

Single Sensor-based Multi-quality Multi-modal Biometric Score Database and Its Performance Evaluation

Takuhiro Kimura, Yasushi Makihara, Daigo Muramatsu, Yasushi Yagi
Osaka University
8-1 Mihogaoka, Ibaraki, Osaka, 567-0047, Japan
{kimura, makihara, muramatsu, yagi}@am.sanken.osaka-u.ac.jp

Abstract

We constructed a large-scale multi-quality multi-modal biometric score database to advance studies on quality-dependent score-level fusion. In particular, we focused on single sensor-based multi-modal biometrics because of their advantages of simple system construction, low cost, and wide availability in real situations such as CCTV footage-based criminal investigation, unlike conventional individual sensor-based multi-modal biometrics that require multiple sensors. As for the modalities of multiple biometrics, we extracted gait, head, and the height biometrics from a single walking image sequence, and considered spatial resolution (SR) and temporal resolution (TR) as quality measures that simultaneously affect the scores of individual modalities. We then computed biometric scores of 1912 subjects under a total of 130 combinations of the quality measures, i.e., 13 SRs and 10 TRs, and constructed a very large-scale biometric score database composed of 1,814,488 genuine scores and 3,467,486,568 imposter scores. We finally provide performance evaluation results both for quality-independent and quality-dependent score-level fusion approaches using two protocols that will be beneficial to the score-level fusion research community.

1. Introduction

Multi-modal biometrics [1] have gained considerable attention because they enhance the accuracy of biometric person authentication as well as security against attack [2]. Combinations of such multi-modal biometrics are diverse: fingerprint and iris [3]; iris and face [4]; face and ocular [5]; face and gait [6, 7]; and face, gait, and the height [8].

An important process in multi-modal biometrics is to appropriately fuse individual modalities. The fusion of multiple modalities can be considered at different levels: sensor, feature, score, rank, and decision levels. Among the given fusion levels, score-level fusion has been studied exten-

sively [9–14] because of its wide application range, where multiple scores are fused into a single score by considering discrimination capabilities and/or joint score distributions of the individual modalities.

In addition to the score itself, we sometimes consider a sort of auxiliary information, or *quality measure* [15], that affects the score but does not identify the subject. A variety of these quality measures have been proposed up to now: the degree of occlusion and blur in iris recognition [16]; signal-to-noise ratio for speaker verification [17]; image size, or spatial resolution (SR), for image-based biometrics such as fingerprint, iris, and face; and frame-rate, or temporal resolution (TR), for video-based biometrics such as gait. For example, SR and TR affect the face and gait recognition score (i.e., lower SRs and TRs make the genuine and imposter score distributions more confusing) even though they do not contain any subject identity information.

The quality measure plays an important role in score-level fusion; the effectiveness of quality-dependent score-level fusion has been proven for various combinations of multiple biometrics [15, 18, 19]. Because multi-modal biometrics are often captured by individual sensors (e.g., for combined face and fingerprint biometrics, face and fingerprint images are captured by a camera and scanner, respectively), quality measures are usually associated with each modality's score.

Conversely, some studies capture multi-modal biometrics using a single sensor [6–8] (e.g., simultaneously capturing face and gait images from a single camera). Single sensor-based multi-modal biometrics are promising because they can be implemented using a low cost, simple system that widens the range of real applications (e.g., a forensic application to verify a perpetrator or suspect captured by CCTVs [20, 21]). Note that the quality measures generally affect all the multi-modal biometric scores in the single sensor-based approaches (e.g., SR affects both face and gait biometric scores), whereas they are associated only with each corresponding modality's score in the individual sensor-based approaches.

In this work, we therefore construct a large-scale multi-quality multi-modal biometric score database, called BS-MultiQ (The OU-ISIR Biometric Score Database, Set 5), to advance the state-of-the-art quality-dependent score-level fusion for promising single sensor-based multi-modal biometrics. More specifically, we consider single sensor-based multi-modal biometrics from gait, head, and the height as well as the SR and TR quality measures that significantly affect individually biometric scores, as reported in many research projects such as the DARPA Human identification at a distance program [22]. The contributions of this paper are two-fold.

1. Construction of single sensor-based multi-quality multi-modal biometric score database¹

We constructed a large-scale database composed of gait, head, and height biometric scores as well as two qualities, SR and TR, drawn from the OU-ISIR Gait Database, the Large Population dataset comprising 1912 subjects [23]. Whereas previously released multi-modal biometric score databases such as BANCA [24], BA-Fusion [25], Biosecure DS2 [26], and NIST BSSR1 [27] consider individual sensor-based approaches, our database focuses on single sensor-based multi-modal biometrics. Moreover, our database contains a very large number of scores because of the diversity of subjects and qualities, i.e., 1,814,488 genuine scores and 3,467,486,568 imposter scores, and hence it will serve the research community as a benchmark for score-level fusion approaches to multi-modal biometrics.

2. Performance evaluation

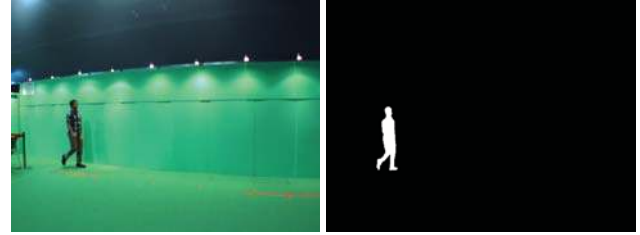
We provide two protocols for the constructed database and conducted a performance evaluation with a variety of score-level fusion approaches: transformation-based approaches [9], classification-based approaches [13], probability density-based approaches [10–12, 14], and some of their quality-dependent versions. The results of this performance evaluation will serve as a baseline for future research on quality-dependent score-level fusion.

2. Single Sensor-based Multi-quality Multi-modal Biometric Score Database

2.1. Overview

We drew walking image sequences from a publicly available gait database, i.e. the OU-ISIR Gait Database, the Large Population data set [23]. Each subject was asked to walk straight along a corridor twice (i.e., the gallery and probe, respectively) and each of the walking image sequences was captured by 640×480 pixels at 30 fps, as shown in Fig. 1(a). A silhouette sequence was then extracted by background subtraction-based graph-cut segmen-

¹The constructed database with detailed protocols (e.g., training/test sets divisions for cross-validation) is available at <http://www.am.sanken.osaka-u.ac.jp/BiometricDB/BioScore.html>.



(a) Original image

(b) Silhouette

Figure 1: Original walking image and extracted silhouette. Although the size of the silhouette region is dependent on the subject, its average is approximately 90×180 pixels.

tation [28], as shown in Fig. 1(b), and gait and height biometrics were extracted after lens distortion removal, rectification, and cropping.

Here, the rectification was done using camera calibration parameters such that the x and y coordinates of the image plane were parallel to the walking and vertical directions, respectively. This rectification process is beneficial for computing the height biometrics for two reasons: (1) we can easily obtain the vertical positions of the bottom of the foot and the top of the head from bounding box information, and (2) we can easily convert the subject’s height in the rectified image into their corresponding height in the real world based on the calibration parameters because the distance between the camera and the walking course is the same for all the subjects in this database.

We describe the details of the subsequent procedures to construct the single sensor-based multi-quality multi-modal biometric score database in the following subsections.

2.2. Matchers

In this subsection, we briefly describe matchers of individual gait, head, and the height biometrics. We refer the reader to [8] for more details.

Gait: We used the averaged silhouette [29], or gait energy image (GEI) [30], as the gait feature since the GEI is the most widely used feature in the gait recognition community. The GEI is a simple yet effective gait representation that has achieved state-of-the-art performance in an evaluation using a large-population data set of more than 4000 subjects [23]. We first constructed a 88×128 pixel-sized normalized silhouette sequence (see Fig. 4) from the silhouette sequence (see Fig. 1(b)) based on the bounding box of the subject, and then averaged them over one gait period to compute the GEI (see Fig. 2(a)). Once the GEIs of a probe and gallery have been computed as feature vectors G_p and G_g , respectively, the dissimilarity score between the probe and gallery is calculated by Euclidean distance as

$$S_{gait} = \|G_p - G_g\|_2, \quad (1)$$

where $\|\cdot\|_2$ denotes the L_2 norm.

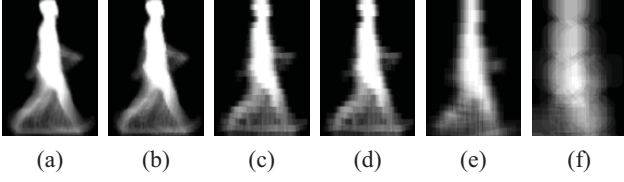


Figure 2: GEs extracted from silhouette sequences at 30 fps of various sizes: (a) 640×320 , (b) 320×240 , (c) 160×120 , (d) 80×60 , (e) 40×30 , and (f) 20×15 pixels. The given sizes are for the original/scaled-down images. Note that the final size of the GEI (and also the normalized silhouette sequence) is consistent (88×128 pixels) throughout the various SRs.

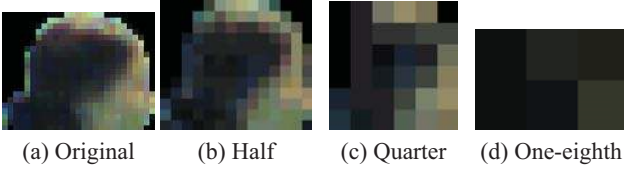


Figure 3: Head templates extracted from original images as well as scaled-down images. The sizes are indicated in reduction rate from the original images. Note that the head template size is dependent on a subject, and the size of this specific subject is 27×22 pixels for (a), 13×11 pixels for (b), 6×6 pixels for (c), and 3×2 pixels for (d).

Head: Although most of the face recognition approaches focus on the interior region of the face excluding the hair and face contour parts, we cannot achieve high recognition accuracy using only the interior region when the SR is extremely small (e.g., less than 25×25 pixels). In an analogous fashion to the extension from iris biometrics to periocular biometrics [31], we extend the region used from the inside face region to the head region, including the hair and face contour parts. More specifically, we keep the color textured image of the head region with the silhouette mask (see Fig. 3(a)) as a gallery template. The size of the head region is dependent on the subject and it ranges from 18×20 pixels to 31×25 pixels for a 640×480 pixel image before it is scaled down.

We calculate the dissimilarity score between two head images based on a conventional template matching algorithm as follows. Let F_{p_i} be a head feature (color texture image) of a probe at the i -th frame, and let $F_{g_{j,k}}$ be a head feature of a gallery at the j -th frame with k -th spatial displacement within the search regions. The dissimilarity score between the probe and the gallery is calculated by correlation-based template matching as

$$S_{head} = \min_{i,j,k} [1 - NCC(F_{p_i}, F_{g_{j,k}})], \quad (2)$$

where $NCC(F_{p_i}, F_{g_{j,k}})$ is the normalized cross correlation (NCC) between F_{p_i} and $F_{g_{j,k}}$ within the region of interest

(silhouette mask).

Note that the other advanced approaches, including the feature extraction process [32], pose normalization [33], and/or face alignment [34] do not consistently work for a wide range of SR: from middle-level SR (see Fig. 3(a)) to extremely low-level SR (see Fig. 3(d)). We therefore adopted the NCC-based template matching because it consistently works for various SRs, even though it is not a state-of-the-art face recognition algorithm.

The height: As described in Section 2.1, we can easily compute the apparent height of the subject in the rectified image based on the bounding box of the silhouette. Because the camera calibration is done in advance and each subject walks along a straight line at a fixed depth from the camera, we convert the apparent height into actual height in the real 3D world simply by multiplying the ratio of the focal length and depth from the subject. Moreover, because the heights are defined frame-by-frame and vary because of the up-and-down motion caused by gait, we define the height of the subject as the average of the heights over the image sequence. In short, height h is calculated as

$$h = \frac{1}{N_f} \sum_{i=1}^{N_f} Z_i, \quad (3)$$

where Z_i is the height of the i -th frame, and N_f is the number of frames in the image sequence.

Given a pair of heights, we then calculate the dissimilarity score as the absolute difference of the heights as

$$S_{height} = |h_p - h_g|, \quad (4)$$

where h_p and h_g are the heights of the subjects in the probe and gallery.

2.3. Quality measures

To simulate the multi-biometric scores under various quality conditions, we scaled down the image sequences in terms of the SRs and TRs and then computed the dissimilarity scores for individual gait, head, and height biometrics in the same way as described above.

SR: We considered a total of 13 variations of SRs including the original image size. For this purpose, we scaled down the original 640×480 pixel images by $3/4$, $1/2$, $1/3$, $1/4$, $1/5$, $1/6$, $1/7$, $1/8$, $1/10$, $1/12$, $1/16$, $1/32$, which correspond to images of dimensions 480×360 , 320×240 , 213×160 , 160×120 , 128×96 , 106×80 , 91×68 , 80×60 , 64×48 , 53×40 , 40×30 , and 20×15 pixels, respectively. These scaled-down images simulated the difference with respect to not only the image size itself but also the distance from the camera to the subject.² Examples of head

²Although the weak perspective projection assumption with respect to the subject needs to hold to accurately simulate different distances, this assumption is almost true in our case since the distance to the subject is sufficiently large compared with the size of the subject.



Figure 4: Normalized silhouette sequences (top: half period at 30 fps, bottom: half period at 5 fps)

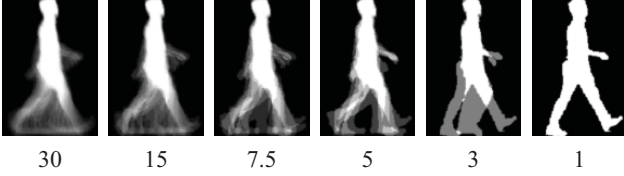


Figure 5: GEIs extracted from silhouette sequences of fixed size (640×480 pixels) at various TRs [fps]

templates and GEIs for the scaled-down images are shown in Figs. 2 and 3.

TR: We considered 10 variations of TRs including the original frame-rate (30 fps). For this purpose, we reduced the TR by selecting images at specific frame intervals from an original image sequences (hereafter called frame-skipped image sequences) at frame rates of 15, 10, 7.5, 6, 5, 3.75, 3, 2, and 1 fps. Normalized silhouette sequences at 30 and 5 fps are shown in Fig. 4. GEIs extracted from the normalized silhouette sequences at various TRs are shown in Fig. 5.

2.4. Constructed database

We drew 1912 subjects from the OU-ISIR Gait Database Large Population Dataset [23] and selected a pair of subsequences from the 85-deg view (almost side view) for two different sessions for each subject, one as a gallery and the other as a probe.

Because we had multiple possible starting frames when reducing the TR (e.g., we could pick image sequences of odd or even number frames, respectively, at 15 fps), we used all of the frame-skipped image sequences starting from various frames for a probe, while we used a single frame-skipped image sequence starting from the first frame for the gallery to avoid a quadratic increase of scores. We therefore prepared 2, 3, 4, 5, 6, 8, 10, 15, and 19 frame-skipped image sequences per probe³ for 15, 10, 7.5, 6, 5, 3.75, 3, 2, and 1 fps, respectively, and denote the number of frame-skipped image sequences as N_{TR} .

We then computed the dissimilarity matrices of the gait, head, and height between 1912 N_{TR} probes and 1912 galleries for all combinations of the 13 variations of SRs and 10

³Although 30 frame-skipped image sequences are available for 1-fps down-sampling from an original 30-fps video in principle, they are limited to 19 frame-skipped image sequences. This is because the number of frames per complete gait cycle is only 19 for some subjects and we wished to maintain consistency among subjects.

variations of TRs including the original resolution, which sums up to 130 combinations of SRs and TRs. The constructed score database comprises these dissimilarity matrices with the subject ID list as well as the quality measures (i.e., SRs and TRs). As a result, we can draw 1912 N_{TR} genuine (true match) scores and $1912 N_{TR} \times 1,911 = 3,653,832 N_{TR}$ imposter (false match) scores for each biometric modality and quality, which results in a very large-scale database, i.e., 1,814,488 genuine scores and 3,467,486,568 imposter scores for each biometric modality.

3. Performance evaluation

3.1. Protocols

Protocol 1: This protocol aims to benchmark quality-independent score-level fusion approaches for each quality, i.e., the performance evaluation is done within each quality independently in protocol 1. We considered four typical qualities by combining high and low SRs and TRs: 640×480 pixels at 30 fps (Q_{HH}), 640×480 pixels at 1 fps (Q_{HL}), 53×40 pixels at 30 fps (Q_{LH}), and 53×40 pixels at 1 fps (Q_{LL}). The score database was then randomly divided into training and test sets, disjoint with respect to subjects, and this two-fold cross validation was repeated 10 times so as to reduce the influence of the random divisions. The scores were normalized based on the average and standard deviation computed with the training set of the quality Q_{HH} for each modality⁴. Finally, performances in verification and identification scenarios were evaluated with typical measures such as ROC curves, EERs, FRRs at specific FARs, area under curves (AUCs), CMC curves, and rank- n identification rates for each quality independently.

Protocol 2: This protocol aims to benchmark quality-dependent score-level fusion approaches where the score database is disjoint in terms of not only subjects but also qualities, unlike protocol 1, where the same qualities are shared between training and test sets. More specifically, the qualities for the training and test sets are defined as shown in Table. 1. Thereafter, we follow the same processes as for protocol 1: 10-times two-fold cross validation and score-normalization with the training set of the quality Q_{HH} . Finally, performances are evaluated in verification scenarios for the test scores including all the test qualities.

3.2. Benchmarks

Protocol 1: Score-level fusion approaches mainly fall into three categories: transformation-based, classification-based, and probability density-based approaches, as summarized in [14]. We therefore provide in total seven benchmarks from a wide variety of score-level fusion ap-

⁴Note that the score normalization does not essentially affect performance for training-based approaches (exceptions are Sum and Min) since differences in score scales are absorbed in the training process.

Table 1: Qualities for training and test sets in protocol 2.

Data set	SR [pixels]	TR [fps]
Training	640 × 480, 320 × 240, 160 × 120, 106 × 80, 80 × 60, 53 × 40	30, 15, 7.5, 5, 3, 1
Test	480 × 360, 213 × 160, 128 × 96, 91 × 68, 64 × 48	10, 6, 3.75, 2

proaches for protocol 1. We provide the sum rule (Sum) and minimum rule (Min) for the transformation-based approaches [9]; a support vector machine (SVM) with a radial basis function kernel [13] for the classification-based approach; and the Gaussian mixture model (GMM)⁵ [12], linear logistic regression (LLR) [11], kernel density estimation (KDE) [35], and non-parametric posterior estimation using floating control point (FCP) [14] for the probability density-based approaches.

Protocol 2: Because protocol 2 aims to evaluate quality-dependent score-level fusions, we provide both quality-independent and quality-dependent score-level fusion approaches to highlight the effectiveness of quality-dependent approaches over quality-independent approaches. More specifically, we first chose the three best approaches with respect to verification performance in protocol 1: GMM, LLR, and FCP (see 3.3) in addition to Sum as a baseline. We then trained the parameters (or posteriors for FCP) with the training set of the highest SR and TR (the same as Q_{HH}) and used them regardless of qualities, resulting in quality-independent approaches.

In addition to these quality-independent approaches, we introduce Q-stack [36], a concatenated vector of the multi-modal biometric scores and the quality measures for quality-dependent approaches. In our database, the Q-stack is defined as a five-dimensional vector composed of three biometric scores, S_{gait} , S_{head} , and S_{height} , as well as two quality measures, SR q_{SR} , and TR q_{TR} . Note that we define q_{SR} and q_{TR} on a logarithm scale of spatial reduction and frame rate for numerical stability. The Q-stack vectors of the training sets are fed into each of the GMM, LLR, and FCP methods to create quality-dependent versions (called GMM (Q-stack), LLR (Q-stack), and FCP (Q-stack), respectively).

Furthermore, we introduce another quality-dependent version of LLR that interpolates the weights for each modality of a test quality measure from those obtained in the training set using a Gaussian process regression [37] (called LLR (GPR), see [19] for details).

3.3. Evaluation

Protocol 1: Performance evaluation results for protocol

⁵The number of mixture components are determined between 1 and 20 so as to optimize minimum description length criterion and random selection process of initial seeds is repeated 10 times for each number of

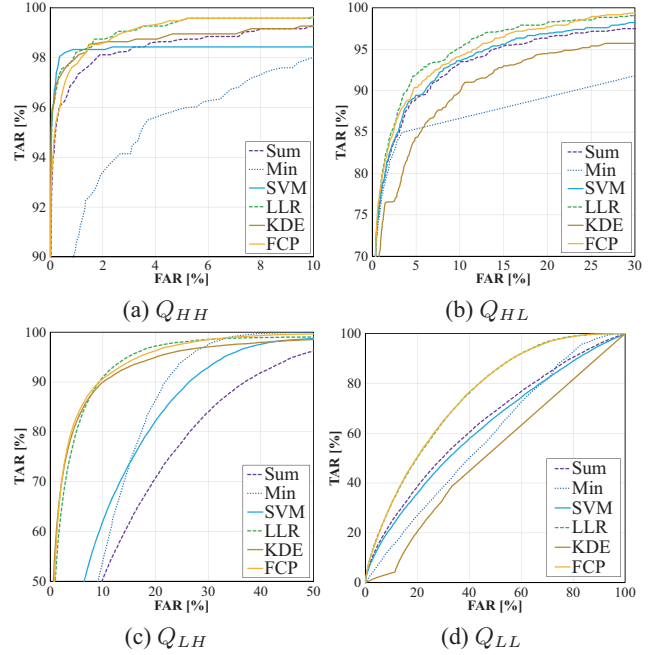


Figure 6: ROC curves for protocol 1 (the first cross-validation set). Note that the scales differ among graphs.

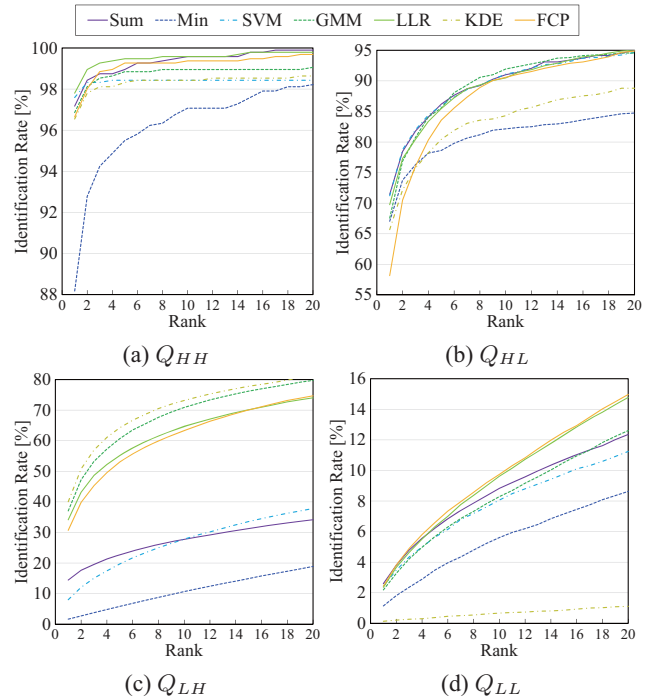


Figure 7: CMC curves for protocol 1. Note that scales differ among graphs.

1 are shown by the ROC and CMC curves in Figs. 6 and 7. In addition, we selected some typical measures: EERs

mixture components to mitigate the effect of randomness.

Table 2: EER, FRRs at 1% and 10% FARs ($FRR_{1\%}$ and $FRR_{10\%}$, respectively), and AUC for protocol 1. Bold and italic bold fonts indicate the best and second best accuracies throughout this paper.

Fusion rule	EER [%]				$FRR_{1\%}$ [%]				$FRR_{10\%}$ [%]				AUC [%]			
	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}
Sum	2.0	23.6	7.6	39.6	2.7	82.6	22.4	94.5	0.7	49.4	6.7	75.6	0.3	14.9	2.3	35.4
Min	4.4	18.1	11.8	44.8	9.3	93.7	23.6	98.5	2.0	45.8	13.3	85.6	0.6	10.8	4.9	41.4
SVM	1.6	17.8	8.4	44.2	1.5	79.2	22.4	94.8	1.5	35.6	6.7	76.9	0.8	10.1	3.3	41.8
GMM	1.5	10.6	6.5	35.5	1.5	44.5	21.4	94.1	0.4	9.0	4.5	74.8	0.3	3.6	1.6	28.8
LLR	1.5	9.9	6.3	32.5	1.9	49.3	21.2	93.3	0.4	9.2	4.8	68.5	0.2	3.8	1.6	25.1
KDE	1.6	11.9	9.4	49.8	1.9	41.7	26.2	97.8	0.6	9.7	8.7	87.7	0.3	5.2	4.1	49.5
FCP	1.5	10.2	6.7	32.3	1.9	44.5	21.5	93.4	0.3	9.8	5.5	68.3	0.2	3.6	1.7	24.9

Table 3: Rank-1/5 identification rates for protocol 1.

Fusion rule	Rank-1 [%]				Rank-5 [%]			
	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}	Q_{HH}	Q_{HL}	Q_{LH}	Q_{LL}
Sum	97.0	14.5	72.2	2.6	98.9	23.2	87.1	6.4
Min	89.8	1.7	68.5	1.3	95.4	6.1	80.5	3.8
SVM	97.4	12.7	71.4	1.6	98.4	25.2	85.4	4.0
GMM	97.1	34.7	67.9	1.9	98.7	58.2	84.7	5.5
LLR	97.9	32.6	69.4	2.5	99.3	53.9	85.1	6.7
KDE	96.8	35.9	62.5	0.6	98.2	58.2	77.4	1.7
FCP	96.6	32.3	59.6	2.6	98.8	55.3	84.0	6.8

and FRRs at 1% and 10% FARs as well as AUC for ROC in Table 2 and rank-1 and rank-5 identification rates for CMC in Table 3.

As a result, we can see that some of the probability density-based methods such as GMM, LLR, and FCP perform well in the verification scenarios (see Fig. 6 and Table 2) because the probability density-based approaches guarantee optimality in terms of ROC [38] as long as the estimated probability densities are correct.

However, the highest accuracies in the identification scenarios (see Fig. 7 and Table 3) are relatively mixed. Note that the identification performances are determined by probe-dependent rank statistics, whereas the verification performances are determined by aggregated score distributions, which may produce inconsistency between verification and identification performances, as reported in [39].

Because the classification-based and density-based approaches to score-level fusion used in this paper also aggregate the score distribution, we chose the three best methods with respect to verification, GMM, LLR, and FCP, and used them in the following performance analysis for protocol 2.

Protocol 2: We evaluated the accuracies of the verification scenarios using ROC curves in Fig. 8 and the EERs and FRRs at 1% and 10% FARs as well as AUCs in Table 4. As a result, we can see that quality-independent approaches such as Sum, GMM, LLR, and FCP do not perform well, with EERs of more than 40%. Conversely, almost all of the quality-dependent approaches such as LLR (Q-stack), FCP (Q-stack), and LLR (GPR) achieve much higher accuracies than the quality-independent approaches (e.g., the EER

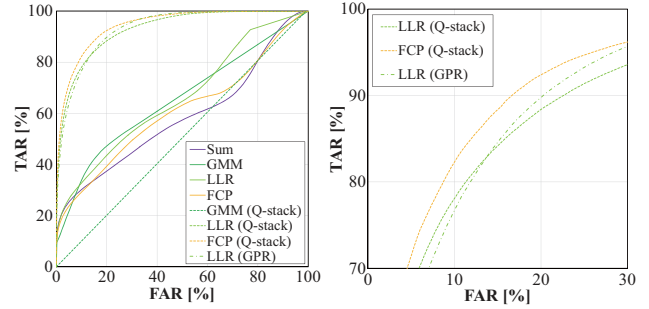


Figure 8: ROC curves for protocol 2 (the first cross-validation set) (Left) and a selected portion of the curve (right).

Table 4: EERs, FRRs at 1% and 10% FARs ($FRR_{1\%}$ and $FRR_{10\%}$, respectively) and AUC for protocol 2.

Fusion rule	EER [%]	$FRR_{1\%}$ [%]	$FRR_{10\%}$ [%]	AUC [%]
Sum	46.0	82.4	69.8	41.9
GMM	41.0	88.6	66.1	35.3
LLR	43.0	83.0	67.2	36.2
FCP	43.2	84.3	70.8	40.5
GMM (Q-stack)	44.1	99.0	89.9	40.8
LLR (Q-stack)	15.3	56.8	21.9	7.0
FCP (Q-stack)	14.4	52.0	17.7	6.5
LLR (GPR)	15.6	61.3	23.3	7.1

for the FCP (Q-stack) is less than 15%), except for GMM (Q-stack). This exceptional worse performance by GMM (Q-stack) is caused by degeneration of covariance matrices of GMM due to discrete training qualities (see Table. 1). As a result, the probability density for the test quality which is different from the training quality, becomes zero both for positives and negatives and therefore the fused score lose its discrimination capability. Apart from this point, as we expect, it is essential to consider quality measures to achieve better performances for protocol 2, and we believe that the quality-dependent approaches used in this paper are appropriate baselines for benchmarking quality-dependent score-level fusion approaches in future.

4. Discussion and future work

Whereas we considered multiple scores derived from multi-modal biometrics and provided a single matcher for each modality in this paper, it is also possible to introduce multiple scores derived from multiple matchers, as in the existing biometric score database NIST-BSSR1 [27], which contains scores derived from multiple face recognition algorithms. From the viewpoint of quality-dependent score-level fusion, it is particularly interesting to introduce matchers with different sensitivities to the quality measures used in this paper (i.e., SR and TR). For example, gait feature representations encoding more temporal and/or motion information (e.g., [40, 41] may be sensitive to TR (i.e., yielding higher accuracy for higher TR and vice versa), while those encoding more static (shape) information (e.g., [42]) may be insensitive to TR.

Moreover, we can improve the sensitivity to SR and TR by incorporating spatial and/or temporal super resolution techniques [43]. In fact, spatially super-resolved face images are used to fuse gait and side-view faces in [44]. It would therefore be one interesting future research avenue to construct such an advanced multi-modal biometric score database that includes multiple matchers as well as super resolution techniques.

In addition, whereas we treat SR and TR as quality measures in the context of fusion of gait, head, and height biometrics, there exist even more qualities to be considered in a real situation. In particular, because our biometric score database was built upon the OU-ISIR Gait Database Large Population Dataset [23] collected in a relatively controlled situation (e.g., indoors, controlled illumination, predefined course), we need to consider a variety of covariate factors that might arise in a real situation. For example, background motion artifacts caused by trees or additional persons may degrade the silhouette qualities⁶ for gait biometrics, while illumination change caused by cloud cover could drastically change head texture. Furthermore, camera calibration errors directly affect the height biometrics. As such, it would be beneficial to collect biometric scores in real situations with additional quality measures (e.g., illumination changes, view changes, clothing, silhouette qualities, etc.) in future research.

Because it is more a challenging task to fuse a high-dimensional Q-stack vector containing a plethora of multi-modal multi-matcher biometric scores as well as the quality measures as above, it is also essential to add a benchmark for score-level fusion that incorporates dimension reduction as preprocessing (e.g., [45]).

⁶Moreover, silhouettes extracted from lower-SR original images may be worse than downscaled silhouettes used in this paper, since foreground/background blending at the boundary may make silhouette extraction difficult. A possible solution to fill this gap is using alpha matting instead of hard segmentation for better silhouette quality [7].

5. Conclusion

We constructed a single sensor-based multi-quality multi-modal biometric score database. We extracted gait, head, and height biometrics from a single walking image sequence, and considered SR and TR as the quality measures. As a result, the database contains a large number of biometric scores: 1,814,488 genuine scores and 3,467,486,568 imposter scores under a total of 130 combinations of SRs and TRs. We defined two protocols: quality-independent and quality-dependent score-level fusions, and also provided performance evaluation results with several benchmarks for each protocol, which will advance studies on quality-dependent score-level fusions.

Acknowledgment

This work was partly supported by JSPS Grants-in-Aid for Scientific Research (S) 21220003, “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, Science and Technology, the Japanese Government, and the JST CREST “Behavior Understanding based on Intention-Gait Model” project.

References

- [1] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multi-biometrics*, ser. Int. Series on Biometrics. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.
- [2] T. Murakami, K. Takahashi, and K. Matsuura, “Towards optimal countermeasures against wolves and lambs in biometrics,” in *Proc. of the 5th IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2012)*, Sept 2012, pp. 69–76.
- [3] F. Besbes, H. Trichili, and B. Solaiman, “Multimodal biometric system based on fingerprint identification and iris recognition,” in *Proc. of the 3rd Int. Conf. on Information and Communication Technologies: From Theory to Applications*, April 2008, pp. 1–5.
- [4] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun, “Discriminative multimodal biometric authentication based on quality measures,” *Pattern Recognition*, vol. 38, no. 5, pp. 777–779, May 2005.
- [5] R. Jillela and A. Ross, “Mitigating effects of plastic surgery: Fusing face and ocular biometrics,” in *Proc. of the 5th IEEE Int. Conf. on Biometrics: Theory, Applications and Systems (BTAS 2012)*, Sept 2012, pp. 402–411.
- [6] X. Zhou and B. Bhanu, “Feature fusion of side face and gait for video-based human identification,” *Pattern Recognition*, vol. 41, no. 3, pp. 778–795, 2008.
- [7] M. Hofmann, S. M. Schmidt, A. Rajagopalan, and G. Rigoll, “Combined face and gait recognition using alpha matte preprocessing,” in *Proc. of the 5th IAPR Int. Conf. on Biometrics*, New Delhi, India, Mar. 2012, pp. 1–8.
- [8] D. Muramatsu, H. Iwama, Y. Makihara, and Y. Yagi, “Multi-view multi-modal person authentication from a single walking image sequence,” in *Proc. of the 6th IAPR Int. Conf. on Biometrics*, 2013, pp. 1–8.
- [9] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, “On combining classifiers,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

- [10] S. C. Dass, K. Nandakumar, and A. K. Jain, "A principled approach to score level fusion in multimodal biometric systems," in *Proc. of the 5th Int. Conf. on Audio- and Video-Based Biometric Person Authentication*, Ny, USA, July 2005, pp. 1049–1058.
- [11] F. Alonso-Fernandez, J. Fierrez, D. Ramos, and J. Ortega-Garcia, "Dealing with sensor interoperability in multi-biometrics: the upm experience at the biosecure multimodal evaluation 2007," in *Proc. of SPIE 6994, Biometric Technologies for Human Identification V*, no. 69440J, Orlando, FL, USA, Mar. 2008, pp. 1–12.
- [12] K. Nandakumar, Y. Chen, S. C. Dass, and A. K. Jain, "Likelihood ratio-based biometric score fusion," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 342–347, 2008.
- [13] M. He, S. J. Horng, P. Fan, R. S. Run, R. J. Chen, J. L. Lai, M. Khan, and K. O. Sentosa, "Performance evaluation of score level fusion in multimodal biometric systems," *Pattern Recognition*, vol. 43, no. 5, pp. 1789–1800, May 2010.
- [14] Y. Makihara, D. Muramatsu, M. Hossain, H. Iwama, T. Ngo, and Y. Yagi, "Score-level fusion by generalized delaunay triangulation," *Proc. of the 2nd IEEE/IAPR Int. Joint Conf. on Biometrics (IJCBC 2014)*, pp. 1–8, 2014.
- [15] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz, "Confidence measures for multimodal identity verification," *Information Fusion*, vol. 3, no. 4, pp. 267–276, 2002.
- [16] E. Krichen, S. Garcia-Salicetti, and B. Dorizzi, "A new probabilistic iris quality measure for comprehensive noise detection," in *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE Int. Conf. on, 27-29 2007*, pp. 1–6.
- [17] A. Harriero, D. Ramos, J. Gonzalez-Rodriguez, and J. Fierrez, "Analysis of the utility of classical and novel speech quality measures for speaker verification," in *Proc. of the 3rd LAPR Int. Conf. on Biometrics*, 2009, pp. 434–442.
- [18] K. Toh, W. Yau, E. Lim, L. Chen, and C. Ng, "Fusion of auxiliary information for multi-modal biometrics authentication," in *Proc. of Int. Conf. on Biometrics*, Hong Kong, 2004, pp. 678–685.
- [19] T. Kimura, Y. Makihara, D. Muramatsu, and Y. Yagi, "Quality-dependent score-level fusion of face, gait, and the height biometrics," *IPSJ Trans. on Computer Vision and Applications*, vol. 6, pp. 543–547, Jul. 2014.
- [20] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *Journal of Forensic Sciences*, vol. 56, no. 4, pp. 882–889, 2011.
- [21] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi, "Gait verification system for criminal investigation," *IPSJ Trans. on Computer Vision and Applications*, vol. 5, pp. 163–175, Oct. 2013.
- [22] "Darpa, human identification at a distance program," <http://www.darpa.mil/iao/HID.htm>.
- [23] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.
- [24] "Banca score database," http://personal.ee.surrey.ac.uk/Personal/Norman.Poh/web/banca_multi/main.php%3Fbodyfile=entry_page.html.
- [25] N. Poh and S. Bengio, "Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication," *Pattern Recogn.*, vol. 39, no. 2, pp. 223–233, Feb. 2006.
- [26] N. Poh, T. Bourlai, and J. Kittler, "A multimodal biometric test bed for quality-dependent, cost-sensitive and clientspecific score-level fusion algorithms," *Pattern Recognition*, vol. 43, no. 3, pp. 1094–1105, Mar. 2010.
- [27] "NIST biometric scores set," <http://www.itl.nist.gov/iad/894.03/biometricscores/1>.
- [28] Y. Makihara and Y. Yagi, "Silhouette extraction based on iterative spatio-temporal local color transformation and graph-cut segmentation," in *Proc. of the 19th Int. Conf. on Pattern Recognition*, Tampa, Florida USA, Dec. 2008, pp. 1–4.
- [29] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *Proc. of the 17th International Conference on Pattern Recognition*, vol. 1, Aug. 2004, pp. 211–214.
- [30] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.
- [31] C. Tan and A. Kumar, "Towards online iris and periocular recognition under relaxed imaging constraints," *IEEE Trans. on Image Processing*, vol. 22, no. 10, pp. 3751–3765, Oct. 2013.
- [32] K. Ito, T. Aoki, T. Hosoi, and K. Kobayashi, "Face recognition using phase-based correspondence matching," in *2011 IEEE Int. Conf. on Automatic Face Gesture Recognition*, March 2011, pp. 173–178.
- [33] A. Asthana, M. Jones, T. Marks, K. Tieu, and R. Goecke, "Pose normalization via learned 2d warping for fully automatic face recognition," in *Proc. BMVC*, 2011, pp. 127.1–127.11.
- [34] G. Tzimiropoulos and M. Pantic, "Gauss-newton deformable part models for face alignment in-the-wild," in *27th IEEE Conf. on Computer Vision and Pattern Recognition*, June 2014, pp. 1851–1858.
- [35] B. Ulery, W. Fellner, P. Hallinan, A. Hicklin, and C. Watson, "Studies of biometric fusion. appendix c. evaluation of selected biometric fusion techniques," *NIST Interagency Report*, vol. 7346, pp. 1–14, 2006.
- [36] K. Kryszczuk and A. Drygajlo, "Improving classification with class-independent quality measures: Q-stack in face verification," in *In 2nd Int. Conf. on Biometrics, Seoul, South Korea*, 2007, pp. 1124–1133.
- [37] C. K. I. W. Carl Edward Rasmussen, *Gaussian Processes for Machine Learning*. The MIT Press, 2006.
- [38] J. Neyman and E. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Philosophical Trans. of the Royal Society of London*, vol. 231, pp. 289–337, 1933.
- [39] B. DeCann and A. Ross, "Relating roc and cmc curves via the biometric menagerie," in *Prof. of the 6th IEEE Int. Conf. on Biometrics: Theory, Applications and Systems*, Sept 2013, pp. 1–8.
- [40] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2164–2176, nov. 2012.
- [41] K. Bashir, T. Xiang, and S. Gong, "Gait representation using flow fields," in *Proc. of the 20th British Machine Vision Conference*, London, UK, Sep. 2009, pp. 1–11.
- [42] Y. Guan, C.-T. Li, and S. Choudhury, "Robust gait recognition from extremely low frame-rate videos," in *Biometrics and Forensics (IWBF), 2013 International Workshop on*, April 2013, pp. 1–4.
- [43] N. Akae, A. Mansur, Y. Makihara, and Y. Yagi, "Video from nearly still: an application to low frame-rate gait recognition," in *Proc. of the 25th IEEE Conf. on Computer Vision and Pattern Recognition (CVPR2012)*, Providence, RI, USA, Jun. 2012, pp. 1537–1543.
- [44] X. Zhou and B. Bhanu, "Integrating face and gait for human recognition at a distance in video," *IEEE Trans. on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 37, no. 5, pp. 1119–1137, oct. 2007.
- [45] K. Kryszczuk and N. Poh, "Handling high dimensionality in biometric classification with multiple quality measures using locality preserving projection," in *IEEE Computer Society and IEEE Biometrics Council Workshop on Biometrics 2010*, San Francisco, CA, USA, Jun. 2010, pp. 1–8.