial orders \( r \in \{1,2,\ldots,6\} \) will be experimented, with regularization parameter being empirically fixed at \( b = 10^{-3} \) based on twofold cross-validation using only the training set. Hence we have a standardized setup to observe performance changes attributed to variation of \( r \) only. The quadratic approximation offset \( \eta \) is fixed at 1, both for the standardized experimentation reason as well as for its minor impact on estimation for positive values from our empirical observation (see Fig. 2).

Apart from adoption of the RM model, we extend our investigation into a specific sigmoidal model from Ref. [30] which is obtained by a replacement of the RM polynomial expansion terms, \( p_k(x) \) of Eq. (1), by randomly placed and fixed sigmoids. In other words, we replace the RM polynomial expansion terms by a fixed number of sigmoids (this number is denoted as \( N_{SIG} \)) given by \( \sigma(\gamma^T \mathbf{x} + \gamma_0) \) (11) with the input weight values \( \{\gamma, \gamma_0\} \) being randomly chosen within \([-1, 1]\) for each sigmoid. Both LSE-based and AACQ-based learning will be studied for such a sigmoidal model (herein abbreviated as LSE-SIG and AACQ-SIG, respectively). The number of sigmoids \( N_{SIG} \in \{1,2,5,10,50,100\} \) will be experimented without regularization following the original setting of Ref. [30]. In view of the random settings of the input weight values, we perform 10 trials (using different random seeds for the random weight values) of the above-mentioned 10 runs of twofold cross-validation which used different partitioning of the training and test sets. In summary, we performed 100 runs of twofold tests for LSE-SIG and AACQ-SIG considering both data partitioning and weight randomness at six different \( N_{SIG} \) settings.
The simple SUM rule was reported to be among the best fusion methods for multi-biometrics when the inputs are normalized appropriately (see e.g. Ref. [44]). For benchmarking purpose, we include in our experiments two important normalization methods namely the Min-Max and the robust hyperbolic tangent (tanh) method (see e.g. Ref. [44]) to precede the simple SUM rule. We will call these methods as SUM(Min-Max) and SUM(Tanh) in brief.

In addition to SUM rule, the support vector machines (SVM) adopting different kernels are also included. The SVM adopting a polynomial kernel will be abbreviated as SVMPoly and the SVM adopting a radial basis function kernel will be abbreviated as SVM-Rbf. This is to cater for nonlinear decision hyper-surfaces as well as to benchmark with the well-acclaimed SVM in classification and biometrics fusion (see e.g. Ref. [6]). The SVMs from Ref. [45] are experimented with different model settings (polynomial order and Gaussian width $\gamma$) to observe the operational performance.

Although the unimodal biometric will be used as performance baseline, our focus here will be on comparison among different fusion methods rather than showing multimodal versus unimodal performances.

7. Results and discussion

7.1. Results: Fusion(VS + IR), Fusion(FP + HD) and NIST-BSSR1

LAUC evaluation: Figs. 8, 9, and 10 show, respectively, for Fusion(VS + IR), Fusion(FP + HD) and NIST-BSSR1, the average (over 10 runs of twofold tests) LAUC values resulted from the use of LSE.
Fig. 12. Fusion(FP+HD): (a) average EER values plotted over different model settings, (b) a zoom-in view.

These results show consistently lower LAUC values of AAC_Q than that of any single biometric over all model order settings in all three fusion experiments. The AAC_Q-SIG shows improved performances over LSE-SIG for all NSIG settings with NSIG ∈ {2, 5, 10, 50, 100} showing a comparable performance with AAC_Q in several instances. The SVMPoly shows a rather good LAUC fusion performance comparing to single biometric except for the case of Fusion(VS + IR) in some order settings. Both LSE and SVMRbf show large LAUC fluctuations between those LAUC values of the single biometric and that of the best fusion result. The SUM methods (Min–Max and tanh) show deterioration in performance comparing to single biometric in the Fusion(VS + IR) case, and with improved performance in Fusion(FP+HD) and NIST-BSSR1 cases. These results show stability of AAC_Q over LSE and other fusion methods, particularly for fusion of a weak biometric with a relatively strong biometric in the Fusion(VS + IR) case where a deterioration in fusion performance can occur.
whereas AAC obtained the best training and test EERs in all case studies. Since the above trials (10 runs of twofold experiments on 6 or 12 model settings) for each method and each case study (Fusion(VS + IR), Fusion(FP + HD) and NIST-BSSR1), the average (over 10 runs of twofold tests) EER values for all compared methods. Except for the fluctuation trend of LSE, most methods show similar EER trends with those in LAUC plots.

**Average error rates:** Table 1 summarizes the average EER from the above trials (10 runs of twofold experiments on 6 or 12 model settings) for each method and each case study (Fusion(VS + IR), Fusion(FP + HD) and NIST-BSSR1). A t-test for statistical significance of difference between the mean values are also included for each compared method with respect to AACQ. Here it is seen that AACQ obtained the best training and test EERs in all case studies. Since some of these results are not of statistical significance, the AACQ can only be said to have either better than or comparable EER performance with respect to those compared methods.

Those bracketed numerics at the rear of test results in Table 1 indicate the percentage change of test performance from training performance based on a paired \( t \)-test comparing with AACQ. Here it is seen that AACQ obtained the best training and test EERs in all case studies. Since some of these results are not of statistical significance, the AACQ can only be said to have either better than or comparable EER performance with respect to those compared methods.

Those bracketed numerics at the rear of test results in Table 1 indicate the percentage change of test performance from training performance based on a paired \( t \)-test comparing with AACQ. Here it is seen that AACQ obtained the best training and test EERs in all case studies. Since some of these results are not of statistical significance, the AACQ can only be said to have either better than or comparable EER performance with respect to those compared methods.

**Computational training time:** As the CPU results show quite similar trends in the above three cases, the average training CPU times for the training-based fusion methods are shown in Fig. 14 for NIST-BSSR1 as a representative. All experiments were ran under the PC Windows Matlab platform using a Pentium-M-1.1 GHz computer. The results show a much higher computational cost of AACQ and AACQ-SIG as compared to other methods. Moreover, the increment of computational cost of AACQ and AACQ-SIG is seen to be nearly proportional to the model order \( r \) and number of sigmoids \( N_{SIG} \), respectively. The main reasons for the high computational cost include two factors, namely the computational complexity as detailed in Section 4.4(1) and the partially vectorized Matlab implementation due to the double summation loops in (16). While the first factor is inherent to the algorithm, the second factor can be overcome by an efficient implementation from software.

---

**Table 1**

<table>
<thead>
<tr>
<th>Method</th>
<th>Ave. Train EER (%)</th>
<th>Ave. Test EER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion(VS + IR)</td>
<td>SUM(Min-Max): 10.7689(N,N)</td>
<td>10.7689(Y,N)(0.00)</td>
</tr>
<tr>
<td></td>
<td>SUM(anh): 9.8962(Y,N)</td>
<td>9.9011(Y,N)(0.05)</td>
</tr>
<tr>
<td></td>
<td>SVMPoly: 7.8400(Y,N)</td>
<td>7.8562(Y,N)(0.21)</td>
</tr>
<tr>
<td></td>
<td>SVMRbf: 19.8550(Y,Y)</td>
<td>20.1045(Y,Y)(1.26)</td>
</tr>
<tr>
<td></td>
<td>LSE: 8.0637(Y,N)</td>
<td>8.1106(Y,N)(0.58)</td>
</tr>
<tr>
<td></td>
<td>LSE-SIG: 17.7103(N,N)</td>
<td>17.7670(N,N)(0.32)</td>
</tr>
<tr>
<td>Fusion(FP + HD)</td>
<td>AACQ−SIG: 7.5846(N,N)</td>
<td>7.6094(0.33)</td>
</tr>
<tr>
<td></td>
<td>LSE−SIG: 11.1031(N,N)</td>
<td>11.1048(N,N)(0.02)</td>
</tr>
<tr>
<td></td>
<td>AACQ: 0.2086(N,N)</td>
<td>0.2073(0.63)</td>
</tr>
<tr>
<td></td>
<td>LSE: 1.0950(N,N)</td>
<td>1.0950(N,N)(0.00)</td>
</tr>
<tr>
<td>NIST-BSSR1</td>
<td>LSE−SIG: 10.950(N,N)</td>
<td>10.950(N,N)(0.00)</td>
</tr>
<tr>
<td></td>
<td>AACQ-SIG: 1.4532(N,N)</td>
<td>1.4548(0.11)</td>
</tr>
<tr>
<td></td>
<td>LSE: 2.6925(N,N)</td>
<td>2.8906(N,N)(7.36)</td>
</tr>
<tr>
<td></td>
<td>LSE−SIG: 1.4548(N,N)</td>
<td>1.4548(0.11)</td>
</tr>
</tbody>
</table>

(Y,Y): Yes, of statistical significance based on a paired \( t \)-test comparing with AACQ at 0.05 significance level. The second Y denotes comparison with AACQ-SIG; (N,N): No, of no statistical significance based on a paired \( t \)-test comparing with AACQ at 0.05 significance level. The second N denotes comparison with AACQ-SIG; (0.00): Percentage change of test performance from training performance based on \( (\text{test-train})/\text{train} \times 100\% \); Bold: Best performance.

---

2 A note for the horizontal scale difference: AACQ is plotted on \( r \in \{1,2,3,4,5,6\} \) whereas AACQ-SIG is plotted on \( N_{SIG} \in \{1,2,5,10,50,100\} \).
memory and architectural aspects which is not the focus of this work.

7.2. Results: 32 publicly available fusion data sets

Here 32 fusion data sets from Ref. [43] are used for experimentation. Fig. 15 shows the average LAUC and average EER values over 32 data sets for different model settings (polynomial order \( r \in \{1, \ldots, 6\} \) for LSE and AAC\(_Q\); \( r \in \{1, 2, 3\} \) for SVMPoly since it encounters convergence problem at higher polynomial order for certain data sets; kernel width \( \gamma \in \{0.1, 1, 5, 10, 50, 100\} \) for SVMRbf). LSE-SIG and AAC\(_Q\)-SIG are plotted over different number of sigmoids (\( NSIG \in \{1, 2, 5, 10, 50, 100\} \)). These results show certain correlation between LAUC and EER values at large but not for fine variations. This is understandable since large variation of ROC area can be related to EER performance variation whereas minor EER difference corresponds to local threshold points on the ROC curve.

From Fig. 15(a), we see fluctuations of LAUC values for SVMPoly, SVMRbf and LSE at different model settings. The AAC\(_Q\)-SIG shows a significantly improved performance as compared to that of LSE-SIG for all model orders. The LAUC performance of AAC\(_Q\) appears to be relatively stable for all model settings. This shows that AAC\(_Q\) maximizes the AUC for all model orders and other methods do not.

Regarding the EER performance in Fig. 15(b), we see similar fluctuation trends to those in LAUC performance for SVMPoly and
SVMRbf. The SVMPoly method shows lower EER values than that of the mean operator at two model settings ($r = 1, 3$). Both LSE (including LSE-SIG) and AAC$_Q$ (including AAC$_Q$-SIG), however, show a somewhat decreasing trend of EER values over increment of model order. At model order $r = 3$ or higher ($N_{SIG} = 50$ for LSE-SIG and AAC$_Q$-SIG), both LSE (LSE-SIG) and AAC$_Q$ (AAC$_Q$-SIG) show lower EER values than that of mean operator. This shows that a complex decision surface is preferred over a simple one when adopting the LSE and AAC$_Q$ methods. The best performed settings among LSE (LSE-SIG), AAC$_Q$ (AAC$_Q$-SIG) and SVMPoly with respect to average EER performance are seen to be comparable. Figs. 16 and 17 show the detailed performances (LAUC and EER) of all individual data sets at these best performed model settings for each fusion classifier.

7.3. Decision landscapes

Fig. 18 shows the score distributions and decision contours plotted over equal and fixed intervals for combining the visual and the infra-red face biometrics. The genuine-user and imposter scores show large overlapping regions and rather indistinguishable
Fig. 16. LAUC values for individual data from the 32 sets: LSE and AACQ at \( r = 5 \), SVMPoly at \( r = 1 \), SVMRbf at \( \gamma = 50 \); LSE-SIG and AACQ-SIG at \( N_{SIG} = 50 \).

Fig. 17. EER values for individual data from the 32 sets: LSE and AACQ at \( r = 5 \), SVMPoly at \( r = 1 \), SVMRbf at \( \gamma = 50 \); LSE-SIG and AACQ-SIG at \( N_{SIG} = 50 \).
Fig. 18. Decision contours of different classifiers when combining two biometrics: (a) least squares error minimization using 3rd-order RM model, (b) SVM using 3rd-order polynomial kernel, (c) SVM using Rbf kernel with gamma = 100, (d) AAC minimization using 3rd-order RM model $\text{AAC}_Q$, (e) LSE-SIG at $N_{\text{SIG}} = 5$, (f) $\text{AAC}_Q$-SIG at $N_{\text{SIG}} = 5$.

distributions. This gives rise to a difficult classification problem. Here, the compared methods include (a) LSE method using the RM model, (b)–(c) SVM using polynomial and RBF kernels, (d) AAC minimization using the RM model $\text{AAC}_Q$, (e) LSE-SIG, and (f) $\text{AAC}_Q$-SIG. Here we see that for LSE methods (LSE and LSE-SIG), the decision landscape is much affected by the density of data since the fixed interval decision contours rarely intersect the rather flat and high density imposter region. For the SVM methods, the decision contours appear to be determined by the distribution structure of data. This is particularly obvious for SVM adopting the RBF kernel.
For all fusion cases using LSE and AAC, two phenomena are observed: (i) the decision contours appear to be unaffected by and run through the high density imposter regions like that in the SVM Poly case, and (ii) the orientations of the contours appear to go along with the two clusters’ distribution direction. This suggests that the decision boundary is determined by classification error distribution rather than by data density apart from the effect of using a less flexible model (low model orders \( r = 1, \ldots, 6 \)).

7.4. Summary of results and comments

- All the four sets of experiments show strong correlation between the LAUC and EER values at large. This suggests that EER could be a good hint for performance even regarding the entire operating range.
- For all fusion cases using LSE and AAC\(_Q\) in the first three sets of experiments, the test results were seen to be consistent to the training results with little over-fitting. The main reason could possibly be due to the low flexibility of models with low polynomial order settings. This enable us to have good confidence to tune the fusion module based on the available training data for prediction of unseen data.
- Fusion of two relatively strong biometrics is seen to have good performance for most fusion classifiers. The EER and LAUC performances of AAC\(_Q\) and AAC\(_Q\)-SIG are seen to improve over that of LSE and LSE-SIG in terms of stability over different model orders.
- Fusion of a relatively strong and a very weak biometric (such as in Fusion\((V5 \rightarrow IR)\), or fusing with a soft biometric) needs particular care. The combined decision may deteriorate if the fusion module is not well trained. Comparing to other training-based methods, the proposed AAC\(_Q\) is relatively less troubled in this aspect as seen from the empirical evidences. The SUM rule using Min-\(\max\) and \(\tan\)h normalization shows deterioration in performance due to unbalanced data.
- The stability of AAC\(_Q\) comes with a higher computational cost than those compared training-based methods. We believe this limitation will be overcome by an efficient software implementation as well as by the advancement of computing technology in future.

8. Conclusion

In this paper, we proposed a direct approach to optimize the ROC performance for multimodal biometric scores fusion. Essentially, the idea is to optimize the area under the ROC curve for decision scores fusion. Utilizing a fusion classifier which is linear in its parameters, an approximate formulation to compute the area under ROC was proposed. This formulation was then solved directly for a classifier which maximizes the area under the ROC curve. When the link–loss functions are matched based on a linear-quadratic pair, the optimal classifier design can be solved in a closed form. We performed extensive experiments on biometrics fusion with practical considerations such as fusing strong and weak cases to observe the effects upon adjusting the design parameters. The proposed algorithm based on ROC optimization was shown to be superior to the LSE density fitting method in many experimented instances particularly the highly overlapping decision cases. Accompanied by a higher computational cost, the proposed method is shown to have comparable or better performance than that of the widely used SUM rule and a well trained SVM adopting different kernels and training settings.

Acknowledgment

The authors would like to thank the following colleagues for their kind assistance in related scores data preparation: Dr. Norman Poh, Sang-Ki Kim, Prof. Xudong Jiang, Lawrence Chen, and Wei Xiong. The provision of fusion database in the public domain from National Institute of Standards and Technology (NIST), USA, and the constructive suggestions by the anonymous reviewers are also acknowledged.

This work was supported by the Korea Science and Engineering Foundation (KOSEF) through the Biometrics Engineering Research Center (BERC) at Yonsei University (Grant Number: R112002105000010(2008)).

References


About the Author—KAR-ANN TOH received the PhD from Nanyang Technological University (NTU), Singapore in 1999. He worked for 2 years in the aerospace industry prior to his post-doctoral appointments at research centres in NTU from 1998 to 2002. He was then affiliated with Institute for Infocomm Research, Singapore from 2002 to 2005. Currently, he is an associate professor with the School of Electrical and Electronic Engineering, Yonsei University, Korea. His research interests include biometrics, pattern classification, and machine learning. He is a co-inventor of a US patent and has made several PCT filings related to biometric applications, apart from having an active publication record in above areas of interest. He has served as a member of technical program committee and as a reviewer for international conferences and journals related to biometrics and pattern recognition.

About the Author—JAIHIE KIM received the PhD degree in electrical engineering in 1984 from the Case Western Reserve University, USA. Since 1984, he has been a professor in the School of Electrical and Electronic Engineering, Yonsei University in Korea, and currently he is also the Director of Biometrics Engineering Research Center in Korea. Dr. Jaihie Kim is a member of the National Academy of Engineering of Korea. His research interests include pattern recognition, computer vision, and biometrics.

About the Author—SANGYOUN LEE received B.S. and M.S. degrees in Electronic Engineering from Yonsei University, Seoul, South Korea in 1987, 1989, respectively. He received the PhD in Electrical & Computer Engineering from Georgia Tech., U.S.A., in 1999. He was a senior researcher in Korea Telecom from 1989 to 2004. He is now an Assistant Professor in Electrical & Electronic Engineering Department, Yonsei University. His research interests include pattern recognition, computer vision, video coding and biometrics.