Modality Selection Attacks and Modality Restriction in Likelihood-Ratio Based Biometric Score Fusion*

Takao MURAKAMI††, Yosuke KAGA††, and Kenta TAKAHASHI††, Members

SUMMARY The likelihood-ratio based score level fusion (LR fusion) scheme is known as one of the most promising multibiometric fusion schemes. This scheme verifies a user by computing a log-likelihood ratio (LLR) for each modality, and comparing the total LLR to a threshold. It can happen in practice that genuine LLRs tend to be less than 0 for some modalities (e.g., the user is a “goat”, who is inherently difficult to recognize, for some modalities; the user suffers from temporary physical conditions such as injuries and illness). The LR fusion scheme can handle such cases by allowing the user to select a subset of modalities at the authentication phase and setting LLRs corresponding to missing query samples to 0. A recent study, however, proposed a modality selection attack, in which an impostor inputs only query samples whose LLRs are greater than 0 (i.e., takes an optimal strategy), and proved that this attack degrades the overall accuracy even if the genuine user also takes this optimal strategy. In this paper, we investigate the impact of the modality selection attack in more details. Specifically, we investigate whether the overall accuracy is improved by eliminating “goat” templates, whose LLRs tend to be less than 0 for genuine users, from the database (i.e., restricting modality selection). As an overall performance measure, we use the KL (Kullback-Leibler) divergence between a genuine score distribution and an impostor’s one. We first prove the modality restriction hardly increases the KL divergence when a user can select a subset of modalities (i.e., selective LR fusion). We second prove that the modality restriction increases the KL divergence when a user needs to input all biometric samples (i.e., non-selective LR fusion). We conduct experiments using three real datasets (NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand), and discuss directions of multibiometric fusion systems.

key words: multibiometric fusion, likelihood ratio, modality selection attack, KL divergence, goat

1. Introduction

A number of multibiometric fusion schemes have been studied to improve the accuracy of biometric authentication systems [2]. These schemes combine multiple sources of biometric information (e.g., fingerprint, face, and vein; left fingerprint and right fingerprint), and can be roughly divided into the following categories: feature level fusion, score level fusion, and decision level fusion. They combine multiple biometric features, multiple matching scores (distances or similarities), and multiple decision results, respectively. Among them, score level fusion have been particularly well studied (e.g., see Chapter 4 of [2]) due to its general applicability; it can be applied to any kind of biometric system that outputs scores. For example, feature level fusion cannot be applied to most commercial biometric systems, since they do not provide access to biometric features. On the other hand, score level fusion can be applied to these systems if they provide access to scores (the one-to-one matching function in BioAPI [3] outputs a score). In addition, score level fusion generally provides much higher accuracy than decision level fusion, since scores contain much richer information than decision results.

The likelihood-ratio based score level fusion scheme (referred to as the LR fusion scheme) [4] is known as one of the most promising score level fusion schemes. This scheme computes a log-likelihood ratio (LLR), which is the logarithm of the ratio between the likelihood of being a genuine user and the likelihood of being an impostor, using scores. Then it compares the LLR with a pre-determined threshold (if the LLR exceeds the threshold, accept; otherwise, reject). In multiple modalities (e.g., fingerprint, face, and vein) or multiple instances (e.g., left fingerprint and right fingerprint) [2], it is often assumed that all scores are independent [5]–[8]. Then, the LR fusion scheme can compute the total LLR by summing up an LLR for each modality (or instance) This scheme is based on the likelihood-ratio test [9], and minimizes, for a given FRR (False Reject Rate), FAR (False Accept Rate) among all score level fusion schemes (that use the same scores as information sources) if the total LLR is accurately estimated (i.e., Neyman-Pearson lemma [9]).

In this paper, we refer to an LLR in the case when a genuine user (resp. impostor) attempts authentication as a genuine (resp. impostor) LLR. Ideally, genuine (resp. impostor) LLRs should be greater than (resp. less than) 0 for all modalities. However, it can happen in practice that genuine LLRs tend to be less than (or equal to) 0 for some modalities. For example, a user can be a “goat” (a person who is inherently difficult to recognize) [10] for some modalities. For another example, a user might not be able to appropriately input some query samples (biometric samples at the authentication phase) due to temporary physical conditions (e.g., finger injuries, sore throats). The LR fusion scheme can handle such cases by allowing the user to select a subset of modalities at the authentication phase and setting the LLRs corresponding to missing query samples to 0 [5]. In this paper, we refer to this mode as a modality selection mode, and the LR fusion scheme that adopts this mode as a selective LR fusion scheme (we use the term “selective” in the same manner as for selective feature level fusion).
way as [11]). The selective LR fusion scheme provides a usable way of authentication for users who have difficulty in achieving high LLRs for some modalities.

However, a recent study [12], [13] proposed a modality selection attack against the selective LR fusion scheme. In this attack, an impostor (attacker) inputs only query samples whose LLRs are greater than 0 to impersonate others. Since the total LLR is maximized, this attack is optimal in the sense that it maximizes FAR among all possible combinations of modalities. The impostor can find such vulnerable modalities by using some background knowledge about the target user’s biometrics. For example, many people make a part of their biometric information (e.g., face, voice) to the public. A residual fingerprint might be found on a sensor surface or glass surface. After obtaining such background information, the impostor computes a score (or LLR) between his/her own biometric sample and the target user’s biometric sample offline (i.e., offline analysis). Since a score (or LLR) in the offline analysis is highly correlated with an LLR output by the target system, the impostor can predict, for each modality, whether an LLR will be larger than 0.

It is worth noting that a genuine user can also takes this optimal strategy; she can input only query samples whose LLRs are greater than 0 to maximize GAR (= 1 – FRR). The genuine user can find such modalities because she knows from experience that some modalities tend to cause false rejection (i.e., goat), or because she knows her own temporary physical conditions (e.g., injuries, illness). However, Murakami et al. [12], [13] proved that the modality selection attack degrades the overall accuracy even if the genuine user also takes this optimal strategy. More specifically, they used the KL (Kullback-Leibler) divergence between a genuine score distribution and an impostor’s one [14], [15], which can be directly compared with password entropy [16], [17], as a measure of the overall accuracy. Then they proved that the KL divergence of the selective LR fusion scheme explained above is smaller than that of the non-selective LR fusion scheme (which requires a user to input all biometric samples enrolled in the database).

The aim of this paper is to investigate the impact of the modality selection attack in more details. More specifically, we consider whether the overall accuracy can be improved by eliminating “goat” templates, whose LLRs tend to be less than (or equal to) 0 for genuine users, from the database. For example, consider a multimodal system that combines fingerprint, face, and voice using the selective LR fusion scheme. Alice knows from experience that she is a goat for voice, and asks a system operator to eliminate her voice template from the database (there are also methods to automatically detect goat templates [10], [18])\(^7\). After the elimination, modalities that can be selected at the authentication phase are restricted to fingerprint and face. In this paper, we refer to the LR fusion scheme that eliminates goat templates in this way as a restricted LR fusion scheme. Since Alice’s LLR tends to be less than (or equal to) 0 for a voice template, her LLR will be greater than 0 for Alice’s voice template, cannot input voice (i.e., she cannot take an optimal strategy). Thus, the elimination of the voice template might mitigate the impact of the modality selection attack (and therefore decrease FAR).

To thoroughly investigate the effect of the modality restriction explained above, we divide the LR fusion scheme into two categories (selective and non-selective) and two sub-categories (restricted and non-restricted) as follows:

- **SLR**: selective and non-restricted LR fusion.
- **SLR\(^\ast\)**: selective and restricted LR fusion.
- **NLR**: non-selective and non-restricted LR fusion.
- **NLR\(^\ast\)**: non-selective and restricted LR fusion.

Figure 1 shows four types of the LR fusion schemes. For example, in **SLR**, the genuine user inputs fingerprint and face, while the impostor inputs face and voice (i.e., both of them take an optimal strategy). In **SLR\(^\ast\)**, since the genuine user eliminated the voice template from the database, the impostor cannot input voice (i.e., she cannot take an optimal strategy). In **NLR**, both the genuine user and the impostor need to input all of the three biometric samples. In **NLR\(^\ast\)**, they need to input fingerprint and face.

We investigate, both theoretically and experimentally, whether the overall accuracy is improved by the modality restriction using the KL divergence explained above. Murakami et al. [12], [13] proved that the KL divergence of **SLR** is smaller than that of **NLR**. In this paper, we compare the KL divergence of **SLR\(^\ast\)** with that of **SLR**, and the KL divergence of **NLR\(^\ast\)** with that of **NLR**. To our knowledge,\(^7\)\footnote{As an alternative, the system can use a user-specific genuine score distribution [11] to more accurately compute an LLR for a goat template. However, since the number of templates (training samples) is generally very small, it is difficult to accurately estimate a user-specific genuine score distribution. In particular, the standard deviation is much harder to reliably estimate than the average [11]. Thus, we do not consider a user-specific genuine score distribution in this paper.}

\(\textbf{Table 1} \quad \texttt{Genuine User} \quad \texttt{Impostor} \quad \texttt{Selective} \quad \texttt{Non-selective} \quad \texttt{Restricted} \quad \texttt{Restricted} \quad \texttt{Non-restricted} \quad \texttt{Non-restricted}\)

\begin{tabular}{|c|c|c|c|c|c|}
\hline
\textbf{Ft} & \textbf{Fa} & \textbf{Vo} & \textbf{Ft} & \textbf{Fa} & \textbf{Vo} & \textbf{Ft} & \textbf{Fa} & \textbf{Vo} \\
\hline
\textbf{LLR} > 0 & \textbf{LLR} \leq 0 & \textbf{LLR} > 0 & \textbf{LLR} > 0 & \textbf{LLR} \leq 0 & \textbf{LLR} > 0 & \textbf{LLR} \leq 0 \\
\hline
\end{tabular}

\(\texttt{SLR} \quad \texttt{SLR} \quad \texttt{NLR} \quad \texttt{NLR}\)

\(\texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓} \quad \texttt{✓}\)
this is the first study to thoroughly investigate the overall accuracy of four types of the LR fusion schemes.

1.1 Our Contributions

The contributions of this paper, which are common to the previously published conference paper [1], are as follows:

- We theoretically prove a negative result that the modality restriction hardly increases the KL divergence in the selective LR fusion scheme. This means that the KL-divergence loss caused by the modality selection attack cannot be prevented by eliminating goat templates from the database (Sect. 3.2).
- We theoretically prove that the modality restriction can increase the KL divergence in the non-selective LR fusion scheme. This means that the KL divergence can be increased by eliminating goat templates when the system does not operate in the modality selection mode (Sect. 3.3).

This paper is an extension of the previously published conference paper [1]. The main enhancements are as follows:

- In [1], we conducted experiments using only the NIST BSSR1 Set1 dataset [19]. In this paper, we add two more real datasets (the Biosecure DS2 dataset [20] and the CASIA-Iris-Thousand dataset [21]) to provide enough evidence (Sect. 4).
- In our theoretical analysis, we assume that a genuine user successfully eliminates goat templates (i.e., there are no goat detection errors). However, there are goat detection errors in practice. We evaluate the effect of the goat detection errors on the accuracy of SLR and NLR in our experiments (Sect. 4.3).
- We show that the impostor can perform the modality selection attack without any background knowledge about the target user’s biometrics in our experiments (Sect. 4.4).

Regarding the third enhancement, it may be very difficult to obtain background knowledge about the target user’s biometrics in practice (especially for modalities such as iris, vein, and retina). It also takes a lot of effort for the impostor to compute a score (or LLR) offline (i.e., offline analysis). However, the impostor may perform the modality selection attack without such background knowledge, especially when the system uses modalities with high accuracy (e.g., vein, iris) along with modalities with low accuracy (e.g., face, voice). For example, consider a multibiometric system that combines vein and face using selective LR fusion. Since the impostor knows that an LLR will be less than 0 for vein in most cases, she would input only face, which is optimal in most cases. In this case, the impostor does not use any background knowledge about the target user’s biometrics. Although such an attack is discussed in [1], [13], they did not perform experiments to show its effectiveness. In this paper, we show that this attack is indeed effective.

Based on the above results, we discuss directions of the multibiometric system using the LR fusion scheme in Sect. 5.

1.2 Paper Organization

The rest of this paper is organized as follows. In Sect. 1.3, we introduce the basic notations and assumptions in this paper. In Sect. 2, we describe the previous work closely related to ours. In Sect. 3, we present theoretical analysis on the four types of the LR fusion schemes. In Sect. 4, we report experimental results. Finally, in Sect. 5, we conclude this paper with directions of the multibiometric system using the LR fusion scheme.

1.3 Basic Notations and Assumptions

In this paper, we focus on a multibiometric system that combines multiple modalities (e.g., fingerprint, face, and voice) or multiple instances (e.g., left fingerprint and right fingerprint) [2] using the LR fusion scheme. Let $M$ be the total number of modalities (or instances) used in the multibiometric system. We assume that a score is discrete$^1$. Let $X$ be a score space, and $x(m) \in X (1 \leq m \leq M)$ be a score between the $m$-th query sample and the corresponding template. We define the following two hypotheses:

$H_1$: The user is genuine.
$H_0$: The user is an impostor.

We then make the following two assumptions:

(i) Scores $x(1), \ldots, x(M)$ are independent of each other.
(ii) Score $x(m)$ is generated from a genuine score distribution $f_{gen}(m)$ if $H_1$ is true, and from an impostor distribution $f_{imp}(m)$ if $H_0$ is true:

$$P(x(m)|H_1) = f_{gen}(m),$$

$$P(x(m)|H_0) = f_{imp}(m).$$

These assumptions are often made in multiple modalities (or instances) [5]–[8], as described in Sect. 1.

Let $y(m) \in \mathbb{R}$ be an LLR at the $m$-th modality. For example, LLR $y(m)$ in NLR can be written as follows:

$$y(m) = \log \frac{P(x(m)|H_1)}{P(x(m)|H_0)} = \log \frac{f_{gen}(m)}{f_{imp}(m)}.$$  

We then denote a genuine distribution of LLR $y(m)$ in SLR, SLR$, NLR$, and NLR by $g_{gen}(m)$, $g_{gen}(m)$, $h_{gen}(m)$, and $h_{gen}(m)$, respectively. Similarly, we denote an impostor distribution of LLR $y(m)$ in SLR, SLR$, NLR$, and NLR by $g_{imp}(m)$, $g_{imp}(m)$, $h_{imp}(m)$, and $h_{imp}(m)$, respectively. Table 1 shows the basic notations used in this paper.

$^1$This assumption is made to simplify our theoretical analysis. In many cases, a score is defined as a continuous value (e.g., a similarity value between 0 and 1). However, this value is generally rounded to a float-point number (e.g., 4-byte or 8-byte floating point), which is discrete. Thus, even when a score is defined as continuous, our assumption is appropriate in practice.
2. Related Work

In this section, we describe the previous work closely related to ours. Specifically, we describe the LR fusion scheme [4], [5] (Sect. 2.1), a measure of the overall accuracy that can be compared with password entropy [14], [15] (Sect. 2.2), and the modality selection attack [12], [13] (Sect. 2.3).

2.1 Likelihood Ratio Based Score Level Fusion

The LR fusion scheme was proposed by Nandakumar et al. [4]. In [4], they assumed that a user inputs all query samples in the database (i.e., NLR). After observing scores \(x^{(1)}, \ldots, x^{(M)}\), this scheme computes a total LLR given by

\[
\zeta^{(tot)} = \log \frac{P(x^{(1)}, \ldots, x^{(M)}|H_1)}{P(x^{(1)}, \ldots, x^{(M)}|H_0)}.
\] (4)

Under the assumptions (i) and (ii) in Sect. 1.3, the total LLR \(\zeta^{(tot)}\) can be decomposed using (3) as follows:

\[
\zeta^{(tot)} = \sum_{m=1}^{M} \log \frac{P(x^{(m)}|H_1)}{P(x^{(m)}|H_0)} = \sum_{m=1}^{M} \zeta^{(m)}.
\] (5)

The LLR \(\zeta^{(m)}\) is estimated from a score \(x^{(m)}\) by using, for example, a logistic regression model [22]. This model estimates the LLR \(\zeta^{(m)}\) as a linear function of the score \(x^{(m)}\), and its effectiveness is shown in score level fusion [8], [23], [24]. After computing the total LLR \(\zeta^{(tot)}\) in (5), the system verifies a user by comparing \(\zeta^{(tot)}\) with a pre-determined threshold \(A\): if \(\zeta^{(tot)} > A\), accept; otherwise, reject.

The LR fusion scheme can also allow a user to select a subset of modalities at the authentication phase (i.e., SLR) by simply setting LLRs \(\zeta^{(m)}\) corresponding to missing query samples to \(\zeta^{(m)} = 0\) [5]. After computing the total LLR \(\zeta^{(tot)}\) in this way, the system verifies a user by comparing \(\zeta^{(tot)}\) with a threshold \(A\) in the same way as NLR.

Many studies have been made on the LR fusion scheme. Nandakumar et al. [5] extended the LR fusion scheme to identification. They also showed in [6] that the assumptions (i) and (ii) in Sect. 1.3 do not adversely affect the verification accuracy even in multiple matchers. Tao and Veldhuis [7] proposed a method to estimate an LLR via the ROC (Receiver Operating Characteristics) curve. Some studies [25], [26] showed that if the impostor perfectly spoofs one of the modalities in the LR fusion scheme, she can significantly increase FAR of the system.

Another research direction related to the LR fusion scheme is a sequential fusion scheme [8], [27] based on SPRT (Sequential Probability Ratio Test) [28]. This scheme compares the total LLR \(\zeta^{(tot)}\) with the threshold \(A\) each time a user inputs a query sample. In [29], they extended the SPRT-based sequential fusion scheme to identification.

2.2 BSE and the KL Divergence

Takahashi and Murakami [14], [15] proposed the BSE (Biometric System Entropy) as a measure of the overall accuracy in biometrics. The feature of the BSE is that it can be directly compared with password entropy [16], [17], which is widely used to measure the strength of passwords or PINs. It is sometimes beneficial to compare the discriminative power of biometrics with that of other authentication factors such as passwords and PINs. For example, a company that is planning to introduce an authentication system may decide what kind of authentication factors to use based on their discriminative power (e.g., if the discriminative power of fingerprint is much larger than that of 4-digit PINs, the company chooses fingerprint; otherwise, she chooses 4-digit PINs). Since the BSE can be compared with password entropy, it is suitable for such a situation.

We begin by explaining password entropy. Let \(\mathcal{A}\) be a set of (discrete) personally identifiable information (e.g., passwords, PINs), and \(A \in \mathcal{A}\) is a random variable representing personally identifiable information. Then, password entropy is defined as \(H(A)\) (i.e., entropy of \(A\)). For example, password entropy of user chosen 4-digit PIN (which is generally non-uniform) is estimated to be 9 bits in [16].

It should be noted that password entropy can also be expressed as the mutual information between a user and personally identifiable information. Let \(\mathcal{U} = \{u_1, \ldots, u_N\}\) be a set of enrollees in a certain application (e.g., physical access control system), and \(U \in \mathcal{U}\) be a random variable representing a user. Then, since personally identifiable information \(A\) is uniquely determined given user \(U\) (e.g., 4-digit PIN of \(u_1\) is 3952), password entropy \(H(A)\) can also be expressed as the mutual information \(I(U; A)\):

\[
H(A) = H(A) - H(A|U) = I(U; A).
\] (6)

Based on this fact, Takahashi and Murakami [14], [15] proposed the BSE as a measure of the overall accuracy in biometrics. We begin by explaining the BSE in a unimodal biometric system (\(M = 1\)). Let \(x_n \in \mathbb{X}\) be a score between a query sample and a template of user \(u_n \in \mathcal{U}\). \(x = \{x_1, \ldots, x_N\}\) be a set of scores for all users \(u_1, \ldots, u_N\), and \(\mathbf{X}\) be a random variable whose realization is \(x\). Then they defined the BSE as the mutual information \(I(U; \mathbf{X})\).
Since the BSE is the mutual information between a user and personally identifiable information in the same way as password entropy, it can be compared with password entropy.

The problem of the BSE is that it is difficult to compute \( I(U; X) \) in practice (especially when the number of enrollees \( N \) is large). Thus, Takahashi and Murakami [14], [15] proposed to use the KL divergence as an approximation of the BSE. Let \( f_{gen} \) (resp. \( f_{imp} \)) be a genuine (resp. impostor) score distribution. Then, the KL divergence between these two distributions can be expressed as follows:

\[
D(f_{gen} \parallel f_{imp}) = \sum_{x \in \mathcal{X}} f_{gen}(x) \log \frac{f_{gen}(x)}{f_{imp}(x)}. \tag{7}
\]

They proved that the BSE \( I(U; X) \) converges to \( D(f_{gen} \parallel f_{imp}) \) as the number of enrollees \( N \) increases:

**Lemma 1 (BSE and the KL divergence) [14], [15]:**

\[
I(U; X) \rightarrow D(f_{gen} \parallel f_{imp}) \quad (N \rightarrow \infty). \tag{8}
\]

Thus, the KL divergence \( D(f_{gen} \parallel f_{imp}) \) can be used as an approximation of the BSE \( I(U; X) \). The KL divergence can be estimated by using, for example, the generalized k-NN (Nearest Neighbor) estimator [30].

The BSE can also be applied to the multibiometric fusion system \((M > 1)\) by regarding \( x_m \in \mathcal{X} \) as an integrated score (i.e., final score). In this case, \( D(f_{gen} \parallel f_{imp}) \) is the KL divergence for all modalities, and can be upper bounded by the summation of the KL divergence \( D(f_{gen}^{(m)} \parallel f_{imp}^{(m)}) \) for each modality \((1 \leq m \leq M)\) (for details, see [15]).

### 2.3 Modality Selection Attacks

Murakami et al. [12], [13] proposed a modality selection attack against the selective LR fusion scheme (SLR). In this attack, an impostor inputs only query samples whose LLRs are greater than 0 (i.e., takes an optimal strategy). Murakami et al. [12], [13] proved that when both genuine users and impostor input only query samples whose LLRs are greater than 0, the KL divergence \( D(g_{gen}^{(m)} \parallel g_{imp}^{(m)}) \) in SLR becomes smaller than the KL divergence \( D(h_{gen}^{(m)} \parallel h_{imp}^{(m)}) \) in NLR:

**Lemma 3 (KL-divergence loss) [12], [13]:**

\[
D(h_{gen}^{(m)} \parallel h_{imp}^{(m)}) - D(g_{gen}^{(m)} \parallel g_{imp}^{(m)}) = \alpha_m \log \frac{1 - \beta_m}{\alpha_m} - \gamma_m > 0, \tag{9}
\]

where

\[
\alpha_m = \sum_{y^{(m)} < 0} h_{gen}^{(m)}(y^{(m)}) \quad (10)
\]

\[
\beta_m = \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)}) \quad (11)
\]

\[
\gamma_m = -\sum_{y^{(m)} < 0} h_{gen}^{(m)}(y^{(m)}) y^{(m)}(> 0). \quad (12)
\]

\(D(h_{gen}^{(m)} \parallel h_{imp}^{(m)}) - D(g_{gen}^{(m)} \parallel g_{imp}^{(m)})\) in (9) is the KL-divergence loss at the \( m \)-th modality caused by the modality selection attack.

Figure 2 shows \( h_{gen}^{(m)}, h_{imp}^{(m)}, \alpha_m \) and \( \beta_m \). Consider a unimodal biometric system that verifies a user by comparing an LLR \( y^{(m)} \) with a threshold \( A = 0 \) (if \( y^{(m)} > 0 \), accept; otherwise, reject). Then, \( \alpha_m \) and \( \beta_m \) are FRR and FAR of the system, respectively. \( \gamma_m \) is a positive value that is generally larger than \( \alpha_m \) and \( \beta_m \), since \( y^{(m)} \) in (12) can take a very large negative value. However, if \( \alpha_m \) and \( \beta_m \) are very close to 0 (i.e., if the biometric system uses a very accurate modality such as iris and finger-vein), \( \gamma_m \) is also close to 0. This is because \( h_{gen}^{(m)} \) in (12) takes a very small value in this case. If \( \alpha_m, \beta_m \), and \( \gamma_m \) are close to 0, the KL-divergence loss at the \( m \)-th modality is also close to 0 (since \( \lim_{x \rightarrow 0} x \log((1 - x)/x) - x = 0 \)). Therefore, if the \( m \)-th modality is accurate (resp. inaccurate), the KL-divergence loss at the \( m \)-th modality is small (resp. large).

### 3. Theoretical Evaluation

In this section, we theoretically investigate whether the overall accuracy can be improved by restricting modality selection. In our theoretical analysis, we use the KL divergence for each modality (i.e., \( D(g_{gen}^{(m)} \parallel g_{imp}^{(m)}), D(g_{gen}^{(m)} \parallel g_{imp}^{(m)}), D(h_{gen}^{(m)} \parallel h_{imp}^{(m)}), \) and \( D(h_{gen}^{(m)} \parallel h_{imp}^{(m)}), \)) instead of the KL divergence for all modalities. This is because the KL divergence for all modalities is upper bounded by the summation of the KL divergence for each modality, as described in Sect. 2.2. Note that our theoretical results hold irrespective of whether scores \( x^{(1)}, \ldots, x^{(M)} \) are independent of each other (i.e., irrespective of the assumption (i) in Sect. 1.3). This is because our theoretical analysis focuses only on the \( m \)-th modality (i.e., we analyze the overall accuracy of the unimodal biometric system that uses the \( m \)-th modality, which is irrelevant to the other modalities).

We first describe the additional assumptions made (along with the assumption (ii) in Sect. 1.3) in our theoretical analysis (Sect. 3.1). We then compare the KL divergence of SLR with that of SLR (Sect. 3.2), and the KL divergence of NLR with that of NLR (Sect. 3.3). We finally discuss our theoretical results (Sect. 3.4).
3.1 Assumptions

In our analysis, we focus on the case when a genuine user attempts verification against herself for once (i.e., she inputs only one query sample for each modality) for simplicity. Then we make the following assumption (along with the assumption (ii) in Sect. 1.3):

(iii) The genuine user successfully eliminates her templates whose LLRs are less than (or equal to) 0 for her query samples in SLR*.

In other words, we assume that the genuine user successfully eliminates her “goat” templates. Note that this assumption may not hold in practice, since it can happen that the genuine user fails to exclude templates whose LLRs are less than (or equal to) 0 (i.e., false negative errors). It can also happen that the genuine user incorrectly excludes templates whose LLRs are greater than 0 (i.e., false positive errors). In other words, there are goat detection errors in practice. We discuss the effect of the goat detection errors on our theoretical analysis in Sect. 3.4.

We also make the following assumption:

(iv) Template’s genuine scores and impostor scores are independent of each other.

Strictly speaking, this assumption may also not hold in practice. Yager and Dunstone [31] introduced four user groups (worms, doves, chameleons, and phantoms) according to the relationship between template’s genuine scores and impostor scores. For example, “worms” are the worst conceivable users who show low similarity scores against themselves and high similarity scores against many others. Yager and Dunstone examined the existence of the four user groups using various datasets, and showed that although such user groups did not exist in many datasets, they existed in some datasets (e.g., they showed a significant presence of worms in one system out of ten systems; see Table 3 in [31] for details). We also discuss the effect of such users on our analysis in Sect. 3.4.

3.2 Selective LR Fusion

We now investigate the effectiveness of the modality restriction in SLR. Under the assumptions (ii), (iii), and (iv), we prove the following theorem:

**Theorem 1 (The difference of the KL divergence between SLR* and SLR):**

\[
D(g_{gen}^{(m)} \| g_{imp}^{(m)}) - D(g_{gen}^{(m)} \| g_{imp}^{(m)}) = \alpha_m + O(\alpha_m^2). \quad (13)
\]

**Proof.** In SLR, both genuine users and impostors input query samples whose LLRs are greater than 0, and LLRs corresponding to missing query samples are set to 0. Since \(\alpha_m\) and \(1 - \beta_m\) are the summation of \(h_{gen}^{(m)}\) and \(h_{imp}^{(m)}\) over the area where \(y^{(m)} \leq 0\), respectively (see Fig. 2), \(g_{gen}^{(m)}\) and \(g_{imp}^{(m)}\) are expressed as follows:

\[
g_{gen}^{(m)}(y^{(m)}) = \begin{cases} h_{gen}^{(m)}(y^{(m)}) & (\text{if } y^{(m)} > 0) \\ \alpha_m & (\text{if } y^{(m)} = 0) \\ 0 & (\text{if } y^{(m)} < 0) \end{cases} 
\]

\[
g_{imp}^{(m)}(y^{(m)}) = \begin{cases} h_{imp}^{(m)}(y^{(m)}) & (\text{if } y^{(m)} > 0) \\ 1 - \beta_m & (\text{if } y^{(m)} = 0) \\ 0 & (\text{if } y^{(m)} < 0). \end{cases} \quad (14)
\]

In SLR*, a genuine user successfully eliminates her templates whose LLRs are less than (or equal to) 0 (under the assumption (iii)). Since a genuine LLR of SLR* is not changed from that of SLR in this case, \(g_{gen}^{(m)}\) can be expressed as follows:

\[
g_{gen}^{(m)}(y^{(m)}) = g_{gen}^{(m)}(y^{(m)}) \quad (\text{for any } y^{(m)}). \quad (16)
\]

Since \(\alpha_m\) is the summation of \(h_{gen}^{(m)}\) over the area where \(y^{(m)} \leq 0\) (as shown in Fig. 2), a template is eliminated with probability \(\alpha_m\). Thus, even if the impostor inputs a query sample whose LLR is greater than 0, the LLR becomes 0 with probability \(\alpha_m\) (note that an impostor LLR is generated independently of a genuine LLR under the assumption (iv)). Therefore, if \(y^{(m)} > 0\),

\[
g_{imp}^{(m)}(0) = 1 - \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)})(1 - \alpha_m) \quad (17)
\]

\[
= 1 - \beta_m(1 - \alpha_m) \quad (18)
\]

\[
\text{(the last equality follows from (15)). Since } \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)}) \text{ is } 1, \text{ we have}
\]

\[
g_{imp}^{(m)}(0) = 1 - \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)}) = 1 - \beta_m(1 - \alpha_m) \quad (19)
\]

\[
= 1 - \beta_m + \alpha_m \beta_m \quad (20)
\]

\[
= 1 - \beta_m + \alpha_m \beta_m \quad (21)
\]

\[
= 1 - \beta_m(1 - \alpha_m) \quad (22)
\]

\[
\text{(from (19) to (20), we used (18); from (20) to (21), we used (11)). Thus, } g_{imp}^{(m)} \text{ can be expressed as follows:}
\]

\[
g_{imp}^{(m)}(y^{(m)}) = \begin{cases} h_{imp}^{(m)}(y^{(m)})(1 - \alpha_m) & (\text{if } y^{(m)} > 0) \\ 1 - \beta_m + \alpha_m \beta_m & (\text{if } y^{(m)} = 0) \\ 0 & (\text{if } y^{(m)} < 0) \end{cases} \quad (23)
\]

By (7), (14), (15), (16), and (23), we have

\[
D(g_{gen}^{(m)} \| g_{imp}^{(m)}) - D(g_{gen}^{(m)} \| g_{imp}^{(m)}) = \sum_{y^{(m)} > 0} h_{gen}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m))}} \right) - \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m))}} \right) + \alpha_m \log \left( \frac{\alpha_m}{1 - \beta_m} \right) - \alpha_m \log \left( \frac{\alpha_m}{1 - \beta_m} \right) \quad (24)
\]

\[
= \sum_{y^{(m)} > 0} h_{gen}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m))}} \right) - \sum_{y^{(m)} > 0} h_{imp}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m))}} \right) + \alpha_m \log \left( \frac{\alpha_m}{1 - \beta_m} \right) - \alpha_m \log \left( \frac{\alpha_m}{1 - \beta_m} \right) \quad (25)
\]
\[-\alpha_m \log \left( \frac{1-\beta_m + \alpha_m \beta_m}{\alpha_m \beta_m} \right) \]
\[
= -(1 - \alpha_m) \log(1 - \alpha_m) - \alpha_m \log \left( 1 + \frac{\alpha_m \beta_m}{1 - \beta_m} \right)
\]  
(26)
(27)

(from 26) to (27), we used (10).

Here we use the Taylor series expansion for \( (1 - x) \log(1 - x) \) and \( \log(1 + x) \), where \( x \ll 1 \) (recall that \( \alpha_m \) and \( \beta_m \) can be regarded as FRR and FAR of the \( m \)-th modality, respectively, and are much smaller than 1 in general). Then, (27) can be expressed as follows:

\[-(1 - \alpha_m) \log(1 - \alpha_m) - \alpha_m \log \left( 1 + \frac{\alpha_m \beta_m}{1 - \beta_m} \right)\]
\[
= \alpha_m + O(\alpha_m^2)
\]
\[
= \alpha_m + O(\alpha_m^2),
\]
(30)
which proves (13). \( \square \)

Note that \( \alpha_m \) can be regarded as FRR (as described in Sect. 2.3), and is much smaller than 1 in general. Thus, Theorem 1 means that although \( D(\gamma_{gen}(m) || \gamma_{imp}(m)) \) is greater than \( D(\gamma_{gen}(m) || \gamma_{gen}(m)) \), they are almost the same (i.e., \( D(\gamma_{gen}(m) || \gamma_{gen}(m)) \approx D(\gamma_{gen}(m) || \gamma_{gen}(m)) \)). In other words, the modality restriction hardly increases the KL divergence in SLR.

3.3 Non-selective LR Fusion

We then investigate the effectiveness of the modality restriction in NLR. Under the assumptions (ii), (iii), and (iv), we prove the following theorem:

Theorem 2 (The difference of the KL divergence between NLR* and NLR):

\[ D(\gamma_{gen} || \gamma_{gen}) - D(\gamma_{gen} || \gamma_{imp}) = \gamma_m + O(\alpha_m). \]  
(31)

Proof. In NLR*, a genuine user successfully eliminates her templates whose LLRs are less than (or equal to) 0 (under the assumption (iii)). By eliminating these templates, a genuine LLR becomes always equal to that of SLR. Thus, \( h_{gen}^{(m)} \) can be expressed as follows:

\[ h_{gen}^{(m)}(y^{(m)}) = g_{gen}^{(m)}(y^{(m)}) \quad \text{(for any } y^{(m)}). \]  
(32)

Recall that a template is eliminated with probability \( \alpha_m \). Thus, even if the impostor inputs a query sample whose LLR is not 0, the LLR becomes 0 with probability \( \alpha_m \) (under the assumption (iv)). Therefore, if \( y^{(m)} \neq 0 \),

\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(y^{(m)})(1 - \alpha_m). \]  
(33)

Since \( \sum_{y^{(m)}} h_{imp}^{(m)}(y^{(m)}) = 1 \) and \( \sum_{y^{(m)}} h_{imp}^{(m)}(y^{(m)}) = 1 \), we have

\[ h_{imp}^{(m)}(0) = 1 - \sum_{y^{(m)} \neq 0} h_{imp}^{(m)}(y^{(m)}) \]
\[
= 1 - \sum_{y^{(m)} \neq 0} h_{imp}^{(m)}(y^{(m)})(1 - \alpha_m)
\]
\[
= 1 - (1 - h_{imp}^{(m)}(0))(1 - \alpha_m)
\]  
(34)
(35)
(36)

(37)

(from 34 to 35), we used (33). Thus, \( h_{imp}^{(m)} \) can be expressed as follows:

\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(0)(1 - \alpha_m) + \alpha_m \]
\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(0)(1 - \alpha_m) + \alpha_m \]
\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(0)(1 - \alpha_m) + \alpha_m \]
\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(0)(1 - \alpha_m) + \alpha_m \]
\[ h_{imp}^{(m)}(y^{(m)}) = h_{imp}^{(m)}(0)(1 - \alpha_m) + \alpha_m \]

(38)

By (7), (14), (32), and (38), we have

\[ D(h_{gen} || h_{gen}) - D(h_{gen} || h_{imp}) \]
\[ = \sum_{y^{(m)} > 0} h_{gen}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m)})} \right) \]
\[ + \alpha_m \log \left( \frac{\alpha_m}{\sum_{y^{(m)} > 0}(1 - \alpha_m) + \alpha_m} \right) \]
\[ - \sum_{y^{(m)} < 0} h_{gen}^{(m)}(y^{(m)}) \log \left( \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m)})} \right). \]  
(40)

Since the first term in (40) is the same as the first term in (26), it can be expressed as \( \alpha_m \) (see (30)). Regarding the third term in (40), Murakami et al. [13] proved that the following equality holds for a genuine distribution \( h_{gen} \) and an impostor distribution \( h_{imp} \) of LLR \( y^{(m)} \):

\[ y^{(m)} = \log \frac{h_{gen}^{(m)}(y^{(m)})}{h_{imp}^{(m)}(y^{(m)})} \]  
(41)

(see Appendix of [13] for details). By (12) and (41), the third term in (40) can be expressed as \( \gamma_m \).

Regarding the second term in (40), we have

\[ \alpha_m \log \left( \frac{\alpha_m}{\sum_{y^{(m)} > 0}(1 - \alpha_m) + \alpha_m} \right) < \alpha_m \log \left( \frac{\alpha_m}{\alpha_m} \right) = 0. \]  
(42)

By (10) and (41), \( h_{imp}^{(m)}(0) = h_{gen}^{(m)}(0) \leq \sum_{y^{(m)} < 0} h_{gen}^{(m)}(y^{(m)}) = \gamma_m \).

Thus, we also have

\[ \alpha_m \log \left( \frac{\alpha_m}{\sum_{y^{(m)} > 0}(1 - \alpha_m) + \alpha_m} \right) \geq \alpha_m \log \left( \frac{\alpha_m}{\alpha_m(1 - \alpha_m) + \alpha_m} \right) \]
\[ = -\alpha_m \log(2 - \alpha_m) \]
\[ \geq -\alpha_m \log 2. \]  
(43)
(44)
(45)

It follows from (42) and (45) that the second term in (40) is a negative value that is greater than or equal to \( -\alpha_m \log 2 \).

Thus, the second term in (40) can be expressed as \( O(\alpha_m) \), which proves (31). \( \square \)

Note that \( \gamma_m \) is a positive value, as described in Sect. 2.3. Thus, Theorem 2 means that the modality restriction can increase the KL divergence in NLR. Figure 3 shows the relationship of the KL divergence among four types of the LR fusion schemes (i.e., SLR, SLR*, NLR, and NLR*).

3.4 Discussions

Discussions on our theorems. We first consider the reason that the modality restriction hardly increases the KL divergence in SLR (i.e., \( D(\gamma_{gen} || \gamma_{imp}) \approx D(\gamma_{gen} || \gamma_{imp}) \)).
In SLR*, a genuine user successfully eliminates templates whose LLRs are less than (or equal to) 0 under the assumption (iii). Since the total LLR is not changed in this case, FRR of SLR* is the same as that of SLR. On the other hand, since an impostor, whose LLR is greater than 0 for the eliminated template, cannot input the corresponding query sample (i.e., she cannot take an optimal strategy), FAR is decreased. However, it rarely happens that a template shows a negative LLR for a genuine user, while showing a positive LLR for an impostor at the same time under the assumption (iv). In other words, “worms” [31] (described in Sect. 3.1) are very rare under the assumption (iv). In this case, FAR of SLR* is almost the same as that of SLR. We consider this is the reason that although $D(h_{\text{gen}}(m) \mid h_{\text{imp}}^*) > D(h_{\text{en}}(m) \mid h_{\text{imp}}^*)$, the genuine user can still significantly decrease FAR, and therefore increase the KL divergence. In our experiments in Sect. 4, we show that this is indeed the case.

We then consider the reason that the modality restriction increases the KL divergence in NLR (i.e., $D(h_{\text{gen}}(m) \mid h_{\text{imp}}^*) < D(h_{\text{en}}(m) \mid h_{\text{imp}}^*)$). In NLR, a genuine user has to input all query samples, including the ones whose LLRs are less than 0. On the other hand, in NLR*, the genuine user does not have to input the query samples whose LLRs are less than 0, since she has successfully eliminated the corresponding templates (under the assumption (iii)). Thus, FRR of NLR* is much smaller than that of NLR. We consider this is the reason that $D(h_{\text{gen}}(m) \mid h_{\text{imp}}^*)$ is greater than $D(h_{\text{en}}(m) \mid h_{\text{imp}}^*)$. One might still wonder why NLR does not provide the best performance even as it uses all scores as information sources. The reason for this is that NLR* uses the whereabouts of goat templates in the database as an additional source of information (while NLR does not use it).

Discussions on our assumptions. We also discuss the relationship between our theoretical results and the assumptions (i), (ii), (iii), and (iv). We first discuss the assumptions (i) and (ii); the LR fusion scheme that we focus on assumes that scores $x^{(1)}, \ldots, x^{(M)}$ are independent of each other, and that a genuine (resp. impostor) score $x^{(m)}$ is generated from $f_{\text{gen}}^{(m)}$ (resp. $f_{\text{imp}}^{(m)}$). In reality, scores $x^{(1)}, \ldots, x^{(M)}$ can be correlated with each other, especially in multiple instances (e.g., left fingerprint and right fingerprint; left iris and right iris). For example, the image quality of each fingerprint can be correlated with each other, and causes the score correlation. It is also shown in [32] that the left iris score is correlated with the right iris score. We also show in Sect. 4 that scores of multiple instances are correlated with each other.

If multiple instances are highly correlated with each other, the overall accuracy (i.e., the KL divergence) is not much improved by fusing these instances. However, we emphasize again that our theoretical results hold for each instance, irrespective of whether scores $x^{(1)}, \ldots, x^{(M)}$ are independent of each other (i.e., irrespective of the assumption (i)), as described in the beginning of Sect. 3. Since the relationship of the KL divergence in Fig. 3 holds for the LLR of each modality/instance, we can expect that it also holds for the integrated LLR. We show that this is indeed the case in our experiments in Sect. 4.

We then discuss the assumption (iii); a genuine user successfully eliminates goat templates, whose LLRs are less than (or equal to) 0. In reality, there are goat detection errors. If the genuine user incorrectly eliminates templates whose LLRs are greater than 0 (i.e., false positive errors), the genuine user cannot input the corresponding query samples to maximize the total LLR (i.e., she cannot take an optimal strategy). Thus, it is reasonable to consider that the goat detection errors increase FRR, and therefore decrease the KL divergence. In our experiments in Sect. 4, we show that this is indeed the case.

We finally discuss the assumption (iv); template’s genuine scores and impostor scores are independent of each other. Strictly speaking, this assumption may also not hold, as described in Sect. 3.1. For example, suppose that there is a negative correlation between the genuine scores and the impostor scores, and there are many “worm” templates, which show negative LLRs for genuine users and positive LLRs for impostors. Then, in SLR, the modality restriction can significantly decrease FAR, and therefore increase the KL divergence (i.e., $D(h_{\text{gen}}(m) \mid g_{\text{imp}}^*) < D(h_{\text{en}}(m) \mid g_{\text{imp}}^*)$). However, it is reasonable to consider that there are few “worm” templates in many practical systems (Yager and Dunstone [31] showed a significant presence of worms in only one system out of ten systems). In our experiments in Sect. 4, we also show that there is a very low correlation between genuine LLRs and impostor LLRs (i.e., (iv) is appropriate), and there are few “worm” templates.

How to counteract the modality selection attack. Our theoretical results showed that the KL divergence loss caused by the modality selection attack cannot be preventing by the modality restriction. Then, the following question arises: is it impossible to counteract the modality selection attack in the selective LR fusion scheme?

We consider it is possible to counteract the modality selection attack. One way to do so is to use only modalities with high accuracy in the selective LR fusion scheme. As described in Sect. 2.3, if the $m$-th modality is accurate, the KL-divergence loss at the $m$-th modality is close to 0 (i.e., $D(h_{\text{gen}}(m) \mid g_{\text{imp}}^*) \approx D(h_{\text{en}}(m) \mid h_{\text{imp}}^*)$). In addition, since $\alpha_m$, $\beta_m$, and $\gamma_m$ are also close to 0 in this case, it follows from Theorem 1 and Theorem 2 that $D(h_{\text{gen}}(m) \mid g_{\text{imp}}^*) \approx D(h_{\text{en}}(m) \mid h_{\text{imp}}^*)$. In other words, all of SLR, SLR*, NLR, and NLR* provide almost the same overall accuracy in this case. In Sect. refsec:exp, we show that this is true when we use left iris
and right iris (CASIA-Iris-Thousand dataset [21]), both of which are accurate.

4. Experimental Evaluation

We performed experiments to validate our theoretical results. We describe the experimental set-up (Sect. ref{setup}), and report the experimental results (Sect. 4.2). We then evaluate the effect of goat detection errors on the accuracy in SLR* and NLR* (Sect. 4.3). We finally show that an impostor can perform the modality selection attack without any background knowledge about the target user’s biometrics (Sect. 4.4).

4.1 Experimental Set-Up

We performed experiments using three datasets: the NIST BSSR1 Set1 dataset [19], the Biosecure DS2 dataset [20], and the CASIA-Iris-Thousand dataset [21]. In the following, we described the details of each dataset:

**NIST BSSR1 Set1 [19]:** The NIST BSSR1 Set1 dataset contains face scores, left fingerprint scores, and right fingerprint scores \((M = 3)\) from 517 subjects. Although there were face scores from two algorithms (“C” and “G”) in this dataset, we used the face scores from “C” and eliminated one subject whose scores were inappropriate (the score values were \(-1\)). Thus, we used \(3 \times 516 \times 516\) scores in total. We randomly divided the 516 subjects into 116 training subjects and 400 testing subjects.

**Biosecure DS2 [20]:** The Biosecure DS2 dataset (cost-sensitive evaluation set) contains scores of face, six fingerprints (left/right thumb, left/right index, and left/right middle fingers), and left iris from 207 subjects. Note that the accuracy of left iris is very low, due to bad iris segmentation, as described in [20]. We used scores of face, left index fingerprint, right index fingerprint, and left iris \((M = 4)\). Although each subject contributed four biometric samples per each modality, we used scores between the first biometric samples and the second biometric samples. Thus, we used \(4 \times 207 \times 207\) scores in total. We randomly divided the 207 subjects into 57 training subjects and 150 testing subjects.

**CASIA-Iris-Thousand [21]:** The CASIA-Iris-Thousand dataset contains left iris images and right iris images \((M = 2)\) from 1000 subjects. We extracted scores from them using the VeriEye SDK 2.8 [33], which is known as the state-of-the-art commercial iris matcher. Due to this sophisticated matcher, the accuracy of these irises is very high (as opposed to Biosecure DS2), as will be shown later. Although each subject contributed ten biometric samples per each iris, we used scores between the first biometric samples and the second biometric samples. Thus, we used \(2 \times 1000 \times 1000\) scores in total. We randomly divided the 1000 subjects into 100 training subjects and 900 testing subjects.

We attempted 100 cases to randomly divide the subjects into training subjects and testing subjects, and performed, for each case, the following experiments.

In our experiments, we evaluated four types of the LR fusion schemes (i.e., SLR, SLR*, NLR, and NLR*). In SLR, we assumed that both genuine users and impostors input only query samples whose LLRs are greater than 0 (i.e., both of them take an optimal strategy). In SLR* and NLR*, we assumed that the genuine users successfully eliminate templates whose LLRs are less than (or equal to) 0 for their own query samples (i.e., the assumption (iii) in Sect. 3.1 holds).

Using the training subjects, we trained a model to estimate an LLR \(y^{(m)}\) from a score \(x^{(m)}\). In NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand, we used 116, 57, and 100 genuine scores, and 116 × 115, 57 × 56, and 100 × 99 impostor scores for each modality, respectively. In NIST BSSR1 Set1 and Biosecure DS2, we assumed the logistic regression model [22], and trained regression coefficients using the bias-reduced maximum likelihood estimation method [34]. In CASIA-Iris-Thousand, we assumed the Gaussian distribution model, and trained parameters using the maximum likelihood estimation method (we did not use logistic regression in this dataset, because the variance of genuine scores is too large and regression coefficients were not accurately estimated).

Using the testing subjects, we evaluated the accuracy of four types of the LR fusion schemes. In NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand, there were 400, 150, and 900 genuine attempts, and 400 × 399, 150 × 149, and 900 × 899 impostor attempts, respectively. As measures of the accuracy, we used FAR, FRR, the DET (Detection Error Tradeoff) curve (i.e., FAR-FRR curve) [35] and the KL divergence between a genuine distribution of the total LLR \(z^{(m)}\) and impostor’s one. To estimate the KL divergence, we used the generalized k-NN (Nearest Neighbor) estimator [30]. We averaged FAR, FRR, and the KL divergence over the 100 cases to randomly divide the subjects into training subjects and testing subjects to stabilize the performance.

4.2 Experimental Results

We first examined EER, \(\alpha_m\), \(\beta_m\), and \(\gamma_m\) for each modality/instance. We also computed the KL-divergence loss in (9) using \(\alpha_m\), \(\beta_m\), and \(\gamma_m\). Table 2 shows the results (we averaged these values over the 100 cases to randomly divide the subjects). It can be seen that EER, \(\alpha_m\), and \(\beta_m\) are very large in the left iris of Biosecure DS2 (due to bad iris segmentation, as described in [20]). On the other hand, they are very small in CASIA-Iris-Thousand, since we used the VeriEye SDK 2.8 [33] (state-of-the-art commercial iris matcher). It can also be seen that when \(\alpha_m\) and \(\beta_m\) are small (resp. large), \(\gamma_m\) and the KL-divergence loss are also small (resp. large), as described in Sect. 2.3.

We second examined the correlation coefficients of scores between two modalities/instances using all scores
Table 2  EER, $\alpha_m$, $\beta_m$, $\gamma_m$, and the KL-divergence loss in (9) for each modality/instance (FA: face, LF: left fingerprint, RF: right fingerprint, LI: left iris, RI: right iris).

(i) NIST BSSR1 Set1

<table>
<thead>
<tr>
<th>Modality</th>
<th>EER[%]</th>
<th>$\alpha_m$</th>
<th>$\beta_m$</th>
<th>$\gamma_m$</th>
<th>KL-divergence loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>4.5</td>
<td>0.062</td>
<td>0.029</td>
<td>0.11</td>
<td>0.090</td>
</tr>
<tr>
<td>LF</td>
<td>8.4</td>
<td>0.13</td>
<td>0.0079</td>
<td>0.20</td>
<td>0.090</td>
</tr>
<tr>
<td>RF</td>
<td>5.0</td>
<td>0.083</td>
<td>0.0031</td>
<td>0.15</td>
<td>0.086</td>
</tr>
</tbody>
</table>

(ii) Biosecure DS2

<table>
<thead>
<tr>
<th>Modality</th>
<th>EER[%]</th>
<th>$\alpha_m$</th>
<th>$\beta_m$</th>
<th>$\gamma_m$</th>
<th>KL-divergence loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>9.5</td>
<td>0.13</td>
<td>0.034</td>
<td>0.25</td>
<td>0.015</td>
</tr>
<tr>
<td>LF</td>
<td>7.8</td>
<td>0.095</td>
<td>0.047</td>
<td>0.18</td>
<td>0.057</td>
</tr>
<tr>
<td>RF</td>
<td>6.8</td>
<td>0.082</td>
<td>0.045</td>
<td>0.16</td>
<td>0.051</td>
</tr>
<tr>
<td>LI</td>
<td>14.0</td>
<td>0.18</td>
<td>0.073</td>
<td>0.26</td>
<td>0.048</td>
</tr>
</tbody>
</table>

(iii) CASIA-Iris-Thousand

<table>
<thead>
<tr>
<th>Modality</th>
<th>EER[%]</th>
<th>$\alpha_m$</th>
<th>$\beta_m$</th>
<th>$\gamma_m$</th>
<th>KL-divergence loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI</td>
<td>1.5</td>
<td>0.015</td>
<td>0.020</td>
<td>0.068</td>
<td>0.0087</td>
</tr>
<tr>
<td>RI</td>
<td>2.0</td>
<td>0.020</td>
<td>0.020</td>
<td>0.092</td>
<td>0.020</td>
</tr>
</tbody>
</table>

Table 3  Correlation coefficients of scores between two modalities/instances (FA: face, LF: left fingerprint, RF: right fingerprint, LI: left iris, RI: right iris).

(i) NIST BSSR1 Set1

<table>
<thead>
<tr>
<th>Modality</th>
<th>F-LF</th>
<th>F-RF</th>
<th>LF-RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genuine</td>
<td>-0.12</td>
<td>-0.015</td>
<td>0.41</td>
</tr>
<tr>
<td>Impostor</td>
<td>0.043</td>
<td>0.047</td>
<td>0.16</td>
</tr>
</tbody>
</table>

(ii) Biosecure DS2

<table>
<thead>
<tr>
<th>Modality</th>
<th>F-LF</th>
<th>F-RF</th>
<th>F-LI</th>
<th>LF-RF</th>
<th>LF-LI</th>
<th>RL-LI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genuine</td>
<td>0.095</td>
<td>-0.15</td>
<td>0.095</td>
<td>0.50</td>
<td>0.028</td>
<td>0.083</td>
</tr>
<tr>
<td>Impostor</td>
<td>-0.013</td>
<td>-0.017</td>
<td>0.074</td>
<td>0.13</td>
<td>-0.025</td>
<td>-0.022</td>
</tr>
</tbody>
</table>

(iii) CASIA-Iris-Thousand

<table>
<thead>
<tr>
<th>Modality</th>
<th>LI-RF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genuine</td>
<td>0.63</td>
</tr>
<tr>
<td>Impostor</td>
<td>0.041</td>
</tr>
</tbody>
</table>

(i.e., in NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand, we used 516, 207, 1000 genuine scores and 516×515, 207×206, 1000×999 impostor scores for each modality/instance). Table 3 shows the results. It can be seen that scores are correlated especially in multiple instances (i.e., LF-RF in NIST BSSR1 Set1 and Biosecure DS2, and LI-RI in CASIA-Iris-Thousand), as discussed in 3.4. This indicates that the overall accuracy is not much improved by fusing these instances. However, we emphasize again that our theoretical results hold, irrespective of whether scores are independent of each other (i.e., irrespective of the assumption (i)), as will be shown in the following.

We third evaluated FAR and FRR. Figure 4 shows the relationship between the threshold $A$ and FAR/FRR. Since some curves overlap, we also show some concrete values of FAR and FRR in Table 4. It can be seen from Fig. 4 and Table 4 that FAR of SLR is about 20 to 100 times larger than that of NLR at the same threshold in NIST BSSR1 Set1 and Biosecure DS2. This means that the modality selection attack significantly increases FAR in these datasets. On the other hand, FAR of SLR is not much different from that of NLR in CASIA-Iris-Thousand. This is because $\beta_m$ (i.e., FAR of the unimodal system that verifies a user by comparing an LLR $y^{(m)}$ with a threshold $A = 0$) is very small for both irises in this dataset, as shown in Table 2. In other words, when we use only modalities with high accuracy, the impact of the modality selection attack is very small, as discussed in Sect. 3.4.

It can also be seen from Fig. 4 and Table 4 that although FAR of SLR* is smaller than that of SLR, they are almost the same, as discussed in Sect. 3.4. This means that the impact of the modality selection attack cannot be mitigated by the modality restriction. It can also be seen from Table 4 that SLR, SLR*, and NLR* achieve completely the same FRR, which is smaller than that of NLR. This is because the genuine users input only query samples whose LLRs are greater than 0 (i.e., they take an optimal strategy) in SLR, SLR*, and NLR*.

We also evaluated the DET curve and the KL divergence. Figure 5 and Table 5 show the results. In Table 5, we also show password entropy of user-chosen PINs that is estimated in [16] (since it is sometimes beneficial to compare the discriminative power of biometrics with that of other authentication factors, as described in Sect. 2.2). It can be seen that in NIST BSSR1 Set1 and Biosecure DS2, SLR* and SLR achieve almost the same overall accuracy, NLR outperforms them, and NLR* outperforms NLR, as shown in Fig. 3. This means that the modality restriction hardly improves the overall accuracy in SLR, while it can improve the overall accuracy in NLR. On the other hand, it can be
Table 4  Concrete values of FAR/FRR in Fig. 4 (A = 2, 4, 6, or 8).

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 2</th>
<th>A = 4</th>
<th>A = 6</th>
<th>A = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>5.01 × 10⁻⁴</td>
<td>7.03 × 10⁻²</td>
<td>1.12 × 10⁻²</td>
<td>1.97 × 10⁻³</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>4.90 × 10⁻¹</td>
<td>6.81 × 10⁻²</td>
<td>1.08 × 10⁻²</td>
<td>1.94 × 10⁻³</td>
</tr>
<tr>
<td>NLR</td>
<td>1.58 × 10⁻²</td>
<td>2.94 × 10⁻³</td>
<td>4.51 × 10⁻⁴</td>
<td>4.39 × 10⁻⁵</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>2.64 × 10⁻²</td>
<td>3.87 × 10⁻³</td>
<td>4.89 × 10⁻⁴</td>
<td>4.39 × 10⁻⁵</td>
</tr>
</tbody>
</table>

(ii) NIST BSSR1 Set1, FRR[%]

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 2</th>
<th>A = 4</th>
<th>A = 6</th>
<th>A = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>5.55 × 10⁻⁴</td>
<td>1.26</td>
<td>2.20</td>
<td>4.11</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>5.55 × 10⁻¹</td>
<td>1.26</td>
<td>2.20</td>
<td>4.11</td>
</tr>
<tr>
<td>NLR</td>
<td>1.61</td>
<td>2.59</td>
<td>4.16</td>
<td>6.76</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>5.55 × 10⁻¹</td>
<td>1.26</td>
<td>2.20</td>
<td>4.11</td>
</tr>
</tbody>
</table>

(iii) Biosecure DS2, FAR[%]

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 2</th>
<th>A = 4</th>
<th>A = 6</th>
<th>A = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>2.67</td>
<td>4.24 × 10⁻³</td>
<td>8.98 × 10⁻³</td>
<td>1.96 × 10⁻³</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>2.28</td>
<td>3.40 × 10⁻¹</td>
<td>6.61 × 10⁻²</td>
<td>1.31 × 10⁻²</td>
</tr>
<tr>
<td>NLR</td>
<td>4.44 × 10⁻²</td>
<td>8.01 × 10⁻³</td>
<td>2.86 × 10⁻³</td>
<td>2.24 × 10⁻⁴</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>6.94 × 10⁻²</td>
<td>1.08 × 10⁻²</td>
<td>2.64 × 10⁻³</td>
<td>1.34 × 10⁻⁴</td>
</tr>
</tbody>
</table>

(iv) Biosecure DS2, FRR[%]

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 2</th>
<th>A = 4</th>
<th>A = 6</th>
<th>A = 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>0</td>
<td>3.67 × 10⁻¹</td>
<td>2.27</td>
<td>5.02</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>0</td>
<td>3.67 × 10⁻¹</td>
<td>2.27</td>
<td>5.02</td>
</tr>
<tr>
<td>NLR</td>
<td>1.86</td>
<td>3.28</td>
<td>5.04</td>
<td>8.39</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>0</td>
<td>3.67 × 10⁻¹</td>
<td>2.27</td>
<td>5.02</td>
</tr>
</tbody>
</table>

(v) CASIA-Iris-Thousand, FAR[%]

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 20</th>
<th>A = 40</th>
<th>A = 60</th>
<th>A = 80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>2.08 × 10⁻¹</td>
<td>1.86 × 10⁻²</td>
<td>3.18 × 10⁻³</td>
<td>6.19 × 10⁻⁴</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>2.06 × 10⁻¹</td>
<td>1.83 × 10⁻²</td>
<td>3.08 × 10⁻³</td>
<td>6.14 × 10⁻⁴</td>
</tr>
<tr>
<td>NLR</td>
<td>1.17 × 10⁻¹</td>
<td>1.23 × 10⁻²</td>
<td>2.35 × 10⁻³</td>
<td>4.68 × 10⁻⁴</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>1.17 × 10⁻¹</td>
<td>1.24 × 10⁻²</td>
<td>2.26 × 10⁻³</td>
<td>5.04 × 10⁻⁴</td>
</tr>
</tbody>
</table>

(vi) CASIA-Iris-Thousand, FRR[%]

<table>
<thead>
<tr>
<th>Scheme</th>
<th>A = 20</th>
<th>A = 40</th>
<th>A = 60</th>
<th>A = 80</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR</td>
<td>5.78 × 10⁻¹</td>
<td>7.84 × 10⁻¹</td>
<td>8.83 × 10⁻¹</td>
<td>9.92 × 10⁻¹</td>
</tr>
<tr>
<td>SLR⁺</td>
<td>5.78 × 10⁻¹</td>
<td>7.84 × 10⁻¹</td>
<td>8.83 × 10⁻¹</td>
<td>9.92 × 10⁻¹</td>
</tr>
<tr>
<td>NLR</td>
<td>5.78 × 10⁻¹</td>
<td>7.88 × 10⁻¹</td>
<td>8.86 × 10⁻¹</td>
<td>1.03</td>
</tr>
<tr>
<td>NLR⁺</td>
<td>5.78 × 10⁻¹</td>
<td>7.84 × 10⁻¹</td>
<td>8.83 × 10⁻¹</td>
<td>9.92 × 10⁻¹</td>
</tr>
</tbody>
</table>

seen that all of SLR, SLR⁺, NLR, and NLR⁺ provide almost the same overall accuracy in CASIA-Iris-Thousand. This is because the accuracy of both iris is very high in this dataset, as discussed in Sect. 3.4.

We finally examined the relationship between template’s genuine LLRs and impostor LLRs to validate the assumption (iv) in Sect. 3.1. Although we attempted 100 cases to randomly divide the subjects (as described in Sect. 4.1), we used the first 10 cases in this analysis. For each template, we randomly chose one impostor, and computed a pair of a genuine LLR and an impostor LLR. Then we plotted these pairs for each modality (there are 10 × 400, 10 × 150, and 10 × 1000 pairs in NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand, respectively). Figure 6 shows the results. It can be seen that the number of templates that show a negative LLR for a genuine user and a positive LLR for an impostor (i.e., “worm” templates [31]) is very small, as discussed in Sect. 3.4. We also show in Fig. 6 the correlation coefficient r between genuine LLRs and impostor LLRs for each modality/instance. Since all of the correlation coefficients are small (less than 0.15), we can say that the assumption (iv) in Sect. 3.1 is (not perfect but) appropriate in our experiments.

4.3 Effect of Goat Detection Errors on the Accuracy

In Sects. 4.1 and 4.2, we assumed that a genuine user successfully eliminates goat templates, whose LLRs are less than (or equal to) 0 for her own query samples, in SLR⁺ and NLR⁺ (i.e., the assumption (iii) in Sect. 3.1 holds). However, since there are goat detection errors in practice, we evaluated the effect of goat detection errors on the accuracy in SLR⁺ and NLR⁺. Note that we used only NIST BSSR1 Set1 and Biosecure DS2 in this analysis. This is because the impact of the modality selection attack was very small in CASIA-Iris-Thousand (we do not have to counteract this attack via the modality selection in this dataset).

As described in Sect. 3.1, there are two types of goat detection error rates: the false positive error and the false negative error. The false positive error is an error where the genuine user incorrectly excludes templates whose LLRs are greater than 0, while the false negative error is an error where the genuine user fails to exclude templates whose
Table 5  KL divergence and password entropy estimated in [16].

(i) NIST BSSR1 Set1

<table>
<thead>
<tr>
<th>Description</th>
<th>KL divergence of SLR</th>
<th>bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL divergence of SLR</td>
<td>16.747</td>
<td></td>
</tr>
<tr>
<td>KL divergence of SLR*</td>
<td>16.749</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR</td>
<td>17.031</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR*</td>
<td>17.274</td>
<td></td>
</tr>
</tbody>
</table>

(ii) Biosecure DS2

<table>
<thead>
<tr>
<th>Description</th>
<th>KL divergence of SLR</th>
<th>bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL divergence of SLR</td>
<td>13.725</td>
<td></td>
</tr>
<tr>
<td>KL divergence of SLR*</td>
<td>13.795</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR</td>
<td>14.195</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR*</td>
<td>14.609</td>
<td></td>
</tr>
</tbody>
</table>

(iii) CASIA-Iris-Thousand

<table>
<thead>
<tr>
<th>Description</th>
<th>KL divergence of SLR</th>
<th>bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>KL divergence of SLR</td>
<td>19.374</td>
<td></td>
</tr>
<tr>
<td>KL divergence of SLR*</td>
<td>19.374</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR</td>
<td>19.376</td>
<td></td>
</tr>
<tr>
<td>KL divergence of NLR*</td>
<td>19.435</td>
<td></td>
</tr>
</tbody>
</table>

(iv) Password entropy

<table>
<thead>
<tr>
<th>Description</th>
<th>bits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Password entropy of user-chosen 8-digit PINs</td>
<td>13</td>
</tr>
<tr>
<td>Password entropy of user-chosen 10-digit PINs</td>
<td>15</td>
</tr>
<tr>
<td>Password entropy of user-chosen 12-digit PINs</td>
<td>17</td>
</tr>
<tr>
<td>Password entropy of user-chosen 14-digit PINs</td>
<td>19</td>
</tr>
<tr>
<td>Password entropy of user-chosen 16-digit PINs</td>
<td>21</td>
</tr>
</tbody>
</table>

LLRs are less than (or equal to) 0. In this paper, we denote the false positive probability and the false negative probability by FPP and FNP, respectively. In our analysis, we set these values to an equal value \( \phi = FPP = FNP \) for simplicity. For the value of \( \phi \), we tested \( \phi = 0\% \), 2\%, 4\%, 6\%, 8\%, or 10\%. Recall that there are 400, 150, and 900 subjects for evaluation in NIST BSSR1 Set1, Biosecure DS2, and CASIA-Iris-Thousand, respectively, and each subject has provided one query sample and one template for each modality/instance. For each template, we randomly caused the goat detection error with probability \( \phi \). That is, if the template’s LLR was less than or equal to 0 for the query sample of the same subject, we excluded the template from the database with probability \( \phi \) (i.e., false positive error) and did not exclude it with probability \( 1 - \phi \). If the template’s LLR was greater than 0 for the query sample, we did not exclude the template with probability \( \phi \) (i.e., false negative error) and excluded it with probability \( 1 - \phi \). Under this setting, we evaluated the performance of SLR* and NLR* (the experimental setting except for the goat detection error probability \( \phi \) is the same as described in Sect. 4.1).

Figures 7 and 8 show FRR and the KL divergence, respectively (we averaged FRR and the KL divergence over the 100 cases to randomly divide the subjects). We do not show FAR, since FAR was hardly changed by increasing the goat detection error probability \( \phi \). (The reason for FAR being hardly changed is that the correlation coefficient \( r \) between genuine LLRs and impostor LLRs is very small, as shown in Fig. 6. In other words, the goat detection error is not much relevant to the increase/decrease of FAR, while it directly affects the increase of FRR.)

It can be seen that as the goat detection error probability \( \phi \) increases in SLR* and NLR*, FRR increases and therefore the KL divergence decreases, as discussed in Sect. 3.4. As a consequence, the KL divergence of SLR* is smaller than or almost equal to that of SLR, irrespective of the value of \( \phi \). This means that the modality restriction is not effective in SLR, irrespective of whether the assumption (iii) holds.
It can also be seen that when \( \phi = 10\% \), the KL divergence of \( \text{NLR}^* \) is smaller than that of \( \text{NLR} \) in both datasets. This means that the modality restriction is also not effective in \( \text{NLR} \), when there are many goat detection errors.

### 4.4 Modality Selection Attacks without Background Knowledge about the Target User’s Biometrics

We have so far assumed that an impostor can find query samples whose LLRs are more than 0 against a target user, which requires a lot of background knowledge about the target user’s biometrics. However, when the system uses modalities with high accuracy along with modalities with low accuracy, the impostor may perform the modality selection attack without such background knowledge, as described in Sect. 4.1. We finally show that this attack is indeed effective through the experiments. Here we use only NIST BSSR1 Set1 and Biosecure DS2 (in the same way as Sect. 4.3), since we have already seen that the impact of the modality selection attack is very small in CASIA-Iris-Thousand, even with the aid of the background knowledge.

We note that the impostor who has no background knowledge would input only one modality whose \( \beta_m \) is the largest. This is because an impostor LLR is most likely to be greater than 0 for such a modality. It can be seen from Table 2 that face (resp. left iris) provides the largest \( \beta_m \) in NIST BSSR1 Set1 (resp. Biosecure DS2). Therefore, the impostor who has no background knowledge about the target user’s biometrics would input only face (resp. left iris) in NIST BSSR1 Set1 (resp. Biosecure DS2).

Taking this into account, we evaluated the performance of SLR in the case when impostors input only face (resp. left iris) in NIST BSSR1 Set1 (resp. Biosecure DS2) (while genuine users input only query samples whose LLRs are greater than 0 in both the datasets). We denote SLR in this case by SLR (w/o BK) (BK is an abbreviation of background knowledge). We compared the performance of SLR (w/o BK) with that of SLR. Figure 9 and Table 6 show FAR, and Figure 8 shows the KL divergence (we do not show FRR, since FRR of SLR (w/o BK) is completely the same as that of SLR). It can be seen that FAR of SLR (w/o BK) is slightly lower than that of SLR, and therefore the KL divergence of SLR (w/o BK) is slightly higher than that of SLR. However, FAR (resp. the KL divergence) of SLR (w/o BK) is still larger (resp. smaller) than that of NLR (see Fig. 4,
Table 4, and Table 5). Thus, as a conclusion, we can say that the impostors can perform the modality selection attack without any background knowledge about the target user’s biometrics in these datasets.

5. Conclusion

In this paper, we investigated the impact of the modality selection attack in [12], [13] in details. Specifically, we divided the LR fusion scheme in two categories and two sub-categories (i.e., SLR, SLR*, NLR, and NLR*), and investigated whether the overall accuracy is improved by restricting modality selection. Our theoretical and experimental results can be summarized as follows:

- The modality restriction hardly increases the overall accuracy in SLR.
- The modality restriction increases the overall accuracy in NLR, when there are few goat detection errors.
- The impostor can perform the modality selection attack without any background knowledge about the target user’s biometrics, when the system uses modalities with high accuracy along with those with low accuracy.
- However, when the system uses only modalities with high accuracy, the impact of the modality selection attack is very small (it hardly increases FAR and hardly decreases the overall accuracy).

As a conclusion derived from these results, we discuss directions of the multibiometric system using the LR fusion scheme. We first consider that the modality selection mode is necessary in practice, since this mode is helpful for a user who is inherently difficult to recognize (i.e., goat) for some modalities, or a user who cannot appropriately input some query samples due to temporary physical conditions (e.g., injuries, illness). Furthermore, selective fusion allows a user to be authenticated with a small number of query samples (in the same way as sequential fusion [8], [27]), while non-selective fusion requires a user to input all query samples. From these perspectives, we can say that SLR is more convenient and realistic than NLR. However, the modality selection attack can cause the loss of the KL divergence. Unfortunately, this KL-divergence loss cannot be prevented by the modality restriction (while the KL divergence can further be increased by the modality restriction in NLR).

To prevent the modality selection attack, it is important to use only modalities with high accuracy. As shown in our experiments, if the system uses modalities with low accuracy along with modalities with low accuracy, the impostor can easily perform the modality selection attack by abusing modalities with low accuracy. On the other hand, if the system uses only modalities with high accuracy, the impact of the modality selection attack is very small (i.e., the KL-divergence loss is close to 0) as shown in our experiments.

It should also be noted that the goal of multibiometric fusion is to generally provide (not the optimal accuracy but) a satisfactory level of accuracy. Even if the system uses modalities with low accuracy and the impostor performs the modality selection attack, SLR has the possibility of satisfying the accuracy requirement. However, when we use the modalities with low accuracy in SLR, we should evaluate FAR and the overall accuracy of the system under the assumption that the impostor performs the modality selection attack (even without the background knowledge about the target user’s biometrics). We should also keep in mind that the overall accuracy cannot be improved by restricting modality selection.

References


Takao Murakami is a Researcher with the Information Technology Research Institute (ITRI), the National Institute of Advanced Industrial Science and Technology (AIST). He received the B.S., M.S., and Ph.D. degrees from the University of Tokyo in 2004, 2006, and 2014, respectively. He joined Hitachi, Ltd. in 2006, and AIST in 2014. His research interests include biometrics, privacy, and information security. He received the Yamashita SIG Research Award from the Information Processing Society of Japan (IPSJ) in 2011, the IEEE TrustCom’15 Best Paper Award in 2015, and the 15th DOCOMO Mobile Science Award (Advanced Technology Award of Excellence) in 2016. He is a member of IEICE, IPSJ and IEEE.

Yosuke Kaga is a Researcher with the Research & Development Group, Hitachi, Ltd. He received the B.S. and M.S. degrees from Hokkaido University in 2006 and 2008, respectively. He joined Hitachi, Ltd. in 2008. Since then, he has worked on research and development of biometric authentication systems and information security technologies. He is a member of IEICE.

Kenta Takahashi is a Senior Researcher with the Research & Development Group, Hitachi, Ltd. He is also a visiting associate professor with the Graduate School of Information Science and Technology, the University of Tokyo. He received the B.S., M.S., and Ph.D. degrees from the University of Tokyo in 1998, 2000, and 2012, respectively. He joined Hitachi, Ltd. in 2000. Since then, he has worked on research and development of biometric authentication systems and information security technologies. He received the Best Paper Award from the Information Processing Society of Japan (IPSJ) in 2009, the Nagao Special Researcher Award from IPSJ in 2015, and the 15th DOCOMO Mobile Science Award (Advanced Technology Award of Excellence) in 2016. He is a member of IEICE and IPSJ.