Gait recognition from ill-posed silhouettes

Submitted to
Graduate School of Information Science and Technology
Osaka University

January 2020

Md. Zasim Uddin

Abstract

The importance of automated visual surveillance in a public space has been increased in response to the recent rising concerns about a safe and secure society. Gait, which can be defined as the manner of walking of a person, has been recognized as a potential biometrics modality, and it enables person recognition from a low-resolution image sequence in a CCTV footage. Image sequence can be captured at a distance without the subject's cooperation, and information associated with shape and his/her walking motion can be obtained from the binary silhouettes from the captured image sequence. Therefore, person recognition using silhouette-based gait features (i.e., gait recognition) is a promising real-life application. However, practical deployment in a real-world scenario is still considered a challenging task because several issues should be managed to achieve efficient gait recognition. Such issues can be related to the presence of external factor to a subject (e.g., carried object, clothing, shoes); the internal factor to a subject (e.g., walking speed, aging, pregnancy); the environmental context (e.g., illumination, walking surface, occlusion) and the quality of captured image sequence from a camera (e.g., spatial resolution, temporal resolution, observation view). These factors make the gait recognition more challenging and prevent the use of gait recognition in the real-world scenario. In order to realize efficient gait recognition in the situation with those factors, this thesis addresses the following three issues: i) carried object(s) (COs) ii) occlusion and iii) data quality.

The factors are considered in this thesis, affect the captured image sequence of a person for gait recognition. Therefore, the obtained silhouettes from the corresponding captured images can be either ill-posed or well-posed based on visual quality. The well-posed silhouettes mean that the obtained silhouettes are of good quality, whereas ill-posed silhouettes are degraded, problematic, and low-quality silhouettes. For example, COs silhouette is added to the person's silhouette, whereas some parts of a person's silhouette are unobservable due to the occlusion. Regarding data quality of the captured image sequence from a CCTV camera, the obtained silhouettes are different depending on several camera settings such as spatial resolution, temporal resolution, and observed view of a person from the camera.

First, in order to tackle the COs issue, we constructed the world's largest gait database with real-life COs and analyzed the performance of silhouette-based gait recognition. Whereas existing databases for gait recognition include at most 306 subjects with COs, we constructed an extremely large-scale gait database that includes 62,528 subjects, with an equal distribution of males and females, and ages ranging from 2 to 95 years old. Moreover, existing gait databases include person image sequences with COs, but the variation of the COs' type, positions, and numbers are limited and predefined. However, our constructed database considered a wide variety of COs' types, positions, and numbers. Besides, the silhouettes data in the database were manually annotated into seven carrying status (CS) labels. The extremely large-scale gait database with CS labels enabled us to evaluate and analyze the impact of the training data size, the recognition difficulty level of the CS labels, and the classification of CS labels.

Research for the second issue is managing occlusion. The presence of occlusion in an image sequence obscures a significant amount of the human body. However, usual gait recognition methods require an unoccluded silhouette sequence. Thus, occlusion handling is an important issue for gait recognition. We, therefore, propose a silhouette sequence reconstruction approach from an occluded sequence (sVideo) based on leveraging the powerful capabilities of conditional deep generative adversarial network (GAN). To reconstruct the occluded silhouette sequence well constrained, we regularize the training of the proposed generative network, based on triplet hinge loss incorporating with Wasserstein GAN (WGAN-hinge). To the best of our knowledge, WGAN-hinge is the first adversarial loss that supervises the generator network during training by incorporating pairwise similarity ranking information. The proposed approach was evaluated on multiple challenging occlusion patterns. The experimental results demonstrate that the proposed approach outperforms existing state-of-the-art benchmarks.

Finally, we considered the quality of the captured image sequence from a CCTV camera; particularly, we analyze the camera-based qualities of the spatial resolution (SR) and temporal resolution (TR) along with the observation view because they affect the recognition accuracy. It is well-known that uni-modal biometrics is usually affected by some practical problems such as poor recognition performance and vulnerability to spoofing attacks. Therefore, we extracted gait, head, and soft biometric height modalities from the captured image sequence obtained from a single camera and introduce SR, TR, and view as the quality that significantly affects biometric system performance. Moreover, data quality (i.e., quality measure) that affects the matching score of each modality is often incorporated as a quality-dependent score-level fusion, which is a popular and promising approach. We considered seven and ten scaling factors for SR and TR, respectively, with four view variations. Therefore, a large database is constructed,

comprising approximately four million genuine and 7.5 billion imposter scores. Later, we analyze the recognition accuracies associated with gait, head, and height modalities in different spatial and temporal resolutions along with the observation views. We observed that TR impacts profoundly on the gait modality, whereas SR for the head modality. These analyses can be useful for forensic analysis, mainly when the quality of the captured image sequences is very poor. Finally, we evaluate and analyze the performance considering all of the modalities in a score-level fusion by designing two different protocols that are the quality-independent and quality-dependent manner using state-of-the-art score-level fusion approaches. These evaluation results will be beneficial for score-level fusion research.

This thesis together with the considered issues could make a significant contribution to improving the gait recognition performance in various real-world scenarios that further enable the adoption of gait-based person recognition into automated visual subservience and forensic applications.

Acknowledgments

It is a great pleasure to mention the people who have made the creation of this thesis possible.

First and foremost, I would like to express my deepest gratitude to my supervisor, Yasushi Yagi, for allowing me to undertake this research program, as well as his invaluable guidance throughout my whole Ph.D.

I am deeply grateful to two of my advisers during Ph.D., Yasushi Makihara and Daigo Muramatsu. I learned a great deal from the discussions with them, and I received lots of advice on the scientific approach, logical way of scientific thinking, designing experiments, writing an academic paper, and presentation. Also, their immense knowledge and inspiration have been of great value to me. Without their contributions, this work would not have been possible.

My special thanks go to Thanh Trung Ngo, Noriko Takemura, and Md. Atiqur Rahman Ahad for their assistance in the last year. I would especially like to thank Al Mansur, a former member of Yagi Lab, who helped me a lot at the initial stage at Yagi lab, as well as in Japan.

I would also like to thank my dearest family and friends. Their emotional supports encourage me not to give up and to continue moving towards my goal. Without their constant support and inspiration, this work probably would never be done.

List of publications

- **Journal (with peer review)**

  1. **Md. Zasim Uddin**, Daigo Muramatsu, Noriko Takemura, Md. Atiqur Rahman Ahad and Yasushi Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion, IPSJ Trans. on Computer Vision and Applications, vol.11, no.9, pp.1-18, Nov., 2019.
  2. **Md. Zasim Uddin**, Thanh Trung Ngo, Yasushi Makihara, Noriko Takemura, Xiang Li, Daigo Muramatsu and Yasushi Yagi. The OU-ISIR Large Population Gait Database with Real-Life Carried Object and its performance evaluation, IPSJ Trans. on Computer Vision and Applications, vol.10, no.5, pp.1-11, May, 2018.
  3. **Md. Zasim Uddin**, Daigo Muramatsu, Takuhiro Kimura, Yasushi Makihara and Yasushi Yagi. MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation, IPSJ Trans. on Computer Vision and Applications, vol.9, no.18, pp.1-25, Jul., 2017.
  4. [1] Daigo Muramatsu, Akira Shiraishi, Yasushi Makihara, **Md. Zasim Uddin**, Yasushi Yagi. Gait-based Person Recognition Using Arbitrary View Transformation Model", IEEE Trans. on Image Processing, vol.24, no.1, pp.140-154, Jan., 2015.

- **Conference (with peer review)**

  1. [2] Yasushi Makihara, Al Mansur, Daigo Muramatsu, **Md. Zasim Uddin** and Yasushi Yagi. Multi-view Discriminant Analysis with Tensor Representation and Its Application to Cross-view Gait Recognition. The 11th IEEE int. conf. on Automatic Face and Gesture Recognition, pp.1-8, Ljubljana, Slovenia, Apr., 2015

- **Conferences and workshop (without review)**

[1] Not included in the thesis.
[2] Not included in the thesis.

1. **Md. Zasim Uddin**, Thanh Trung Ngo, Yasushi Makihara, Noriko Takemura, Xiang Li, Daigo Muramatsu and Yasushi Yagi. Construction and Performance Evaluation of the OU-ISIR Large Population Gait Database with Real-life Carried Object. The 22nd SANKEN int. Symposium, Suita, Japan, Jan. 2019
2. **Md. Zasim Uddin**, Thanh Trung Ngo, Yasushi Makihara, Noriko Takemura, Xiang Li, Daigo Muramatsu and Yasushi Yagi. The OU-ISIR Large Population Gait Database with Real-Life Carried Object and its performance evaluation. The 13th int. Workshop on Robust Computer Vision, Beijing, China, Jan. 2019
3. **Md. Zasim Uddin**, Daigo Muramatsu, Takuhiro Kimura, Yasushi Makihara and Yasushi Yagi. MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. The 12th int. Workshop on Robust Computer Vision, Nara, Japan, Jan. 2018
4. **Md. Zasim Uddin**, Daigo Muramatsu, Takuhiro Kimura, Yasushi Makihara and Yasushi Yagi. MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. The Seventh Symposium on Biometrics, Recognition and Authentication, Tokyo, Japan, Nov. 2017

Contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Biometrics and gait

Biometrics is the science of recognizing the subject identity, based on their inherent physiological and/or behavioral traits associated with a person [1]. Physiological traits include the face, iris, DNA, fingerprint, palm print, finger veins, while behavioral traits include gait, keystroke, signature. A biometric trait needs to satisfy the certain conditions [2], such as *Universality*: each subject should have the trait; *Distinctiveness*: a subject should be separated from other subjects by the trait; *Permanence*: the trait should be invariant for a subject over a certain period of time; *Collectability*: the trait can be measured or calculated quantitatively. Besides, practical biometric systems have to manage the issues of performance, acceptability, circumvention, data security, and privacy.

Person authentication using biometrics is becoming increasingly important for various applications, such as access control, visual surveillance, law enforcement, and forensics [3, 4]. A biometric system may have either verification and/or identification mode, depending on the application scenario. Regarding verification, the biometric system validates the claimed identity for a subject by comparing the biometric query sample (i.e., probe) against the biometric trait associated with the claimed identity stored in a database (i.e., gallery), that is also known as so-called one-to-one matching. Particularly, verification has widely been used in physical access control, cellular phone, forensic analysis. By contrast, in identification, the biometric system recognizes a subject by comparing all of the stored biometric samples in the database, which is also known as a so-called one-to-many matching. It is used more frequently in the applications of forensic analysis, missing children's identification. An illustration of a biometric verification and identification system is shown in Fig. 1.1.

Gait biometrics can be defined as recognizing people (i.e., gait recognition) by the way they walk. Initially, Johansson [5] showed that a subject could be recognized with different types

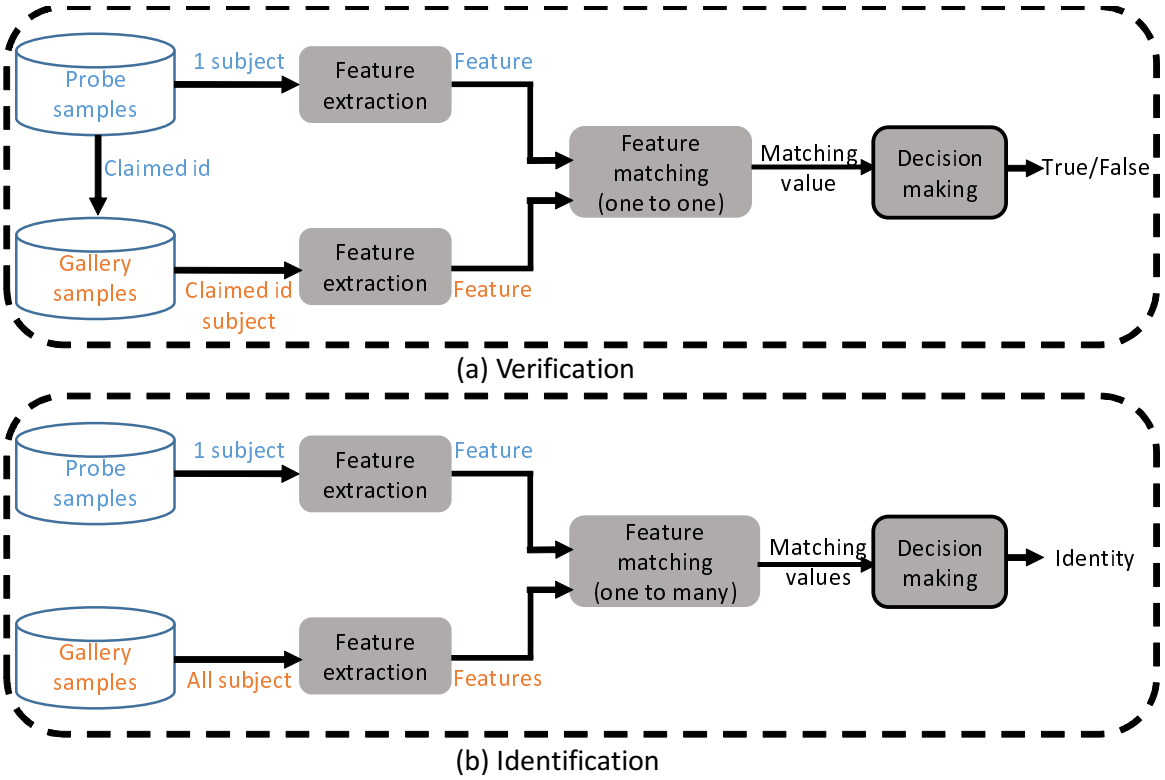


Figure 1.1: The framework of a biometric system: (a) the verification; (b) the identification.

of biological activities (e.g., walking, running, bicycling, dancing) by observing the motion of Moving Light Displays which are attached to the human body joints. Afterward, Cutting and Kozlowski [6] modified this capability to showed that a person could be recognized by their gait. Later, gait recognition is one of the topics of active interest in the biometric research community because it provides some unique advantages over other biometric features, such as the face, iris, and fingerprints. For example, it can be captured without the subject's cooperation at a distance and has a discriminative capability from relatively low-resolution image sequences from CCTV footage [7]. Since the gait image sequence can be captured without subject cooperation, therefore, it is hard to conceal or disguise. Although gait analysis for human recognition is not yet as mature as fingerprint, iris, or face, it can be a useful biometric tool to conviction. Recently, gait has been used as a forensic feature, and there has already been a conviction produced by gait analysis, for example, gait from image sequences has been recognized as a piece of valuable evidence for convictions in criminal cases (a burglary case in the U.K. [3]).

1.2 Performance metrics

As mentioned in the previous section, a biometric system can operate in verification and identification modes. For verification, there are different types of errors concerning with a biometric system. The widely used terms are the False Accept Rate (FAR) and False Reject Rate (FRR). The FAR is the probability of falsely accepting an imposter, while the FRR is the probability of falsely rejecting a true (genuine) user. The FAR and FRR are calculated using the imposter and genuine scores distribution with a threshold value. The FAR and FRR can be calculated as follows:

$$FAR = \frac{\text{Number of attempts for accepted imposter}}{\text{Number of total imposter attempts}}, \tag{1.1}$$

$$FRR = \frac{\text{Number of attempts for rejected genuine}}{\text{Number of total genuine attempts}}. \tag{1.2}$$

An illustration of the genuine and imposter score distribution with FAR, and FRR is shown in Fig. 1.2. A more specific measure for verification can also be used for biometric performance evaluation. For example, the equal error rate (EER) when the value of FAR and FRR are equal based on a threshold value, that is $EER = FAR = FRR$, while the half total error rate (HTER) can be calculated as $HTER = \frac{1}{2}(FAR + FRR)$. A graphical plot for FAR and FRR is called the receiver operating characteristic curve (ROC) to indicates the trade-off between the FRR of genuine samples and FAR of imposter samples with varying thresholds.

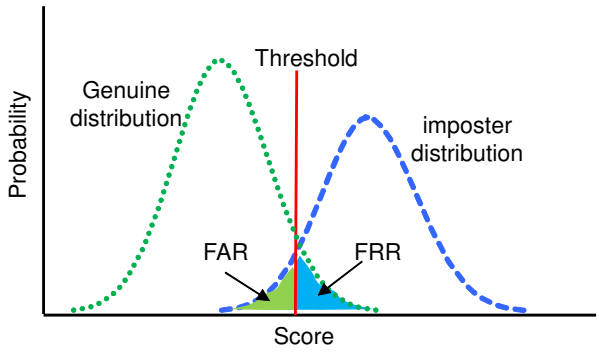


Figure 1.2: An illustration of the FAR and FRR with the genuine and imposter score distribution.

Regarding identification, a cumulative matching curve (CMC) is used. The CMC is a graphical plot of correct identification with rank. Moreover, a more specific measure, for example, the Rank-1, can also be used to measure the error in identification mode. Example of curves for CMC and ROC is shown in Fig. 1.3.

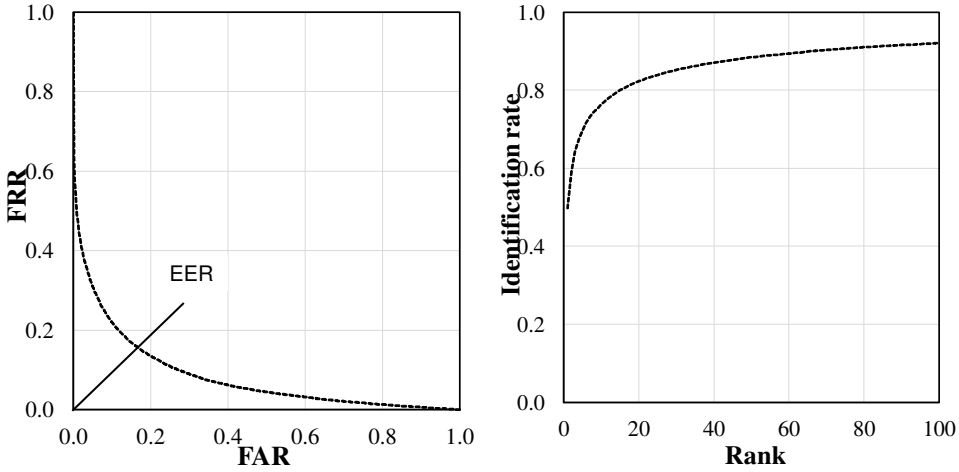


Figure 1.3: A typical ROC (left) and CMC (right) curves.

1.3 Gait feature representation

In the last few decades, various gait recognition approaches have been studied, and they can be divided into two main categories: model-based and model-free approach.

1.3.1 Model-based approach

In a model-based approach, a model is constructed from various motion cures or other parameters such as the shape of a person from the captured image sequence and/or corresponding silhouette sequence (i.e., background-subtracted binary image sequence), or skeleton data acquired from a depth sensor.

Some approaches developed models considering partial body regions. For example, Cunado et al. [8] extracted the periodical angular information from the upper leg using Fourier analysis. In contrast, Bouchrika et al. [9] exploited the information from both the knee and hip during a gait cycle at different phases. In contrast, most of the model-based approaches considered the whole body region for a person. For example, approach in [10] employed the stick-like model where sticks are connected with articulated body joints, whereas approaches in [11] divided the person body into elliptical regions and analyzed the parameters of the fitted ellipses.

In addition to the above mentioned image-based approaches, human skeleton data can also be employed for gait recognition that is directly acquired from any depth sensor, such as Microsoft Kinect or by using pose estimation algorithms [12]. For example, approaches in [13, 14] extracted static and dynamic, or relative geometric features from skeleton data of Kinect, whereas approach in [15] employed body joint heatmap that was extracted by a pose

estimation algorithm. Although model-based approaches seem to view-invariant, it is challenging to locate body joints due to body parts are occluded by the carried object(CO) or other parts of the person's body during walking or even occluded by other objects. Moreover, these types of approaches are computationally expensive and require a high spatial resolution (SR) image sequence. As for the skeleton data from the Kinect sensor, it can operate only on a limited range.

1.3.2 Appearance-based approach

The appearance-based approach represents gait features by directly analyzing the shape and motion of a person's image sequence and/or corresponding silhouette sequence without constructing a model. Thus, they are widely used in gait recognition due to its low computational cost and robustness to the noise. As a result, a vast number of approaches in this category have been proposed to date from a direct comparison of silhouettes to energy-based approaches. As for the direct comparison of silhouettes, Sarkar et al. [16] proposed a baseline algorithm using the spatial-temporal correlation between time-normalized silhouettes for a probe and gallery, whereas the approach in [17] used the key frames from a sequence.

Regarding the energy-based approach, Hoffman et al. [18] proposed an approach to extract gradient histogram in an image sequence and averaged them, which is called gradient histogram energy images (GHEI). Bobick et al. [19] proposed Motion Energy Image (MEI) and Motion History Image (MHI), where MEI represents the location of motion in a sequence by differentiating silhouette images, and MHI represents the history of motion at each pixel. Although MHI can capture the motion, it suffers to discriminate the motion direction due to self-occlusion (i.e., the person body part is occluded with other parts). Later, a simple but very effective approach was proposed by Han et al. [20] to average the silhouette sequence to accumulate the shape and motion into a single energy image, which is called gait energy image (GEI). Given a silhouette sequence, a GEI can be calculated by the following formula:

$$G(a,b) = \frac{1}{N}\sum_{t=1}^{N} I_t(a,b), \tag{1.3}$$

where I_t is a silhouette with t frame index, a and b are pixel values in the 2D image coordinate and N is the number of silhouettes in the gait cycle, where a gait cycle covers two strides: the left foot forward and the right foot forward.

It is reported in the literature that gait recognition from the silhouette-based appearance features (e.g., GEI) has shown outstanding performance in a controlled environment with normal walking pattern [21, 22] for publicly available dataset [23]. Although the silhouette-based

appearance approaches are sensitive to carried object (CO), view variation, and occlusion, they are invariant to color, texture and capable of discriminating in a very low SR image sequence [7].

Silhouette-based gait recognition based on GEI follows the basic five steps of operation, as illustrated in Fig 1.4. In the first step, the video data of moving subjects is captured by a CCTV camera. In the second step, the binary images of a moving subject are extracted by background subtraction or segmentation technique. In the third step, the height and the size normalization are performed; a gait cycle is calculated from this normalized silhouette because the gait of a person is a periodic activity. Then, the GEI feature is extracted from the normalized silhouette sequence corresponding to a gait cycle. In the last step, the extracted GEI is compared with the stored database.

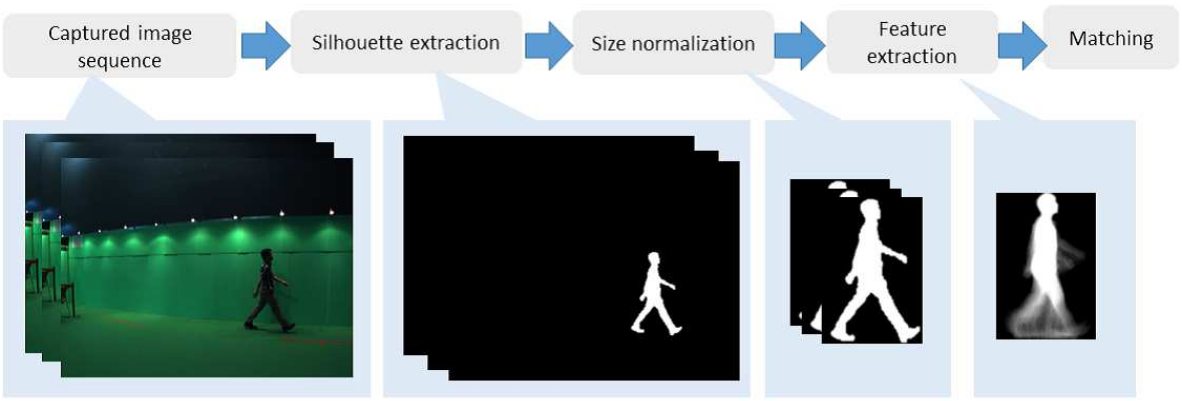


Figure 1.4: A general flow diagram for silhouette-based gait recognition.

1.4 Challenge in gait recognition

Already mentioned in the previous section, silhouette-based gait recognition achieved outstanding performance in a controlled environment. However, these results significantly degrade in the real-world scenario because the person walking patterns and/or captured image sequences for gait recognition change depending on multiple issues, and the change becomes large with the unconstrained environment. Therefore, the walking pattern between the probe and gallery gait silhouette images are not always the same. These issues can be related to the following categories:

1. related to environmental context,

2. related to the external factor,

3. related to the internal factor,

4. related to the camera setting.

Examples illustrating these issues that affect gait (i.e., the walking pattern for a subject) and/or captured image sequence for gait recognition are shown in Fig. 1.5. The environmental factor can change the walking pattern for a subject and/or affect the captured image sequence for recognition. For example, the walking surface [16] changes the walking pattern, whereas sudden lighting or illumination changes affect the captured image sequence; and a subject is not clearly visible due to occlusion [24]. Regarding external factors to a subject can influence the gait and captured image sequence for a person, which is usually referred to as covariate in the gait recognition. Some examples of covariate are carried object [25], clothing [26], and shoes [16]. As for the internal factors to a subject, such as physiological changes (e.g., aging, and pregnancy) or sickness (e.g., foot injury) change the usual walking pattern; some other examples for the internal factors are walking speed [27], and age [28]. The factors related to camera setting includes the relative position of the camera to the target person that impact the size of the target person in the image, the observed view of a person [23, 29] from which the person is observed; and the camera capturing setting such as spatial resolution (SR) and temporal resolution (TR) [30].

Among the factors mentioned above, the occlusion [24, 31], and carried object (CO) [24, 31] are most common and frequently occur factors in the real-world scenario. Besides, the quality of the captured image sequence due to the camera setting and/or relative distance of the person to the camera is not always the same. As a result of these issues, the obtained silhouettes from the corresponding captured images can be degraded, and we can define the silhouettes either well-posed or ill-posed based on the visual quality. Well-posed silhouettes mean that the obtained silhouettes are of good quality whereas ill-posed silhouettes are degraded, problematic, and low-quality silhouettes. More specifically, the presence of occlusion in an image sequence obscures a significant amount of the human body. Therefore, some parts of the human body are missing into the obtained silhouette, and it leads to the degraded silhouette (i.e., ill-posed silhouette). On the other hand, COs silhouette is added to the person silhouette, after obtaining the silhouette of a person with COs and separation of these are difficult; thus, the extracted silhouettes from a walking person with COs become ill-posed. Similarly, the observed appearance of a person differs depending on the observation view [32, 33], and gait features are

Figure 1.5: Common challenges in gait recognition. The face is masked due to privacy.

affected by spatial resolution and temporal resolution [7, 30]. Therefore, the silhouette leads to ill-posed.

Furthermore, the occlusions in the real-world may occur for several reasons, such as, with static obstacle objects (e.g., pillar, bench, tree); or with a dynamic object (e.g., a car, dog, and another walking person). Therefore, the occluded position and portion of the person's body differs depending on the occluder objects' properties such as shape, size, and motion; this leads to gait recognition more difficult. Regarding CO, people usually carried objects in their daily life with various shapes and sizes. Examples for the daily-life COs are books, umbrella, handbag, backpack, luggage, and travel bag. Some of them impact the pose and motion for a walking person while others change the silhouette. Therefore, gait recognition against CO with unconstrained shape and size becomes more difficult.

The main challenge for gait recognition is to make a robust approach that is invariant to these ill-posed silhouettes and realize in the real-world scenario. Then, it would be a significant contribution to gait biometrics and other associated communities such as forensics and medical applications.

1.5 Motivation

Extensive researches have already been conducted to realize silhouettes-based gait recognition against the factors mentioned in the previous section. For example, observation view in [22, 33, 34], carried object in [35, 36] and occlusion in [24]. Nonetheless, there are still some significant key challenges that remained unsolved for silhouette-based gait recognition. For example, occlusion issues have not been solved; the CO issue is explored by considering a very few predefined CO; a very few approaches tackled the quality of the captured image sequence from a camera. However, these issues are most influential and frequently observed in a real-world scenario, and therefore, the application of gait recognition is hindered in real-life. Besides, all of the existing approaches were developed based on a limited number of training subjects and also tested a probe subject against a small number of gallery subjects (e.g., a few hundred subjects). However, the identification rate significantly drops when the number of subject increase in a gallery.

Motivated by these unresolved issues and limitations for gait recognition, particularly for the real-world scenario where extracted silhouettes are ill-posed, we first look into the important issue the diversity and the number of subjects for a database to develop an approach for gait recognition during training and also testing. Furthermore, we considered the most common and frequently occurred issue in daily-life the carried object(s) (COs). Then, we tackle another challenging environmental factor in the real-world scenario, the occlusion for gait recognition. Finally, we look forward to the quality of the captured data, i.e., the spatial resolution and temporal resolution along with an observation view.

1.6 Contributions

We address the gait recognition using the following major ill-posed silhouettes that are related to three issues from different categories of factors in this thesis. The contributions of this thesis are summarized as follows:

1. **Construct an extremely large scale database with an unconstrained variation of CO covariate, and analyze its performance**

   With the growing data science trend, we always need a large-scale dataset to solve a problem efficiently. Recently, many sophisticated machine learning techniques, such as deep learning (DL), have been developed, and they require a large number of training samples because more data are more important than a better algorithm [37]. Though, several gait databases have been constructed to date with or without considering CO covariate [38, 39, 40, 41]. Although these datasets for gait recognition seem to be sufficient for a conventional machine learning algorithm (e.g., without DL), they are not sufficiently large to efficiently conduct a study using a DL-based approach.

   In this study, we first construct an extremely large population gait database with an unconstrained variation of CO that will encourage the gait recognition community to research this practical covariate deeply. The constructed database is the largest gait database in the world and included 62,528 subjects with an equal distribution of males and females with a wide range of ages. To the best of our knowledge, it is more than six times the size of the existing largest dataset for gait recognition without covariate and 200 times with CO covariate. There is no constraint on the type, quantity, and position of the CO. We considered some real-life COs that are used in daily life (e.g., handbag, vanity bag, book, notepad, and umbrella) or when traveling (e.g., backpack, luggage, and travel bag). Additionally, the typical position labels of the COs are manually annotated to analyze gait recognition with unconstrained COs in a different position. Second, we analyze the performance for gait recognition by considering a set of experiments using existing state-of-the-art appearance-based gait representation. Additionally, we analyze the classification and gait recognition difficulty concerning these manually annotated typical position labels.

2. **Spatio-temporal silhouette sequence reconstruction with occlusion using deep learning for gait recognition**

   In this study, we explore the deep learning-based approaches and adopt them for gait recognition under the challenging environmental factor associated with occlusion. Recently, deep convolutional neural networks (CNNs) and generative adversarial networks (GANs) [43] are employed in many research areas of computer vision and biometrics.

The outstanding capacity of GAN in modeling data distribution has significantly advanced many reconstruction problems as a conditional GAN (CGAN) such as inpainting, image, or video completion [44, 45, 47, 48, 49, 50, 52].

We proposed an effective feed-forward conditional deep generative network for silhouette sequence reconstruction. To make the silhouette sequence reconstruction well constrained, we regularize the training process by incorporating triplet hinge loss into the so-called Wasserstein GAN (WGAN) loss [53, 54] as adversarial loss and reconstruction loss in pixel space. To the best of our knowledge, this is the first adversarial loss that supervises the generator network during training by incorporating pairwise similarity ranking information, and the entire network is trained end to end with the reconstruction and proposed adversarial losses.

Compared with existing reconstruction-based approaches, one of the significant advantages of our proposed approach is that it does not require occluded position information as input for reconstruction. Therefore, it can be applied to an arbitrarily structured occluded silhouette sequence during reconstruction. Another advantage is that we can reconstruct the silhouette sequence without knowing the gait cycle in advance, while the existing approaches necessitate gait cycle information for reconstruction. Finally, we demonstrate the stability of the proposed approach to reconstruct the silhouette sequence by designing a set of experiments and present superior results for gait recognition compared with the state-of-the-art methods.

3. **Multi-quality and multi-modal biometrics**

In this study, we explore another challenging issue for gait recognition related to the quality of the captured image sequence. Particularly, we analyze the camera-based qualities of the spatial resolution (SR) and temporal resolution (TR) along with the quality of the observation view. We already defined gait in Section 1.1 as a manner of walking for a person, and the gait feature is derived from the shape and walking motion corresponding to a gait cycle. Therefore, gait recognition accuracy is influenced by the quality of SR and TR for the silhouette sequence. A usual solution to suppress the impact of degraded accuracy caused by different quality is to add additional complementary information [56]. For example, face modality can be one of the complementary information combined with gait modality because, as a dynamic feature, gait suffers low TR while

face still works cause of static feature. Moreover, uni-modal biometrics usually suffers some practical problems such as poor recognition performance and vulnerability to spoofing attacks [57]. For a solution to these problems, the system may use multiple biometric traits for recognition. A wide variety of biometric modality combinations are available for recognition with gait and face [55, 56, 58]. Besides, the auxiliary quality information or quality measures that are associated with the quality of samples affect the system's discrimination capabilities, even though they do not provide the capability to identify the subject by themselves. To use multiple modalities, score-level fusion is the most popular approach because it offers reasonable recognition accuracy and comparatively simple implementation [59, 60, 61, 62, 63, 64], and therefore, we considered score-level fusion in this research.

To analyze the camera-based qualities of the SR, TR, and observed view, there are no matching score databases that exist in the research community. Therefore, in this work, we first constructed an extremely large-scale matching score database from gait and head modality along with soft biometric height, including three types of qualities (SR, TR, and view). Secondly, we analyze how quality measures affect the recognition accuracy of each modality. Finally, we fuse the modalities in the score-level, and evaluate in two protocols that are the quality dependent and independent manner; then analyze the performance using a wide variety of state-of-the-art score-level fusion approaches.

Chapter 2

Literature review

2.1 Introduction

The research for vision-based gait recognition started with a very small dataset with tens of subjects showed that a high recognition rate could be achieved [8]. Later, a bit larger gait datasets, including a more or less 100 subjects were captured considering a wide range of factors (e.g., clothing, CO, occlusion, view, surface, shoes, and elapsed time) [16, 65].

In this chapter, a survey of literature related to the contributions made in this thesis is reported, which is the detailed descriptions of significant gait recognition datasets and gait recognition approaches to tackle the carried object (CO) and occlusion, fusing gait biometric with face and height. Moreover, we also describe the state-of-the-art deep learning-based approaches for image and video reconstruction for gait recognition.

2.2 Gait recognition databases

Several scientific institutions have been capturing and developing their gait datasets to compare and evaluate the performance of gait recognition fairly. This section will introduce existing major gait datasets and summarize them in Table 2.1.

1. **The USF dataset [16]:** This dataset is considered one of the most widely used gait dataset and captured outdoors under different walking conditions considering various environmental and external factors (e.g., view, surface, shoes, CO, elapsed time). This dataset consists of persons walking in elliptical paths in front of two cameras (i.e., left and right) and composed of 122 subjects. This dataset considers a briefcase as a CO, and as a result, at most two options for samples (i.e., with or without a CO) are available.

2. **The SOTON datasets [25, 65]:** These datasets are composed of a small population dataset and a large population dataset. The small database was created to probe the robustness of gait recognition in various covariate conditions (i.e., CO, cloth, and shoe). It contains subjects walking around an indoor track, with each subject filmed wearing a variety of COs, cloth, and shoes. It considers three types of bags (i.e., barrel bag, handbag, and rucksack) as COs, and the subject carries these bags in four ways (i.e., barrel bag slung over the shoulder or carried by hand on the shoulder, handbag held in hand, and rucksack). Because this dataset contains a larger variety of CO covariates than that of the USF dataset, it can be used for exploratory CO covariate analysis for gait recognition [67]. The other dataset, the large population datasets specially designed for the study of view-invariant gait recognition. Although it was considered a large population dataset, it includes only 115 subjects, that are not suitable for modern machine learning.

3. **The TUM-IITKGP [41]:** This dataset contains unique environmental factors, such as dynamic and static occlusion, along with CO and clothing variation. They defined, dynamic occlusion means when a person is occluded by other walking persons coming from the opposite direction; on the contrary, the static occlusion means when a person is occluded by a standing person. It is constructed from 35 subjects, and therefore, this dataset can be used as an initial benchmark for occlusion-handling gait recognition. Sample with both occluded and non-occluded situation is captured. Thus, it can be used for exploratory analysis for different factors for gait recognition.

4. **The TUM-GAID [40]:** This dataset is the first multi-signal gait dataset to contain audio signals, RGB images, and depth images by Microsoft Kinect from 305 subjects. Particularly, it was designed to evaluate the robustness of covariate factors like CO and shoe, and elapsed time.

5. **CASIA dataset B [38]:** This dataset is an extensive multi-view gait database that contains 11 views from the front view to rear view for an 18-degree interval, and constructed from 124 subjects. Besides, they considered CO along with clothing variation. Before capturing the sequences with a CO, each subject choses a bag from a set of the knapsack, satchel, or handbag that he/she liked. As a result, there are at most four options of samples available regarding COs (no bag, knapsack, satchel, and handbag).

6. **CASIA dataset C [39]:** This dataset considers only a backpack as a CO, and data was captured from 153 subjects using a thermal infrared camera designed for the study of night gait recognition.

7. **FVG dataset C [42]:** This dataset is constructed from 226 subjects and considered the walking speeds, CO, clothing, background, and time elapses along with various variations of the front view. Although this dataset considered different internal, external factors, it included only 226 subjects with a fixed type of bag as CO.

8. **The OU-ISIR, Large Population datasets [23, 29]:** These datasets are composed of limited view variations and large view variations with large population datasets. The OU-ISIR, Large Population datasets (OULP) contains the largest number of subjects (4007) with limited view variations from 55 to 85 degrees for a 15-degree interval. Later, the OU-ISIR, Multi-View (OUMV) dataset [29] is constructed by considering large view variation (14 views for an interval of 15-degree) with 10,307 subject. Although the number of subjects variation for these datasets is high compare to other datasets, there is no covariate in these datasets. Therefore, it is not usable for studying a covariate. Samples gait images for OULP and OUMV dataset are shown in Fig. 2.1.

(a) OULP dataset

(b) OUMV dataset

Figure 2.1: Sample gait images for the OULP (top) [23] and OUMV (bottom) [29] datasets. The face is masked due to privacy.

Table 2.1: Existing major gait recognition databases

Database	#Sub.	Convariates	Types of CO	#Options for CO positions
SOTON small [25]	12	multi view, CO, shoe, clothing	Handbag, barrel bag, rucksack	Four
USF [16]	122	multi view, CO, surface, shoe, time	Briefcase	One
CASIA, B [38]	124	multi view, CO, clothing	Knapsack, satchel, handbag	Three
CASIA, C [39]	153	view, CO	Bag	One
Mobo [68]	25	multi view, CO, surface, Speed	Ball	One
TUM-IITKGP [41]	35	single view, clothing, CO	Backpack	One
TUM-GAID [40]	305	single view, CO, time	Backpack	One
FVG [42]	226	single view, clothing, CO, speed	Bag	One
OU-ISIR, LP [23]	4,007	multi view	N/A	N/A
Proposed [116]	**62,528**	Single view, CO	**Unconstrained**	**Unconstrained**

2.3 Gait recognition approaches

2.3.1 Approaches to tackle with carried object

As mentioned in section 1.3, the gait energy image (GEI) [20] is the most prevalent and frequently used feature for gait recognition [21, 22, 23, 73, 74, 75]; however, this feature is sensitive to CO. To mitigate this limitation, some modified GEIs have been introduced. For example, Bashir et al. proposed Gait Entropy Image (GEnI) [69] and Masked GEI (MGEI) [35]. GEnI is computed by calculating the Shannon entropy for every pixel of the GEI to enhance the dynamic information while attenuating the static information for gait recognition. Given a GEI image $g = G(a,b)$, GEnI is calculated by the following formula:

$$GEnI = -g * \log_2 g - (1-g) * \log_2 (1-g). \tag{2.1}$$

The gait energies are masked out when entropy is smaller than a certain threshold for MGEI. Furthermore, some approaches employed Gabor filter with GEI for gait recognition against CO; for example, Tao et al. [70] directly applied Gabor filter to GEI, whereas the approach in [36] employed on transformed GEI [36].

Appearance-based features, however, often suffer from large intra-subject appearance changes because of COs. To gain more robustness, the most popular way is to incorporate spatial metric learning-based approaches, such as linear discriminant analysis (LDA) [71] and a ranking support vector machine (RankSVM) [72]. Additionally, with the great successes of employing DL-based approaches in face recognition and many other areas of computer vision, it is started to use in gait recognition. In particular, CNN-based approaches which consider spatial proximity using a convolution operation, and significantly improves the accuracy. For example, Wu

et al. [21] employed CNN for gait recognition, and demonstrated better recognition accuracy than those of benchmarks with cross-view and also CO; Li et al. [76] proposed a CNN-based approach and demonstrated its effectiveness for gait recognition with CO, and it outperforms on other benchmarks.

2.3.2 Approaches to tackle with occlusion

In this section, we review some works related to gait recognition to occlusions as two categories: reconstruction-free approaches and reconstruction-based approaches.

Regarding reconstruction-free approaches, Zhao et al. [78] extracted features based on fractal scale wavelet analysis for each silhouette from a sequence of a gait cycle and then averaged them. They evaluated their approach against occluded data by adding a vertical bar in the silhouette sequence. Chen et al. [79] extracted features from the frame difference energy image (FDEI) representation to mitigate the problem of spatial and temporal silhouette incompletion caused by imperfect silhouette segmentation and occlusion. The FDEI representation follows the four steps: in the first step silhouette sequence of a gait cycle is divided into clusters, and the GEI feature is calculated for each cluster; in the second step denoise it as:

$$D_c(a,b) = \begin{cases} G_c(a,b), & \text{if } G_c(a,b) \geq T \\ 0, & \text{otherwise} \end{cases} \tag{2.2}$$

where $G_c(a,b)$ is the GEI for c_{th} cluster, and T is the threshold. The denoised image is called the dominant energy image (DEI). In the third step, the positive difference of a silhouette ($I_t(a,b)$) from the previous one ($I_{t-1}(a,b)$) is calculated as:

$$F_t(a,b) = \begin{cases} 0, & \text{if } I_t(a,b) \geq I_{t-1}(a,b) \\ I_{t-1}(a,b) - I_t(a,b), & \text{otherwise} \end{cases} \tag{2.3}$$

where t is the frame number, in the last step, FDEI representation is computed as the summation of a silhouette difference $F_t(a,b)$ with its corresponding cluster's DEI $D_c(a,b)$.

Ortells et al. [80] proposed a statistical framework to minimize the influence of silhouette defects. The authors evaluated gait recognition on GEIs and gradient histogram energy images (GHEI) by adding artificial occlusion and noise into a silhouette sequence. A different technique to handle the problem of occlusion was addressed in [81], in which a GEI was separated into four modules, and a module was excluded for gait recognition if occlusion was identified.

Regarding reconstruction-based approaches, Roy et al. [24] proposed an approach in which a silhouette sequence was first divided into a few subsequences of the gait cycle(s) based on

key poses. It first identified whether a silhouette of a gait cycle was occluded. Then, the occluded silhouettes were then reconstructed using a balanced Gaussian process dynamical model (BGPDM). The effectiveness of the proposed approach was evaluated based on reconstruction accuracy; they, however, did not evaluate gait recognition using the reconstructed silhouette sequence. Hofmann et al. [31] proposed a very simple method to detect partially occluded gait subsequences from a sequence using the number of foreground pixels. The occluded silhouettes were then replaced by similar-pose clean silhouettes from other gait cycles. In [82], a complete GEI was regenerated from a partially observable GEI using the subspace-based method. Gait recognition was evaluated according to whether a matching pair did not share a common visible region.

2.4 Deep generative approach for image and video reconstruction

The purpose of the discriminator network is to distinguish whether the content is generated by a generator network or is real. In contrast, the generator network is trained to fool the discriminator network. Especially, G and D are trained by solving the minimax problem as:

$$\min_G \max_D \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} [\log D(x)] + \mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g} [\log (1 - D(G(z)))], \tag{2.4}$$

where $\mathbb{E}(\cdot)$ indicates the expectation operator, and $\mathbb{P}_r$ and $\mathbb{P}_g$ are the real and generated data distributions, respectively. Generator G transforms input sample z to mimic a real sample. However, one of the main issues of GAN is instability during training. Several works have addressed improving the training stability. Radford et al. [83] proposed deep convolutional GANs (DCGAN) that imposed empirical constraints on the architecture of the GAN and optimized the hyperparameters. More recently, Arjovsky et al. [53] introduced Wasserstein GAN (WGAN) [54], which minimizes the Earth-Mover distance (a.k.a Wasserstein-1) between the generator and real data distribution. Specifically, the objective function was constructed by applying the Kantorovich–Rubinstein duality:

$$\min_G \max_{D\in\mathscr{D}} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r} [D(x)] - \mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g} [D(G(z))], \tag{2.5}$$

where $\mathscr{D}$ is the set of 1-Lipschitz functions. To enforce the Lipschitz constraint on the critic function, Gulrajani et al. [54] introduced an improved version of WGAN with a gradient penalty term with respect to the input. The new objective are as follows:

$$\min_{G}\max_{D} \mathop{\mathbb{E}}_{x\sim\mathbb{P}_r}[D(x)] - \mathop{\mathbb{E}}_{G(z)\sim\mathbb{P}_g}[D(G(z))] + \lambda L_{GP}, \tag{2.6}$$

where $L_{GP} = \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right]$, $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}$, and λ is a gradient penalty coefficient and $\varepsilon \sim U[0,1]$. The authors called the auxiliary network a critic instead of discriminator because it is not a classifier.

Numerous works exist for image and video reconstruction in the literature from traditional approaches to learning-based approaches (i.e., deep learning). Traditional approaches include patch-based [84] and diffusion-based [85] techniques. The patch-based approach extracts patches from a source image and then pastes them into a target image. The patch-based techniques are also used for video completion [86] by replacing image patches with Spatio-temporal synthesis across frames, whereas the diffusion-based approach propagates the image appearance around the target position. However, these types of methods can only fill a minimal and homogeneous area, and one main limitation is the repetition of content.

Recently, conditional GAN-based [87] approaches have emerged as promising tools for image and video completion. Regarding image completion, a Context Encoder (CE) [44] was the first attempt to train deep neural networks for image completion. It is trained to complete the center region using pixel-wise reconstruction and single discriminator loss. Some conditional GAN-based approaches exist in the literature introduce two discriminators/critics [45, 47, 48] networks as adversarial losses, where one discriminator/critic considers the whole image while the other focuses on a target area to enforce local consistency. However, the main issue for these types of approaches is that they assume the occluded/inpainting position is known in advance during training and also testing. The generator takes the masked image as input and outputs the generated image. Finally, it replaces pixels in the non-masked region of the generated image with the original pixels. On the other hand, there are very few works in the literature for video completion. First, Vondrick et al. [88] introduced a generative video network for video generation. They also predicted the future frame using the DCGAN model [83] and Spatio-temporal three-dimensional (3D) convolutions [89, 90]. Afterward, Kratzwald et al. [52] improved the video generative network using WGAN with a gradient penalty critic network and extend it applied multiple applications.

2.5 Multi-modal biometric system

A system that combines the evidence from multiple sources of the biometric trait to reliably recognize an individual is known as multi-modal systems [91]. Multi-modal biometric systems

can alleviate many of the limitations of a uni-modal biometric system (i.e., systems that use a single biometric trait for recognition), which is commonly affected by practical problems such as non-universality, vulnerability to spoofing and poor recognition performance [57].

The most crucial concern in multi-modal biometrics is how to fuse these individual modalities. Fusion can be done at different levels, including raw sensors, extracted features, matching score, rank, and decision levels. In sensor-level fusion, the raw acquired sample from multiple sensors are combined (i.e., by *mosaicing*, where multiple individual 2D images are combined to generate a single image); in feature-level fusion, the extracted features from multiple modalities are concatenated into a single large feature vector; finally, it is used for classification; in score-level fusion, the multiple matching scores are fused into a single score, and a decision is then made on the basis of the fused score; in decision-level fusion, multiple binary outputs are merged to produce a final decision using a specified rule such as a majority voting rule. Among these methods, score-level fusion is the most popular approach because it offers reasonable recognition accuracy and comparatively simple implementation. Thus many researchers in the multi-modal research community work in the score-level fusion field [59, 60, 61, 62, 63, 64].

2.5.1 Existing score-level fusion database

At the beginning of multi-modal biometric, researchers developed fusion approaches using *chimeric datasets* because the capture of multi-modal data from the same subjects is a laborious and challenging task. Because data of different modalities are independent of each other, it believes that the properties of chimeric datasets are similar to those of real data sets. Therefore, multiple modality features that are captured from different subjects are used as features from a single virtual subject. For example, a face feature from one person and a fingerprint feature from another person are used as a multi-modal feature for a virtual subject [92]. Although chimeric datasets are accepted to some degree in score-level fusion research, however, Poh et al. demonstrated experimentally that the performance of a dataset of real multi-modal users is not equivalent to that of a database of chimeric users [93]. Moreover, Wayman mentioned that multi-modal data might necessarily be correlated [94], and therefore chimeric databases should be avoided to use.

Recently, several real multi-modal biometric databases and score databases have been released for research purposes. The major real multi-modal score databases that exist at present are summarized in Table 2.2. These databases are briefly described here.

The BioSecure DS2 score database [95] was constructed using the desktop scenario dataset of the BioSecure DS2 database [99] with face, fingerprint and iris modalities. Originally, the

Table 2.2: Existing major score databases.

Database	#Subjects	Modality	Quality
BioSecure DS2	333 [95]	Face, fingerprint and iris	Yes
BA-Fusion	295 [96]	Face and speech	No
BANCA	52 [97]	Face and speech	No
NIST-Multimodal	517 [98]	Face and fingerprint	No

desktop scenario dataset contained voice, face, signature, fingerprint, hand, and iris modalities, and data were acquired in a desktop-based office environment. A total of 333 subjects with equal male and female distributions are included in this score database, although the age distribution is somewhat biased; two thirds of the subject's ages are in the 18 - 40 years old range, while the others are over 40 years old. The main characteristic of the BioSecure DS2 score database is that it considers 14 quality measures for face modality; six of these measures are face-related quality measures, and these qualities are set on the basis of face detection (e.g., detection reliability, number of pixels between the eyes, face with or without glasses, rotation in the plane, rotation in depth, and degree of frontal face); the remaining eight measures are related to the image quality. Additionally, the database considers one quality measure for the fingerprint modality (e.g., texture richness), and three quality measures for the iris modality (e.g., texture richness, difference between the iris and pupil diameters, and proportion of the iris used for matching). In [95], multiple evaluations were reported, including quality-dependent, client-specific, and cost-sensitive fusion.

The BA-Fusion score database [96] was built from the XM2VTS database [100] using face and speech modalities. This score database is composed of eight matching scores; five of these scores are related to the face, while the remaining three are related to speech. For face score calculations, multiple feature extractors and multiple classifiers are used, while for speech scores, multiple feature extractors and only single classifiers are used. This score database is composed of scores from 295 subjects; both genders were included, but the age distributions are biased (in that the subjects are all adults). No quality measures were provided in this database.

The BANCA score database [97] is composed of face and speech modalities from 52 subjects in two groups. This score database was generated using a set of state-of-the-art baseline classifiers along with template-based approaches. While this score database does consider controlled (clean), adverse (under challenging conditions) and degraded scenarios, no quality measures are provided in this database.

The NIST-Multimodal score database [98] is composed of two face and two fingerprint scores from 517 subjects. Two fingerprint scores were obtained by comparing a pair of left

index finger and a pair of right index finger. Two face scores were generated using two separate face matchers. The numbers of generated genuine and imposter scores are 517 and 517 $\times$ 516=266,772, respectively. Again, no quality measures were provided in this database.

2.5.2 Score-level fusion approaches

In the score-level fusion of multi-biometric systems, the most important issue is how to fuse the scores of the different modalities. Many fusion techniques have been proposed in the literature to date. As shown in Table 2.3, these approaches can be classified into three generic categories: (1) transformation-based, (2) classification-based, and (3) probability density-based approaches.

Table 2.3: Fusion approaches at the various multi-modal biometric system levels.

Level of fusion	Fusion approach
Sensor-level fusion	Mosaicing
Feature-level fusion	Feature concatenation
Score-level fusion	Transformation-based approaches [59, 101, 102, 103]
	Classification-based approaches [63, 72, 104]
	Probability density-based approaches [60, 61, 62, 64, 105]
Decision-level fusion	Majority voting

In a transformation-based approach, the scores are usually normalized to a common domain by one of several normalization techniques (e.g., z-normalization [101], F-normalization [102], and EER-normalization [103]) and the normalized scores are then combined. Kittler et al. focused on classifier combination and developed a theoretical framework for classifier combination [59]. They used a sum rule, a product rule, a minimum rule, a maximum rule, a median rule, and majority voting as the basis for the classifier combination scheme.

In the classification-based approach, multiple scores that are derived from multiple matchers are treated as a feature vector and the classifier is then constructed to discriminate genuine scores from imposter scores. A support vector machine (SVM) is one such classifier and the signed distance from the decision boundary is usually regarded as a fused score [63, 104]. Additionally, the ranking SVM (RankSVM) [72] is a well-known extension of the conventional SVM that focuses more on the relative distance between two classes. RankSVM has been used in many research fields, including person re-identification and gait recognition [74]. Because ranking statistics play an important role in identification scenarios [106], RankSVM can be considered to be a promising approach to score-level fusion for identification scenarios.

The probability density-based approach is further divided with respect to two specific aspects: generative vs. discriminative approaches, and parametric vs. nonparametric approaches. Parametric and generative approaches explicitly or implicitly model the distributions for each class separately and subsequently estimate the model parameters from the training data. Nandakumar et al. proposed a framework for a likelihood ratio-based fusion rule and estimated the genuine and imposter distributions in the form of a finite Gaussian mixture model (GMM). It was shown that a likelihood ratio-based approach led to high performance for quality-based biometrics when using three multibiometric databases [62]. In contrast, discriminative approaches model the posterior probabilities directly. Linear logistics regression (LLR) combines discriminative and parametric approaches. In [61], the log likelihood ratio of genuine and imposter scores is expressed as a linear combination of these scores, and the weight of each modality score is then optimized to minimize the loss function that is derived from the logistic functions of the log likelihood ratio. The main limitation of these parametric approaches is that they can produce poor predictive performances if the chosen model does not fit the actual distribution well.

The nonparametric approaches represent distributions that use histogram bins or control points. Kernel density estimation (KDE) is one of generative and nonparametric approaches. Dass et al. proposed an approach that computed the generalized densities that were estimated from genuine and imposter training samples for each modality and combined them using product rule or copula methods [60]. For discriminative and nonparametric approaches, lattice-type control points are used to represent the Bayes error gradient distribution in a nonparametric manner [105]. In another approach, floating control points are used in conjunction with generalized Delauney triangulation for a more efficient representation [64]. In both methods, the estimation of individual genuine and imposter densities is bypassed and the discriminative function is trained directly.

2.5.3 Quality-dependent approaches

Quality can be considered to be auxiliary information that affects the matching score. Generally speaking, if biometric samples are of good quality, matching scores of the genuine and imposters are more easily separable. Therefore, quality measurement/assessment algorithms and recognition accuracy improvement using the quality measures, enjoy a large body of literatures in biometrics community [107, 108, 109].

Quality measures can be used at various stages in the recognition pipeline to improve the recognition accuracy. During the enrollment phase, a quality measure is used as the criterion for

sample recapture [110]. In the preprocessing phase, both quality-dependent feature enhancement and quality-dependent target region selection are considered [108]. In the matching phase, different matching algorithms are used to calculate the scores for uni-modal and multi-modal biometrics. In this phase, classifier or distance metrics are selected adaptively depending on the sample quality [111]. Another direction is to directly stack the quality measures into a score vector, i.e., Q-stack vector and to treat it as a feature vector for classification [112, 113]. Moreover, in [114, 115], biometric samples are classified into clusters based on the sample quality, and score normalization or fusion are done in a cluster-dependent way.

2.6 Conclusion

In this chapter, existing methods and dataset that are related to the work presented in this thesis have been reviewed.

From the analysis of the existing gait recognition datasets, it is noticed that existing datasets are unsuitable not only for studying CO covariates but also taking advantage of modern machine learning (e.g., DL) approaches. This motivated us to present a database that contains unconstrained variations of COs and the largest number of subjects, which is approximately 200 times larger than the largest existing gait database with COs, that is, TUM-GAID, and six times larger than that without COs for gait recognition, that is, the OUMV dataset. Additionally, we can observe that appearance-based features such as GEI and modification of GEI are dominant for gait recognition. To gain more robustness, the spatial metric learning-based method is incorporated. In this work, we analyze the gait recognition with CO covariate considering GEI features with state-of-the-art metric learning-based approaches.

From the survey on the gait recognition approaches with occlusion, we can observe that some approaches tackle occlusion directly on pre-processed feature GEI for a gait cycle. Therefore, they assume that the gait cycle is known in advance. The remaining approaches estimate the gait cycle from the occluded silhouette sequence, which is very difficult or error-prone when all frames are occluded in a sequence. Moreover, they consider a very large silhouette sequence where multiple gait cycles are available for gait recognition. However, many scenarios in real-world applications where only a few frames (i.e., not more than a gait cycle) are available in a sequence, and all are partially or totally occluded. In those scenarios, existing approaches are not applicable. In this thesis, we look into this challenging problem of gait recognition against occlusion. Reviewing the deep learning-based approach for image and video reconstructed, we can observe that conditional GAN has emerged as a promising approach for image and video

completion. This motivates us to present a novel approach to tackle gait recognition using a conditional generative network for silhouette sequence reconstruction.

After the literature reviewed of multi-modal biometrics, we can observe that fusion in score-level is the most popular and straightforward approach, and it offers reasonable recognition accuracy. Additionally, some auxiliary or quality measures can improve recognition accuracy, although they do not have discrimination ability. This motivates us to propose a score-level fusion database drawn from gait, head and height modalities and analyze the accuracy by designing a quality-independent and quality-dependent experiments.

Chapter 3

Construction of gait database with real-life carried object and its performance analysis

3.1 Introduction

The carried object (CO) is one of the most important and frequently occurred factors because people often need to carry objects in their daily lives, such as a handbag, briefcase on the way to work, or multiple bags after shopping. Some of them impact the pose and motion for a walking person while others change the silhouette depending on the location of COs are being carried. Therefore, extracted silhouettes of a sequence with unconstrained variations of COs being carried in unconstrained positions of the human body become ill-posed.

To address gait recognition with CO, first and foremost, a common gait database that considers the CO is essential. Based on the literature review in the chapter 2, there are some existing gait databases in the research community that consider COs, yet they have a certain number of limitations. For example, they contain a limited number of subjects with predefined COs, and lack of information about the positions and types of COs. Moreover, with the growing data science trend, we always need a large-scale dataset to solve a problem efficiently. Recently, many sophisticated machine learning techniques, such as deep learning (DL), have been developed, and they require a large number of training samples because more data are more important than a better algorithm [37]. To overcome the issues as mentioned earlier, in this chapter, we first propose an extremely large population gait database [116] with a large variation of CO covariate that will encourage the gait recognition community to research this practical covariate deeply. Second, we analyze gait recognition accuracy by employing state-of-the-art appearance-based gait representation.

3.2 Construction of gait database with real-life carried object

3.2.1 Capture system

The proposed database was constructed from sequence automatically collected by a gait collecting system named Gait Collector [117]. The gait data were collected together with an experience-based demonstration of video-based gait analysis in a science museum (Miraikan), and electronic informed consent was obtained for research. An overview of the data capture system is illustrated in Fig. 3.1. The camera was set at a distance of approximately 8 m from the straight walking path, and a height of approximately 5 m. The resolution and frame rate for the image sequence was 1280×980 pixels and 25 fps, respectively. Green background panels and carpet were arranged along the walking course for the extraction of the clear silhouette. The camera continuously captured video during the museum opening hours, photo-electronic sensors were used for detecting a subject walking past, and a sequence of a subject was extracted from the entire video sequence.

Each subject was asked to walk the predefined walking course three times at his/her preferred speed. First, the subject walked to the other side of the course with his/her COs, and then placed these COs into a storage box. Subsequently, he/she walked two more times without COs in the same direction and then picked up the COs from the storage box and left the walking course. Thus, we obtained three sequences for each subject. The first sequence with or without COs (if he/she did not have COs) is called the A_1 sequence in this work, and the second and third sequences without COs are called A_2 and A_3 sequences, respectively.

3.2.2 Gait feature generation

We used the GEI feature for gait recognition. To this end, a silhouette sequence of a subject was extracted using a chroma-key [118] (i.e., removal of the green background area using HSV color space). Then, registration and size normalization of the silhouette images were performed as follows: First, the subject's silhouette images were localized by detecting the top, bottom, and horizontal center (i.e., median) positions. Then, a moving-average filter was applied to smooth these positions. Finally, the sizes of the silhouette images for a subject were normalized according to the average positions so that his/her height was 128 pixels and maintained the aspect ratio. As a result, we generated the subject's silhouette images of 88 x 128 pixels. Finally, a gait cycle was determined using Normalized Auto Correlation (NAC) [23] of the subject's silhouette image sequence along the temporal axis for GEI feature generation using Eq. 1.3.

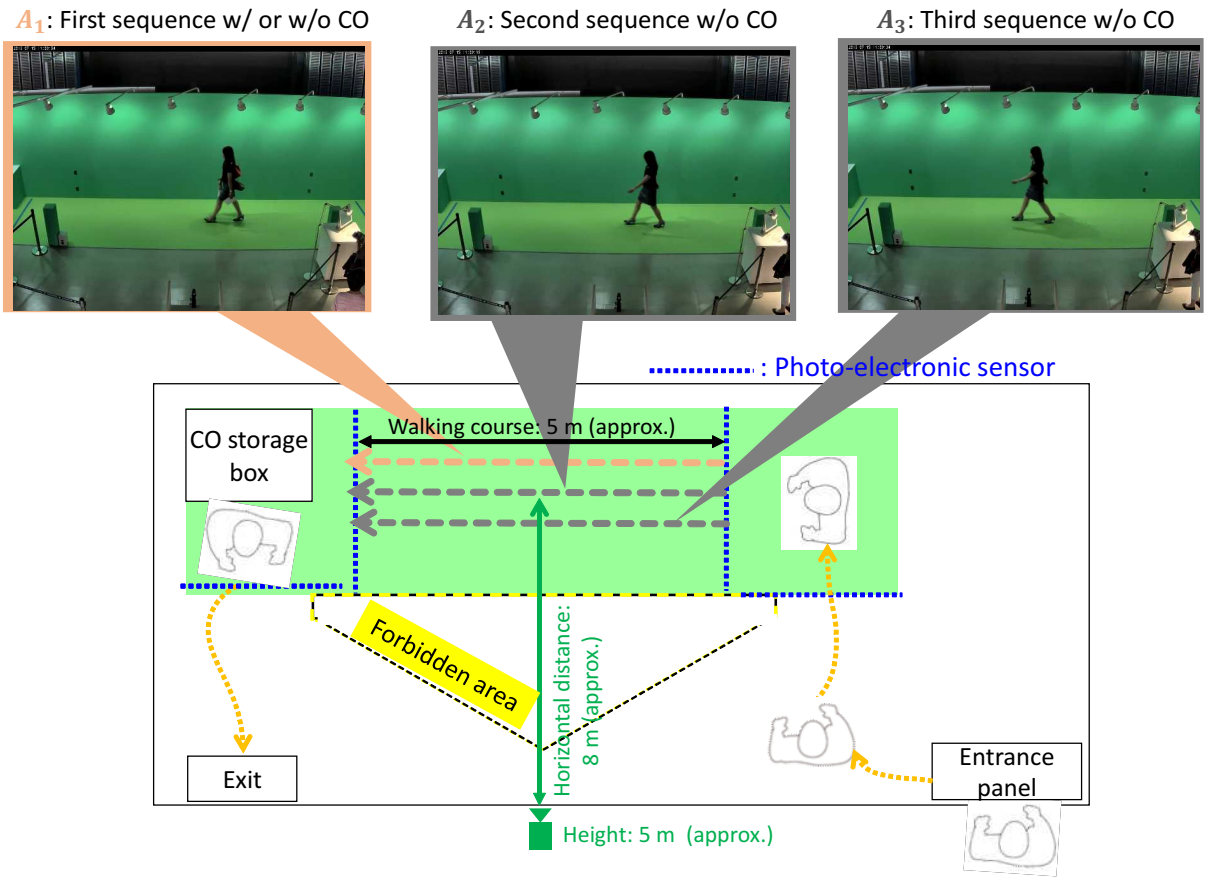


Figure 3.1: Illustration of the data collection system to capture gait images with real-life carried object.

3.2.3 Annotation of the carrying status

Because we did not constrain the subject regarding the type and number of COs, or where and how it was carried, therefore, it could be carried in a variety of positions and orientations as well as any number of COs. Thus, it was difficult and challenging to categorize the position accurately. For simplicity, we divided the area in which the COs could be carried into four regions concerning the human body: side bottom, side middle, front, and back, as shown in Fig. 3.2. However, some subjects did not carry any object, some carried multiple COs in multiple regions, and others changed a CO position within a gait cycle.

For each GEI, every fourth silhouette image within a gait cycle was manually checked to annotate the carrying status (CS). As a result, a total of seven distinct labels for the CS were annotated in our database. A summary of the denotation of the CS labels is explained in Table 3.1, and some examples of CS labels are shown in Fig. 3.3. It should be noted that the annotation process was only applied to the A_1 sequence for each subject because only the

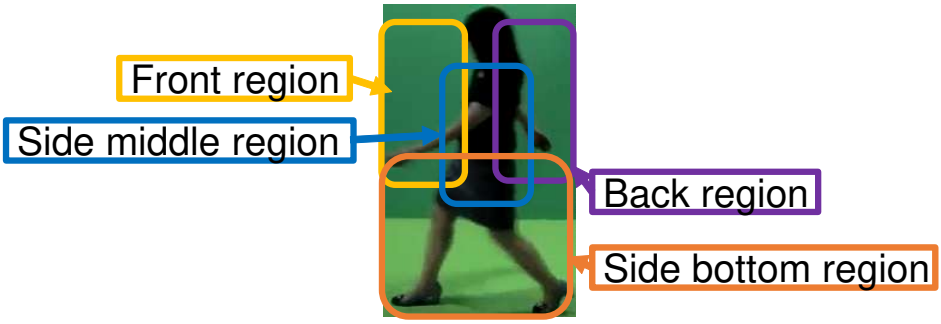


Figure 3.2: Four approximating regions for a person in which a carried object is being carried.

samples for the A_1 sequence may have contained COs.

Table 3.1: Carrying status label.

CS label	Explanation
NoCO	No carried object
SbCO	CO(s) being carried in the side bottom region
SmCO	CO(s) being carried in the side middle region
FrCO	CO(s) being carried in the front region
BaCO	CO(s) being carried in the back region
MuCO	COs being carried in multiple regions
CpCO	CO(s) with position being changed from one region to another within a gait period

3.2.4 Database statistics

With the good design of the system, the world's largest database for gait recognition with COs was constructed. It was composed of 62,528 subjects with ages ranging from 2 to 95 years, detailed distributions of the subjects' genders by age groups are shown in Fig. 3.4. We can see that the gender distribution is well-balanced for each age group, which is a desirable property for the comparison of gait recognition accuracy in terms of genders [119].

Improper GEIs were excluded from the final database for the following cases: (a) if a subject stopped walking for a while at the center of the walking course; (b) If a subject changed walking direction before the end of the walking course; (c) If a subject continued to carry COs in the A_2 and A_3 sequences; (d) If a subject exited from the walking course after finishing the first sequence, A_1. Thus, each subject had a maximum of three sequences. We, therefore constructed a database that included 60,450 subjects for the A_1 sequence, and 58,859 and 58,709 subjects for A_2 and A_3 sequences, respectively.

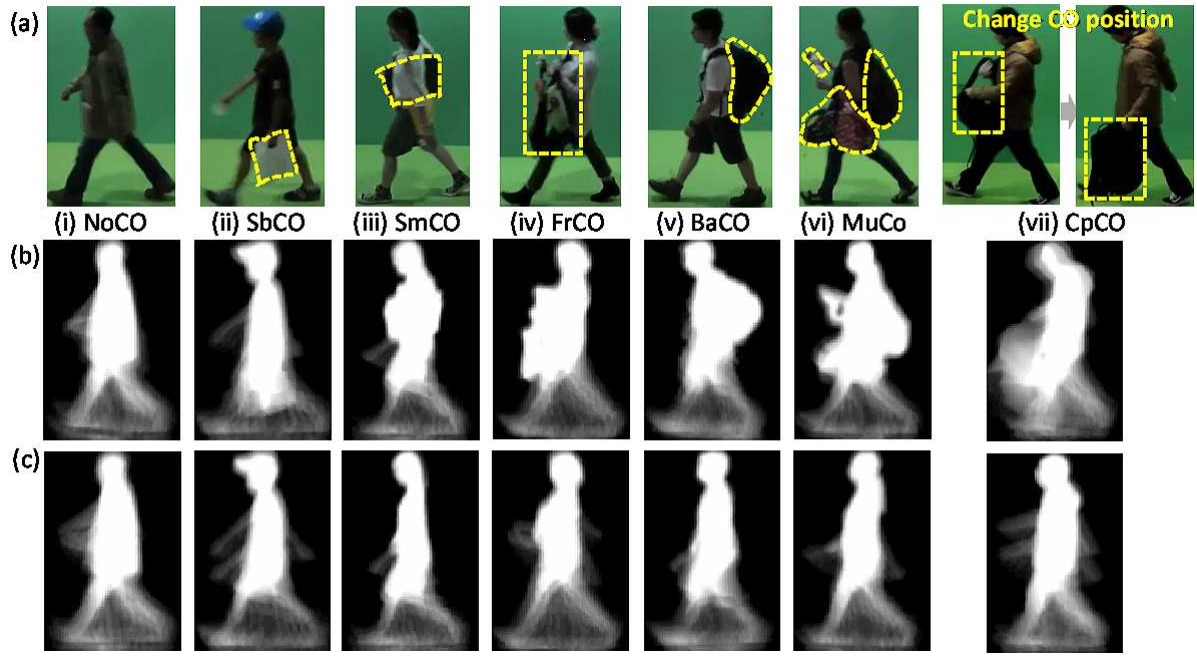


Figure 3.3: Examples of carrying status labels: (a) sample RGB image within a gait period with COs (circled in yellow) in their A_1 sequence; (b) corresponding GEI feature; (c) GEI feature of the same subject without a CO in another captured sequence (A_2 or A_3); for reference.

The distributions of the CS labels are shown in Fig. 3.5. We can see that, most of the subjects carried multiple COs in multiple regions (i.e., with MuCO) and the subjects liked to carry COs at the front (i.e., with FrCO), and back regions (i.e., with BaCO) along with the subjects did not like to carry COs (i.e., with NoCO) equally. Furthermore, we can also observe that few subjects liked to changed their CO positions from one region to another (i.e., with CpCO); similarly, few subjects carried COs in the side middle region. Meanwhile, the number of subjects who carried COs in the side bottom region (i.e., with SbCO) was approximately double of those who carried COs in the side middle region (i.e., with SmCO).

3.3 Experiments

3.3.1 Overview

In order to analyze the gait recognition with CO covariate, we considered a set of experiments on the proposed database. These experiments were designed to address a variety of challenges for gait recognition against COs and provided benchmark results for a competitive performance comparison of the state-of-the-art algorithms. Notably, we considered two sets of well-known experiments for gait recognition: *cooperative and uncooperative settings* and *impact of the*

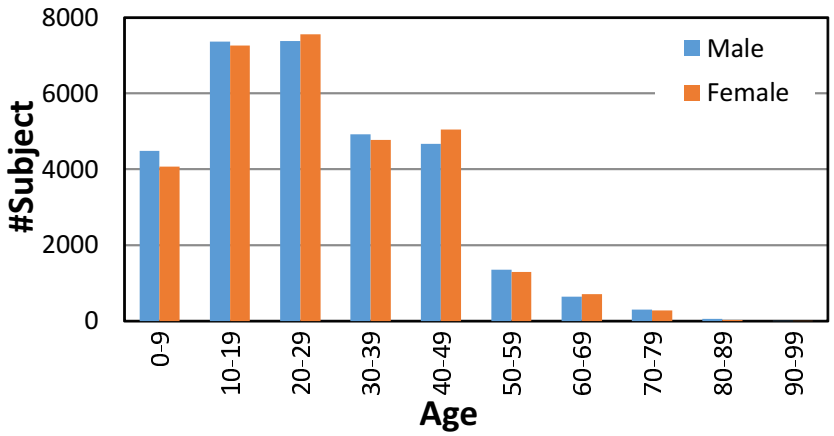


Figure 3.4: Distribution of genders by age group.

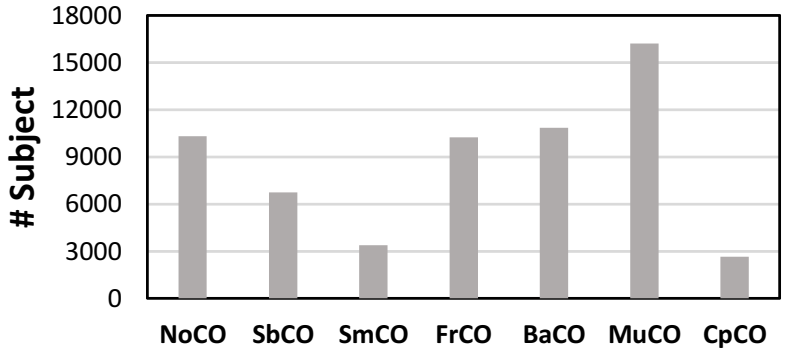


Figure 3.5: Distribution of the carrying status label.

number of training subjects. Additionally, we designed two more sets of original experimental settings to analyze the impact of COs: *difficulty level of the CS labels* and *classification of the CS labels*. To the best of our knowledge, they have not been studied before.

3.3.2 Benchmarks

There are several state-of-the-art appearance-based approached available for gait recognition in the literature, as discussed in subsection 2.3.1. We selected total seven benchmark approaches from the wide variety of approach to analyze the gait recognition accuracy with CO covariate, which are summarized as follows:

- The first benchmark is a non-training-based direct matching method [23], which calculates the dissimilarity using the euclidean distance (i.e., L_2 distance) between two GEIs. The method is denoted by DM in this thesis.

- The second benchmark employed LDA [71], which is widely explored in gait recognition [36, 120]. Mainly, we first applied principal component analysis (PCA) to an unfolded feature vector for a GEI to reduce its dimensions and subsequently utilized LDA to obtain a metric to recognize an unknown sample. PCA_LDA denotes the benchmark in this thesis.

- Gait energy response function (GERF) [36] transforms a GEI into a better discriminative feature. Then a Gabor filter was employed to the modified GEI. Subsequently, LDA was applied, followed by PCA for dimensionality reduction. This approach was used as a third benchmark and denoted as GERF.

- A support vector machine (SVM) [121] is a widely used state-of-the-art method for multi-class classification. We, therefore, selected SVM as a benchmark with a polynomial kernel (third-degree) for the classification of the CS labels. Later, this benchmark is denoted by mSVM in this thesis.

- RankSVM [72] is an extension of a SVM that is used for gait recognition in the literature [33, 74, 75] as a metric learning-based approach. In the training stage, we set the positive and negative feature vectors as the absolute difference between the genuine and impostor pair of GEIs, respectively. To reduce the computational and memory cost, we selected randomly nine impostor pairs against a genuine pair. The benchmark is denoted by RSVM in this thesis.

- GEINet [73] is a deep learning-based approach with a simple CNN network architecture for gait recognition. Single input GEI feature is fed to the network and the soft-max value from the output of the final layer (fc4). The number of nodes for soft-max is equal to the number of training subjects, which is considered as the probability that the input matches a corresponding subject. This benchmark is denoted by GEINet in the experiment discussions of this thesis.

- Similar to GEINet, Siamese [122] is also deep learning-based approach. However, the main difference is that two input GEI features are fed to train the two parallel CNN networks with shared parameters [22, 123]. The output of the final layer (fc4) is a feature vector. A contrastive loss was used for the genuine pair, while so-called hinge loss for the imposter pair. Similar to RSVM, we set nine imposter pairs for a genuine pair during training. The benchmark is denoted by SIAME in this thesis.

3.3.3 Cooperative and uncooperative settings

In this section, we investigated the impact of the cooperative and uncooperative settings for recognition accuracy. The crucial assumption for the cooperative setting is that the covariate condition is uniform in a gallery set. However, it is challenging to collect such types of data in a real-world scenario because of the uncooperative and non-intrusive nature of gait biometrics. In addition to the cooperative setting, we, therefore, selected a more natural uncooperative setting where the covariate condition was inconsistent in the gallery set [75].

For the experimental settings, a subject list was prepared in which, information who had a sample A_1 sequence (i.e., with CO) and a sample in either the A_2 or A_3 sequences (i.e., without CO) are included. As a result, the subject list included 58,199 subjects. Afterward, the subject list was divided randomly by subject identity into two sets: a training set and test equally for each CS label. The training set included 29,097 subjects, and the test set included 29,102 subjects. Finally, the test set was divided into two subsets: a probe set and a gallery set. We used samples from the A_2 or A_3 sequences (i.e., without COs) in the gallery, whereas the sample from the A_1 sequence was used as a probe for the cooperative setting. On the other hand, samples of each subject were randomly divided into a probe set, and gallery set so that the gallery included a mix of samples (i.e., with and without CO) that is the A_1 and A_2 or A_3 sequences for the uncooperative setting. The training sets for the cooperative and uncooperative settings were prepared in the way to reflect the corresponding test sets.

The results for CMC and ROC curves are shown in Fig. 3.6, and Rank-1, Rank-5, $\text{FRR}_{1\%}$, EER, and area under curve(AUC) are shown in Table 3.2. From these results, we can see that the accuracy for the cooperative setting is better than that of the uncooperative setting for most of the benchmarks.

Among the benchmark methods, DM achieved the worst performance. Because DM is the non-training-based approach and did not apply a technique to tackle the CO covariate, it was, therefore, directly affected by the spatial displacement of the corresponding body parts in GEIs caused by the CO difference. On the other hand, the accuracy of the training-based approaches was better than that of DM because the similarity or dissimilarity metrics were optimized using the training data.

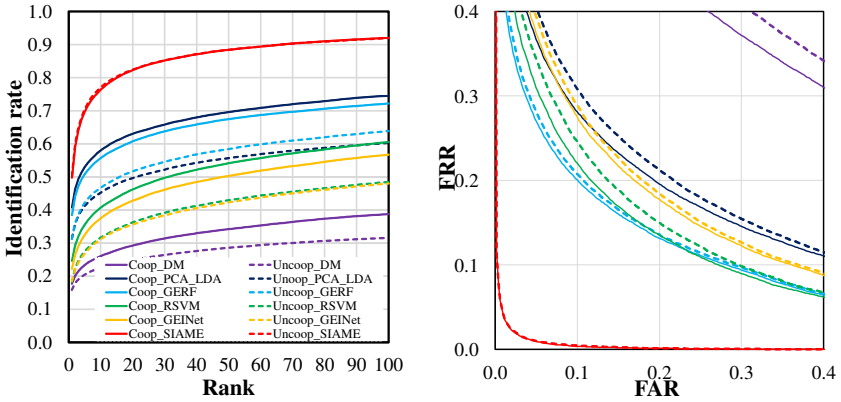
Regarding the LDA-based metric learning benchmarks, both PCA_LDA and GERF worked reasonably well and their performances were very similar. However, GERF was slightly better for the uncooperative setting, whereas PCA_LDA was slightly better for the cooperative setting, as shown in Fig. 3.6 and Table 3.2. We believe that LDA performed better recognition for both benchmarks by reducing intra-subject appearance variation while increasing inter-subject

variations. Furthermore, in GERF, before applying LDA and PCA, a pre-processing technique was performed on GEI, for example, transforming a pixel value for a better discriminative feature. This transformation in GERF was not effective for the cooperative setting; however, it worked well for the uncooperative setting. As a result, the performance of GERF was better for the uncooperative setting.

Regarding the LDA-based metric learning approaches, both GERF and PCA_LDA worked reasonably well, and their performances were comparable. GERF was slightly better for the uncooperative setting, while PCA_LDA was slightly better for the cooperative setting, as shown in Fig. 3.6 and Table 3.2. We think that LDA performed better for both of these benchmarks by reducing intra-subject appearance variation while increasing inter-subject variations. Moreover, in GERF, before applying LDA, a pre-processing was performed on GEI, for example, transforming a pixel value for a better discriminative feature. This transformation, therefore, worked well for the uncooperative setting. However, it was not useful for the cooperative setting. As a result, the performance of GERF was slightly worse for the cooperative setting.

Regarding RSVM, it is reported that RankSVM works better in identification for gait recognition [75] because it emphasizes more on the relative distance between two classes along with considers the probe-dependent rank statistics. However, it did not work well in our experiment. We think the cause of this weak performance was that we could only set the number of impostor pairs at nine against a genuine pair (see Section 3.3.2) to reduce computation and memory cost. Therefore, RankSVM could not efficiently maximize inter-subject variation. This is one of the important limitations of the RankSVM method for an extremely large training dataset.

As for the CNN-based benchmarks, we can see that, GEINet did not work well; however, SIAME achieved the best recognition accuracy with a large margin compared with other benchmarks. We believe the cause of the weak performance for the one-input GEINet was that the parameter of CNN architecture was trained to maximize the soft-max of the output layer (fc4) node for the same subject's input GEIs. Therefore, it focuses on minimizing only intra-subject variation. However, only two sample GEIs for each subject were used in our experiments, which was not sufficient to train a useful parameter for CNN. By contrast, the Siamese the two-input CNN architecture in SIAME was trained so that it minimized the variation between the intra-subject and maximized the variation between inter-subject GEIs. Moreover, there was no accuracy difference between the cooperative and uncooperative settings for the SIAME benchmark. We believe that the deep network architecture of the Siamese was sufficiently powerful to manage CO covariates given an extreme large dataset for training.



(a) CMC curves (b) ROC curves

Figure 3.6: CMC and ROC curves for cooperative and uncooperative settings. Legend marks are common in all graphs.

Table 3.2: Rank-1/5 [%], $FRR_{1\%}$, EER [%], and AUC [%] for cooperative (Coop) and uncooperative (Uncoop) settings. Bold and italic bold fonts indicate the best and second-best benchmarks, respectively.

Benchmark	Rank-1		Rank-5		$FRR_{1\%}$		EER		AUC	
	Coop	Uncoop	Coop	Uncoop	Coop	Uncoop	Coop	Uncoop	Coop	Uncoop
DM	17.7	15.9	23.4	20.5	71.7	74.6	34.4	36.5	27.0	29.4
PCA_LDA	***40.8***	***31.4***	***53.0***	41.3	*54.7*	*59.1*	*19.7*	20.7	*12.4*	13.1
GERF	38.5	31.2	50.9	***42.2***	***42.6***	***44.5***	***15.6***	***15.9***	***7.7***	***8.0***
RSVM	24.7	18.3	35.6	27.6	49.1	48.6	16.1	16.2	8.4	8.2
GEINet	22.3	18.5	32.5	26.9	57.8	60.7	18.7	19.1	10.8	11.2
SIAME	**49.8**	**50.3**	**69.7**	**70.5**	**4.1**	**4.1**	**2.2**	**2.2**	**0.2**	**0.2**

3.3.4 Difficulty level of the CS labels

In this experiment, we analyze the difficulty level for the CS labels based on gait recognition performance. To achieve the goal, we selected the same protocol as the cooperative setting, except the probe set was divided into seven subsets according to the CS label. In contrast, the gallery was unchanged for a fair comparison.

The results for the Rank-1 rate and EERs, respectively, for identification and verification, are shown in Fig. 3.7. NoCO and CpCO achieved the best and worse CS labels, respectively, whereas the remaining labels (i.e., SbCO, SmCO, FrCO, BaCO, and MuCO) were approximately at the middle level difficult. We can discuss the results by considering the static shape and dynamic motion of the gait feature.

NoCO was the best CS label for any benchmark, and this is reasonable because there was no CO covariate between the probe and gallery of the same subject. Therefore, shape and motion

for gait features were stable.

The motion and shapes are deviated by different amounts for middle-level difficulty labels. For example, for SbCO and SmCO CS labels, subjects frequently carried small and lightweight COs, which were occluded by the subject's body very often, as shown in Fig. 3.3. Therefore, the COs did not have much of an influence on the shape. For the case of BaCO, subjects typically carried a large CO, such as a backpack, rucksack that was secured by two straps that fit over the shoulders. Therefore, the position of the CO was fixed and stable within a gait cycle. However, the large CO heavily affected the shape and posture, as shown in Fig. 3.3. Similarly, MuCO, for which subjects typically carried a large backpack-type CO together with other COs that were carried in other human body regions. Though the CO position of the back region was fixed, other CO positions were random. Thus, GEI samples for MuCO were heavily affected not only by shape but also by motion. Therefore, the recognition performance of MuCO label was worse than that of BaCO. Regarding FrCO, the subjects typically carried a lightweight object in hand in the front human body region. Particularly, both hands were required to hold the CO in the front region; thus, the CO position was not stable. Therefore, the GEI samples of FrCO were affected slightly by shape and fairly affected by motion.

As for CpCO, the CO position was random in any region within a gait cycle because of the randomly changing position from one region to another. Thus, GEI samples for CpCO were severely affected by the motion feature along with shape. As a result, CpCO was the most difficult CS label.

3.3.5 Impact of the number of training subjects

It is well-known that the performance of a modern deep machine learning-based CNN approach depends on a variety of training samples. In a specific scenario, for example, in our case, this variety can be expressed by the number of training subjects. Therefore, in this section, the impact of the number of training subjects for gait recognition is investigated.

To investigate the impact of the number of training subjects for recognition, we chose the cooperative setting of section 3.3.3 and chose the CNN-based benchmark SIAME, which was achieved the best accuracy. Afterward, we prepared the training set included 100, 200, 500, 1,000, 2,000, 5,000, and 10,000 subjects randomly selected from the entire training set (29,097), and the test set was kept unchanged for a fair comparison. Note the smaller training sets are not independent; they are the subset of the largest training set (29,097 subjects), and we did not select training sets multiple times to reduce randomness.

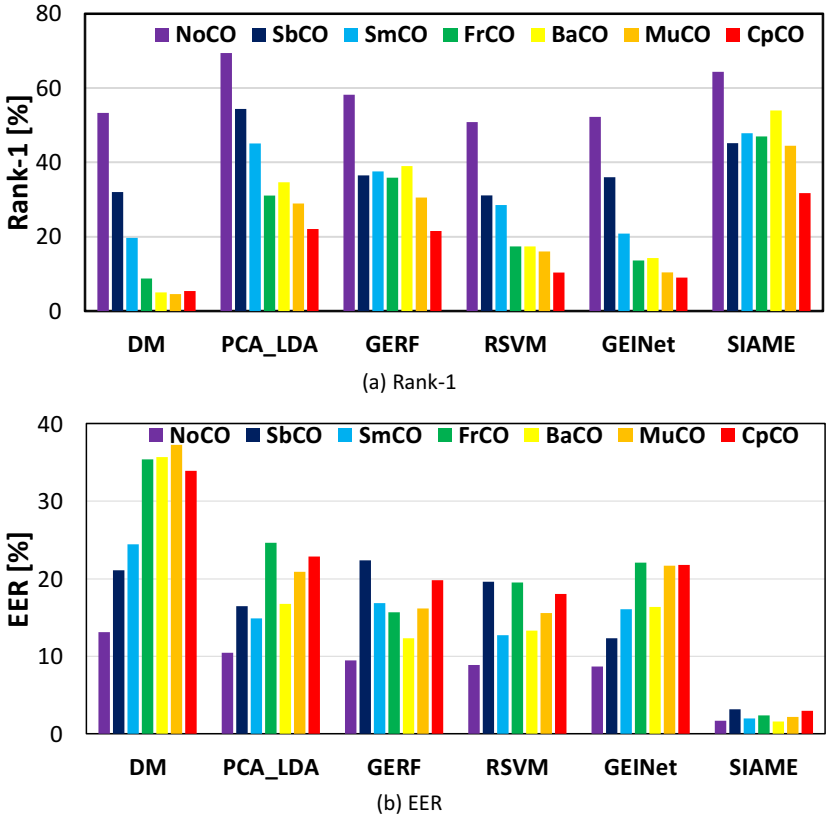


Figure 3.7: Rank-1 identification rate and EERs for the difficulty level of CS labels.

The results for Rank-1 and EERs are shown in Fig. 3.8. It was clearly showed the accuracy was better for a larger number of training subjects. For example, EER reduced by approximately 13% when the number of training subjects increased from 100 to 29,097, while the rank-1 increased by approximately 44%. The above results demonstrate the importance of the number of training subjects for a deep learning-based approach. Therefore, a database for gait recognition with a large number of subjects is essential.

3.3.6 Classification of the CS labels

In previous sections, we analyzed gait recognition. In this section, we investigate a different recognition problem, that is, the classification of the CS labels based on the gait feature. These could have numerous applications, for example, the detection of suspicious events, such as bag-prohibited area incursion and identifying the person with a backpack. However, to the best of our knowledge, there is no standard gait-based database with CO covariate, including labeling information about the position and type of CO. Thus, existing work in the literature of gait community detects a CO using the gait feature [70, 124, 125]. Additionally, they can only

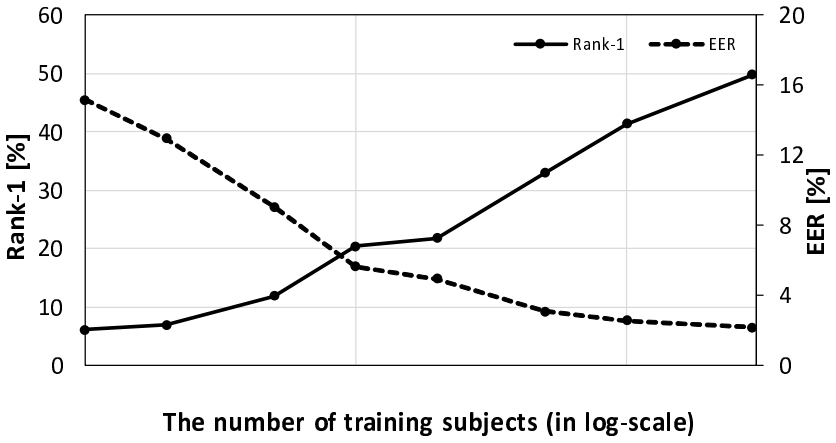


Figure 3.8: Relationship between the number of training subjects and recognition accuracy for SIAME.

classify a subject with or without a CO. We strongly believe that to overcome such a limitation, our proposed database with labeling information can be used as a benchmark database for the detection and classification of CO positions.

To evaluate the performance of the classification of the CS labels, we prepared the subject list for training and testing; for this purpose, the subject ids for each label are divided into a training set, and test set equally. To equalize the number of training subjects for each label, we considered the smallest number of training subjects for a label, that is, for CpCO label (1,300 subjects). Because CS labels are annotated from the location of CO is being carried for a person, and all CO in a location is not the same color and texture. For example, the backpack can be a different color, and texture can be carried in the back region (i.e., BaCO label). Therefore, we considered the color and texture invariant silhouette-based features (i.e., GEI). Because the annotation was performed on A_1 sequence, and we considered A_1 GEI for this experiment. Then, the training-based benchmarks are trained by the prepared equalized training set. Regarding testing, each sample of a CS label was matched with all the available samples of the training set. To predict the CS label for a test sample, the mean distance to a class was used all benchmarks except mSVM [121]; whereas majority voting was used for mSVM. More specifically, each sample was matched against all the samples of the training set and then calculated the mean value for available samples of the corresponding class (i.e., 1300 subjects); finally, the decision was made based on the mean values for all classes.

The result for correct classification rate (CCR) of all CS labels for each benchmark are shown in Fig. 3.9. In addition, the confusion matrices for the best and second-best benchmarks that are the SIAME and mSVM shown in Table 3.3 for all labels as an average accuracy. The accuracy for each label was quite different and depended on the benchmark.

Table 3.3: Confusion matrix for the classification of the CS labels.

(a) SIAME

Actual label \ Predicted label	NoCO	SbCO	SmCO	FrCO	BaCO	MuCO	CpCO
NoCO	76.8	11.1	5.7	3.1	1.6	0.5	1.1
SbCO	13.5	73.0	6.1	2.5	0.9	2.1	2.1
SmCO	11.1	11.8	32.7	13.5	12.4	10.2	8.3
FrCO	3.2	2.6	7.4	72.8	0.3	4.2	9.5
BaCO	2.4	0.8	4.5	0.2	78.9	12.1	1.0
MuCO	0.9	3.2	5.9	7.6	15.9	62.6	3.9
CpCO	2.9	6.6	6.2	20.3	4.4	13.3	46.2

(b) mSVM

Actual label \ Predicted label	NoCO	SbCO	SmCO	FrCO	BaCO	MuCO	CpCO
NoCO	72.6	9.4	11.5	2.3	1.8	0.6	1.8
SbCO	24.5	58.4	11.0	1.6	1.2	0.9	2.4
SmCO	18.1	11.7	36.0	9.8	12.2	4.2	8.0
FrCO	10.7	2.4	9.7	60.2	0.5	1.7	14.7
BaCO	5.3	1.3	6.8	0.2	74.7	10.6	1.0
MuCO	3.3	4.3	10.2	7.9	27.3	40.1	6.9
CpCO	8.2	7.0	11.8	30.4	7.9	8.2	26.5

From these result, we can see that SIAME and mSVM consistently worked well for each label, as shown in Fig. 3.9. For SIAME, as already discussed in Section 3.3.3, the Siamese network architecture was trained by minimizing the distance between intra-labels and maximizing the distance between inter-labels. Therefore, it achieved the best classification accuracy. Although mSVM used a shallow traditional machine learning approach (i.e., SVM), it worked well. We believe the cause is that multi-class SVM [121] constructed multiple binary classifiers (e.g., $K(K-1)/2$ classifiers for K classes), one for each pair of classes, and finally identified a class based on majority voting. By contrast, the remaining benchmarks had a similar trend to the cooperative and uncooperative settings, such as GERF, and PCA_LDA achieved nearly equal accuracy.

Regarding the classification accuracy of each label, NoCO and BaCO worked well because there was no CO in NoCO, and the shape and position of the CO were fixed and stable in BaCO. For SIAME, the CCRs were 76.8% and 78.9% for NoCO and BaCO, respectively, as shown in Table 3.3. For the case of SbCO and FrCO, the position and shape of the COs were fairly distinguished compared to other labels. Therefore, the classification accuracy of these labels was reasonable and nearly equal. However, SbCO was slightly confused with NoCO because of

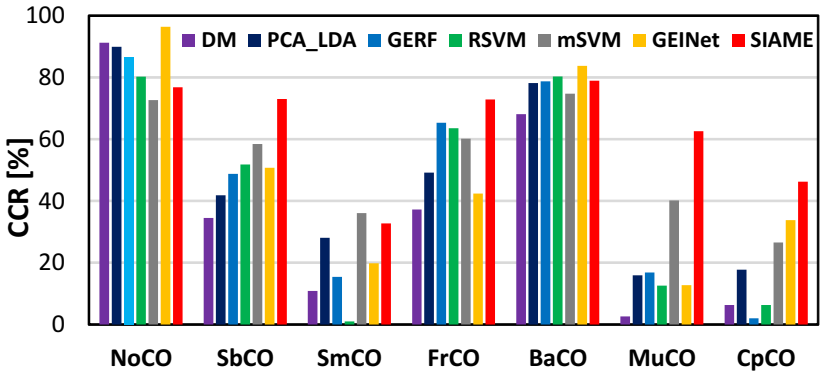


Figure 3.9: CCRs of the CS labels.

the shape similarity with respect to the upper part of the GEIs. Therefore, sometimes samples of SbCO were misclassified as NoCO; see Table 3.3.

Regarding SmCO, MuCO, and CpCO, the GEI features were not stable, and therefore, samples of these labels were sometimes misclassified as other labels. Because of the occlusion of COs with the subject's body for SmCO, the GEI feature was confused with that of SbCO, NoCO, and BaCO, depending on the part of the COs that was occluded, as shown in Fig. 3.3. Therefore, sometimes samples of SmCO label were misclassified as SbCO, NoCO, and BaCO; see Table 3.3. Similarly, for the case of MuCO, it was confused with BaCO, because, as already mentioned in Section 3.3.4, subjects typically carried, for example, a backpack type object in the back region together with a small object in other regions in MuCo, as shown in Fig 3.3. Besides, for the case of CpCO, subjects usually changed the CO position from one region to another through the front using the hands. As a result, the GEI feature of CpCO was slightly confused with that of FrCO.

3.4 Discussion

Extension to other factors: While we constructed the dataset by including CO with single view variation, there are even more challenging factors that can be encountered in the real scenario. In addition, the dataset was captured in a relatively controlled situation (e.g., controlled illumination and a predefined walking course). Therefore, the construction of a gait database considering the other challenging issues, which includes both the external factors (e.g., clothing) and environmental factor (e.g., illumination, walking surface and occlusion) along with view variation, are necessary for the development gait-based human recognition in the real-world environment. Furthermore, although the constructed dataset included 62,528 subjects, it

is not large enough in many applications, for example, the identification at million scales (i.e., finding a person in a database from millions of people). Moreover, to leverage the full capabilities of a deep learning-based approach, it requires more training data similar to the domain of face recognition [51].

Evaluation experiments: Further analysis of gait recognition performance using our database is still needed. While the existing work in the literature using the attribute information for disentangled representation and multi-task learning, for example, pose or view variation information in [126, 127]. Therefore, it can be added to the CS labels information for disentangled representation to improve gait recognition accuracy. In addition, the generative adversarial network (GAN) [83] can be used to remove the carried object (CO) to reconstruct the silhouette sequence with CO to silhouette sequence without CO or preprocess GEI images with CO to GEI without CO for gait recognition.

3.5 Conclusion

In this chapter, we presented a gait database that consisted of an extremely large number of subjects with unconstrained types and positions of COs. Furthermore, we analyzed silhouette-based gait recognition using state-of-the-art methods. This database had the following advantages over the existing gait databases in the literature: (1) the database included 62,528 subjects, which was more than six times greater than the existing largest database for gait recognition without CO and 200 times with CO; and (2) it has manually annotated CO position and gait sequence with CO were classified as seven distinct CS labels. Furthermore, we conducted four experiments to analyze the gait recognition with CO covariate using the proposed database. The results provided several insights, such as estimating the difficulty level among annotated CS labels based on recognition performance and the classification accuracy for CS labels.

Chapter 4

Spatio-temporal silhouette sequence reconstruction for gait recognition with occlusion

4.1 Introduction

In chapter 3, we tackle the problem of the carried object, which is an essential covariate in real-world scenarios. In contrast, in this chapter, we consider another important factor for environmental context, i.e., the occlusion for gait recognition. The presence of occlusion in silhouette sequence missing a significant amount of the human body, lead to an ill-posed because unoccluded silhouette reconstruction from an occluded silhouette, an infinite number of solutions exist that produce from the same silhouette. Occlusion can be one of two types based on the relative position between the occluder and the target subject in an image sequence: relative dynamic occlusion and relative static occlusion. For relative dynamic occlusion, the occluded portion of the target subject changes continuously over an image sequence, whereas, for relative static occlusion, the occluded portion does not change. An example of relative dynamic occlusion is shown in Figs. 4.1 (a) and (b), in which the person is occluded at different positions in each frame, and the occluded portion of the person's body gradually changes in the video sequence during the person's gait cycle. For the example of relative static occlusion shown in Fig. 4.1 (c), the person is occluded at a fixed portion of the body in each frame in the video sequence during the person's gait cycle.

Approaches to gait recognition against occlusion can be roughly grouped into two categories, as discussed in section 2.3.2. The first category is reconstruction-free approaches [78, 79, 80, 81], which focus on extracting features from a silhouette sequence of a gait cycle or an average of them, such as the gait energy image (GEI) [20]. Because gait features are extracted by considering the static shape and dynamic motion information from a silhouette

sequence for a gait cycle, approaches of this type can achieve good performance for a very low degree of occlusion. However, the obvious limitation of this type of approach is that it cannot be applicable to cases in which the gait cycle is difficult to estimate.

The second category is reconstruction-based approaches [24, 31], approaches in this category focus on reconstructing occluded silhouettes. In these approaches, occluded silhouettes are identified and a sequence is separated into occluded and unoccluded gait cycles, and then silhouettes of occluded gait cycles are reconstructed. These approaches showed good silhouette reconstruction. However, these were applied on long sequences that consisted of multiple gait cycles in which some frames were partially occluded. These approaches are difficult to apply in the case in which all frames are severely occluded in a sequence, for example, the occlusion shown in Figs. 4.1 (a) and (b). One of the major limitations of reconstruction-based approaches is that the reconstructed silhouette sequence sometimes deteriorates the discrimination ability of the individual after reconstruction. Therefore, it can negatively influence gait recognition performance after reconstruction [128].

Reviewing the work of CNN and generative adversarial networks (GANs) [43] in section 2.4, the reconstruction-based approaches have been formulated as a conditional image or video generation problem. Thus GAN significantly advances many ill-posed problems such as image inpainting [44, 45, 47, 48, 49], video inpainting [50, 52] and future prediction [52, 88, 129, 130]. Although these works have been shown to generate very good looking realistic images, such as faces, objects, and scenes, they sometimes lost subject identity [131]. An approach that can generate not only good looking samples but also samples with the discrimination ability of an individual is necessary for biometric-based person recognition.

We present an effective feed-forward conditional deep generative network [160] for silhouette sequence reconstruction considering dilated convolution [48, 133] and a skip connection [134]. Dilated convolutional kernels are spread out in the spatial and temporal directions, which allows us to reconstruct each pixel by covering a large Spatio-temporal input area. This is important for silhouette sequence reconstruction because each input pixel is essential for reconstruction, whereas a skip connection allows us to retain unoccluded input pixels as output. The input to the encoder network that maps hidden representations is the occluded silhouette sequence, and the output of the decoder is the reconstructed silhouette sequence. We regularize the training process of the generator network by incorporating triplet hinge loss into Wasserstein GAN (WGAN) loss [53, 54] as adversarial loss and reconstruction loss in pixel space. A triplet contains a query sequence, a positive sequence, and a negative sequence, where the query

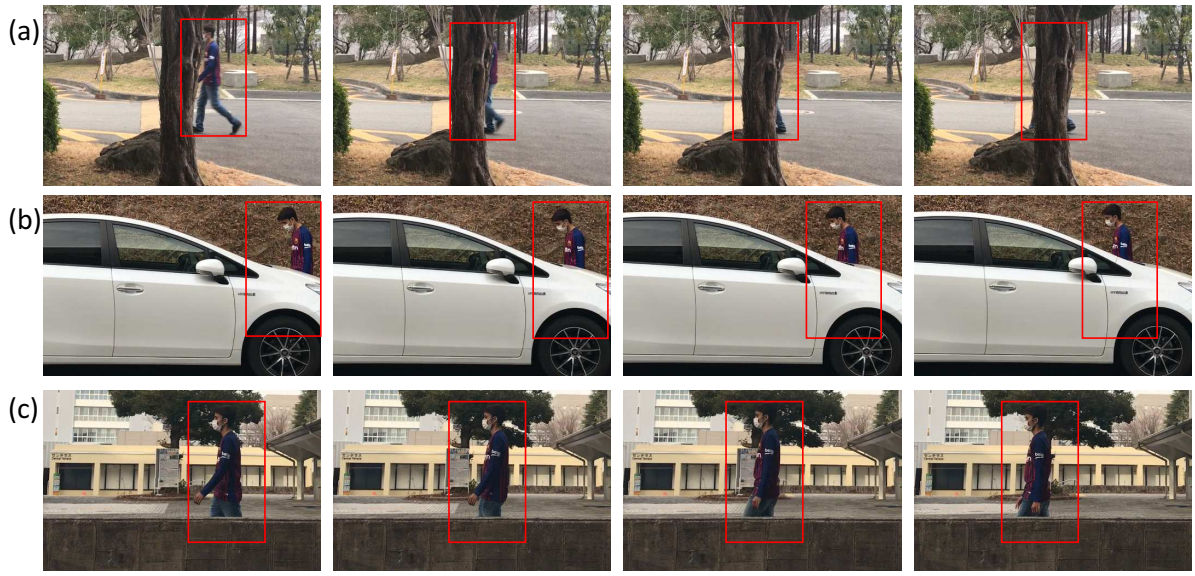


Figure 4.1: Examples of occlusion in real-life applications (every fifth frame of a sequence): (a) relative dynamic occlusion where the subject is occluded by a tree and continuously changes the occluded portion from left to right; (b) relative dynamic occlusion where the subject is occluded by a parked car and continuously changes the occluded portion from bottom to top; and (c) relative static occlusion where the subject is occluded by wall in a fixed position.

sequence is the reconstructed silhouette sequence, the positive sequence is the unoccluded silhouette sequences of the same subject as the query subject, and the negative sequence is of a different subject. The relative distance in the triplet characterizes the similarity relationship.

The entire network is trained end to end with the reconstruction and proposed adversarial losses. Compared with existing inpainting or reconstruction-based approaches, one of the significant advantages of our proposed approach is that it does not require occluded or inpainting position information (i.e., a mask) for reconstruction. Therefore, it can be applied to an arbitrarily structured occluded silhouette sequence during reconstruction. Because of the silhouette sequence reconstruction approach, we can evaluate gait recognition without knowing the gait cycle in advance because the gait cycle can be estimated from the reconstructed silhouette sequence.

4.2 Spatio-temporal silhouette sequence reconstruction

The purpose of the proposed approach is to reconstruct a silhouette sequence from an occluded sequence based on conditional GANs. An overview of the proposed approach is shown in Fig. 4.2. It utilizes a generator G and critic D networks. The generator network is used for the silhouette sequence reconstruction. In contrast, the additional network critic is used to

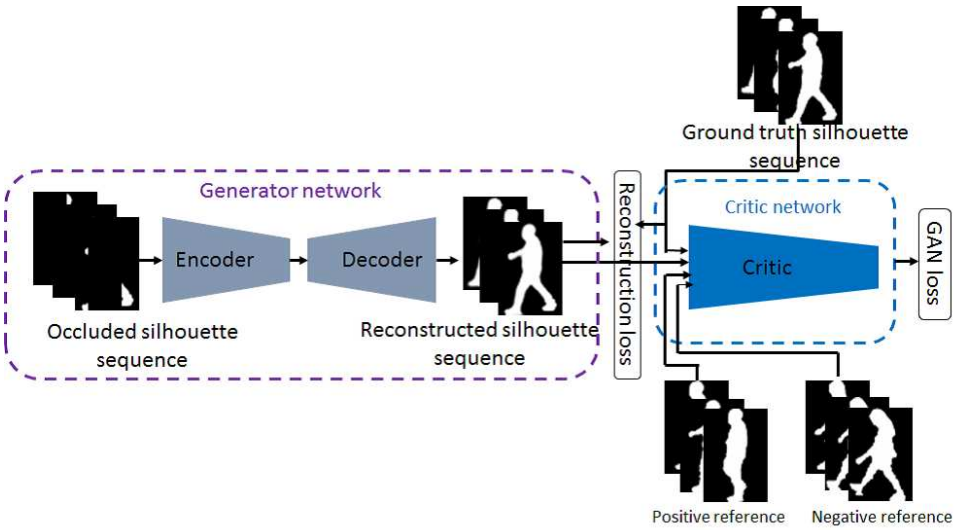


Figure 4.2: Overview of our silhouette sequence reconstruction framework. It consists of a generator (encoder and decoder) and a critic network. The generator takes the occluded silhouette sequence as input and outputs the reconstructed silhouette sequence. The critic is used to supervise the generator network during training (i.e., positive reference is unnecessary for the target subject reconstruction during testing).

supervise the generator network during training to realistically reconstruct as well as preserve subject identity. After training, the generator network can take an occluded silhouette sequence and reconstruct it.

Different from existing video generative approaches [52, 88], we propose to design an architecture for the generator network considering the spatio-temporal 3D convolution with small kernels along with dilated convolution and skip connections; we will describe in detail in section 4.2.1. For the case of the critic network, we chose popular critic architecture to [52]. Nevertheless, the training procedures are different; and we will explain it in detail in section 4.2.2.

4.2.1 Generator network

Generator network architecture is designed as an encoder-decoder pipeline. The occluded silhouette sequence is given to the encoder to map into hidden low dimensional representations, which allows low computational and low memory costs by reducing the spatial and temporal resolutions. Unlike a pooling layer, the encoder decreases the spatial and temporal resolution twice by stridden convolutions to avoid a blurred texture in the occluded regions. Afterward, the decoder takes this low-dimensional feature and restores it to the original spatial and temporal resolution through a series of the convolutional layers with fractional strides [135]. Unlike

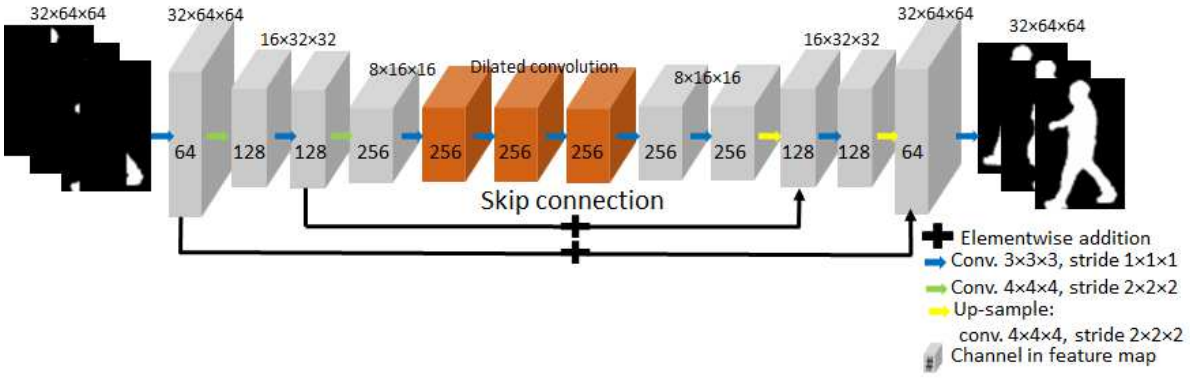


Figure 4.3: Illustration of the architecture of the generator network. The silhouette sequence and feature dimensions are shown in the figure, and denoted as "time×width×height".

[48, 88], we employed convolutional kernels of $3 \times 3 \times 3$ (time $\times$ width $\times$ height) and $4 \times 4 \times 4$ because it is demonstrated that small size kernels perform better in a deep 3D network [90]. An illustration of the generator network is shown in Fig. 4.3.

We employed dilated convolution [133] in the mid-layer, whereas skip connections [134] in the top layers. The dilated kernels are spread out through the spatio-temporal directions, which allows us to compute each output pixel by considering a much larger input area, while the number of parameters and computational cost remains constant. This is very important for the silhouette sequence reconstruction from a partially observable occluded sequence because the spatial context and the neighbor frame are crucial for reconstruction. To keep unoccluded input pixels in the reconstructed sequence, we use a U-shape-like network with skip connections (i.e., the feature of the encoder are combined with the decoder) because the decoder path is more or less symmetric to the encoder path.

We initialize the convolutional weights for stable training and faster convergence as [136]. We perform batch normalization [137] to zero mean and unit variance followed by rectified linear unit (ReLU) activation functions after each layer, except the final output layer. A hyperbolic tangent function is employed in the last layer, which is helpful for normalizing the reconstructed sequence within the range $[-1, 1]$.

4.2.2 Critic network

Different from existing GANs [52, 53, 54] in which a discriminator/critic discriminates generated samples from ground truth samples and supervise generator network adversarially. However, considering a different direction, we propose exploring a modified WGAN. Our proposed critic network, D, can discriminate a reconstructed silhouette sequence of a subject from ground

truth, and simultaneously use the pairwise similarity ranking, where the critic network provides a smaller distance to a silhouette sequence of the same subject and larger distance to a different subject, and it is realized using hinge loss. Employing the hinge loss along with WGAN loss, we use the adversarial loss so that the generator can maximally fool the critic.

The architecture and layer settings are similar to [52]. More specifically, we consider five convolutional layers, followed by a downsampling layer with $4 \times 4 \times 4$ convolutional kernels with a stride of $2 \times 2 \times 2$. We set the number of output channels for the first layer to 64 and twice the values as the layer goes into deeper. Similar to DCGAN [83], we employ LeakyReLU [138] with threshold a of 0.2. As in [54], we use layer normalization [139] instead of batch normalization. Because our proposed critic is not trained to classify the reconstructed silhouette sequence from the ground truth, we exclude softmax or any other activation in the final layer and instead train the network to give good gradient information to the generator updates.

4.2.3 Training objective

To train networks, we use objective functions comprise of silhouette sequence reconstruction loss, WGAN loss along with hinge loss as an adversarial loss. Given occluded z and corresponding ground truth silhouette sequences x along with positive reference $\bar{x}$ and negative reference $\bar{\bar{x}}$, respectively, as the same and different subject as ground truth, our proposed approach is trained to minimize the generative loss for generator network G:

$$L_{gen} = L_{adv} + \gamma L_{img}, \tag{4.1}$$

where γ is a weighting hyper-parameter to control the trade-off between adversarial L_{adv} and image loss L_{img}.

L_{img} is the image loss, calculates the mean squared error, which tries to minimize the pixel-wise error between the reconstructed ($\tilde{x} = G(z)$) and ground truth silhouette sequence. It is well-known that stabilizing the adversarial training is a significant issue in GANs. A loss in image space is added with adversarial loss, and the loss in image space can contribute to stabilizing the training [140]. We, therefore, use the image loss L_{img} with adversarial loss in our proposed approach, which can be defined as follows:

$$L_{img} = \underset{\tilde{x},x\sim\mathbb{P}_g,\mathbb{P}_r}{\mathbb{E}} \left[(\tilde{x} - x)^2 \right], \tag{4.2}$$

where $\mathbb{P}_g$ and $\mathbb{P}_r$ represent the distributions of reconstructed silhouette sequence $\tilde{x}$ and ground truth x, respectively.

Adversarial loss L_{adv} is the generator loss in adversarial training, which is the combination of WGAN loss and triplet ranking hinge loss, which can be defined as follows:

$$L_{adv} = L_{WGAN} - \kappa L_{hinge}, \tag{4.3}$$

where $L_{WGAN} = -\mathop{\mathbb{E}}_{\tilde{x}\sim\mathbb{P}_g}[D(\tilde{x})]$ is the WGAN loss and L_{hinge} is the hinge loss for pairwise similarity ranking, and κ is the coefficient to control the trade-off between WGAN and the proposed hinge loss. The output of the critic network D is a real-valued scalar, and the hinge loss is calculated using the relative distance of the output of the reconstructed silhouette sequence with the positive reference and negative reference. The positive reference silhouette sequence is the same subject to the reconstructed silhouette sequence, whereas a different subject for negative reference. More specifically, the triplet pairwise ranking hinge loss function can be defined as follows:

$$\begin{aligned} L_{hinge} = \max(margin &- \mathop{\mathbb{E}}_{\tilde{x},\bar{\bar{x}}\sim\mathbb{P}_g,\mathbb{P}_{\bar{\bar{x}}}}[|D(\tilde{x}) - D(\bar{\bar{x}})|] \\ &+ \mathop{\mathbb{E}}_{\tilde{x},\bar{x}\sim\mathbb{P}_g,\mathbb{P}_{\bar{x}}}[|D(\tilde{x}) - D(\bar{x})|], 0), \end{aligned} \tag{4.4}$$

where $\mathbb{P}_{\tilde{x}}$, $\mathbb{P}_{\bar{x}}$ and $\mathbb{P}_{\bar{\bar{x}}}$ represent the distributions of reconstructed $\tilde{x}$, positive reference $\bar{x}$ and negative reference silhouette sequence $\bar{\bar{x}}$, respectively.

Similar to the generator network G, we train critic network D using the framework of the improved WGAN with a gradient penalty coefficient [54] together with the proposed hinge loss. Especially, the critic network D is trained to minimize the following loss function:

$$L_{critic} = \mathop{\mathbb{E}}_{\tilde{x},x\sim\mathbb{P}_g,\mathbb{P}_r}[D(\tilde{x}) - D(x)] + \lambda L_{GP} + \kappa L_{hinge}, \tag{4.5}$$

where $L_{GP} = \mathop{\mathbb{E}}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1)^2\right]$, $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}$, and λ is a gradient penalty coefficient and $\varepsilon \sim U[0,1]$. We used Adam optimizer[141] to update both networks G and D with batch size and learning rate α are respectively as 32 and 0.0001 for a fixed number of iterations n for the generator network. The other hyperparameters for the Adam optimizer were set to $\beta_1 = 0.5$ and $\beta_2 = 0.99$. Algorithm 1 shows the complete algorithm for training our proposed framework in this chapter. We used default $\lambda = 10$, as suggested in [54], and $\gamma = 1000$ according to [52]. The values of the coefficients κ and *margin* were determined empirically as 20 and 3, respectively, for each experiment. All the networks were implemented in Python with the Tensorflow library, and every experiment was trained from scratch. We normalized all silhouette sequences to be in the range $[-1,1]$.

Algorithm 1 Training of our proposed framework. We use default values $n_{critic} = 4$, $\alpha = 0.0001$, $\lambda = 10$, $margin = 3$, $\gamma = 1000$, $\kappa = 20$, $\beta_1 = 0.5$, $\beta_2 = 0.99$

Require: Batch size b, training iterations n, gradient penalty coefficient λ, number of critic iterations per generator iteration n_{critic}, coefficient κ, width W, height H of silhouette, Adam hyperparameters α, β_1, β_2

Require: Initial critic parameter W_{D_0}, initial generator parameter W_{G_0}

1: **for** $iter \leftarrow 1$ to n **do**

2: **for** $i = 1, ..., n_{critic}$ **do**

3: Sample batches for occluded silhouette sequences z, ground truth silhouette sequences x, positive reference $\bar{x}$ and negative reference $\bar{\bar{x}}$, a random number $\varepsilon \sim U[0,1]$

4: Update the weight W_D of critic network D using Eq.(4.5) :
$\tilde{x} = G(z)$, $\hat{x} = \varepsilon x + (1-\varepsilon)\tilde{x}$
$L_{Wdist} = \frac{1}{b}\sum_{j=1}^{b} D(\tilde{x_j}) - D(x_j)$,
$L_{GP} = \frac{1}{b}\sum_{j=1}^{b}(\|\nabla_{\hat{x_j}} D(\hat{x_j})\|_2 - 1)^2$,
$L_{hinge} = \max(margin - \frac{1}{b^2}\sum_{j=1}^{b}\sum_{k=1}^{b}\left|D(\tilde{x_j}) - D(\bar{\bar{x}}_k)\right| + \frac{1}{b}\sum_{j=1}^{b}\left|D(\tilde{x_j}) - D(\bar{x_j})\right|, 0)$
$W_D \leftarrow Adam(\nabla_{W_D}(L_{Wdist} + \lambda L_{GP} + \kappa L_{hinge}), W_D, \alpha, \beta_1, \beta_2)$

5: **end for**

6: Sample batches for occluded silhouette sequences z, ground truth silhouette sequences x, positive reference $\bar{x}$ and negative reference $\bar{\bar{x}}$

7: Update the weight W_G of generator network G using Eq.(4.1):
$L_{img} = \frac{1}{bWH}\sum_{j=1}^{b}(\tilde{x_j} - x_j)^2$, $L_{adv} = \frac{1}{b}\sum_{j=1}^{b} -D(\tilde{x_j}) - \kappa L_{hinge}$
$W_G \leftarrow Adam(\nabla_{W_G}(L_{adv} + \gamma L_{img}), W_G, \alpha, \beta_1, \beta_2)$

8: **end for**

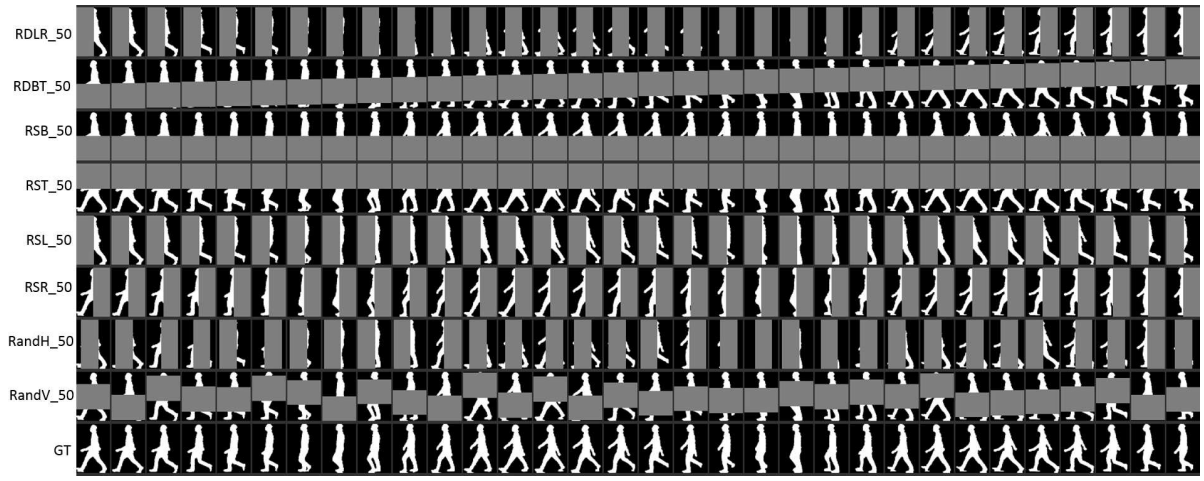


Figure 4.4: Example of different simulated occlusion patterns for a subject. The left-hand side of the figure: labels for the occlusion pattern, where the first term indicates the type of occlusion and the second term shows the degree of occlusion. The occluded area is shown as gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background.

4.3 Experiments

4.3.1 Overview

To evaluate the accuracy of the proposed approach against a wide variety of occlusion patterns, we artificially simulated several occlusion patterns because there is no publicly available large-scale gait recognition database with occlusion variation. Moreover, a systematic analysis of multiple occlusion patterns is necessary for gait recognition. Regarding evaluation, we considered three sets of original experiments to validate the proposed approach. These experiments were meant to address a variety of challenges for various occlusion patterns and different training settings. The goal of these experiments was to evaluate gait recognition for the following conditions:

1. the occlusion pattern was known and the same for a matching pair (probe against gallery);
2. the occlusion pattern was known and different for a matching pair; and
3. the occlusion pattern was unknown for a matching pair.

4.3.2 Dataset

We chose the OUMV dataset [29], which is included gait image sequences with multiple views variation from 10,307 subjects. The data was captured in a controlled environment with a green background for 25 fps temporal resolution and $1,280 \times 980$ spatial resolution. Cameras are placed approximately 8 m from the course at the height of 5 m. The silhouette sequence was extracted using a chroma-key technique. Then the size was normalized by considering the top, bottom, and horizontal center of the silhouette regions for the subject of interest such that the height was 64 pixels and the aspect ratio of each region was maintained. Finally, 44×64 pixels silhouette images were generated. For our experiments, we selected a subset from the dataset (i.e., the side gait sample). Moreover, we considered the subjects (9,001) that had at least two sequences. To artificially simulate occlusion pattern, 32 contiguous normalized silhouettes of a sequence were used. If a sequence had fewer than 32 samples, we repeated the last frame to make it uniform to fit into the network.

Occlusion pattern: We considered two categories of real-world occlusion patterns that could happen in daily life, that is, relative dynamic and relative static occlusion, along with a random occlusion pattern. Regarding relative dynamic occlusion, we simulated an occlusion type in which a person walked from right to left occluded by a beam, pillar, or tree covering

the entire height (e.g., Fig. 4.1 (a)). Therefore, we can imagine that occluder objects move in a continuous motion from left to right within the subject of interest in an image sequence. To realize this occlusion pattern, we added a background rectangle mask (i.e., set to zero in the occluded position) to cover a certain portion against the entire silhouette in the left-most position of first frame of a sequence, and gradually changed the position of the mask toward the end of the frame with the right-most position. Later in this chapter, we named this type of occlusion pattern as a relative dynamic occlusion from left to right (RDLR). Similarly, we simulated relative dynamic occlusion from bottom to top (RDBT) when an occluder occluded a person from bottom to top (e.g., Fig. 4.1 (b)).

As for relative static occlusion, we added a background mask in a fixed position for each frame in a sequence. Thus, we simulated relative static occlusion in the bottom (RSB), top (RST), left (RSL), and right (RSR) positions. Regarding random occlusion, we added a background mask in a random position in horizontal and vertical directions across the silhouette sequence. Later in this chapter, we refer to this as random occlusion horizontally (RandH) and random occlusion vertically (RandV), respectively. For each type of occlusion pattern, we added 30%, 40%, and 50% degrees of occlusion for a silhouette against the full area. As a result, we simulated a total of 24 (i.e., 8×3) occlusion patterns. Fig. 4.4 shows the simulated occluded silhouette sequence for a subject.

4.3.3 Experimental settings

We divided the total subjects list randomly into three disjoint sets of approximately equal size: training (i.e., subjects 3,001); validation (i.e., subjects 3000) and test (i.e., 3,000 subjects). Then, the test and validation sets were divided into two subsets: the gallery set and probe set. The validation set was used to select the best iteration number n for experiments, whereas the test set was used to evaluate the accuracy of our proposed approach and other state-of-the-art approaches considered in this chapter. Because the number of samples was large for the experiments of unknown occlusion patterns compared with the experiments of known occlusion patterns, it took more iterations to converge. We, therefore, trained the proposed approach using a validation dataset for up to 30,000 iterations for experiments for known occlusion patterns, whereas we used 60,000 iterations for unknown occlusion patterns. It saved the learned parameter for every 3,000 iterations to select the best iteration using the validation dataset for testing. We followed the same settings for each benchmark for a fair comparison to select the best-learned model.

The OUMV dataset included multiple subsequences of more or fewer than 32 silhouette frames; therefore, we selected all the subsequences of 32 silhouette frames for training to increase the training sample, and the centered subsequences of 32 frames were used for the validation and test sets where the starting pose was not the same between the probe and gallery. We padded both sides of the width with zeros for each silhouette to make a 64×64 pixels resolution from a 44×64 to fit the network. After reconstructing a sequence, we padded it out to make it the original size (44×64) of the silhouette.

Unlike existing conditional video generative networks [52, 88], those quantitatively evaluate test samples by rating manually. We, however, evaluate the reconstructed silhouette sequence as gait recognition by using the GEI feature, we followed the same preprocessing as mentioned in section 3.2.2. If several gait cycles for a sequence of 32 frames were detected, then we chose the first gait cycle. Finally, we calculated the dissimilarity using the L_2 distance between two GEIs (i.e., probe and gallery).

4.3.4 Comparison methods

In this section, we describe the existing methods used for the evaluation of the experiments, and compare performance with our proposed approach. Each of them is a state-of-the-art approach for the generative approach. To fair comparison, we retrained the model using our simulated dataset from scratch to determine the best performing model. We used the same hyperparameters as those mentioned in the original papers.

Context Encoder (CE) [44]: we compared the results of our proposed approach with those obtained from the CE, which is a state-of-the-art method for image inpainting. The network architecture is similar to DCGAN [83]; that is, the encoder and auxiliary discriminator architecture are similar to that of the discriminator of DCGAN, and the decoder is also similar to the decoder of DCGAN. However, the dimension of the bottleneck layer is 4,000 instead of 100. We evaluated the CE by post-processing the restoration of pixels outside the occluded position for the experiment where the occlusion pattern is known.

Video GAN (VideoGAN) [88]: VideoGAN is the first generative adversarial network for video generation from random noise. The model is also capable of predicting a future frame given a conditional input frame in the encoder network. Therefore, we selected it as silhouette sequence reconstruction by changing its input to the occluded silhouette sequence into the encoder network. The architecture of the decoder is similar to that of DCGAN [83], except it is extended to the temporal direction.

Improved video GAN (iVideoWGAN) [52]: iVideoWGAN is the updated version of VideoGAN. The significant modification is that the discriminator network is replaced by a critic network, and it is trained using the framework of WGAN with gradient penalty [54].

In addition to the methods as mentioned above, we evaluated our proposed generator network using the training of a critic network with WGAN and WGAN-hinge loss. Later in this chapter, we refer to them as sVideoWGAN and sVideoWGAN-hinge, respectively. Similarly, we evaluated the proposed critic network (WGAN-hinge) with the generator networks of iVideoWGAN [52]. We analyzed how the proposed critic could supervise the generator to update the parameter to reconstruct the silhouette sequence. Later, we refer to it as iVideoWGAN-hinge.

4.3.5 Experiment for the known and same occlusion pattern

In this section, we analyze accuracy for gait recognition using the reconstructed silhouette sequence in which the occlusion pattern is the same between a matching pair (i.e., the probe and gallery). To prepare such experiments, we selected typical occlusion patterns from artificially simulated relative dynamic-type occlusion, such as RDLR and RDBT, with the highest and lowest degrees of occlusion (i.e., 30% and 50%). Therefore, we consequently prepared four subsets of occlusion patterns, denoted by RDLR_30, RDLR_50, RDBT_30, and RDBT_50, where the first and second subscripts indicate respective for the type of occlusion and degree of occlusion. For the evaluation, the training sets for each subset were prepared in the same manner to reflect the corresponding test sets, so that the occlusion pattern is known.

Figs. 4.5 and 4.6 show the reconstructed silhouette sequences for the occlusion patterns of RDLR_50. From these reconstructed silhouette, sVideoWGAN-hinge, VideoWGAN-hinge and iVideoWGAN-hinge could reconstruct the silhouette sequence well. Moreover, we can observe that the reconstructed silhouette sequence by comparing with ground truth, sVideoWGAN-hinge is similar with that of sVideoWGAN. We explain the causes in section 4.3.7.

The results for CMC and ROC are shown in Fig. 4.7 while Rank-1, Rank-5 and EER are shown in Table 4.1. From these results, we can see that our proposed generator with the proposed critic (i.e., sVideoWGAN-hinge) outperformed the existing state-of-the-art benchmark methods in all settings. Besides, the proposed generator and proposed critic improved accuracy individually. For example, if we compare the proposed generator and the generator for VideoGAN [52] with the critic of WGAN, referred to as sVideoWGAN and iVideoWGAN, respectively, then accuracy improved from 80.8% to 81.9% and 6.2% to 6.1% (see Table 4.1) for the Rank-1 and EERs, respectively, for the occlusion pattern of RDLR_30, and 71.3% to

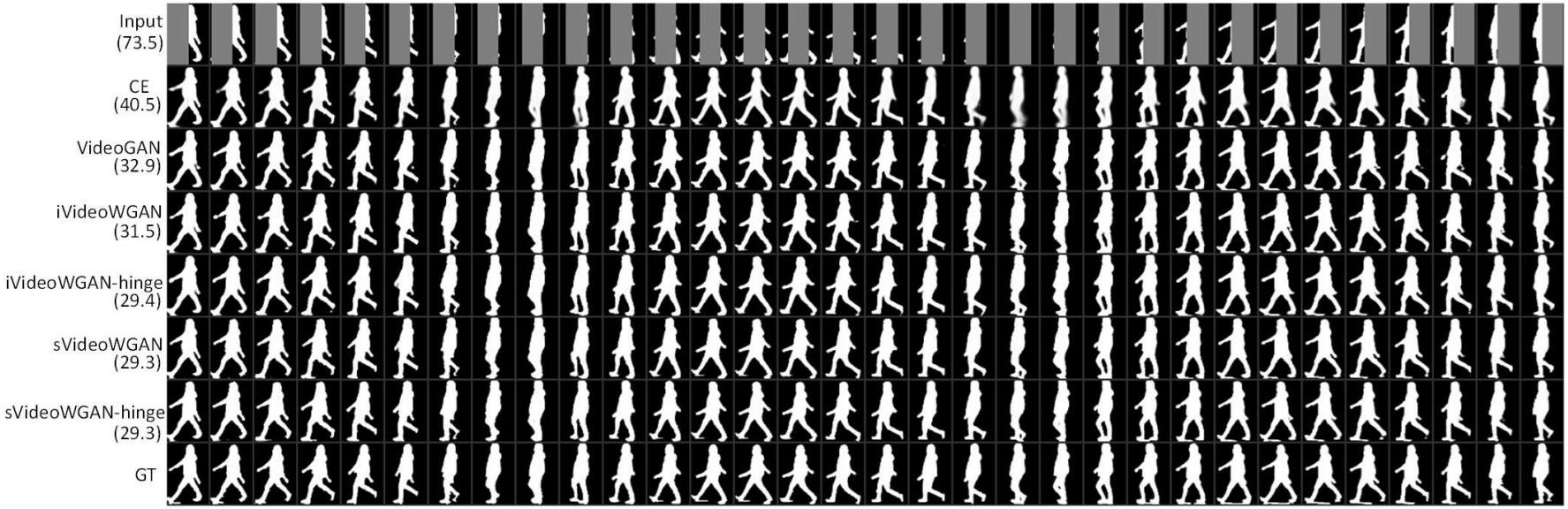


Figure 4.5: Reconstructed silhouette sequence (all frames) for the experiment for the known and same occlusion pattern for RDLR_50. The left-hand side of the figure: second to seventh rows show the labels for the benchmark used to reconstruct the silhouette sequence, whereas the first and last rows show the input and GT, respectively. Values in the parentheses under each label show the average L_2 distance for the reconstructed and the ground truth sequence. Occluded areas are gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background.

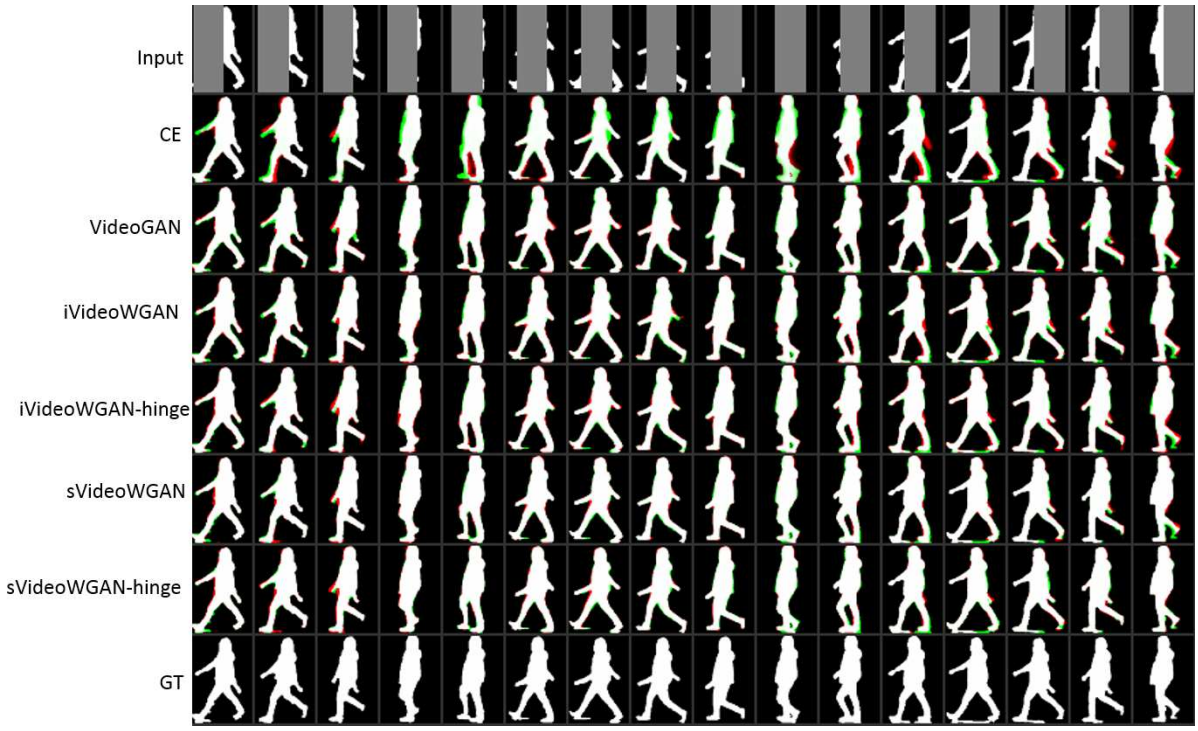


Figure 4.6: Reconstructed silhouette sequence (every second frame) for the experiment for the known and same occlusion pattern for RDLR_50 to show how a benchmark can reconstruct silhouette sequence. Green and red color indicate falsely reconstructed and falsely unreconstructed pixels, respectively compared with GT. The left-hand side of the figure: second to seventh rows show the labels for the benchmark used to reconstruct the silhouette sequence, whereas the first and last rows show the input and GT, respectively. Occluded areas are gray only for visualization purposes; in the experiment, we masked the occluded area with black, namely the values of the masked area are set to zero; this value is the same for the background.

74.7% and 7.4% to 6.8% for RDLR_50. Similarly, accuracy improved for the proposed generator network from 81.4% to 82.4% and 6.1% to 6.0% for Rank-1 and EERs, respectively, for RDLR_30, and 73.2% to 75.9 and 6.8% to 6.6% for RDLR_50 while the critic was trained with WGAN-hinge. By contrast, the proposed critic WGAN-hinge also (i.e., incorporating hinge loss in WGAN) improved accuracy independently, for example, 81.9% to 82.4% and 6.1% to 6.0%, for Rank-1 and EERs, respectively, while the generator network was proposed for the type of occlusion pattern of RDLR_30.

Regarding existing benchmarks, the reconstructed silhouette sequence from the CE benchmark looks blurred and easy to identify occluded area because it reconstructed only the occluded area frame by frame, which led to a bad recognition accuracy compared with other benchmarks. Notably, it achieved worse accuracy for a high degree of occlusion. Although iVideoWGAN used an identical generator network to VideoGAN, it improved accuracy for

each experiment because the WGAN loss supervised the generator network better than that of the discriminator of DCGAN.

Table 4.1: Rank-1/5 [%] and EER [%] for the experiment for the known and same occlusion pattern. Bold and italic bold fonts indicate the best and second best accuracies throughout the work in this chapter, respectively.

Reconstruction method	RDLR_30 vs RDLR_30			RDLR_50 vs RDLR_50			RDBT_30 vs RDBT_30			RDBT_50 vs RDBT_50		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	57.4	70.0	9.5	30.9	39.6	18.6	66.4	79.4	8.4	52.2	66.8	10.8
CE	72.6	83.8	7.2	54.3	70.5	9.5	76.1	84.9	7.0	66.0	77.6	8.5
VideoGan	79.7	87.5	6.3	70.9	81.8	7.3	80.7	87.9	6.0	74.7	84.6	6.7
iVideoWGAN	80.8	87.8	6.2	71.3	82.0	7.4	80.5	88.0	6.0	75.2	84.8	6.7
iVideoWGAN-hinge	81.4	88.1	***6.1***	73.2	83.8	***6.8***	81.4	88.0	**5.8**	75.9	84.9	6.7
sVideoWGAN	***81.9***	***88.5***	***6.1***	***74.7***	***84.5***	***6.8***	***82.0***	***88.7***	6.0	***76.8***	***85.4***	***6.6***
sVideoWGAN-hinge	**82.4**	**88.6**	**6.0**	**75.9**	**85.4**	**6.6**	**82.5**	**89.0**	***5.9***	**77.3**	**86.3**	**6.2**

4.3.6 Experiment for the known but different occlusion pattern

In this section, we analyze the gait recognition accuracy using the reconstructed silhouette sequence in which the occlusion pattern is different for a probe sample from gallery samples. To prepare such experiments, we selected occlusion patterns with the same occlusion type but different degrees of occlusion, and different occlusion types with different degrees of occlusion. Specifically, we compared the accuracy of gait recognition of RDLR_30 against RDLR_50 and RDLR_30 against RDBT_50. For the performance evaluation, in the same way as the previous experiments, the training sets for each experiment were prepared to reflect the corresponding test sets.

Fig. 4.8 shows the CMC and ROC, and Rank-1, Rank-5 and EER are shown in Table 4.2. From these results, we can observe that the recognition accuracy w/o reconstruction drastically changed because of the appearance change between different occlusion patterns. However, the tendency of recognition accuracy for other benchmarks was the same as the known and same occlusion pattern experiments.

4.3.7 Experiment for the unknown occlusion pattern

In the previous sections, we analyzed the gait recognition accuracy from the reconstructed silhouette sequence for the same and different occlusion patterns, and the parameters are trained using the occlusion pattern to reflect the test sample. Therefore, we assume that the occlusion pattern is known in advance. However, it is difficult to collect such data from a real-world perspective because of the uncooperative and non-intrusive nature of gait biometrics. We, therefore, analyze the accuracy of gait recognition when the occlusion pattern is unknown. For this

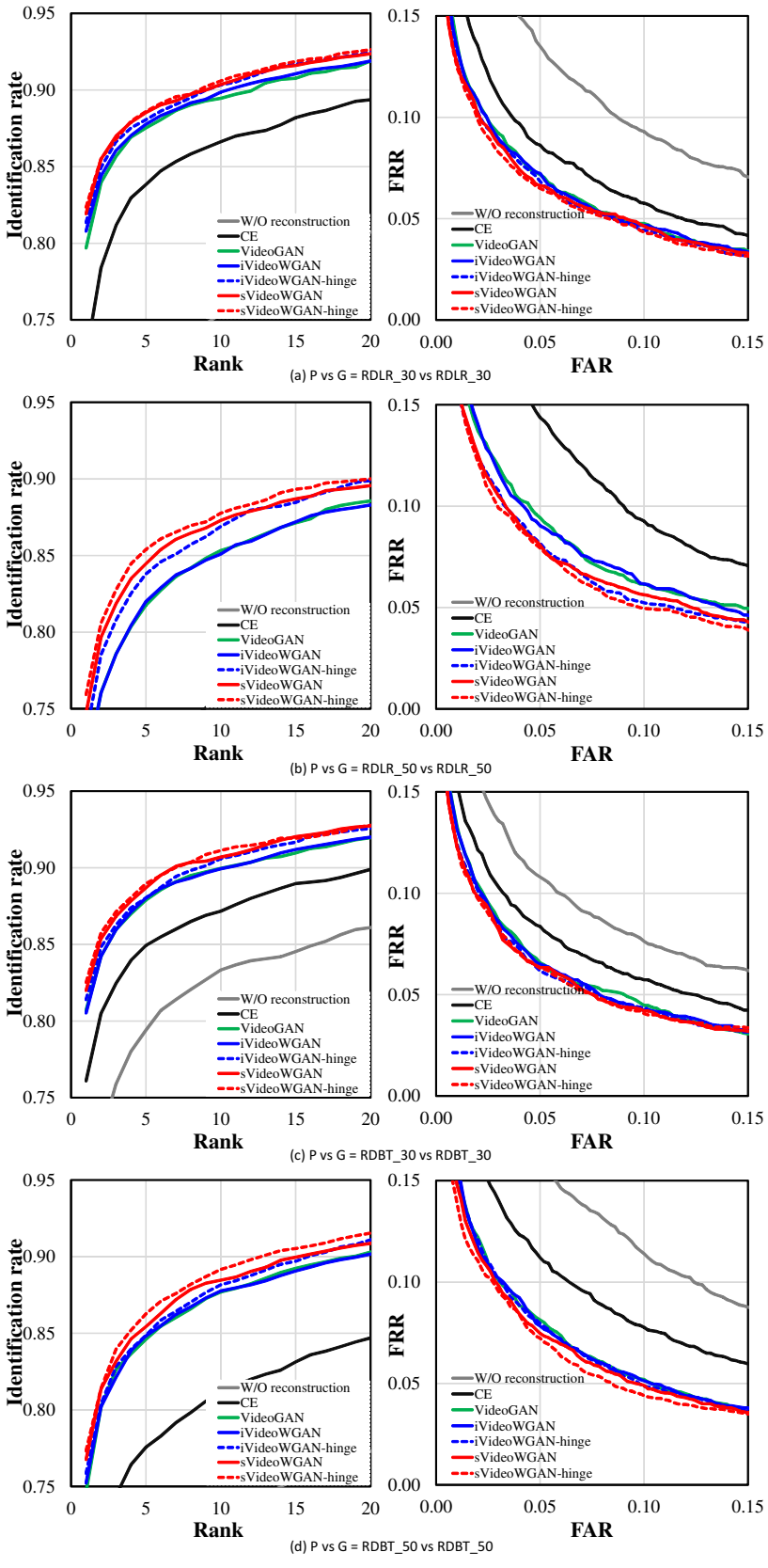
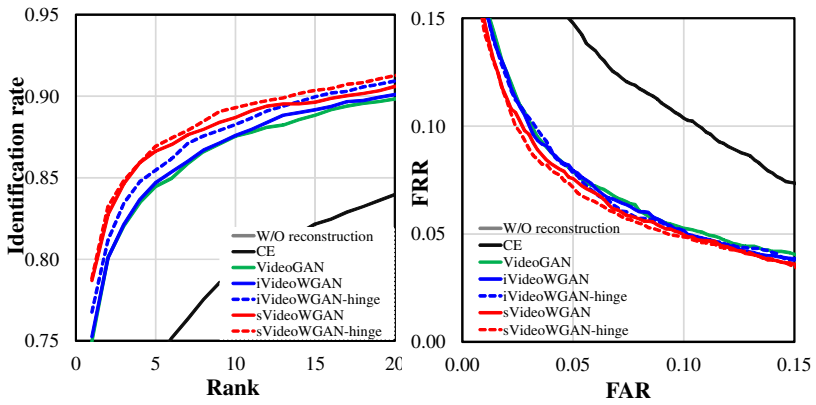
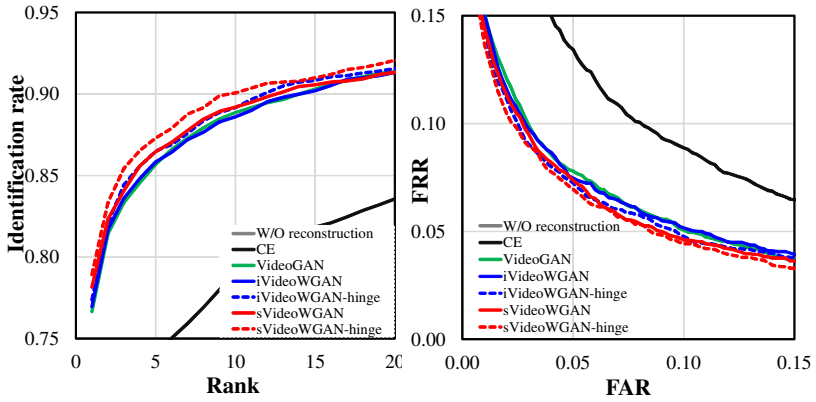


Figure 4.7: CMC and ROC curves for the different experiments for the known and same occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means that the occlusion pattern of the probe and gallery whereas RDLR_XX and RDBT_XX indicate that the relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves.



(a) P vs G = RDLR_30 vs RDLR_50



(b) P vs G = RDLR_30 vs RDBT_50

Figure 4.8: CMC and ROC curves for the different experiments for the known but different occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means that the occlusion pattern of the probe and gallery whereas RDLR_XX and RDBT_XX indicate that the relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves.

purpose, we trained the parameter of our proposed approach as well as other benchmark networks by considering all occlusion patterns of training sets to make a robust model that was capable of reconstructing any occlusion pattern. For testing, we employed the *cooperative and uncooperative setting*, and the *unknown but the same and different occlusion patterns*.

In this section, we investigated the impact of the cooperative and uncooperative settings for recognition accuracy. The implicit assumption for the cooperative setting is that the covariate condition is uniform in a gallery set. However, it is challenging to collect such types of data in a real-world scenario because of the uncooperative and non-intrusive nature of gait biometrics. In addition to the cooperative setting, we, therefore, selected a more natural uncooperative setting where the covariate condition was inconsistent in the gallery set [75].

Cooperative and uncooperative setting

We have already mentioned in the section 3.3.3 the implicit assumption of the cooperative and uncooperative setting is the covariate condition, respectively, as consistent and inconsistent for gallery samples. However, in this experiment, we considered the occlusion as a covariate. More specifically, the uncooperative setting is that the occlusion pattern is inconsistent for all samples throughout the probe and gallery sets [75] (i.e., the occlusion pattern is unknown), whereas, for the cooperative setting, the occlusion pattern is consistent for all samples in a gallery set. To create such an uncooperative setting, occlusion patterns were randomly selected for each subject for the probe and gallery sets. In contrast, for the cooperative setting, ground truth samples were used in the gallery set.

The results for the cooperative and uncooperative settings for CMC and ROC are shown in Fig. 4.9 while Rank-1, Rank-5 and EER are shown in Table 4.3. The accuracy of the cooperative setting was better than that for the uncooperative setting for each of the benchmarks. We can observe that the accuracy of CE degraded drastically from the cooperative to uncooperative settings compared with other benchmarks. For example, CE degraded the Rank-1 identification by 12%, whereas the maximum degradation for a benchmark was 8.2% (e.g., for iVideoWGAN-hinge). We believe that CE reconstructed the silhouette sequence frame by frame and therefore lost the motion information, especially when a silhouette was completely occluded, as shown in Figs. 4.5 and 4.6. As a result, CE lost subject discrimination.

In addition, we can see that sVideoWGAN-hinge did not improve accuracy from sVideoWGAN for the cooperative setting. We believe that the proposed generator network used the element-wise addition of the encoder with the decoder to keep the unoccluded silhouette in the reconstructed silhouette as much as possible, and WGAN supervised the generator to reconstruct by comparing the reconstructed sequence with the ground truth sequence. However, the proposed critic (WGAN-hinge) guided the generator by comparing not only the ground truth but also positive and negative reference sequences. Therefore, the reconstructed silhouette sequence by comparing with ground truth, sVideoWGAN-hinge is similar or slightly worse than that of sVideoWGAN, as shown in Figs. 4.5 and 4.6.

Unknown but the same and different occlusion pattern settings

Because the learned parameter for the unknown occlusion pattern experiments can reconstruct any occlusion pattern, we choose the same and different occlusion patterns between the probe and gallery for evaluation. Thus, we selected the RDLR_30 occlusion pattern as the probe; two

Table 4.2: Rank-1/5 [%] and EER [%] for the experiment for the known but different occlusion pattern.

Reconstruction method	RDLR_30 vs RDLR_50			RDLR_30 vs RDBT_50		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	0.8	2.2	39.6	2.0	5.5	29.2
CE	55.2	73.7	10.3	58.3	73.8	9.2
VideoGan	75.0	84.5	6.8	76.7	85.7	6.7
iVideoWGAN	75.3	84.7	6.6	77.0	85.8	6.6
iVideoWGAN-hinge	76.8	85.5	6.6	77.4	***86.5***	6.4
sVideoWGAN	***78.7***	***86.6***	***6.5***	***78.2***	***86.5***	***6.3***
sVideoWGAN-hinge	**78.8**	**86.9**	**6.3**	**78.9**	**87.3**	**6.2**

Table 4.3: Rank-1/5 [%] and EER [%] for the experiment for cooperative and uncooperative settings for the unknown occlusion pattern.

Reconstruction method	Uncooperative			Cooperative		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	3.9	5.0	46.4	7.4	11.0	45.5
CE	42.1	59.2	11.5	54.1	68.9	9.8
VideoGan	56.8	71.4	8.9	64.1	75.6	8.3
iVideoWGAN	58.4	72.8	8.8	65.4	77.4	8.1
iVideoWGAN-hinge	59.7	73.8	**8.3**	67.9	78.7	7.7
sVideoWGAN	***63.4***	***75.6***	8.5	***69.8***	**80.1**	**7.6**
sVideoWGAN-hinge	**64.7**	**76.5**	***8.4***	**70.5**	***79.9***	**7.6**

typical occlusion patterns for each type of relative dynamic occlusions, such as RDLR_30 and RDLR_50, and RDBT_30 and RDBT_50, together with the ground truth silhouette sequence as the gallery. Therefore, we could compare accuracy for the unknown occlusion patterns with known occlusion patterns.

Fig. 4.10 shows the results for CMC and ROC, while Table 4.4 shows the Rank-1, Rank-5 and EER. From these results, we can see that the recognition accuracy for CE degraded for each combination when compared with that of the combination from the known occlusion pattern. For example, Rank-1 and EER were 72.6% and 7.2%, respectively, when the occlusion pattern was known for RDLR_30 versus RDLR_30, and 70.7% and 7.4% for the unknown occlusion pattern. We believe that, because the occlusion pattern was unknown and we, therefore, did not know the occlusion position to replace the original unoccluded input pixel in the output as post-processing, the reconstructed silhouette sequence for the experiment for the unknown occlusion pattern is worse than that of known occlusion pattern. Similar to the results for the experiment of a cooperative setting, sVideoWGAN-hinge did not improve accuracy from sVideoWGAN for RDLR_30 versus GT (See Table 4.4).

In addition, identification accuracy was degraded for VideoGAN, iVideoWGAN, and iVideoWGAN-hinge when compared with the same combination for the known occlusion pattern; however, the verification accuracy improved. We think that those benchmarks used the same generator network of comparatively shallow architecture and therefore lost inter-subject discrimina-

tion ability because identification is performed based on one-to-many matching and accuracy depend on the relative distance; whereas they retain the intra-subject discrimination ability because verification is performed based on one-to-one matching and accuracy depend on absolute distance.

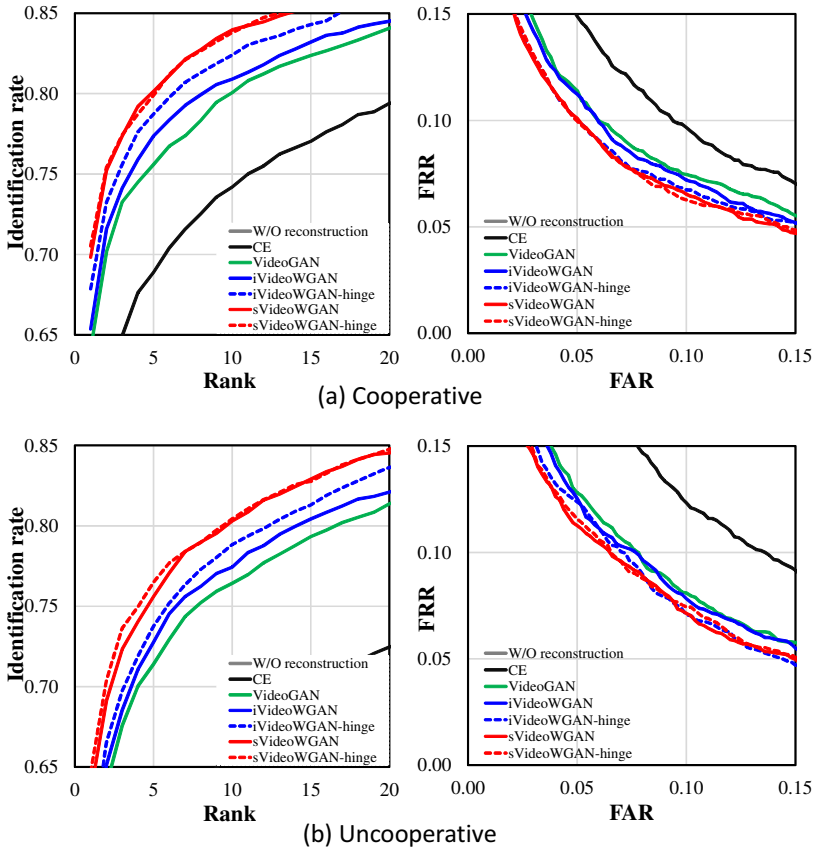


Figure 4.9: CMC and ROC curves for the experiment for cooperative and uncooperative settings for the unknown occlusion pattern. The left side shows the CMC curves, and the right side shows the ROC curves. Note that some benchmarks do not provide curves.

Table 4.4: Rank-1/5 [%] and EER [%] for the experiment for the unknown but same and different occlusion pattern settings.

Reconstruction method	RDLR_30 vs RDLR_30			RDLR_30 vs RDLR_50			RDLR_30 vs RDBT_30			RDLR_30 vs RDBT_50			RDLR_30 vs GT		
	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER	Rank-1	Rank-5	EER
W/O reconstruction	57.4	70.0	9.5	0.8	2.2	39.6	6.1	15.4	25.6	2.0	5.5	29.2	1.3	2.6	36.7
CE	70.7	82.4	7.4	51.3	70.5	10.7	66.4	78.7	8.1	52.5	69.9	9.9	62.9	77.4	8.1
VideoGan	78.0	86.3	6.2	72.9	83.5	6.8	78.0	86.2	6.2	74.3	84.6	6.7	75.5	85.0	6.5
iVideoWGAN	79.4	87.2	5.9	74.5	84.7	6.5	78.9	87.5	6.0	74.2	85.0	6.5	78.1	85.7	6.4
iVideoWGAN-hinge	80.0	87.8	***5.8***	76.3	84.9	6.5	79.8	87.4	6.0	76.3	85.6	6.3	78.6	86.6	**6.2**
sVideoWGAN	***82.8***	***89.3***	***5.8***	***78.2***	***86.6***	***6.3***	**82.8**	***89.2***	***5.8***	***79.4***	**88.2**	***6.0***	**80.6**	**88.0**	**6.2**
sVideoWGAN-hinge	**83.1**	**89.4**	**5.5**	**78.5**	**86.8**	**6.2**	*82.7*	**89.3**	**5.7**	**79.8**	*87.5*	**5.9**	*80.5*	**88.0**	**6.2**

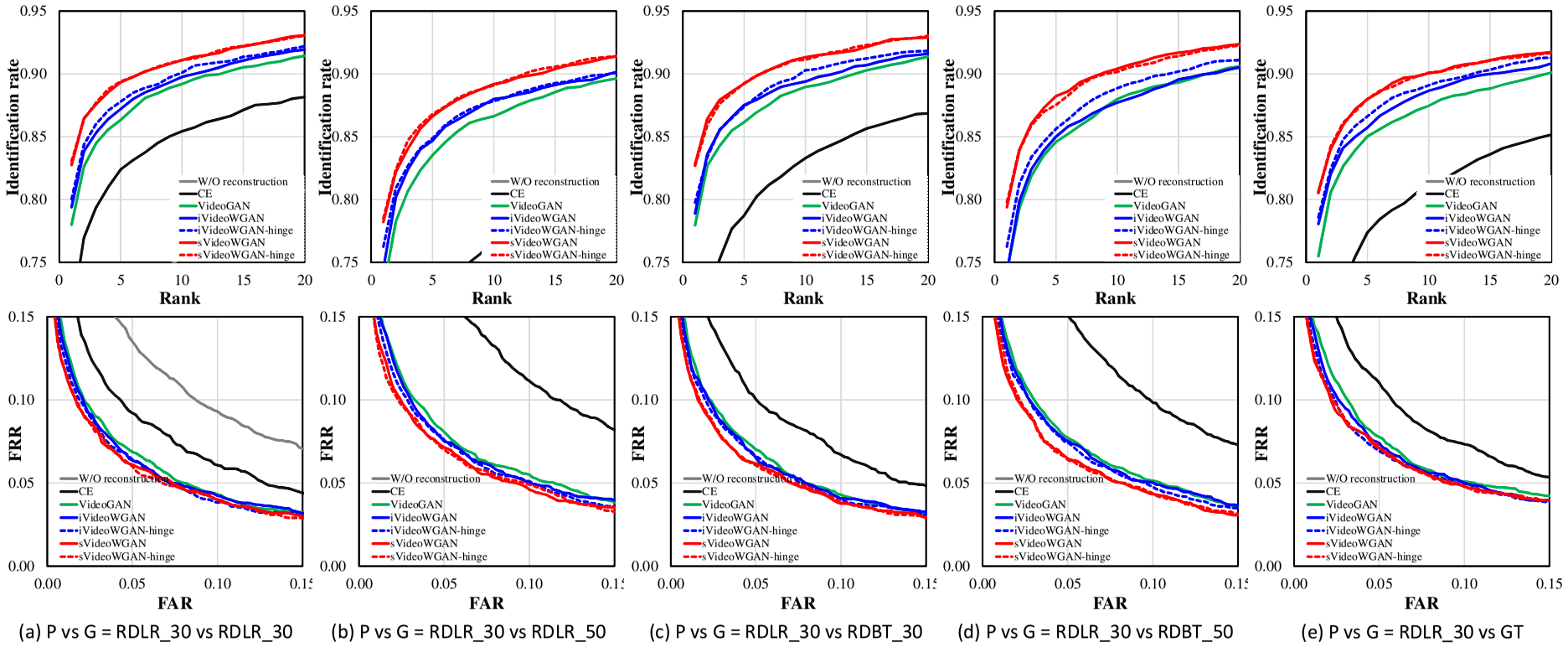


Figure 4.10: CMC and ROC curves for the experiment for the unknown but same and different occlusion pattern settings. The left side shows the CMC curves, and the right side shows the ROC curves; P vs G means that the occlusion pattern of the probe and gallery whereas RDLR_XX and RDBT_XX indicate that the relative dynamic occlusion left to right and relative dynamic occlusion from bottom to top, respectively along with the degree of occlusion (XX%). Note that some benchmarks do not provide curves.

4.4 Conclusion

In this chapter, we focused on gait recognition, where all frames in a sequence were occluded by various patterns of occlusions. For this purpose, we presented an approach based on deep conditional GAN that consisted of a generator network and a critic network. It allowed us to reconstruct an unoccluded image from an occluded silhouette sequence for gait recognition. We demonstrated that our proposed triplet hinge loss along with WGAN regularized the training of the generative network and reconstructed the silhouette sequence with a high discrimination ability, which led to better accuracy for gait recognition. To demonstrate the effectiveness of the proposed approach, we considered 24 occlusion patterns. These are divided into relative dynamic and relative static occlusions for different degrees of occlusions that are common in real-world scenarios. We designed a set of experiments in which the occlusion patterns between the probe and gallery were the same/different and known/unknown. The experimental results demonstrated that the reconstructed silhouette sequence of the proposed approach achieved state-of-the-art accuracy.

While we simulated occlusion patterns artificially and were derived from a single view dataset in this work, it is necessary to collect a dataset in a real-world considering different occlusions, as mentioned in section 3.4. Then, we need to explore the proposed network to overcome the effect of occlusion in the real-world image sequence. In addition, we assumed that the bounding box of the occluded subjects is given when simulated the occlusion pattern. However, for the localization, a deep learning-based approach can be applicable for real-world image sequences with various occlusions.

Chapter 5

Multi-quality and multi-modal biometrics

5.1 Introduction

In chapters 3 and 4, we investigated the gait recognition against carried object (CO) and occlusion by considering unimodal biometric system (i.e., a single biometric trait is used for recognition). Although unimodal biometrics are the mainstream biometric system because of their simple configurations and high usability. These systems are, however, commonly affected by practical problems such as noisy sensor data, non-universality, and vulnerability to spoofing attacks and poor recognition accuracy [57]. One possible solution to these problems is to use multiple biometric traits for recognition (i.e., multi-modal biometrics) [91]. A wide variety of biometric modality combinations have been considered for recognition in the literature, including iris and face [142], face and ocular [143], face and gait [55, 56], and face with gait and height [58].

Additionally, some of the quality information or *quality measures* [144] that are associated with the biometric samples affect the system's discrimination capabilities, even though they do not provide the capability to identify the subject by themselves. Many quality measures that significantly affect the recognition accuracy have been reported in the literature to date. These quality measures mainly fall into two families: sample-based quality, and sensor-based quality. Examples of sample-based quality measures include the degree of occlusion and/or blur, and texture richness in iris recognition [95, 145]; brightness, contrast, and illumination for fingerprint and face recognition [95]; and the view for face and gait biometrics [33]. Corresponding examples for a camera sensor-based quality include image size or spatial resolution (SR) for image-based biometric systems such as face recognition systems [95], and frame rate (i.e., temporal resolution) for video-based biometrics such as gait recognition systems [146].

Multi-modal biometrics fusion can be done at various level as discussed in section 2.5 and summarized in Table 2.3. Among these methods, score-level fusion is the most popular ap-

proach because it offers reasonable recognition accuracy and comparatively simple implementation, and thus many researchers in the multi-modal research community work in the score-level fusion field [59, 60, 61, 62, 63, 64]. Moreover, the quality measures play an important role in score-level fusion systems and have been used to improve recognition accuracy [144, 147].

In the multi-modal research community, there is no multi-quality score database available from a single camera sensor considering gait, face, and height biometrics as discussed in section 2.5.1 and summarized in Table 2.2. Therefore, in this chapter, we introduce a very-large-scale multi-quality multi-modal biometric score database [132] to analyze quality-dependent score-level fusion considering the gait, head, and height modality from a single camera sensor. Besides, we consider a spatial resolution (SR), temporal resolution (TR), and view, which all significantly affect the matching scores. An earlier preliminary version of this work was published in [7] and the extensions from it can be summarized in the following points:

- We constructed our database by introducing the multi-view concept, whereas the previously released database [7] considered only the side-view, and discussed use of the view as a quality measure; in addition, we applied some pre-processing steps on the feature level and have called this database "MultiQ Score Database version 2".
- We have added an analysis of how the quality measures affect the recognition accuracy of each modality.
- We have added the RankSVM as a new benchmark to enable further discussion of the evaluation results. Consideration of the RankSVM enables analysis of the recognition accuracy in greater depth.

5.2 Multi-quality and multi-modal biometric score database

5.2.1 Overview of the OULP dataset to extract multi-modal matching score

The OULP dataset [23] was collected with the intent to statistically reliable performance evaluation of gait recognition. To capture gait image sequence, each subject was asked to walk along a predetermined walking course naturally twice, and two walking image sequences were captured for each subject using a single camera placed at a distance of 5 m. The spatial and temporal resolution and spatial resolutions were 640 $\times$ 480 pixels and 30 fps, respectively. Each captured image sequence was divided into four segments based on observation azimuth

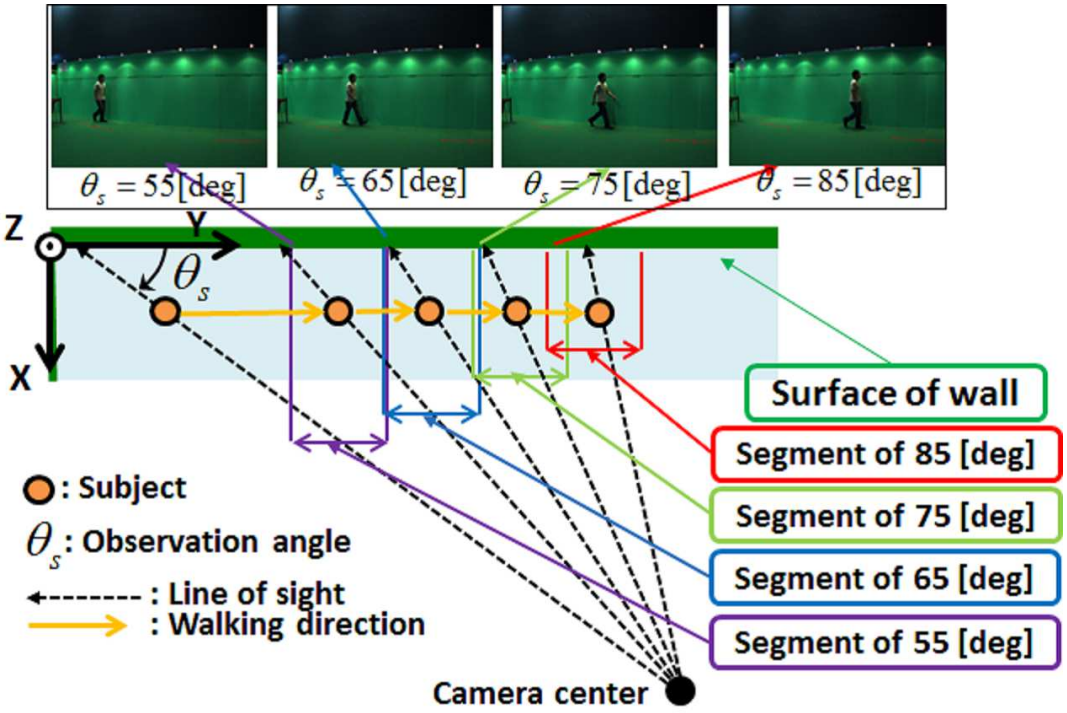


Figure 5.1: Illustration of the data collection system in the OULP Dataset.

angles of 55, 65, 75, and 85 deg. An overview of the captured images is shown in Fig. 5.1. We choose a subset of 1,912 subjects to generate matching scores for this research.

Silhouette sequences were extracted initially from the captured image sequences using a background subtraction-based graph-cut segmentation method [148]. Then, each silhouette was checked manually and modify if necessary. Head region image features were extracted from the captured image sequences using the information from the silhouette. Lens distortion removal, rectification, and cropping processes were used to generate the gait and height features. Rectification was performed using the parameters of camera calibration such that the horizontal and vertical axes (i.e., x and y) of the image plane are located parallel to the walking for a subject and vertical directions of the real 3D world, respectively. Therefore, calculation of the subject's height becomes simple; the vertical foot positions and the head top can be obtained from the bounding box, and the actual height of the subject in the real 3D world can then be easily estimated by a conversion using the parameter of the camera calibration because the distance between the subject and the camera on the walking course was the same for each subject in the dataset.

5.2.2 Quality measures

We focus on the spatial resolution (SR), the temporal resolution (TR), and the view as factors that affect the matching score of an algorithm for each feature, and generate a score database with different qualities using image sequences with different values of these factors. To generate such a database, we first generated datasets with various qualities by downsampling the image sequences down in terms of their SRs and TRs. Then, we extracted the required gait, head, and height features. Finally, we calculated the matching scores for each modality(i.e., the gait, the head, and the height) using these extracted features. To generate different qualities, we considered the following SRs and TRs.

SR: To simulate the image sequences with different spatial resolution and/or image sequences of subjects at different distances from a camera[1], We set scaling factors to downsample the originally captured image sequences such that the average height of all subjects does not fall below 20 pixels because silhouette extraction is difficult when the SR is extremely low. More specifically, we downscaled the original images as the factors of 1/2, 1/3, 1/4,s 1/5, 1/6, and 1/8. We therefore, prepared image sequences with sizes of 640×480, 320×240, 213×160, 160×120, 128×96, 106×80, and 80×60 pixels. Consequently, image sequences with seven distinct SRs were used to construct the score database.

We followed a few preprocessing to get better-simulated image sequences. Firstly, the previously release score database [7] used nearest-neighbor interpolation for down-sample, whereas we used area interpolation implemented in OpenCV [149]; area interpolation with thresholding is more appropriate down-sampling than nearest-neighbor. Thus, we get gray-scale values instead of binary values (e.g., background or foreground) because of the area interpolation. So, we applied thresholding to get it to binary silhouette image sequence after down-sampling. Moreover, because the boundary of down-sampling by the area interpolation (e.g., pixels whose horizontal or vertical position is a multiple of k for down-sampling with factors of $1/k$) does not necessarily coincide with the bottom of foot or the top of the head, we randomly shift-up/down the boundary of down-sampling for each subject to better simulate the walking position differences among subjects. More specifically, we generated a random number for each subject to shift the silhouette image sequences up or down such that the top of the subject's head or the bottom parts of their feet are not moved outside the image. The entire silhouette image sequence for each subject was then shifted up/down using the set value, and the same process was applied to each subject.

[1]Because the distance from the camera to the subject is adequately large compared with the subject size, the assumption of weak perspective projection concerning the subject can be almost true. Thus, we can simulate image sequences of the same subjects at different distances.

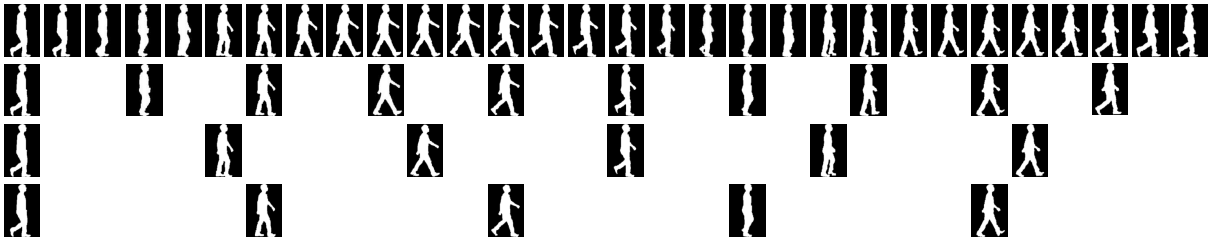


Figure 5.2: Full cycle normalized silhouette sequences at various TRs from 85 deg view. The top row shows normalized silhouette sequences at 30 fps; the second row shows normalized silhouette sequences at 10 fps; the third row shows normalized silhouette sequences at 6 fps; and the bottom row shows normalized silhouette sequences at 5 fps. The silhouette sequences were taken from frame-skipped image sequences that started from the first frame.

TR: To simulate image sequences at different TRs, we prepared ten scaling factors, including the original frame rate. We, therefore, chosen images at specific frame intervals from the original 30 fps image sequences, we called frame-skipped sequences in this chapter. We subsequently generated image sequences with frame rates of 15, 10, 7.5, 6, 5, 3.75, 3, 2, and 1 fps. Examples of these silhouette sequences with 30, 10, 6, and 5 fps frame rates are shown in Fig. 5.2.

View: The image sequences of the OULP dataset are divided into four observation azimuth angle, as shown in Fig. 5.1. We, therefore, used the image sequences with observation azimuth angles of 55, 65, 75, and 85 deg, and use the view information as a quality.

5.2.3 Matching algorithm for score calculation

We calculated matching scores for the gait, head, and height features to construct the score database. We briefly explain the feature extraction and score calculation processes for the gait, head, and height features.

Gait matching: As we mentioned in section 3.2.2, GEI [20] is the most widely used feature in gait recognition, we used the GEI as a gait feature. For GEI extraction, we followed the same procedure, as mentioned in section 3.2.2. Some examples of GEIs with different views and SRs are shown in Fig. 5.3 and GEIs with different TRs are shown in Fig. 5.4. For the gait-based matching scores, we used the DM method, as discussed in section 3.3.2.

Head matching: Many sophisticated face recognition approaches have been proposed in the literature, and they generally consider the inner region of the face for recognition. However, these approaches often cannot achieve reasonable accuracy when the targeted face is extremely small [150, 151]. We, therefore, considered the texture information from the head region,

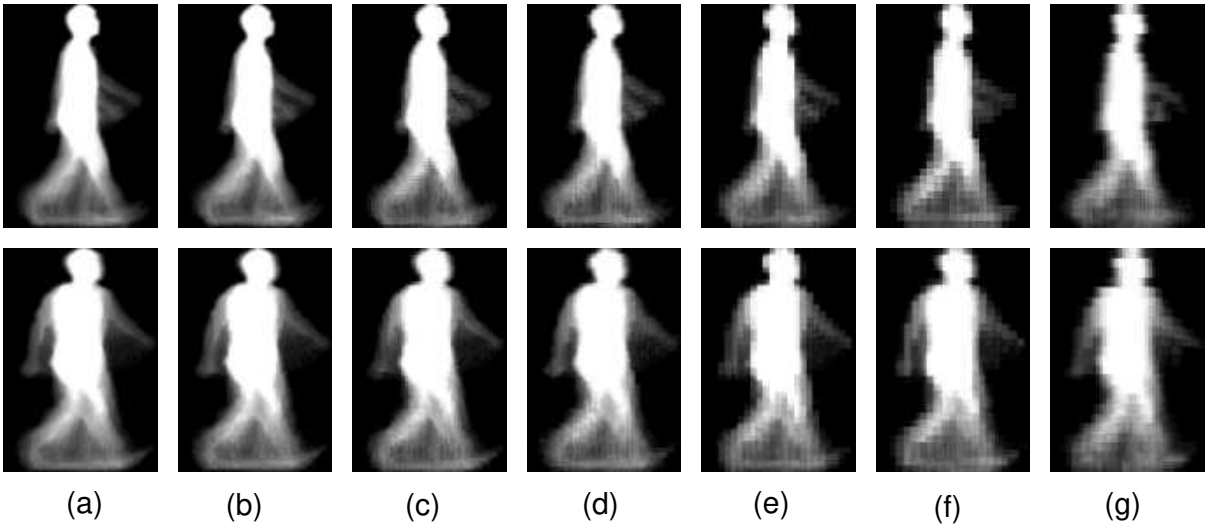


Figure 5.3: GEIs extracted from 30 fps images of various sizes from 85 and 55 deg views. The top row show GEIs with the 85 deg view and the bottom row shows GEIs with the 55 deg view. Image sizes: (a) 640 $\times$ 480, (b) 320 $\times$ 240, (c) 213 $\times$ 160, (d) 160 $\times$ 120, (e) 128 $\times$ 96, (f) 106 $\times$ 80, and (g) 80 $\times$ 60 pixels. The sizes given are not for the bounding box, but are for the original/scaled-down images.

including the inner face region, the hair and the face contour parts in this research, and call it the head feature of the target subject because our feature differs from the general face feature. Specifically, we have defined the upper region above the neck of a target person as the head region; this concept is similar to that of [152], which extends the iris feature into a periocular feature.

We calculated the head matching score using the following steps. Firstly, we used the silhouette image associated with the target image as a mask and localized the head region as aforementioned. Then, we extracted the image of the head region and set it as a template for the head feature of the image. Later, we extracted a template from each frame of the probe image

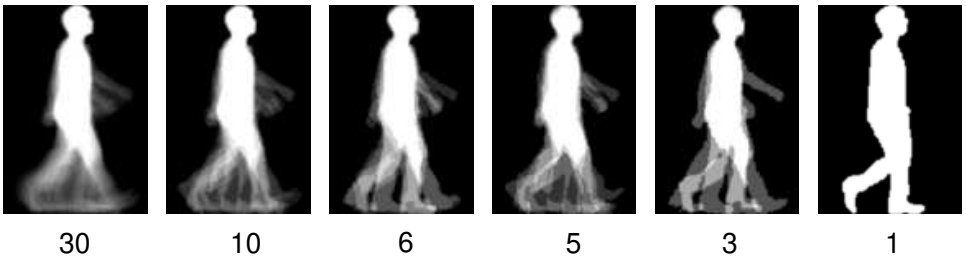


Figure 5.4: GEIs extracted from fixed size (640$\times$480 pixels) normalized silhouette sequences at various TRs [fps].

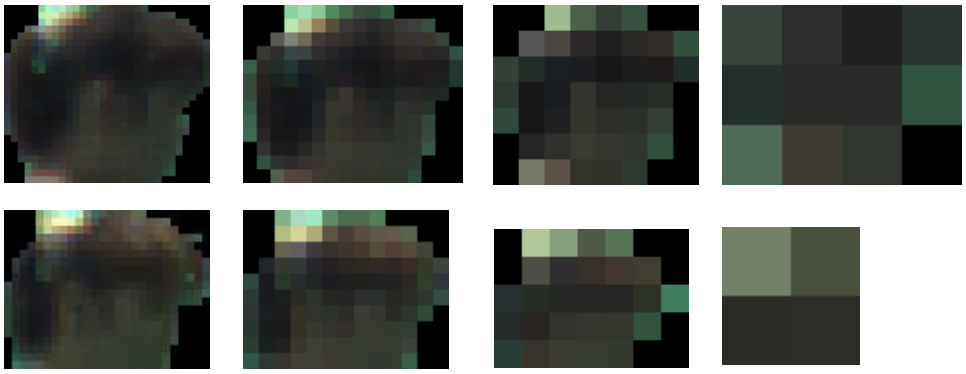


Figure 5.5: Head templates extracted from both the original images and the scaled-down images. The top row shows template images with the 85 deg view and the bottom row shows template images with the 55 deg view. The head template size is dependent on both subject and view; in this figure, the template sizes of the original image are 30×26 and 26×20 pixels for the 85 and 55 deg views, respectively.

sequences separately for each quality. Finally, we applied the conventional template matching algorithm using the template to the gallery image sequences and then calculate a head matching score. For the score calculation process, we use the color texture information. Let F_{p_i} be the template of the head feature associated with the i-th frame of the probe. Let $F_{g_{j,k}}$ be a masked image of the same size that is associated with the j-th frame and the k-th spatial displacement within the gallery search regions. We then calculate the head matching score using correlation-based template matching using

$$S_{head} = \min_{i,j,k}[1 - f_{NCC}(F_{p_i}, F_{g_{j,k}})], \tag{5.1}$$

were, $f_{NCC}(F_{p_i}, F_{g_{j,k}})$ is an operator that is used to calculate the normalized cross-correlation (NCC) between F_{p_i} and $F_{g_{j,k}}$, and template matching was performed within the region of interest that is defined by the silhouette mask of the gallery.

Note here that we did not apply the advanced techniques that are associated with face recognition in the literature, including the feature extraction process [153], pose normalization [154], and/or face alignment [155]. In this work, we considered multiple image sequences over a wide SR range, from mid-level SR (see Fig. 5.5(a)) to extremely low-level SR (see Fig. 5.5(d)). The advanced techniques do not work consistently well on image sequences with wide SR ranges, but the NCC-based technique works stably for these image sequences. While the NCC-based technique is simple and not a state-of-the-art algorithm, it is working stability property that is necessary for the work in this research.

Height matching: Because of the rectification and each subject walked in a straight walking course at a fixed depth from the calibrated camera, the actual height of each subject in the real 3D world can easily be computed from the apparent height from the image by simply multiplying the depth from the subject with the ratio of the focal length. Actual height was computed on a frame-by-frame basis, and therefore, the height of the subject can vary with the up-and-down motion caused by walking. Thus, we define the height feature as an average of the height values calculated over an image sequence. Let Z_i be the height from the ith frame of a target image sequence, and let h be the height feature (scalar). If the target image sequence is composed of N_f frames, and then the height feature can be calculated as follows:

$$h = \frac{1}{N_f} \sum_{i=1}^{N_f} Z_i. \tag{5.2}$$

Let h_p and h_g be the heights of the subjects in the probe and in the gallery, respectively. The height-based matching score S_{height} was calculated to be

$$S_{height} = |h_p - h_g|. \tag{5.3}$$

5.2.4 Constructed score database

Because we used three independent quality measures in this research, we could then set several different quality settings by different combinations of the SR, TR, and view. As a result, we constructed a multi-quality, multi-modal score database.

We generated image sequences with different TRs by a selection of specific frame intervals. Therefore, multiple image sequences were generated from a single image sequence because the selected frames must be different and are dependent on the starting frames. For example, if we consider two different image sequences of 15 fps, the image sequences an odd and even number of frames can be generated from a single image sequence with 30 fps. In this research, all frame-skipped image sequences based on different starting frame were considered as probes. In contrast, a frame-skipped image sequence associated with the first frame is considered as the gallery to avoid increase extremely large number matching scores. As a result, we used 2, 3, 4, 5, 6, 8, 10, 15, and 19 [2] frame-skipped image sequences per probe for frame rates of 15, 10,

[2]While 30 frame-skipped image sequences can be generated from 1 fps downsampling of an original 30 fps video in principle, only 19 frame-skipped image sequences are used. Because the minimum number of frames, including one gait image sequences of some subjects, is only 19, we maintain consistency among all subjects by limiting the number to 19.

7.5, 6, 5, 3.75, 3, 2, and 1 fps, respectively. Later, we mention the number of frame-skipped image sequences as N_{TR}.

Thus, we generated features for each of the N_{TR} probes and galleries separately for each modality. More specifically, we generated normalized silhouette sequences for each N_{TR} probe and the first frame-skipped image sequences for the gallery. The first frame-skipped image sequences for a subject with different TRs are shown in Fig. 5.2. Finally, we calculated the score matrices for the gait, the head, and the height between 1912 N_{TR} probes and 1912 galleries for combinations of the seven SRs, the 10 TRs variations, and the four views variations. As a result, a total of 280 different quality settings were available with the different combinations of the SRs, TRs, and views. It should be noted that the score matrices were calculated between probes and galleries with the same quality setting. As a result, we can generated on N_{TR} score distance matrices with 1912 N_{TR} genuine scores and 1912 N_{TR} $\times$ 1,911 = 3,653,832 N_{TR} imposter scores for each modality and quality, which constitute an extremely large-scale score database containing for each modality the 3,908,128 genuine scores and 7,468,432,608 imposter scores.

5.3 Experiments

5.3.1 Overview

We analyze the performance separately for each modality separately and fused in a score-level. The purpose of evaluating for each modality is to analyze the properties of the corresponding modality against the various quality conditions. Therefore, we evaluated the recognition accuracy for each modality independently for different quality settings. To analyze the score-level fusion, we design two different protocols, i.e., the quality-independent and the quality-dependent protocols, and analyze the recognition accuracies of the benchmarks under both protocols.

5.3.2 Accuracy analysis of each modality and impact analysis of each quality

In this section, we analyze the recognition accuracy of each modality under different quality conditions. Following the other works in this thesis, we evaluate the recognition accuracy in two different modes: verification and identification. We only prepared the ROC and CMC curves of all modalities under typical settings for the quality considered in Figs. 5.6 and 5.7.

In addition, we summarize EERs and Rank-1 in Table 5.1 with Fig. 5.8 and Table 5.2 with Fig. 5.9, respectively. Each quality affects the accuracy of each modality.

As we already discussed before, gait recognition is affected by dynamic motion and static shape, which incur respectively by spatial SR temporal TR. For example, we can see that, when the TR is 30 fps, gait recognition accuracy degrades as the SR decreases; the degradation is, however, moderate as verification. We believe that the temporal motion provides a reasonable discrimination ability. Therefore, a reasonable accuracy is maintained even for a low SR. Regarding middle range TR; the gait recognition accuracy degrades as the SR decreases much more clearly than that at 30 fps both for identification and verification. In this range for TR, some motion information is missing, and the shape information plays a more important role than it does in the high TR. The shape information is missing when the SR is low, and this missing information affects the recognition accuracy directly. In contrast, for low TR, the gait recognition accuracy does not vary as the SR changes. In this case, the gait feature only provides low-level information for discrimination. Therefore, this information does not decrease as the SR decreases for verification. However, some useful information is lost for identification.

Regarding TR, the gait recognition accuracy generally degrades as the TR decreases, but there are some exceptions. We can see that the Rank-1 rate at 6 fps was lower than that at 5 fps, and the rate at 3.75 fps was lower than that at 3 fps for the 85 deg view. This happened cause temporal aliasing when an image is generated by downsampling from the original image sequences. More specifically, we downsampled the original image sequences to simulate the image sequences with many fps. The gait biometric is an almost bilaterally symmetric motion. Therefore, the downsampled image sequence of the side view (i.e., 85 deg) with a specific frame interval, appears the same stance. As shown in Fig. 2, the first, second, and third silhouette images were similar to the fourth, fifth, and sixth images, respectively, for the 6 fps image sequences. Therefore the gait information in the GEI at 6fps is smaller than that at 5 fps. These are the causes of inconsistency. As for the view, significant accuracy differences cannot be observed, but accuracy at 55 deg seems to be slightly worse than that of the other view.

Head features are mostly static information. Therefore, head recognition accuracy severely affected by SR. For example, when TR is 1 fps, head recognition accuracy degrades drastically when SR becomes low. Nevertheless, head feature is also affected by TR for example, when SR is 640$\times$480 pixels, recognition accuracy improved as TR becomes high. Additionally, head modality is affected by a number of factors, such as illumination and pose. Because head features were extracted from walking image sequences, illumination and pose can be different among frames. Consequently, the sample with high TR can include more head features with

a variety of illumination and pose than data with low TR. This leads to improving accuracy. Moreover, head recognition accuracy is different depending on the view. We believe three factors can be the cause. Firstly, the texture information included in the head modality is changed depending on the observation view. Secondly, the shape of the extracted head region is also different depending on the observation azimuth angle, lastly, the size of the extracted head. Because data with different views were captured from a person using a single camera in this data set, the distance between the camera and the subject is different for each view, as shown in Fig. 5.1. As a result, this leads to the size difference.

Regarding height modality, the accuracy of height is affected not only by SR but also by TR. Because the height of the subject is calculated by averaging the height of all frames for a sequence, and the height of each frame is affected by the posture. For low TR, the height is profoundly affected by the postures of the sampled frames, and this can lead to accuracy degradation. As for view quality, the impact of view difference is not so large, because essentially, height is a view-invariant feature.

5.3.3 Experimental protocols for score-level fusion

Protocol 1: We can evaluate and analyze quality-independent score-level fusion because the constructed score database included many subsets with different qualities. Therefore, we can to compare multiple algorithms on these subsets with different qualities. This is the main reason we set Protocol 1 for the constructed score database. To prepare subsets with different qualities, we selected two typical settings for each quality; i.e., we choose high and low SRs of 640 $\times$ 480 pixels and 80 $\times$ 60 pixels, respectively, high and low TRs of 30 fps and 1 fps, respectively, and two views of 85 deg (i.e., almost a side view) and 55 deg (i.e., an oblique view). We consequently prepared eight subsets with these typical settings, denoted by Q_{HH85}, Q_{HL85}, Q_{LH85}, Q_{LL85}, Q_{HH55}, Q_{HL55}, Q_{LH55}, and Q_{LL55}, where the first and the second subscripts denote high (H) or low (L) for the SR and TR, respectively, and the third subscript denotes the view (e.g., Q_{HH85} corresponds to 640 $\times$ 480 pixels at 30 fps from the 85 deg view). For the evaluation, each subset was randomly divided into training and test sets for the subject ids, and two-fold cross-validation was performed. It was repeated ten times to reduce the effects of randomness. Note that, the scores were normalized before fusion.[3]

A variety of normalization are exist in the literature, including min-max normalization, z-score normalization [101], and F-normalization [102]. However, some of these schemes (e.g.,

[3]Score normalization does not impact the performance of training-based approaches.

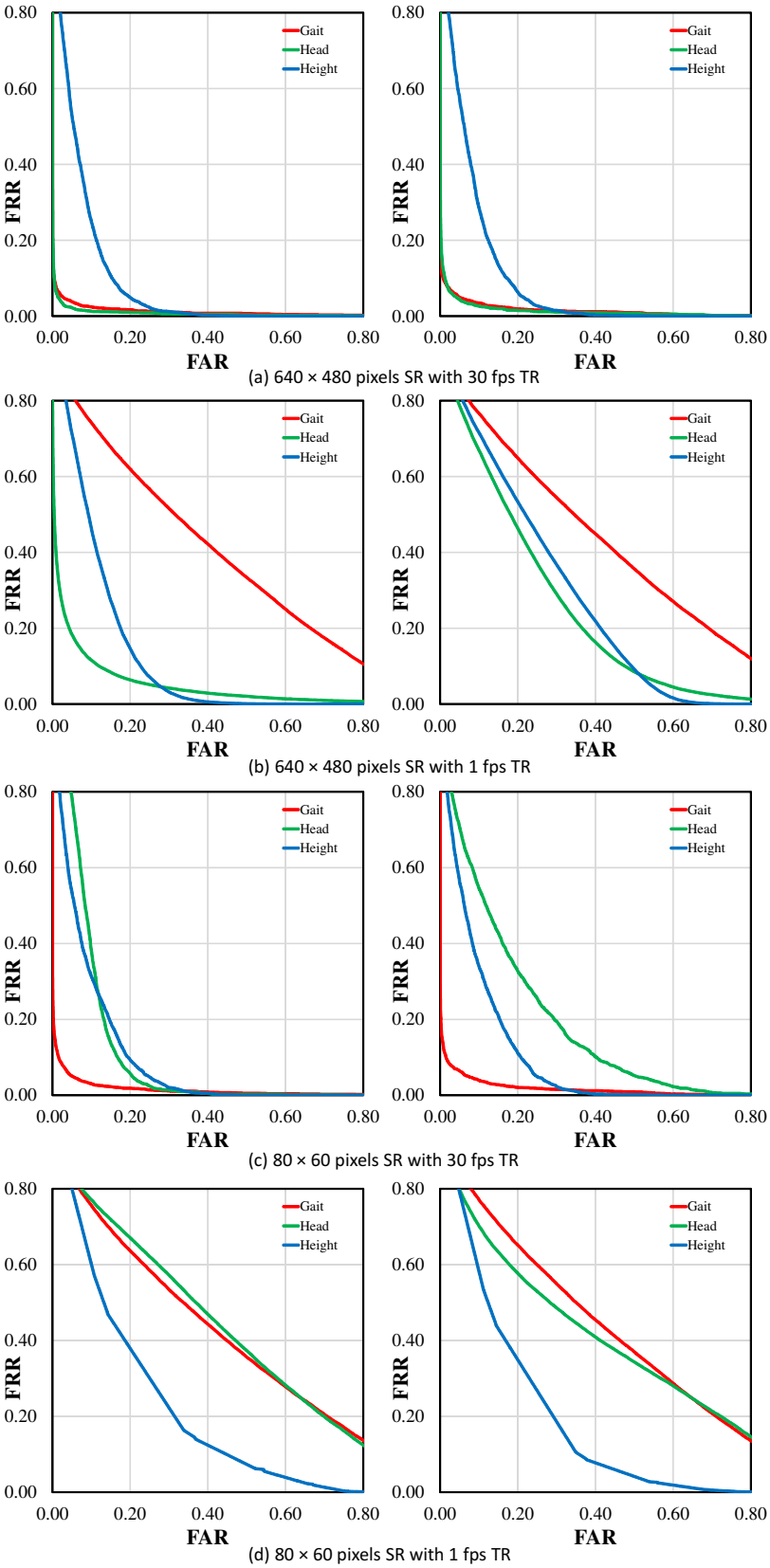


Figure 5.6: ROC curves for individual modalities. The left column is for the 85 deg view and the right is for the 55 deg view.

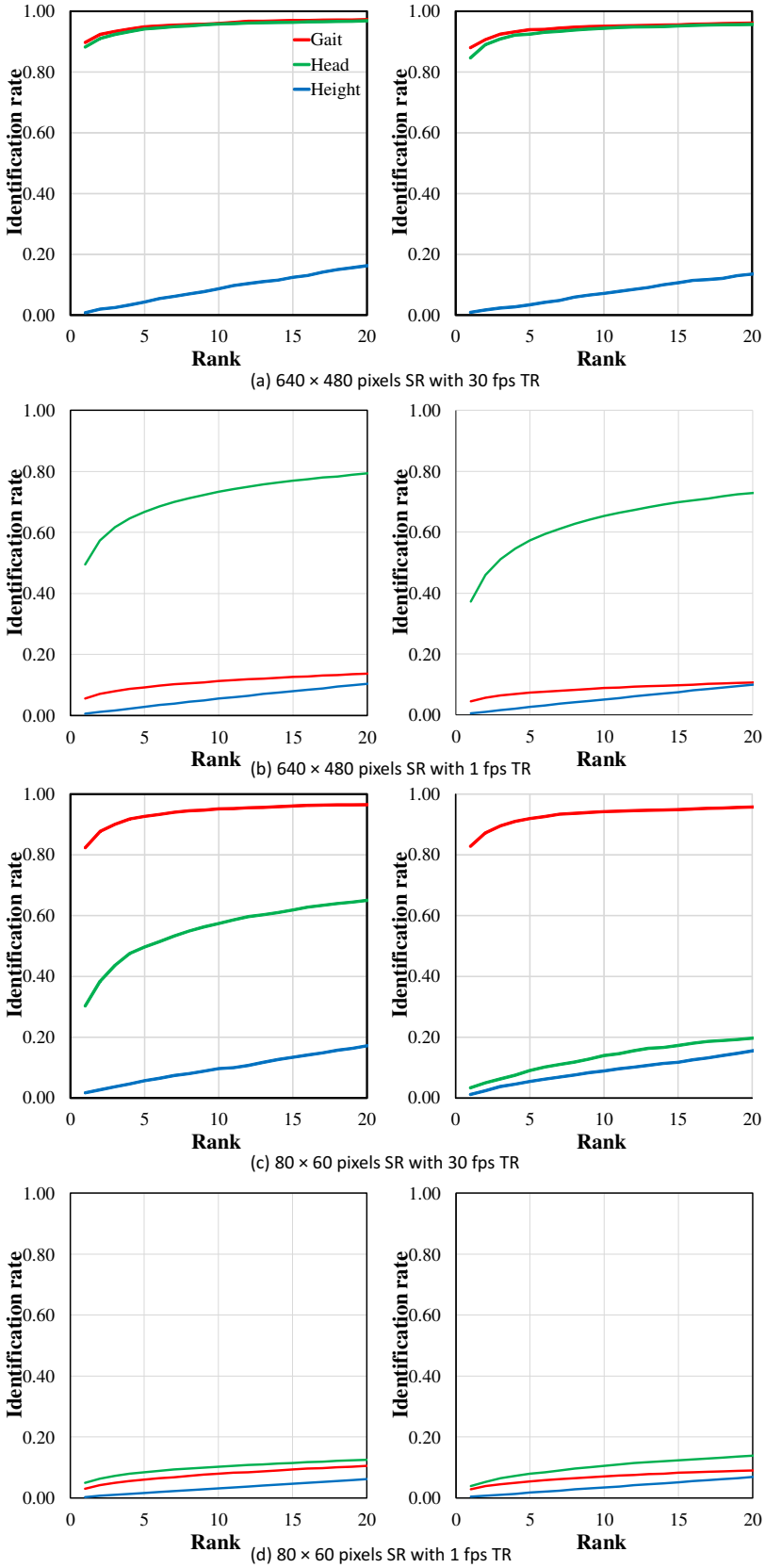


Figure 5.7: CMC curves for individual modalities. The left column is for the 85 deg view and the right is for the 55 deg view. Legend marks are common in all graphs.

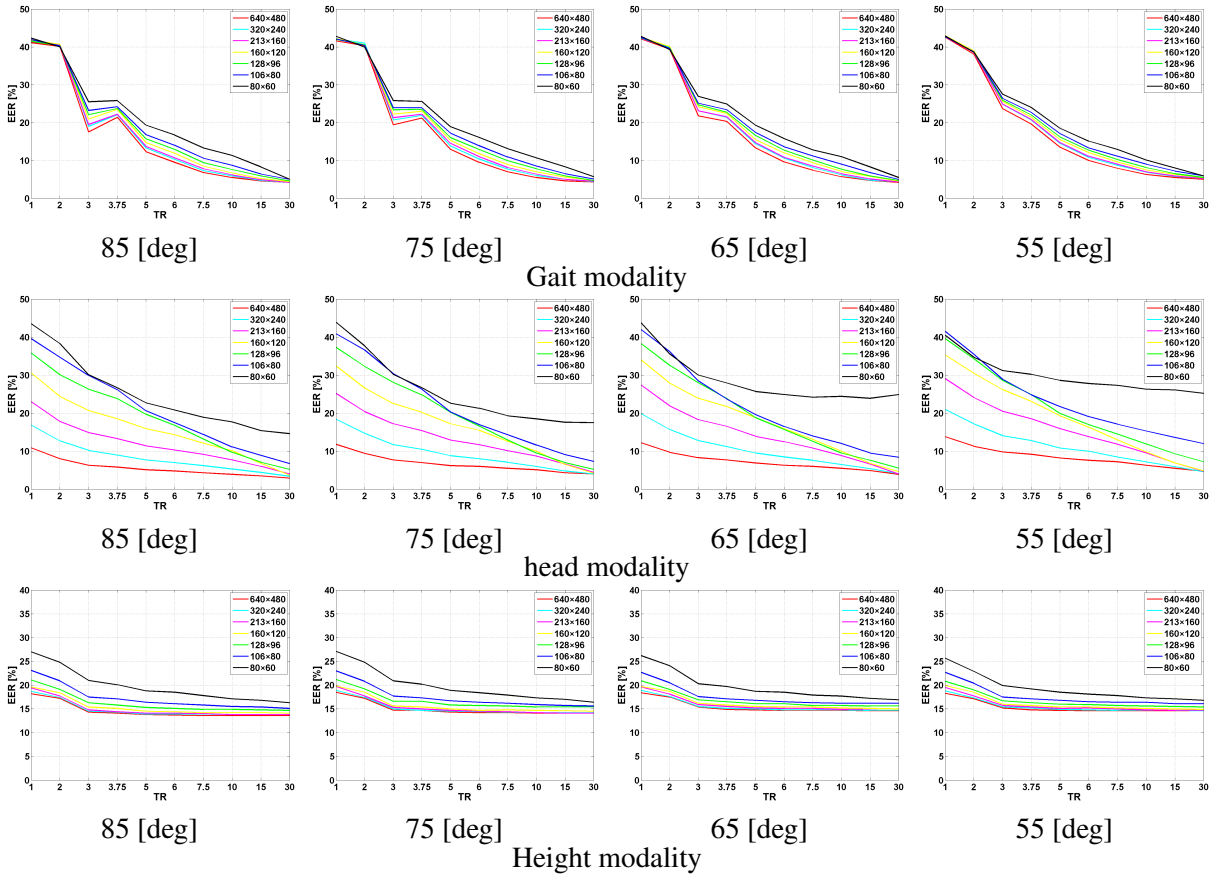


Figure 5.8: EERs [%] for the individual modalities and qualities. Note that the scales used differ among the graphs.

min-max normalization) are sensitive to outliers [63]. We, therefore, used a more statistically-reliable normalization, i.e., z-score normalization, because it is widely used in multi-modal biometrics and score-level fusion [58]. More specifically, we computed an average μ_m and a standard deviation σ_m for each modality $m \in \{gait, head, height\}$ among the training set, and computed a normalized score $\bar{S}_m$ from a raw score S_m as

$$\bar{S}_m = \frac{S_m - \mu_m}{\sigma_m}. \tag{5.4}$$

Because we aim to perform normalization with respect to the modality, we picked a specific quality, i.e., Q_{HH85}, to compute the average and the standard deviation for the normalization process. Furthermore, these average and the standard deviation are used for other subsets also.

To evaluate the score-level fusion, we selected two settings: multi-modal fusion (for gait, head, and height) and bi-modal fusion (for gait and head). Then, we evaluated the accuracy in both verification and identification scenarios with typical measures. ROC curves, EERs, FRRs at specific FARs, AUCs, the half total error rates (HTERs), And CMC curves, and rank-n for identification. Here, the HTER is calculated based on [96].

Table 5.1: EERs [%] for individual modalities and qualities.

(a) 85 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	4.2	4.6	5.5	6.8	9.5	12.3	21.4	17.5	40.1	41.1	2.9	3.5	3.9	4.3	4.8	5.1	5.8	6.3	8.0	10.9	13.6	13.6	13.6	13.6	13.7	13.8	14.1	14.3	17.2	18.2
320×240	4.2	4.7	5.9	7.3	10.3	13.3	22.1	19.0	40.7	41.7	3.4	4.4	5.3	6.2	7.0	7.7	9.0	10.2	12.7	16.9	13.8	13.8	13.8	13.9	13.9	13.9	14.3	14.5	17.4	18.7
213×160	4.2	4.9	6.2	7.8	10.7	13.7	22.2	19.5	40.4	41.4	4.0	6.0	7.7	9.2	10.3	11.4	13.3	14.9	17.8	23.0	13.8	13.8	13.8	14.0	14.1	14.1	14.4	14.8	17.7	19.4
160×120	4.4	5.2	6.7	8.5	11.9	14.7	23.5	20.9	40.7	41.8	3.7	6.8	10.1	12.1	14.3	15.9	18.5	20.7	24.4	30.6	14.1	14.2	14.2	14.2	14.4	14.5	15.1	15.4	18.4	19.6
128×96	4.6	5.9	7.6	9.4	12.9	15.7	23.7	22.1	40.3	41.5	5.2	7.1	9.7	13.2	16.8	19.7	23.8	26.3	30.1	35.8	14.7	14.8	14.9	14.9	15.1	15.3	15.8	16.3	19.1	21.1
106×80	5.0	6.4	8.7	10.6	14.0	16.8	24.2	23.2	40.3	41.9	6.7	8.9	11.1	14.4	17.5	20.6	26.2	29.9	34.7	39.6	15.1	15.4	15.5	15.8	16.1	16.4	17.1	17.5	20.9	23.1
80×60	5.0	8.3	11.3	13.3	16.7	19.3	25.8	25.5	40.0	42.3	14.6	15.4	17.7	18.9	20.8	22.7	26.7	30.1	38.3	43.5	16.3	16.8	17.1	17.8	18.5	18.8	20.1	21.0	24.8	27.0

(b) 75 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	4.3	4.6	5.5	7.0	9.5	12.9	21.2	19.4	40.3	41.6	4.0	4.3	5.1	5.5	6.0	6.2	7.0	7.7	9.4	11.8	14.1	14.1	14.1	14.2	14.2	14.3	14.7	14.7	17.2	18.5
320×240	4.4	5.0	6.0	7.7	10.4	13.8	21.9	20.7	41.0	42.1	4.0	4.8	6.0	7.1	8.0	8.8	10.5	11.7	14.7	18.4	14.2	14.2	14.2	14.3	14.4	14.5	14.7	15.0	17.4	18.9
213×160	4.5	5.0	6.4	8.1	11.0	14.5	22.2	21.3	40.5	42.0	4.4	6.6	8.7	10.1	11.7	12.9	15.4	17.2	20.4	25.2	14.1	14.1	14.2	14.4	14.5	14.7	15.0	15.3	17.7	19.7
160×120	4.7	5.6	7.0	8.9	11.8	15.3	23.0	22.2	40.6	42.2	4.2	6.5	10.0	12.6	15.5	17.2	20.2	22.5	26.6	32.4	14.5	14.6	14.7	14.8	14.7	15.0	15.4	15.6	18.5	19.8
128×96	4.7	5.9	7.8	9.9	12.9	16.0	23.5	23.4	40.6	42.0	5.2	7.0	9.5	12.9	16.6	20.2	24.8	28.1	32.3	37.3	15.4	15.4	15.4	15.6	15.7	15.8	16.6	16.6	19.2	21.2
106×80	5.1	6.5	8.5	10.9	13.9	17.2	23.9	23.9	40.4	42.1	7.3	9.1	11.6	14.3	17.0	20.3	26.3	30.4	36.6	40.9	15.6	15.7	15.9	16.2	16.4	16.7	17.3	17.7	20.8	23.0
80×60	5.7	8.3	10.7	13.1	16.1	18.9	25.6	25.8	39.9	42.8	17.5	17.6	18.5	19.3	21.3	22.6	26.7	30.2	37.7	43.9	16.4	17.0	17.3	17.9	18.4	18.9	20.2	20.9	24.8	27.1

(c) 65 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	4.2	4.7	5.7	7.4	9.6	13.3	20.3	21.8	39.4	42.1	3.9	4.9	5.5	6.0	6.3	6.9	7.7	8.3	9.7	12.2	14.6	14.6	14.7	14.7	14.7	14.8	14.9	15.4	17.5	18.4
320×240	4.4	4.8	6.2	8.1	10.6	14.4	21.4	23.1	40.0	42.5	4.1	5.4	6.5	7.6	8.5	9.5	11.2	12.8	15.7	19.9	14.7	14.7	14.8	14.8	14.8	15.0	15.1	15.5	17.6	18.9
213×160	4.3	5.1	6.5	8.5	10.8	14.7	21.5	23.1	39.7	42.2	4.0	6.6	8.8	10.8	12.5	13.9	16.5	18.3	21.9	27.4	15.0	15.0	14.9	15.1	15.1	15.2	15.5	15.9	18.1	19.6
160×120	4.5	5.8	7.1	9.4	12.0	15.5	22.4	24.2	40.1	42.5	4.6	6.8	10.0	13.1	15.9	18.8	21.7	23.9	27.9	34.0	15.0	15.1	15.2	15.4	15.4	15.5	15.8	16.2	18.7	19.7
128×96	4.7	5.9	7.7	10.1	12.7	16.6	22.7	24.7	39.7	42.4	5.5	7.6	9.5	12.6	15.7	18.8	23.8	28.0	32.6	38.3	15.6	15.6	15.7	15.7	16.1	16.1	16.5	16.9	19.1	20.9
106×80	5.0	6.8	9.0	11.2	13.6	17.4	23.4	25.1	39.6	42.5	8.4	9.5	12.0	14.0	16.5	19.6	23.7	28.5	36.2	42.0	16.2	16.2	16.2	16.3	16.5	16.8	17.1	17.6	20.5	22.7
80×60	5.5	8.3	11.0	12.8	15.8	19.3	24.9	26.9	39.3	42.8	24.9	23.9	24.4	24.2	24.9	25.7	27.9	30.0	35.0	43.8	16.9	17.2	17.7	17.9	18.5	18.7	19.7	20.3	24.1	26.2

(d) 55 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	5.0	5.5	6.3	7.9	10.0	13.5	19.6	23.7	38.1	42.6	4.7	5.5	6.3	7.2	7.6	8.2	9.2	9.8	11.3	13.8	14.6	14.6	14.6	14.6	14.6	14.7	14.8	15.2	17.1	18.3
320×240	5.2	5.8	6.9	8.7	10.8	14.5	20.7	24.8	38.9	43.0	4.6	5.9	7.2	8.5	10.0	10.8	12.8	14.1	17.2	21.0	14.6	14.7	14.7	14.7	14.8	14.9	15.2	15.4	17.3	18.8
213×160	5.2	5.9	7.0	9.0	11.1	14.7	20.8	24.9	38.5	42.7	4.7	6.9	9.5	11.8	13.8	15.9	18.5	20.5	24.5	29.1	14.7	14.7	14.8	15.0	15.2	15.1	15.4	15.7	17.7	19.5
160×120	5.4	6.3	7.5	9.6	12.1	15.5	21.4	25.6	39.0	43.0	4.8	6.9	9.8	13.0	16.3	19.3	23.1	26.2	30.5	35.3	14.9	15.0	15.1	15.2	15.4	15.4	15.7	16.0	18.5	19.9
128×96	5.5	6.5	8.1	10.2	12.7	16.2	22.0	25.9	38.6	42.8	7.2	9.3	11.9	14.5	17.0	19.9	24.9	28.6	34.3	39.6	15.4	15.5	15.6	15.7	15.9	16.0	16.3	16.7	19.0	20.8
106×80	5.9	7.2	9.0	11.1	13.3	17.1	22.7	26.4	38.7	42.7	12.0	13.6	15.3	17.1	19.1	21.7	24.8	28.9	35.5	41.5	16.1	16.1	16.4	16.4	16.5	16.8	17.1	17.5	20.4	22.7
80×60	5.9	8.0	10.1	12.9	15.1	18.5	24.0	27.5	38.6	42.9	25.2	26.1	26.3	27.3	27.8	28.6	30.2	31.2	34.7	40.5	16.8	17.1	17.3	17.8	18.1	18.5	19.2	19.9	22.9	25.7

Protocol 2: An important property of the constructed score database is that it composed of multi-modal scores with multiple qualities. Therefore, It is appropriate for the evaluation of quality-dependent score-level fusion approaches. The intention of this protocol to analyze quality-dependent multi-modal score-level fusion where the score database is partitioned into training and test set with respect to both subjects and qualities. The experimental settings for the training and test sets are defined as shown in Table 5.9. More specifically, we selected three settings for the SRs: "640 × 480, 213 × 160, 128 × 96, 80 × 60"; "640 × 480, 80 × 60"; and "213 × 160, 128 × 96". Regarding to "640 × 480, 213 × 160, 128 × 96, 80 × 60", and "640 × 480, 80 × 60", test set qualities lie between the training qualities. For the TRs, we considered four settings: "30, 10, 7.5, 5, 3, 1", "30, 1", "10, 7.5, 5, 3", and "7.5, 5". For the settings of "30, 10, 7.5, 5, 3, 1" and "30, 1", the test set qualities to lie between the training

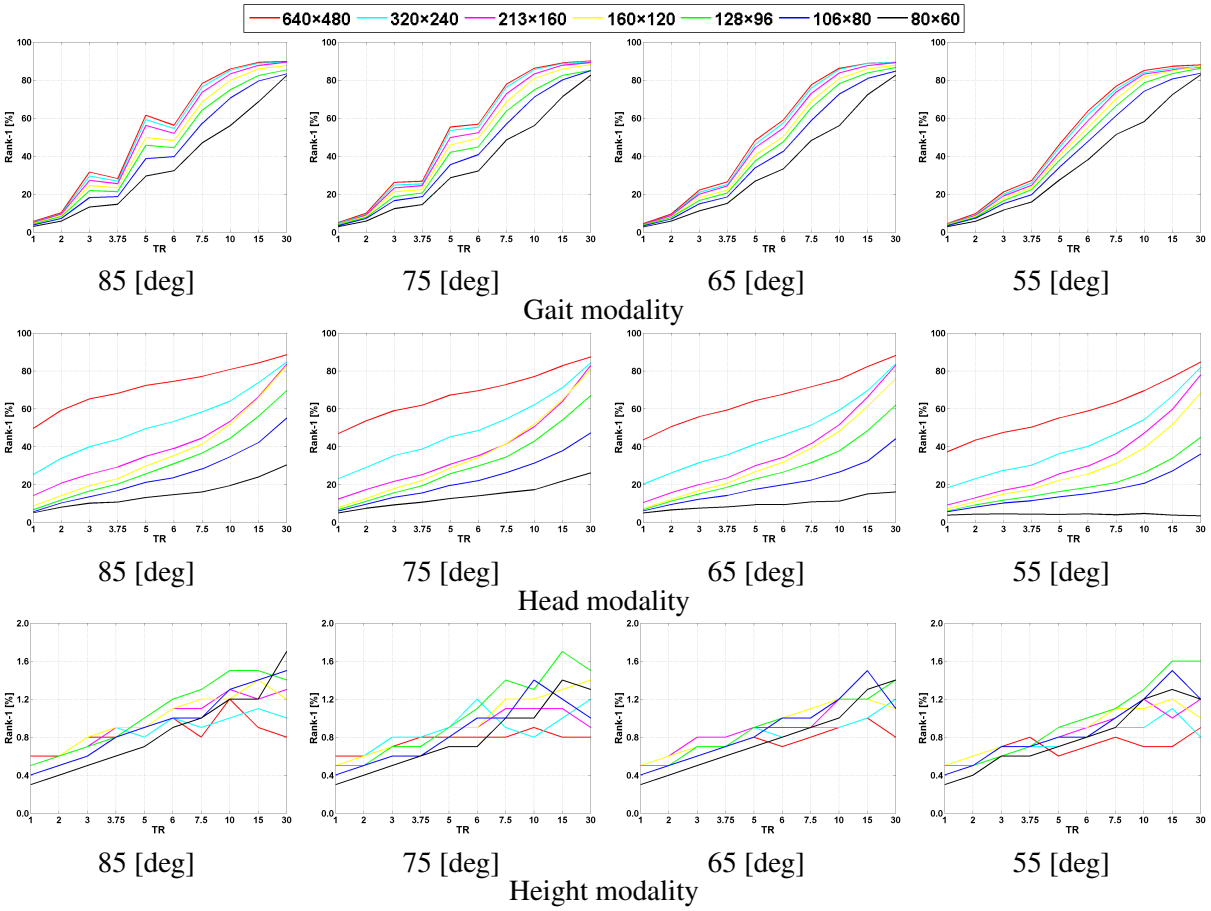


Figure 5.9: Rank-1 identification rate [%] for individual modalities and qualities. Note that the scales differ among the graphs.

qualities. However, some test qualities do not lie between the training qualities in "10, 7.5, 5, 3", and "7.5, 5". We considered the combination of these settings the experiments for protocol 2. In this protocol, the test sets are composed of scores with different quality settings. For the accuracy evaluation, we do not report on the accuracy of each test set separately but report on the total accuracy based on the merging of the test scores with different quality settings.

5.3.4 Benchmarks for score-level fusion

Protocol 1: Score-level fusion approaches classified into three generic categories: transformation-based, probability density-based, and classification-based approaches, as described in section 2.5.2. A total of seven benchmarks considering a wide area of score-level fusion approaches to analyze the experiments for Protocol 1. More specifically, as transformation-based approaches [59], we provide the sum rule (Sum) and the minimum rule (Min); an SVM with a radial basis function kernel [63] and RankSVM [72] for the classification-based approach;

Table 5.2: Rank-1 identification rates [%] for individual modalities and qualities.

(a) 85 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	89.7	89.3	85.9	78.2	56.2	61.5	28.2	31.6	10.3	5.6	88.5	84.2	80.8	77.1	74.5	72.3	68.1	65.2	59.2	49.5	0.8	0.9	1.2	0.8	1.0	0.9	0.8	0.8	0.6	0.6
320×240	89.4	88.9	85.1	76.4	54.5	59.2	26.9	29.5	9.8	5.2	84.6	73.8	64.1	58.3	53.3	49.5	43.8	39.9	33.8	25.2	1.0	1.1	1.0	0.9	1.0	0.8	0.9	0.8	0.6	0.5
213×160	89.4	87.7	83.2	73.6	51.9	56.1	25.5	27.2	9.5	5.0	83.4	66.3	53.4	44.5	39.0	34.8	29.2	25.4	20.7	14.1	1.3	1.2	1.3	1.1	1.0	0.9	0.7	0.6	0.5	
160×120	87.6	85.9	79.8	68.5	48.2	49.8	23.4	24.5	8.8	4.6	82.8	65.8	51.9	41.3	35.3	29.8	23.1	19.3	14.1	8.6	1.2	1.4	1.2	1.2	1.0	0.9	0.8	0.6	0.5	
128×96	85.4	82.4	74.9	64.2	44.4	45.7	21.3	21.9	8.2	4.3	69.5	56.0	44.4	36.5	31.0	25.6	20.2	16.5	11.8	6.6	1.4	1.5	1.5	1.3	1.2	1.0	0.8	0.7	0.6	0.5
106×80	83.2	79.5	70.5	57.2	39.7	38.7	18.7	18.2	7.2	3.8	55.0	42.1	34.7	28.2	23.6	21.1	16.7	13.5	10.2	5.5	1.5	1.4	1.3	1.0	1.0	0.9	0.8	0.6	0.5	0.4
80×60	82.4	68.6	56.0	47.0	32.3	29.5	14.7	13.2	5.8	3.0	30.3	23.9	19.4	16.0	14.6	13.1	10.7	10.1	8.0	5.0	1.7	1.2	1.2	1.0	0.9	0.7	0.6	0.5	0.4	0.3

(b) 75 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	90.0	89.1	86.3	77.8	56.8	55.3	26.8	26.2	10.0	5.0	87.3	82.7	77.0	72.8	69.5	67.3	61.9	59.0	53.6	46.8	0.8	0.8	0.9	0.8	0.8	0.8	0.8	0.7	0.6	0.6
320×240	89.7	88.7	85.5	76.3	55.2	53.5	25.4	24.8	9.4	4.6	84.3	71.2	62.2	54.6	48.5	45.2	38.7	35.3	29.0	23.0	1.2	1.0	0.8	0.9	1.2	0.9	0.8	0.8	0.6	0.5
213×160	89.2	87.8	83.3	72.8	52.3	49.8	24.5	23.3	9.0	4.3	82.8	63.9	50.5	41.5	35.2	30.6	25.2	21.6	17.3	12.2	0.9	1.1	1.1	1.0	0.9	0.9	0.7	0.7	0.6	0.5
160×120	88.2	85.8	80.9	68.5	49.4	45.9	22.4	21.3	8.5	4.1	80.8	64.8	51.9	41.6	34.3	28.6	22.1	17.8	12.6	7.7	1.4	1.3	1.2	1.2	0.9	0.9	0.7	0.7	0.6	0.5
128×96	85.1	82.5	74.9	63.7	44.9	42.0	20.6	18.7	7.8	3.8	66.9	54.1	42.9	34.5	29.7	25.7	19.3	15.6	11.0	6.6	1.5	1.7	1.3	1.4	1.1	0.9	0.7	0.7	0.5	
106×80	84.9	80.1	71.1	56.9	40.8	35.5	18.7	16.6	7.2	3.4	47.2	37.8	31.3	26.2	22.0	19.5	15.5	13.2	9.5	5.9	1.0	1.2	1.4	1.0	1.0	0.8	0.6	0.6	0.5	0.4
80×60	82.5	71.3	56.1	48.5	32.2	28.6	14.5	12.4	5.8	2.9	26.0	21.7	17.2	15.7	14.0	12.6	10.6	9.2	7.4	5.0	1.3	1.4	1.0	1.0	0.7	0.7	0.6	0.5	0.4	0.3

(c) 65 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	89.3	88.8	86.3	77.5	59.0	48.4	26.4	22.1	9.5	4.5	88.1	82.3	75.5	71.6	67.7	64.3	59.3	55.8	50.5	43.5	0.8	1.0	0.9	0.8	0.7	0.8	0.7	0.7	0.6	0.5
320×240	89.3	88.7	85.7	75.9	57.4	46.3	25.0	21.0	8.9	4.1	83.7	69.6	59.3	51.4	46.2	41.3	35.6	31.5	26.1	20.1	1.2	1.0	0.9	0.9	0.8	0.9	0.7	0.7	0.6	0.5
213×160	89.2	87.6	83.8	73.1	54.7	44.5	24.3	20.0	8.7	4.0	82.9	66.2	51.6	41.6	34.3	29.9	23.5	19.9	15.9	10.4	1.1	1.2	1.2	0.9	0.9	0.9	0.8	0.8	0.6	0.5
160×120	87.8	85.8	80.8	69.0	50.2	40.6	22.1	18.1	8.0	3.7	76.0	61.3	48.1	39.2	31.8	26.7	20.6	16.8	12.1	7.3	1.1	1.2	1.2	1.1	0.9	0.9	0.7	0.7	0.6	0.5
128×96	86.6	83.8	78.2	65.5	47.5	37.5	20.6	16.6	7.7	3.6	61.9	48.4	37.8	31.5	26.5	22.9	18.5	15.0	11.1	6.5	1.4	1.2	1.2	1.0	0.9	0.9	0.7	0.7	0.5	
106×80	84.7	80.8	72.7	58.8	42.5	33.9	18.5	14.9	6.9	3.3	44.1	32.6	26.6	22.3	19.9	17.5	14.2	12.2	9.4	6.0	1.1	1.5	1.2	1.0	1.0	0.8	0.7	0.6	0.5	0.4
80×60	82.5	72.3	56.1	48.3	33.3	26.9	15.1	11.2	5.8	2.7	16.0	15.0	11.2	10.8	9.3	9.3	8.1	7.5	6.5	4.9	1.4	1.3	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3

(d) 55 [deg]

Modal	Gait										Head										Height									
SR\TR	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1	30	15	10	7.5	6	5	3.75	3	2	1
640×480	88.0	87.3	85.1	76.8	63.7	46.4	27.2	21.1	9.7	4.5	84.7	76.8	69.6	63.4	58.8	55.2	50.3	47.5	43.4	37.2	0.9	0.7	0.7	0.8	0.7	0.6	0.8	0.7	0.6	0.5
320×240	87.1	86.2	84.0	75.1	61.9	44.9	26.0	19.9	9.0	4.2	81.9	66.8	54.3	46.9	40.1	36.4	30.2	27.4	23.0	18.1	0.8	1.1	0.9	0.9	0.8	0.7	0.7	0.7	0.6	0.5
213×160	86.9	85.4	83.2	73.7	58.7	42.3	24.7	19.0	8.7	4.0	77.9	59.7	47.1	36.3	29.7	25.7	19.7	16.9	12.9	9.1	1.2	1.0	1.2	1.0	0.9	0.8	0.7	0.7	0.6	0.5
160×120	87.2	85.0	81.6	70.1	55.1	40.4	23.4	17.8	8.3	3.8	68.3	51.7	39.3	31.1	25.5	22.2	17.5	14.9	10.9	7.1	1.0	1.2	1.1	1.0	0.9	0.9	0.7	0.7	0.6	0.5
128×96	86.3	83.4	78.5	66.4	52.1	38.0	22.4	16.5	7.8	3.6	44.8	33.9	26.2	21.0	18.5	16.3	13.7	11.7	9.2	6.0	1.6	1.6	1.3	1.1	0.9	0.9	0.7	0.6	0.5	
106×80	83.6	80.6	74.2	61.3	47.7	34.1	19.7	15.0	7.0	3.3	36.0	27.1	20.6	17.5	15.1	13.5	11.4	10.2	8.0	5.6	1.2	1.5	1.2	1.0	0.8	0.8	0.7	0.7	0.5	0.4
80×60	82.8	72.4	58.2	51.4	38.1	27.6	15.9	11.6	5.7	2.9	3.4	3.8	4.6	4.0	4.4	4.2	4.3	4.4	4.3	3.8	1.2	1.3	1.2	0.9	0.8	0.7	0.6	0.6	0.4	0.3

and the GMM[4] [62], LLR [61], and KDE [156] for the probability density-based approaches.

Protocol 2: We consider both quality-dependent and quality-independent score-level fusion approaches to analyze the effectiveness of the quality-dependent fusion by comparing with the quality-independent approaches. We, therefore, selected the three approaches based on the best verification accuracy in a quality-independent protocol: That are the GMM, LLR, and RankSVM, while Sum is selected as a baseline. To estimate the parameters of these approached in a quality-independent manner, we trained the methods using the training set with the highest SR and TR (i.e., Q_{HH85}). Thus, we can then realize quality-independent approaches. We employed the Q-stack [113], which is a concatenated of matching scores with the quality

[4]The number of the mixture components are set to be between 1 and 20.

measures. Thus, we can realize quality-dependent approaches. In this research, we defined Q-stack as a six-dimensional vector that is composed of three biometric scores, i.e., the S_{gait}, the S_{head}, and the S_{height}, along with three quality measures: q_{SR} for SR, q_{TR} for TR and q_{View} for view. We define q_{SR} and q_{TR} based on a log scale of spatial reduction and a frame rate as a direct numerical value, respectively, and q_{View} is based on a view in units of radians. Then, the training sets for the Q-stack vectors are fed into the GMM, LLR, and RankSVM methods to create quality-dependent versions, which are called GMM (Q-stack), LLR (Q-stack) and RankSVM (Q-stack), respectively.

5.3.5 Experimental result for score-level fusion

Protocol 1: The results CMC and ROC for Protocol 1 are shown in Figs. 5.10 and 5.11. Additionally, the EERs, FRRs at 1% and the 10% FARs, and the HTERs along with the AUC for the ROC are summarized in Tables 5.3, 5.4, 5.5, 5.6 and 5.7. Rank-1 and Rank-5 for the identification rate are summarized in Table 5.8 for both the multi-modal and bi-modal fusion types.

From these results, we can see that the accuracy of multi-modal is better than that of bi-modal in all cases of verification and majority cases of identification. Moreover, we see that some of the probability density-based methods such as GMM and LLR and classification-based approach such as RankSVM perform stably and well in verification scenarios (see Fig. 5.10) because probability density-based approaches guarantee optimality in terms of the ROC [157] on the condition that the estimated probability densities are correct. One interesting observation is that the efficiency of the KDE, which is also a probability density-based approach, is not stable and is heavily dependent on the quality settings. The KDE assumes that each score is independent. Therefore, we believe that this independent assumption may be a cause of unstable efficiency. From the viewpoint of the quality settings, we can observe that both the SRs and the TRs have major effects on the accuracy of the fusion approaches. In the case of both high SR and high TR (e.g., Q_{HH85}), because all the modalities work relatively well, the improvements in accuracy when compared with the Sum as a baseline are not as large (e.g., 1.9% EER for Sum and 1.4% EER for the LLR for multi-modal fusion). In the case of high SR and low TR (e.g., Q_{HL85}), because the head modality still works well, while the gait modality does not work because of the low frame rate, the accuracy improvement when compared with the Sum is significant. While the EER of the Sum is 23.0%, the EERs of the RankSVM and the LLR are 7.0% and 7.3%, respectively, for multi-modal fusion. These results indicate the importance of adaptive weighting for the modalities. Similarly, we can see

Table 5.3: EERs [%] for Protocol 1.

Fusion rule	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	2.3	29.0	4.4	40.1	3.3	32.7	5.6	41.4	1.9	23.0	3.9	33.9	2.8	26.4	4.7	34.6
Min	4.2	10.8	4.9	43.4	4.5	13.8	5.7	40.5	4.2	17.0	4.9	31.7	4.5	15.8	5.7	36.5
SVM	1.9	32.9	3.3	47.8	2.9	28.6	7.2	50.2	1.6	18.0	2.7	38.1	2.7	17.3	6.6	41.8
GMM	**1.6**	**10.6**	***2.6***	**39.1**	**2.5**	**13.5**	**4.4**	**37.2**	***1.5***	8.5	***2.1***	**25.6**	***2.4***	10.3	4.2	25.1
LLR	***1.7***	**10.6**	**2.5**	39.8	***2.7***	***13.6***	**4.4**	40.4	**1.4**	***7.3***	**2.0**	**25.6**	**2.2**	***8.8***	**3.8**	**24.0**
KDE	1.8	10.8	3.7	***39.2***	***2.7***	15.6	5.7	***37.3***	1.6	8.3	3.6	48.8	***2.4***	11.7	4.9	44.2
RankSVM	***1.7***	10.7	***2.6***	39.8	2.8	***13.6***	4.5	40.4	1.6	**7.0**	***2.1***	**25.6**	2.7	**8.7**	***4.1***	***24.1***

Bold and italic bold fonts indicate the best and second best accuracies throughout the work in this paper, respectively.

Table 5.4: FRR [%] at 1% FAR ($FRR_{1\%}$) for Protocol 1.

Fusion rule	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	3.2	82.7	10.3	92	5.5	86.6	11.3	92.4	2.5	78.0	10.4	89.9	4.3	83.2	11.9	90.8
Min	7.4	38.4	12.3	92.9	9.5	51.5	11.8	92.8	7.4	89.9	12.3	92.9	9.5	82.7	11.8	92.8
SVM	***2.0***	87.5	5.1	95.5	**3.6**	84.9	10.9	96.6	**1.7**	74.6	3.8	93.4	**3.0**	76.3	9.7	95.6
GMM	**1.9**	**36.6**	4.7	**89.9**	***3.7***	**49.7**	**9.3**	**90.2**	1.8	45.0	3.6	**86.0**	3.3	53.3	9.3	89.5
LLR	2.2	37.8	**4.3**	92.2	3.9	***51.4***	9.4	92.2	**1.7**	29.1	**3.2**	88.5	3.5	***43.3***	**8.6**	**87.7**
KDE	2.1	***37.1***	6.7	***90.7***	3.9	99.7	10.7	***91.6***	1.9	***26.8***	6.9	95.1	***3.2***	99.0	11.3	92.4
RankSVM	2.2	37.9	***4.4***	92.1	4.4	51.9	**9.3**	92.2	1.9	**26.5**	***3.5***	***88.3***	4.1	**40.6**	**8.6**	***88.0***

accuracy improvements in the case of low SR with high TR, in which case the head modality does not work, but the gait modality still works. Also, the view affects the accuracy of the fusion approaches. When compared with the EERs of views of 85 deg and 55 deg under low SR and low TR conditions for multi-modal fusion, the GMM, LLR, and RankSVM achieve better accuracy for the 85 deg view, but the LLR, and RankSVM obtain better accuracy for the 55 deg view.

Regarding identification scenarios (see Fig. 5.11 and Table 5.8), the accuracy trend for each benchmark is, however, slightly different from that for the verification scenario. For example, the RankSVM achieves the best or second-best results for all qualities except for the rank-1 identification rate of Q_{LL85} for multi-modal fusion. This point will be discussed in greater depth in the discussion section. We, therefore, select the best methods concerning both verification and identification in the subsequent evaluation of Protocol 2. Specifically, we selected the three best methods, GMM, LLR, and RankSVM, and used them in the following accuracy analysis for the quality-dependent settings, i.e., for Protocol 2.

Protocol 2: At first, we evaluated the accuracies of quality-independent and dependent approaches (e.g., using training Set 1) and drew the ROC and CMC curves, as shown in Fig. 5.12 and summarized the EERs and FRRs at FARs of 1% and 10%, along with the AUCs, HTER, and Rank-1 and Rank-5, as shown in Table 5.10. Regarding verification, the quality-independent approaches such as Sum, GMM, LLR, and RankSVM do not perform well, with

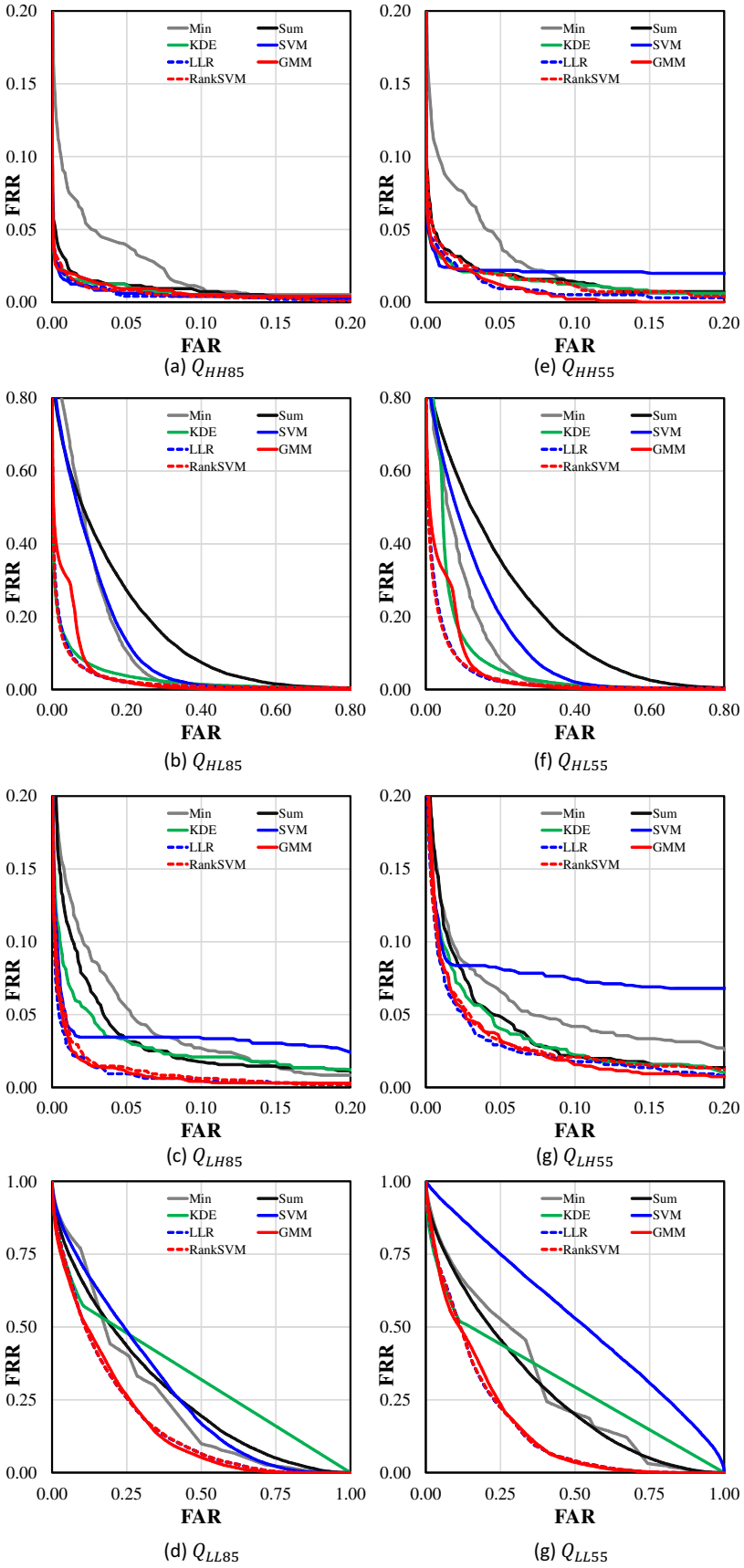


Figure 5.10: ROC curves for Protocol 1 (the first cross-validation set) for multi-modal fusion. Note that the scales differ from graph to graph.

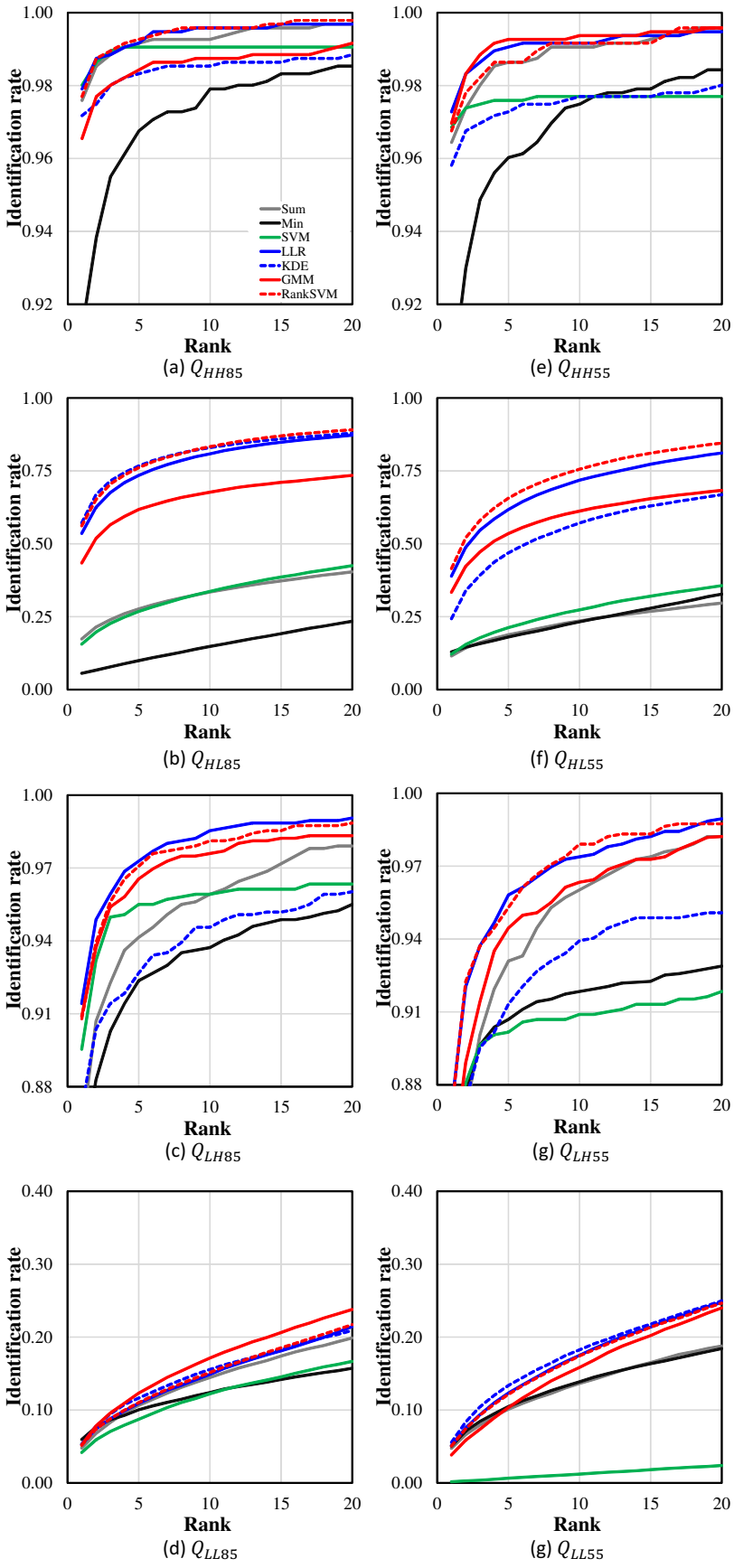


Figure 5.11: CMC curves for Protocol 1 (the first cross-validation set) for multi-modal fusion. Legend marks are common in all graphs. Note that the scales differ from graph to graph.

Table 5.5: FRR [%] at 10% FAR ($FRR_{10\%}$) for Protocol 1.

Fusion rule	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	1.0	56.6	2.5	73.3	1.6	64.4	3.7	75.5	0.6	45.9	1.7	65.8	1.3	54.7	2.1	68.0
Min	0.9	11.5	2.2	77.1	**1.1**	17.9	4.4	70.2	0.9	40.2	2.2	75.4	1.1	32.9	4.4	70.2
SVM	1.6	61.6	2.5	82.5	2.4	54.3	6.3	86.2	1.3	34.2	2.3	75.4	2.3	34.2	6.0	79.3
GMM	**0.6**	**11.0**	**0.9**	**69.2**	**1.1**	**17.2**	***2.6***	***66.2***	0.6	6.5	0.9	**53.0**	***0.9***	11.6	2.0	**52.1**
LLR	***0.7***	***11.2***	**0.9**	73.1	**1.1**	***17.4***	**2.5**	74.0	**0.5**	**5.0**	**0.6**	***53.5***	**0.7**	***7.5***	**1.9**	55.0
KDE	0.9	11.4	2.3	***69.3***	1.6	24.5	4.1	**66.1**	0.6	7.1	2.0	82.8	1.2	14.4	3.1	73.9
RankSVM	0.8	***11.2***	**0.9**	73.1	1.2	17.5	***2.6***	74.0	**0.5**	***5.1***	**0.6**	***53.5***	***0.9***	**7.4**	**1.9**	***54.7***

Table 5.6: AUC [%] for Protocol 1.

Fusion rule	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	0.5	20.1	1.1	35.6	0.7	25.2	1.5	37.3	0.3	13.9	0.8	27.0	0.4	17.7	0.9	27.8
Min	0.6	4.6	0.9	39.6	0.7	6.7	1.5	36.8	0.5	9.7	0.9	24.5	0.6	8.4	1.5	29.6
SVM	0.8	26.6	1.0	46.5	1.4	22.4	3.4	50.0	0.8	10.5	0.8	32.4	1.3	9.6	3.1	38.4
GMM	**0.3**	**4.5**	0.6	**33.7**	**0.5**	**6.4**	***1.1***	**32.1**	0.3	3.1	0.4	**16.6**	**0.3**	4.3	***0.8***	15.9
LLR	**0.3**	**4.5**	**0.4**	35.0	**0.5**	***6.6***	**1.0**	36.1	**0.2**	**2.2**	**0.3**	***17.3***	**0.3**	**3.2**	**0.7**	**15.8**
KDE	**0.3**	4.8	0.7	***33.9***	0.6	10.6	1.6	***32.3***	0.3	2.8	0.7	48.9	0.5	7.6	1.1	41.6
RankSVM	0.4	**4.5**	**0.4**	35.0	**0.5**	***6.6***	***1.1***	36.1	**0.2**	***2.3***	**0.3**	***17.3***	**0.3**	**3.2**	***0.8***	**15.8**

EERs of more than 25%. By contrast, the quality-dependent approaches, such as LLR (Q-stack) and RankSVM (Q-stack) achieve much higher accuracies than the quality-independent approaches (e.g., the EERs for the LLR (Q-stack) and the RankSVM (Q-stack) are 13.0% and 20.9%, respectively), while the EERs of the LLR and RankSVM are 25.8% and 28.8%, respectively.) However, the accuracy of GMM (Q-stack) is worse than that is a quality-independent approach. This abysmal performance by the GMM (Q-stack) is caused by covariance matrices are degenerated of the GMM due to the discrete qualities (see Table. 5.9). Therefore, the probability density for the test quality is different from the training quality. It becomes zero for both positives and negatives, therefore, discrimination capability loses in the fused score.

Regarding identification, the quality-independent approach works better than the quality-dependent approach in general. These results can be related to experimental settings. In this work, although biometric data with many qualities were prepared, cross-quality matching was

Table 5.7: HTER [%] for Protocol 1.

Fusion rule	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	2.1	28.6	4.3	39.9	3.3	32.6	5.3	41.1	1.8	22.6	4.0	33.9	2.6	26.1	4.7	34.6
Min	3.9	10.7	5.1	43.5	4.4	13.9	5.4	38.9	3.9	15.0	5.1	30.1	4.4	14.4	5.4	32.7
SVM	**1.5**	32.8	2.8	45.1	**2.3**	28.4	5.2	46.5	**1.3**	17.2	2.2	37.2	**2.0**	16.3	4.9	40.0
GMM	**1.5**	**10.5**	**2.5**	**38.5**	**2.3**	**13.5**	**4.3**	**36.3**	***1.4***	8.1	***2.1***	**24.4**	***2.2***	9.7	***4.1***	23.6
LLR	1.6	***10.6***	**2.5**	39.6	2.5	***13.6***	**4.3**	40.0	***1.4***	***7.3***	**2.0**	***24.9***	2.3	***8.8***	**3.7**	**23.1**
KDE	**1.5**	10.7	3.4	**38.5**	2.5	15.6	5.2	***36.4***	***1.4***	8.2	3.5	50.0	***2.2***	11.7	4.8	50.0
RankSVM	1.6	***10.6***	2.6	39.6	2.6	13.7	4.5	40.0	1.5	**7.0**	2.2	25.0	2.5	**8.7**	***4.1***	***23.2***

Table 5.8: Rank-1/5 identification rates for Protocol 1.

	Bi-modal (gait and head)								Multi-modal (gait, head, and height)							
Fusion rule	Rank-1[%]															
	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}	Q_{HH85}	Q_{HL85}	Q_{LH85}	Q_{LL85}	Q_{HH55}	Q_{HL55}	Q_{LH55}	Q_{LL55}
Sum	96.3	14.3	88.0	4.1	94.9	9.2	86.6	4.0	97.2	17.5	85.4	4.8	96.1	11.7	82.3	4.7
Min	91.8	**54.9**	85.0	**6.0**	90.7	**43.6**	84.3	***5.0***	91.7	5.4	85.0	**5.9**	90.7	12.1	84.3	5.0
SVM	96.8	10.1	90.9	2.2	95.4	10.0	84.7	1.5	97.3	16.0	91.3	3.4	96.1	14.0	85.4	1.7
GMM	**97.2**	***54.5***	91.0	***5.8***	***96.2***	***41.4***	87.4	**5.3**	97.0	41.3	90.3	5.0	96.3	30.9	84.6	4.1
LLR	**97.2**	53.0	***92.5***	3.9	**96.4**	40.5	**88.8**	4.0	***97.7***	52.8	***91.4***	***5.2***	**96.8**	***39.1***	***86.8***	**5.2**
KDE	96.5	51.3	88.1	5.3	94.9	15.9	84.1	4.1	97.1	**58.1**	86.5	2.0	95.8	20.4	82.5	2.8
RankSVM	**97.2**	52.8	**92.6**	4.0	***96.2***	39.7	***88.7***	4.0	**97.7**	***56.2***	**91.6**	5.1	***96.5***	**42.4**	**87.4**	***5.1***
	Rank-5[%]															
Sum	98.2	22.2	95.8	8.3	97.5	14.7	94.1	7.7	98.9	28.0	94.3	***10.6***	98.4	19.3	93.5	10.2
Min	96.6	**72.9**	93.2	10.2	96.0	**64.8**	91.0	***10.4***	96.4	9.9	93.2	9.9	96.0	17.2	91.0	10.4
SVM	97.9	17.2	95.4	4.5	97.0	18.9	90.5	3.1	98.4	29.4	96.0	7.1	97.1	26.8	91.9	3.8
GMM	**98.7**	***72.7***	96.1	**11.6**	**98.1**	***63.6***	94.0	**11.3**	98.6	58.5	96.1	**12.1**	98.3	50.0	94.0	10.9
LLR	**98.7**	71.1	**97.0**	8.1	**98.1**	61.4	**95.1**	8.0	**99.1**	73.1	***97.1***	10.5	**98.9**	***62.0***	**95.8**	**12.1**
KDE	97.9	70.5	93.1	***10.9***	96.7	35.6	90.5	9.9	98.2	**77.2**	92.9	4.4	97.1	42.4	90.7	6.8
RankSVM	**98.7**	70.9	***96.9***	8.2	**98.1**	60.4	***95.0***	8.0	**99.1**	***76.3***	**97.2**	***10.6***	***98.6***	**66.6**	**95.8**	***11.9***

not considered. In other words, the qualities of probe gait features are always the same as those of gallery gait features. Therefore, quality information given by Q-stack cannot be useful in identification, because all gallery gait features matched to a probe gait features have the same quality measures. Therefore, identification accuracy by only employing Q-stack under this setting is not improved. This implies that the additional quality-dependent approaches are expected to improve the identification accuracy in the scenario where the qualities of the probe and the gallery are the same.

Later, we evaluated the quality-dependent fusion performance using different training sets to analyze how the performance varied. Therefore, we selected the best approach LLR (Q-stack) for Set 1 to act as a benchmark. Figure 5.13 shows the ROC and CMC curves and the EERs and FRRs at FARs of 1% and 10%, along with the AUCs, the HTER, and Rank-1 and Rank-5, as shown in Table 5.11.

From these results, we can find some interesting observations. When we fix the TR quality setting for training, SR quality setting for training does not have much impact on accuracy for both verification and identification. For example, the EERs of Set 1, Set 5 and Set 9 are 13.0%, 13.1% and 12.9%, respectively and Rank-1 of Set 1, Set 5 and Set 9 are 31.1%, 32.2%, and 29.5%, respectively. Regarding when we fix the SR quality setting for training, TR quality settings for training have much impact on recognition accuracy, and impact on verification and identification are different. For verification, in cases where test TR qualities lie between the training TR qualities, the EERs of Set 1 and Set 2 are both 13.0%, and the EERs of Set 9 and Set 10 are almost the same (i.e., 12.9% and 13.0%) even though the densities of TR training

Table 5.9: Qualities for training and test sets in Protocol 2.

Data set		SR [pixels]	TR [fps]	View [deg]
Training	Set 1	640 × 480, 213 × 160, 128 × 96, 80 × 60	30, 10, 7.5, 5, 3, 1	85, 55
	Set 2	640 × 480, 213 × 160, 128 × 96, 80 × 60	30, 1	
	Set 3	640 × 480, 213 × 160, 128 × 96, 80 × 60	10, 7.5, 5, 3	
	Set 4	640 × 480, 213 × 160, 128 × 96, 80 × 60	7.5, 5	
	Set 5	640 × 480, 80 × 60	30, 10, 7.5, 5, 3, 1	
	Set 6	640 × 480, 80 × 60	30, 1	
	Set 7	640 × 480, 80 × 60	10, 7.5, 5, 3	
	Set 8	640 × 480, 80 × 60	7.5, 5	
	Set 9	213 × 160, 128 × 96	30, 10, 7.5, 5, 3, 1	
	Set 10	213 × 160, 128 × 96	30, 1	
	Set 11	213 × 160, 128 × 96	10, 7.5, 5, 3	
	Set 12	213 × 160, 128 × 96	7.5, 5	
Test		320 × 240, 160 × 120, 106 × 80	15, 6, 3.75, 2	75, 65

Table 5.10: EERs, $FRR_{1\%}$,$FRR_{10\%}$, AUC, HTER and Rank-1/5 identification rates of quality-independent and quality-dependent approaches for Protocol 2.

Fusion rule	Training set	EER [%]	$FRR_{1\%}$ [%]	$FRR_{10\%}$ [%]	AUC [%]	HTER [%]	Rank-1 [%]	Rank-5 [%]
Sum	Q_{HH85}	33.2	74.5	50.4	23.3	41.2	38.5	50.6
GMM		26.4	76.4	40.1	19.5	32.5	26.0	38.7
LLR		25.8	71.1	43.5	16.8	34.2	**41.8**	**56.1**
RankSVM		28.8	72.2	46.1	18.4	37.2	***40.9***	54.4
GMM (Q-stack)	Set 1	49.9	99.0	89.9	49.9	49.9	0.1	0.5
LLR (Q-stack)		**13.0**	***69.8***	**18.9**	**5.5**	**12.5**	31.1	48.1
RankSVM (Q-stack)		***20.9***	**68.4**	***37.4***	***11.0***	***26.1***	38.7	***54.6***

$FRR_{1\%}$ and $FRR_{10\%}$ are FRR at 1% FAR and FRR at 10% FAR respectively.

qualities are different. However, in cases where some test TR qualities lie outside the training qualities, verification accuracy becomes worse. For example, the EERs of set 4 and set 12 are 15.8% and 15.2%, respectively. By contrast, for identification, in the case where TR qualities of "7.5 and 5" are used for training, we observed that Rank-1 becomes good even though some test TR qualities lie outside the training TR qualities. As discussed in the first experiment of Protocol 2, this may be related to the evaluation settings where the qualities of the gallery and the probe are the same. Nevertheless, we guess that some specific fusion parameters may work well on many quality settings for identification.

5.4 Discussion

Inconsistency between verification and identification scenarios: As written in the section on the experimental results for score-level fusion, the benchmark performance is dependent on the

Table 5.11: EERs, $FRR_{1\%}$, $FRR_{10\%}$, AUC, HTER and Rank-1/5 identification rates of the quality-dependent approach using the different training sets for Protocol 2.

Fusion rule	Training set	EER [%]	$FRR_{1\%}$ [%]	$FRR_{10\%}$ [%]	AUC [%]	HTER [%]	Rank-1 [%]	Rank-5 [%]
LLR (Q-stack)	Set 1	***13.0***	69.8	18.9	***5.5***	***12.5***	31.1	48.1
	Set 2	***13.0***	73.9	19.3	5.7	14.6	29.1	45.8
	Set 3	13.8	53.7	***18.2***	5.9	15.7	40.2	55.9
	Set 4	15.8	***50.6***	21.8	7.3	19.2	***41.7***	***56.7***
	Set 5	13.1	69.2	18.9	***5.5***	12.6	32.2	49.3
	Set 6	13.3	74.3	19.7	5.8	13.5	30.0	47.0
	Set 7	14.1	52.2	18.6	6.1	16.6	41.2	56.9
	Set 8	16.2	**49.3**	22.4	7.7	20.3	**42.3**	**57.5**
	Set 9	**12.9**	69.3	18.8	**5.4**	**12.4**	29.5	46.3
	Set 10	***13.0***	73.6	19.0	5.7	14.0	27.5	43.6
	Set 11	13.4	53.6	**17.7**	5.6	15.0	39.3	55.1
	Set 12	15.2	50.9	20.8	6.9	18.0	41.1	56.2

$FRR_{1\%}$ and $FRR_{10\%}$ are FRR at 1% FAR and FRR at 10% FAR respectively.

specific verification or identification scenario. Nandakumar et al. [158] proposed an approach to extend the fusion for the verification scenario (i.e., using a likelihood ratio-based approach) into that for the identification scenario, it may be expected that the best benchmark for the verification will also be the best benchmark in the identification. Their method [158] imposes an assumption that the genuine and imposter scores are drawn from the same distribution, regardless of the subjects. However, the subject dependence of the score distribution has been observed in our score database, and good benchmarks are therefore shown not to be consistent among the verification and identification scenarios in our experimental result. In fact, a recent study [159] performed a preliminary experiment to show that a low-performance verification system may still achieve good performance in an identification. Moreover, DeCann and Ross demonstrated in [106] that sets of genuine and imposter scores that generate the same ROC curve can generate different CMC curves. This is because the verification performance is dependent on the aggregated distributions of the genuine and imposter scores, while the identification scenario is dependent on probe-dependent ranking statistics. In principle, suitable approaches for the verification and identification scenarios can be different. It was also reported in [106] that this type of difference between the ROC and CMC curves tends to be particularly outstanding for soft biometrics, for example, gait biometric. It is, therefore, convincing in this case that the different benchmarks yielded higher accuracies for each of the verification and identification scenarios. Specifically, the probability density-based approaches consider the aggregated score distributions directly and thus yielded higher accuracies for verification. On the other hand, because the RankSVM considers the probe-dependent rank statistics directly, it yielded the

best or second-best accuracies for the identification scenarios for multi-modal fusion, which is consistent with the discussion above.

Extension to real scenes: While we considered SR and TR as quality measures for the fusion of gait, head, and height biometrics, in the real-world scenario, it is possible to use more qualities. For example, background motion artifacts, occlusion, carried object degrades the silhouette and lead to ill-posed silhouettes. Therefore, it would be useful to collect biometric scores in real-world scenarios and considered additional quality measures (e.g., illumination changes, view changes, clothing change, CO, occlusion) to achieve a more sophisticated fusion approach.

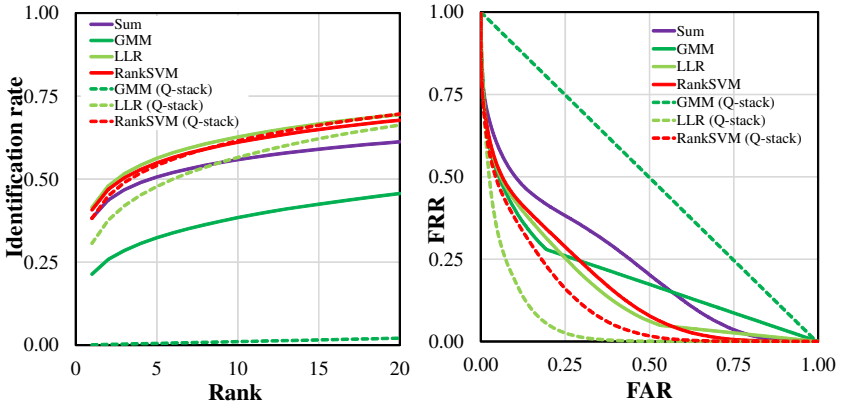


Figure 5.12: CMC (left) and ROC (right) curves for the quality-independent and quality-dependent approaches for Protocol 2.

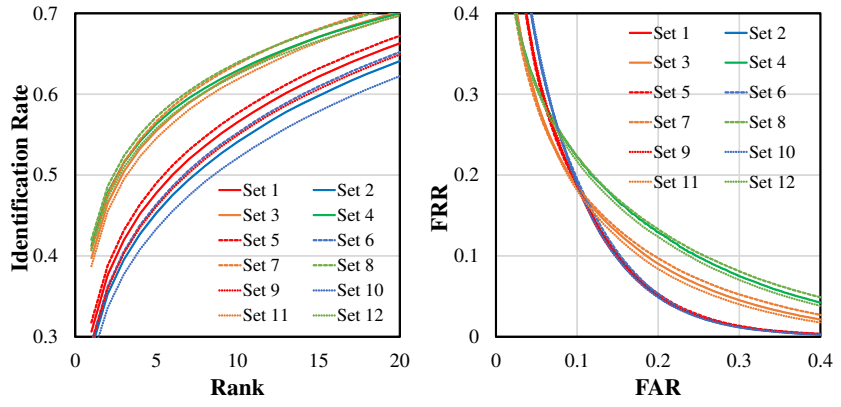


Figure 5.13: CMC (left) and ROC (right) curves for the quality-dependent approach LLR (Q-stack) when using different training sets for Protocol 2.

5.5 Conclusion

In this chapter, we presented a multi-quality, multi-modal biometric score database from the captured image sequence for a single camera and analyzed the accuracy for each modality and fused score level. We extracted the gait, head, and height biometrics from a captured image sequence, analyzed the SR, the TR, and the view as quality measures. Thus, the database contains an extremely large number of biometric scores, including approximately four million genuine scores and 7.5 billion imposter scores. We proposed two original protocols for quality-independent and quality-dependent for score-level fusion, and perform the experiment using a wide area of score-level fusion approach.

Chapter 6

Conclusion and future work

6.1 Conclusion

Gait is a relatively new and prominent biometric that pertains to the use of a subject's walking manner. In recent years, human recognition based on gait has received enormous interest due to its ability to apply on an automated visual surveillance system. The unique advantage of gait biometrics is that it can be used for passive human recognition (i.e., it can be captured data at a distance at low spatial resolution). These yield a strong potential ability for gait recognition to forensic analyses in the visual surveillance system. However, practical deployment in a real-world scenario is still considered a challenging task because several issues should be managed to achieve efficient gait recognition results. Such issues can be related to the presence of external factors to a subject (e.g., carried object, clothing, shoes); the internal factor to a subject (e.g., walking speed, aging, pregnancy); the environmental context (e.g., illumination, walking surface, occlusion) and the quality of captured image sequence from a camera (e.g., spatial resolution, temporal resolution, observation view). These factors make the gait recognition more challenging and prevent the use of gait recognition in a real-world scenario. This thesis has focused on three major issues for gait recognition in more challenging conditions.

First, we address one of the most important external factors to a subject, i.e., the carried object (CO) because people often need to carry objects in their daily lives. A gait database, having an unconstrained variation of CO with sufficient diversity, and an extremely large number of subjects is necessary to research this practical covariate using deep learning. In this work, we presented a gait database that consisted of an extremely large number of subjects with unconstrained types and positions of COs. Furthermore, we analyzed silhouette-based gait recognition using state-of-the-art traditional and deep learning-based approaches. This database had the following advantages over the existing gait databases in the literature: (1)

The database included 62,528 subjects, which was more than six times greater than the existing largest database for gait recognition without CO whereas 200 times with CO; and (2) To analyze the ill-posedness nature of silhouettes with unconstrained variations of COs being carried in unconstrained positions, the silhouettes were manually classified into seven carrying status (CS) labels. In addition, we conducted several experiments to analyze the gait recognition with CO using the proposed database. The results gave several insights, such as estimating the difficulty level among annotated CS labels based on gait recognition performance and the classification accuracy for CS labels.

Second, a very important issue for gait recognition related to the environmental context, i.e., the occlusion was addressed. Occlusion in the real-world is inevitable, and human recognition by gait biometric in the presence of occlusion is a very challenging research problem because the presence of occlusion in an image sequence obscures a significant amount of the human body. Therefore, the extracted occluded silhouettes lead to ill-posed silhouettes, and the gait feature cannot be correctly extracted from such an occluded case. In this work, we presented an approach based on the deep conditional generative adversarial network that consisted of a generator and critic networks. It allowed us to reconstruct an unoccluded image from an occluded silhouette sequence for gait recognition. We demonstrated that triplet hinge loss along with WGAN regularized the training of the generative network and reconstructed the silhouette sequence with a high discrimination ability. Therefore, we achieved better accuracy for gait recognition. To demonstrate the effectiveness of the proposed approach, we considered several occlusion patterns with relative dynamic and relative static occlusion for different degrees of occlusion that were quite common in real-world scenarios. In our experiments, more than 9,000 subjects were considered, along with 24 different occlusion patterns. We designed a set of experiments where the occlusion pattern between the probe and gallery were the same/different and known/unknown. The experimental results demonstrated that the reconstructed silhouette sequence of the proposed approach achieved state-of-the-art accuracy.

Third, we addressed another issue related to the quality of the captured data. In this work, we analyze how the quality of the captured image sequence, i.e., the spatial and temporal resolution affects gait recognition. Besides, we studied the multi-modal biometric fusion of gait, head, and a soft biometric height to tackle the ill-posedness nature of extracted silhouette, and the practical problems of uni-modal biometrics (non-universality, spoofing attacks, and poor recognition performance). We extracted the required gait, head, and height biometrics from a single walking image sequence from a camera considered the spatial resolution (SR), the temporal resolution (TR), and the view as quality. We proposed two original protocols for

quality-independent and quality-dependent for score-level fusion, and experiment using a wide area of score-level fusion approaches.

6.2 Future work

We focused on the three major issues in this thesis; however, there are some remained critical issues to be solved. The works presented in this thesis open up several future directions. Some of them are discussed below.

We studied gait recognition from ill-posed silhouettes, and ill-posed silhouettes were introduced due to the presence of carried object, occlusion, and low quality captured image sequence. However, there are other reasons that can make silhouettes as ill-posed. Some traditional clothing patterns may produce ill-posed silhouettes, thereby making gait recognition more challenging or even almost impossible. For example, long skirt or similar dress like saree (i.e., dress wrapped around the body, worn especially by South Asian females), and the dress that hides the leg fully until toe (for example, Sarong or lungi which is a long piece of thin cloth type dress wrapped around the waist until toe). In addition, the silhouette can also be ill-posed due to the error of background subtraction, particularly in the real-world environment. These are important future research directions for gait recognition from ill-posed silhouettes.

In this thesis, we tackle the gait recognition from ill-posed silhouettes separately for both carried object and occlusion. However, a person can carry an object and be occluded at the same time, and the captured image sequence may not be a good quality sequence. For this type of situation, only a limited portion of a person's silhouette can be available. Therefore, the current approaches may find difficulties to decipher good silhouettes. Hence, an approach is needed to tackle this type of condition. One of the candidates can be a deep generative adversarial network-based approach to reconstruct better silhouettes. It would be future research direction, to reconstruct the silhouette sequence from a piece of partial silhouette information. Moreover, we need large datasets having real-life scenarios in order to explore challenging conditions.

Reference

[1] A. K. Jain, A. Ross, and S. Pankanti. Biometrics: A tool for information security. *IEEE Trans. on Information Forensics and Security*, 1(2):125–143, November 2006.

[2] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, 14(1):4–20, January 2004.

[3] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56(4):882–889, 2011.

[4] H. Iwama, D. Muramatsu, Y. Makihara, and Y. Yagi. Gait verification system for criminal investigation. *IPSJ Trans. on Computer Vision and Applications*, 5:163–175, Oct. 2013.

[5] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, 14(2):201–211, Jun 1973.

[6] James E Cutting and Lynn T. Kozlowski. Recognizing friends by their walk: Gait perception without familiarity cues. *Bulletin of the psychonomic society*, 9(5):353–356, 1977.

[7] T. Kimura, Y. Makihara, D. Muramatsu, and Y. Yagi. Single sensor-based multi-quality multi-modal biometric score database and its performance evaluation. In *Proc. of the Int. conf. on Biometrics*, pages 1–8, 2015.

[8] D. Cunado, M.S. Nixon, and J.N. Carter. Automatic extraction and description of human gait models for recognition purposes. *Computer Vision and Image Understanding*, 90 (1):1–41, 2003.

[9] I. Bouchrika, J. Carter, and M. Nixon. Towards automated visual surveillance using gait for identity recognition and tracking across multiple non-intersecting cameras. *Multimedia Tools and Applications*, 75(2):1201–1221, 2016.

[10] J. Yoo and M. Nixon. Automated markerless analysis of human gait motion for recognition and classification. *Etri Journal*, 33(2):259–266, 2011.

[11] L. Lee and W. Grimson. Gait analysis for recognition and classification. In *Proc. of the 5th IEEE Conf. on Face and Gesture Recognition*, volume 1, pages 155–161, 2002.

[12] Z. Cao, T. Simon, S. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proc. of the IEEE int. conf. on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[13] J. Preis, M. Kessel, M. Werner, and C. Linnhoff-Popien. Gait recognition with kinect. In *proc. of the 1st int. Workshop Kinect Pervas. Comput.*, pages 1–4, New Castle, U.K., 2012.

[14] A. Bari and M. Gavrilova. Artificial neural network based gait recognition using kinect sensor. *IEEE Access*, 7:162708–162722, 2019.

[15] Y. Feng, Y. Li, and J. Luo. Learning effective gait features using lstm. In *Proc. of the int. conf. on Pattern Recognition (ICPR)*, pages 325–330. IEEE, 2016.

[16] S. Sarkar, J.P. Phillips, Z. Liu, I.R. Vega, P. Gro ther, and K.W. Bowyer. The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE Trans. of Pattern Analysis and Machine Intelligence*, 27(2):162–177, 2005.

[17] R. Collins, R. Gross, and J. Shi. Silhouette-based human identification from body shape and gait. In *Proc. of IEEE int. conf. on automatic face gesture recognition*, pages 366–371. IEEE, 2002.

[18] M. Hofmann, and G. Rigoll. Improved Gait Recognition using Gradient Histogram Energy Image. In *Proc. of the IEEE Int. Conf. on Image Processing*, pages 1389-1392, Orlando, Florida, USA., Sept. 2012.

[19] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, (3):257–267, 2001.

[20] J. Han and B. Bhanu. Individual recognition using gait energy image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(2):316– 322, 2006.

[21] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(2):209–226, Feb 2017. ISSN 0162-8828.

[22] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. On input/output architectures for convolutional neural network-based cross-view gait recognition. *IEEE Trans. on Circuits and Systems for Video Technology*, PP(99):1–1, 2017. ISSN 1051-8215.

[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 Trans. on Information Forensics and Security*, 7(5):1511–1521, Oct. 2012.

[24] A. Roy, S. Sural, J. Mukherjee, and G. Rigoll. Occlusion detection and gait silhouette reconstruction from degraded scenes. *Signal, Image and Video Processing*, 5(4):415, Jul 2011.

[25] M.S. Nixon, J.N. Carter, J. Shutler, and M. Grant. Experimental plan for automatic gait recognition. Technical report, Southampton, 2001.

[26] M. A. Hossain, Y. Makihara, J. Wang, and Y. Yagi. Clothing-invariant gait identification using part-based clothing categorization and adaptive weight control. *Pattern Recognition*, 43(6):2281–2291, Jun. 2010.

[27] A. Mansur, Y. Makihara, R. Aqmar, and Y. Yagi. Gait recognition under speed transition. In *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition*, pages 2521–2528, 2014.

[28] Chi Xu, Y. Makihara, Y. Yagi, and J. Lu. Gait-based age progression/regression: a baseline and performance evaluation by age group classification and cross-age gait identification. *Machine Vision and Applications*, 30(4):629–644, Jun 2019.

[29] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Trans. on Computer Vision and Applications*, 10(1):4, Feb 2018. ISSN 1882-6695.

[30] N. Akae, Y. Makihara, and Y. Yagi. Gait recognition using periodic temporal super resolution for low frame-rate videos. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–7, Washington D.C., USA, Oct. 2011.

[31] M. Hofmann, D. Wolf, and G. Rigoll. Identification and reconstruction of complete gait cycles for person identification in crowded scenes. In *VISAPP 2011 - Proc. of the Sixth Int. conf. on Computer Vision Theory and Applications, Vilamoura, Algarve, Portugal, 5-7 March, 2011*, pages 594–597, 2011.

[32] Y. Makihara, A. Mansur, D. Muramatsu, Z. Uddin, and Y Yagi. Multi-view discriminant analysis with tensor representation and its application to cross-view gait recognition. In *2015 11th IEEE Int. conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–8. IEEE, 2015.

[33] D. Muramatsu, A. Shiraishi, Y. Makihara, M.Z. Uddin, and Y. Yagi. Gait-based person recognition using arbitrary view transformation model. *IEEE Trans. on Image Processing*, 24(1):140–154, Jan. 2015.

[34] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi. Gait recognition using a view transformation model in the frequency domain. In *Proc. of the 9th European conf. on Computer Vision*, pages 151–163, Graz, Austria, May 2006.

[35] K. Bashir, T. Xiang, and S. Gong. Gait recognition without subject cooperation. *Pattern Recognition Letters*, 31(13):2052–2060, Oct. 2010.

[36] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren. Gait energy response function for clothing-invariant gait recognition. In *Proc. of the 13th Asian Conf. on Computer Vision (ACCV 2016)*, pages 257–272, Taipei, Taiwan, Nov. 2016.

[37] P. Domingos. A few useful things to know about machine learning. *Commun. ACM*, 55 (10):78–87, October 2012. ISSN 0001-0782.

[38] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 441–444, Hong Kong, China, Aug. 2006.

[39] D. Tan, K. Huang, S. Yu, and T. Tan. Efficient night gait recognition based on template matching. In *Proc. of the 18th Int. conf. on Pattern Recognition*, volume 3, pages 1000–1003, Hong Kong, China, Aug. 2006.

[40] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, and G. Rigoll. The tum gait from audio, image and depth (gaid) database: Multimodal recognition of subjects and traits. *J. Vis. Comun. Image Represent.*, 25(1):195–206, January 2014. ISSN 1047-3203.

[41] M. Hofmann, S. Sural, and G. Rigoll. Gait recognition in the presence of occlusion: A new dataset and baseline algorithms. In *Proc. of Int. conf. on Computer Graphics, Visualization and Computer Vision, Plzen, Czech Republic*, pages 99–104, 2011.

[42] Zhang, Ziyuan and Tran, Luan and Yin, Xi and Atoum, Yousef and Liu, Xiaoming and Wan, Jian and Wang, Nanxin. Gait Recognition via Disentangled Representation Learning. In *Proc. of Int. conf. on Computer Vision and Pattern Recognition*, pages 4710–4719, 2019

[43] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. of the 27th Int. conf. on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680, Cambridge, MA, USA, 2014. MIT Press.

[44] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *2016 IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2536–2544, 2016.

[45] Y. Li, S. Liu, J. Yang, and M. Yang. Generative face completion. In *2017 IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5892–5900, 2017.

[46] S. Zhang, J. Yang and B. Schiele. Occluded pedestrian detection through guided attention in CNNs. In *Proc. of the IEEE int. conf. on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018.

[47] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. Huang. Generative image inpainting with contextual attention. In *2018 IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5505–5514, 2018.

[48] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4):107:1–107:14, 2017.

[49] Raymond A. Yeh, Chen Chen, Teck-Yian Lim, Alexander G. Schwing, Mark Hasegawa-Johnson, and Minh N. Do. Semantic image inpainting with deep generative models. In *2017 IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6882–6890, 2017.

[50] C. Wang, H. Huang, X. Han, and J. Wang. Video inpainting by jointly learning temporal structure and spatial details. *CoRR*, abs/1806.08482, 2018.

[51] Daniel Miller and Ira Kemelmacher-Shlizerman and Steven M. Seitz. MegaFace: A Million Faces for Recognition at Scale. *CoRR*, abs/1505.02108.

[52] B. Kratzwald, Z. Huang, D. Pani Paudel, and L. Van Gool. Improving video generation for multi-functional applications. *CoRR*, abs/1711.11453, 2017.

[53] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017.

[54] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems 30: Annual conf. on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5769–5779, 2017.

[55] M. Hofmann, S. M. Schmidt, AN. Rajagopalan, and G. Rigoll. Combined face and gait recognition using alpha matte preprocessing. In *Proc. of the 5th IAPR Int. Conf. on Biometrics*, pages 1–8, New Delhi, India, Mar. 2012.

[56] X. Zhou and B. Bhanu. Feature fusion of side face and gait for video-based human identification. *Pattern Recognition*, 41(3):778–795, 2008.

[57] A. K. Jain and A. Ross. Multibiometric systems. *Communications of the ACM*, 47(1): 34–40, January 2004. ISSN 0001-0782.

[58] D. Muramatsu, H. Iwama, Y. Makihara, and Y. Yagi. Multi-view multi-modal person authentication from a single walking image sequence. In *Biometrics (ICB), 2013 Int. conf. on*, pages 1–8, 2013. doi: 10.1109/ICB.2013.6612979.

[59] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar. 1998.

[60] 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*, pages 1049–1058, Ny, USA, July 2005.

[61] 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 IV*, Orlando, FL, USA, Mar. 2008.

[62] 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*, 30(2):342–347, 2008.

[63] M. He, S. J. Horng, P. Fan, R. S. Run, R. J. Chen, J. L. Lai, M.K. Khan, and K. O. Sentosa. Performance evaluation of score level fusion in multimodal biometric systems. *Pattern Recognition*, 43(5):1789–1800, May 2010.

[64] Y. Makihara, D. Muramatsu, M.A. Hossain, H. Iwama, T.T. Ngo, and Y. Yagi. Score-level fusion by generalized delaunay triangulation. *Proc. of the 2nd Int. Joint conf. on Biometrics*, 2014.

[65] J.D. Shutler, M.G. Grant, M.S. Nixon, and J.N. Carter. On a large sequence-based human gait database. In *Proc. of the 4th Int. Conf. on Recent Advances in Soft Computing*, pages 66–71, Nottingham, UK, Dec. 2002.

[66] Z. Liu and S. Sarkar. Outdoor recognition at a distance by fusing gait and face. *Image Vision Comput.*, 25(6):817–832, June 2007. ISSN 0262-8856.

[67] I. Bouchrika and M. Nixon. Exploratory factor analysis of gait recognition. In *Proc. of the 8th IEEE Int. conf. on Automatic Face and Gesture Recognition*, pages 1–6, Amsterdam, The Netherlands, Sep. 2008.

[68] R. Gross and J. Shi. The cmu motion of body (mobo) database. Technical report, CMT, Jun. 2001.

[69] K. Bashir, T. Xiang, and S. Gong. Gait recognition using gait entropy image. In *Proc. of the 3rd Int. Conf. on Imaging for Crime Detection and Prevention*, pages 1–6, Dec. 2009.

[70] D. Tao, X. Li, X. Wu, and S. Maybank. Human carrying status in visual surveillance. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 1670–1677, New York, USA, Jun. 2006.

[71] N. Otsu. Optimal linear and nonlinear solutions for least-square discriminant feature extraction. In *Proc. of the 6th Int. Conf. on Pattern Recognition*, pages 557–560, 1982.

[72] Olivier Chapelle and S Sathiya Keerthi. Efficient algorithms for ranking with svms. *Information Retrieval*, 13(3):201–215, 2010.

[73] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi. GEInet: View-invariant gait recognition using a convolutional neural network. In *Proc. of the 8th IAPR Int. Conf. on Biometrics (ICB 2016)*, number O19, pages 1–8, Halmstad, Sweden, Jun. 2016.

[74] R. Martín-Félez and T. Xiang. Gait recognition by ranking. In Andrew W. Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV (1)*, volume 7572 of *Lecture Notes in Computer Science*, pages 328–341. Springer, 2012.

[75] R. Martín-Félez and T. Xiang. Uncooperative gait recognition by learning to rank. *Pattern Recognition*, 47(12):3793–3806, 2014.

[76] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Joint intensity transformer network for gait recognition robust against clothing and carrying status. *IEEE Trans. on Information Forensics and Security*, 2019.

[77] Z. Wu, Y. Huang, and L. Wang. Learning representative deep features for image set analysis. *IEEE Trans. on Multimedia*, 17(11):1960–1968, Nov 2015. ISSN 1520-9210. doi: 10.1109/TMM.2015.2477681.

[78] G. Zhao, L. Cui, and H. Li. Gait recognition using fractal scale. *Pattern Analysis and Applications*, 10(3):235–246, Aug 2007.

[79] C. Chen, J. Liang, H. Zhao, H. Hu, and J. Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977 – 984, 2009.

[80] J. Ortells, R. A. Mollineda, B. Mederos, and R. Martín-Félez. Gait recognition from corrupted silhouettes: a robust statistical approach. *Machine Vision and Applications*, 28(1):15–33, Feb 2017.

[81] P. Nangtin, P. Kumhom, and K. Chamnongthai. Gait identification with partial occlusion using six modules and consideration of occluded module exclusion. *Journal of Visual Communication and Image Representation*, 36:107 – 121, 2016. ISSN 1047-3203.

[82] D. Muramatsu, Y. Makihara, and Y. Yagi. Gait regeneration for recognition. In *2015 Int. conf. on Biometrics (ICB)*, pages 169–176, May 2015.

[83] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Int. conf. on Learning Representations*, 2016.

[84] A. A. Efros and T. K. Leung. Texture synthesis by non-parametric sampling. In *Proc. of the Seventh IEEE Int. conf. on Computer Vision*, volume 2, pages 1033–1038 vol.2, Sept 1999.

[85] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proc. of the 27th Annual conf. on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 417–424, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co. ISBN 1-58113-208-5.

[86] Y. Wexler, E. Shechtman, and M. Irani. Space-time completion of video. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(3):463–476, March 2007. ISSN 0162-8828.

[87] M. Mirza and S. Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[88] C. Vondrick, H. Pirsiavash, and A. Torralba. Generating videos with scene dynamics. In *Advances in Neural Information Processing Systems 29: Annual conf. on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 613–621, 2016.

[89] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(1):221–231, January 2013. ISSN 0162-8828.

[90] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proc. of the 2015 IEEE Int. conf. on Computer Vision (ICCV)*, ICCV '15, pages 4489–4497, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2.

[91] A. Ross, K. Nandakumar, and A. K. Jain. *Handbook of Multibiometrics*. Int. Series on Biometrics. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387222960.

[92] A. K. Jain, K. Nandakumar, and A. Ross. Score normalization in multimodal biometric systems. *Pattern Recognition*, 38(12):2117–2127, December 2005.

[93] N. Poh and S. Bengio. Can chimeric persons be used in multimodal biometric authentication experiments? In *Proc. of the Int. Workshop on Machine Learning for Multimodal Interaction*, pages 87–100. Springer, 2005.

[94] J. L. Wayman. A path forward for multi-biometrics. In *Proc. of the IEEE Int. conf. on Acoustics Speech and Signal Processing Proc.*, pages 1069–1072, 2006.

[95] 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*, 43 (3):1094–1105, Mar. 2010.

[96] N. Poh and S. Bengio. Database, protocols and tools for evaluating score-level fusion algorithms in biometric authentication. *Pattern Recognition*, 39(2):223–233, 2006.

[97] BANCA. Banca score database. http://personal.ee.surrey.ac.uk/Personal/ Norman.Poh/web/ banca_multi/main.php%3Fbodyfile=entry_ page.html.

[98] NIST Biom. NIST biometric scores set. http://www.itl.nist.gov/iad/894.03/biometricscores/l.

[99] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. Freire, J. Gonzalez-Rodriguez, C.Garcia-Mateo, J.-L.Alba-Castro, E.Gonzalez-Agulla, E.Otero-Muras, S.Garcia-Salicetti, L.Allano, B.Ly-Van, B.Dorizzi, J.Kittler, T.Bourlai, N.Poh, F.Deravi, M.Ng, M.Fairhurst, J.Hennebert, A.Humm, M.Tistarelli, L.Brodo, J.Richiardi, A.Drygajlo, H.Ganster, F.M.Sukno, S.-K.Pavani, A.Frangi, L.Akarun, and A.Savran. The multi-scenario multi-environment biosecure multimodal database (BMDB). *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 32(6):1097–1111, June 2010.

[100] K. Messer, J. Matas, J. Kittler, J. Lüttin, and G. Maitre. XM2VTSDB: The Extended M2VTS Database. In *In Second Int. Conf. on Audio and Video-based Biometric Person Authentication*, pages 72–77, 1999.

[101] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independant speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[102] N. Poh and S. Bengio. F-ratio client-dependent normalisation on biometric authentication tasks. In *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, pages 721–724, Philadelphia, 2005.

[103] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target dependent score normalisation techniques and their application to signature verification. In *In LNCS 3072, Int'l Conf. on Biometric Authentication (ICBA)*, pages 498–504, Hong Kong, 2004.

[104] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzalez-Rodriguez. Target dependent score normalization techniques and their application to signature verification. *IEEE Trans. on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 35(3):418 –425, Aug. 2005.

[105] Y. Makihara, M.A. Hossain, D. Muramatsu, and Y. Yagi. Score-level fusion based on the direct estimation of the bayes error gradient distribution. In *Proc. of the Int. Joint Conf. on Biometrics (IJCB2011)*, pages 1–8, Washington D.C., USA, Oct. 2011.

[106] B. DeCann and A Ross. Relating roc and cmc curves via the biometric menagerie. In *Biometrics: Theory, Applications and Systems (BTAS), 2013 IEEE Sixth Int. conf. on*, pages 1–8, Sept 2013. doi: 10.1109/BTAS.2013.6712705.

[107] N. Poh and J. Kittler. A unified framework for biometric expert fusion incorporating quality measures.

[108] P. Grother and E. Tabassi. Performance of biometric quality measures. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(4):531–543, 2007.

[109] F. Alonso-Fernandez, J. Fierrez, and J. Ortega-Garcia. Quality measures in biometric systems. *Security & Privacy, IEEE*, 10(6):52–62, 2012.

[110] R. Wong, N. Poh, J. Kittler, and D. Frohlich. Interactive quality-driven feedback for biometric systems. In *Proc. of the IEEE 4th Int. Conf. on Biometrics: Theory, Applications, and Systems*, pages 1–6, Washington D.C, USA, Sep. 2010.

[111] H. S. Bhatt, S. Bharadwaj, M. Vatsa, R. Singh, A. Ross, and A. Noore. A framework for quality-based biometric classifier selection. In *Proc. of the Int. Joint conf. on Biometrics*, pages 1–7, 2011.

[112] K. Nandakumar, Y. Chen, S.C. Dass, and A.K. Jain. Quality-based score level fusion in multibiometric systems. In *Proc. of the 18th Int. Conf. on Pattern Recognition*, volume 4, pages 473–476, 2006.

[113] 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*, pages 1124–1133, 2007.

[114] N. Poh, J.V. Kittler, and T. Bourlai. Improving biometric device interoperability by likelihood ratio-based quality dependent score normalization. In *Prof. of the IEEE 3rd Int. conf. on Biometrics: Theory, Applications and Systems*, pages 1–5, 2007.

[115] N. Poh, G. Heusch, and J. Kittler. On combination of face authentication experts by a mixture of quality dependent fusion classifiers. In *Proc. of th 7th Int. Workshop on Multiple Classifier Systems*, pages 344–356, Prague, Czech Republic, 2007.

[116] M.Z. Uddin, TT Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, Y. Yagi. The OU-ISIR Large Population Gait Database with real-life carried object and its performance evaluation. *IPSJ Trans. on Computer Vision and Applications* 10(1):5, May 2018.

[117] Y. Makihara, T. Kimura, F. Okura, I. Mitsugami, M. Niwa, C. Aoki, A. Suzuki, D. Muramatsu, and Y. Yagi. Gait collector: An automatic gait data collection system in conjunction with an experience-based long-run exhibition. In *Proc. of the 8th IAPR Int. Conf. on Biometrics (ICB 2016)*, number O17, pages 1–8, Halmstad, Sweden, Jun. 2016.

[118] C. Schultz. Digital keying methods. *University of Bremen Center for Computing Technologies, Tzi*, 4(2):3, 2006.

[119] Y. Makihara, H. Mannami, and Y. Yagi. Gait analysis of gender and age using a large-scale multi-view gait database. In *Proc. of the 10th Asian Conf. on Computer Vision*, pages 975–986, Queenstown, New Zealand, Nov. 2010.

[120] X. Hongye and H. Zhuoya. Gait recognition based on gait energy image and linear discriminant analysis. In *2015 IEEE Int. conf. on Signal Processing, Communications and Computing (ICSPCC)*, pages 1–4, Sept 2015.

[121] C. Chang and C. Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2(3):27:1–27:27, May 2011. ISSN 2157-6904.

[122] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society conf. on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546 vol. 1, June 2005.

[123] C. Zhang, W. Liu, H. Ma, and H. Fu. Siamese neural network based gait recognition for human identification. In *2016 IEEE Int. conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2832–2836, March 2016.

[124] B. DeCann and A. Ross. Gait curves for human recognition, backpack detection, and silhouette correction in a nighttime environment. volume 7667, pages 7667 – 7667 – 13, 2010. doi: 10.1117/12.851296.

[125] M. Lee, M. Roan, B. Smith, and T. E. Lockhart. Gait analysis to classify external load conditions using discriminant analysis. *Human Movement Science*, 28(2):226 – 235, 2009. ISSN 0167-9457.

[126] L. Tran, Xi Yin, and X. Liu. Disentangled representation learning gan for pose-invariant face recognition. In *Proc. of the IEEE int. conf. on Computer Vision and Pattern Recognition*, pages 1415–1424, 2017.

[127] Yiwei He, Junping Zhang, Hongming Shan, and Liang Wang. Multi-task gans for view-specific feature learning in gait recognition. *IEEE Trans. on Information Forensics and Security*, 14(1):102–113, 2018.

[128] Z. Liu and S. Sarkar. Effect of silhouette quality on hard problems in gait recognition. *Trans. of Systems, Man, and Cybernetics Part B: Cybernetics*, 35(2):170–183, 2005.

[129] C. Lu, M. Hirsch, and B. Schölkopf. Flexible spatio-temporal networks for video prediction. In *2017 IEEE conf. on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2137–2145, 2017.

[130] H. Cai, C. Bai, Y. Tai, and C. Tang. Deep video generation, prediction and completion of human action sequences. In *Computer Vision - ECCV 2018 - 15th European conf., Munich, Germany, September 8-14, 2018, Proc., Part II*, pages 374–390, 2018.

[131] S. Yu, H. Chen, E. B. García Reyes, and N. Poh. Gaitgan: Invariant gait feature extraction using generative adversarial networks. In *2017 IEEE conf. on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 532–539, 2017.

[132] M.Z. Uddin, D. Muramatsu, T. Kimura, Y. Makihara, Y. Yagi MultiQ: Single sensor-based multi-quality multi-modal large-scale biometric score database and its performance evaluation. *IPSJ Trans. on Computer Vision and Applications* 9(1):18, Jul. 2017.

[133] F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.

[134] O. Ronneberger, P. Fischer, editor="Navab N. Brox, T.", J. Hornegger, W. M. Wells, and A. F. Frangi. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer Int. Publishing. ISBN 978-3-319-24574-4.

[135] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus. Deconvolutional networks. In *2010 IEEE Computer Society conf. on Computer Vision and Pattern Recognition*, pages 2528–2535, June 2010.

[136] . He, X. Zhang, S. Ren, and J Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proc. of the 2015 IEEE Int. conf. on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2.

[137] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of the 32Nd Int. conf. on Int. conf. on Machine Learning - Volume 37*, ICML'15, pages 448–456. JMLR.org, 2015.

[138] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *CoRR*, abs/1505.00853, 2015.

[139] L. J. Ba, R. Kiros, and G. E. Hinton. Layer normalization. *CoRR*, abs/1607.06450, 2016.

[140] A. Dosovitskiy and T. Brox. Generating images with perceptual similarity metrics based on deep networks. In *Advances in neural information processing systems*, pages 658–666, 2016.

[141] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[142] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, and J. Bigun. Discriminative multimodal biometric authentication based on quality measures. *Pattern Recognition*, 38(5):777–779, May 2005.

[143] 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)*, pages 402–411, Sept 2012. doi: 10.1109/BTAS.2012.6374607.

[144] S. Bengio, C. Marcel, S. Marcel, and J. Mariethoz. Confidence measures for multimodal identity verification. *Information Fusion*, 3(4):267–276, 2002.

[145] 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*, pages 1–6, 27-29 2007. doi: 10.1109/BTAS.2007.4401906.

[146] 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)*, pages 1537–1543, Providence, RI, USA, Jun. 2012.

[147] K.A. Toh, W.Y. Yau, E. Lim, L. Chen, and C.H. Ng. Fusion of auxiliary information for multi-modal biometrics authentication. In *Proc. of Int. Conf. on Biometrics*, pages 678–685, Hong Kong, 2004.

[148] 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.

[149] Open CV. Intel open source computer vision library. http://opencv.org.

[150] W. Zou and P. Yuen. Very low resolution face recognition problem. *IEEE Trans. on Image Processing*, 21(1):327–340, 2012.

[151] Z. Wang, Z. Miao, QM Jonathan Wu, Y. Wan, and Z. Tang. Low-resolution face recognition: a review. *The Visual Computer*, 30(4):359–386, 2014.

[152] C.W. Tan and A. Kumar. Towards online iris and periocular recognition under relaxed imaging constraints. *IEEE Trans. on Image Processing*, 22(10):3751–3765, Oct. 2013.

[153] K. Ito, T. Aoki, T. Hosoi, and K. Kobayashi. Face recognition using phase-based correspondence matching. In *Proc. of the IEEE Int. conf. on Automatic Face Gesture Recognition and Workshops (FG 2011)*, pages 173–178, March 2011. doi: 10.1109/FG.2011.5771393.

[154] A. Asthana, M. Jones, T. Marks, K. Tieu, and R. Goecke. Pose normalization via learned 2d warping for fully automatic face recognition. In *Proc. of the British Machine Vision conf.*, pages 1–11, 2011. ISBN 1-901725-43-X.

[155] G. Tzimiropoulos and M. Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proc. of the IEEE conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1851–1858, June 2014.

[156] 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*, 7346:1–14, 2006.

[157] J. Neyman and E. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Trans. of the Royal Society of London*, 231:289–337, 1933.

[158] K. Nandakumar, A. K. Jain, and A. Ross. Fusion in multibiometric identification systems: What about the missing data? In *Proc. of the 3rd Int. conf. on Biometric*, pages 743–752, 2009. ISBN 978-3-642-01792-6.

[159] B. DeCann and A. Ross. Can a "poor" verification system be a "good" identification system? a preliminary study. In *Proc. of the IEEE Int. Workshop on Information Forensics and Security*, pages 31–36, Dec 2012. doi: 10.1109/WIFS.2012.6412621.

[160] M.Z. Uddin, D. Muramatsu, N. Takemura, MAR Ahad, Y. Yagi. Spatio-temporal silhouette sequence reconstruction for gait recognition against occlusion. *IPSJ Trans. on Computer Vision and Applications* 11(1):9, Nov. 2019.