Computer Vision Techniques for Gait-based Visual Surveillance

July 2012

Haruyuki IWAMA

Computer Vision Techniques for Gait-based Visual Surveillance

Submitted to
Graduate School of Information Science and Technology
Osaka University

July 2012

Haruyuki IWAMA

Abstract

The importance of automated visual surveillance in public space has been increased in response to the recent rising concerns about safe and security, and computer vision-based person identification techniques play a key role in it. Gait as a biometric cue has received much attention in recent years due to the ability of identifying individuals at a distance, and gait-based person identification technique could contribute much to crime investigation and safety confirmation through wide-area surveillance. Although gait-based person identification has such a promising ability, several critical issues need to be sufficiently considered when applying it to real visual surveillance tasks. Among them, this thesis addresses following three issues, each of which is corresponded to a primal step in gait-based person identification: i) accuracy of foreground segmentation in preprocessing step, ii) robustness to intra-subject variations in identification step, and iii) statistical reliability in performance evaluation step.

First, a research for the first issue is described. We propose a method for accurate foreground segmentation in the presence of strong shadow. For the separation of foreground and shadow, the homography constraint in binocular system is used. In addition, while existing homography-based methods often suffer from the occlusion relationship, the proposed method takes such relationship into account explicitly by using a homography correspondence pair-based symmetric labeling scheme. The scheme is formulated in the form of energy minimization problem and optimized by an α–β swap algorithm. The experimental results demonstrate that the proposed method realizes more accurate segmentation than the existing methods in the presence of strong shadow and occlusion.

Next, we propose a novel person identification framework where the identification performance could be enhanced against intra-subject variations. We pay attention to the fact that people often act in groups such as friends, family, and co-workers in social living and we utilize this as a cue for identifying individuals to improve the identification performance. The individual cues and the group cue are integrated in the form of conditional random field model, and the identities of individuals are optimized via belief propagation algorithm. The comparison experiments with the straightforward identification scheme show the effectiveness of the proposed method.

Finally, we construct the world's largest gait database. The database includes 4,007 subjects (2,135 males and 1,872 females) with ages ranging from 1 to 94 years. The database enables the statistically reliable performance comparison among state-of-the-art gait features for person identification. Also, we investigate the dependences of the identification performance on gender and age group, and several novel insights are provided such as the gradual change in identification performance with human growth.

Together with the considering of these issues, this thesis could make a large contribution to the development of more accurate and practical gait-based visual surveillance.

Acknowledgments

I want to express my gratitude to all those people who have technically and emotionally supported me to accomplish this thesis. First and foremost, I would like to express my heartfelt appreciations and gratitude to my excellent and supportive supervisors, Professor Yasushi Yagi. It was my good fortune to have an acceptance from Professor Yagi as my supervisor. Throughout my research, he patiently provided the vision, philosophy, encouragement and a great deal of advice necessary for me. Above all, he offered me a great and challenging topic, and a wonderful and conducive environment for developing my doctoral work in his lab.

I am deeply grateful to my respectable adviser, Assistant Professor Yashushi Makihara. For three years and half studying under his guidance, I learned a great deal from the discussions with him and I received lots of abundant advice on scientific approach, logical way of thinking, experiments, paper writing, and presentation. Also, his immense knowledge and inspiration have been of great value for me. Without his contributions, this work would not have been possible.

The years studying in Yagi Lab, The Institute of Scientific and Industrial Research, Osaka University, has been a great time to learn experience from excellent teachers. I wish to express my sincere thanks to Associate Professor Yasuhiro Mukaigawa for his valuable advice, warm encouragement, and friendly help. I would also like to gratitude Assistant Professor Ikuhisa Mitsugami, Specially Appointed Associate Professor Daigo Muramatsu, Specially Appointed Assistant Professor Hirotake Yamazoe, and Guest Associate Professor Ryusuke Sagawa. They never hesitated to give their experience and constructive suggestions throughout my studies.

I want to express my gratitude to all reviewers, Professor Toshimitsu Masuzawa, Professor Shinji Kusumoto, and Professor Haruo Takemura, for their kind efforts to review my dissertation.

I have had the support and encouragement of the secretaries and technical staffs in Yagi Lab. They have been very understanding and friendly advisers on non-scientific side of my campus life. I greatly appreciate Noriko Yasui, Masako Kamura, and Makiko Fujimoto, who gave me many types of detailed assistance, such as that on accounting work, paperwork, and

English proofreading. Thanks to their works, I could concentrate on my research. I owe a debt of gratitude to technical staffs, Aya Iiyama, Yoko Irie, Yoshiko Matsumoto, Mika Iguchi, and Yoshimi Ohkohchi for their patient and tireless efforts and careful works. Many experiments in my research owed to their technical supports.

The members of gait research group in Yagi Lab including present and past students have contributed immensely to my personal and professional time. My heartfelt appreciation goes to Mayu Okumura. She was a pioneer and center of our gait database project, and my trusty work partner in the project. Throughout the project, I shared good times and bad times with her, and the project could never be achieved without her conscientious effort and fantastic work. Also, she was always great sources of laughter, joy, and encouragement, and her genial and pleasing character gave me a warm fuzzy feeling at all times. Special thanks to my early neighbor in the lab, Naoki Akae. I usually enjoyed technical discussion and desultory chat with him. He joined all the experiments in my study and made a careful effort to fulfill my many demands without hesitation. In addition, my desire for improvement had been fueled by his prominent skill in software development and his smart working method. I would like to express my gratitude to Atsushi Mori. He was the first friend in this lab and he not only technically supported me, but also helped me as a good guide of Yagi Lab. Owing to his kindness and friendliness, I could get used to life in the lab smoothly. I wish to express my warm thanks to Akira Shiraishi, a beloved clown of the lab, and Ryo Kawai, a diligent student with polite and sincere manner, and other gait research group members for their generous contributions to my work.

I have greatly benefited from other group members. I would particularly like to thank Barbie (Yoko Baba). She was a queen of pleasure in the lab with wild and innocent mind, and her inspiring ideas and cheerful words and actions were amusing and exciting, and they constantly prompted a smile from me. Moreover, her ingenious proposals of parties, excursions, and other events made my time at the lab more enjoyable and precious. I am indebted to Sakatch (Kazuhiro Sakashita), who has been a loose idol in the lab with resilient, artistic, charming, and unhealthy belly fat. He often made my feel relaxed by providing a very laid-back atmosphere. I am grateful to Inotch (Chika Inoshita), who is an undisputed ace of the young Japanese researcher, and is seen as a future empress of the computer vision community. I greatly relished various discussions with her, and she frequently listened to my whining and encouraged me as my trusted counselor. In addition, her ambitious spirit and brilliant capability as a researcher considerably motivated me.

Finally, I would like to thank all the other members in Yagi Lab. They have been very nice to me and make my campus life happy.

List of Publications

A. International Journal Papers (Full paper reviewed)

[1] Haruyuki Iwama, Yasushi Makihara, Yasushi Yagi, "Group Context-aware Person Identification in Video Sequences", IPSJ Transactions on Computer Vision and Applications (accepted).

[2] Haruyuki Iwama, Mayu Okumura, Yasushi Makihara, Yasushi Yagi, "The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait Recognition", IEEE Transactions on Information Forensics and Security (accepted).

B. Domestic Journal Papers (Full paper reviewed)

[1] 岩間 晴之, 槇原 靖, 八木 康史, "ホモグラフィ対応ペアに基づく前景・影・背景のセグメンテーション", 電子情報通信学会論文誌 D, Vol. J94-D, No. 8, pp. 1300-1313 , Aug. 2011.

C. International Conference (Full paper reviewed)

[1] Haruyuki Iwama, Yasushi Makihara, Yasushi Yagi, "Foreground and Shadow Segmentation based on a Homography-Correspondence Pair", Proceeding of the 10th Asian Conference on Computer Vision (ACCV2010), pp.2790-2802, Queenstown, New Zealand, Nov. 2010.

[2] Mayu Okumura, Haruyuki Iwama, Yasushi Makihara, Yasushi Yagi, "Performance Evaluation of Vision-based Gait Recognition using a Very Large-scale Gait Database", Proceeding of the IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems (BTAS2010), pp.1-6, Washington D.C., USA, Sep. 2010.

[3] Yasushi Makihara, Mayu Okumura, Haruyuki Iwama, Yasushi Yagi, "Gait-based Age Estimation using a Whole-generation Gait Database", Proceeding of the International Joint Conference on Biometrics (IJCB2011), no.195, pp.1-6, Washington D.C., USA, Oct. 2011.

[4] Haruyuki Iwama, Daigo Muramatsu, Yasushi Makihara, Yasushi Yagi, "Gait-based Person-Verification System for Forensics", Proceeding of the IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS2012), Washington D.C., USA, Sep. 2012. (accepted).

D. International Conference (Without review)

[1] Haruyuki Iwama, Yasushi Makihara, Yasushi Yagi, "Homography-correspondence Pair Based Foreground and Shadow Segmentation", The 15th SANKEN International Symposium, Jan. 2012.

E. Domestic Conference (Full paper reviewed)

[1] 岩間 晴之, 槇原 靖, 八木 康史, "ホモグラフィ対応ペアに基づく前景・影・背景のセグメンテーション", 第 13 回 画像の認識・理解シンポジウム (MIRU2010), 7 月, 2010.

[2] 奥村 麻由, 岩間 晴之, 槇原 靖, 八木 康史, "大規模データベースを用いた歩容認証手法の性能評価", 第 13 回 画像の認識・理解シンポジウム (MIRU2010), 7 月, 2010.

F. Domestic Conference (Without review)

[1] 岩間 晴之, 槇原 靖, 八木 康史, "ホモグラフィ対応ペアに基づいた前景と影のセグメンテーション", 情報処理学会研究報告 UC 研究会, 情報処理学会, pp.1-8, 6 月, 2010.

Contents

Chapter 1

Introduction

Realization of the safe and secure life is always social demand, and the visual surveillance has been contributed to it. Together with the increase in fears of violent crimes and terrors such as the events of September 11, 2001, USA and the bomb attacks in London on the July 2005, today the importance of visual surveillance, especially the surveillance from a distance, has been definitely increased worldwide to observe the broad area activities of people and vehicles with minimum blind area. In fact, an enormous number of surveillance cameras are deployed in a wide range of public places (e.g., airport, underground station, street, school, shopping mall, parking lot, and sports arena) for crime reduction and risk management [1]. On the other hand, massive surveillance cameras have posed expensive cost for manual operation required to manage them. As a result, the video from the cameras cannot be always monitored and is often used only as a record for post investigation of an incident. For crime prevention and investigation and safety confirmation, however, real-time event detection from lots of video data is still needed and also, comprehensive analysis for widely distributed cameras is desired for such purposes. To meet such needs, computer-assisted surveillance technology has been developed in recent years [1] with the great advance of computer processing power.

Computer vision technology takes the central role in automated surveillance [2, 3, 4, 5, 6]. For example, the techniques of person detection, tracking, and action recognition from videos ensure the automatic alert ability for intruders and suspicious individuals with abnormal behavior. In addition, the techniques of character recognition realize the timely detection and online tracking of the wanted vehicles based on their number plates. Finally, biometric-based person identification techniques make the surveillance system to be more intelligent, that is, the system acquires the ability to know that "*who exists or dose not exist in the area*". This enables the automatic detection of suspected persons, unwelcome strangers, and protected persons (e.g.,

[1] More than 4.2 million cameras are deployed in UK.

children and seniors). Consequently, this significantly contributes to realization of the safety and security in our society.

There are two major biometric cues for person identification in visual surveillance, face and gait[2]. As for the face-based person identification from the videos or still images, a considerable number of techniques have been developed [7, 8, 9, 10, 11]. As a result, the techniques are now used commonly not only for visual surveillance [12, 13, 14], but also for access control [15, 16, 17], image searching [18, 19], and consumer photo management [20, 21]. Also, many commercial softwares of face recognition are available (e.g., FaceIt–SDK [22], FaceVACS–SDK [23], OKAO Vision [24], and NeoFace [25]). Automatic face recognition mainly uses 2D front-face pattern and texture information and it essentially requires the high-resolution image, though there exists some techniques of obtaining high resolution image from low resolution images. This restricts its range of application in visual surveillance in terms of not only the observation view, but also the distance to the subjects. Thus, face-based identification is not suitable for many of the videos from surveillance cameras deployed in public space, in which the subjects are captured from a distance. In addition, the individual face can easily be altered or concealed by dark glasses and mask.

On the other hand, gait-based biometrics is a relatively new area of study within the community of computer vision research [26]. The gait has attractive advantages including the difficulty to disguise and the ability of identifying individuals at a distance without their cooperation. Therefore, it is expected to be applied to the long-distance surveillance in public space. In fact, gait-based verification from public CCTV images has been admitted as evidence in UK [27], and gait evidence has been used as a cue for criminal investigations in Japan.

In the last decade, various gait-based person identification techniques have been studied, and there are two different approaches: model-based approach and model-free approach. Model-based approaches fit a model to input images and represent gait features such as shape and motion by the parameters of the model. Some methods [28, 29] extracted periodical features of leg motion by Fourier analysis. Bobick et al. [30] extracted parameters of shape and stride. Wagg et al. [31] extracted static shape parameters and gait period with an articulated body model, and Urtasun et al. [32] extracted joint angles with an articulated body model. Although model-based approaches are generally view-invariant and scale-invariant and these advantages are important for practical applications, the approaches tend to be more complex and computationally more expensive than model-free approaches. In addition, model-based

[2]Manner of walking.

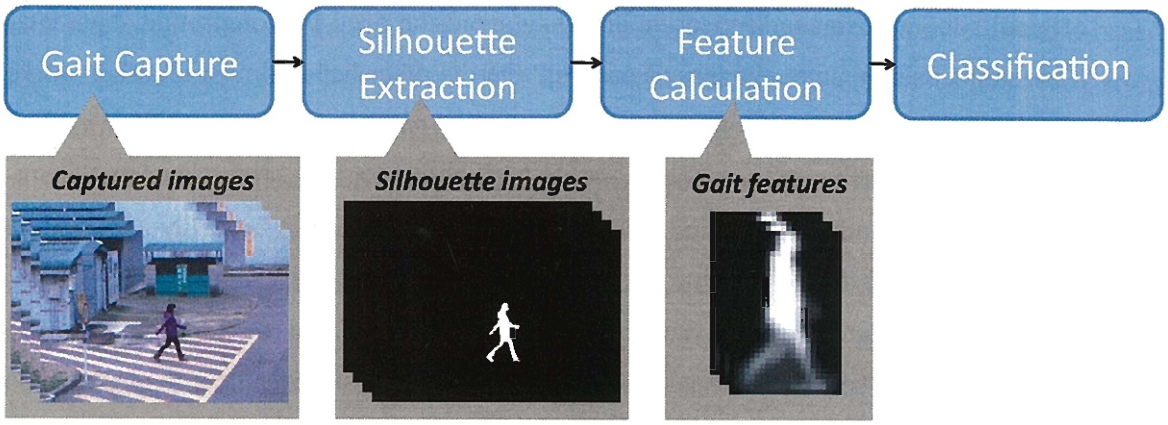


Figure 1.1: A general processing flow of model-free gait identification (the GEI [36] is illustrated as gait feature).

approaches often require good quality images to calculate the model parameters with high accuracy from a gait sequence, which may be difficult to obtain in a real surveillance system.

Recent trends in gait-based person identification are model-free approaches due to the low computational cost and robustness to the noise. Model-free approaches represent gait features by directly analyzing the subject's binary silhouettes without the use of a model. A general framework of model-free approach is shown in Fig. 1.1, which consists of silhouette extraction including subject segmentation, feature extraction, and classification procedures. To date, a variety of mode-free methods have been proposed ranging from a method of direct matching of silhouette images, which is called as a baseline algorithm [33], to hidden Markov model-based method [34]. Among them, a periodic template-based methods seem to be the mainstream due to its simplicity and high performance, as typified by a method based on averaged silhouette [35] (which is also known as Gait Energy Image (GEI) [36]). In recent works, it is reported that, when there is no subject's condition change, nearly 100 % identification performance is achieved by state-of-the-art methods [37, 38] for public gait databases [39, 40] which provide relatively clear silhouette sequences of over 100 subjects.

Although such successful performance has been achieved by model-free approaches, there are still some significant issues posed to the practical use of gait-based person identification for visual surveillance[3]. Of these, we address the following major three issues in this thesis, each of which corresponds to a primal step in gait-based person identification.

1. **Accuracy of foreground segmentation in preprocessing step:** Many of the existing works use the public gait databases such as [39, 33, 40] which provide normalized or

[3]To the best of our knowledge, there is no study which evaluates the gait-based person identification performance with real surveillance videos automatically.

original silhouette image sequences, and they avoid the need for foreground segmentation (subject's silhouette extraction) process. On the other hand, almost all other works use original gait dataset captured in rather controlled environment only for their evaluation. Typically, each sequence contains only one subject without strong shadow. Therefore, the foreground segmentation task is not so serious and it can be easily achieved by straightforward background subtraction technique. In the real surveillance videos, however, the task is not always easy, especially in outdoor scene due to strong shadow. The failure in shadow removal causes not only the distortion of a silhouette's shape, but also the mergence of silhouettes of two or more subjects, and consequently, the performance could be significantly decreased. Although shadow removal is major problem in computer vision and many color-based techniques have been proposed, it is still difficult to separate shadow from the foreground accurately in the presence of strong shadow.

2. **Robustness to intra-subject variations in identification step:** Gait is behavioral characteristic and tends to be more fluctuated for each attempt (walking) than physiological characteristic such as face. The fluctuation often arises in arm swing and head pose, and especially, it might be notably appeared in children's walk due to the immaturity of their walking. Generally, identification performance is decreased by such fluctuation. In addition, various condition changes such as observation view, clothing, and carrying condition changes cause the identification performance decrement as reported in [33]. Although, there are some works which aim to construct robust scheme for such intra-variations (e.g., [41] addresses the view-invariant scheme and clothing change is considered in [42]), the performance cannot be fully recovered.

3. **Statistical reliability of performance in performance evaluation step:** For the statistically reliable evaluation of gait identification approaches, the construction of a common gait database is essential. Though several gait databases have been constructed to date [43, 44, 39, 45, 46, 47, 33, 40, 48, 49, 50, 42, 51, 52], these databases include at most 185 subjects [51] and the subjects' genders and ages are biased in many of the databases. Therefore, these are insufficient for the performance evaluation especially in terms of the number and diversity of the subjects.

The thesis is organized as follow. Chapter 2 presents a novel framework of foreground and shadow segmentation with a static binocular system is described for the first issue. Homography constraint is one of geometric constraints in a multi-camera system, and it is often

used for foreground separation from shadow. Though existing homography constraint-based segmentation approaches can work well in the presence of strong shadow, it often suffer from occlusion problems between foreground and shadow. In our approach, to explicitly take the occlusion relationship into account, we treat a homography-correspondence pair symmetrically. Also, we regard the segmentation problem as a multi-labeling problem for each homography-correspondence pair. We then formulate the problem as an energy minimization problem, and get the pair-wise labeling results by minimizing it via an α–β swap algorithm. Experimental results show that accurate segmentation is obtained in the presence of the occlusion region in each side image.

Next, we describe a framework of group context-aware person identification for the second issue in Chapter 3. In social living scenarios, people often act in groups composed of friends, family, and co-workers. We utilize this as a cue for person identification to improve identification performance in the presence of the attenuation of gait-based identity caused by intra-subject variations. The relationships between the people in an input sequence are modeled using a graphical model. The identity of each person is then propagated to their neighbors in the form of message passing in the graph via belief propagation, depending on each person's group affiliation information and their characteristics, such as spatial distance and velocity vector difference, so that the members of the same group with similar characteristics enhance each other's identities as group members. The proposed method is evaluated through gait-based person identification experiments using both simulated and real input sequences. Experimental results show that the identification performance is considerably improved when compared with that of the straightforward method based on the gait feature alone.

Then, we describe the construction of the world's largest gait database—the "*OU-ISIR Gait Database, Large Population Dataset*"—and its application to a statistically reliable performance evaluation of gait-based person identification for the third issue in Chapter 4. Whereas existing gait databases include at most 185 subjects, we construct a larger gait database that includes 4,007 subjects (2,135 males and 1,872 females) with ages ranging from 1 to 94 years. The dataset allows us to determine statistically significant performance differences between currently proposed gait features. In addition, the dependences of identification performance on gender and age group are investigated and the results provide several novel insights, such as the gradual change in identification performance with human growth.

Finally, conclusions are drawn and future work is discussed in Chapter 5.

Chapter 2

Foreground and Shadow Segmentation based on Homography-correspondence Pair

2.1 Introduction

Foreground segmentation is crucial preprocessing for gait-based person identification. For extracting foreground, background subtraction has been widely used [53] for surveillance. The methods, however, often extract not only the objects but also their shadows, which can be problematic. The false detection of shadow as foreground causes the distortion of foreground appearance and mergence of the areas of foreground, and this impairs the original feature of foreground. As a result, the performance of gait-based person identification is significantly decreased. Therefore, shadow segmentation, detection, or removal is also important problem, and many techniques have been proposed for the purpose. Although, color-based method is the most popular [54], the method tends to be unstable in real environment.

On the other hand, multiple cameras-based surveillance with overlapped fields of view has attracted increasing interest in recent years, due to the demands of accurate detection [55] and tracking [56, 57, 58, 59, 60] of multiple people occluded by other people and analysis of their activities [61, 62] in a complex environment.

In such a multi-view framework, several geometric approaches have been applied to the foreground/shadow segmentation [63][64] taking advantage of the framework. One well known approach is foreground separation from shadow based on disparities [63]. However, it often suffers from mis-correspondence problems and cannot be applied to scenes with no texture.

Alternatively, a homography constraint is also popular as a geometric constraint between multiple viewpoints. Approaches based on homography aim mainly to distinguish standing

objects from ground plane objects including shadow [64]. Existing homography-based approaches, however, do not consider the occlusion relationship between foreground and shadow, and they tend to fail at the region of occlusion.

In the field of stereo correspondence problems, symmetric correspondence based approaches have been proposed to handle the occlusion appropriately [65]. These approaches explicitly take the occlusion relationship into account by treating a stereo correspondence pair in a symmetric way.

Inspired by the symmetric approaches, we propose a symmetric segmentation framework based on a homography constraint with occlusion handling between foreground and shadow. Our goal is *"how to segment foreground, shadow, and background"*, and we regard this segmentation problem as a homography-correspondence pair labeling problem. Then, we solve this in an energy minimization framework together with a graph-cut algorithm [66]. Considering the homography-correspondence symmetrically, we cannot only segment the occluded region correctly, but also acquire additional information about the occluded region, such as, what label is assigned to the occluded region, *shadow* or *background*. This kind of information is valuable for many multi-view applications.

The remainder of this paper is organized as follows. Section 2.2 describes related work. Section 2.3 introduces our segmentation framework. Section 2.4 describes the detailed implementation of the proposed method. Section 2.5 demonstrates the effectiveness of the proposed method using experiments and the limitation and our discussions are presented in Section 2.6. Finally, Section 2.7 concludes our work.

2.2 Related Work

Color-based approach: Most of color-based approaches are based mainly on the following two properties of shadow color: (a) The shadow region is darker than the original background region, (b) The color vector direction of the shadow region is similar to that of the original background region. These properties are considered in various color spaces such as RGB color space [54, 67] and HSV color space [68, 69]. Color-invariant feature is also used for shadow removal [70, 71]. The performance comparison result among these methods is reported in [72].

Recently, many learning-based approaches are actively investigated [73, 74, 75]. In these methods, first, the shadow candidate pixels are detected by weak shadow detector based on shadow color properties as described above, and then, a statistical shadow color model is constructed from the candidate pixels. Finally, each pixel is determined whether it belongs to

shadow, based on the goodness of fit of its color on the learned shadow color model. Any state-of-the-art color-based approach, however, could fail at the region with the same color with shadow color (e.g., head region of black haired person), and this is an essential problem of color-based approach.

Texture-based approach: Texture-based approach distinguishes shadow based on the following property: the texture of shadow region is the same with that of corresponding background region, while the texture of foreground region is different from that of corresponding background region. In the methods of Javed et al. [76] and Sanin et al. [77], first, shadow candidate regions are decided via color-based segmentation and blob analysis, and then, the similarities of gradient feature between shadow candidate regions and corresponding background region are calculated. Finally, shadow regions are determined by thresholding the similarities. Besides, some methods use both color and texture. A step-wise shadow detection and removal scheme based on color and texture features is proposed in [78]. In addition, learning-based methods are proposed in [79, 80], where both color and texture features are statistically modeled by learning. Texture-based approaches tend to be unstable for the region of weak texture and the scene with strong shadow such as outdoor scene in daylight.

Disparity-based approach: Disparity is often used for accurate foreground extraction rather than shadow detection. In [63], disparity-based background subtraction scheme is proposed for accurate and stable foreground extraction in the presence of rapid illumination change. In addition, some works integrate color and disparity [81, 82]. In [83], the foreground extraction problem is regarded as a graph-based energy minimization problem and disparity is used for the robust estimation of foreground *Seed*. Disparity-based approaches, however, often suffer from mis-correspondence problems and cannot be applied to scenes with no texture.

Homography-based approach: In the problem of trajectory estimation of soccer player via a static binocular system, Kasuya et al. [84] utilize the homography constraint for the separation of player and shadow region. In the method, first, foreground (player) candidate regions are extracted by background subtraction in each side image, and next, each candidate region is projected onto the field plane by homography transformation. Then, shadow regions are distinguished as the logical product regions in the plane between projected foreground candidate regions of each side camera. Hamid et al. [85] further consider the color similarity in such logical product regions in the plane to extract shadow. In the same manner, the homography constraint is used for objects detection (including shadow detection) in the ground plane [86] and the obstacle detection problem of mobile robot system [87]. In [64], both shadow color property and homography constraint are used for shadow detection. The method is composed

of learning phase and shadow detection phase. In learning phase, first, the foreground candidate pixels determined by background subtraction are divided into foreground and shadow classes based on homography constraint, and then, a shadow color model is constructed as mixture Gaussian distribution from the color information of pixels of the shadow class. In shadow detection phase, the shadow pixels are decided based on both homography constraint and the goodness of fit on the shadow color model.

In this way, homography constraint is often used for filtering the false-positives of shadow. Existing homography-based approaches, however, do not consider the occlusion relationship between two cameras explicitly, and therefore, they fail in separating foreground and shadow at the region of occlusion relationship. The detailed mechanism and the issue are presented in the next section.

2.3 Homography-Correspondence Pair-based Segmentation

2.3.1 Problem setting

In this paper, the following conditions are assumed in our segmentation problem.

- A scene is captured by a static calibrated binocular camera system.
- The background of the scene is modeled as a pixel-wise Gaussian distribution.
- An object in the foreground stands on the ground plane and its shadow appears on the ground plane.

Our goal is to segment the target region as *foreground* (*"F"*) or *shadow* (*"S"*) or *background* (*"B"*), that is to say, to assign one of the three labels *"F"*, *"S"*, or *"B"* to each pixel in both side images in a conformal manner. Note that *"S"* and *"B"* lie on the ground plane while *"F"* stands on the ground plane. Note that shadowed foreground is regarded as foreground because a primal aim of this work is foreground extraction.

2.3.2 Asymmetric treatment of homography constraint

Let us consider the homography-correspondence pair on the ground plane in the binocular camera. According to the homography constraint, if a pixel belongs to the ground plane on one side image, the color of the pixel is strictly consistent with that of the homography-correspondence pixel in the other side image under the condition that ideally any standing object does not exist

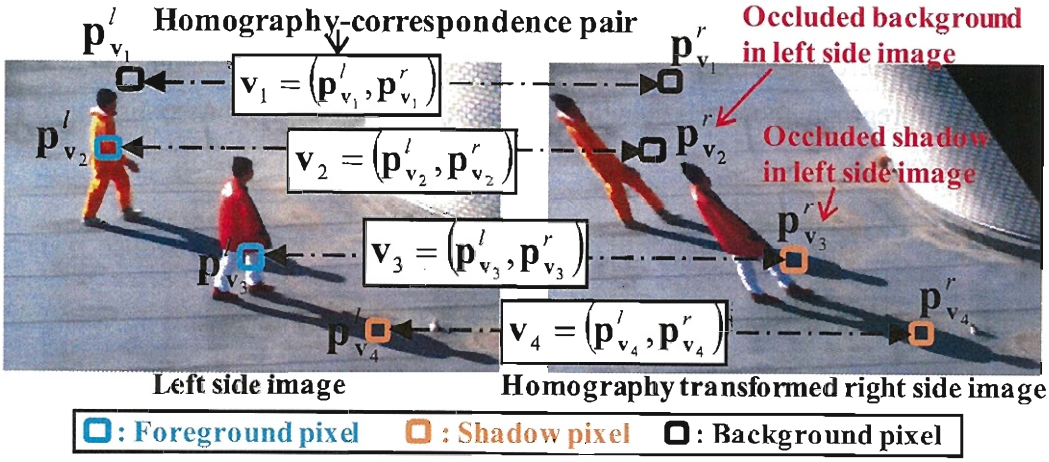


Figure 2.1: Homography-correspondence pair

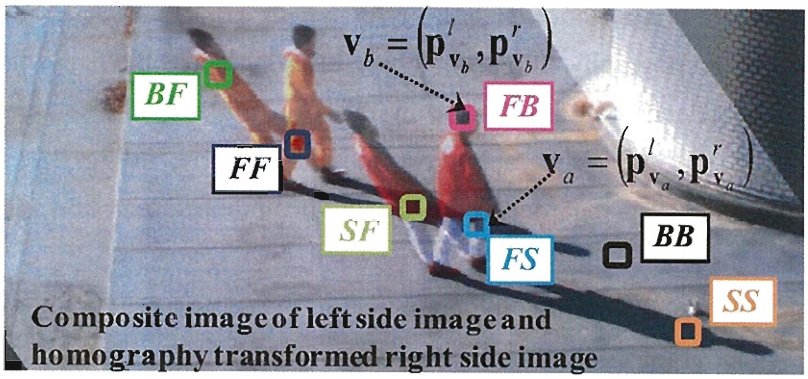


Figure 2.2: Labeling examples

on the ground plane. This is a very useful property to distinguish the standing objects on the ground plane from the ground plane objects.

In segmentation problems, this property is also useful when assigning a label to each pixel. Some examples of the homography-correspondence pairs are shown in Fig. 2.1. First, suppose that the left side image is a base image to be segmented. Because $\mathbf{v}_1$ and $\mathbf{v}_4$ have similar colors between each correspondence pixel, $\mathbf{p}^l_{\mathbf{v}_1}$ and $\mathbf{p}^l_{\mathbf{v}_4}$ are labeled as *"S"* or *"B"* in the left side image. On the other hand, because the pixel pairs $\mathbf{v}_2$ and $\mathbf{v}_3$ have different colors between each pair of correspondence pixels, $\mathbf{p}^l_{\mathbf{v}_2}$ and $\mathbf{p}^l_{\mathbf{v}_3}$ are labeled as *"F"* in the left side image. Next, supposed that the right side image is a base image to be segmented in turn, pixel pairs $\mathbf{p}^r_{\mathbf{v}_1}$ and $\mathbf{p}^r_{\mathbf{v}_4}$ are labeled as *"S"* or *"B"*, and the pixel pairs $\mathbf{p}^r_{\mathbf{v}_2}$ and $\mathbf{p}^r_{\mathbf{v}_3}$ are labeled as *"F"* in the right side image in the same way. The true labels of $\mathbf{p}^r_{\mathbf{v}_2}$ and $\mathbf{p}^r_{\mathbf{v}_3}$ are, however, not *"F"* but *"B"* and *"S"*.

This mislabeling often arises in cases where a pixel belongs to the ground plane in one side image and where the corresponding pixel's ground plane point in the other side image is occluded by a foreground object as shown in this example. Therefore, the existing asymmetric homography-based approaches suffer from the mislabeling due to occlusion.

2.3.3 Symmetric approach based homography-correspondence pair

In our framework, the homography-correspondence is treated symmetrically to cope with the occluded regions and to segment them correctly.

Taking the occlusion relationship into consideration, the labeling strategy is as follows. If the pixels are labeled *"S"* or *"B"* in one side image, their homography-correspondence pixels in the other side image are given either the same label (not the occluded case) or *"F"* (the occluded case). If the pixels in one side image are labeled *"F"*, their homography-correspondence pixels in the other side image are possibly labeled *"F"*, *"S"*, or *"B"*, because the standing object is not constrained by homography. From this observation, the possible pair-wise label for the homography-correspondence pair are defined in Tab. 2.1. In this label set, for example, the pair-wise label *"FS"* (e.g., the homography-correspondence pair $\mathbf{v}_a$ in Fig. 2.2) represents that the left side pixel of the pair is regarded as the shadow-occluding foreground and the right side pixel is regarded as the shadow occluded by foreground.

Note that the label set is not mere the combination of possible labels in each side image, we introduce the homography constraint in the form of prohibiting the pair-wise labels, *"SB"* and *"BS"*. These labels are never occurred because it is not possible for a ground plane object to occlude another ground plane object.

Let us consider the advantage of the prohibition by taking for example the homography-correspondence pair $\mathbf{v}_b$ in Fig. 2.2. This pair is composed of $\mathbf{p}^l_{\mathbf{v}_b}$ and $\mathbf{p}^r_{\mathbf{v}_b}$, and the color of $\mathbf{p}^l_{\mathbf{v}_b}$ is very similar with that of shadow, and the color of $\mathbf{p}^r_{\mathbf{v}_b}$ is almost the same with that of background. Therefore, the label *"S"* is possibly assigned to $\mathbf{p}^l_{\mathbf{v}_b}$ and the label *"B"* might be assigned to $\mathbf{p}^r_{\mathbf{v}_b}$ when color-based labeling is applied for each side image, though the correct label of $\mathbf{p}^l_{\mathbf{v}_b}$ is *"F"*. On the other hand, due to the prohibition of the label *"SB"*, the proposed labeling scheme could assign the correct label *"FB"* to $\mathbf{v}_b$.

Thus our segmentation problem is regarded as a multi-labeling problem for homography-correspondence pair pixels, and the multi-labeling results provide all the relationships between homography-correspondence pair of pixels. For example, the label *"FS"* means the foreground occludes the shadow in the left side image, and also means the shadow in the right side image is occluded by the foreground in the left side image. Example of pair-wise labeling are shown in Fig. 2.2.

Table 2.1: The pair-wise label sets for a homography-correspondence pair

Left-side label	Right-side label		
	F	*S*	*B*
F	*FF*	*FS*	*FB*
S	*SF*	*SS*	–(prohibited)
B	*BF*	–(prohibited)	*BB*

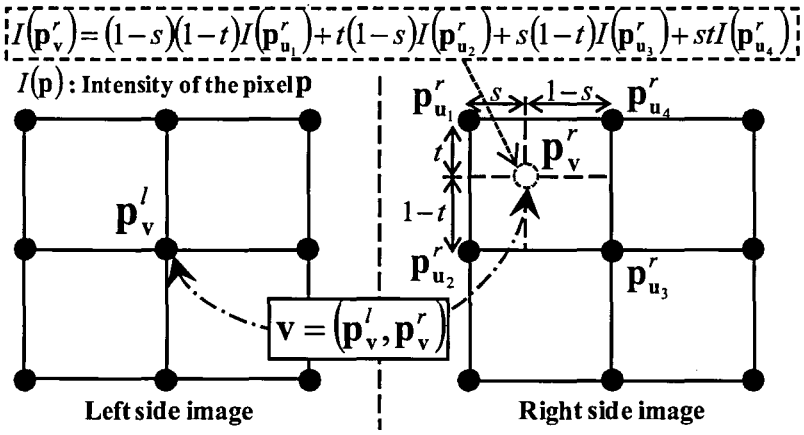


Figure 2.3: Homography-correspondence detail

2.3.4 Problem formulation

We formulate the pair-wise multi-labeling problem in a framework that minimizes energy. Let us define the site $\mathbf{v}=(\mathbf{p}_v^l,\mathbf{p}_v^r)$ which represents a homography-correspondence pair as described in the previous subsection. Then, the label set is defined as,

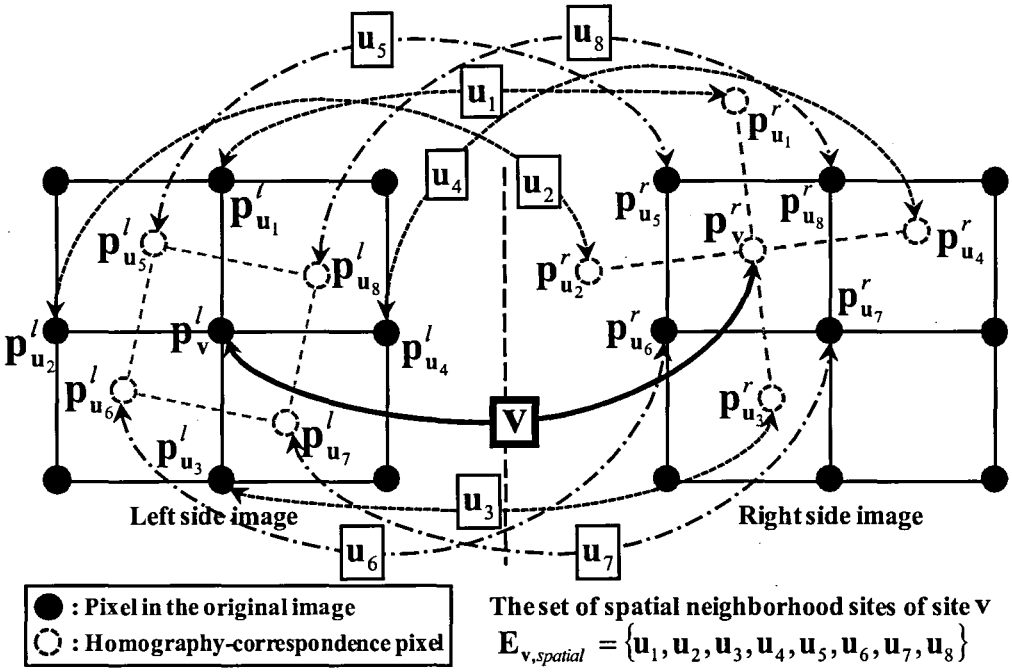
$$\mathbf{L}=\{FF,FS,FB,SF,SS,BF,BB\}, \tag{2.1}$$

and the label assigned to a site $\mathbf{v}$ as $\mathbf{x_v}\in\mathbf{L}$. Then our goal is to assign each site $\mathbf{v}$ a label $\mathbf{x_v}$ from the set $\mathbf{L}$. Generally, this problem is formulated in an energy minimization framework as follows,

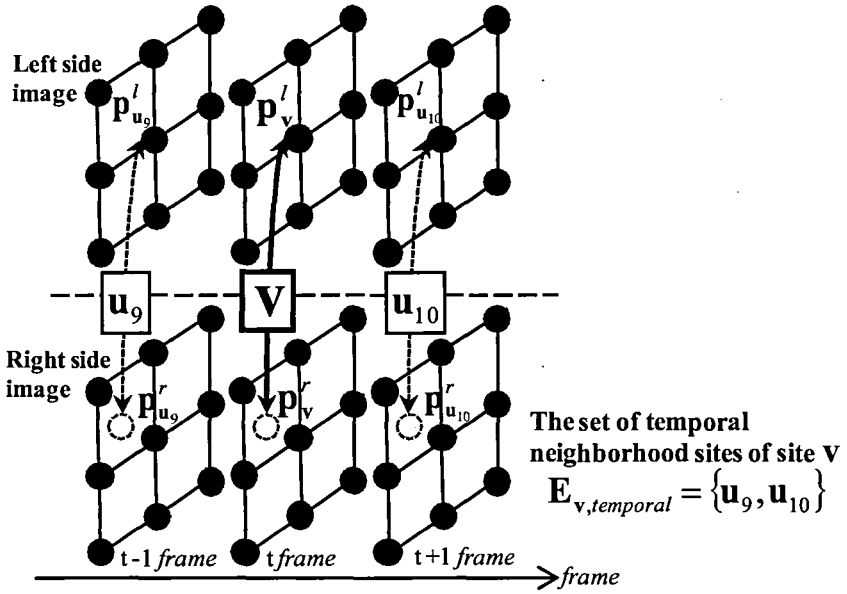
$$E(\mathbf{x})=w_g\sum_{\mathbf{v}\in\mathbf{V}}\mathbf{g}(\mathbf{x_v})+w_h\sum_{(\mathbf{u},\mathbf{v})\in\mathbf{E}}\mathbf{h}(\mathbf{x_u},\mathbf{x_v}) \tag{2.2}$$

where the first and the second terms are data and smoothness terms, w_g and w_h are the weights of each term, $\mathbf{x}$ is a configuration (label combination), $\mathbf{V}$ is a set of all sites, and $\mathbf{E}$ is all the combinations of the neighborhood sites. This energy function is minimized via graph-cut algorithms such as the α-expansion or α-β swap algorithms [66].

Note that, the homography-correspondence positions are calculated using sub-pixel order and the color of the sub-pixel position is spatially interpolated by their 4-neighborhood pixels



(a)Spatial neighborhood



(b) Temporal neighborhood

Figure 2.4: Spatio-temporal neighborhood sites for smoothness term

as shown in Fig. 2.3. In addition, as shown in Fig. 2.4, we consider 10-neighborhood sites in a spatio-temporal 3D domain composed of spatial 8-neighborhood, and temporal 2-neighborhood sites.

2.4 Implementation

2.4.1 Seed generation

Given background subtraction regions as potential regions of shadow and foreground, the foreground seed is provided as the union of the following two regions; one is the intersection of the potential region and background region projected by homography from the other image, and the other is the region which has a largely different color direction from the background one. Then, the shadow seed is decided based on homography consistency and color-based shadow likelihood (see Chapter 2.4.2 for detail).

2.4.2 Data term

The data term is defined by the log of the likelihood as,

$$\mathbf{g}(\mathbf{x_v}) = -\log\Big(P(\mathbf{x_v}|\mathbf{c}(\mathbf{v}))\Big) = -\log\left(\frac{P(\mathbf{c}(\mathbf{v})|\mathbf{x_v})P(\mathbf{x_v})}{\sum_{l_i\in\mathbf{L}} P(\mathbf{c}(\mathbf{v})|\mathbf{x_v}=l_i)P(\mathbf{x_v}=l_i)}\right), \tag{2.3}$$

where $P()$ is probability and $\mathbf{c}(\mathbf{v})$ is a six dimensional color vector at site $\mathbf{v}$ composed of a pair of RGB vectors in each image as where $P()$ is probability and $\mathbf{c}(\mathbf{v})$ is a six dimensional color vector at site $\mathbf{v}$ composed of a pair of RGB vectors in each image as $\mathbf{c}(\mathbf{v}) = [\mathbf{c}(\mathbf{p}_\mathbf{v}^l), \mathbf{c}(\mathbf{p}_\mathbf{v}^r)]^T$, and $\mathbf{c}(\mathbf{p})$ is color vector at pixel $\mathbf{p}$. Then the pair-wise color observation model $P(\mathbf{c}(\mathbf{v})|\mathbf{x_v})$ is decomposed into $\prod_i P(\mathbf{c}(\mathbf{p}_\mathbf{v}^i)|\mathbf{x}_\mathbf{v}^i)$, where $\mathbf{x}_\mathbf{v}^i$ is the one side label and i $(i = l, r)$ is the camera identifier.

Foreground model

The foreground color is approximated by a pixel-wise GMM which is trained by k-means clustering from *foreground seed* pixels, and the foreground observation model is expressed as,

$$P(\mathbf{c}(\mathbf{p}_\mathbf{v}^i)|\mathbf{x}_\mathbf{v}^i = F) = \mathcal{N}(\mathbf{c}_f^{k^*}, \Sigma_f^{k^*}) \tag{2.4}$$

$$k^* = \arg\min_k \left(\left(\mathbf{c}(\mathbf{p}_\mathbf{v}^i) - \mathbf{c}_f^k\right)^T \Sigma_f^{k\,-1} \left(\mathbf{c}(\mathbf{p}_\mathbf{v}^i) - \mathbf{c}_f^k\right) \right), \tag{2.5}$$

where $\mathbf{c}_f^k$ and Σ_f^k are a mean vector and a covariance matrix of the kth cluster, and $\mathcal{N}$ is the Gaussian distribution.

Shadow-Background model

First, a linear color transformation matrix from the background color to the shadow color is estimated from the *shadow seed* colors and their modeled background colors. This matrix is modeled as following a finite-dimensional linear model [88],

$$\mathbf{c}_s(\mathbf{p}) = \mathbf{A}\tilde{\mathbf{c}}_{bg}(\mathbf{p}), \tag{2.6}$$

where $\mathbf{c}_s$ is a color vector of a shadow seed, $\tilde{\mathbf{c}}_{bg}$ is an extended color vector of a modeled background, $\tilde{\mathbf{c}}_{bg} = [\mathbf{c}_{bg}{}^T, 1]$, and $\mathbf{A}$ is a 3 by 4 shadow transformation matrix. Then, the color transformation matrix $\mathbf{A}$ is obtained by minimizing the following objective function S,

$$\mathbf{e}(\mathbf{p}) = \mathbf{A}\tilde{\mathbf{c}}_{bg}(\mathbf{p}) - \mathbf{c}_s(\mathbf{p}) \tag{2.7}$$

$$S = \sum_{\mathbf{p}\in\mathbf{P}_s} \mathbf{e}(\mathbf{p})^T \left(\Sigma_{bg}(\mathbf{p})\right)^{-1} \mathbf{e}(\mathbf{p}), \tag{2.8}$$

where $\mathbf{e}$ and Σ_{bg} are the color transformation error vector and covariance matrix of the modeled background color, and $\mathbf{P}_s$ is a set of shadow seed pixels.

Next we define the vector $\mathbf{c}_r$ which is the nearest color to an input color $\mathbf{c}$ on the line segment between the modeled background color $\mathbf{c}_{bg}$ and the estimated shadow color $\hat{\mathbf{c}}_s = \mathbf{A}\tilde{\mathbf{c}}_{bg}$ in RGB color space as shown in Fig. 2.5. Then, the vector $\mathbf{c}_r$ is expressed as

$$\mathbf{c}_r(\mathbf{p}_\mathbf{v}^i) = \hat{t}\hat{\mathbf{c}}_s(\mathbf{p}_\mathbf{v}^i) + (1-\hat{t})\mathbf{c}_{bg}(\mathbf{p}_\mathbf{v}^i) \tag{2.9}$$

$$t = \frac{\left(\hat{\mathbf{c}}_s(\mathbf{p}_\mathbf{v}^i) - \mathbf{c}_{bg}(\mathbf{p}_\mathbf{v}^i)\right)^T \left(\mathbf{c}(\mathbf{p}_\mathbf{v}^i) - \mathbf{c}_{bg}(\mathbf{p}_\mathbf{v}^i)\right)}{\| \hat{\mathbf{c}}_s(\mathbf{p}_\mathbf{v}^i) - \mathbf{c}_{bg}(\mathbf{p}_\mathbf{v}^i) \|} \tag{2.10}$$

$$\hat{t} = \min\{1, \max\{t, 0\}\}, \tag{2.11}$$

Finally the background and shadow observation models are introduced based on the interpolation on the line segment as,

$$P\left(\mathbf{c}(\mathbf{p}_\mathbf{v}^i) \,|\, \mathbf{x}_\mathbf{v}^i = S\right) = \hat{t}\,\mathcal{N}\left(\mathbf{c}_r(\mathbf{p}_\mathbf{v}^i), \Sigma_r{}'\right) \tag{2.12}$$

$$P\left(\mathbf{c}(\mathbf{p}_\mathbf{v}^i) \,|\, \mathbf{x}_\mathbf{v}^i = B\right) = (1-\hat{t})\,\mathcal{N}\left(\mathbf{c}_r(\mathbf{p}_\mathbf{v}^i), \Sigma_r{}'\right) \tag{2.13}$$

$$\Sigma_r'(\mathbf{p}_\mathbf{v}^i) = \Sigma_r(\mathbf{p}_\mathbf{v}^i) + \Sigma_e(\mathbf{p}_\mathbf{v}^i), \tag{2.14}$$

where Σ_r and Σ_e are covariance matrices of the reference color $\mathbf{c}_r$ and color transformation error $\mathbf{e}$.

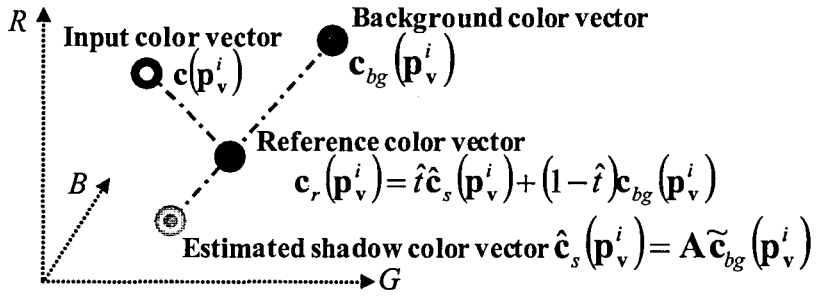


Figure 2.5: Shadow-background model

2.4.3 Smoothness term

The smoothness term considering intensity value normalization is defined as,

$$\mathbf{h}(\mathbf{x}_v, \mathbf{x}_u) = \begin{cases} \exp\left(-\kappa d_e^l(\mathbf{x}_v, \mathbf{x}_u)\right) & \text{Left side label is different} \\ \exp\left(-\kappa d_e^r(\mathbf{x}_v, \mathbf{x}_u)\right) & \text{Right side label is different} \\ \exp\left(-\kappa\sqrt{d_e^l(\mathbf{x}_v, \mathbf{x}_u) d_e^r(\mathbf{x}_v, \mathbf{x}_u)}\right) & \text{Both side labels are different} \\ 0 & \text{Otherwise} \end{cases}, \tag{2.15}$$

where d_e^i is an edge intensity criteria given by,

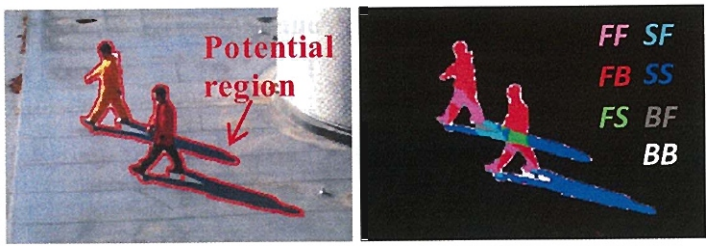
$$d_e^i(\mathbf{x}_v, \mathbf{x}_u) = \frac{1}{D_{\mathbf{p}_u^i \mathbf{p}_v^i}} \left(\frac{\| \mathbf{c}(\mathbf{p}_v^i) - \mathbf{c}(\mathbf{p}_u^i) \|^2}{\| \mathbf{c}(\mathbf{p}_v^i) + \mathbf{c}(\mathbf{p}_u^i) \|^2 + \varepsilon} \right), \tag{2.16}$$

where $D_{\mathbf{p}_u^i \mathbf{p}_v^i}$ is the pixel distance between $\mathbf{p}_u^i$ and $\mathbf{p}_v^i$ (as for the temporal distance, all the distances are set to 1). Also, κ and ε are coefficients for this term.

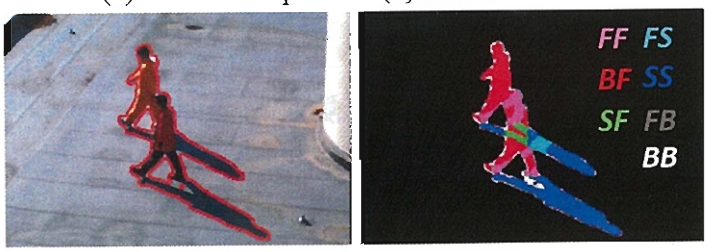
2.5 Experiments

2.5.1 Data set and parameters

We carried out experiments using sequences of people walking outdoors. Tab. 2.2 shows the details of the data set. Every sequence contains some men or women with strong shadows. Of these, *Seq A* (Fig. 2.6 (a) and (b)) is captured at our university as test data, and *Seq B* (Fig. 2.7 (a) and (b)) and *Seq C* (Fig. 2.8 (a) and (b)) are extracted from the videos of surveillance cameras deployed at an elementary school in Japan. The background data for each dataset is generated from the other sequences captured at a different time. A total of 3 images were provided for graph-cut segmentation in a block. Note that in some figures in this section, the

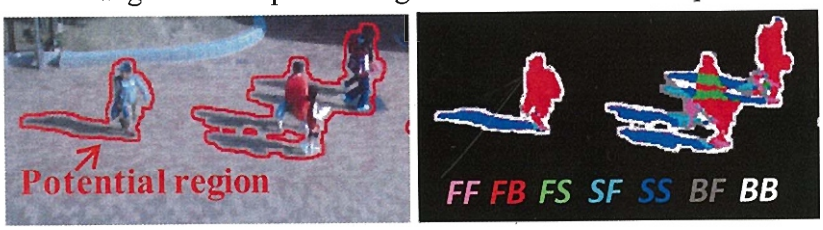


(a) Left side input (b) Left side result



(c) Right side input (d) Right side result

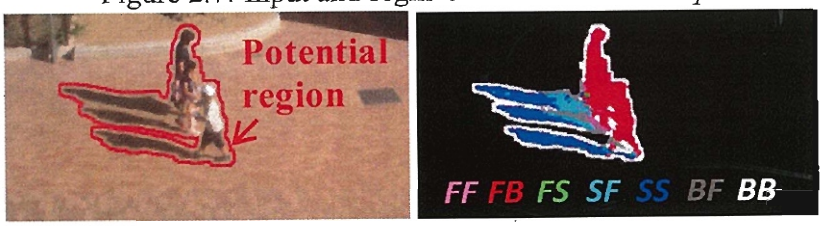
Figure 2.6: Input and segmentation results of *SeqA*



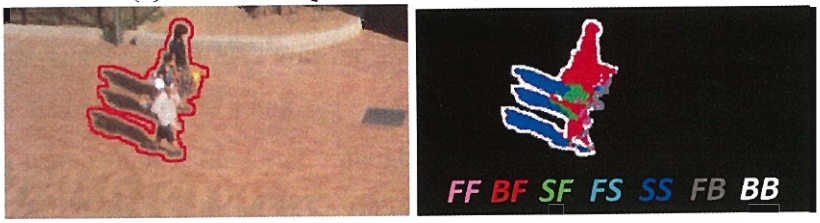
(a) Left side input (b) Left side result

(c) Right side input (d) Right side result

Figure 2.7: Input and segmentation results of *SeqB*



(a) Left side input (b) Left side result



(c) Right side input (d) Right side result

Figure 2.8: Input and segmentation results of *SeqC*

Table 2.2: Data set for experiments

Sequence set	Image size	Image number	Frame rate
SeqA	640×480	32	30 fps
SeqB	620×280	12	9 fps
SeqC	620×280	24	9 fps

results of the experimental images are trimmed around the segmentation target region because page space is limited.

In these experiments, the data terms were spatially smoothed in response to the magnitude of the edge pixels. Because the pixel color is quite variable, and it is unstable near the edge, the reliability of the data terms is very low for such pixels. The segmentation process was done iteratively, and there were 2 iterations. The parameters of the proposed method were experimentally set at $w_g = 3.0$, $w_h = 0.2$, $\kappa = 4.0$, and $\varepsilon = 10^{-7}$. Initially the prior of each label is set as follows: $P(FB)=P(BF)=P(SS)=0.16, P(FS)=P(SF)=P(FF)=0.14, P(BB)=0.1$. In addition, the distribution number of GMM was set at 6 for *SeqA* and at 10 for *SeqB* and *SeqC*. We adopted the α-β swap algorithm [66] to minimize our energy function Eq. (2.2).

2.5.2 Benchmark

We compared the segmentation performance of the proposed method with eight approaches: three existing approaches [54], [69] (color-based method), and [64] (homography-based method), five energy minimization-based approaches (as described later). While a labeling problem for each homography-correspondence pair is considered in the proposed method, other methods take a labeling problem for each pixel in each side image in consideration, and therefore, the labeling was independently-executed in each side image in them. Besides, there are two kinds of methods in comparative methods: one considers three labels of foreground, shadow, and background, and the other considers two labels of foreground and shadow. As for the former, first, we dilated the regions extracted via background subtraction and set the regions as potential region. Then, we regarded the segmentation problem as the multi-labeling problem of foreground, shadow, and background labels for the potential region. As for the latter, we considered the problem as the binary labeling problem where either of foreground and shadow labels is assigned to a pixel in the regions extracted by background subtraction. For fair comparison, we tuned the parameters of each comparative method so that the method achieved best performance in total.

Each of the following five comparative approaches is the energy minimization-based framework where an energy function composed of a data term and a smoothness term based on edge magnitude (Eq. (2.16)) is minimized via graph-cut.

Color: the color-based method where the labels of *"F"*, *"S"*, and *"B"* are considered. The *foreground seed* and *shadow seed* are generated based on the shadow color properties [73, 75]. Then, a foreground color model (see Section 2.4.2) and the shadow-background models (see 2.4.2) are constructed from the *foreground seed* and *shadow seed*, respectively. Finally, we label each pixel in one side image as *"F"*, *"S"*, or *"B"*.

Disparity: the disparity-based method where the labels of *"F"* and *"S"* are considered. The *foreground seed* and *shadow seed* are generated by thresholding disparity. Also, the data term is defined based on disparity. The method in [89] is used for disparity calculation.

Color + Disparity: the integrated method of *Disparity* with *Color*, where the labels of *"F"*, *"S"*, and *"B"* are considered. The data term is defined as a weighted sum of the data term of *Color* and that of *Disparity*.

Homography (asymmetric): the homography-based method where the labels of *"F"* and *"S"* are considered. The *foreground seed* and *shadow seed* are generated by homography constraint, and also, the data term is defined by the color similarity between each homography-correspondence pair.

Color + Homography (asymmetric): the integrated method of *Homography (asymmetric)* with *Color*, where the labels of *"F"*, *"S"*, and *"B"* are considered. The data term is defined as a weighted sum of the data term of *Color* and that of *Homography (asymmetric)*.

2.5.3 Results

First, the multi-labeling results of the proposed method for each data set are shown in Fig. 2.6, Fig. 2.7, and Fig. 2.8. In each result, the labeling results are good even for the occlusions.

Second, the quantitative performance comparisons are shown in Tab. 2.3. The performance of each method is evaluated by *F-measure*, which is defined as,

$$F = \frac{2PR}{P+R} \tag{2.17}$$

$$P = \frac{N_{tp}}{N_{tp}+N_{fp}} \tag{2.18}$$

$$R = \frac{N_{tp}}{N_t} \tag{2.19}$$

where F is *F-measure*, P and R are *precision* and *recall*, N_{tp} and N_{fp} are the number of true positive pixels and that of false positive pixels, and N_t is the number of ground truth pixels.

Table 2.3: Quantitative evaluation results

Method	*SeqA*		*SeqB*		*SeqC*	
	F	*S*	*F*	*S*	*F*	*B*
Horprasert et al. [54]	0.870	0.837	0.816	0.516	0.627	0.653
Sun et al. [69]	0.927	0.851	0.817	0.696	0.659	0.523
Jeong et al. [64]	0.897	0.897	0.878	0.791	0.868	0.764
Color	0.890	0.863	0.896	0.833	0.817	0.776
Disparity	0.889	0.856	0.758	0.643	0.822	0.708
Color + Disparity	0.932	**0.923**	0.906	0.843	0.897	0.857
Homography (asymmetric)	0.893	0.872	0.872	0.779	0.877	0.770
Color + Homography (asymmetric)	0.938	**0.923**	**0.921**	0.854	0.874	0.824
Proposed method	**0.940**	0.902	0.920	**0.874**	**0.900**	**0.865**

F: Foreground, *S*: Shadow

In the tables, we see that the *Color + Disparity*, *Color + Homography (asymmetric)* and the proposed method show better results than other methods, especially, the proposed method achieves the best performance of all in total.

The results of these three methods for the left side inputs shown in Fig. 2.7 and 2.8 are shown in Fig. 2.9. As for the results for *Seq B*, the mis-labeled foreground region as shadow near the head of right side person in the result of *Color + Disparity* is relatively larger than those of the other two methods. This may be because that the disparity is not calculated correctly (calculated disparity is too small) and the region has color like shadow. As for the results for *Seq C*, we can see the occlusion problem in the result of *Color + Homography (asymmetric)* as described in Section 2.3. Although this mis-labeled region can be recovered to some extent when the weight of a data term about *Homography (asymmetric)* is set much smaller than that about *Color*, the problem of color-based method is no longer ignore in such weight setting. This trade-off problem is inevitable as long as a simple combination scheme of color and homography is used. Note that the trade-off is considered and the weight setting is optimized in this experiment as mentioned earlier. On the other hand, the results of the proposed method are relatively better than the other two methods for both datasets.

2.6 Discussions

Effective use of extracted shadow

By making effective use of extracted shadow, our approach can obtain consistent labeling as well as information as to whether the occluded region belongs to the *shadow* or *background*.

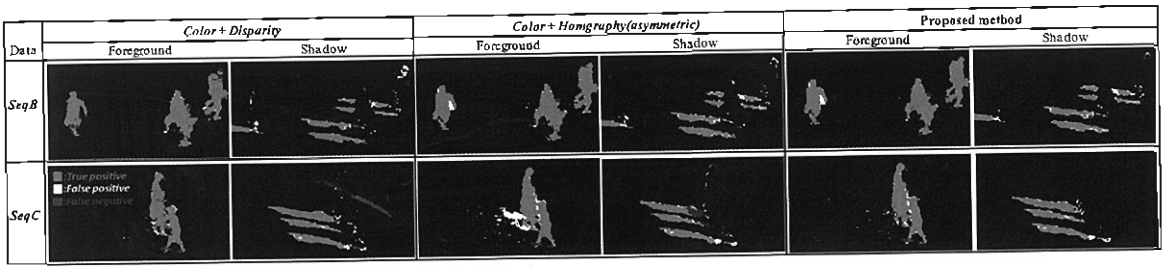
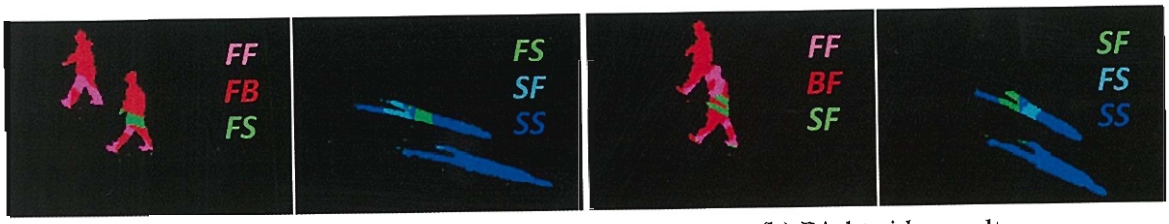


Figure 2.9: Comparison Results



(a) Left side result (b) Right side result

Figure 2.10: Extracted foreground and a whole shadow including occluded shadow for *SeqA*

This means that we can get additional scene information. For example, because a whole shadow silhouette including the occluded shadow, can be seen as another projection from the viewpoint of a light source, we can say that one more different-view of the whole silhouette of the target foreground objects is extracted as shown in Fig. 2.10. This is quite valuable not only for gait-based person identification, but also for many other computer vision applications, especially silhouette based applications, like gesture recognition and 3D reconstruction by shape from silhouettes and so on. As for gait-based person identification, it is reported in [90] that the different views of silhouettes improve identification performance, and more, shadow gait-based person identification scheme is proposed in [91].

In addition, homography-based object localization techniques have been proposed [84], where the position of the object is localized by estimating the intersection point of the object region and the shadow region. Hence, if the occluded shadow region is also extracted by the proposed method, the object localization accuracy is improved.

Extension to more complex scene or moving platform

Although the assumption that the shadow appears on the ground plane may seem to be a heavy constraint, our method can be extended to more complex scenes by modeling scenes as piece-wise facets and by calibrating the homography for each facet.

Furthermore, our method can be applied to a mobile platform such as a vehicle binocular video system, and an intelligent robot with a combination of state of the art dynamic back-

ground modeling, ego-motion, and image stabilizing techniques. For example, we can acquire a background model for each frame of the image sequence by using dynamic background modeling, and we can calibrate the geometric relationship between the binocular camera system and the target plane by using ego-motion and image stabilizing techniques.

2.7 Conclusions

In this chapter, we propose a homography-correspondence pair based segmentation framework. We treat homography-correspondence pairs symmetrically, and formulate the segmentation problem as a multi-labeling problem for a homography-correspondence pair to explicitly take the occlusion relationship into account. Then we obtain the segmentation result by minimizing the energy function via the α-β swap algorithm. In our experiments, it turns out that the segmentation results of the proposed method outperform the existing color-based and asymmetric homography-based methods.

Chapter 3

Group Context-aware Person Identification in Video Sequences

3.1 Introduction

Many gait-based person identification techniques have been developed to date, mostly from the viewpoints of discrimination capability and stability [3, 26]. In all of these techniques, however, the identification performance often decreases due to changes in the walking condition of individuals (e.g., clothing and carrying conditions) and their surroundings (e.g., surface and observation view of camera), and the identification performance may consequently decrease [33], particularly in real environments. Also, as the number of individuals increases, the misidentification rate generally increases due to the growth in ambiguity.

An example of misidentification in a straightforward gait-based person identification framework is shown in Fig. 3.1. Because the gait feature of probe #1 has changed slightly from that in gallery #a (the same subject), particularly in the arm swing, the feature similarities between probe #1 and gallery #a are smaller than those between the probe and other galleries (e.g., #x and #y).

However, it is useful to take into account the characteristics of human activities to provide context for person identification. In social living situations, people often act in groups, as shown in Fig. 3.2, which are composed using social relationships in most cases, such as family, friends, and co-workers. It is assumed, therefore, that a person is likely to be observed close to other persons of the same group in a video sequence. This observation serves as a contextual cue to improve the identification performance for individuals, i.e., the identity of each person can be inferred not only from their biometric cues alone, but also from the identities of other people in their neighborhood and their group affiliations.

This kind of group context can be used in many places, such as amusement or theme parks, airports, factories, and schools, where many tasks based on person identification techniques are

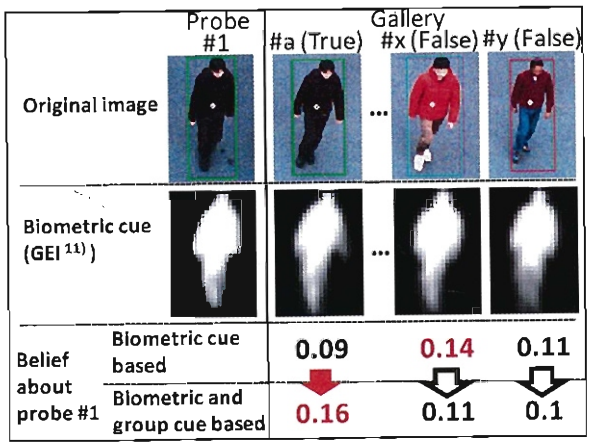


Figure 3.1: Biometric cues and belief.

performed. Examples of these tasks include the detection of a lost child in an amusement park, the detection of intruders who enter the amusement park, airport, or factory without passing regular entrance procedures, and the safety confirmation (or attendance checking) of children at the entrance to the school (in particular, there is a rule for going to school in a group composed of community children for almost all Japanese elementary schools). The group context is also useful for person re-identification across multiple non-overlapping network cameras.

Recently, some works have integrated such kind of group context with face-based person identification in photo collection to improve the identification performance [92, 93, 94]. In these methods, the person-to-person relations are modeled in terms of co-occurrence among persons in photos as group prior. Differently from the photo collection, however, a group is often observed with non-group members at a time in the video sequences of surveillance camera and the spatial relations among them are dynamically changed with time. Therefore, the identity of individual should be inferred not only from the viewpoint of co-occurrence among persons, but also from that of behavioral differences among persons through the sequence.

In this research, we propose a group context-aware framework for person identification in video sequences that unifies the group context with the individual biometric cues. In terms of the group (inter-person) context for person identification, the proposed method take the behavioral differences such as spatial distance and the differences of walking speed and direction among persons through the sequence into account, and this is a primal contribution of this work.

Our key observation is as follows. We assume the group walking situation in a video sequence which includes two different groups and an unregistered person as shown in Fig. 3.2

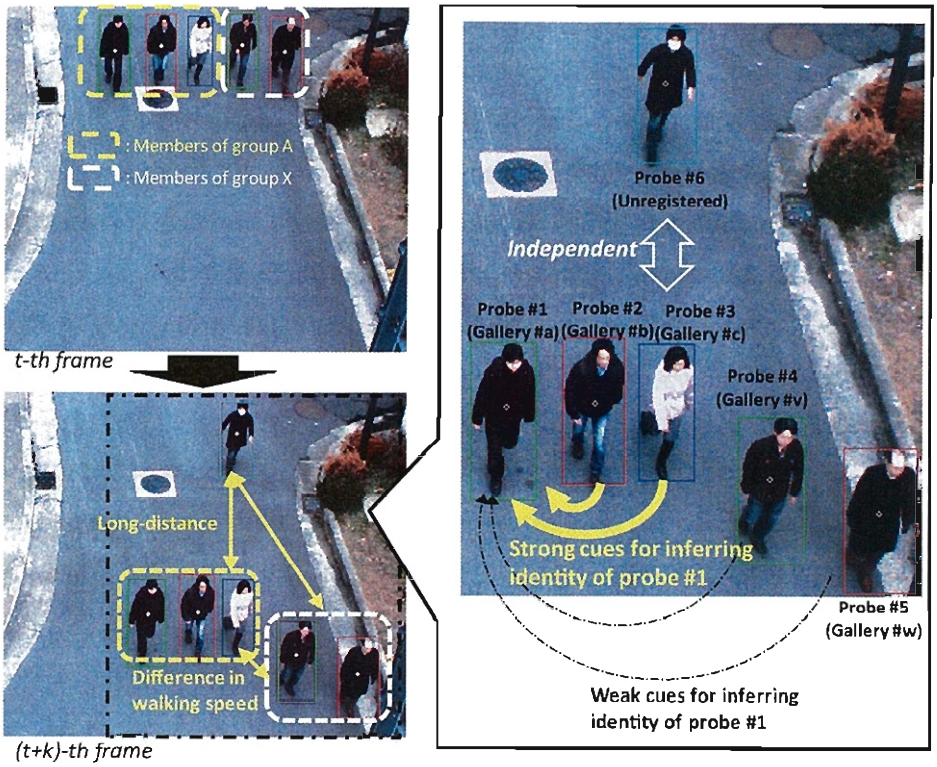


Figure 3.2: An example of group walking in video sequence.

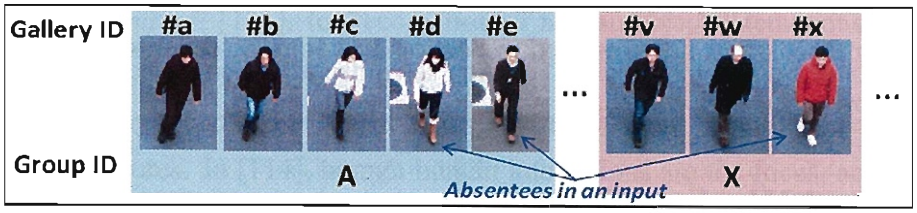


Figure 3.3: An example of group affiliation.

and consider the identity of the probe #1 within the group context. We assume that the gallery subjects #a, #b, #c, #d, and #e belong to the group A and the #v, #w, and #x belong to the other group X, as shown in Fig. 3.3. Also, we assume that the identity of probe #6 is not matched with any of gallery subjects and probe #2, probe #3, probe #4, and probe #5 are confidently inferred to be #b, #c, #v, and #w, respectively, while the identity of probe #1 is mis-inferred to be #x (a member of group X) as shown in Fig. 3.1. If only co-occurrence is used as group context, the identity of probe #1 can be inferred from not only the identities of probes #2 and #3 as #a, but also those of probes #4 and #5 as #x based on their group affiliation information. Consequently, the identity of probe #1 possibly remains to be mis-inferred as #x in this case. In addition, the identity of probe #6 (unregistered person) which appears in the scene from the

middle of the sequence is possibly mis-inferred from the identities of other subjects.

On the other hand, focusing on the behavioral relations among probe subjects through a sequence, we see that probe #6 obviously walks at a distant from all the other subjects, and thus probe #6 can be regarded as an independent subject from all the other subjects. At the same time, we also see that there exists an apparent difference of walking speed between the group of #1, #2 and #3 and the group of #4 and #5. Accordingly, the weights of the inference cues from the probes #2 and #3 come to be able to be distinguished from those of the inference cues from the probes #4 and #5. The identity of probe #1 as #a is then definitely enhanced and the mis-identification of #1 can be recovered as a result.

We realize this idea in the form of a message passing in a graph, where each node corresponds to each probe subject and each edge corresponds to the relationship between each pair of probe subjects. In the iteration of the message passing process, the identity confidence for each probe subject is propagated to the identities of the surrounding probe subjects based on their biometric cues and group information, so that the same group members with similar characteristics (spatial proximity and similar velocity vector) enhance each other's identities.

The remainder of this paper is organized as follows. Section 3.2 introduces related work. Section 3.3 describes our problem formulation, and the detailed implementation is described in Section 3.4. Section 3.5 presents experimental testing of the effectiveness of the proposed method and our discussions are presented in Section 3.6. Finally, conclusions are drawn and future work is proposed in Section 3.7.

3.2 Related Work

In recent years, many researchers have paid considerable attention to the use of context in traditional computer vision problems, such as object detection and categorization, action recognition, and person identification, to improve performance. In this section, we review such context-based approaches briefly.

Object detection/recognition: In the task of object detection, context is mainly used to limit the area in which objects are likely to appear, to reduce false positives. Torralba et al. [95] exploited a global image feature called *gist*, which was a low level representation of an image. Hoiem et al. [96] used the 3D geometrical information of the scene, such as the surfaces, the camera viewpoint, and object positions and sizes as context. While these approaches focused on global scene information, some works instead focused on local information [97, 98]. In [97], the spatial relations between an object of interest and its surroundings are modeled as

a visual context feature composed of geometrical and textural features, and are used to extract prior instances of the object's presence from a scene. In this method, object co-occurrence and bottom-up saliency were also used for context. Heitz and Koller [99] modeled the spatial relationships between an object ("thing") and the surrounding regions ("stuff"), which were the results of unsupervised image clustering, as the *TAS model* ("thing" and "stuff" model). The effect of the use of context in object detection is empirically evaluated in [100].

In recent works in object recognition [101, 102, 103], inter-object relationships, such as co-occurrence, relative location, and scale, are used as context to resolve object appearance ambiguities. Besides those given above, a number of context-based techniques have been discussed and summarized in [104], [105], and [106].

Action/Interaction recognition: Many of these works have indicated that modeling of human-object relationships is useful for the understanding of human actions/interactions [107, 108, 109, 110, 111]. Wu et al. [107] proposed an object-use based action recognition framework, in which the relationships between an action and the object-use events data during that action were used as context, and the relationships were learned automatically using RFID sensors and a common-sense knowledge database. Yao and Fei-Fei [110, 111] proposed two types of approach; one is based on a model of the spatial relationships between human poses (positions of body parts) and objects [110], while the other is based on a structured appearance feature called "*Grouplet*" [111], for recognition of human-object interactions. Marszalek et al. [112] used action-scene relationships as context, which were derived automatically from training videos using video scripts, and in [113], both scene and object features are integrated with the action features. In [114], human-human interactions are the focus, and it was shown that spatio-temporal observations of the surrounding people which represent the actions of the surroundings helped with action recognition. In a similar manner, Choi et al. [115] used the spatio-temporal distribution of multiple people, which included their relative motion and locations, to classify collective activities, such as "queueing" and "talking".

Person identification: The automatic annotation, organization, and retrieval of still images, in particular in personal digital photo collections, have been active research topics in recent years. In these tasks, face-based person identification is crucially important and many context-aware methods have been developed. As mentioned in [116], there are three types of context information: appearance-based, metadata-based, and logic-based context information. In [117] and [116], appearance-based context, such as body parts and clothes, are combined with facial features. Stone et al. [118] used metadata-based context derived from the social network *Facebook*. Gallagher and Chen [93] used co-occurrence between each person as a logic-based context, which indicated how often a pair of faces appeared together in images. In some

works [92, 94], such co-occurrence of persons is also integrated together with other types of contexts such as events (time stamp) and locations which are rather peculiar to the field of photo collection.

In a scenario of person re-identification across multiple non-overlapping cameras, Zheng et al. [119] and Cai et al. [120] proposed a solution to the problem of associating groups of people between the different camera views and demonstrated that group information helped to resolve the ambiguities in individual appearances. For person identification, they simply combined group cues with individual cues in the form of a weighted sum of each score. They considered a group as a small number of people walking in close proximity in spatial domain, and quantified the group cue by measurement of the spatial appearance features. In these methods, although their group representations are designed to be invariant to positional changes of the group members between the different camera views, the fluctuations in the numbers of observed members, which are caused by absentees, isolation of group members, or the proximities of non-group members, lead to significant changes in the spatial appearance of the group. Accordingly, the effectiveness of the group cue is degraded. For instance, if a certain group is composed of 5 members in the gallery image and only 3 members of the group are observed in a probe image, the observed group tends to be matched with other groups composed of 3 members by mistake. Furthermore, these methods do not consider the behavioral relations among persons such as velocity vector difference through the walking.

Our work is inspired by the related work described above and we propose a unified framework for the person identification problem in video sequences, in which group context is integrated with individual biometric observations by using CRF model. Though, the CRF-based framework is similar to the existing context-assisted person identification schemes formulated by MRF/CRF model such as [93], the major difference of this work is that we use the behavioral relations as group context including spatial distance and velocity vector difference among persons through the video sequences, while existing frameworks used co-occurrences among persons as group context. This also differentiate the proposed method from other group context-based person re-identification methods such as [119] and [120]. Though, similar kinds of behavioral relations are utilized for the problem of trajectory prediction of pedestrians in some works [121, 122] and these are also related to our work, we apply such kind of context to person identification problem, and this is a primal contribution of this paper.

3.3 Group Context-aware Person Identification in Video Sequences

We regard the person identification problem as a many-to-many matching problem for a given image sequence. The task we consider is assignment of a registered person's label to each person that is observed in an input sequence. In this work, *group* is not only explicitly-defined as a unit of people that is composed on the basis of social relations, such as family, friends, and co-workers, but is also implicitly-defined as the result of manual or automatic clustering. Then, the following prerequisites are assumed.

- Each registered person belongs to one of the predefined groups.
- Group affiliation and biometric cues of each registered person are given as gallery data in advance.
- Segmentation and tracking of each subject in an input sequence are obtained in advance.
- Each registered person appears at most once and is likely to appear with group members, in detail, in close vicinity and with similar velocity in an input sequence.

Also, the following conditions are considered:

- Unregistered persons also appear in an input sequence randomly.
- Absence and isolation of a registered person in an input sequence are allowed.

Note that for a registered person who does not belong to any group, an expedient group whose only member is that person is defined, while the "*unregistered*" label is only assigned to actual unregistered persons.

3.3.1 Problem formulation

In the labeling task, we must take account of the relationship between the observed characteristics of each probe, such as spatial position and velocity, and the group affiliation of each label in addition to the biometric cue for each probe. The preferred label assignment, therefore, is one where the same group members are likely to appear in a group, and the biometric cues of each probe are given substantial consideration.

We use a pair-wise CRF (conditional random field) model in a manner similar to [123, 124] for our labeling problem. Let each node in a graph represent a person who appears in

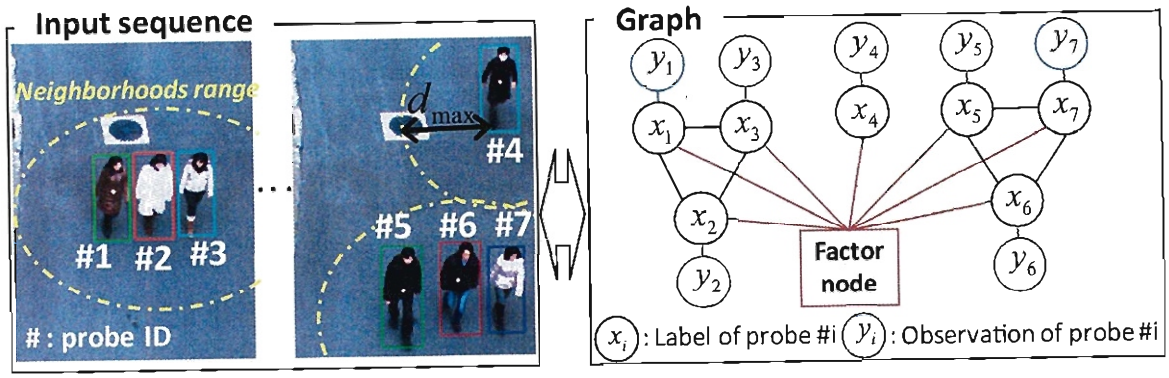


Figure 3.4: Our graphical representation: An example of the input sequence (left) and the corresponding graph (right).

an input sequence, and the label for the i-th node x_i represents the index of the registered person or "*unregistered*" label. The label set is defined as $\mathbf{L} = \{l_1, l_2, \cdots, l_n, l_{un}\}$, where l_k $(k = 1, 2, \cdots, n)$ is the label of the k-th registered person and l_{un} is the label for an unregistered person. A mapping from an individual label to a group is then defined as $g(l_k) \in \mathbf{G}$, where $\mathbf{G} = \{G_1, G_2, \cdots, G_{n_G}, G_{un}\}$ is a group identifier set, G_k $(k = 1, 2, \cdots, n_G)$ is a group identifier for each registered person, and G_{un} is a identifier for an expedient group for any unregistered person.

The graphical representation is shown in Fig. 3.4. In this example, there are seven probe subjects in an input sequence, and each node is connected to neighbor nodes which correspond to the persons within a set spatial distance d_{max}, which is set to 3 [m] in this work, in the input sequence. Also, as described later in detail, all of the nodes are connected by a factor node, which controls the exclusion of each label.

We then let $\mathbf{x}$ be the label assignment for all the nodes and $\mathbf{y}$ be the set of biometric cues for all the nodes, and then the conditional probability of an assignment $\mathbf{x}$ is formulated as,

$$P(\mathbf{x}|\mathbf{y}) \propto \left\{ \prod_i \phi_i(x_i) \prod_{j \in N(i)} \psi_{i,j}(x_i, x_j) \right\} E(\mathbf{x}), \tag{3.1}$$

where $\phi_i(x_i)$ is the local evidence term for node i, $\psi_{i,j}(x_i, x_j)$ is the compatibility term between node i and node j, and $N(i)$ represents a neighbor node set around the node i. $E(\mathbf{x})$ is a label exclusion term, which becomes zero if any registered person label is used more than once and is otherwise one (the label for unregistered persons l_{un} can be used more than once).

The local evidence ϕ_i is defined based on the observed biometric cues for each person. The compatibility $\psi_{i,j}$ corresponds to the group context. The magnitude of the compatibility, therefore, depends on a pair of group identifiers for the label that is assigned to the i-th person

and the j-th person and their spatial distance and velocity vector difference, which are defined in Section 3.4.2 in detail.

3.3.2 Approximate solution via loopy belief propagation

LBP (Loopy belief propagation) [125] is used as an approximate solver to find the assignment $\mathbf{x}$ that maximizes the probability $P(\mathbf{x}|\mathbf{y})$. Ignoring the exclusion term $E(\mathbf{x})$ at this stage, the message $m_{ij}(x_j)$ from node i to node j for each label is defined as,

$$m_{ij}(x_j) \propto \sum_{x_i} \psi_{i,j}(x_i, x_j)\phi_i(x_i) \prod_{k \in N(i)\setminus j} m_{ki}(x_i). \tag{3.2}$$

The belief $b_i(x_i)$ at the node i for each label is found as a marginal probability by gathering messages from its neighbor nodes and from the local evidence,

$$b_i(x_i) = k\phi_i(x_i) \prod_{j \in N(i)} m_{ji}(x_i), \tag{3.3}$$

where k is a normalization constant (summation of belief is normalized to 1). The label assignment of the node i is,

$$x_i^* = \arg\max_l b_i(x_i = l). \tag{3.4}$$

Note that each message is initialized to 1, normalized local evidence is given as the initial belief value, and that the upper limit of iteration of LBP was set to 10 in this work.

3.3.3 Handling the exclusion term

The label exclusion term $E(\mathbf{x})$ is defined such that it forbids the use of a registered person's label more than once, i.e., to suppress the use of the label l_k if another node already has high belief about l_k. Since the label exclusion term is a global function, we can represent it using a factor node that is connected to all of the nodes. In terms of the message passing scheme, the message from a factor node f to a node i is,

$$m_{fi}(x_i = l) \approx \prod_{t \in S\setminus i} \left(1 - m_{tf}(x_t = l)\right), \tag{3.5}$$

where S is the set of all nodes and m_{tf} is defined as,

$$m_{tf}(x_t = l) = \left(b_t(x_t = l)\right)^{\alpha}, \tag{3.6}$$

where α is the message attenuation parameter, and is set to 2 in this work.

Actually, label exclusion via the above message does not completely control the one-time use of the label of a registered person, because the belief of each node for a certain label does not always become 1.0 after message passing. To complete the exclusion control, we therefore execute the *Greedy Algorithm* in terms of the belief score for finalization of the label assignment after the convergence of LBP.

3.4 Implementation

3.4.1 Local evidence

Label of registered person

An observed biometric feature of each person, such as their face or gait, is a crucial clue in itself for person identification, as numerous previous works have demonstrated. We therefore use such a feature as the local evidence for the label of a registered person and it define as,

$$\phi_i(x_i=l_k) \propto p(x_i=l_k|\mathbf{y}_i), \tag{3.7}$$

where $\mathbf{y}_i$ is the observed feature vector of the i-th person and l_k is the label of the k-th registered person. Actually, we regard the prior $p(x_i=l_k)$ as constant for all k, then Eq. (3.7) can be described as,

$$\phi_i(x_i=l_k) \propto p(\mathbf{y}_i|x_i=l_k). \tag{3.8}$$

The probabilistic observation models of the feature vector for each label of each registered person are constructed from gallery feature vectors such as the Gaussian distribution model in advance.

However, since the gallery feature vector of each registered person cannot be captured a number of times, but at most once or twice in most cases, such as real surveillance scenarios, it is difficult to construct the probabilistic model properly in practice. For instance, in the case where only one gallery feature vector is given, it makes no sense to construct a Gaussian distribution as it is. In such a case, therefore, we regard the variation of each feature vector element to be common for all elements and for all persons, and we set the probability model to be,

$$p(\mathbf{y}_i|x_i=l_k) \propto \exp\left(-\frac{{D_{i,k}}^2}{2}\right) \tag{3.9}$$

$$D_{i,k} = \frac{|\mathbf{y}_i-\bar{\mathbf{y}}_k|}{\sigma}, \tag{3.10}$$

where $\bar{\mathbf{y}}_k$ is the average vector of the gallery of the k-th registered person and σ is standard deviation of the feature vector element, which is given as a hyper-parameter.

Label of unregistered person

For the label of an unregistered person, the model cannot be constructed, because the feature vector which represents the "*unregistered person*" can be never captured as gallery data. We thus give a constant value C_{un} as the local evidence for the label l_{un} instead,

$$\phi_i(x_i = l_{un}) = C_{un}. \tag{3.11}$$

3.4.2 Compatibility

The compatibility score for a pair of labels is required to be high only if the group affiliations of the two labels are the same and the corresponding persons appear in close proximity and with similar velocities in an input sequence.

We quantify this using two terms: the distance term E_d and the velocity term E_v. One is based on the spatial distance between the two persons and the other is based on the velocity vector difference between them in the world coordinates. Compatibility for a pair of labels, l_s and l_t, is then defined as,

$$\psi_{i,j}(x_i = l_s, x_j = l_t) \propto \begin{cases} C & (l_s = l_{un} \text{ or } l_t = l_{un}) \\ (1 - \delta_{l_s,l_t}) \left(E_d(d_{i,j}) E_v(v_{i,j}) \delta_{g(l_s),g(l_t)} + C \right) & (\text{otherwise}) \end{cases}, \tag{3.12}$$

where δ is the *Kronecker delta*, $d_{i,j}$ and $v_{i,j}$ are the spatial distance and the velocity vector difference between the i-th person and the j-th person in the world coordinates, and C is a constant value. The distance term $E_d(d_{i,j})$ and the velocity term $E_v(v_{i,j})$ are designed as,

$$E_d(d_{i,j}) = -\frac{(d_{i,j} - d_{max})}{d_{max} - d_{min}} \tag{3.13}$$

$$E_v(v_{i,j}) = -\frac{(v_{i,j} - v_{max})}{v_{max} - v_{min}}, \tag{3.14}$$

where d_{max} and d_{min} are the upper and lower limits of the spatial distance (d_{max} is equal to the one described in Section 3.3.1), and v_{max} and v_{min} are these limits for the velocity difference. In all of our experiments, the parameters are set as $C = 0.1$, $d_{max} = 3$ [m], $d_{min} = 0.5$ [m], $v_{max} = 1$ [km/h], and $v_{min} = 0$ [km/h].

To use the spatial distance and the velocity information in the world coordinates, we need to estimate them from an input video sequence. One of the most reasonable ways to do this is a method based on ground constraints. If the homography correspondences between the ground plane in the world coordinates and the image plane are calibrated in advance, the foot's position trajectory on the ground plane can be estimated from the bottom coordinate of the

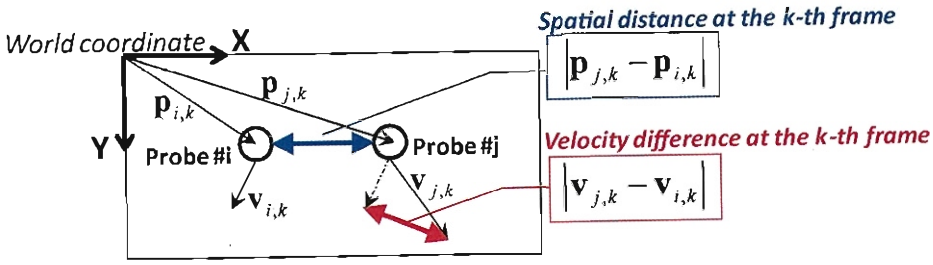


Figure 3.5: Spatial distance and velocity vector difference between a pair of probe subjects (#i and #j) at the k-th frame.

corresponding person in the images. Subsequently, we can derive the spatial distance $d_{i,j}$ and the velocity vector difference $v_{i,j}$ between the i-th and the j-th person as follows,

$$d_{i,j} = \max_{t_s \le k \le t_e} |\mathbf{p}_{i,k} - \mathbf{p}_{j,k}| \tag{3.15}$$

$$v_{i,j} = \max_{t_s \le k \le t_e} |\mathbf{v}_{i,k} - \mathbf{v}_{j,k}|, \tag{3.16}$$

where $\mathbf{p}_{i,k}$ is the smoothed 2D position in world coordinates of the i-th person at the k-th frame, $\mathbf{v}_{i,k}$ is the smoothed 2D velocity vector of the i-th person at the k-th frame (both are illustrated in Fig. 3.5), and t_s and t_e are the first and last frame identifiers for the frames where the i-th person and the j-th person appear together in an input video. Note that, in this case, if a pair of persons does not appear together in any frame, they are not considered to be in the same neighborhood as each other.

3.4.3 Seed node selection

While the ambiguity of a biometric-based identity is solved by messages, it is desirable that a node with confident local evidence for a certain label is then unchanged by messages, to avoid unreasonable belief variation.

For this purpose, we fix the labels of the nodes to such persons with confident local evidence at the first stage. We denote this label-fixed node and the fixed label as the *seed node* and *seed label*, respectively. The seed node is decided using the thresholding mahalanobis distance (Eq. (3.10)) with threshold T_s. More specifically, when only the k-th node has a lower mahalanobis distance than T_s about a certain label l, the k-th node and the label l are regarded as the seed node and the seed label. We then set the belief of the other nodes about the label l to 0 and set the messages to the k-th node from the other nodes and the belief of the k-th node as

$$m_{ik}(x_k = l_j) = b_k(x_k = l_j) = \delta_{l_j,l}, \tag{3.17}$$

where δ is the *Kronecker delta*. Also, we set the message from the seed node (the k-th node) to the other node as,

$$m_{ki}(x_i = l_j) \propto \psi_{k,i}(x_k = l, x_i = l_j). \tag{3.18}$$

Note that the local evidence for the seed label l is regarded as 1 $\left(\phi_k(x_k = l_k) = \delta_{l,l_k}\right)$ in this equation. In addition, in the message passing process, if the belief of a node about a certain label reaches a predefined criterion, which is set to 0.9 in this work, we set the node to be the seed node at that stage.

3.4.4 Relaxation of a biased message caused by an imbalance in the number of group members

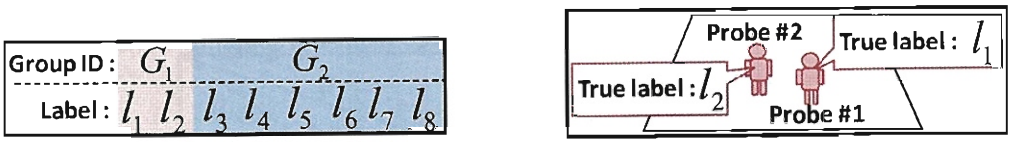
In the presence of an imbalance in the number of group members, the message magnitude is biased by this imbalance. We illustrate this with examples of the gallery set and the situation in an input video as shown in Fig. 3.6.

Consider the messages from probe #2 (the true label is l_2) to probe #1 (the true label is l_1) at the first iteration. As long as the local evidence of probe #2 about the label l_2 is higher than the local evidence about the other labels, the message to enhance the belief for the label l_1 at probe #1 is preferred, because probe #1 and probe #2 belong to the same group G_1 in this situation. For simplicity, suppose that the compatibility between probes #1 and #2 is approximated to $\psi_{2,1}(x_2 = l_s, x_1 = l_t) = \left(1 - \delta_{l_s,l_t}\right) \delta_{g(l_s),g(l_t)}$, where δ is the *Kronecker delta*. The message about the label l_k is then described as,

$$m_{21}(x_1 = l_k) = \sum_{l \in L_{g(l_k)} \setminus l_k} \phi_2(x_2 = l), \tag{3.19}$$

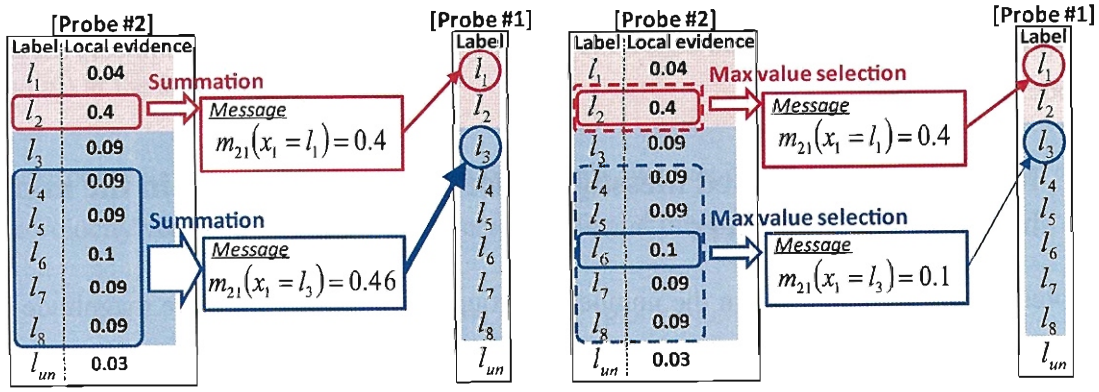
where L_G is a label set of group G members, defined as $L_G = \{l | g(l) = G\}$. Consequently, the magnitude of the message depends not only on the local evidence $\phi_2(x_2 = l)$, but also on the number of group members $|L_{g(l_k)}|$. This may cause an undesired reversal of the message magnitude when the local evidence of probe #2 is given as shown in Fig. 3.7(a). In this case, because the number of group G_2 members is higher than that of the group G_1 members, the summation of the local evidence for the labels of group G_2 becomes higher than that for the labels of group G_1, despite the fact that the local evidence for the label l_2 is the highest, and that the evidence about each label of group G_2 is low. As a result, the message about the label l_3 becomes higher than that about the label l_1, as illustrated in Fig. 3.7(a).

To avoid such undesirable message effects, we propose an alternative message form based on the exclusion of within-group labels via a max selection scheme in message formula (Eq. (3.2))



(a) Gallery set (b) Input situation

Figure 3.6: An example of the gallery and the input situation.



(a) Reversal effect in the standard form (b) Exclusion of within-group labels via the max selection

Figure 3.7: An example of the reversal effect of the message magnitude caused by the bias for the number of group members in the standard message form, and the concept of exclusion of within-group labels via the max selection as a solution to the problem.

as,

$$m_{ij}^{max}(x_j = l_k) \propto \sum_{g \in \mathbf{G}} \max_{l \in L_g} \psi_{i,j}(x_i = l, x_j = l_k)\, \phi_i(x_i = l) \prod_{k \in N(i) \setminus j} m_{ki}(x_i = l) \tag{3.20}$$

In this form, the number of group members no longer influences the message magnitude, because we exclude all of the labels of the group G_k other than the within-group maximum in marginalization of the message, as illustrated in Fig. 3.7(b). The intuitive interpretation of this form is that we model a person-to-group relationship in this message form, rather than a person-to-person relationship, i.e., from the standpoint of probe #1, the magnitude of the message from probe #2 is based not on "who is the probe #2", but "to what group does the probe #2 belong".

3.5 Experiment

In this experiment, the effectiveness of the proposed method was examined first using real video sequences, and the performance for a massive data set was then explored using simulation data sets. We chose gait as the biometric cue and used GEI [36] (22 pixels × 32 pixels) as the gait feature, because it achieved the best performance in [126]. The group affiliation of each

gallery is manually assigned in these experiments. The performance of the proposed method was compared with straightforward local evidence-based labeling via the *Greedy Algorithm*. We evaluated the labeling accuracy R_l as $R_l = \frac{N_t}{N_p}$, where N_p and N_t were the probe number and the correctly labeled probe number, respectively.

3.5.1 Experiment with real image data

We conducted the experiments for two types of real image sequences, one is captured at our campus for preliminary performance evaluation, and the other is obtained from the surveillance cameras installed in a Japanese elementary school.

Preprocessing

We obtained the blob information of each subject in image sequences as follows. First, the foreground regions are extracted via graph-cut-based segmentation [127] in conjunction with background subtraction. Second, each blob is extracted from the foreground regions based on connectivity and the blob statistics, such as area, gravity position, and bounding box are then obtained for each blob. In this process, blobs of different persons may be merged in case where a person is closely-attached to the other person. To avoid such merge, we set the upper limits for the height and width of bounding box respectively, and we split the blob based on the limits if necessary. For example, if the blob has larger height than its upper limit, we count the number of foreground pixels for each height and split the blob at the height with the minimum pixel count within a certain height range.

As for tracking, each bounding box in the current frame is corresponded to the nearest bounding box in the next frame, and the foot's position trajectory of each individual is obtained as a result[1]. Finally, the gait feature of each individual is extracted from the corresponding blob sequence. The bounding box and trajectory contain errors in some degree, and these also decrease the quality of gait feature.

Note that we omitted the occlusion situation among persons in this experiment, because we focus on the evaluation of the effectiveness of the proposed inference algorithm.

Preliminary evaluation

Gallery and probe data set: We used an input sequence (640 pixels × 480 pixels / 15 fps / bmp format) which includes 18 probe subjects, as shown in Fig. 3.8. In this sequence, the walking

[1] We calculated the foot's position on the ground plane from the bottom center coordinate of the bounding box.

Figure 3.8: Snapshots of the input sequence for preliminary experiment.

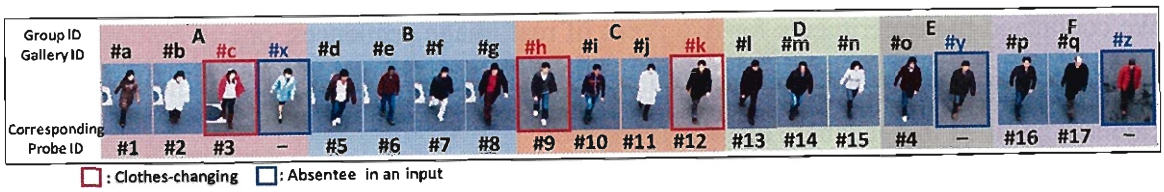


Figure 3.9: Gallery set for the input shown in Fig. 3.8.

directions of all subjects are almost the same. Then, we arranged the gallery set, which includes 20 subjects, as shown in Fig. 3.9. In this setting, the clothes of gallery members #c, #h, and #k are changed at the time of the input sequence to make the person identification problem setting more difficult, which is intentional so that biometric cues alone cannot perfectly identify the subjects. Three absentees (#x, #y, and #z) and one unregistered person (probe #18) are arranged to demonstrate that the proposed method can handle such situations.

In this experiment, the label for each gallery is denoted by a corresponding gallery ID for convenience as $\mathbf{L} = \{\#a, \#b, \cdots, \#un\}$, where #un is the label for an unregistered person.

Parameters: The standard deviation of the feature vector element was set at $\sigma = 394.5$, which is determined from the other preliminary experiment. Local evidence for the label of the unregistered person was set at $C_{un} = \frac{1}{N_l}$, where N_l is the number of gallery labels.

Results: Table 3.3 shows the initial label correspondence via straightforward labeling. In this table, seven probe subjects (#3, #6, #7, #12, #13, #15, and #18) are initially mislabeled because of the within-class variation of the gait features caused by walking manner variations, clothes changes, and silhouette noise.

We illustrate the message effect on improving the belief from the initial state by taking probe #3 as an example. As shown in Fig. 3.10, probe #3 is connected to probes #1 and #2, which truly belong to group A (the same group as probe #3), and probe #4, which truly belongs to group E. Initially, probe #3 is mislabeled as #z and probes #1, #2, and #4 are correctly labeled, as shown in Table. 3.3. The received message and the belief of probe #3 after the first message passing is then shown in Fig. 3.11. In this figure, we see that the messages from probe

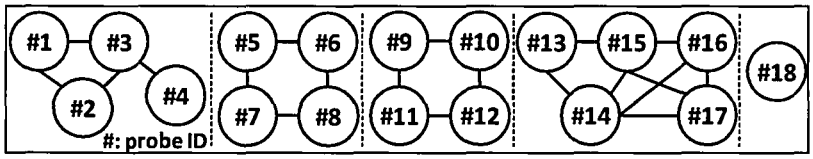


Figure 3.10: Node connection between the probe subjects in Fig. 3.8.

Table 3.1: Compatibility for a pair of labels in the same group between probe subjects in Fig. 3.8.

Probe pair (#i, #j)	$\psi_{ij}(x_i=l_k, x_j=l_l)$ $l_k \neq l_l$ and $g(l_k)=g(l_l)$
(#1, #3)	0.42
(#2, #3)	0.73
(#4, #3)	0.16

Table 3.2: Labeling accuracy for the input shown in Fig. 3.8.

Method	Labeling accuracy
Straightforward	0.61
Proposed	1.0

#1 and probe #2 contribute much to boost the belief for the label #c. This is because probes #1 and #2 have high initial beliefs (local evidence) for their true labels, and high compatibilities for a pair of labels which belong to the same group as probe #3, as shown in Table. 3.1.

On another note, in the message shown in Fig. 3.11, the message about the label #x (absentee) is relatively high because #x is also a member of group A. The belief of probe #3 for the label #x, however, does not exceed that for the true label #c because local evidence for the true label #c is essentially higher than that for the label #x, even though the message magnitude for label #x is nearly equal to that for the label #c.

In this way, the initial mislabel assignments gradually improve with iteration of the message passing. Note that probe #18, which is an unregistered person and is initially mislabeled as #k, is not connected to any probe subject, as shown in Fig. 3.10, but is only connected to the factor node in this case. The assigned label to probe #18 is therefore changed only by exclusive force with an increase in the beliefs of the other labels.

The labeling accuracy of the proposed method under no seed node and of the straightforward method are shown in Table. 3.2 (in this experiment, the result of proposed method is unchanged with or without seed nodes). In this table, we can see that the proposed method significantly improves the labeling accuracy.

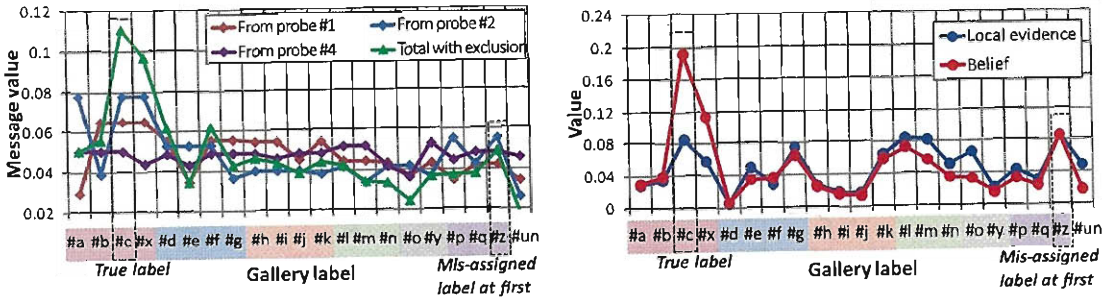
Evaluation for the dataset from the real surveillance camera

Gallery and probe data set: We arranged the real image sequences (320 pixels × 240 pixels / 9 fps / jpeg format) which are obtained from the surveillance cameras installed in a Japanese

Table 3.3: Initial label correspondence for an input shown in Fig. 3.8. The numerical value in the table represents the belief.

Probe ID	Gallery ID																				
	#a	#b	#c	#x	#d	#e	#f	#g	#h	#i	#j	#k	#l	#m	#n	#o	#y	#p	#q	#z	#un
#1	0.177	0.023	0.016	0.011	0.002	0.120	0.025	0.031	0.027	0.057	0.115	0.077	0.054	0.018	0.058	0.051	0.023	0.044	0.012	0.011	0.048
#2	0.014	0.171	0.060	0.032	0.032	0.027	0.041	0.087	0.041	0.004	0.018	0.046	0.029	0.051	0.013	0.037	0.052	0.060	0.096	0.045	0.048
#3	0.029	0.034	0.085	0.058	0.005	0.050	0.029	0.074	0.027	0.017	0.017	0.065	0.084	0.082	0.051	0.066	0.021	0.043	0.030	0.085	0.048
#5	0.001	0.077	0.028	0.003	0.564	0.005	0.038	0.081	0.017	0.000	0.001	0.007	0.006	0.015	0.001	0.005	0.024	0.010	0.050	0.021	0.048
#6	0.099	0.001	0.007	0.057	0.000	0.065	0.003	0.009	0.004	0.124	0.040	0.024	0.052	0.017	0.354	0.075	0.003	0.011	0.001	0.009	0.048
#7	0.028	0.047	0.028	0.027	0.005	0.045	0.091	0.053	0.077	0.017	0.011	0.049	0.062	0.098	0.010	0.049	0.131	0.053	0.049	0.023	0.048
#8	0.008	0.089	0.055	0.032	0.028	0.020	0.044	0.136	0.029	0.003	0.005	0.045	0.038	0.079	0.008	0.029	0.079	0.031	0.126	0.066	0.048
#9	0.043	0.048	0.057	0.018	0.006	0.097	0.037	0.044	0.082	0.027	0.033	0.084	0.054	0.054	0.031	0.058	0.051	0.075	0.032	0.021	0.048
#10	0.095	0.005	0.007	0.023	0.000	0.117	0.016	0.009	0.038	0.230	0.049	0.068	0.040	0.026	0.071	0.079	0.022	0.048	0.005	0.004	0.048
#11	0.112	0.017	0.010	0.038	0.001	0.050	0.019	0.014	0.024	0.068	0.238	0.078	0.048	0.015	0.066	0.053	0.025	0.054	0.016	0.007	0.048
#12	0.048	0.015	0.036	0.110	0.001	0.053	0.024	0.045	0.021	0.043	0.018	0.067	0.116	0.079	0.079	0.076	0.028	0.028	0.019	0.045	0.048
#13	0.048	0.031	0.053	0.020	0.007	0.106	0.044	0.065	0.057	0.028	0.039	0.075	0.087	0.045	0.030	0.054	0.036	0.053	0.033	0.038	0.048
#14	0.018	0.046	0.117	0.035	0.007	0.041	0.026	0.054	0.067	0.013	0.012	0.042	0.082	0.139	0.024	0.046	0.044	0.061	0.034	0.045	0.048
#15	0.114	0.007	0.014	0.032	0.000	0.059	0.007	0.010	0.014	0.091	0.089	0.063	0.051	0.020	0.247	0.069	0.010	0.039	0.005	0.010	0.048
#4	0.045	0.012	0.023	0.099	0.000	0.065	0.018	0.018	0.027	0.063	0.042	0.039	0.054	0.046	0.136	0.157	0.017	0.058	0.008	0.025	0.048
#16	0.031	0.035	0.031	0.045	0.002	0.077	0.044	0.027	0.074	0.036	0.031	0.062	0.037	0.060	0.034	0.114	0.052	0.112	0.028	0.019	0.048
#17	0.014	0.092	0.033	0.051	0.006	0.028	0.048	0.050	0.032	0.006	0.017	0.063	0.038	0.057	0.010	0.039	0.081	0.056	0.183	0.049	0.048
#18	0.095	0.026	0.009	0.038	0.001	0.071	0.036	0.014	0.039	0.077	0.084	0.091	0.033	0.032	0.061	0.087	0.044	0.085	0.021	0.007	0.048

: True correspondence : Assigned correspondence



(a) Received messages

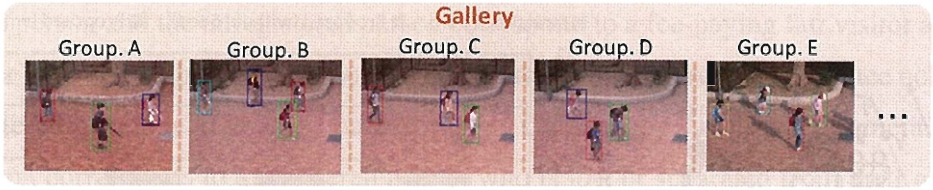
(b) Local evidence and belief

Figure 3.11: Received messages and belief of probe #3 at the first message passing.

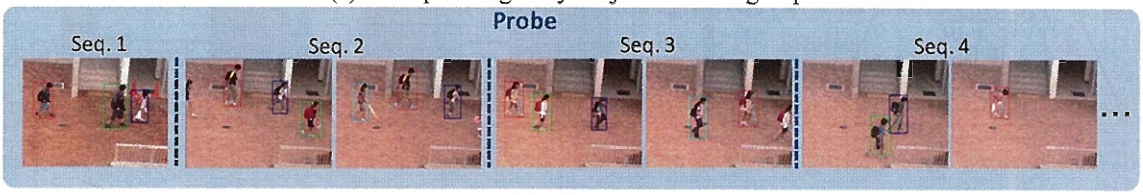
elementary school. In this experiment, a scenario of person re-identification across two non-overlapping cameras is assumed and we collected gallery and probe subsequences from the two different cameras. The numbers of gallery and probe subjects are shown in Table. 3.4, and the examples are shown in Fig. 3.12.

In this dataset, the observation angle of each subject is different to some extent between gallery and probe sequences and the trajectory and walking manner of each subject are more fluctuated than those in the dataset used in previous section.

Parameters: The standard deviation of the feature vector element was set at $\sigma = 1071.1$ and the seed decision threshold was set at $T_s = 0.7$. Both of these values were determined based on the training dataset composed of 40 subjects which are also extracted from the same cam-



(a) Examples of gallery subjects and the groups



(b) Examples of probe subjects

Figure 3.12: Examples of gallery and probe subjects in the dataset from the real surveillance camera.

Table 3.4: Gallery and probe settings in the dataset from the real surveillance camera.

Gallery setting				Probe setting						
Number of group	Number of member	Subject number		Absentee	Number of group	Subject number				
						Registered			Unregistered	Total
		Group belonging	Stand alone			Group belonging		Stand alone		
						In a group	In isolation			
14	2 to 5	39	1	0	16	37	2	1	7	47

eras. Local evidence for the label of the unregistered person was set in the same way as the preliminary experiment.

Result: Table 3.5 shows the labeling accuracy. In this table, we can see that the proposed method improves the labeling accuracy even for the real situation.

3.5.2 Experiment with simulation data

Settings

Observed space and trajectory: We assumed an input video sequence in which each walking person is captured by a surveillance camera in a virtually constructed space. We set the whole space to be 10 [m] $\times$ 2000 [m] and the observed space to be 10 [m] $\times$ 20 [m] as shown in Fig. 3.13. In such a space, we arranged the initial position for each person, gave them

Table 3.5: Labeling accuracy for the dataset from the real surveillance camera.

Method	Labeling accuracy
Straightforward	0.70
Proposed	0.87

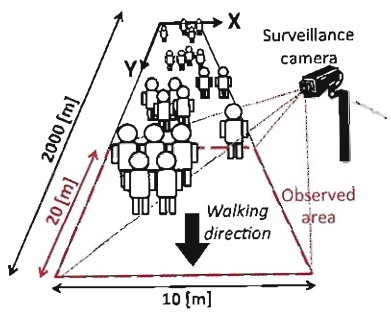


Figure 3.13: Assumed environment in simulation experiments.

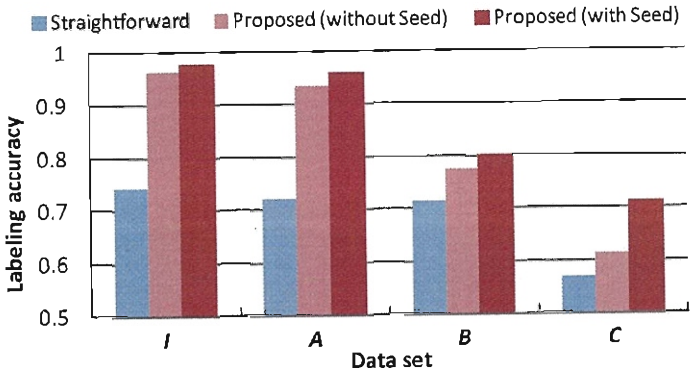


Figure 3.14: Results for simulation data set.

Table 3.6: Gallery and probe settings in simulation experiments.

Data set	Gallery setting					Probe setting						
	Number of group	Number of member	Subject number			Absentee	Number of group	Subject number			Unregistered	Total
								Registered				
			Group belonging	Stand alone	Total			Group belonging		Stand alone		
								In a group	In isolation			
I	125	6 to 10	1000	0	1000	0	125	1000	0	0	0	1000
A						10		980	10		10	510
B	100	2 to 10	500	500		500	50	200	50	250	500	1000
C												

velocities, and then moved them. For simplicity, we assumed that each person walked with constant velocity and that the walking direction was only the Y-direction, as shown in Fig. 3.13.

Gallery and probe data set: In all the simulation experiments, the number of gallery subjects (registered persons) is set to 1000, and gait features for all of the gallery and probe subjects are randomly chosen from the gait database proposed in [126]. Note that the gait database [126] has expanded and includes 1,580 subjects at time of writing. We used two side-view sequences as the probe and gallery sequences.

We then considered the following three scenarios, and we defined the gallery and probe settings for each scenario as shown in Table. 3.6.

Set *A*: *Person identification when going to elementary school in a group*: All of the gallery subjects are grouped. The registered person and the unregistered person correspond to a school student and an intruder, respectively. Absentees and isolated persons correspond to absent students and early or late arrival students. There can be a small number of unregistered persons, absentees, and isolated persons.

Set *B*: *Person identification in amusement theme parks*: Substantial numbers of the gallery subjects are assumed to be standalone (persons who belong to groups of only one member). The

registered person and the unregistered person correspond to a fee-paying fair visitor and an unfair visitor who enters the park without the due entrance procedure. The absentee corresponds to a registered person who is in the park but is not captured by surveillance camera. The isolated person corresponds to a registered person who is lost or separated from their group with another objective. There can be a small number of unregistered persons and isolated persons in addition to some absentees.

Set *C*: *Person re-identification in network cameras*: We assume that there are two cameras which have different fields of view, and regard one side camera as the gallery-side camera and the other as the probe-side camera. Some of the gallery subjects are assumed to be standalone. A registered person corresponds to a person who is captured by the gallery-side camera, and an unregistered person corresponds to a person who is captured only by the probe-side camera. An absentee corresponds to a registered person who is not captured by the probe-side camera. An isolated person corresponds to a registered person who is separated from their group with another objective. There can be some unregistered persons and absentees, and a small number of isolated persons.

We also arranged the ideal scenario, where all gallery subjects are grouped and there are no absentees, isolated persons, or unregistered persons (denoted as set I in Table. 3.6). We arranged 10 different sets randomly for each scenario. The performance for each data set is evaluated by averaging their results.

Parameters: The standard deviation of the feature vector element was set at $\sigma = 366.2$ and the seed decision threshold was set at $T_s = 0.8$. Both of these values were determined based on the gait database used. The local evidence for the label of an unregistered person C_{un} significantly influences the performance of the many-to-many labeling scheme in the presence of an unregistered person, particularly in the presence of a relatively large number of unregistered persons in an input sequence such as set C. Thus, we set the parameter at $C_{un} = 0$ for set I, $C_{un} = 0.002$ for sets A and B, and $C_{un} = 0.005$ for set C, so that the performance of the straightforward method for each data set becomes the best. Note that we also conducted the same experiments under no seed node ($T_s = 0.0$) to verify the effectiveness of seed node.

Results

Figure 3.14 shows the labeling accuracy. In this figure, we see that the proposed method discernibly improves the labeling accuracy for each data set and the introduction of seed node contributes the performance improvement. In particular, when the ratio of the number of persons in a group is high, the effectiveness of the proposed method is greatest, as shown in the

results for sets I and A, while the performance improvements for sets B and C are relatively low.

Basically, the belief values for isolated persons, standalone persons, and unregistered persons for their own true labels are not expected to be directly boosted by the messages. Thus, in the case where such a person has the highest belief for a wrong label about another person at the first stage, it is difficult to recover the true label, except in the case where the wrong label is a label about a person in a group in an input sequence; that is, the exclusive force for the wrong label is expected (as the label change of probe #18 shows in the experiment in Section 3.5.1). This is one of the major reasons why the performances of the proposed method for sets B and C are lower than those of sets I and A.

3.6 Discussion

3.6.1 Limitation

While the proposed method significantly improves the labeling performance, there are still some subjects who are mislabeled, and subjects whose labels are negatively changed via message passing, even for the ideal set I in the simulation experiments. We list the typical cases of failure for the proposed method as follows.

Mislabel within the same group members

When a person in a group is mislabeled as another person in the same group at first, it is difficult to recover the true label because the belief for the true label and the wrongly assigned label are boosted to the same degree. Mislabeling within the same group members is, however, relatively rare compared with mislabeling between different groups. The rate of this kind of mislabel is relatively low.

Negative label change in the presence of an absentee or an isolated person in a group

As shown in Fig. 3.15, when the following three incidents occur simultaneously, where i) an absentee or an isolated person exists in a group (the gallery subject with the label l_4 in group G_1), ii) *another person*[2] (probe #4) comes close to the group members (probe #1, #2, and #3) with similar velocity vector in an input sequence, and iii) *another person* is not set as a seed. Then, *another person* may possibly be mislabeled as an absentee or an isolated person by messages from the group members. At the same time, if *another person* is mislabeled as

[2]Not only a standalone person or an unregistered person, but also a person from another group.

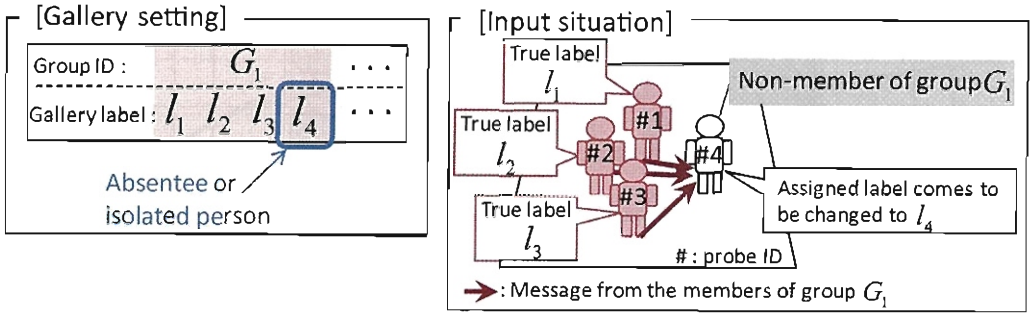


Figure 3.15: Example situation of negative label change.

an isolated person in such a case, the initial correct label assignment for the identical isolated person is excluded by *another person* and changed to the other incorrect label. Note that this often occurs in the presence of a number of standalone persons, unregistered persons, and isolated persons, such as our simulation sets B and C, because the event probability of the above incident increases.

Though the initial mislabel assignment and negative label changes as listed above possibly cause other negative label changes through the propagation of an undesirable message using the proposed method, the impact of such a negative effect is basically smaller than that of the positive effects in total, as shown in the results of the proposed method (Fig. 3.14).

3.6.2 Effect of the seed node on performance

The contribution of seed node to the performance improvement of the proposed method is demonstrated in the simulation results (Fig. 3.14). The advantages of introducing seed node in graph are considered as followings.

- The avoidance of negative label change: As discussed in Section 3.6.1, the negative label change is not occurred if *another person* (which is described in Section 3.6.1) is set as a seed.

- The enhancement of message effect: According to the Eq. (3.18), a seed node can send more discriminative messages for the labels which belong to the same group of the assigned seed label as following example. We consider again the situation shown in Section 3.4.4 (Fig. 3.6 and Fig. 3.7 (b)), and let assume that the probe #2 is set as a seed with seed label l_2, that is, the local evidence for the label l_2 is set to 1 and that for each of all the other label is set to 0 in Fig. 3.7 (b). In this case, the messages from probe #2 to probe

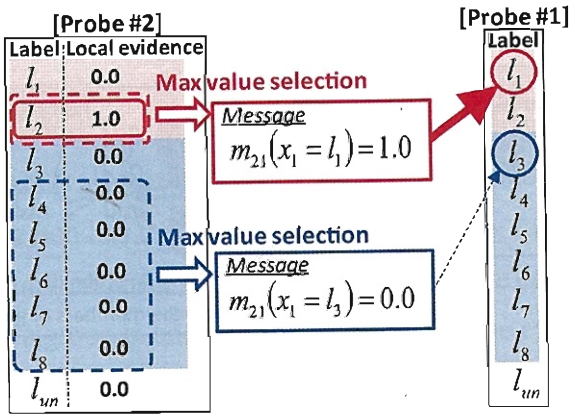


Figure 3.16: Messages from a seed node (probe #2) to the other node (probe #1) under the setting shown in Fig. 6 (Section 4.4).

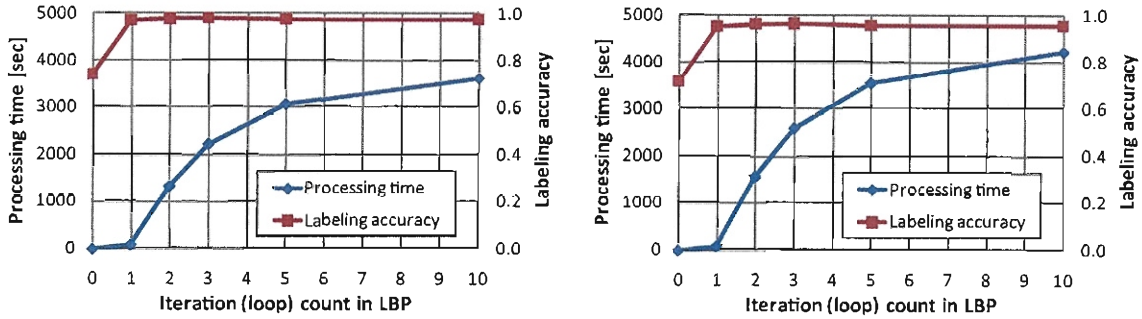
#1 for the labels l_1 and l_3 become as, $m_{21}(x_1 = l_1) = 1.0$ and $m_{21}(x_1 = l_3) = 0.0$, respectively[3] as shown in Fig. 3.16. Therefore, the messages from a seed node promote the belief updates of its neighbor nodes, and positive label changes of them are also expected to be promoted as a result. Though, the negative label changes are possibly promoted, in particular, in the case that a seed node is assigned false label as seed label, such negative case is assumed to be occurred less often than positive case.

3.6.3 Effect of the absence of homography calibration on performance

We assume the homography calibration for the calculation of the position of each subject as described in Section 3.4.2. The cost of calibration is, however, expensive in some practical systems. One of the alternative ways is a direct use of the image pixel coordinate system instead of the world coordinate system to represent the trajectory of each individual. In many of practical surveillance systems, the camera captures the scene from near the top view or oblique view just like the scene used in our experiments. In such views, it is assumed that the direct use of image pixel coordinate does not have a serious impact on the performance of the proposed method.

To examine this, we conducted an additional experiment for the dataset used in Section 3.5.1 and we used the image pixel coordinate directly for the calculation of the positions of individuals. The parameters are set in pixel units, and we decided the parameters $d_{max} = 160$ [pixel] and $d_{min} = 30$ [pixel] based on the road width (approx. 320 pixels) and human width (approx. 30

[3]This is an extreme case and the degree of magnitude relation between these messages are biased by constant value C in actual.



(a) Dataset *I* (b) Dataset *A*

Figure 3.17: The relationships of iteration count in LBP with the labeling accuracy and processing time for the dataset *I* and *A*.

Table 3.7: Processing times [sec] for the dataset *I* and *A*.

Method		Dataset *I*	Dataset *A*
Straightforward		2.5	2.5
Proposed	*after the first iteration*	88.5	86.9
	after the last iteration	3638.7	4248.1

pixels), and the $v_{max} = 15$ [pixel/sec] and $v_{min} = 0$ [pixel/sec] based on the average velocity 60 [pixel/sec] which roughly estimated from the dataset. As a result, we get the same result with that shown in Table 3.2, though the neighbor relationships among probe subjects are slightly changed.

3.6.4 Relationship between labeling accuracy and computational cost

The computational cost of the proposed method is largely dependent on the calculation of messages (message update procedure) and the iteration count of the message passing in LBP. The time complexity of a message update procedure is roughly estimated as, $O\big(N_n(N_p - N_s)(N_g - N_s)^2\big)$ at the first message update, and $O\big(N_n^2(N_p - N_s)(N_g - N_s)^2\big)$ at the second and subsequent message update, where N_p, N_g, N_n, and N_s are the number of probe subject, that of gallery subjects, average number of neighbor nodes, and that of seed nodes.

First, the resultant processing times in simulation experiments are investigated, and the relationships of iteration count in LBP with the labeling accuracy and processing time for the dataset *I* and *A* are shown in Fig. 3.17. Note that the experiments are done on a 2.20 GHz AMD Opteron(tm) Processor 6174 PC running Microsoft Windows Server 2008 operation system, and the message passing scheme is parallelized via multi-thread processing of 24 threads. In this figure, we can see that the labeling accuracy is almost saturated after the first iteration,

though the processing time is gradually and largely increased. These results indicate that only one-time iteration of message passing is enough in terms of labeling accuracy. Then, detailed processing times for these datasets are shown in Tab. 3.7. As for the proposed method, both of the processing time at the end of the first iteration (indicated as *after the first iteration*) and that at the end of the last iteration (indicated as *after the last iteration*) are shown in this table. From these results, we can see that all the processing times of the proposed method after the first iteration are less than 90 [sec], which seems to be reasonable to some extent for practical use, though they are still far from real-time even with the use of high-performance PC as described above.

3.6.5 Issues toward the practical system

The proposed method is based on some assumptions as described in Section 3.3. In terms of the total system (practical surveillance system), however, the following challenging issues are required to be addressed in future work.

Obtaining the group affiliation

In practice, we need some kinds of registration procedures to associate the group affiliations with the individuals in advance. This is not such a serious problem in surveillance systems at factories and schools, where the potential observed persons are well-known in advance, i.e. the school children and the factory workers. Also, the registration can be achieved relatively easily with a system constructed at a place where the entrance and exit are controlled, i.e., where the group affiliation of each person can be easily checked and registered at the entrance gate, as in amusement or theme parks, stadiums, theaters, and airports. Alternatively, group affiliations can be derived by manual annotation (by user interaction) of the video sequence, and also inferred automatically by means of grouping techniques, such as data mining and clustering. In particular, social behavior-based group finding techniques have been developed in recent years [128][129]. In these methods, the group is estimated based on trajectory, distance, and velocity of pedestrians. Thus, these methods bear affinity with the proposed method in terms of focusing such kinds of social behaviors, and the integration with these techniques is future work for the practical use of the proposed method.

Handling of more detailed relationship among individuals

As shown in Eq. (3.12), we formulate the compatibility uniformly for each pair of persons in the same group, and also do that for each pair of persons in different groups. This means that

we assume the uniform strength of intra-group relationship and that of inter-group relationship. In practice, however, such strengths might not be equivalent, and rather more complex in some cases. For example, in the case of a group of friends, a person in the group might be especially friendly with a certain member of the group compared with other members, and the person might also be friendly with persons in different groups. Ideally, in such case, the difference in relational strength among pairs of persons is desired to be reflected in the compatibility. In the proposed method, this is possibly realized by introducing layered representation of the group affiliation, corresponding layered conditional branching of the compatibility function, and corresponding parameter settings of d_{max}, d_{min}, v_{max}, v_{min}, and C. Besides, to put it in an extreme way, these parameters could be tuned for each pair of persons without deterministic group affiliation, when the strength of each pair-wise relation is well known. Of course, such design of compatibility leads to explosion of the number of parameters and it is almost impossible to tune the parameters by manual. Therefore, automatic tuning or learning of the parameters in conjunction with automatic obtaining of group affiliation as above mentioned is required when we introduce such extended compatibility.

Obtaining the trajectory and biometric cue

Segmentation and tracking of each person are essential for the acquisitions of the trajectory and biometric cue, and these are not easy tasks when the scene is crowed, in particular, in the presence of occlusion among persons. To evaluate the proposed method for more practical scenes including such occlusion relationships, state-of-the-art techniques of segmentation and tracking, such as [130], [131], and [132] are required to be applied for this problem. Moreover, cross-view matching of biometric cue is also essential and in the case of gait-based identification, the view transformation model [41] can be applied for this issue. The integration of these techniques with the proposed method also remains in future work.

3.7 Conclusion

In this chapter, we proposed the behavior-based group context for person identification in video sequences and integrated it in the framework of CRF. In the proposed method, by means of message passing, the belief of individual identity is propagated to neighborhoods based on their group affiliation information and their behavioral differences, such as the spatial distance and the velocity vector difference in an input sequence, so that the same group members enhance one member's belief as those group members enhance each others' beliefs. In our experiments,

we showed that the proposed method significantly improves the performance compared with the straightforward method based on biometric cues alone.

Our future work includes construction of the model for optimal selection of local evidence for the label of an unregistered person C_{un}. This is a rather general issue for many-to-many matching problems when considering an unregistered person.

Chapter 4

The OU-ISIR Gait Database Comprising the Large Population Dataset and Performance Evaluation of Gait-based Person Identification

4.1 Introduction

For the development and statistically reliable evaluation of gait-based person identification approaches, the construction of a common gait database is essential. There are two considerations in constructing a gait database: (1) the variation in walking conditions (e.g., view, speed, clothing, and carrying conditions), and (2) the number and diversity of the subjects. The first consideration is important in evaluating the robustness of the gait-based person identification, because walking conditions depend on the time and circumstances and often differ between gallery and probe. For instance, the clothing and carrying conditions when walking along a street in a suit with a bag while on business can differ from those when strolling empty-handed in casual clothes during leisure time. The second consideration is important to ensure statistical reliability of the performance evaluation. Moreover, if the database is used for soft biometric applications such as gait-based gender and age classification [133, 134], the diversity of subjects in terms of gender and age plays a significant role in the performance evaluation.

Although several gait databases have been constructed [43, 44, 39, 45, 46, 47, 33, 40, 48, 49, 50, 42, 51, 52], with most of these taking good account of the first consideration, the second consideration is still insufficiently addressed since these databases include at most 185 subjects [51] and the subjects' genders and ages are biased in many of the databases. The exceptions are the large-scale datasets introduced in [126] and [135], which do address the second consideration and include respectively, 1,035 and 1,728 subjects with ages ranging

Table 4.1: Existing major gait databases

Database	#Subjects	Data covariates
Soton database	12 [45]	4 views, 5 shoes, 3 clothes, 5 bags (including w/o), 3 speeds
	115 [39]	3 scenarios (outdoor, indoor track, treadmill), 2 views per scenario
	25 [136]	Time (0, 1, 3, 4, 5, 8, 9, and 12 months), 12 views
USF dataset	122 [33]	2 views, 2 shoes, 2 surfaces, baggage (w/ and w/o), time (6 months)
CASIA dataset	20 [46]	3 views
	124 [40]	11 views, clothing (w/ and w/o coat), baggage (w/ and w/o)
	153 [48]	3 speeds, baggage (w/ and w/o),
OU-ISIR Gait Database, Treadmill dataset	34 [50]	9 speeds (2, 3, 4, 5, 6, 7, 8, 9, and 10 km/h)
	68 [42]	32 clothes combination at most
	185 [51]	Gait fluctuation among periods
	168 [52]	25 views

from 2 to 94 years. In these datasets, however, the gait images are captured using cameras with varying poses (e.g., a camera's pose on one day differs slightly from that on another day, or some subjects are captured using first one camera and then another with a slightly different pose) and this could introduce bias into the evaluation results.

In this study, we focus on the second consideration and introduce a large population dataset that is a major upgrade to previously reported large-scale datasets in [126] and [135]. The extensions of this dataset are as follows.

1. The number of subjects is considerably greater in the dataset; i.e., there are more than thrice the number of subjects in the dataset in [126] and more than twice the number in the dataset in [135].

2. All silhouette images are normalized with respect to the image plane to remove the bias of camera rotation for more equitable performance evaluation.

3. The observation angle of subjects in each frame is specifically defined for the sake of fair analysis in terms of the observation angle, whereas previous works merely defined the angle as a *side view*.

Our dataset is the largest gait dataset in the world, comprising over 4,000 subjects of both genders and including a wide range of ages. Although the dataset does not include any variations

in walking conditions, it allows us to investigate the upper limit of identification performance in a more statistically reliable way and to reveal how gait-based person identification performance differs between genders and age groups. Thus, our dataset can contribute much to the development of gait-based applications, and we demonstrate its validity through experiments with state-of-the-art gait representations.

The outline of the paper is as follows. Section 4.2 introduces existing gait databases, while Section 4.3 addresses the construction of the dataset. The gait-based person identification approach for performance evaluation is described in Section 4.4, and various performance evaluations using our dataset are presented in Section 4.5. Section 4.6 presents our conclusions and discusses future work.

4.2 Related Work

Existing major gait databases are summarized in Table 4.1. Here, we briefly describe these databases.

The Soton database is composed of a small population dataset [45] and a large population dataset [39]. The small dataset contains subjects walking around an indoor track, with each subject filmed wearing a variety of footwear and clothing, carrying various bags, and walking at different speeds. Hence, the database is used for exploratory factor analysis of gait-based person identification [137]. The large dataset was the first gait database to contain over 100 subjects and has contributed to the study of gait-based person identification mainly in terms of inter-subject variation. The recently published Soton Temporal database [49] contains the largest time variations; up to 12 months to date [136]. It enables the investigation of the effect of time on the performance of gait biometrics, allowing the use of 3D volumetric data.

The USF dataset [33] is one of the most widely used gait datasets and is composed of a gallery and 12 probe sequences captured outdoors under different walking conditions including factors such as view, shoes, surface, baggage, and time. As the number of factors is the largest of all existing databases, despite there being only two variations for each factor, the USF database is suitable for the evaluation of the inter-factor effect, as opposed to the intra-factor effect, on identification performance.

The CASIA database, Dataset A [46] contains image sequences from three views and can be used for the analysis of the effect of the view angle on identification performance. The CASIA database, Dataset B [40] consists of multi-view (11 views) walking sequences and includes variations in the view angle, clothing, and carrying conditions. Since it contains the

finest azimuth view variations, it is useful for the analysis and modeling of the effect of view on gait-based identification [138]. The CASIA database, Dataset C [48] was the first database to include infrared gait images captured at night, thus enabling the study of gait-based person identification at night.

The OU-ISIR Gait Database, Treadmill Dataset [50, 42, 51, 52] contains gait images of subjects on a treadmill with the largest range of view variations (25 views: 12 azimuth views times 2 tilt angles, plus 1 top view), speed variations (9 speeds: 1 km/h intervals between 2 and 10 km/h), and clothing variations (up to 32 combinations), and as such, it can be used to evaluate view-invariant [41], speed-invariant [50] and clothing-invariant [42] gait-based person identification. In addition, it is used to analyze gait features in gender and/or age-group classification [52], since the diversities of gender and age of the subjects are greater than those in currently available gait databases.

Next, we review the number and diversity of subjects. Table 4.1 shows that existing major databases include more than 100 subjects. Although these databases are statistically reliable to some extent, the number of subjects is insufficient when compared with databases of other biometrics such as fingerprints and faces. In addition, the populations of genders and ages are biased in many of these databases; e.g., there are no children in the USF dataset with most of the subjects in their twenties and thirties, while the ratio of males to females is 3 to 1 in the CASIA dataset (Dataset B). Such biases are undesirable in performance evaluation of gait-based gender and age-group estimation and in performance comparison of gait-based person identification between genders and age groups.

4.3 The OU-ISIR Gait Database, Large Population Dataset

4.3.1 Capture System

An overview of our capture system is illustrated in Fig. 4.1. Each subject was asked to walk at his or her own preferred speed through a straight course (red arrows) at most twice under the same conditions. The length of the course was approximately 10 m, with approximately 3 m (at least 2 m) sections at the beginning and end regarded as acceleration and deceleration zones, respectively. Two cameras were set approximately 4 m from the walking course to observe (1) the transition from a front-oblique view to a side view (camera 1), and (2) the transition from a side view to a rear-oblique view (camera 2). We used Flea2 cameras manufactured by Point Gray Research Inc. with HF3.5M-2 lenses manufactured by SPACE Inc. The image size and frame rate were, respectively, 640 $\times$ 480 pixels and 30 fps. The recorded image format

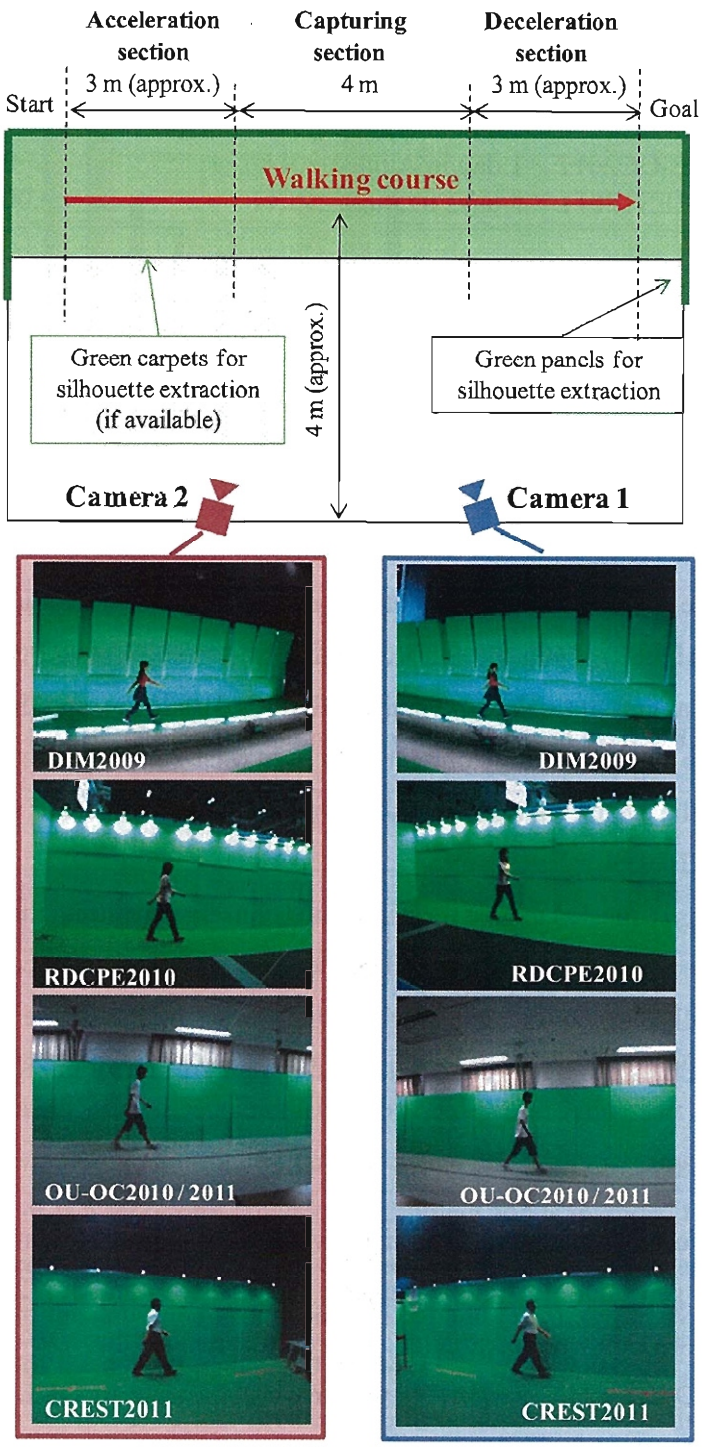


Figure 4.1: Overview of capture system and captured images.

Table 4.2: Visitors at events

Event	Term	#Visitors (approx.)
DIM2009	3 days in March 2009	1,600
RDCPE2010	2 days in June 2010	280
OU-OC2010	1 day in August 2010	70
OU-OC2011	1 day in August 2011	90
CREST2011	5 days in August 2011	2,000

was uncompressed bitmap. Moreover, green background panels and carpet (if available) were arranged along the walking course for the purpose of clear silhouette extraction.

4.3.2 Data Collection

The dataset was collected during entertainment-oriented demonstrations of an online gait personality measurement [139] at outreach activity events in Japan, including the Dive Into the Movie project (DIM2009) [140], the 5th Regional Disaster and Crime Prevention Expo (RDCPE2010), Open Campus at Osaka University (OU-OC2010/2011), and the Core Research for Evolutional Science and Technology project (http://www.jst.go.jp/kisoken/crest/en/index.html, CREST2011). All the events were held at indoor halls and the numbers of visitors at each event are summarized in Table 4.2.

Each subject was requested to give their informed consent permitting the use of the collected data for research purposes. Also, the age and gender of each subject were collected as metadata. All the subjects walked empty-handed, wearing their own clothing (some subjects wore a hat) and footwear. Examples of images captured at each event are shown in Fig. 4.1.

4.3.3 Statistics

From the data collected by camera 1 (images were taken with two cameras at the events), the world's largest gait dataset of 4,007 subjects (2,135 males and 1,872 females) with ages ranging from 1 to 94 years was constructed. We call this dataset the "*OU-ISIR Gait Database, Large Population Dataset C1 Version1*"[1], which we abbreviate to **OULP-C1V1**[2]. Detailed distributions of the subjects' gender and age are shown in Fig. 4.2, while example images of the subjects are shown in Fig. 4.3. Almost all the subjects are of Asian descent.

[1]To be prepared for publication. The data will be published in the form of normalized silhouette image sequences in PNG format, with a total data size of about 1.5 GB.

[2]The naming format is OULP-[*camera ID*][*version ID*]-[*header1*]-[*header2*]-....

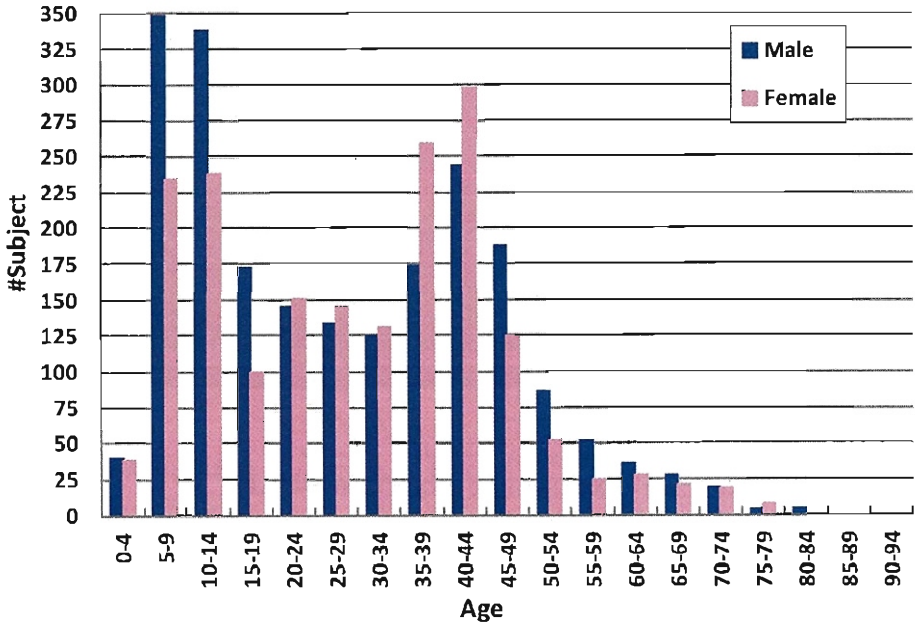


Figure 4.2: Distributions of the subjects' gender and age in **OULP-C1V1**.

Table 4.3: Breakdown of the number of subjects in **OULP-C1V1**

Dataset	Observation angle					Total
	55 [deg]	65 [deg]	75 [deg]	85 [deg]	All	
LP-C1V1-A	3,706	3,770	3,751	3,249	3,141	**3,835**
LP-C1V1-B	3,998	4,005	4,002	3,923	3,904	**4,007**

The dataset comprises two subsets, which we call **OULP-C1V1-A** and **OULP-C1V1-B**. **OULP-C1V1-A** is a set of two sequences (gallery and probe sequences) per subject and is intended for use in evaluating identification performance under almost constant walking conditions. **OULP-C1V1-B** is a set of one sequence per subject and is intended for use in investigating gait-based gender classification and age estimation. **OULP-C1V1-A** and **OULP-C1V1-B** are major upgrades to the datasets introduced in [126] and [135], respectively. For brevity, we omit the description of the dataset header "**OULP-C1V1-**".

Each of the main subsets is further divided into five subsets based on the observation angle (55 [deg], 65 [deg], 75 [deg], 85 [deg], and including all four angles) of each subject. We call these subsets **A/B-55**, **A/B-65**, **A/B-75**, **A/B-85**, and **A/B-ALL**, respectively, with each subject belonging to at least one of these subsets. The observation angle θ_s of each subject in each frame is defined by the y-axis of the world coordinate system (which is parallel to the walking

Figure 4.3: Examples of subjects in **OULP-C1V1**.

direction) and the line of sight of the camera as illustrated in Fig. 4.4.

A subject is included in a bin of a subset if one gait period occurs in the range of angles (as illustrated in Fig. 4.4) corresponding to that subset. For example, if a subject is recorded twice (both gallery and probe sequences) with a complete gait period in the range of 55 [deg], the subject is included in a bin of **A-55** and one of **B-55**. Moreover, if a subject is recorded twice with a complete gait period covering all the angle ranges, the subject is included in a bin of all the subsets. A gait period is calculated from the whole sequence (see Section 4.4.2 for details on the calculation of the gait period).

An example image for each observation angle is shown in Fig. 4.4, while a breakdown of the number of subjects is given in Table 4.3. In this table, the values in the "Total" column represent the number of subjects included in at least one of the subsets of 55 [deg], 65 [deg], 75 [deg], and 85 [deg]. As mentioned above, the numbers of subjects for dataset **A** represent those that have been recorded twice. Also, the differences between datasets **A** and **B** for each subset represent the numbers of subjects recorded only once. Take for example, the subset of 55 [deg] in Table 4.3 (**A-55** and **B-55**) where 3,706 subjects are recorded twice and 292 subjects are recorded only once. Note that there are also differences in the numbers of subjects between subsets, because the sequence length and observation angles for each subject are not exactly the same.

4.3.4 Advantages

Compared with existing gait databases, our dataset has the following strengths.

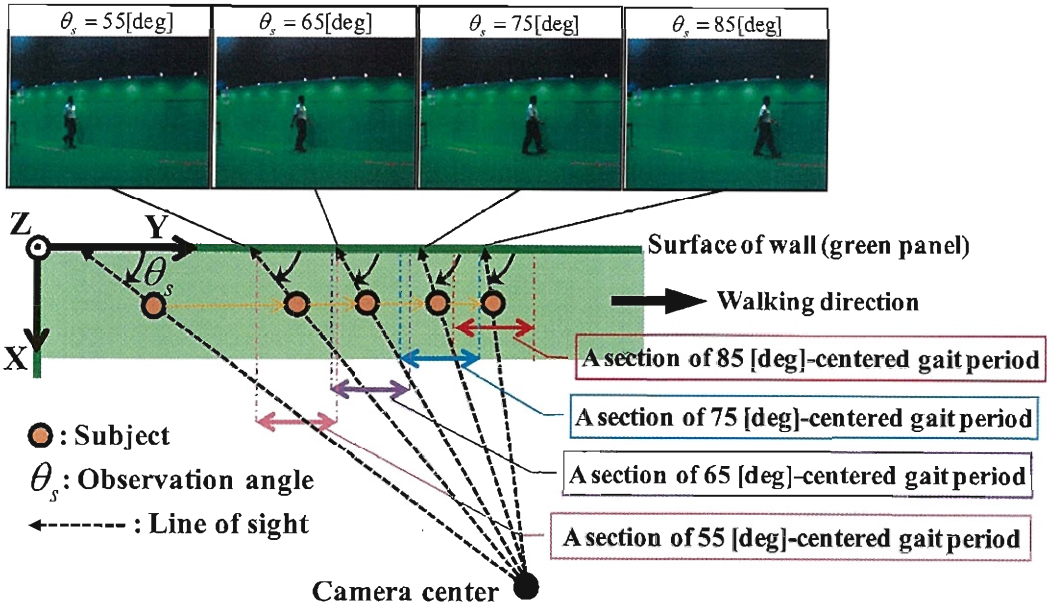


Figure 4.4: Definitions of the world coordinate system and the observation angle of a subject, and an example image at each observation angle. The Y–Z plane corresponds to the background wall behind the walking subjects, while the X–Y plane corresponds to the ground plane.

1. **Large population**: The number of subjects is more than 20 times that in publicly available large-scale gait databases. This improves the statistical reliability of various performance evaluations such as the comparison of gait-based person identification.

2. **Gender balance**: The ratio of males to females is close to 1. This is a desirable property for more reliable performance evaluation of gait-based gender classification and for comparison of identification performance between genders.

3. **Whole generation**: The age range is from 1 to 94 years with each 10-year interval up to 49 years of age containing more than 400 subjects (even in the smallest subset **A-ALL**). In addition, it is noteworthy that our dataset includes a sufficient number of children at all stages of growth, whereas other large-scale gait databases are mainly composed of adult subjects. This provides more statistically reliable results for gait-based age-group classification and comparisons of the difficulties in gait-based person identification among age groups.

4. **Silhouette quality**: The quality of each silhouette image is relatively high because we visually checked each silhouette more than twice and made manual modifications if necessary. This enables the elimination of silhouette quality problems from gait analysis.

On the contrary, the silhouette images in most of the existing public databases are automatically extracted and often include significant over/under-segmentation. Although manually modified silhouettes were created in the investigation of the effect of silhouette quality on gait-based person identification in [141] and [142], these have not been published.

4.3.5 Preprocessing

This section briefly describes the method used for size-normalized silhouette extraction.

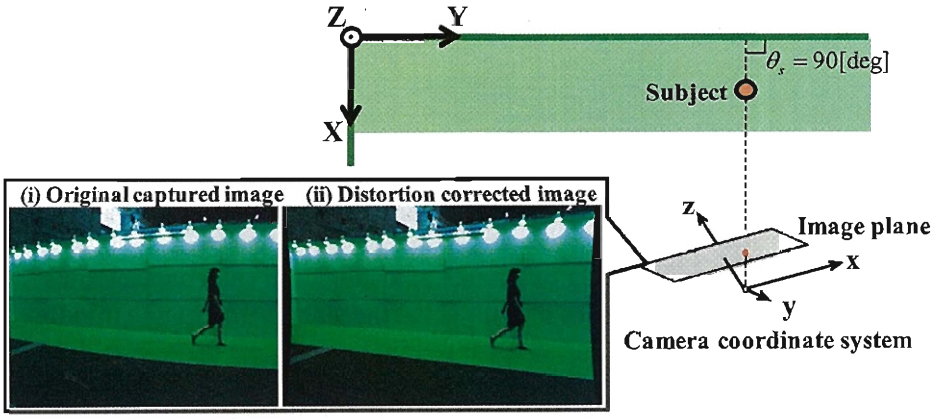
Silhouette extraction

The first step involved extraction of gait silhouette images via graph-cut-based segmentation [127] in conjunction with background subtraction. Of course, over/under-segmentation errors appeared in some extracted silhouette images. Hence, as described above, we visually checked all silhouette images at least twice and then manually modified under/over-segmentation if necessary. In more detail, a silhouette was shown to the observer in the form of a composite image in which the silhouette contour was overlaid on the corresponding original image. The observer checked whether the silhouette contour fitted the visually perceived human contour and if not, modified it using a GUI tool specially developed for this purpose.

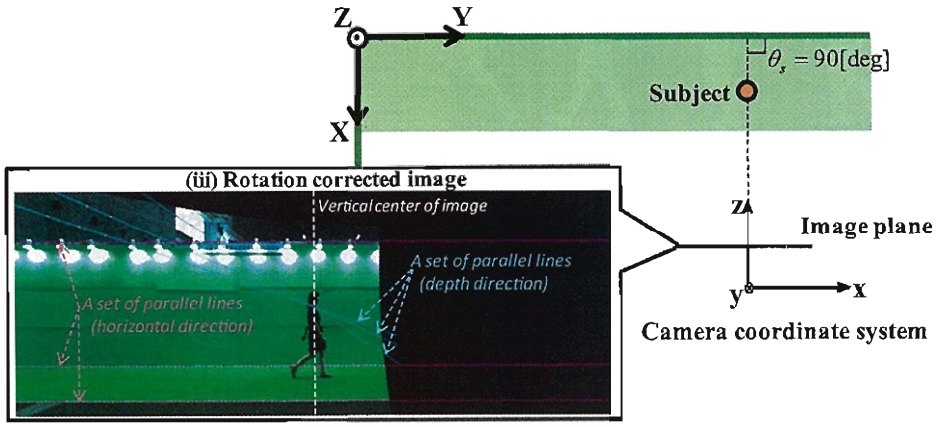
Correction of camera rotation

In the second step, image normalization, including the correction of distortion and camera rotation, was carried out. Because the camera pose in the world coordinate system for each day/event was not strictly the same, we normalized the camera rotations in all silhouette images such that the image plane in each is parallel with the Y–Z plane in the world coordinate system as shown in Fig. 4.5. First, the intrinsic parameters of the camera and coefficients of lens distortion were estimated [143][3] and distortion corrected. An example of an undistorted image is shown in Fig. 4.5(a). The transformation matrix of camera rotation from the original pose (shown in Fig. 4.5(a)) to the target pose (shown in Fig. 4.5(b)) was then estimated for each day/event from the undistorted image using a pair of vanishing points [144] (i.e., horizontal and vertical vanishing points), estimated from the sets of parallel lines in the scene [145]. Finally, all the image pixels in the original image plane were reprojected onto the normalized image plane. An example of a camera rotation corrected image is shown in Fig. 4.5(b). Also, examples of a subject in each dataset after rotation correction are shown in Fig. 4.6.

[3]Calibration procedures were implemented using OpenCV version 1.1 functions.



(a) Original image plane

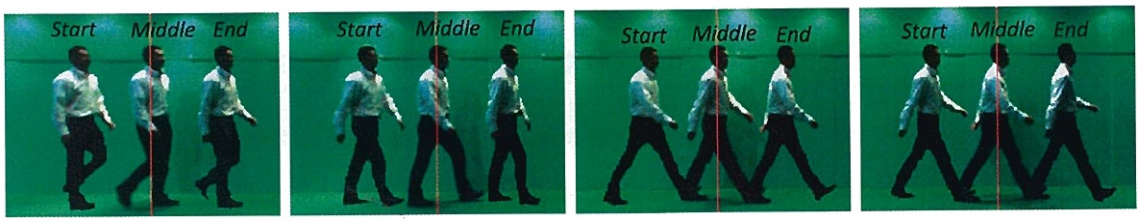


(b) Normalized image plane

Figure 4.5: Examples of the original and normalized camera pose, image plane, and images. In the rotation-corrected image in (b), the set of cyan lines and set of magenta lines represent the sets of parallel lines in the scene used to determine the vanishing points, while the white dashed line represents the vertical center line of the image. The observation angle is 90 [deg] at this line.

Registration and size normalization

The third step involved registration and size normalization of the silhouette images [41]. First, the top, bottom, and horizontal center of the silhouette regions were obtained for each frame. The horizontal center was chosen as the median of the horizontal positions belonging to the region. Second, a moving-average filter was applied to these positions. Third, we normalized the size of the silhouette images such that the height was just 128 pixels according to the average positions, and the aspect ratio of each region was maintained. Finally, we produced an 88 $\times$ 128 pixel image in which the average horizontal median corresponds to the horizontal center of the image. Examples of size-normalized silhouettes are shown in Fig. 4.7.



(a) **A-55** (b) **A-65** (c) **A-75** (d) **A-85**

Figure 4.6: Composite images showing examples of a subject in each dataset after rotation correction. Each composite image includes the subject at the start ("*Start*"), the middle ("*Middle*"), and the end ("*End*") of the section. The vertical red line represents the center of the section.

Figure 4.7: Examples of size-normalized gait silhouettes (every four frames).

4.4 Gait-based Person Identification

This section describes a framework for performance evaluation of gait-based person identification.

4.4.1 Gait Features

The current trend in gait representation is appearance and period-based representation, such as the averaged silhouette [35], also known as the Gait Energy Image (GEI) [36]. In this paper, we deal with six such state-of-the-art gait features: GEI, Frequency-Domain Feature [41] (referred to as FDF in this paper), Gait Entropy Image (GEnI) [146], Masked GEI based on GEnI [37] (referred to as MGEI in this paper), Chrono-Gait Image (CGI) [147], and Gait Flow Image (GFI) [148].

The GEI is obtained by averaging silhouettes over a gait cycle, while the FDF is generated by applying a Discrete Fourier Transform of the temporal axis to the silhouette images in a gait cycle. In this work, 0, 1, and 2 times frequency elements are used. The GEnI is computed by calculating Shannon entropy for every pixel over a gait cycle, where the value of the GEI is regarded as the probability that the pixel takes the binary value. The MGEI is computed by masking the GEI with a pair-wise mask generated by each pair of probe and gallery GEnIs. The GEnI and MGEI aim to select the dynamic area from the GEI. The CGI is a temporal

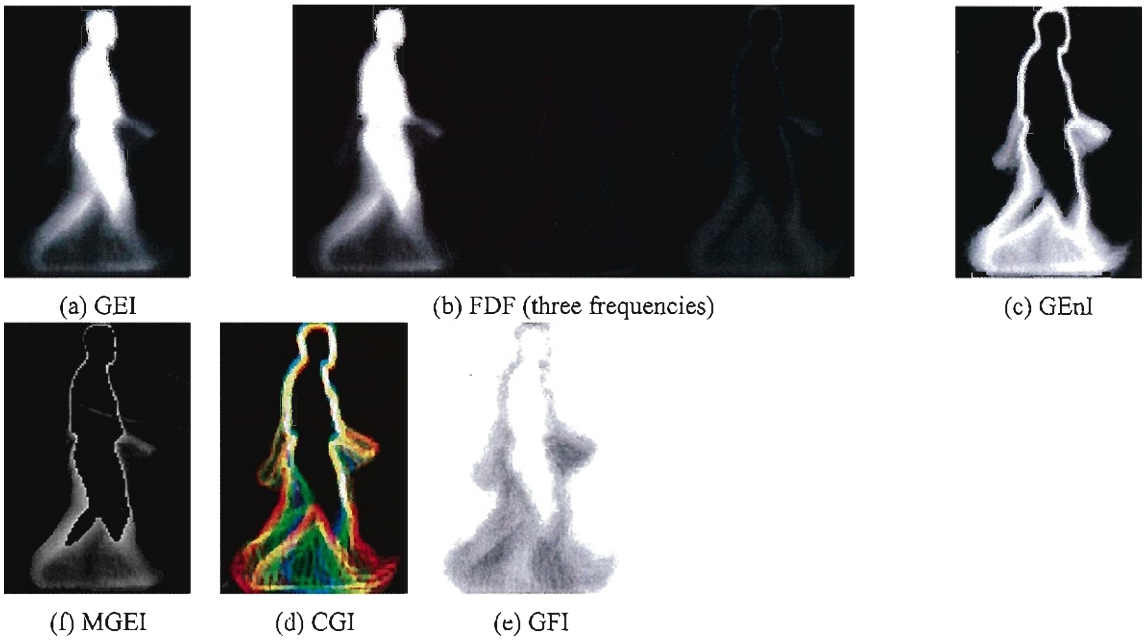


Figure 4.8: Examples of gait features.

template in which the temporal information among gait frames is encoded by a color mapping function, and is obtained by compositing the color encoded gait contour images in a gait cycle. The GFI is based on an optical flow field from silhouettes representing motion information and is created by averaging the binarized flow images over a gait cycle. An example of each feature is shown in Fig. 4.8.

4.4.2 Gait Period Detection

For the quantification of periodic gait motion, we adopted the Normalized Auto Correlation (NAC) of the size-normalized silhouette images for the temporal axis:

$$C(N) = \frac{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n)g(x,y,n+N)}{\sqrt{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n)^2}\sqrt{\sum_{x,y}\sum_{n=0}^{T(N)} g(x,y,n+N)^2}} \tag{4.1}$$

$$T(N) = N_{total} - N - 1, \tag{4.2}$$

where $g(x,y,n)$ is the silhouette value at position (x,y) of the n-th frame, $C(N)$ is the autocorrelation for the N-frame shift, and N_{total} is the total number of frames in the sequence. Because gait is a symmetrical motion to some extent, peaks of the NAC were assumed to appear for all half periods on the temporal axis. Thus, we determined the gait period N_{gait} as the frame shift corresponding to the second peak of the NAC. An example of the relation between NAC and frame shift is shown in Fig. 4.9.

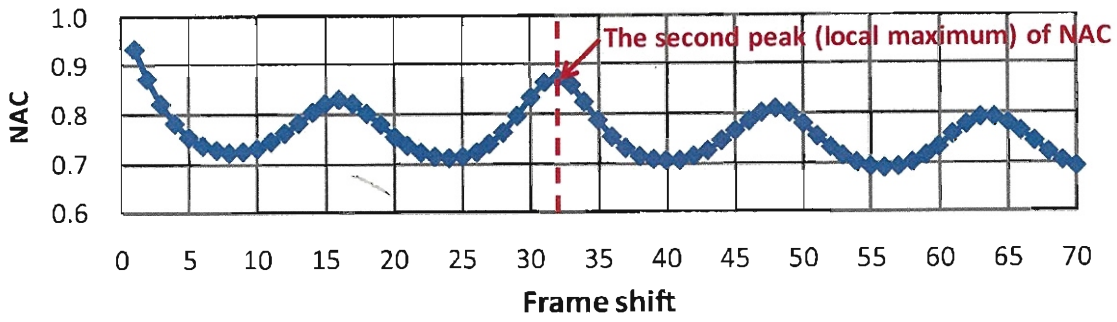


Figure 4.9: Example of the relation between NAC and frame shift. In this figure, the frame shift corresponding to the second peak of the NAC is 32.

4.4.3 Distance Matching

In the evaluation of datasets **A-55**, **A-65**, **A-75**, and **A-85**[4], a gait feature for a subject was created from a section of a dataset (as illustrated in Fig. 4.6) that includes one gait period. Note that there is some area of overlap for some subjects between sections as shown in Fig. 4.4. All pairs of features (gallery and probe features) were then directly matched[5].

The distance $D_{ij,K}$ between the i-th probe subject and the j-th gallery subject in dataset $K \in \{\textbf{A-55}, \textbf{A-65}, \textbf{A-75}, \textbf{A-85}\}$ was measured as,

$$D_{ij,K} = \| \mathbf{P}_{i,K} - \mathbf{G}_{j,K} \|_2, \tag{4.3}$$

where $\mathbf{P}_{i,K}$ and $\mathbf{G}_{j,K}$ are feature vectors of the i-th probe and j-th gallery in dataset K, respectively, and $\| \cdot \|_2$ is the Euclidean distance. In addition, we exploited z-normalization [149] of the distance among galleries for each probe to improve the performance in a one-to-one matching scenario.

For dataset **A-ALL**, we first calculated z-normalized distances for each section of the four abovementioned datasets and then averaged them as a total distance. Note that this averaging is equivalent to combining the normalized scores via the sum rule [150].

[4]Because two sequences (gallery and probe sequences) are required for person identification, dataset **A** is used hereafter.

[5]Since only a single gait feature was obtained for each dataset, statistical discriminant analysis considering within-class variance such as linear discriminant analysis could not be applied.

4.5 Performance Evaluation of Gait-based Person Identification

Despite the recent welcome development in gait-based person identification in the research community, the following open issues still remain.

1. An evaluation of gait-based person identification with statistical reliability has not been carried out owing to the lack of a large population dataset.

2. Also, to the best of our knowledge, the effects of gender and age on identification performance have not been explored because of the lack of a dataset with sufficient subject diversity.

Therefore, we address the above issues using our dataset. In this section, we first show the statistical reliability of the evaluation using our database. The upper limits of identification performance of state-of-the-art gait representations introduced in the previous section are then demonstrated. Finally, we reveal the effects of age and gender on identification performance.

4.5.1 Effect of the Number of Subjects

First, the effect of the number of subjects is demonstrated by means of a Receiver Operating Characteristic (ROC) curve. The ROC curve is a common tool for performance evaluation in biometrics and denotes the trade-off between the False Rejection Rate (FRR) and False Acceptance Rate (FAR) when the acceptance threshold is changed by a receiver in a one-to-one matching scenario.

From statistical analysis of ROC curves [151], the standard deviation of the FRR with a single probe for each subject is estimated as

$$\hat{\sigma}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \tag{4.4}$$

where $\hat{p}$ is the observed FRR and n is the number of subjects. This indicates that the obtained FRR becomes more reliable as the number of subjects increases.

To validate the estimation, we repeated the experiments with randomly chosen subsets with fewer subjects and compared the actual standard deviation of the performance and that estimated from Eq. (4.4) using the GEI as the gait feature. First, we prepared 100 subsets comprising 100 subjects randomly chosen from dataset **A-65** (which comprises 3,770 subjects) and obtained 100 ROC curves from the experimental results. We then calculated the average and

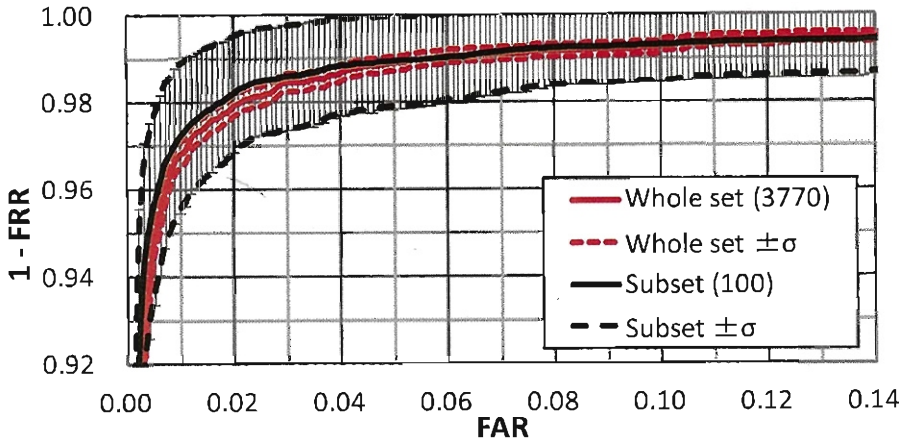


Figure 4.10: ROC curves of gait-based person identification using GEI with a varying number of subjects. Black and red indicate, respectively, smaller subsets and the whole set of **A-65**. The bold line and two bounding dashed lines indicate, respectively, the average μ and standard deviation range $\mu \pm \sigma$ derived from Eq. (4.4). Gray bars denote the standard deviation ranges $\mu \pm \sigma$ obtained in the experiments.

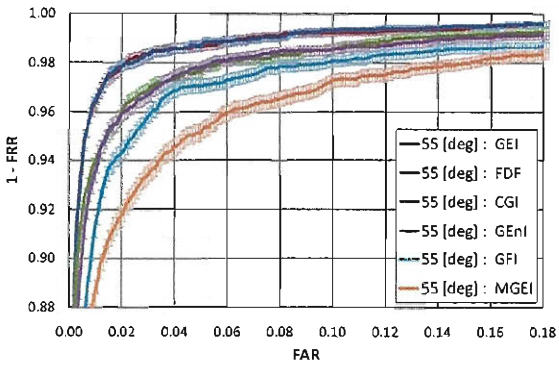
standard deviation of the FRR for each FAR, depicted as an averaged ROC curve (bold black line) and standard deviation range bar (gray bar) in Fig. 4.10. Additionally, the estimated standard deviation range is depicted as two dashed black lines. From the graph, we see that the standard deviation ranges derived from the experimental results correspond well with those estimated from Eq. (4.4).

In addition, the results for the whole set are superimposed as the bold red line, while the standard deviation range estimated from Eq. (4.4) is depicted as two dashed red lines in Fig. 4.10. We see that the standard deviation range is significantly narrower than that of subsets with fewer subjects.

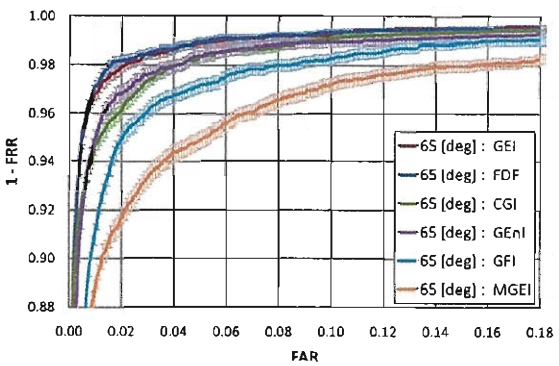
4.5.2 Comparison of the Gait Feature

Performance comparison

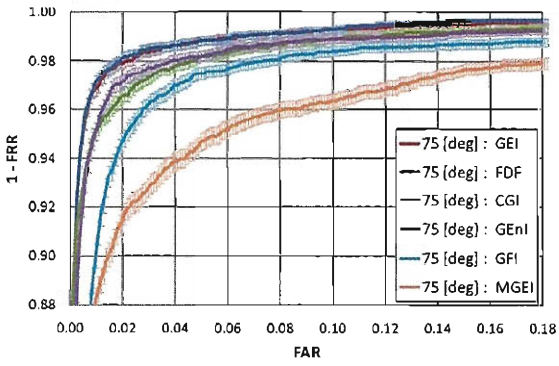
This section compares the identification performance of the six gait features described in Section 4.4.1. The identification performance was evaluated using two metrics: (1) the ROC curve, and (2) the rank-1 and rank-5 identification rates. The rank-1 and rank-5 identification rates, which are common evaluation measures in a one-to-N matching scenario, denote the percentages of correct subjects out of all the subjects appearing within the first and fifth ranks, respectively. Note that the rank-1 and rank-5 identification rates depend on the gallery size, whereas the ROC curve is essentially independent of the gallery size.



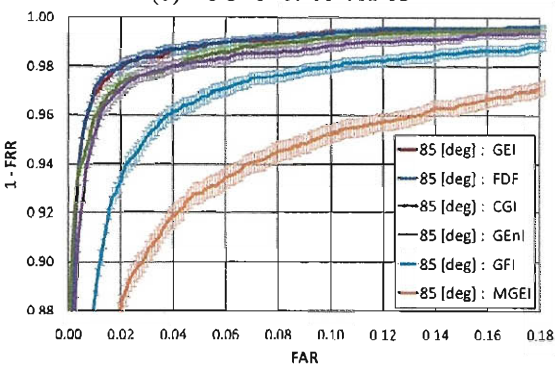
(a) ROC for dataset **A-55**



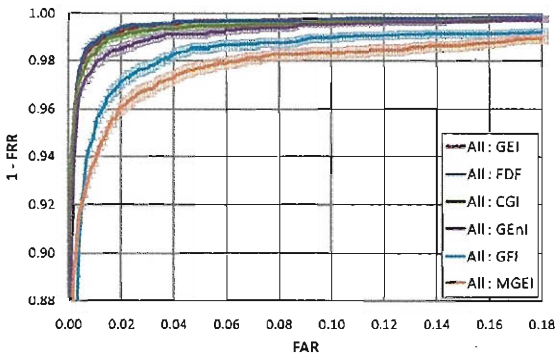
(b) ROC for dataset **A-65**



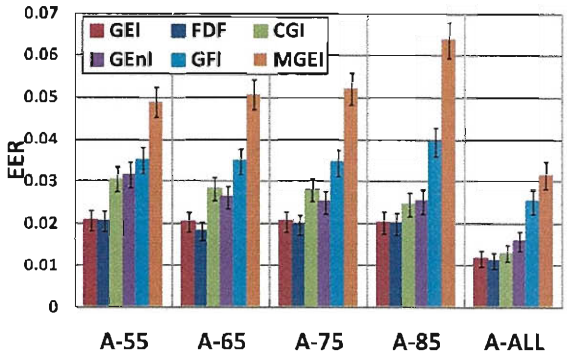
(c) ROC for dataset **A-75**



(d) ROC for dataset **A-85**



(e) ROC for dataset **A-ALL**



(e) EERs for all the datasets

Figure 4.11: Performance comparison of six gait features in terms of the ROC curve and EER. Each bar represents a standard deviation range derived from Eq. (4.4).

Table 4.4: Performance comparison of six gait features in terms of the rank-1 identification rate.

Dataset	GEI	FDF	GEnI	CGI	GFI	MGEI
A-55	**84.70**	83.89	76.42	75.58	75.15	68.35
A-65	**86.63**	85.49	78.65	78.97	77.11	68.91
A-75	**86.91**	86.59	79.95	81.58	76.54	67.10
A-85	85.72	**85.90**	80.95	83.35	74.92	61.19
A-All	**94.24**	94.17	90.93	91.60	87.46	84.18

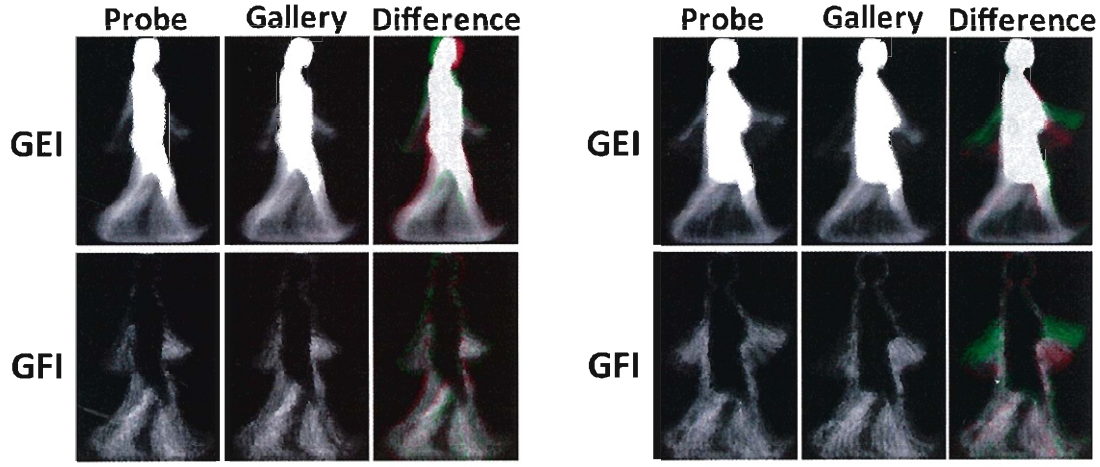
Table 4.5: Performance comparison of six gait features in terms of the rank-5 identification rate.

Dataset	GEI	FDF	GEnI	CGI	GFI	MGEI
A-55	**92.39**	91.53	86.67	86.02	85.83	80.09
A-65	**92.84**	92.81	88.14	88.06	87.32	79.71
A-75	92.78	**92.88**	89.23	89.28	85.84	78.41
A-85	**93.01**	92.83	89.60	90.80	84.73	73.19
A-All	**97.13**	97.10	95.35	95.32	92.84	90.58

First, the performance is compared for each observation angle using datasets **A-55**, **A-65**, **A-75**, and **A-85**, since the gait feature property is dependent on the observation angle[6]. The ROC curves with standard deviation range bars for each dataset are shown in Figs. 4.11(a), (b), (c), and (d), while the Equal Error Rate (EER) is summarized in Fig. 4.11(f). In addition, rank-1 and rank-5 identification rates are given in Table 4.4 and Table 4.5. From the results, although the performances of the GEI and FDF are nearly equal and the performances of the GEnI and CGI are nearly equal, we see that there is a statistically significant performance difference between the GEI (or FDF), GEnI (or CGI), GFI, and MGEI, and the performance order of these techniques is almost independent of the observation angle.

Next, we compare the total performance using dataset **A-ALL**, with the results shown in Fig. 4.11(e), Table 4.4, and Table 4.5 (bottom row). As for the results for **A-ALL**, the following reasons are suggested for the improvement in identification performance: a) the effect of gait fluctuations, which notably appears on the arm swing and head pose, was decreased by combining the scores of each observation angle, and b) the variations in the gait feature property caused by the observation angle improved the identification performance, as reported in [90]. From these results, it can be seen that the GEI and FDF achieve the best performance overall.

[6]For example, static features such as body shape are clearly seen in front-view gait images, while dynamic features such as the step and arm swing are clearly seen in side-view gait images.



(a) A subject with large head pose fluctuation

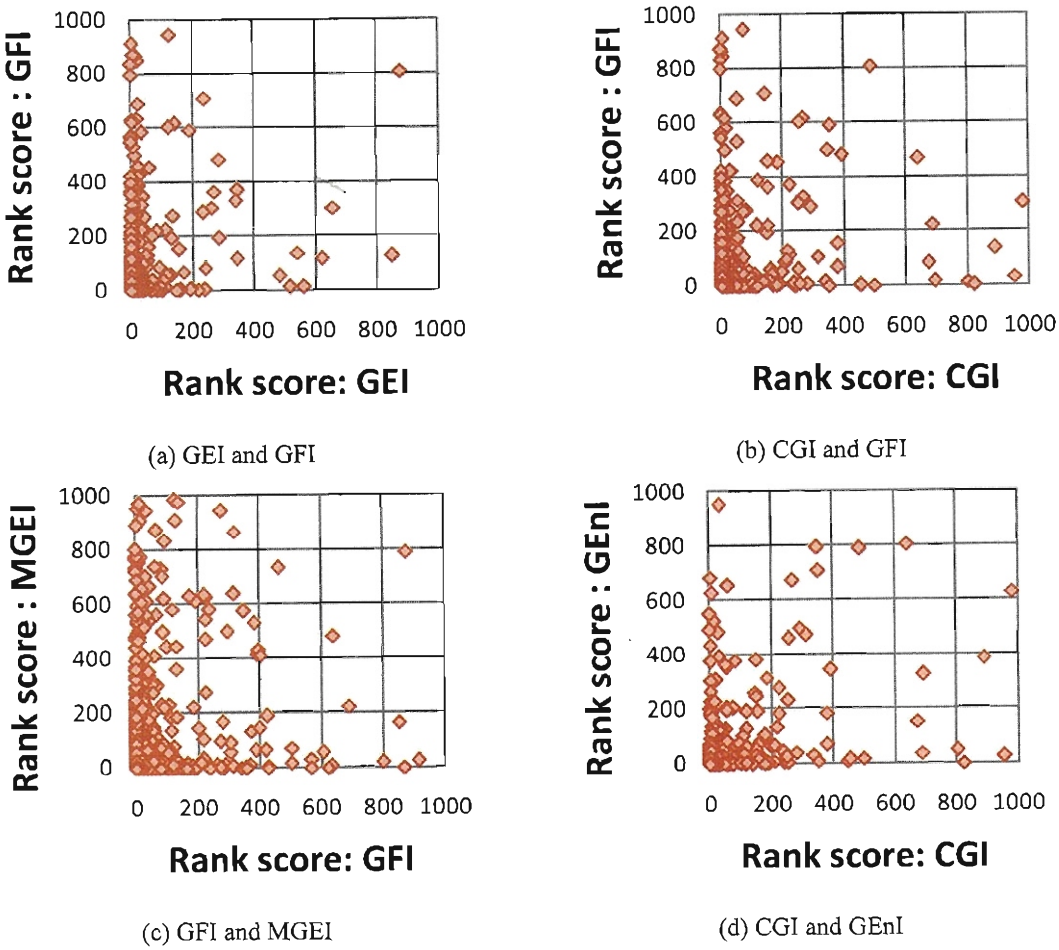
(b) A subject with large arm swing fluctuation

Figure 4.12: Examples of subjects in **A-85**. Note that the value of the GFI is inverted for visibility. All feature differences between gallery and probe features are visualized by colors (green and red) in the corresponding difference image. Green indicates that the probe feature appears more strongly, while red depicts the opposite. Regarding the subject in (a), the rank score using the GEI is 216, while that using the GFI is 1 (also, the rank scores are 79, 239, 7, and 19 using the FDF, CGI, GEnI, and MGEI, respectively). On the other hand, for the subject in (b), the rank score using the GEI is 1, while that using the GFI is 567 (also, the rank scores are all 1 using the other features).

Note that these comparison results are partly inconsistent with the results in previous works, for example, [147] (GEI vs. CGI) and [148] (GEI vs. GFI). The differences between the databases used for the evaluations (e.g., subject diversity, silhouette quality, sequence length, and intra-subject variations) are considered to be the cause of the inconsistencies. For example, according to the latest evaluation results of the CGI reported in [38], GEI performance is superior to that of CGI only if there is no intra-subject variation and only a single gait period occurs in a sequence. Both these conditions are true in our dataset.

Correlation among features

Although some kind of upper limit on identification performance using state-of-the-art gait features has been shown in the previous section, investigating the correlation among gait features is still meaningful for the design of a feature fusion scheme [152] to further improve identification performance. Each gait feature has a unique property and is considered to be independent of other features to some extent. For example, Fig. 4.12(a) shows a subject in **A-85** whose rank score is 216 using the GEI and 1 using the GFI. On the other hand, Fig. 4.12(b) shows a subject in **A-85** whose rank score is 1 using the GEI and 567 using the GFI. These typical examples



(a) GEI and GFI

(b) CGI and GFI

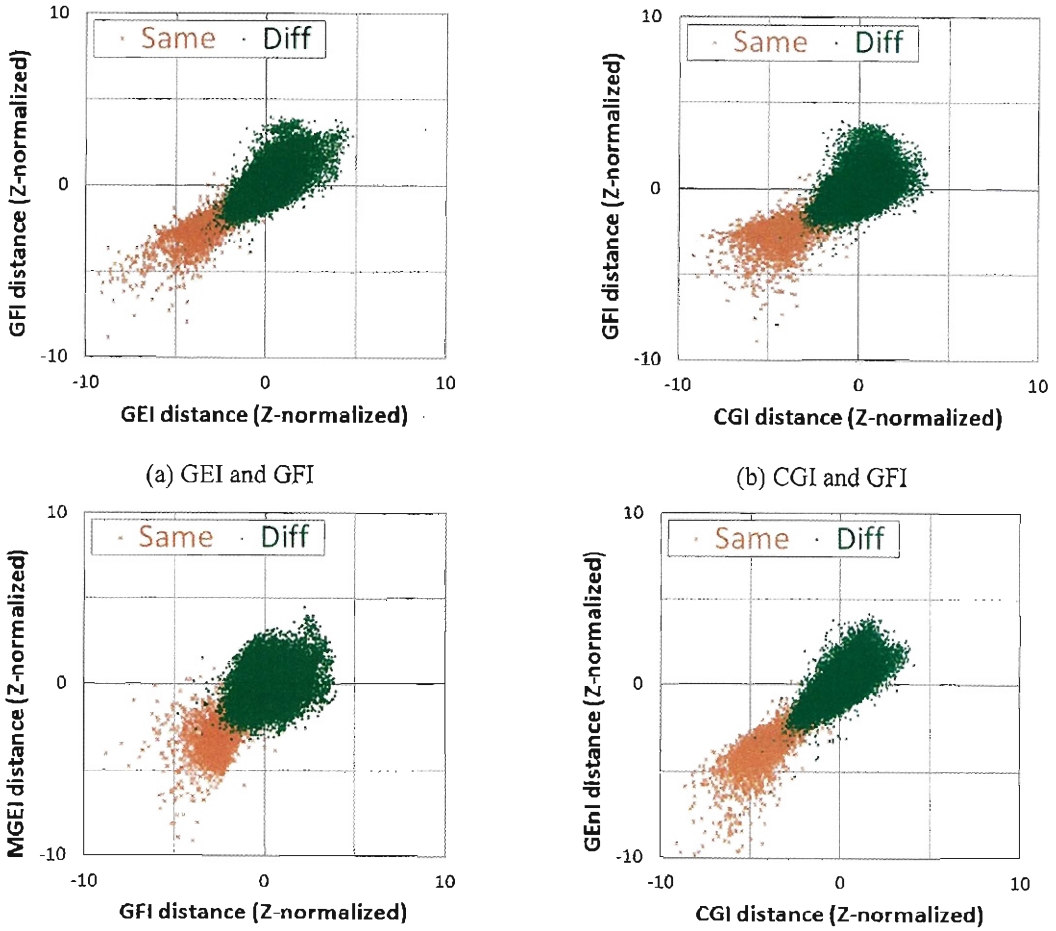
(c) GFI and MGEI

(d) CGI and GEnI

Figure 4.13: Examples of the rank score correlations between pairs of gait features.

indicate that the GEI is relatively sensitive to static pose fluctuations and robust to motional fluctuation, while the GFI is the exact opposite.

To reveal which pair of features has a weak correlation, that is, is suitable for fusion, the rank score relations among gait features for each subject were analyzed. The results show that the GFI has relatively weak correlation with all the other features except the GEnI, and the CGI has the same with the GEnI, MGEI, and GFI. In addition, the GEnI has the same with the GEI, CGI, and MGEI. Some notable relations of rank scores for dataset **A-85** among these features are shown in Fig. 4.13, while the relations of distances of the same subjects and different subjects are shown in Fig. 4.14. In the distance distributions shown in Fig. 4.14, though we can see that each distance relation between each pair of features is correlated as a whole, dispersal exists at a certain level. Therefore, these figures indicate that there is room for improvement in the identification performance by fusing these gait features. Demonstration of this through fusion is one of our future works.



(a) GEI and GFI

(b) CGI and GFI

(c) GFI and MGEI

(d) CGI and GEnI

Figure 4.14: Examples of the distance correlations between pairs of gait features.

4.5.3 Effects of Gender and Age

This section investigates the difference in gait-based person identification performance between genders and age groups. Our gait database is suited to this purpose because the age distribution of each gender is much wider than that in existing gait databases as mentioned in Section 4.3.3. In this experiment, we adopted the GEI as the gait feature and carried out the evaluation on subset **A-65**, since it has the largest number of subjects in dataset **A**[7].

Ages were grouped in 5-year intervals up to 20 years of age and in 10-year intervals from 20 to 60 years of age for each gender[8]. Ages over 60 years were treated as one age group because of the shortage of subjects. The numbers of subjects in each gender/age group are given in

[7]We also carried out this experiment using all the other gait features described in Section 4.4.1 on another subset **A-85**, but the results showed similar trends.

[8]Taking the rapid physical growth rate into consideration, we used 5 year intervals up to 20 years to reveal more detailed changes in identification performance during the growing process.

Table 4.6: Numbers of subjects of each gender and age group in **A-65**

	Age								
Gender	0–4	5–9	10–14	15–19	20–29	30–39	40–49	50–59	Over 60
Male	32	312	323	150	267	288	412	134	89
Female	29	213	226	90	285	369	403	73	75

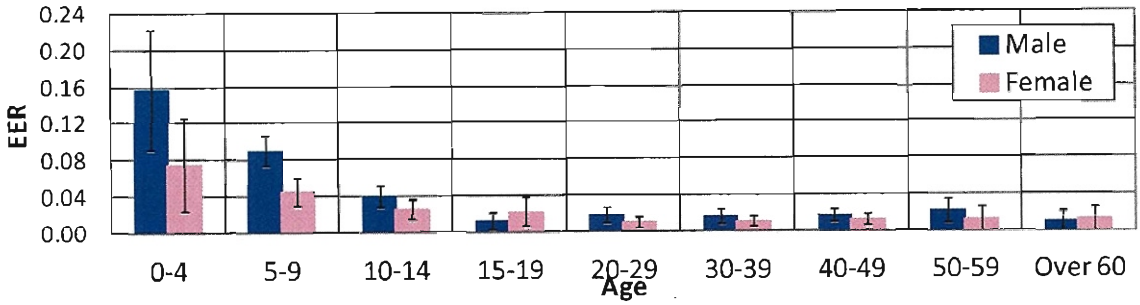


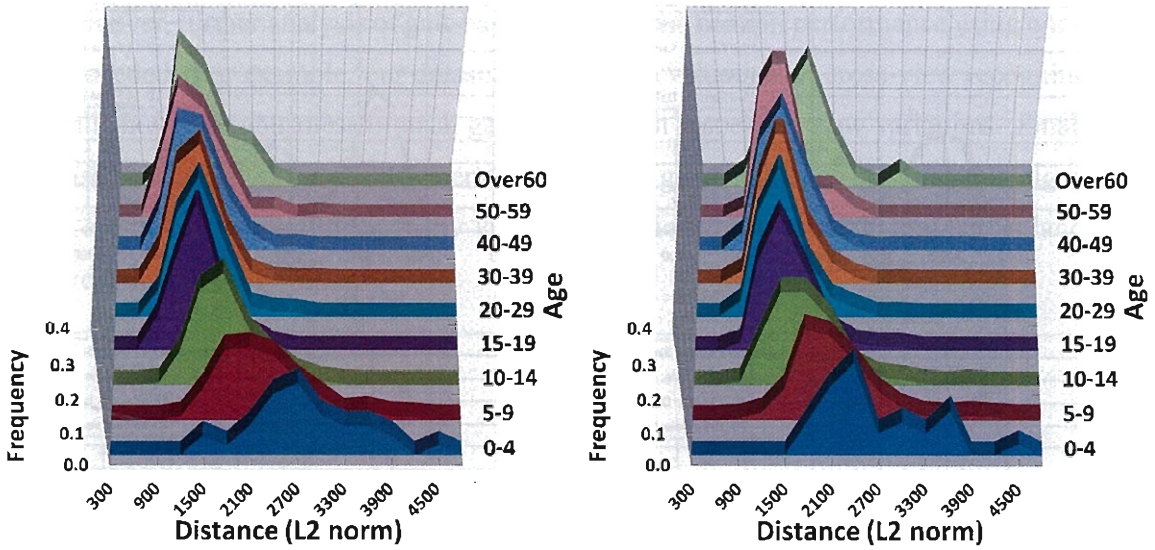
Figure 4.15: EERs among genders and age groups. The black bars represent the standard deviation ranges derived from Eq. (4.4).

Table 4.6.

The EER for each gender/age group is depicted in Fig. 4.15, while the distance distributions of the same subjects (true attempts) for each group are shown in Fig. 4.16. A comparison of the distance distributions of the same subjects and different subjects (imposters) for four typical age groups—under 10s (5 to 9 years old), early 10s, 20s, and 40s—are depicted in Fig. 4.17. Note that the original L2 norm (Eq. (4.3)) is shown in these distributions.

Effect of gender

First, we focus on the difference in gait-based person identification performance between males and females. According to the results in Fig. 4.15, identification performance for females tends to be better than that for males in almost all the age groups. Additionally, Fig. 4.17 implies that the inter-subject variation in females' gait is greater than that in males' gait, while intra-subject variations are not that different between males and females in each age group. The difference in intra-subject variation is assumed to be due to the fact that the range of appearance variation in females, which mainly comes from variations in hair style, clothes, and shoes, is greater than that in males.



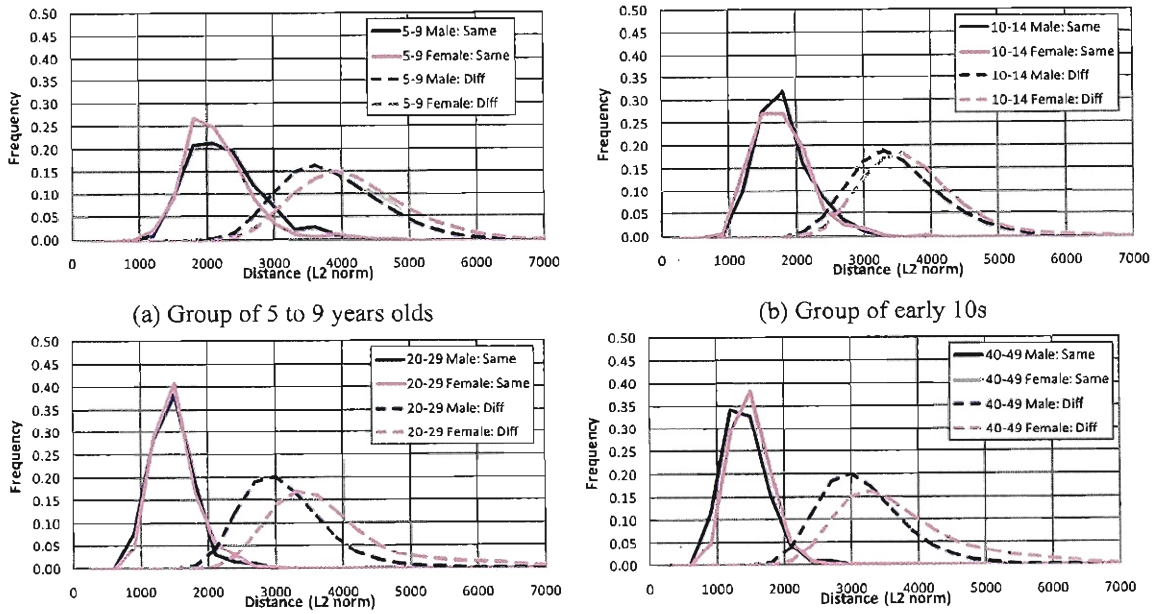
(a) Male (b) Female

Figure 4.16: Distance distributions of the same subjects in each gender/age group.

Effect of age

Next, we focus on the difference in gait-based person identification performance between age groups. From the results in Fig. 4.15, we see that identification performance for the group of very young children (0 to 4 years old) is worse than that for the other age groups, and this gradually improves with the slightly older groups up to the group of late 10s. This result is intuitively understandable because the intra-subject gait fluctuation for children is greater owing to the immaturity of their walking, as illustrated in Fig. 4.16. On the other hand, the fluctuation in gait for adults is small as shown in Fig. 4.16. This indicates that adults have established their own walking style; in other words, they have a fixed gait pattern. In this regard, however, the intra-subject variation in the over-60 female group is slightly larger than that in other adult age groups, and this is assumed to be due to a decline in physical strength with aging. Further study of elderly groups (over 60 years old), together with the additional data collection required, is considered as future work. In addition, a more detailed analysis of gait properties among age groups, such as investigation of the differences in the effects of body parts among age groups, is one of our future works.

The above observations indicate that the dependence of gait fluctuation on the age group implies that gait fluctuation can be a useful cue for age classification according to gait. In addition, the age group can be regarded as a so-called quality measure [153] for gait-based identification, which is one of the interesting future directions of this study.



(a) Group of 5 to 9 years olds

(b) Group of early 10s

(c) Group of 20s

(d) Group of 40s

Figure 4.17: Comparison of distance distributions of the same subjects and different subjects in four typical age groups.

4.6 Conclusions

This paper described the construction of a gait database comprising a large population dataset and presented a statistically reliable performance evaluation of gait-based person identification. This dataset has the following advantages over existing gait databases: (1) the number of subjects is 4,007, which is more than 20 times greater than the number in existing public large-scale databases, (2) the male-to-female ratio is close to 1, (3) the age distribution is wide, ranging from 1 to 94 years, and (4) the quality of all silhouettes is guaranteed by visual confirmation. Using our dataset, we carried out a statistically reliable performance comparison of gait-based person identification using state-of-the-art gait features. Moreover, the dependence of identification performance on gender and age group was analyzed with the results providing several new insights, including the performance difference between males and females, and the gradual change in identification performance with human growth.

Although our dataset has the largest population of all databases till now, there is still an insufficient number of very young children and elderly persons when compared with the numbers of other generations. Therefore, we need to collect the required gait datasets by taking advantage of various events, such as outreach activities, in the future. Additionally, the construction of another dataset using images taken with camera 2 is a future work.

Moreover, further analysis of gait-based person identification performance using our dataset is still needed. For example, our dataset enables the evaluation of cross-view recognition and this will show the robustness of each gait feature with respect to view variations. Finally, our database is suitable for the development of gait-based gender and age classification algorithms, which are quite meaningful for many vision applications such as intelligent surveillance, and these remain as future works.

Chapter 5

Conclusion

This thesis describes the techniques for visual surveillance using gait-based person identification. Gait as a biometric cue has the ability of identifying individuals at a distance and is difficult to disguise. Due to such advantages, gait-based person identification could enhance the value of surveillance in public space, furthermore, it could contribute a lot to the realization of safe and secure society. In this research, we tackled the three major issues of gait-based person identification toward the practical use in visual surveillance.

First, we addressed an issue of accurate foreground segmentation which is primal preprocessing of gait-based person identification, and we proposed a novel framework of foreground and shadow segmentation with a static binocular system. We aimed more accurate segmentation in the presence of strong shadow and occlusion relationship between two cameras. The homography constraint was utilized to distinguish the foreground and shadow. In addition, while existing homography-based approaches did not consider the occlusion relationship between foreground and shadow and often failed at such regions, it was taken into account by treating homography-correspondence pairs symmetrically in the proposed method, and the segmentation problem was regarded as a multi-labeling problem for a homography-correspondence pair. In the labeling strategy, the labels which represent the correspondence of shadow and background were prohibited by homography constraint. The multi-labeling problem was formulated in the framework of energy minimization in which the energy function was composed of a data term and a smoothness term. The data term contributed to the label assignment in terms of the color of each pair of homography-correspondence pair, and the smoothness term did in terms of spatio-temporal label continuity, and the energy minimization problem was optimized by α–β swap algorithm. The performance of the proposed method was examined through the comparison experiments with color-based approach, disparity-based approach, existing homography-based approach, and their integration approaches. Three different real image sequences with

strong shadow and occlusion relationship were used as the evaluation data. The results showed that the proposed method overall outperformed the existing methods.

Second, an issue of identification performance decrement caused by intra-subject variations such as gait fluctuation and condition changes (e.g., view, clothing, and carrying conditions) is addressed. We focused on the social context that we often act in groups (e.g., friends, family, and co-workers). Such group context was used as a cue for identifying individuals to enhance the identification performance in the presence of the degradation of individual cue caused by intra-subject variations. In surveillance videos, a group often observed with non-group members at a time and the spatial relations among them are dynamically changed with time. To consider such dynamic relations among persons, we used the behavioral relations as group context including spatial distance and velocity vector difference among persons through the video sequences, while existing group context-assisted person identification approaches just used co-occurrences among persons as group context. The group context was integrated with individual cue in the form of a pair-wise CRF model. In the model, the person-to-person relationships were represented as a graph, where each node corresponds to each person and each edge corresponds to the relationship between each pair of persons. A person identification problem was then formulated as a maximization problem of the conditional probability of the label assignment for all the nodes, and the problem was approximately solved by LBP algorithm. In the iteration of the message passing process of LBP, the identity confidence for each person was propagated to the identities of the surrounding persons based on their individual cues and group information, so that the same group members with similar characteristics (spatial proximity and similar velocity vector) enhanced each other's identities. We conducted the experiments to confirm the effectiveness of the proposed method by using both a real image dataset composed of 47 subjects and the simulation datasets composed of a thousand subjects. In the experiments, gait is used as a biometric cue, and the significant improvement in identification performance was demonstrated by the proposed method through all the experiments compared with the straightforward method based on individual cues alone.

Third, we addressed an issue of the performance evaluation of gait-based person identification with statistical reliability. To solve the issue, we constructed the world's largest gait database composed of 4,007 subjects (2,135 males and 1,872 females) with ages ranging from 1 to 94 years, while existing major databases included at most 185 subjects with biased gender and age distributions. In addition, the quality of each silhouette was ensured by visual check and manual modification, and the observation angle of each subject in each frame was

specifically defined in our database to enable the fair performance evaluation in terms of silhouette quality and observation angle. By using the database, first, we demonstrated that the dataset ensured the statistical reliability of an performance evaluation result, comparing with that from the dataset comprising 100 subjects. Then, we carried out the performance comparison of state-of-the-art model-free gait representations: GEI, FDF, GEnI, MGEI, CGI, and GFI. The results showed a kind of their upper limits of identification performance and the significant performance differences among them. The results also showed that the GEI and FDF achieved the best performance of all in total. Finally, we analyzed the dependence of identification performance on gender and age group. As the results, several new insights such as the performance difference between males and females and the gradual change in performance with human growth were provided.

In the future, the construction of a total scheme of gait-based person identification for visual surveillance is required for the practical use, by integrating each of the techniques proposed in this thesis and other state-of-the-art techniques of segmentation, tracking, and gait representation. Although we focused on the three issues, there are some remained critical issues to be solved. Of these, the occlusion among persons, which often arises in the surveillance videos, especially in crowded scenes, is one of the most challenging problems not only from the view point of preprocessing, but also from that of identification.

Segmentation and tracking of persons as preprocessing in such occlusion relationship is inevitable to obtain each gait of them. Some recent works [130, 131, 154] address the issue, and it is considered that the use of behavioral context among persons as described in this thesis can enhance their performance of segmentation and tracking, as demonstrated in the studies of context-based tracking [121, 122]. On the other hand, even if the segmentation and tracking are done well by such state-of-the-art techniques, an issue of identification via incomplete gait features created from the partial silhouettes of the occluded persons is still remained. For the issue, gait-based person identification approaches using partial gait representations [155, 156, 42, 157] could ensure the identification performance to some extent. Moreover, the incorporation of these methods into the group context-aware framework proposed in this thesis will much improve the identification performance in such case.

As for the performance evaluation, though the world's largest gait database in terms of the number and diversity of subjects is constructed in this thesis, it includes no walking variation. Therefore, the other challenging issue is the construction of more expansive gait database which includes both a number of walking variations (e.g., views, clothing, shoes, speeds, surfaces, and carrying conditions) and a number of subjects with wide ranging ages like that in our database.

Such database will encourage the development of the walking condition-invariant gait-based person identification scheme with statistical reliable robustness toward practical use.

Only after gathering these techniques, gait-based person identification can be applied for various practical scenes in visual surveillance and will be able to contribute to the safe and secure life in a real sense.

Reference

[1] V. Gouaillier and A. Fleurant, "Intelligent video surveillance: Promises and challenges," *Technological and Commercial Intelligence Report*, 2009. [Online]. Available: http://www.crim.ca/fr/r-d/vision_imagerie/documents/CRIM_TDS_TechnoCommercialIntelligenceReportVideoSurveillance.pdf

[2] R. Collins, A. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 745–746, 2000.

[3] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 34, pp. 334–352, 2004.

[4] M. Valera and S. Velastin, "Intelligent distributed surveillance systems: a review," in *IEE Proceedings Vision, Image and Signal Processing*, vol. 152, 2005, pp. 192–204.

[5] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking," *IEEE SIGNAL PROCESSING MAGAZINE*, vol. 22, pp. 38–51, 2005.

[6] S. Gong, C. C. Loy, and T. Xiang, "Security and surveillance," in *Visual Analysis of Humans*. Springer, 2011, ch. 23, pp. 455–472.

[7] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[8] S. G. Kong, J. Heo, B. R. Abidi, J. Paik, and M. A. Abidi, "Recent advances in visual and infrared face recognition: a review," *Computer Vision and Image Understanding*, vol. 97, pp. 103–135, 2005.

[9] A. F. Abate, M. Nappi, D. Riccio, and G. Sabatino, "2d and 3d face recognition: A survey," *Pattern Recognition Letters*, vol. 28, pp. 1885–1906, 2007.

[10] X. Zhang and YongshengGao, "Face recognition across pose: A review," *Pettern Recognition*, vol. 42, pp. 2876–2896, 2009.

[11] A. K. Jain, B. Klare, and U. Park, "Face matching and retrieval in forensics applications," *Multimedia, IEEE*, vol. 19, pp. 20–28, 2012.

[12] "Faceit argus." [Online]. Available: http://www.l1id.com/pages/71-facial-screening

[13] "Checkpoint.s facial surveillance." [Online]. Available: http://www.omniperception.com/products/checkpoints-facial-surveillance/

[14] "Facevacs-videoscan." [Online]. Available: http://www.cognitec-systems.de/FaceVACS-VideoScan.20.0.html

[15] H. Imaoka, A. Hayasaka, Y. Morishita, A. Sato, and T. Hiroaki, "Nec's face recognition technology and its applications," *NEC Technical Journal*, vol. 5, pp. 28–33, 2010.

[16] "Checkpoint access control." [Online]. Available: http://www.omniperception.com/products/checkpoint-access-control/

[17] "Sy face." [Online]. Available: http://www.synel.com/access-control/sy-face.html

[18] "Facevacs-dbscan with examiner." [Online]. Available: http://www.cognitec-systems.de/FaceVACS-DBScan-with-Examiner.21.0.html

[19] "Colossus facial search engine." [Online]. Available: http://www.omniperception.com/products/colossus-image-search-engine/

[20] "iphoto." [Online]. Available: http://www.apple.com/ilife/iphoto/

[21] "Picasa." [Online]. Available: http://picasa.google.com/

[22] "Faceit sdk." [Online]. Available: http://www.l1id.com/pages/101-faceit-sdk

[23] "Facevacs-sdk." [Online]. Available: http://www.cognitec-systems.de/ FaceVACS-SDK.19.0.html

[24] "Okao vision." [Online]. Available: http://www.omron.com/r_d/coretech/vision/okao.html

[25] "Neoface." [Online]. Available: http://www.nec.com/en/global/solutions/ security/ product/neoface.html

[26] M. S. Nixon and J. N. Carter, "Automatic recognition by gait," in *Proceedings of the IEEE*, vol. 94, no. 11, 2006, pp. 2013–2024.

[27] "How biometrics could change security." [Online]. Available: http://news.bbc.co.uk/2/hi/programmes/click_online/ 7702065.stm

[28] D. Cunado, M. Nixon, and J. Carter, "Automatic extraction and description of human gait models for recognition purposes," *Computer Vision and Image Understanding*, vol. 90, no. 1, pp. 1–41, 2003.

[29] C. Yam, M. Nixon, and J. Carter, "Automated person recognition by walking and running via model-based approaches," *Pattern Recognition*, vol. 37, no. 5, pp. 1057–1072, 2004.

[30] A. Bobick and A. Johnson, "Gait recognition using static activity-specific parameters," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 423–430.

[31] D. Wagg and M. Nixon, "On automated model-based extraction and analysis of gait," in *Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 11–16.

[32] R. Urtasun and P. Fua, "3d tracking for gait characterization and recognition," in *Proc. of the 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2004, pp. 17–22.

[33] S. Sarkar, J. Phillips, Z. Liu, I. Vega, P. Grother, and K. Bowyer, "The humanid gait challenge problem: Data sets, performance, and analysis," *Trans. of Pattern Analysis and Machine Intelligence*, vol. 27, no. 2, pp. 162–177, 2005.

[34] A. Sundaresan, A. R. Chowdhury, and R. Chellappa, "A hidden markov model based framework for recognition of humans from gait sequences," in *Proc. on International Conference on Image Processing*, 2003, pp. 93–96.

[35] Z. Liu and S. Sarkar, "Simplest representation yet for gait recognition: Averaged silhouette," in *Proc. of the 17th Int. Conf. on Pattern Recognition*, vol. 1, Aug. 2004, pp. 211–214.

[36] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 2, pp. 316–322, 2006.

[37] K. Bashir, T. Xiang, and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognition Letters*, vol. 31, no. 13, pp. 2052–2060, 2010.

[38] C. Wang, J. Zhang, L. Wang, J. Pu, and X. Yuan, "Human identification using temporal information preserving gait template," *IEEE Transactions on Pattern Analysis and Machine Intelligence (Epub ahead of print)*, 2011.

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

[40] 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*, vol. 4, Hong Kong, China, Aug. 2006, pp. 441–444.

[41] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," in *European conference on Computer vision*, vol. 3, 2006, pp. 151–163.

[42] 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*, vol. 43, no. 6, pp. 2281–2291, Jun. 2010.

[43] R. Gross and J. Shi, "The cmu motion of body (mobo) database," CMT, Tech. Rep., Jun. 2001.

[44] T. Chalidabhongse, V. Kruger, and R. Chellappa, "The umd database for human identification at a distance," University of Meryland, Tech. Rep., 2001.

[45] J. B. Hayfron-Acquah, M. S. Nixon, and J. N. Carter, "Automatic gait recognition by symmetry analysis," *Pattern Recognition Letters*, vol. 24, pp. 2175–2183, 2003.

[46] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1505–1518, 2003.

[47] R. Tanawongsuwan and A. Bobick, "A study of human gaits across different speeds," Georgia Tech, Tech. Rep., 2003.

[48] D. Tan, K. Huang, S. Yu, and T. Tan, "Efficient night gait recognition based on template matching," in *Proceedings of the 18th International Conference on Pattern Recognition*, vol. 3, 2006, pp. 1000–1003.

[49] D. Matovski, M. Nixon, S. Mahmoodi, and J. Carter, "The effect of time on the performance of gait biometrics," in *IEEE Fourth Conf. on Biometrics: Theory, Applications and Systems*, Washington DC, USA, Sep. 2010, pp. 1–6.

[50] Y. Makihara, A. Tsuji, and Y. Yagi, "Silhouette transformation based on walking speed for gait identification," in *Proc. of the 23rd IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, USA, Jun 2010.

[51] A. Mori, Y. Makihara, and Y. Yagi, "Gait recognition using period-based phase synchronization for low frame-rate videos," in *Proc. of the 20th Int. Conf. on Pattern Recognition*, Istanbul, Turkey, Aug. 2010, pp. 2194–2197.

[52] H. Mannami, Y. Makihara, 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*, Queenstown, New Zealand, Nov. 2010.

[53] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 1999, pp. 246–252.

[54] T. Horprasert, D. Harwood, and L. S. Davis, "A robust background subtraction and shadow detection," in *Proc. of the 4th Asian Conference on Computer Vision*, 2000, pp. 983–988.

[55] T. T. Santos and C. H. Morimoto, "People detection under occlusion in multiple camera views," in *Proceedings of the 2008 XXI Brazilian Symposium on Computer Graphics and Image Processing*, 2008, pp. 53–60.

[56] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 663–671, 2006.

[57] W. Du and J. Piater, "Multi-camera people tracking by collaborative particle filters and principal axis-based integration," in *Proceedings of the 8th Asian conference on Computer vision*, vol. 1, 2007, pp. 365–374.

[58] S. Calderara, R. Cucchiara, and A. Prati, "Bayesian-competitive consistent labeling for people surveillance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 354–360, 2008.

[59] S. Calderara, A. Prati, and R. Cucchiara, "Hecol: Homography and epipolar-based consistent labeling for outdoor park surveillance," *Computer Vision and Image Understanding*, vol. 111, pp. 21–42, 2008.

[60] S. M. Khan and M. Shah, "Tracking multiple occluding people by localizing on multiple scene planes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 505–519, 2009.

[61] S. Park and M. M. Trivedi, "Homography-based analysis of people and vehicle activities in crowded scenes," in *Proceedings of the Eighth IEEE Workshop on Applications of Computer Vision*, 2007, p. 51.

[62] B. Song, A. T. Kamal, C. Soto, C. Ding, A. K. Roy-Chowdhury, and J. A. Farrell, "Tracking and activity recognition through consensus in distributed camera networks," *IEEE Transactions on Image Processing*, vol. 19, pp. 2564–2579, 2010.

[63] Y. Ivanov, A. Bobick, and J. Liu, "Fast lighting independent background subtraction," *International Journal of Computer Vision*, vol. 37, no. 2, pp. 199–207, 2000.

[64] K. Jeong and C. Jaynes, "Moving shadow detection using a combined geometric and color classification approach," in *Proc IEEE Workshop on Motion and Video Computing 2005*, vol. 2, 2005, pp. 36–43.

[65] V. Kolmogorov and R. Zabih, "Graph cut algorithms for binocular stereo with occlusions," *Mathematical Models in Computer Vision: The Handbook*, pp. 423–438, 2005.

[66] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 11, pp. 1222–1239, 2001.

[67] J. Choi, Y. Jun, and J. Y. Choi, "Adaptive shadow estimator for removing shadow of moving object," *Computer Vision and Image Understanding*, vol. 114, no. 9, pp. 1017–1029, 2010.

[68] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, "Detection moving objects, ghosts, and shadows in video streams," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 10, pp. 1337–1342, 2003.

[69] Y. Sun, B. Li, B. Yuan, Z. Miao, and C. Wan, "Better foreground segmentation for static cameras via new energy form and dynamic graph-cut," in *Proc. of the 18th Int. Conf. on Pattern Recognition*, vol. 4, Aug. 2006, pp. 49–52.

[70] E. Salvador, A. Cavallaro, and T. Ebrahimi, "Cast shadow segmentation using invariant color features," *Computer Vision and Image Understanding*, vol. 95, no. 2, pp. 238–259, 2004.

[71] T. Kakuta, L. B. Vinh, R. Kawakami, T. Oishi, and K. Ikeuchi, "Detection of moving objects and cast shadows using a spherical vision camera for outdoor mixed reality," in *Proc. on 15th ACM Symposium on Virtual Reality Software and Technology*, Oct 2008, pp. 219–222.

[72] A. Prati, I. Mikic, M. M. Trivedi, and R. Cucchiara, "Detecting moving shadows: Algorithms and evaluation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 918–923, 2003.

[73] F. Porikli and J. Thornton, "Shadow flow: A recursive method to learn moving cast shadows," in *Proc. IEEE Int. Conf. on Computer Vision*, vol. 1, 2005, pp. 891–898.

[74] N. Martel-Brisson and A. Zaccarin, "Learning and removing cast shadows through a multidistribution approach," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 29, no. 7, pp. 1133–1146, 2007.

[75] J.-B. Huang and C.-S. Chen, "Moving cast shadows detection using physics-based features," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 2310–2317.

[76] O. Javed and M. Shah, "Tracking and object classification for automated surveillance," in *Proc. on 7th European Conference on Computer Vision*, 2002, pp. 343–357.

[77] A. Sanin, C. Sanderson, and B. C. Lovell, "Improved shadow removal for robust person tracking in surveillance scenarios," in *Proc. of the 20th Int. Conf. on Pattern Recognition*, 2010, pp. 141–144.

[78] I. Huerta, M. Holte, T. Moeslund, and J. Gonzalez, "Detection and removal of chromatic moving shadow in surveillance scenarios," in *Proc. IEEE Int. Conf. on Computer Vision*, 2009, pp. 1499–1506.

[79] A. J. Joshi and N. P. Papanikolopoulos, "Learning to detect moving shadows in dynamic environments," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 30, pp. 2055–2063, 2008, 11.

[80] R. Qin, S. Liao, Z. Lei, and S. Z. Li, "Moving cast shadow removal based on local descriptors," in *Proc. of the 20th Int. Conf. on Pattern Recognition*, 2010, pp. 1377–1380.

[81] G. Gordon, T. Darrell, M. Harville, and J. Woodfill, "Background estimation and removal based on range and color," in *Proc. of the 18th Int. Conf. on Pattern Recognition*, 1999, pp. 459–464.

[82] A. Moro, K. Terabayashi, and K. Umeda, "Detection of moving objects with removal of cast shadows and periodic changes using stereo vision," in *Proc. of the 20th Int. Conf. on Pattern Recognition*, 2010, pp. 328–331.

[83] J.-H. Ahn, K. Kim, and H. Byun, "Robust object segmentation using graph cut with object and background seed estimation," in *Proc. of the 18th Int. Conf. on Pattern Recognition*, vol. 2, 2006, pp. 361–364.

[84] N. Kasuya, I. Kitahara, Y. Kameda, and Y. Ohta, "Robust trajectory estimation of soccer players by using two cameras," in *Proc. of the 19th Int. Conf. on Pattern Recognition*, 2008, pp. 1–4.

[85] R. Hamid, R. KrishanKumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins, "Player localization using multiple static cameras for sports visualization," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2010, pp. 731–738.

[86] P. Kelly, P. Beardsley, E. Cooke, N. O'Conner, and A. Smeaton, "Detecting shadows and low-lying objects in indoor and outdoor scenes using homographies," in *Proc. IEEE Int. Conf. on Visual Information Engineering, Convergence in Graphics and Vision*, April 2005, pp. 393–400.

[87] P. H. Batavia and S. Singh, "Obstacle detection using adaptive color segmentation and color stereo homography," in *Proc. of IEEE Int. Conf. on Robotics and Automation*, vol. 1, May 2001, pp. 705–710.

[88] D. H. Marimont and B. A. Wandell, "Linear models for surface and illumination spectra," *Journal of the Optical Society of America*, pp. 1905–1913, 1992.

[89] V. Kolmogorov, "Graph based algorithms for scene reconstruction from two or more views," Ph.D. dissertation, Cornell University, Sep 2003.

[90] K. Sugiura, Y. Makihara, and Y. Yagi, "Gait identification based on multi-view observations using omnidirectional camera," in *Proc. on 8th Asian Conference on Computer Vision*, Nov 2007, pp. 452–461.

[91] Y. Iwashita and A. Stoica, "Gait recognition using shadow analysis," in *2009 Bio-inspired Learning and Intelligent Systems for Security*, 2009, pp. 26–31.

[92] M. Naaman, R. B. Yeh, H. Garcia-Molina, and A. Paepcke, "Leveraging context to resolve identity in photo albums," in *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, 2005, pp. 178–187.

[93] A. C. Gallagher and T. Chen, "Using group prior to identify people in consumer images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[94] D. Lin, A. Kapoor, G. Hua, and S. Baker, "Joint people, event, and location recognition in personal photo collections using cross-domain context," in *Proceedings of the 11th European conference on Computer vision*, 2010, pp. 243–256.

[95] A. Torralba, K. P. Murphy, and W. T. Freeman, "Using the forest to see the trees: exploiting context for visual object detection and localization," *Communications of the ACM*, vol. 53, no. 3, pp. 107–114, 2010.

[96] D. Hoiem, A. A. Efros, and M. Hebert, "Putting objects in perspective," *International Journal of Computer Vision*, no. 1, pp. 3–15, 2008.

[97] R. Perko and A. Leonardis, "A framework for visual-context-aware object detection in still images," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 700–711, 2010.

[98] W.-S. Zheng, S. Gong, and T. Xiang, "Quantifying contextual information for object detection," in *IEEE International Conference on Computer Vision*, 2009, pp. 932–939.

[99] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *European Conference on Computer Vision*, no. 30–43, 2008.

[100] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[101] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[102] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, no. 2, 2008, pp. 1–8.

[103] D. Parikh, C. L. Zitnick, and T. Chen, "From appearance to context-based recognition: Dense labeling in small images," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[104] A. T. A. Oliva, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, no. 12, pp. 520–527, 2007.

[105] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, pp. 712–722, 2010.

[106] O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: techniques, implications, and applications," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 303–339, 2011.

[107] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg, "A scalable approach to activity recognition based on object use," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[108] A. Gupta, A. Kembhavi, and L. S. Davis, "Observing human-object interactions: Using spatial and functional compatibility for recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1775–1789, 2009.

[109] D. Han, L. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *IEEE International Conference on Computer Vision*, 2009, pp. 1933–1940.

[110] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 9–16.

[111] ——, "Modeling mutual context of object and human pose in human-object interaction activities," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 17–24.

[112] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[113] N. Ikizler-Cinbis and S. Sclaroff, "Object, scene and actions: Combining multiple features for human action recognition," in *European Conference on Computer Vision*, vol. 1, 2010, pp. 494–507.

[114] T. Lan, Y. Wang, G. Mori, and S. Robinovitch, "Retrieving actions in group contexts," in *ECCV Workshop on Sign, Gesture and Activity*, 2010.

[115] K. S. W. Choi and S. Savarese, "Learning context for collective activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[116] Y. Song and T. Leung, "Context-aided human recognition- clustering," in *European Conference on Computer Vision*, 2006.

[117] L. Zhang, L. Chen, M. Li, and H. Zhang, "Automated annotation of human faces in family albums," in *Proceedings of the eleventh ACM international conference on Multimedia*, 2003, pp. 355–358.

[118] Z. Stone, T. Zickler, and T. Darrell, "Autotagging facebook: Social network context improves photo annotation," in *In Proceedings of CVPR Workshop on Internet Vision*, 2008, pp. 1–8.

[119] T. X. Wei-Shi Zheng, Shaogang Gong, "Associating groups of people," in *British Machine Vision Conference*, vol. 5, 2009.

[120] Y. Cai, V. Takala, and M. Pietikainen, "Matching groups of people by covariance descriptor," in *Proceedings of the 20th International Conference on Pattern Recognition*, 2010, pp. 2744–2747.

[121] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," in *IEEE 12th International Conference on Computer Vision*, 2009, pp. 261–268.

[122] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg, "Who are you with and where are you going?" in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1345–1352.

[123] T. S. Cho, S. Avidan, and W. T. Freeman, "A probabilistic image jigsaw puzzle solver," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 183–190.

[124] ——, "The patch transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 8, pp. 1489–1501, 2010.

[125] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Understanding belief propagation and its generalizations," in *Exploring artificial intelligence in the new millennium.* Morgan Kaufmann Publishers Inc. San Francisco, CA, USA, 2003, ch. 8, pp. 239–269.

[126] M. Okumura, H. Iwama, Y. Makihara, and Y. Yagi, "Performance evaluation of vision-based gait recognition using a very large-scale gait database," in *IEEE Fourth International Conference on Biometrics: Theory, Applications and Systems*, 2010.

[127] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary and rrgion segmentation of objects in n-d images," in *Proc. of Int. Conf. on Computer Vision*, July 2001, pp. 105–112.

[128] W. Ge, R. T. Collins, and R. B. Ruback, "Automatically detecting the small group structure of a crowd," in *IEEE Workshop on Applications of Computer Vision*, 2009, pp. 1–8.

[129] J. Sochman and D. C. Hogg, "Who knows who - inverting the social force model for finding groups," in *IEEE International Workshop on Socially Intelligent Surveillance and Monitoring*, 2011, pp. 1–8.

[130] T. Zhao, R. Nevatia, and B. Wu, "Segmentation and tracking of multiple humans in crowded environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1198–1211, 2008.

[131] M. Wu, X. Peng, Q. Zhang, and R. Zhao, "Segmenting and tracking multiple objects under occlusion using multi-label graph cut," *Computers and Electrical Engineering*, vol. 36, no. 5, pp. 927–934, 2010.

[132] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, "Online multiperson tracking-by-detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[133] S. Yu, T. Tan, K. Huang, K. Jia, and X. Wu, "A study on gait-based gender classification," *IEEE Trans. on Image Processing*, vol. 18, no. 8, pp. 1905–1910, Aug. 2009.

[134] J. Lu and Y.-P. Tan, "Gait-based human age estimation," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 761–770, 2010.

[135] Y. Makihara, M. Okumura, H. Iwama, and Y. Yagi, "Gait-based age estimation using a whole-generation gait database," in *Proc. of the International Joint Conference on Biometrics*, no. 195, 2011, pp. 1–6.

[136] D. Matovski, M. Nixon, S. Mahmoodi, and J. Carter, "The effect of time on gait recognition performance," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 543–552, 2011.

[137] I. Bouchrika and M. Nixon, "Exploratory factor analysis of gait recognition," in *8th IEEE International Conference on Automatic Face and Gesture Recognition*, Amsterdam, The Netherlands, 2008.

[138] S. Yu, D. Tan, and T. Tan, "Modelling the effect of view angle variation on appearance-based gait recognition," in *Proc. of 7th Asian Conf. on Computer Vision*, vol. 1, Jan. 2006, pp. 807–816.

[139] M. Okumura, Y. Makihara, S. Nakamura, S. Morishima, and Y. Yagi, "The online gait measurement for the audience-participant digital entertainment," in *Invited Workshop on Vision Based Human Modeling and Synthesis in Motion and Expression*, Xi'an, China, Sep. 2009.

[140] S. Morishima, Y. Yagi, and S. Nakamura, "Instant movie casting with personality: Dive into the movie system," in *Proc. of the 2011 international conference on Virtual and mixed reality: systems and applications*, 2011, pp. 187–196.

[141] Z. Liu, L. Malave, and S. Sarkar, "Studies on silhouette quality and gait recognition," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 704–711.

[142] Z. Liu and S. Sarkar, "Effect of silhouette quality on hard problems in gait recognition," *Trans. of Systems, Man, and Cybernetics Part B: Cybernetics*, vol. 35, no. 2, pp. 170–183, 2005.

[143] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[144] S. Tsuji, Y. Yagi, and M. Asada, "Dynamic scene analysis for a mobile robot in a man-made environment," in *Proc. of the IEEE the International Conference on Robotics and Automation*, 1985, pp. 850–855.

[145] M. Magee and J. Aggarwal, "Determining vanishing points from perspective images," *Computer Vision, Graphics, and Image Processing*, vol. 26, no. 2, pp. 256–267, 1984.

[146] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proceedings of the 3rd International Conference on Imaging for Crime Detection and Prevention*, 2009.

[147] C. Wang, J. Zhang, J. Pu, X. Yuan, and L. Wang, "Chrono-gait image: A novel temporal template for gait recognition," in *Proceedings of the 11th European conference on Computer vision*, 2010, pp. 257–270.

[148] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognition*, vol. 44, no. 4, pp. 973–987, 2011.

[149] P. Grother, "Face recognition vendor test 2002 supplemental report," Tech. Rep., 2004. [Online]. Available: http://www.face-rec.org/vendors

[150] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226–239, 1998.

[151] G. Snedecor and W. Cochran, *Statitical methods*. Iowa State University Press, 1967.

[152] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics (International Series on Biometrics)*. Springer-Verlag New York, Inc., 2006.

[153] P. Grother and E. Tabassi, "Performance of biometric quality measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 4, pp. 531–543, 2007.

[154] M. O. Alam and B. Boufama, "Occlusion handling based on sub-blobbing in automated video surveillance system," in *Proceedings of The Fourth International C* Conference on Computer Science and Software Engineering*, 2011, pp. 139–143.

[155] Y. Chai, Q. Wang, J. Jia, and R. Zhao, "A novel human gait recognition method by segmenting andextracting the region variance feature," in *Proc. of the 18th Int. Conf. on Pattern Recognition*, vol. 4, Hong Kong, China, Aug. 2006, pp. 425–428.

[156] N. Boulgouris and Z. Chi, "Human gait recognition based on matching of body components," *Pattern Recognition*, vol. 40, no. 6, pp. 1763–1770, 2007.

[157] I. Venkat and P. DeWilde, "Robust gait recognition by learning and exploiting sub-gait characteristics," *International Journal of Computer Vision*, vol. 91, pp. 7–23, 2011.