Score-level Fusion based on the Direct Estimation of the Bayes Error Gradient Distribution

Yasushi Makihara, Daigo Muramatsu, Yasushi Yagi
Osaka University
{makihara, muramatsu, yagi}@am.sanken.osaka-u.ac.jp

Md. Altab Hossain
University of Rajshahi
altab@ru.ac.bd

Abstract

This paper describes a method of score-level fusion to optimize a Receiver Operating Characteristic (ROC) curve for multimodal biometrics. When the Probability Density Functions (PDFs) of the multimodal scores for each client and imposter are obtained from the training samples, it is well known that the isolines of a function of probabilistic densities, such as the likelihood ratio, posterior, or Bayes error gradient, give the optimal ROC curve. The success of the probability density-based methods depends on the PDF estimation for each client and imposter, which still remains a challenging problem. Therefore, we introduce a framework of direct estimation of the Bayes error gradient that bypasses the troublesome PDF estimation for each client and imposter. The lattice-type control points are allocated in a multiple score space, and the Bayes error gradients on the control points are then estimated in a comprehensive manner in the energy minimization framework including not only the data fitness of the training samples but also the boundary conditions and monotonic increase constraints to suppress the over-training. The experimental results for both simulation and real public data show the effectiveness of the proposed method.

1. Introduction

Recently, multimodal biometrics [31] has attracted much attention because multimodal cues realize better performance than individual single cues. Several combinations of biometrics have been attempted up to now, e.g., fingerprint and iris [21], fingerprint and face [3], fingerprint and voiceprint [34], and face and voiceprint [4][32]. Moreover, various public databases for multimodal biometrics are available.

Fusion approaches fall into four groups: sensor-level fusion, feature-level fusion, score-level fusion, and decision-level fusion. Among these, score-level fusion has been widely studied because scores basically exclude privacy information and hence are relatively easily available as a public database [28][2][30][20][19][27][24][10][7][23].

One of the popular score-level fusion approaches is a sum rule where single-modal scores are simply summed up to provide a final score. Moreover, given training samples of scores, score normalization is often incorporated to achieve better performance [1][26].

Another approach is a classification-based method where classification techniques such as Linear Discriminant Analysis (LDA) or Support Vector Machine (SVM) find a decision boundary between a client and imposter based on training samples [9][17][11][6][35].

Although the sum rule and classification-based method do not guarantee optimality for all the operating points on the Receiver Operating Characteristic (ROC) curve, it is well known that Bayes probability density-based methods enable the ROC optimization as long as the Probability Density Functions (PDFs) of client and imposter are identical between training and evaluation [31]. In these methods, the functions of probability density (e.g. the likelihood ratio of client to imposter, posterior of client/imposter, and Bayes error gradient to acceptance rate [18]) are computed from the estimated PDFs of client and imposter over a multi-score space, and isolines of the function are regarded as threshold lines for each operation point on the ROC curve.

The remaining and challenging issue for probability density-based methods is how to accurately estimate the PDFs of a client and an imposter. Approaches to PDF estimation are mainly divided into two groups: parametric and non-parametric methods. The parametric methods assume a certain PDF model such as the Gaussian Mixture Model (GMM) and estimate the parameters of the model from the training sample. While parametric methods have advantages in terms of computational efficiency owing to their compact forms, the methods fail when the selected model does not fit actual PDFs. For example, heavy-tail distributions, which frequently occur with biometrics scores, typically cannot be represented accurately by used GMMs.

On the other hand, non-parametric methods overcome this disadvantage because they bypass the model selection process and estimate the probability densities directly for each control point allocated over a multi-score space. The
methods, however, still suffer from over-training as a result of small sample sizes and troublesome hyper parameter settings (e.g. the kernel size for Gaussian kernel-based estimation methods and the number of nearest neighbors considered in kNN-based estimation).

Rethinking the basics of the problem, it is noticeable that not two PDFs of a client and an imposter but a single function of probability densities such as the Bayes error gradient is sufficient for ROC optimization. Therefore, we propose a method of direct estimation of the Bayes error gradient distribution from the training sample that bypasses the troublesome PDF estimation of both the client and imposter. The proposed approach first allocates lattice-type control points in the multi-score space in the same way as the previous work, and then estimates the Bayes error gradients on the control points in the energy minimization framework, including not only the data fitness of the training samples but also boundary conditions and monotonically increasing constraints as prior knowledge to suppress over-training if necessary.

The remainder of this paper is organized as follows. Section 2 reviews related work on score-level fusion. Section 3 describes a method of score-level fusion based on the direct estimation of the Bayes error gradient. Section 4 presents experiments using both simulation and real public data to evaluate the proposed fusion method, and Section 5 discusses the limitations of the proposed method. Finally, Section 6 gives conclusions and outlines future work for this research.

2. Related work

Transformation-based approaches: In transformation-score fusion, the match scores are first normalized to a common domain and then combined. Choice of the normalization scheme and combination weights are data-dependent in this fusion technique [13][33]. Kittler et al. [14] develop a theoretical framework for combining the evidence obtained from multiple classifiers using the sum rule, product rule, minimum rule, maximum rule, median rule, and majority voting. To apply these schemes, they converted the matching score into posteriori probabilities for a client and imposter. They found that the sum rule outperformed the other methods.

Classification-based approaches: Scores from multiple matchers are treated as feature vectors and a classifier is constructed to discriminate genuine and impostor scores in classifier-based score fusion [9][17][11]. Chen et al. [6] applied a wavelet probabilistic neural network classifier for the combination of the face and iris. The features of the face and iris are extracted using the 1D energy signal and 1D wavelet transform. In [35], two matching scores from a Laplacian face-based verifier and phase information-based iris verifier are combined to form a feature vector and the SVM-based fusion rule is then applied.

Probability density-based approaches: Probability density-based score fusion methods rely on score distribution estimation such as the well-known naive Bayesian [25] and the GMM [22]. In [22], Nandakumar et al. proposed a framework for the combination of match scores based on likelihood ratio tests. The distributions of genuine and imposter match scores are modeled as a finite Gaussian mixture model. They show that the method achieved comparable performance on three multimodal biometric databases with face, fingerprint, iris and speech modalities.

Quality measure-dependent approaches: In addition, a number of quality-based fusion algorithms were recently proposed [29][21][8]. Fierrez-Aguilar [8] proposed a number of fusion schemes based on adaptation of the score fusion functions for multimodal biometric authentication. The adapted fusion approaches were trained using background information (e.g. a pool of users) and then adjusted by considering input information such as the user-dependent scores or test-dependent quality measures. In [29], Poh and Kittler proposed a framework for quality-based fusion of multimodal biometrics from a Bayesian perspective. The framework could be implemented using both generative and discriminative classifiers. Kryszczuk and Drygajlo [15] presented the notion of the Q-stack, namely, a concatenated vector of multiple scores and quality measures, and they defined a multi-dimensional space of the Q-stack as evidence space.

Performance estimation of fusion approaches: In [12], Hube introduced two methods, the root power rule and the exponential convolution rule, for estimating the score level fusion performance of independent modalities. They addressed the problem when there was a need for fusion estimates and when actual score level data were not available or insufficient in quantity. From the input ROC curve that represents the performance for each of the modalities, they rescaled the correct acceptance rate and false acceptance rate to new input values at each threshold. They compared their score level fusion estimation methods with the decision level fusion it AND and OR rule.

3. Score-level fusion

3.1. Functions of probability densities

In this section, we introduce several functions of probability densities used in ROC optimization. In a person authentication scenario, given a pair of an enrollment (template) and a probe (query), the output of multimodal biometrics is either multiple scores (similarity) or distances (dissimilarity). In this paper, we adopt distance-based representation and hence change the sign given to scores-based data.

First, we assume that the system installs $M$-modal biometrics authentication functions and returns an $M$-dimensional distance vector $t = [t_0, \ldots, t_{M-1}]^T \in \mathbb{R}^M$ given a gallery and probe pair. The PDF of the distance
vector \( t \) for each client and imposter are then defined as 

\[ p(t|X = C) \text{ and } p(t|X = I) \], respectively, where \( X \) is 

the label of a client or imposter. We consider next a fusion 

function of the distance vector \( t \) to generate a fused 

scalar distance \( t_F = f(t) \). Given an acceptance threshold 

for the fused distance as \( T_F \), the False Acceptance Rate 

(FAR) \( R_{FA} \) and False Rejection Rate (FRR) \( R_{FR} \) are 

defined as 

\[
R_{FA}(T_F; f) = \int_{f(t) \leq T_F} p(t|X = I) dt 
\]

\[
R_{FR}(T_F; f) = 1 - \int_{f(t) \leq T_F} p(t|X = C) dt. 
\]

The ROC curve in multimodal biometrics is regarded as a 
tradeoff curve of the above FAR and FRR when the threshold 

of the fused distance changes.

Moreover, the ROC curve is also regarded as a tradeoff 
curve of the Half Bayes Error Rate (HBER) and Half Ac- 
cceptance Rate (HAR), which are defined as 

\[
R_{HBER}(T_F; f) = \frac{R_{FA}(T_F) + R_{FR}(T_F)}{2} 
\]

\[
R_{HAR}(T_F; f) = \frac{R_{FA}(T_F) + 1 - R_{FR}(T_F)}{2}. 
\]

The ROC curve optimization problem is equivalent to 
finding the optimal fusion function to realize a set of mini- 
mized pairs of FAR and FRR at each operation point on the 
ROC or to minimize HBER for all HAR. It is well known 
that certain functions of probability densities can be solutions 
of the fusion function. A typical solution is the likelihood ratio of 
imposter to client derived from the so-called likelihood ratio test [31], which is defined as 

\[
l(t) = \frac{p(t|X = I)}{p(t|X = C)} 
\]

Note that the BEGs are close to \(-1\) and \(+1\) in the client-
dominant area \( p(t|X = I) \ll p(t|X = C) \) and imposter-
dominant area \( p(t|X = I) \gg p(t|X = C) \), respectively. 

Moreover, the BEG is related to the posteriors under constant priors as 

\[
g(t) = 2P(X = I|t) - 1 
\]

Here, we consider the property for each function. Whereas the domain of the likelihood ratio is unbounded 
\([0, \infty)\), the domains of the posterior and the BEG are 
bounded at \([0, 1]\) and \([-1, 1]\), respectively, which are suitable for direct estimation. In addition, the BEG and its related ROC representation by HBER and HAR have two tractable properties: (1) the decision boundary \( g(t) = 0 \) 
gives the minimal Bayes error, and (2) the HBER at \( R_{HAR} \) is 
1/2 is the Equal Error Rate (EER). In particular, the first property is significant because it can be used as a pre-
determined operating point to suppress both FAR and FRR, i.e. the Bayes error rate. Therefore, we try to estimate the 
BEG distribution in the following sections.

### 3.2. Direct estimation of BEG distribution

In this section, we describe a direct estimation of the BEG distribution based on the training samples. First, \( N \)
training samples are given as pairs of the \( M \)-dimensional distance vectors and client/imposter labels \( \{t_i, y_i\}\) \((i = 0, \ldots, N - 1)\), where \( y_i = 1 \) for the client and \( y_i = -1 \) for the imposter. In addition, a weight \( w_i \) for \( i \)th training sample is set as 

\[
w_i = \begin{cases} 
\frac{1}{2N_C} & (y_i = 1) \\
\frac{1}{2N_I} & (y_i = -1)
\end{cases}
\]

where \( N_C \) and \( N_I \) are the numbers of the client and imposter training samples, respectively.

Next, lattice-type control points are allocated in \( M \)-
dimensional multi-score space as 

\[
t^{cp}_{i_0, \ldots, i_{M-1}} = [t^\min_m + \ell_0 s_m, \ldots, t^\min_m + i_{M-1} s_m] 
\]

\((i_m = 0, \ldots, N^cp_m)\),

where \( t^\min_m \), \( s_m \), and \( N^cp_m \) are the minimum score, the quantization step, and the number of quantizations decided in order to cover the maximum score for the \( m \)th-modal score, respectively. An example of lattice-type control points in the two-dimensional case is shown in Fig. 1. Consequently, the number of total control points is \( N^cp = \prod_{m=0}^{M-1} (N^cp_m + 1) \) and the BEGs on all the control points are vectorized as \( g^{cp} \). Moreover, given the BEG on the control point \( t^{cp}_{i_0, \ldots, i_{M-1}} \), a BEG on an arbitrary point \( t \) in multi-score space is expressed by a multi-linear interpo-
Figure 1. Illustration of lattice-type control points in a two-dimensional case.

Sponds to the BEG. Hence, the data fitness terms drive the rate, respectively, and for which the resultant ratio corresponds to the Bayes error rate and acceptance at the control point to the addition of the summed weights, traction of the summed weights for the imposter and client.

Finally, the energy function is a quadratic form of $g^{cp}$, hence the optimal BEG distribution $g^{cp}$ is easily obtained by linear solution.

3.3. Incorporation of prior knowledge

In this section, a method of incorporating prior knowledge to prevent over-estimation is addressed. Because, in general, the number of training samples is limited, the BEG distribution may sometimes be over-estimated, which results in large Bayes generalization errors. While the over-training problem can be solved to some extent by strengthening the smoothness coefficient $\lambda$, too large a smoothness coefficient may corrupt the original distribution.

Another solution is the use of prior knowledge on the BEG distribution. In almost every case, we can observe that clients are dominant in small distance areas while imposters are dominant in large distance areas. In particular, the BEGs on the two extreme control points, that is, the point of the minimum distance set $t_0, \ldots, 0$ and the point of the maximum distance set $t_{N^p}, \ldots, t_{N^p}^{M-1}$, should be subject to the following boundary conditions.

$$g^{cp}_{0, \ldots, 0} = -1$$
$$g^{cp}_{N^p, \ldots, N^p} = 1$$

In addition, because BEGs are expected to increase as the distances of each dimension increase, we use the following monotonically increasing constraint on the BEG distribution.

$$g^{cp}_{i_0, \ldots, i_{M-1}} \geq g^{cp}_{i_0, \ldots, i_{M-1}} \forall i_0, \ldots, i_{M-1}$$
$$\vdots$$

$$g^{cp}_{i_0, \ldots, i_{M-1}+1} \geq g^{cp}_{i_0, \ldots, i_{M-1}} \forall i_0, \ldots, i_{M-1}$$

Therefore, we incorporate the above boundary conditions and monotonically increasing constraints on the BEGs on the control points into the energy minimization framework in Eq. (13). Consequently, the whole optimization framework is expressed in a quadratic form accompanied by linear equality and inequality constraints in terms of the BEGs on the control points $g^{cp}$, and the solution is given by convex quadratic programming.

Note that these types of prior knowledge are based on the relationship between imposter and client PDFs, and hence it cannot be applied to the independent estimation problem of each client and imposter PDF.

4. Experiments

4.1. Simulation data

In this section, we describe experiments using two-dimensional simulation data that assume a score-level fusion problem of bi-modal biometrics. First, the imposter samples of distances are drawn from a single two-dimensional $\chi^2$ distribution with a pair of $(8,8)$ Degrees of Freedom (DoF) for the two dimensions. On the other hand, the client samples of distances are drawn from a
mixture of two-dimensional $\chi^2$ distributions with a pair of $(3, 3)$, $(3, 6)$, and $(6, 3)$ DoF for the two dimensions, and the weight for each distribution is set at $1/3$. The pairs of $(3, 6)$ and $(6, 3)$ assume that the client distances of either the first or second modality may enlarge owing to low-quality measures (e.g. low-quality sensors, and large differences in observation conditions between enrollment and trial). In addition, note that the $\chi^2$ distribution is more suitable for the evaluation of the biometrics score-level fusion because it has a heavier tail than is frequently used in Gaussian distribution, and to some extent the biometrics score has a heavy tail property.

Next, 1,000 client and 10,000 imposter training samples are drawn from the distributions, as shown in Fig. 2(a). 10,000 client and 10,000 imposter test samples are then drawn from the same distribution.

The benchmark methods are the sum rule as a non-training method, GMM as a parametric PDF estimation (denoted as GMM later), and a Gaussian kernel method as a nonparametric PDF estimation (denoted as Kernel later). Moreover, the proposed methods without and with prior knowledge are denoted as Direct and Direct + Prior, respectively.

With regard to the parameters, the step of the control points for each dimension is predetermined as 1.0, and the standard deviation for the Kernel is manually chosen as half a step, that is, 0.5 in these simulation experiments. The C0 and C1-smoothness coefficients $\lambda_0$ and $\lambda_1$ in the energy minimization framework are predetermined as 0.0 and 1.0, respectively. The numbers of mixtures for the client and imposter in the GMM are predetermined as 3 and 1, respectively, which are the same as the numbers of mixtures for the original PDFs.

First, the estimated BEG distributions for probability density-based methods are shown with the ground truth in Fig. 3. Because the BEG can be seen as a fused distance and its isolines serve as threshold lines in two-dimensional distance space as previously discussed, the summed distances and its isolines for the sum rule are also illustrated in Fig. 3(a) for comparison.

We can see the isoline patterns of the summed distance (Fig. 3(a)) are clearly different from those of the ground truth BEG distribution (Fig. 3(f)). The BEG distribution of the GMM (Fig. 3(b)) is also slightly different from the ground truth (Fig. 3(f)) because it cannot accurately express the heavy tail of the $\chi^2$ distribution, particularly in the area illustrated by the red dot circle in Fig. 3(b). The BEG distribution for the Kernel (Fig. 3(c)) is globally similar to that of the ground truth (Fig. 3(f)). It is, however, jagged to some extent and also partly different from the ground truth, particularly in the sparsely sampled areas illustrated by the red dot circle in Fig. 3(c).

On the other hand, the proposed Direct (Fig. 3(d)) can reconstruct a considerably smoother BEG distribution similar to the ground truth. Moreover, the proposed Direct + Prior (Fig. 3(e)) corrects several bumps of the Direct illustrated by the red dot circle in Fig. 3(d) thanks to prior knowledge of the boundary conditions and monotonically increasing constraints. The resultant ROC curves after score-level fusion are shown in Fig. 2(b). As a result, the proposed Direct and Direct + Prior achieve almost the same performance as the fusions based on the ground truth BEG distribution (BEG (GT)).

Finally, because the number of training samples, particularly the client training samples, is usually limited, we investigate the impact of the number of client training samples on the EER as shown in Fig. 4. As a result, the EER for the Kernel increases as the number of client training samples decreases and it becomes worse than the sum-rule in a domain of less than 100 client training samples. The EER for the Direct slowly increases as the number of client training samples decreases.
The number of training samples is smaller than that of the simulation data. Because of useful prior knowledge, and is competitive to the sum rule in the case of quite small numbers of client training samples.

4.2. Real data

In this section, experiments using real public data, Biosecur DS2 dataset [28][5] are addressed. The dataset includes scores of four modalities: face, thumb (fin1), index finger (fin2), and middle finger (fin3) with associated quality measures. Because the focus of this paper is score-level fusion, only the scores are used in this experiment. In addition, the proposed method was applied to each pair of the four modalities, and a total of six bi-modal score-level fusions were evaluated.

The dataset was divided into training and test sets. Outliers of measure failures are included in both sets, and those in the training set were removed manually for better training, while those in the test set remained. Consequently, approximately 200 client and 42,000 imposter training samples were used for estimating the BEG distributions, and 1,256 client and 307,680 imposter test samples were used for evaluation. For example, the training samples for face and thumb are shown in Fig. 5(a).

As for the parameters, the step of the control points for the simulation data, we make not only the C0-smoothness coefficient \( \lambda_0 \) but also the C0-smoothness coefficient \( \lambda_1 \) valid, namely, we predetermined the coefficients as 1.0 and 1.0, respectively, so as to suppress over-training. On the other hand, the number of mixtures for client and imposter in the GMM is determined within a range 1 to 5 for each dimension so as to minimize the Bayes error rate for the training sample when the acceptance threshold 0.0 for the BEG is applied.

First, the resultant ROC curves for a pair of a face and thumb are shown in Fig. 5(b). We can see that the proposed Direct and Direct + Prior achieve better performance than the other probability density-based methods: GMM and Kernel. On the other hand, the sum rule is fairly competitive with the proposed methods.

The EERs and Half Total Error Rate (HTER) for all the pairs of the four modalities are shown in Fig. 6. The HTER is half of the Bayes error rate when a predetermined acceptance threshold is applied to the test set. The threshold is usually decided so as to minimize the Bayes error rate for the training set with the exception that the GMM employs a theoretical threshold 0.0.

As a result, we can see that either the proposed method Direct + Prior or the sum-rule gives the best performance.

One of the reasons for the competitiveness is that the isolines of the BEG as acceptance threshold curves are close to lines in several areas and the isolines of the summed distance give the better approximation of the isolines of the BEG for this dataset.

Another reason is the difference between the PDFs in the training and test samples, an issue of the generalized Bayes error. The problem of the generalized Bayes error is, to a greater or lesser extent, basically related to all the training-based score-level fusion approaches including transformation, classification, and probability density-based approaches. The proposed method Direct + Prior, however, can to some extent still suppress such a generalized Bayes error by incorporating widely applicable prior knowledge, and hence it almost always gives better performance than the proposed method without prior knowledge Direct.

5. Discussion

As discussed in the previous section, the generalization capability is a significant issue. It is observed that generalization capability of the proposed method decreases while the sum rule becomes superior as the number of training samples decreases as shown in Fig. 4. In addition, as the coefficient of the C0 and C1-continuity smoothness \( \lambda_0 \) and \( \lambda_1 \) increase in the proposed energy minimization framework, the estimated BEG distribution gets closer to a planar form.
similar to the summed distance (Fig. 3(a)). Therefore, one possible strategy is adaptive smoothness coefficient control based on the number of training samples.

Another point to note is a limitation of the proposed method in the case of much higher-dimensional score-level fusion. As mentioned in Section 3.2, the number of lattice-type control points \( N_{\text{cp}} \) increases exponentially with dimension, that is, the number of modalities. To extend the proposed method to a much higher dimension, one possible choice is the incorporation of a dimension reduction technique such as Kryszczuk and Poh [16] exploited with a locality preserving projection to handle high dimensionality. The other choice is to replace the lattice-type control points with floating control points, which increases not exponentially with the dimension but proportionally to the number of training samples in the future.

Although fusion methods of multiple biometric scores were the focus of this paper, quality measures should also be considered for better performance. Kryszczuk and Drygajlo [15] proposed the notion of a Q-stack, namely, a concatenated vector of multiple scores and quality measures, and they defined a multi-dimensional space of the Q-stack as evidence space. The proposed method is in principle applicable to such evidence space after replacing the multi-dimensional distance vector \( t \) with the Q-stack, although countermeasures to handling the high dimensionality discussed above are essential.

Finally, missing data are often the case with multimodal biometrics; typically in a public database [28][5]. We need to take the treatment of missing data into consideration for better performance as a whole in future work.

6. Conclusions

We proposed a framework of score-level fusion for multimodal biometrics based on the direct estimation of the BEG distribution, which enables us to bypass the troublesome PDF estimation for each client and imposter. The BEGs on the lattice-type control points are in the energy minimization framework including the data fitness and smoothness terms in conjunction with widely applicable prior knowledge, namely, boundary conditions and monotonically increasing constraints to enhance the generalization capability. The experimental results for both simulation and real public data showed the effectiveness of the proposed method.

Future directions are in adaptive smoothness control to cope with the generalized Bayes error problem and the introduction of the dimension reduction techniques or the floating control points for the extension to higher-dimensional score-level fusion.

Acknowledgement

This work was partially supported by “R&D Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Society”, Strategic Funds for the Promotion of Science and Technology of the Ministry of Education, Culture, Sports, and Technology, the Japanese Government, Grant-in-Aid for Scientific Research (S) 21220003, CREST (Core Research for Evaluation Science and Technology) of JST (Japan Science and Technology).

References


