博士論文

A Study of Person Recognition Using Body Sway Observed from Overhead Camera

(頭上カメラから観測された身体動揺を用いた人物認識の研究)

2021年1月

工学研究科 情報エレクトロニクス専攻

D18T2102M 神谷 卓也

指導教員 岩井儀雄　教授

Abstract

Person recognition using a camera is an important technique for developing large-scale automated computer systems. There are two interesting and important tasks in person recognition: specifically, person re-identification and gender classification. Existing methods use gait features—mainly those that represent large swings of the limbs—for both identifying people and classifying their gender. The problem with the use of gait features is that the performance of identification and classification decrease when a person stops walking and maintains an upright posture. To extract informative features for person re-identification and gender classification, it is important to measure small swings of the body, which are referred to as body sway. This thesis reports three techniques for identifying people and classifying their gender using body sway observed from a camera.

The first part describes the extraction of features from the body sway observed from an overhead camera for the purpose of identifying people. To represent identity from body sway, bodies are spatially divided into regions in a video sequence and local movements are temporally measured in the body regions. The power spectral density is estimated from the local movements as features for identifying people. To evaluate the identification performance when using the body sway features, three original video datasets of body sway sequences were collected. The first dataset contains a large number of participants in an upright posture. The second dataset includes variation over the long term. The third dataset represents body sway in different postures. The results on these datasets confirm that the local movements can represent features that are informative for person re-identification.

The second part describes person re-identification in the case of self-occlusion, by using body sway measured at the head using an overhead camera. To represent the identity of people, as reflected in body sway, it is important to estimate appearances of a person accurately from images. Defects caused by self-occlusion in such images frequently degrade the performance of one of the existing methods of identifying people because that method uses whole-body regions to identify people. To solve the problem of self-occlusion in this context, silhouette sequences of regions at the head are computed by applying a segmentation technique. To reflect people's identities using body sway, the head region is spatially divided into local blocks and movements inside the blocks are temporally measured. The results of experiments show that the proposed method can improve the performance of the existing method of identification from 17.3% to 57.9%.

The third part discusses whether it is possible to classify the gender of a standing person from a video sequence containing body sway, observed from an overhead camera. A spatiotemporal feature is designed for representing body sway using the frequency analysis of time-series signals derived from the local movements. To evaluate the classification accuracy of the proposed method, video sequences of body sway were acquired from 30 females and 30 males using an overhead camera. The proposed method achieved $90.3 \pm 1.3\%$ accuracy for the gender classification of a standing person. The accuracy of the proposed method was compared with that of existing methods that use other spatiotemporal features. The proposed spatiotemporal feature extracted from body sway significantly improved the accuracy of gender classification.

Throughout the research described in this thesis, body sway was used for person re-identification and gender classification and enabled significantly improved performance, compared with existing methods, when people maintain an upright posture. The proposed techniques help with recognizing and understanding people. In the future, the author expects that the techniques will contribute to the development of security systems and marketing analysis systems in research and business fields.

Acknowledgments

First and foremost, I would like to express my gratitude to my advisor, Professor Yoshio Iwai, for his excellent guidance on my research. His suitable advice and continuous support have helped me to complete this thesis. I would like to thank him for the opportunity to work with him. I would like to express my great appreciation to Associate Professor Masashi Nishiyama, for his polite and enthusiastic guidance for how to progress with research and to write papers. His support has greatly advanced my research. I am deeply grateful to Assistant Professor Hiroki Yoshimura for his many helpful advice and supports regarding my research. He gave me the opportunity to discuss my research deeply. Furthermore, I would like to thank Professor Takayoshi Yokota who was a vice-chair in my dissertation committee. His great feedback allowed me to rethink my research from a new perspective.

I would like to express my gratitude to laboratory members. In particular, I am happy to have been able to work on the research with my colleagues on the team of body sway. I would also like to thank the participants who helped me collect the experimental data. Thanks to their cooperation, we were able to obtain excellent experimental results. I received great ideas and advice for my research from many people that I met at conferences and technical exchanges. They contributed to the development of my research. I would like to express my gratitude to them.

Finally, I would like to express my heartfelt gratitude to my family. They expressed an understanding of my advance to the doctoral degree and encouraged me continuously. Thanks to their supports, I was able to make great progress in my research.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, there has been a high demand for technologies [1, 2, 3, 4] for recognizing people. Technologies for recognizing people accurately are expected to be applied to security enhancement or marketing analysis. The specific tasks involved in recognizing people, such as identifying people and classifying their attributes, have been studied for many years. In the case of person re-identification, the author searches a database for the same person as the recognized person. The person re-identification task [5, 6, 7, 8, 9, 10, 11] is expected to be applied to searching for lost children or suspicious persons. In the case of attribute classification, the attributes describe properties, such as the gender or age, of a person, and the author predicts a label (e.g., male or female, young or old) for each attribute. In this thesis, the author focuses specifically on the gender attribute. The gender classification task [12, 13, 14, 15, 16, 17, 18] is expected to be applied to analyzing advertisements and merchandise in which males and females have different interests. Figure 1.1 provides an overview of person re-identification and gender classification. It is necessary to obtain informative characteristics that represent differences between individuals and differences between genders, to perform person re-identification and gender classification, respectively.

Sensors are commonly used to obtain characteristics that represent people's identity and gender. There are two types of sensors. The first type is a contact sensor, such as a fingerprint sensor [19, 20, 21, 22, 23], vein sensor [24, 25, 26, 27], or iris sensor [28, 29, 30, 31, 32]. The author can identify people and classify their gender with high accuracy by obtaining the charac-

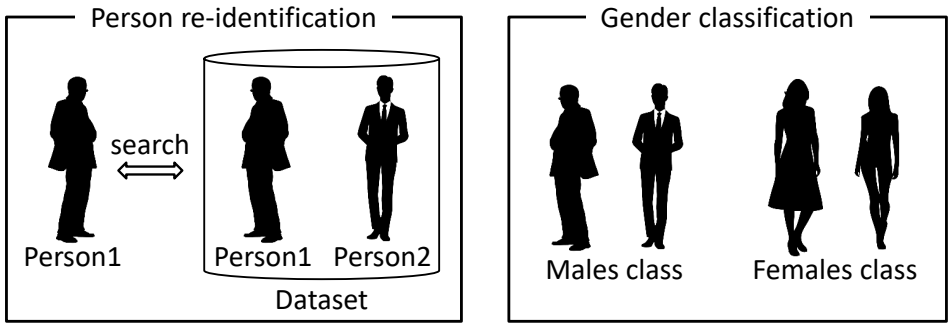


Figure 1.1: Overview of person re-identification and gender classification. In person re-identification, the database is searched for the same person as the recognized person. In gender classification, the author determines whether the target person is male or female.

teristics of individuals or genders using a contact sensor. However, the target person needs to be close to a contact sensor and use it voluntarily. The second type is a non-contact sensor, such as an RGB camera [5, 6, 12, 13] or depth camera [33, 34, 35, 36]. It is possible to obtain the characteristics by using a non-contact sensor even when the target person is some distance away. In this thesis, the author considers the use of a camera, which is a non-contact sensor, to obtain the characteristics to be used for person re-identification and gender classification.

Person re-identification and gender classification using a video sequence from a surveillance camera is a key technology for the development of various authentication systems [37, 38, 39]. To achieve a high performance in person re-identification and gender classification, it is important to design methods of extracting informative features from the video sequence. Recently, soft biometrics [40, 41, 42, 43, 44, 45] that represent human characteristics have been an active topic in pattern recognition research because of their ability to extract informative features for person re-identification and gender clas-

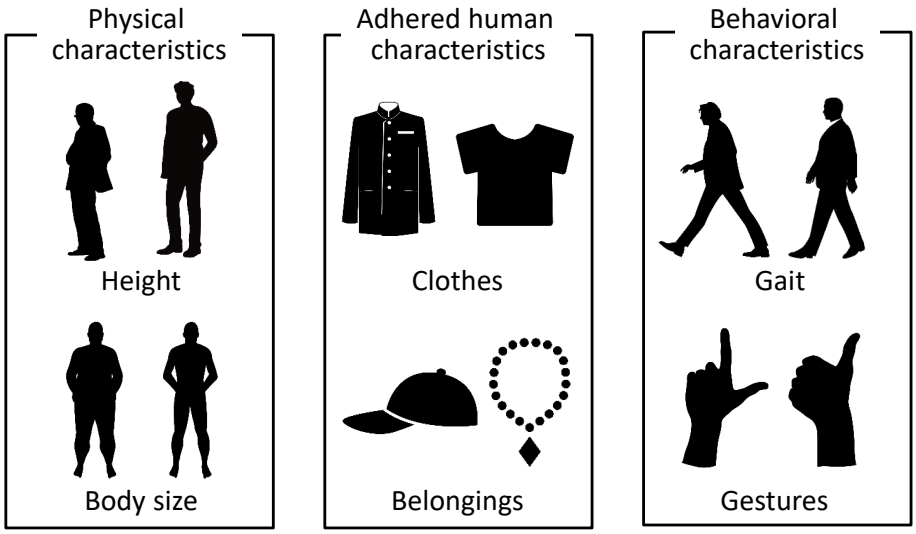


Figure 1.2: Examples of human characteristics that can be acquired from a video sequence. The left panel represents physical characteristics (e.g., height and body size). The center panel represents adhered human characteristics (e.g., clothes and belongings). The right panel represents behavioral characteristics (e.g., gestures and gait).

sification. Human characteristics can be split intuitively into three classes: physical characteristics [46, 47, 48, 49] (e.g., height and body size), adhered human characteristics [50, 51, 52, 53] (e.g., clothes and belongings), and behavioral characteristics [54, 55, 56, 57, 58] (e.g., gestures and gait). Figure 1.2 depicts examples of human characteristics. In particular, behavioral characteristics have the advantage that they can be used for person re-identification and gender classification even when characteristics such as their age, height, and clothing are the same. For instance, there are situations in which many office workers in a building wear a suit or workers in a factory wear a uniform.

In this thesis, the author focuses on how to design a method of extracting and using behavioral characteristics. Existing methods [54, 55, 56, 57, 58]

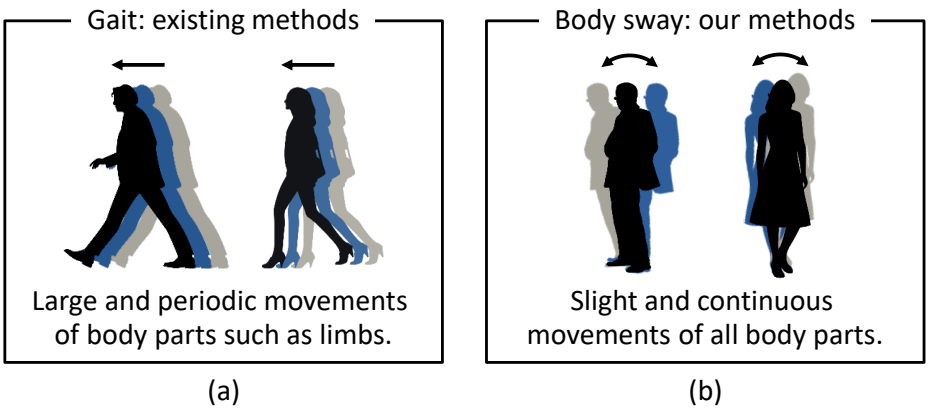


Figure 1.3: Examples of gait and body sway for person re-identification and gender classification. (a) explains gait, as used in existing methods. (b) explains body sway, as used in the proposed methods.

that use behavioral characteristics generally exploit gait features acquired from a video sequence. Gait features represent large and periodic movements of body parts such as limbs, as shown in Figure 1.3(a). However, people frequently stop walking, for example, while waiting for a security gate to open. Therefore, because periodic movements of body parts do not always occur, gait features do not sufficiently represent behavioral characteristics and are not informative for person re-identification or gender classification. Indeed, when people stop walking, the use of gait features sometimes causes a decrease in the performance of person re-identification and gender classification.

The author considers person re-identification and gender classification systems that require people to maintain an upright posture for several tens of seconds. When people maintain their posture, their bodies do not remain completely still, but move slightly and continuously in all directions. This body movement occurs naturally to maintain a person's posture, as shown in Figure 1.3(b), and is called body sway. The author considers an upright posture as a specific example of the posture of a person who has stopped walking.

In the field of medical science, many researchers [59, 60, 61] have attempted to measure the center of gravity of body sway using force plates embedded in the floor. These methods are not intended for person re-identification or gender classification purposes but can be used to classify people with lower-back pain [59], women with morning sickness [60], and patients with neuropathy [61]. The author thus assumes that body sway contains information about the identity of people and gender, and is a behavioral characteristic that can be used to characterize humans in soft biometrics.

In this thesis, the author tackles the challenging task of extracting an informative feature for person re-identification and gender classification from a video sequence featuring body sway. Body sway has the advantage that an overhead camera can be used to observe people passively because body sway movement can be measured from the upper half of the body. Hence, it is not necessary to locate a camera to the side of the person, which is commonly required for extracting gait features [54, 55, 56, 57, 58]. The use of an overhead camera avoids the occlusion of people as the number of people increases.

To accomplish this task, the author proposes a method of identifying people and a method of classifying the gender of a person using a video sequence of body sway acquired from people in an upright posture. The author expects to apply the proposed methods to the use cases shown in Figure 1.4. The proposed methods compute the center of body sway from a video sequence and spatially divide the body into small local regions using the center of body sway. They then measure the temporal and spatial changes in local movements in these regions and conduct a frequency analysis for feature extraction. The main purposes of these studies are to confirm whether the body sway observed when people maintain an upright posture can be used for person re-identification and gender classification. The author collected several original datasets of body sway. The experimental results show that

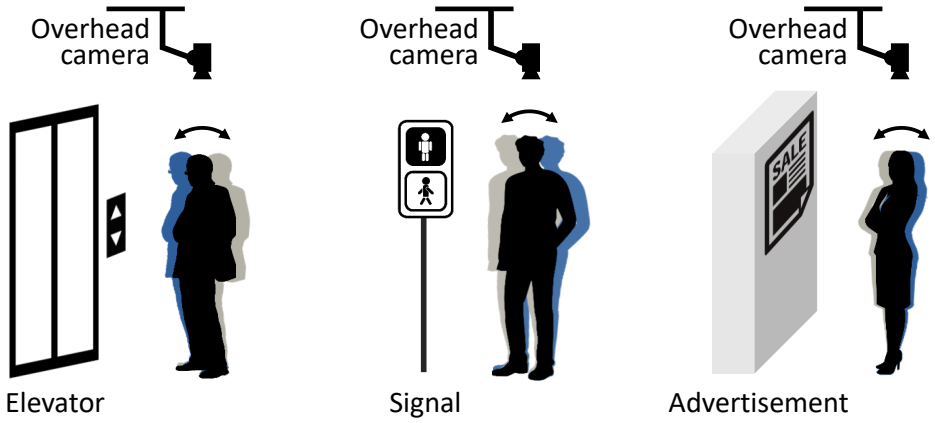


Figure 1.4: Examples of the use cases in which the author applies person re-identification and gender classification using body sway observed from an overhead camera. The left panel shows the use case in which a person waits for an elevator to arrive. The center panel shows the use case in which a person waits for a signal to turn green. The right panel shows the use case in which a person reads an advertisement displayed on a station platform.

the proposed methods of person re-identification and gender classification performed better than existing methods that use the gait feature.

Chapter 2

Temporal and spatial analysis of local body sway movements for the identification of people

2.1 Design of features in terms of body sway

2.1.1 Overview

We consider the informative features extracted from a video sequence of body sway acquired from people in an upright posture. The first feature is the temporal and spatial swinging movement of the body. The movement contains information about identity differences, including gender, age, chronic disease, how muscles are attached, and the sense of balance. The second feature represents the body shape, such as whether the person is obese or thin. The third feature represents the body posture, such as a stooping or slouching posture.

The features of existing gait recognition methods [54, 55] mainly represent the body shape together with the temporal swinging movements, and action recognition methods [54, 62] are similar. Researchers [63, 64] exploited temporal movements as features for gaze authentication. The existing methods have been designed to represent the features of gait, action or gaze. In preliminary experiments, we used the existing methods to extract features from video sequences of body sway. However, we could not obtain high iden-

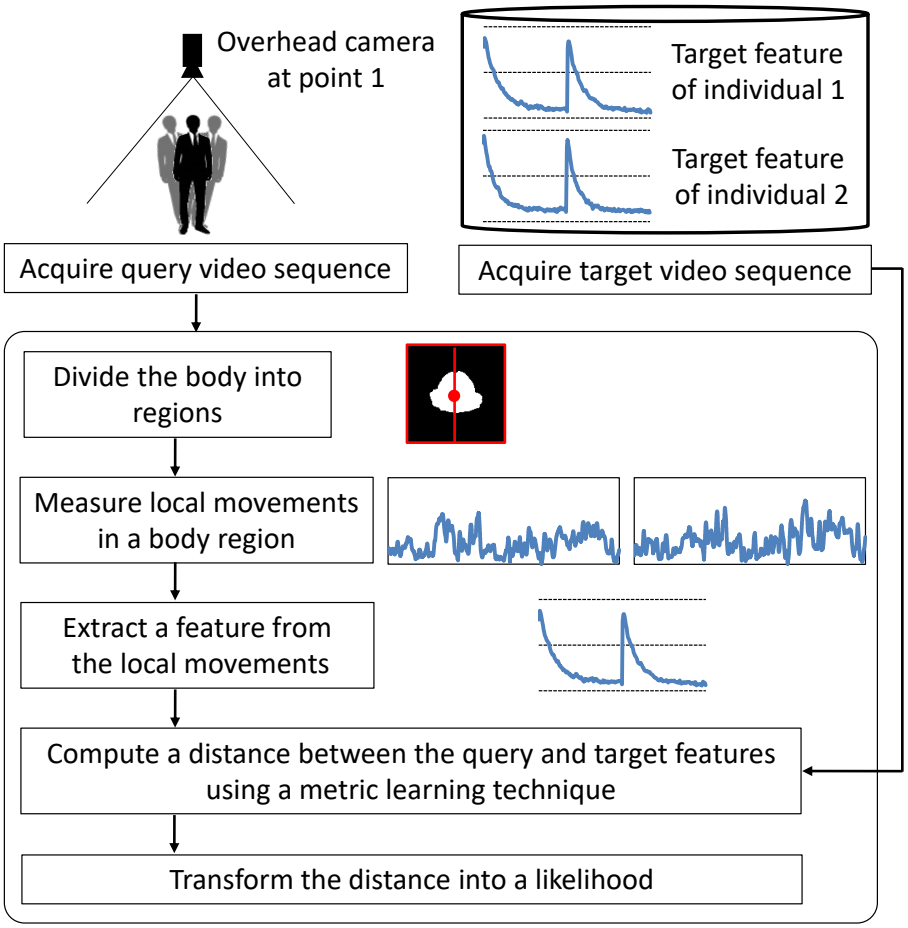


Figure 2.1: Overview of our method.

tification performance using these methods.

We focus on how to represent identity using temporal and spatial changes in movements due to body sway. Figure 2.1 provides an overview of our method. We divide the body into small local regions to represent the spatial movements of body sway. We measure temporal changes in local movements for each local region and compute a query feature from measured temporal changes. We store target features in an authentication system in advance. Our method computes the distance between query and target features using

a metric learning technique [65]. Finally, the distance is transformed into the likelihood of the query and target features belonging to the same person using the technique described in [66]. The details of our method are described below.

2.1.2 Measuring temporal and spatial changes in local movements

We describe a method of measuring temporal and spatial changes in local movements from a video sequence of body sway. Movements of whole body occur around a central position. An existing method [67] measures the movements using the body regions, under the assumption that all body parts move synchronously in the same direction. Although the method considers the temporal changes in movements, it ignores the spatial changes. We thus extend the method to represent spatial changes in movement and extract informative features measured from body sway.

From a frame of the video sequence at time $t \in 1, ..., T$, we compute a mask image $\boldsymbol{m}_t$ in which a pixel takes a value of 1 if it is within a body region and 0 otherwise. The original method [67] uses algorithm 1 to infer the reference time r, which represents the temporal center of swings, from the whole body region in the mask image.

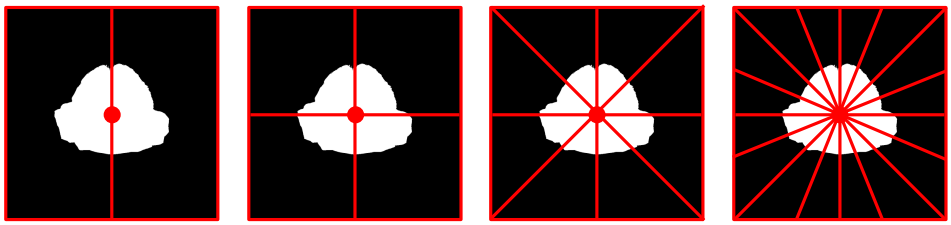


Figure 2.2: Examples of local regions radially divided from the body region. I is the number of local regions.

Algorithm 1 Determining reference time

Input: Mask images $\{\boldsymbol{m}_t | t \in 1, ..., T\}$

Output: Reference time r

```
for r̃ = 1 to T do
    Initialize D_r̃ ← 0
    for t = 1 to T do
        compute d̃ = ||m_r̃ − m_t||_1
        D_r̃ ← D_r̃ + d̃
    end for
end for
r ← arg min D_r̃
```

To consider the spatial change in movement, our method divides the body region into numerous local regions and computes the local movement in each region. The simplest method is to divide the body region into a lattice. However, we cannot stably measure movements around the center of the body region when the lattice cells are small. Therefore, we divide the body radially into local regions using the position of the center of the body, as illustrated in Figure 2.2. We compute the center position using the body

Algorithm 2 Computing local movement

Input: Reference time r, mask images $\{\boldsymbol{m}_t | t \in 1, ..., T\}$, the length of the video sequence T, the number of local regions I

Output: The local movement $\{d_{i,t} | t \in 1, ..., T, i \in 1, ..., I\}$

```
compute the position of the center of body region in m_r
for i = 1 to I do
    set the i-th region using the computed center
    for t = 1 to T do
        compute d_{i,t} using Equation (2.1)
    end for
end for
```

region of the mask image $\boldsymbol{m}_r$ acquired at reference time r. Note that we assume that the center is in the same position at all times $t \in 1, ..., T$. We measure the temporal changes in local movements from the spatially divided local regions using Algorithm 2. We aim to represent the spatial changes in movement in more detail by increasing the number of divisions. The local movement $d_{i,t}$ in a local region $i \in 1, ..., I$ is computed as

$$d_{i,t} = \sum_{\boldsymbol{x} \in \text{region}(i)} \|\boldsymbol{m}_r(\boldsymbol{x}) - \boldsymbol{m}_t(\boldsymbol{x})\|_1 \tag{2.1}$$

where $\boldsymbol{m}_r(\boldsymbol{x})$ and $\boldsymbol{m}_t(\boldsymbol{x})$ are pixel values indicated by $\boldsymbol{x}$, and region(i) is the i-th local region. We use the $\boldsymbol{L}_1$-norm because the mask images are binary.

2.1.3 Extracting the feature for identification

We describe a method of extracting the feature for identification from the temporal and spatial changes in local movements. The identification performance decreases when directly using the changes in local movements because the direction of body sway varies randomly. We thus need to consider a feature that is invariant to the randomness of movements.

Algorithm 3 Extracting the feature using local movements

Input: The local movement $\{d_{i,t}|t \in 1, ..., T, i \in 1, ..., I\}$, the length of the video sequence T, the number of local regions I

Output: The feature $\boldsymbol{f}$

1: **for** $i = 1$ to I **do**
2: compute the PSD with L from $\{d_{i,t}|t \in 1, ..., T\}$
3: compute a value by taking the logarithm of the PSD for each frequency
4: set $\boldsymbol{f}_i$ using the values of all frequencies
5: **end for**
6: concatenate $\{\boldsymbol{f}_i|i \in 1, ..., I\}$ to $\boldsymbol{f}$

In the signal processing field, frequency analysis techniques are widely used to extract informative features from time series signals. Because the changes in local movements are also time series signals, we assume that the frequency analysis techniques are adequate for achieving high performance. We assume that the phase components are shifted each time when we measure the local movements. To alleviate the randomness of swings, we do not use the phase components.

Our method estimates the power spectral density (PSD) from the local movements $d_{i,t}$ using Welch's method [68], and extract the feature $\boldsymbol{f}$ for identification using Algorithm 3. To compute the PSD, we divide the local movements into small segments by convoluting a Hann window. We denote the length of each segment L. If L takes a large value, the frequency resolution increases. We believe that features can capture the details of movements when using a large L when there is no influence of noise. However, the appropriate value of L needs to be chosen experimentally because we cannot ignore noise. We use all of the values from the DC component to the $L/2$-th component computed by the PSD for identification. The dimension of $\boldsymbol{f}_i$ is $L/2$. The feature for identification is represented as $\boldsymbol{f} = [\boldsymbol{f}_1^{\mathrm{T}}, ..., \boldsymbol{f}_I^{\mathrm{T}}]^{\mathrm{T}}$. The

dimension of $\boldsymbol{f}$ is $IL/2$. We expect $\boldsymbol{f}$ to represent the identify of a person while alleviating the random swings in the temporal and spatial changes of local movements.

2.2 Experiments with upright postures

2.2.1 Dataset of video sequences of body sway

We evaluated whether the features extracted from the video sequences of body sway contained identities. We collected video sequences of body sway from 118 participants (average age 22.1 ± 4.3 years; 83 males and 35 females). Each participant maintained an upright posture while standing with their heels aligned as shown in Figure 2.3 (a). We asked all participants to wear the same dark-blue nylon outerwear, similar to a uniform worn by factory workers. We set an overhead camera at a height of 2.3 m. We applied a camera calibration technique such that the optical axis coincided with the normal direction of the floor. Each participant stood under the camera as shown in Figure 2.3 (b). A marker was set to indicate the position of the participant's heel in the standing position. We asked each participant to look at a timer placed 3 m away. We displayed the time lapse on the timer. We used video sequences comprising images of 1920×1080 pixels captured at 30 fps by a Microsoft Kinect V2. The highest PSD frequency was 15 Hz. The time length of a video sequence was 120 s, and the number of sampled movements was $T = 120 \times 30 = 3600$ for each local region. We used a fixed 1000×1000 bounding box to measure local movements. We observed each participant three times. The participant sat and rested between each sequence. To generate the mask images of body regions, we applied a background subtraction technique using images without participants.

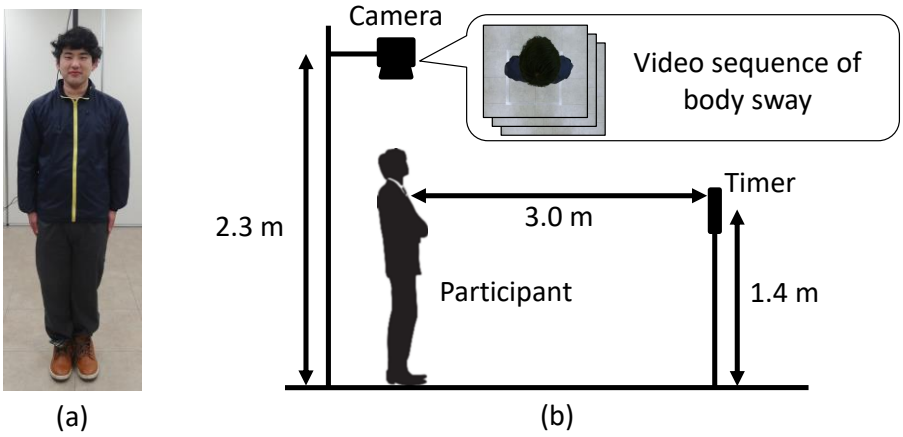


Figure 2.3: Setup for acquiring video sequences of body sway.

2.2.2 Evaluation of the parameters of our method

Figure 2.4 shows examples of our local movement features for two participants. The acquired video sequences are shown in (a) and (e), and the four local regions are presented in (b) and (f). The features in (c) and (g) were extracted from the temporal and spatial changes of the local movements in (d) and (h). The features differed between participants while they maintained an upright posture even though the video sequences appeared to be almost the same.

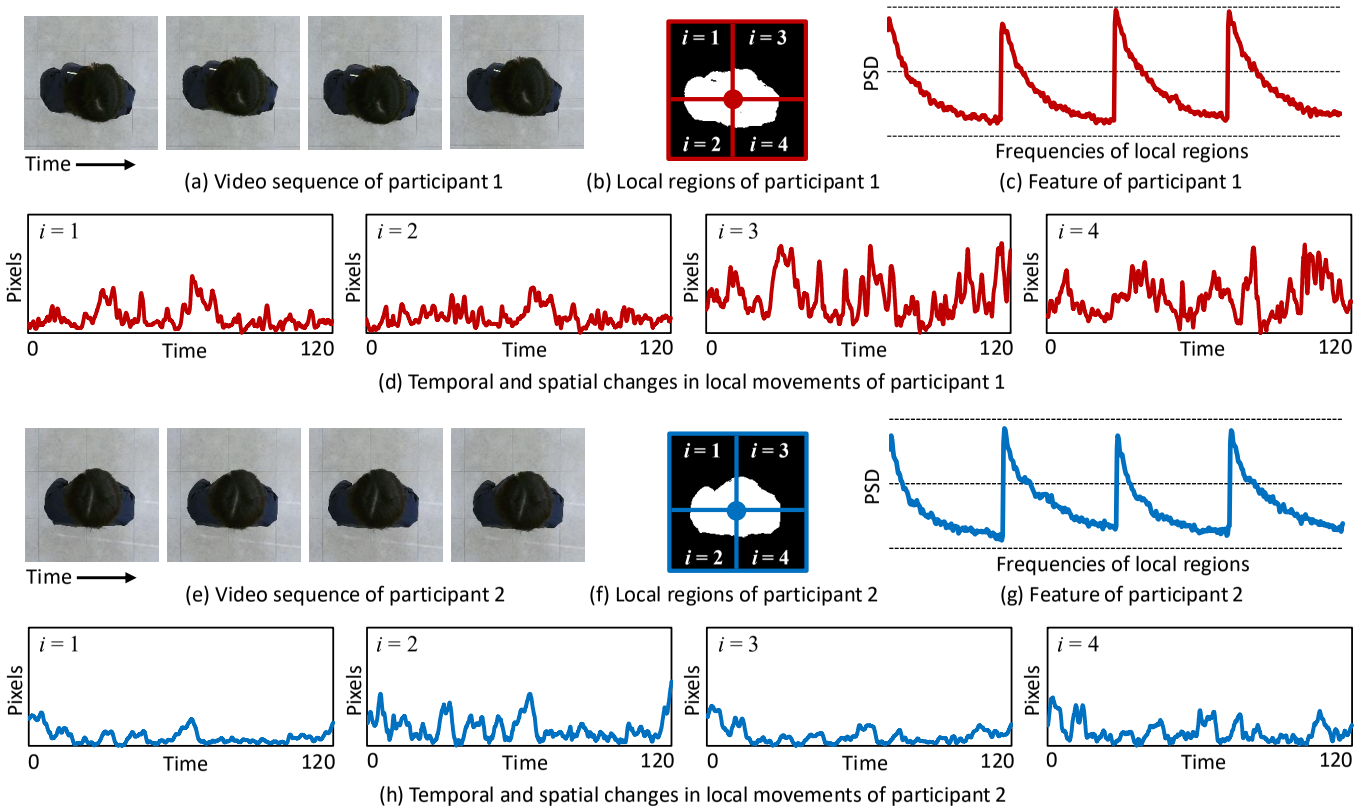


Figure 2.4: Examples of temporal and spatial changes in local movements measured from the video sequences of two participants.

We evaluated the identification performance while changing the number of local regions I, the length of the video sequence T and the length of each segment L, separately. Our method used a metric learning technique, the large margin nearest neighbor (LMNN) method [65]. We randomly selected 59 participants from the dataset described in Section 2.2.1. The remaining 59 participants were used to train a metric matrix for LMNN. We repeated the random selection five times to generate different test sample sets. In each set, we tested ${}_3\mathrm{P}_2 = 6$ times for each participant by selecting a single video sequence as a target and a single video sequence as a query. We used the first matching rate for the identification performance. We used a nearest-neighbor algorithm for identification.

Figure 2.5 shows the mean and standard deviation of the correct matching rate when a certain parameter was fixed and other parameters were changed. The best identification performance was $88.8 \pm 3.8\%$ using $I = 25$, $L = 256$ and $T = 3600$. Figure 2.5 (a) shows that the identification performance was improved by increasing the number of local regions I. When the number of local regions exceeded 15, the identification performance was almost constant. Figure 2.5 (b) shows that the identification performance was stable when the length of each segment L was set between 64 and 512. Figure 2.5 (c) shows the identification performance when the time length of the video sequences was less than 120 s ($T = 3600$). When the length of the video was reduced to 3/4 or 1/2, the performance was reduced by 3.6 and 10.7 points, respectively. We believe that the degradation in performance is related to the periodicity of body sway. A short time length is preferable for the development of practical applications. We will conduct further experiments to reduce the time length in future work.

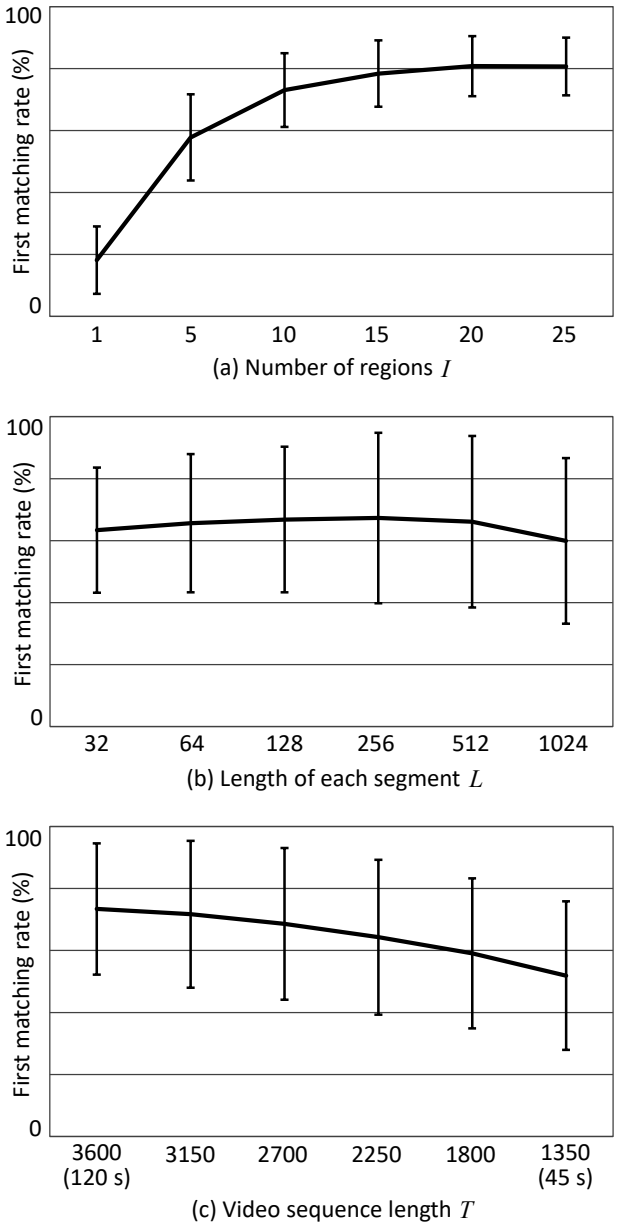


Figure 2.5: Identification performance when changing the parameters of our method: (a) number of local regions, (b) length of each segment, (c) video sequence length.

2.2.3 Comparison with features extracted using existing methods

We compared the identification performance between the local movement features obtained using our method with those obtained using existing methods.

- **LM** (Local Movements): We computed a feature using our method with the parameters set as $I = 25$, $L = 256$, $T = 3600$.
- **GEI** (Gait Energy Image) [54]: We assumed a walking cycle T. We computed a feature by averaging the mask images as $\Sigma_{t=1}^{T} \boldsymbol{m}_t / T$. Figure 2.6 (a) shows an example of a GEI.
- **MHI** (Motion History Image) [62]: We assigned a weight $\tau = t/T$ for each time at a position where movement was generated. We added the temporal weights for each position. Figure 2.6 (b) shows an example of an MHI.
- **MEI** (Motion Energy Image) [62]: We set the positions where movements were generated as $\cup_{t=2}^{T} |\boldsymbol{m}_t - \boldsymbol{m}_{t-1}|$. Figure 2.6 (c) shows an example of an MEI.
- **C** (Cepstrum) [63]: We applied cepstrum analysis to the temporal change in local movements. We used frequencies from the DC component to the 1100-th component for a feature.
- **MFCC** (Mel-frequency Cepstrum Coefficients) [64]: We computed a feature using 40 coefficients.

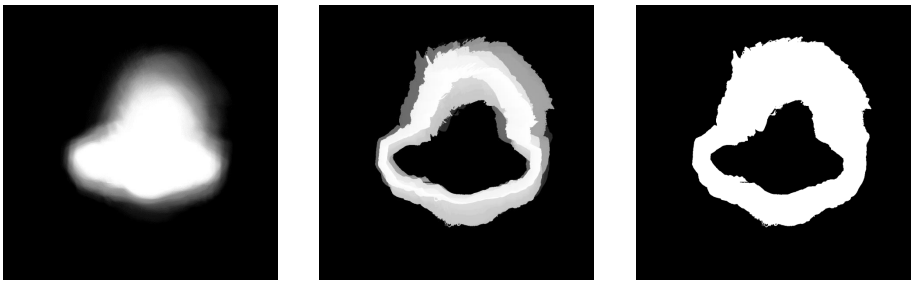


Figure 2.6: Examples of GEI, MHI and MEI computed from the video sequence of the same participant.

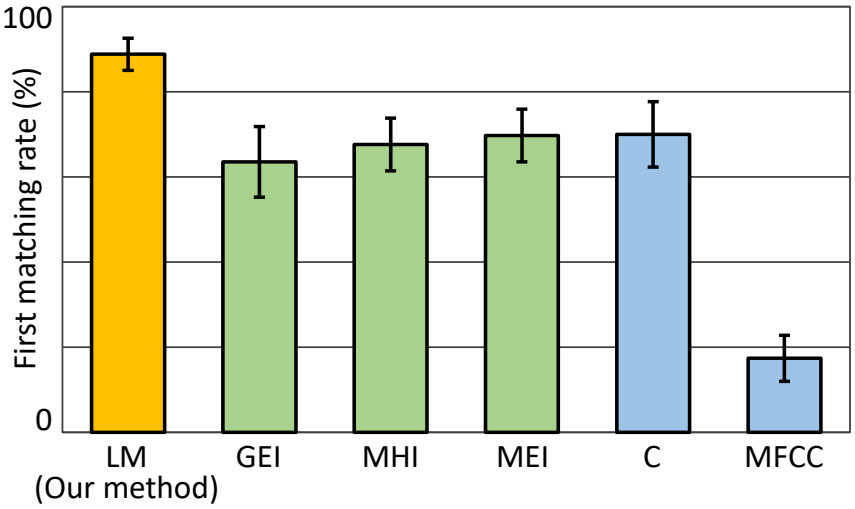


Figure 2.7: Comparison of the first matching rate achieved with the local movement feature obtained using our method and those obtained using existing methods.

The same experimental conditions were used for the query and target sequences as described in Section 2.2.2. Note that C and MFCC were computed from 25 local regions in our method.

Table 2.1: Comparison of the numbers of correctly identified queries. The total number of queries was 1770 (59 individuals, 5 sets, 6 permutations).

	GEI (Correct)	GEI (Wrong)
LM (Correct)	1017	**553**
LM (Wrong)	**107**	93

Figure 2.7 compares the identification performance achieved using features extracted using our method and existing methods; the figure shows that LM outperformed GEI, MHI, and MEI. We believe that the performances of the other methods were lower because they cannot represent small movements of the body: GEI was designed for gait recognition, which involves large limb movements, while MHI and MEI were designed for action recognition with dynamic movement of the body. The performances of MHI and MEI were almost equivalent, while the performance of GEI was lower. Table 2.1 shows that LM correctly identified more queries than GEI. Returning to Figure 2.7, we see that LM outperformed C and MFCC. We believe that the lower performances of C and MFCC were because these methods were designed for gaze authentication with abrupt and rapid movements of the eyes, whereas body sway is characterized by low-frequency components over a longer time. We also evaluated the identification performance using the CMC (Cumulative Match Characteristic) curve, which represents the j-th matching rate. The results shown in Figure 2.8 confirm that our method outperforms existing methods in identifying people using body sway.

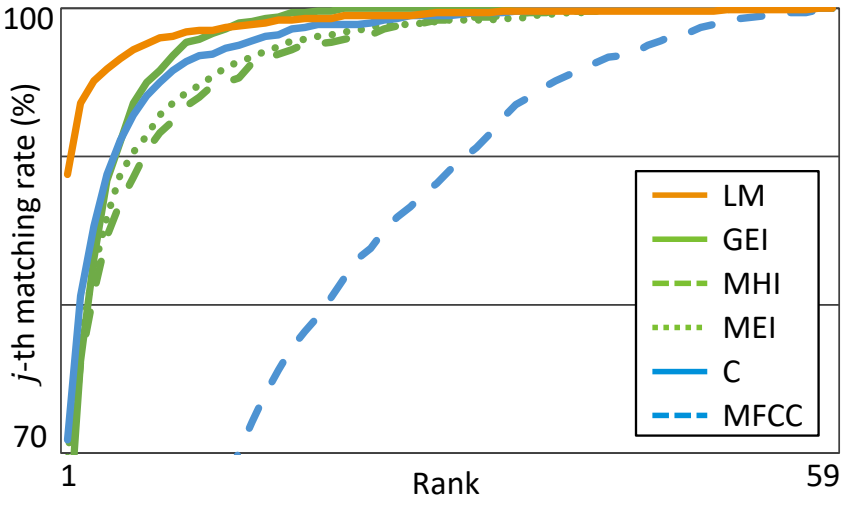


Figure 2.8: Comparison of CMC curves achieved with local movement features obtained using our method and those obtained using existing methods.

2.2.4 Frequency analysis of temporal changes in local movements

We evaluated the identification performance using each frequency band obtained from temporal changes in local movements. We used the frequency band at intervals of 3 Hz. We set the same parameters for LM as described in Section 2.2.3.

Figure 2.9 shows the identification performance using each frequency band. We can see that the identification performance using frequency band $[0, 3)$ Hz was higher than that using the different frequency bands. We believe that the low frequency components of temporal changes in local movements contain more identity information than the high frequency components.

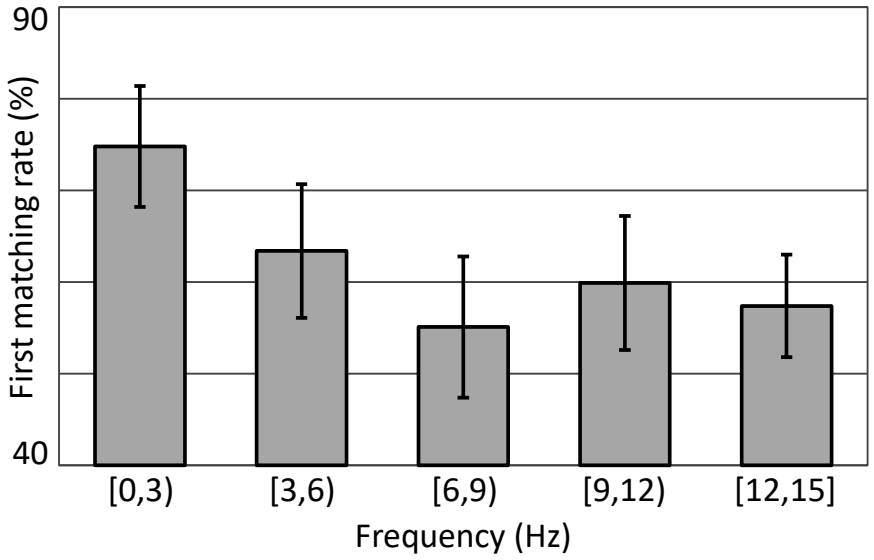


Figure 2.9: Comparison of the identification performance achieved using each frequency band at intervals of 3 Hz.

2.2.5 Improvement of the identification performance by combining likelihoods.

To improve the identification performance, we combined two different characteristics: a likelihood obtained by LM and a likelihood obtained by GEI, MHI, MEI, C or MFCC. We applied the weighted linear sum for a combination of likelihoods. We used the weight α for the likelihood of LM and the weight $1 - \alpha$ for the likelihood of another feature. We set the same feature parameters as described in Section 2.2.3 and $T = 1350$ (45 s) instead of $T = 3600$ (120 s).

Figure 2.10 shows the identification performance using a combination of a likelihood of LM and a likelihood of GEI while changing α. When α was 0 or 1, a single likelihood of GEI or LM was used. The highest

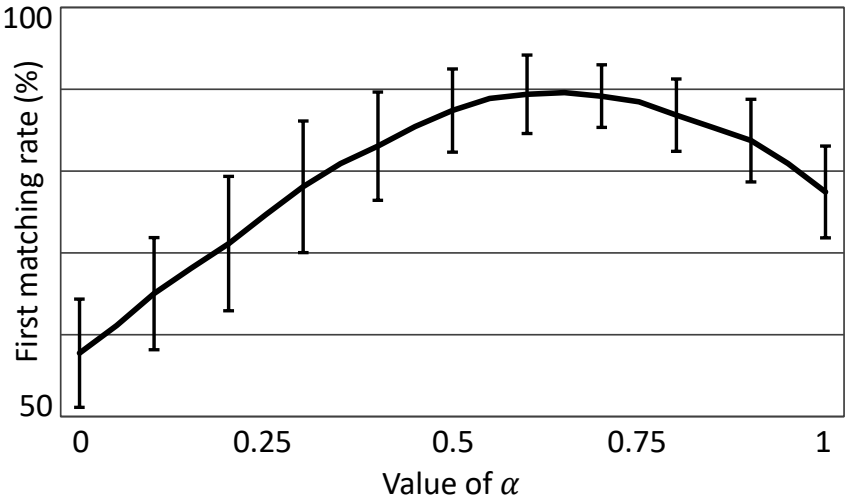


Figure 2.10: Identification performance using the combination of the likelihood of LM and the likelihood of GEI while changing α.

identification performance was $89.6 \pm 4.2\%$ using $\alpha = 0.65$. The performance using a combination of likelihoods was higher than that using each likelihood. Interestingly, a combination of likelihoods using $T = 1350$ obtained almost the same performance as a single likelihood of LM using $T = 3600$.

Figure 2.11 shows the identification performance using a combination of a likelihood of LM and a likelihood of MHI, MEI, C or MFCC. We set $\alpha = 0.65$ for HEI and MEI and $\alpha = 0.9$ for C and MFCC. The identification performance using a combination with a likelihood of LM was higher than that using any other likelihood. We confirmed that LM improved the identification performance by combining different characteristics.

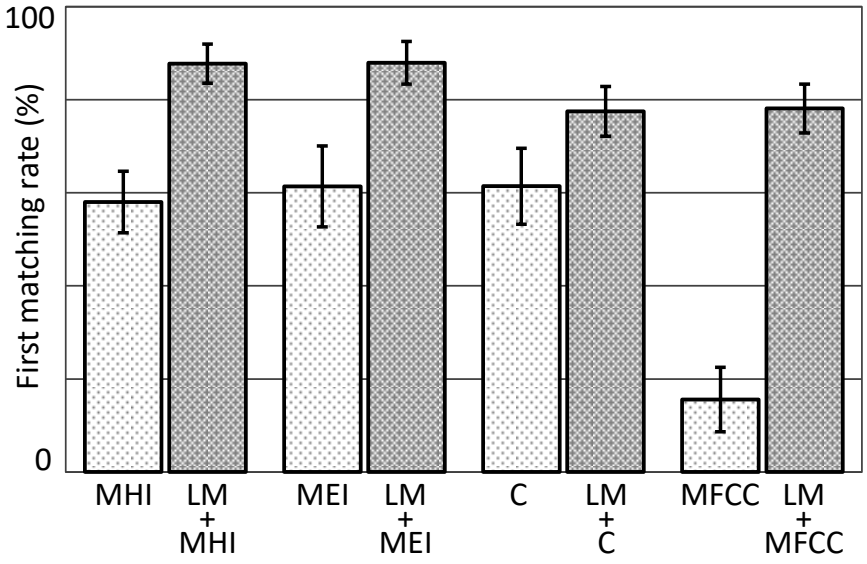


Figure 2.11: Comparison of the identification performance using a combination of likelihoods and using each likelihood. We set $T = 1350$ (45 s).

2.2.6 Evaluation of the variation in identification performance over the long term

We checked the variation in the identification performance over the long term. We collected video sequences for 10 participants (average age 22.6 ± 1.3 years; 9 males and 1 female) using the camera setup in Figure 2.3. We acquired three target video sequences for each participant. After 128 days, we acquired three query video sequences for each participant. We used a single video sequence as a target and a single video sequence as a query for each participant. We generated a metric matrix using 108 participants by removing the 10 test participants from the dataset described in Section 2.2.1. We compared the identification performance of our method with those of GEI, HMI, MEI and C, which obtained better performance in Section 2.2.3.

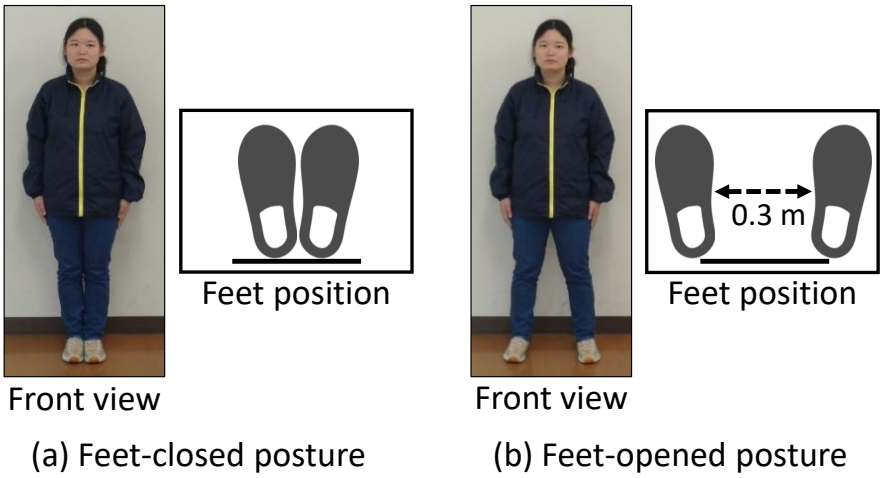


Figure 2.12: Different postures of each participant.

The number of participants correctly classified by LM was 6.7 ± 0.9, compared with 5.4 ± 0.7, 4.1 ± 1.3, 4.8 ± 1.3 and 5.1 ± 1.2 identified by GEI, MHI, MEI and C, respectively. Although this result confirmed that our method performed better than the existing methods, the variation over the long term was still too high. We need to improve the performance to construct a practical application in future work.

2.3 Experiments with different postures

2.3.1 Datasets

We evaluated the identification performance under the condition that the postures of each participant were different between the query and target video sequences. We collected video sequences of body sway from 31 participants (average age 22.2 ± 1.2 years; 22 males and 9 females) using the camera setup in Figure 2.3 (b). Each participant maintained a feet-closed posture as in Figure 2.12 (a) and a feet-open posture as in Figure 2.12 (b), respectively.

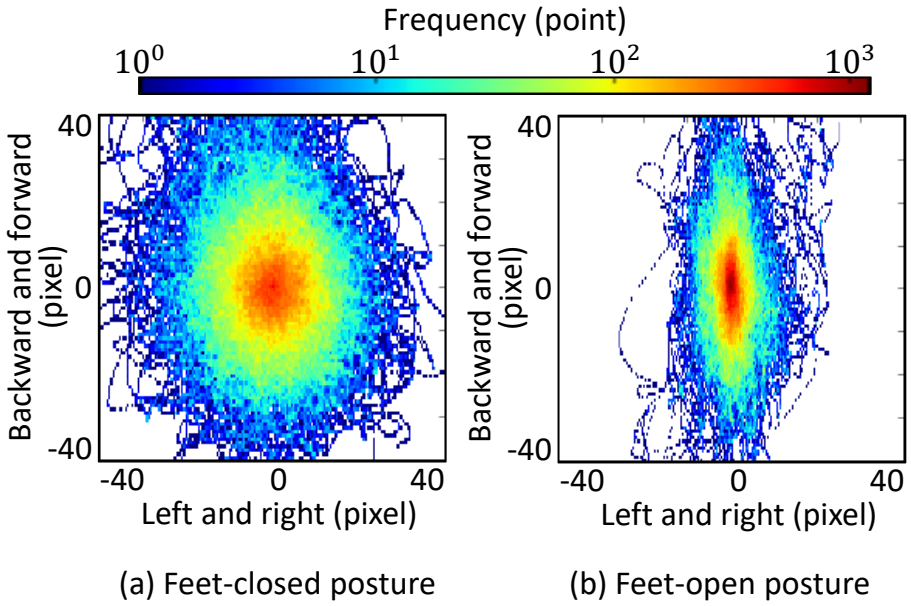


Figure 2.13: Distributions of the positions of the center of the body. The vertical axis shows backward and forward movements. The horizontal axis shows left and right movements. The color bar shows the frequency of appearance.

We observed each participant three times in each posture.

Figure 2.13 shows the distributions of the positions of the centers computed from the body regions of the mask images $\boldsymbol{m}_t$ (t = 1,..., 3600). To determine the origin of the distribution, we used the position of the center of the body region in the reference mask image $\boldsymbol{m}_r$. Each distribution in Figure 2.13 followed a normal distribution. We regarded the distribution in (a) as isotropic and that in (b) as anisotropic. The movements became small in the direction of the straight line between the feet. There seemed to be variation in body sway between the feet-closed and feet-open postures.

Table 2.2: Comparison of the first matching rate (%) of LM, GEI and a combination of LM and GEI when the participants' postures differed between the query and target video sequence.

LM	GEI	LM+GEI
55.0 ± 11.8	62.5 ± 8.1	64.9 ± 11.3

2.3.2 Comparison of the identification performance between feet-closed and feet-open postures

We used a video sequence of the feet-closed posture as the query and a video sequence of the feet-open posture as the target, and vice versa. We removed the 31 test participants from the dataset described in Section 2.2.1 to generate a metric matrix of 87 participants. We used LM and GEI to evaluate the performance.

Table 2.2 shows the identification performance between feet-closed and feet-open postures. The performance of LM was 7.5 points lower than that of GEI. The difference in postures clearly influenced the identification performance of LM. However, because the highest performance was achieved using a combination of LM and GEI, we believe that LM helped to improve the identification performance. However, the performances were not high because of the variation between feet-closed and feet-open postures. We will work on improving the performance to construct a practical application in future work.

2.4 Conclusions

We proposed a method of identifying people using video sequences of body sway. We designed a feature extraction method for identification by measuring temporal and spatial changes in local movements. To evaluate our method, we originally collected three novel datasets containing video sequences of body sway. The first dataset included 118 participants in an upright posture, the second included the variation over 128 days and the third included variation in feet-closed and feet-open postures. We confirmed that our method can extract informative features from video sequences of body sway for the identification of people.

As part of our future work, we intend to increase the tolerance of our method when there is variation over the long term and variation in postures. Furthermore, we plan to develop a practical application by reducing the time required to measure body sway.

Chapter 3

Identifying people using body sway in case of self-occlusion

3.1 Introduction

The widespread use of surveillance cameras is expected to help further develop biometric authentication systems [38, 69]. To identify people accurately from images captured through such cameras, behavioral characteristics [54, 55, 70] have been considered in research on biometrics as they can be used to identify people based on their movements. Features of the gait [54, 55] represent identities as reflected in periodic movements of such parts of the body as the limbs, and have been used as representative behavioral characteristics for identifying people with high accuracy. However, gait features do not adequately represent identities encapsulated in body movements in certain cases, e.g., when people are stationary, because periodic movements of the body parts no longer occur. Therefore, body sway [70] has been recommended for use in identifying people when they are not moving. Body sway is defined as continuous, slight, and unconscious movements of the body to maintain pose even when a person is otherwise not moving. People can be identified using these slight movements. Note that we consider an upright pose to be a typical example of the pose of a person who had been walking but has now stopped. Body sway can be used to identify people who maintain an upright pose, say, in front of a security gate or an automatic door. People who work in factories, for one, appear very similar because they wear

a uniform. The aim in such cases is to accurately identify people using body sway when their appearances are similar.

To the above end, we need to extract appropriate features contained in body sway in both the spatial and the temporal domains. The identity in the spatial domain lies in the shape of the body and that in the temporal domain in the movement of the entire body. In the following, we consider how to obtain identities using body sway in the spatial and temporal domains by using images acquired from surveillance cameras. In this scenario, we observe the shape of the body in spatial domain as a person's appearance, and the movement of the entire body in the temporal domain as sequential changes in their appearance. To appropriately represent identity as reflected by body sway, a person's accurate appearance needs to be acquired in images from the camera. However, defects in this appearance are common when occlusion occurs, and depend on the relationship between the position of the camera and that of the person being photographed. This problem needs to be solved.

We examine why occlusion occurs when we measure body sway. There are two main types of occlusion. The first type occurs when an individual stands in front of another. In this case, part of the appearance of the person far from the camera is hidden by the one close to it. This phenomenon is called mutual occlusion, and occurs when in case of a large number of people. The use of a top-view camera reduces the occurrence of mutual occlusion. The second type of occlusion is one where part of a person's own body obstructs the sight of him/her. This phenomenon is called self-occlusion, and occurs even when the top-view camera is used. Therefore, we need to consider how to reduce the influence of self-occlusion for identifying people using body sway.

The region around the head is the most robust against the influence of self-occlusion when using a top-view camera. Some prevalent methods use regions of the head acquired using a top-view camera to count the number

of people in a given image [71, 72, 73] or to track people's walking routes [74, 75]. However, regions around the head have not been used to aim to identify people. Another such method [70] does not use the region around the head, although it is designed to identify people using body sway. This method causes the accuracy of the identification to decrease dynamically in case of self-occlusion because it uses whole-body regions to obtained features used to identify people.

To this end, we propose a method of representing identities as reflected in body sway by using the region around the head acquired using a top-view camera to accurately identify people in case of self-occlusion. Our method computes silhouette images around the head regions by applying a segmentation technique. To represent identities contained in body sway, we spatially divide the head regions into local blocks and temporally measure movements in these blocks. In this way, we can appropriately represent identities reflected in body sway in the spatial and temporal domains. We formed a dataset of images of body sways of 50 participants with self-occlusion. The results of experiments to verify the proposed method show that it can improve the accuracy of identification of prevalent methods, which use images of whole-body regions, from 17.3% to 57.9% by using only images of regions around the head. The remainder of this paper is organized as follows: Section 2 explains the influence of self-occlusion, and Section 3 describes our method of extracting features for identification contained in the body sway using images of regions of the head, Section 4 details identification performance when using body sway, and Section 5 presents the conclusions of this study.

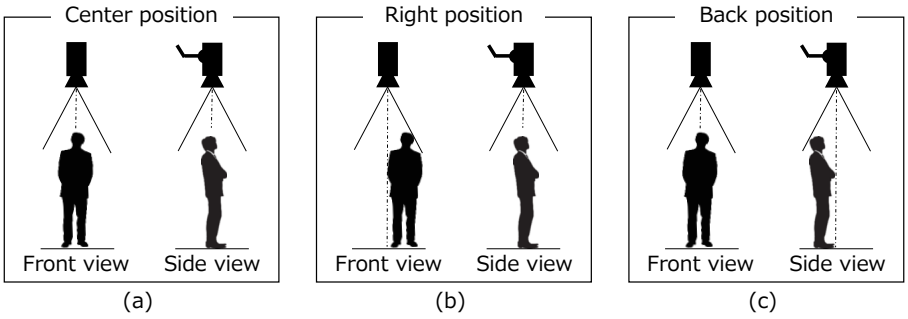


Figure 3.1: The standing positions of a person used to investigate the influence of self-occlusion.

3.2 The influence of self-occlusion

The appearances of an individual acquired from a top-view camera depend on his/her standing position when self-occlusion occurs. In a preliminary experiment, we compared the appearances of an individual in different standing positions. The position of the top-view camera was fixed, as shown in Figure 3.1. We defined the point where the optical axis of the camera was orthogonal to the floor as the center. Figure 3.1 (a) shows the condition when the person standing at the center was observed, and Figures 3.1 (b) and (c) show conditions of observation of people standing to the right and behind the center, respectively.

Figure 3.2 shows examples of the appearance of the entire bodies of two people acquired in three standing positions, where the upper row shows individual 1 and the lower row shows individual 2. In comparison with Figures 3.2 (a), (b), and (c), we see that the appearances of the entire body acquired from each standing position were different. We also describe the head regions used in this paper. Figure 3.3 shows examples of head regions acquired under the same observation conditions as in Figure 3.2. The

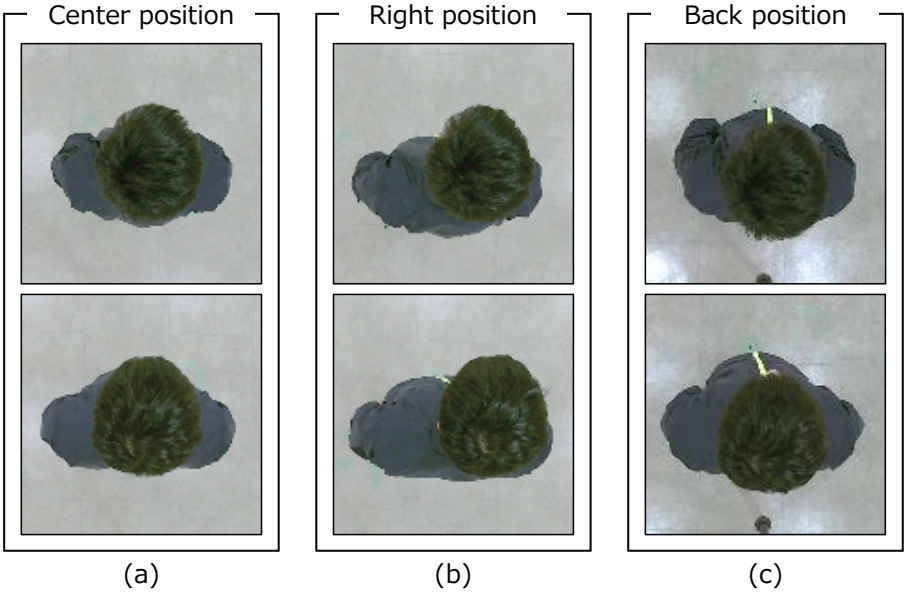


Figure 3.2: Examples of the appearance of entire body acquired from two people standing in three different positions.

green pixels in the images represent the head regions. A comparison of Figures 3.3 (a), (b), and (c) shows that the head regions acquired from each standing position were similar. We also examined regions of the shoulders, which changed in each standing position due to self-occlusion as shown in Figure 3.3. Regions of the left and right shoulders were symmetrically at the center as shown in Figure 3.3 (a). However, in Figure 3.3 (b), the parts of regions of the right shoulder acquired at the center are hidden by the head regions, and parts of regions of the left shoulder hidden by the head regions at the center are acquired. The same tendency can be observed in Figure 3.3 (c). Therefore, the head regions are more robust against the influence of self-occlusion than any other region of the body. We thus use them for identifying people based on body in case of self-occlusion.

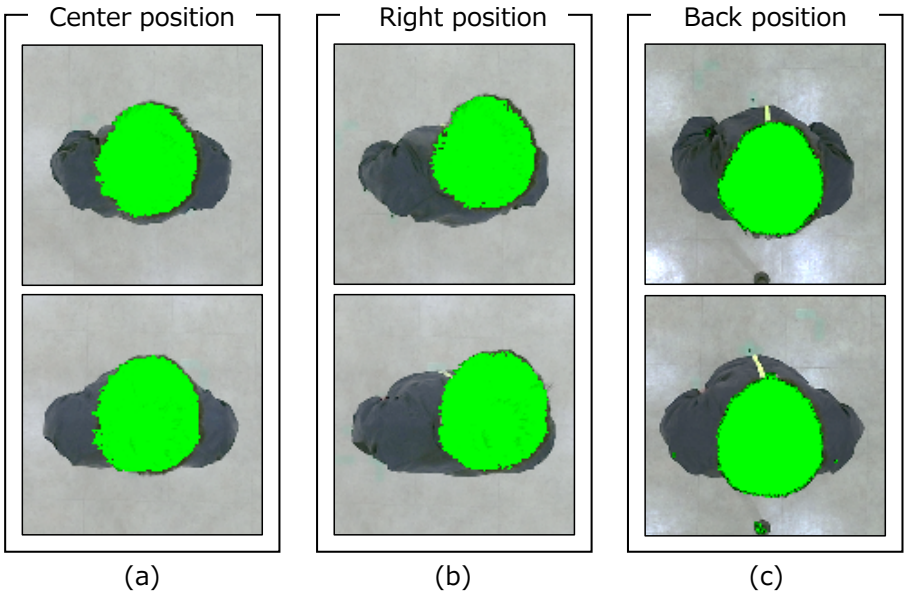


Figure 3.3: Examples of head regions acquired from two people in three different standing positions. The green pixels represent the head regions.

3.3 Our method

3.3.1 Overview

We propose a method to extracting spatio-temporal features from images using region of the head to identify people based on body sway. Figure 3.4 provides an overview of our method. We acquire a set of images of a person by using a top-view camera while he/she maintains an upright pose. To reduce the influence of self-occlusion, we compute silhouette images of the head regions from this set by applying a segmentation technique. To extract features for identification, we spatially divide the head regions into local blocks and temporally measure movements in each local block. The details of our method are described below.

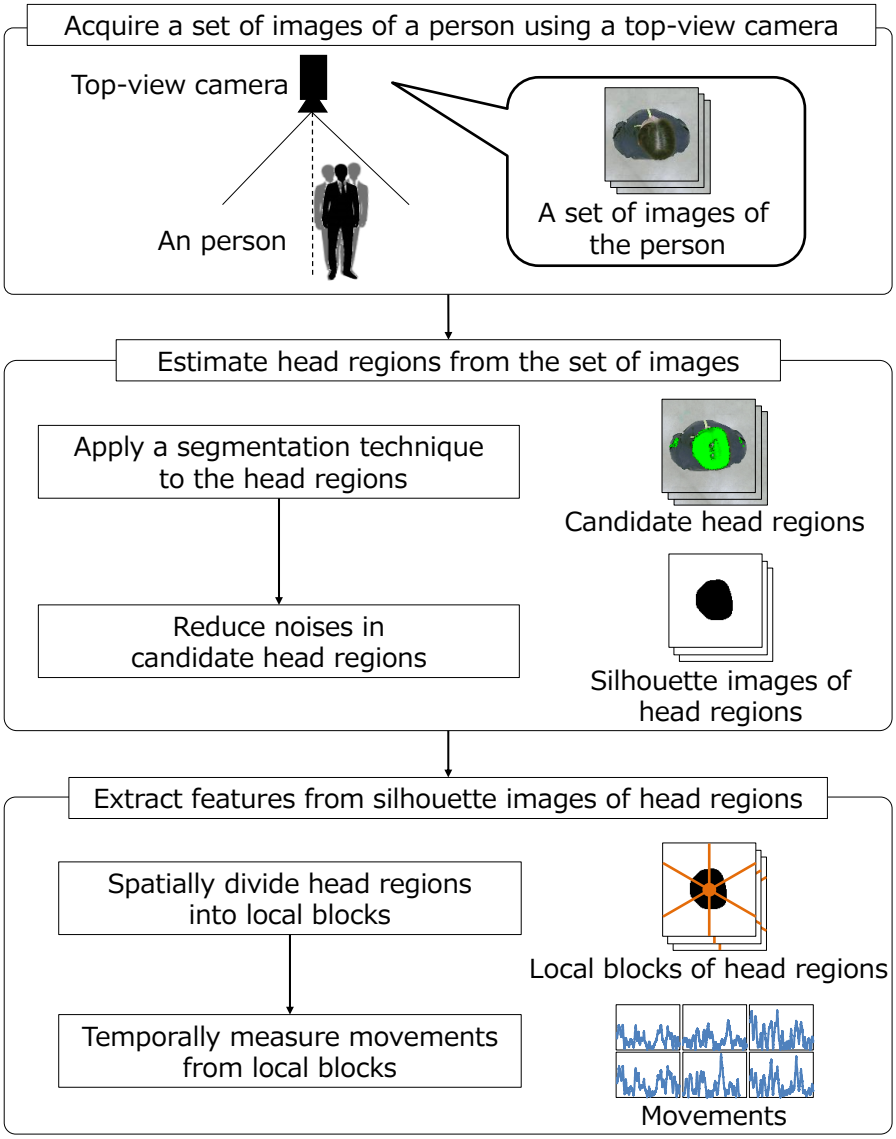


Figure 3.4: Overview of our method.

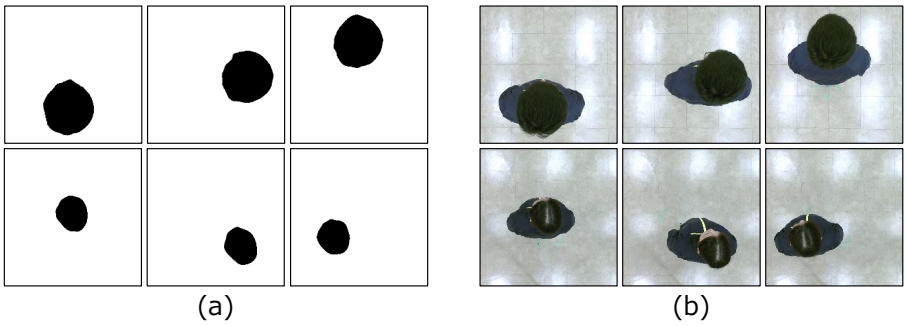


Figure 3.5: Examples of annotation labels of head regions and images of people used to train a network model for head segmentation.

3.3.2 Estimating head regions from a set of images of a person

The head regions can be estimated accurately by statistically learning using a large number of training images featuring variations in the appearance of people. Various segmentation techniques are available based on statistical approaches [76, 77, 78, 79, 80, 81]. Segmentation methods that use deep learning techniques [82, 83, 84] have been popular in recent years as they are highly accurate. We prepared a large number of pairs of images of people with the head regions annotated to train a network model for segmentation. Figure 3.5 (a) shows examples of the annotation labels of the head regions, and Figure 3.5 (b) shows examples of the images of people that were used. The trained network model output candidate head regions.

The candidate head regions estimated by deep learning techniques contained noise. Some pixels of the head regions were incorrectly identified as pixels belonging to other regions of the body, and some belonging to other regions were incorrectly identified as belonging to the head region. Figure 3.6 (a) shows examples of candidate head regions containing noise. To

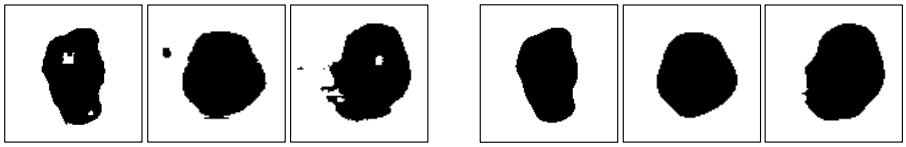


Figure 3.6: Examples of candidate head regions containing noise, and silhouette images of these regions having reduced reducing noise.

reduce it, we selected the largest regions from the candidate head regions in a single image and corrected all pixels in them. We reduced noise around a boundary between the head region and background regions by using a median filter. Figure 3.6 (b) shows examples of silhouette images of the head regions after noise had been reduced.

3.3.3 Extracting a spatio-temporal feature from silhouette images of head regions

The proposed method to extract spatio-temporal features from silhouette images of the head regions extends our previous method [70]. The head slightly moves as a person maintains an upright pose, where this movement occurs around a center acquired at a reference time. To obtain this reference time, we select the silhouette image of a person most similar to each silhouette image of the same person, and set a time acquired it as the reference time. To represent identity in the spatial domain, we radially divide each silhouette image into local blocks using the central position of the head at the reference time. To represent identity in the temporal domain, we compute movements over time from the local blocks. To extract features for identification, we estimate the power spectral density (PSD) [68] of each movement.

3.4 Experiments

3.4.1 Dataset

To evaluate the validity of our method, we collected sets of images of the body sways of 50 participants (average age, 22.7 $\pm$ 3 years; 42 males and eight females) using a top-view camera as they stood in different positions. Each participant maintained an upright pose (Romberg posture) with the limbs aligned. We asked all participants to wear the same dark-blue nylon outerwear similar to the uniform worn by factory workers. Figure 3.7 (a) shows the examples of poses and clothes. We set-up a top-view camera at a height of 2.5 m from the ground, and calibrated it such that the optical axis coincided with the direction normal to the floor. We used a set of images captured at 30 fps by Microsoft Kinect V2, where each image size was 1920×1080 pixels.

Figure 3.7 (b) shows five standing positions set on the floor. We set as center the point where the optical axis of the top-view camera was orthogonal to the floor. We set the remaining four standing positions as points that were shifted to the front, back, left, and right from the center by 0.15 m, respectively. Circle markers were set on the floor to indicate each standing position. We asked all participants to stand so that the center of his/her feet corresponded to the circle marker as shown in Figure 3.7 (c). Figure 3.7 (d) shows the setup for acquiring a set of images of the body sway when a participant stood in the front. We asked each participant to look at a target point 3 m away to fix the direction of the head. We set the target point in front of the participant in each standing position. The time needed to acquire a set of images was 60 seconds for each standing position. We observed each participant two times in five standing positions. They were allowed to sit and rest between observations. The order of standing positions was random.

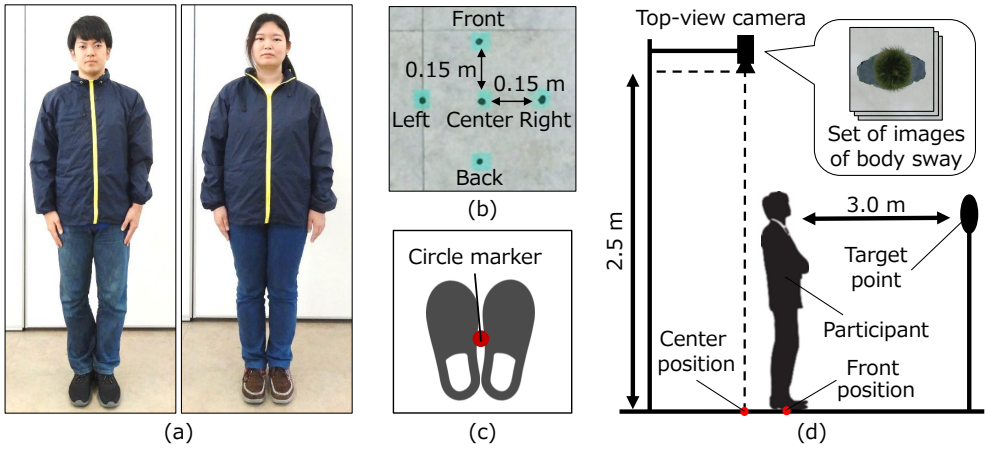


Figure 3.7: The conditions under which each participant was observed. (a) shows their poses and clothes, and (b) shows their standing positions set on the floor. (c) shows the circle marker to align the position of the feet of the participants, and (d) shows the setup used to acquire a set of images of the body sway.

We cropped the 1920×1080-pixels images of all participants to 1080×1080 pixels, and resized them to 256×256 pixels.

3.4.2 Assessing accuracy of estimating head regions

We evaluated the accuracy of estimating head regions from images of people using a top-view camera. We applied U-net [83], which was used in research [74], to estimate the head regions. We set eight down-sampling layers and eight up-sampling layers in the U-net architecture. To train the U-net, we randomly selected 25 participants from the dataset described in Section 3.4.1. Data for the remaining 25 participants were used to test the performance of the proposed method. We repeated the random selection five times, and used 45,000 pairs of images and annotation labels of head regions

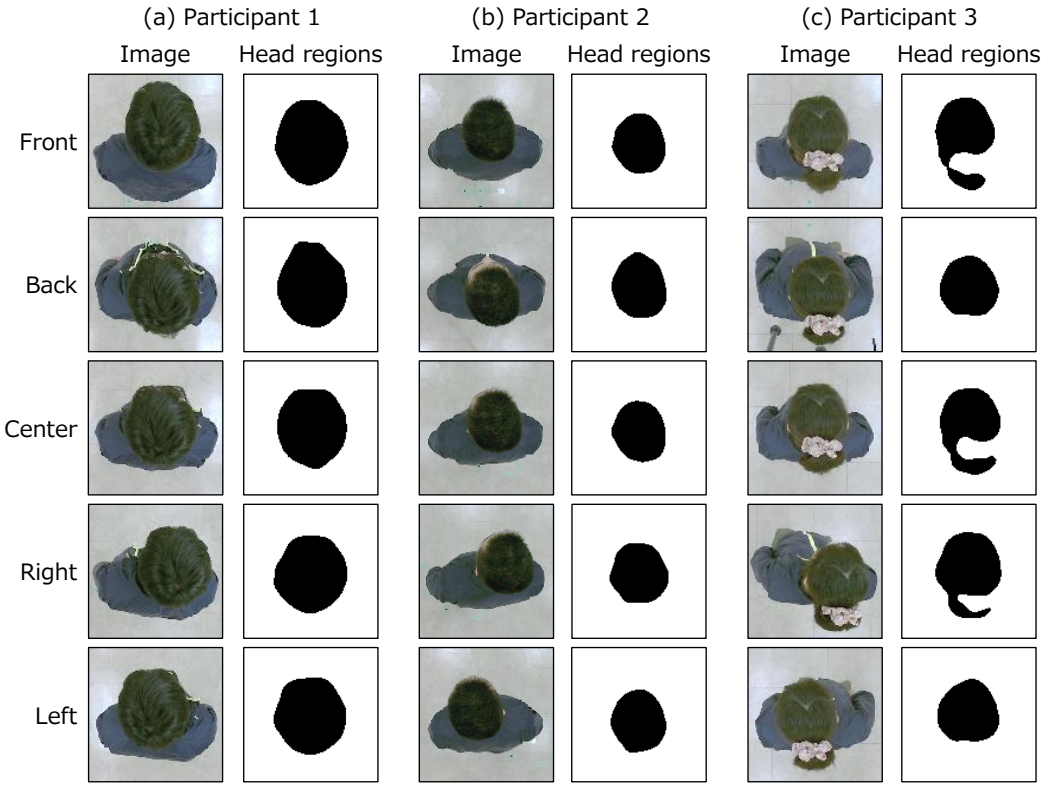


Figure 3.8: Examples of head regions estimated from images of three participants in five standing positions using U-net.

to train the U-net. The sizes of both the images and the annotation labels were set to 256×256 pixels, and the number of epochs of training was set to 200. To evaluate the accuracy of the proposed method to estimate head regions, we used the F-measure, which is the harmonic mean of precision and recall. A value of 1 indicates the best result that that of 0 the worst.

The proposed method recorded an accuracy of 0.96 ± 0.03 in terms of estimating the head regions. Figure 3.8 shows examples of head regions estimated for images of three participants in five standing positions using U-net. It is clear that the head regions in Figures 3.8 (a) and (b) were estimated

with high accuracy in all positions. The appearances of the head regions in Figure 3.8 (c) were different in each position. Although a part of the head regions was incorrectly estimated, the mean value of the F-measure was close to 1. Thus, the results were accurate.

3.4.3 Evaluation of identification performance

We assessed whether our method can be used to identify people from images of the head regions in case of self-occlusion. We compared the head regions obtained using it with other regions of the body to this end. The experimental conditions were as follows.

Head: We used head regions estimated by our method. Figure 3.9 (a) shows examples of them.

Whole body: We used whole-body regions as used in a prevalent method [70]. Figure 3.9 (b) shows examples of them.

Shoulder: We used shoulder regions excluding the head regions from entire-body regions. Figure 3.9 (c) shows examples of them.

We estimated the whole-body regions and shoulder regions by applying the method described in Section 3.3.2. We extracted features to identify people from silhouette images of each body part by applying the method described in Section 3.3.3. We set the number of blocks to spatially divide regions of each body part to 25. We selected a set of silhouette images at the center as query, and a set at a position other than the center as target. We also evaluated the performance of the proposed method when switching the query with the target. We used the nearest-neighbor algorithm to identify people from the images and the first matching rate to assess

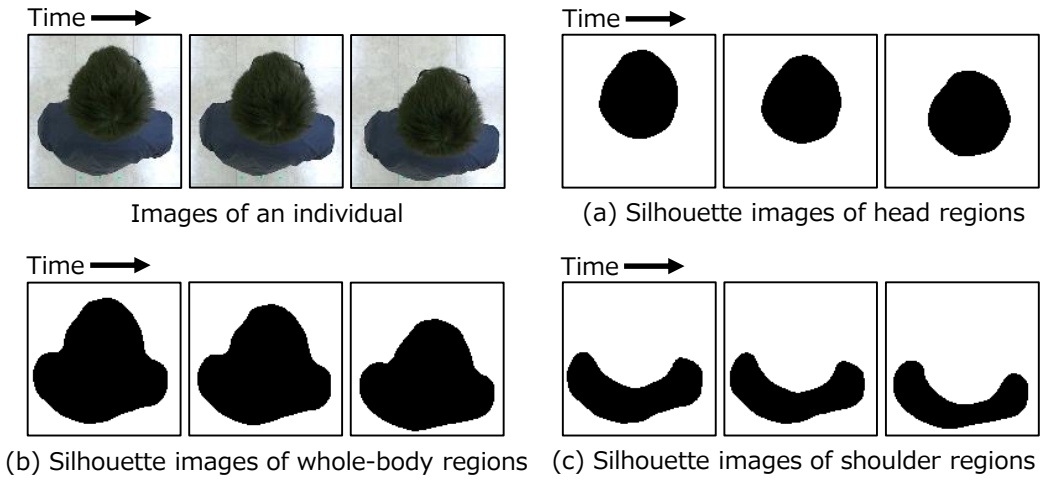


Figure 3.9: Examples of the silhouette images of each body part.

performance. The proposed applied a metric learning technique, the large-margin nearest-neighbor (LMNN) method [65]. We randomly selected 25 participants not from the target and the query, and used them for LMNN. Data for the remaining 25 participants were used for identification. We repeated the random selection five times.

Table 3.1 shows the identification performance of the proposed method when it used regions of each body part. Using the head regions as in our method yielded better performance than whole-body regions and shoulder regions. The worst performance was obtained when using shoulder regions (that excluded the head regions). Therefore, the head regions were more robust to self-occlusion than whole-body regions and shoulder regions when using a top-view camera to identify people.

Table 3.1: Comparison of identification performance using regions of each body part.

Region	First matching rate (%)
Head	**57.9 ± 11.1**
Whole body	17.3 ± 6.8
Shoulder	9.8 ± 5.3

3.4.4 Performance comparison when using spatial features and temporal features

To determine whether the spatio-temporal features extracted by our method were valid, we compared its performance when using spatio-temporal features with the results obtained when using only spatial features and only temporal features. We extracted each set of features from the same head regions. The experimental conditions were as follows.

Spatio-temporal features: We extracted the spatio-temporal features from the set of silhouette images of the head regions using our method.

Spatial features: To extract features in the spatial domain from the head regions, we selected a single silhouette image from the set of silhouette images, and used it at the reference time as described in Section 3.3.3.

Temporal features: To extract features in the temporal domain from the head regions, we computed the central position of the head regions in a silhouette image and measured the central positions of the entire set of silhouette images. We used the temporal changes in the central positions as features for identification.

Table 3.2: Comparison of the identification performance of the proposed method when using spatio-temporal features, only temporal features, and only spatial features.

Feature	First matching rate (%)
Spatio-temporal	**57.9 ± 11.1**
Spatial	33.8 ± 10.7
Temporal	40.2 ± 9.9

The experimental conditions except for the features used were the same as described in Section 3.4.3.

Table 3.2 shows the performance of the proposed method when using spatio-temporal features, only temporal features, and only spatial features. It is clear that it performance in terms of identification was superior when using only temporal features than when using only spatial features. It is also evident that the spatio-temporal features extracted by the proposed method yielded the best performance. Thus, extracting features from both the spatial and the temporal domains is the best means of accurately reflecting features of body sway in the head regions.

3.4.5 Comparison of proposed method with prevalent methods

We compared the performance of the proposed method with that of prevalent methods in terms of identification. The GEI [54] and STHOG [85] methods are widely used to authenticate gait, and were chosen along with the dynamic image method [86], which is used in action recognition, for comparison with the proposed method. To extract the GEI, we computed the average image from the silhouette images for 60 seconds. To extract the STHOG, we set

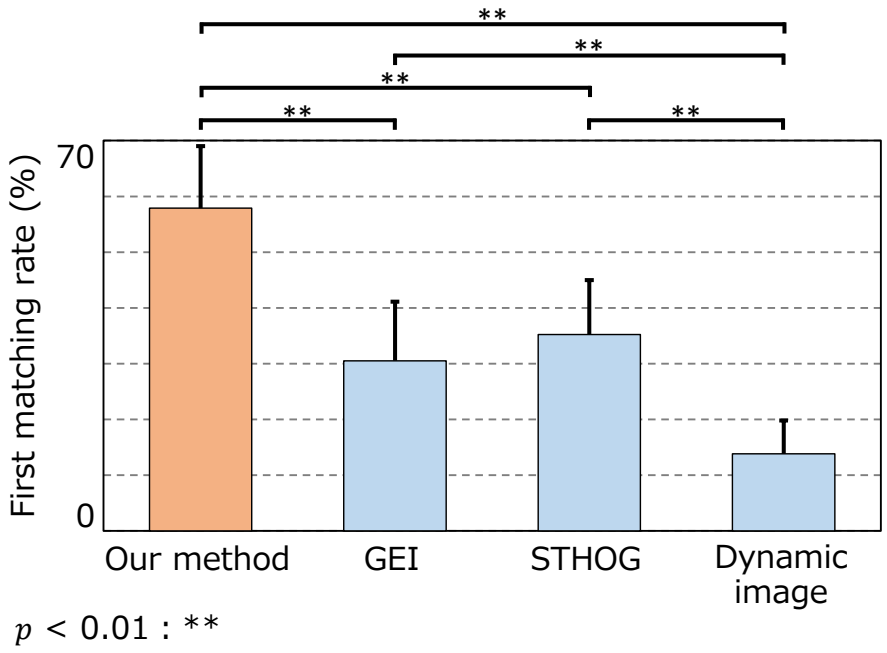


Figure 3.10: Comparison of the identification performance of our method, which uses spatio-temporal features, with prevalent methods.

the number of spatio-temporal blocks to $6 \times 6 \times 6 = 216$, and computed the gradients. To extract the dynamic image, we applied rank SVM to the silhouette images of the head regions. The experimental conditions, except for the spatio-temporal features used, were the same as described in Section 3.4.3.

Figure 3.10 compares the identification performance of all methods. It is clear that our method outperformed all other methods. The results of the Wilcoxon signed-rank test and the Bonferroni correction verify this. A significant difference was observed between our method and GEI. The same tendencies were observed for the STHOG, and dynamic images.

3.5 Conclusions

In this paper, we proposed a method to identify people using their body sway in the spatial and temporal domains by using head regions acquired from a top-view camera in case of self-occlusion. To estimate head regions from the set of images of a person, we applied a method of segmentation using deep learning technique and reduced noise in the candidate head regions chosen. To represent identity-related information reflected in the body sway, we spatially divided the head regions into local blocks and temporally measured movements in these blocks. To evaluate our method, we formed a dataset containing images of people, with a focus on their body sways in the presence of self-occlusion. The results of experiments showed that the proposed method, using head regions, outperforms a prevalent method, which uses the whole-body region.

In future work, we intend to represent identities reflected in the body sways of people in spite of occlusion due to headwear, such as a hat or helmet. Furthermore, we plan to reduce the time needed to observe the body sway.

Chapter 4

Gender classification using video sequences of body sway recorded by overhead camera

4.1 Introduction

There is high demand for technology that can classify the gender of a person based on a video sequence [87, 88, 89]. Such gender classification has various applications, such as security surveillance and marketing planning. To accurately classify the gender of a person, the characteristics that distinguish between females and males must be obtained. The movements of a person in a video sequence have recently been considered for representing such characteristics.

In general, the movements of a person can be divided into walking movements (gait) and standing movements (body sway). Below, we review methods that classify the gender of a person based on walking or standing movements in a video sequence. We first consider gender classification based on gait. To distinguish between females and males based on gait, some methods [56, 57, 58] extract the gait energy image (GEI) as a feature for training a gender classifier. It has been reported that GEIs can be used to classify the gender of a walking person with high accuracy. However, methods based on GEIs are designed for classifying the gender of a walking person. To the best of our knowledge, there are no existing methods for gender classification

based on body sway. Here, we propose a method for extracting a feature from body sway and investigate whether it can be used for gender classification.

We discuss whether body sway can be used to distinguish between females and males. Analytical research in the medical field has shown that there are differences between standing females and males in terms of body sway. Analytical studies [90, 91, 92] have used time-series signals of the center positions of the pressure of the feet acquired from a force plate placed on the floor. They demonstrated that there are significant differences between females and males in terms of the frequency characteristics and trajectories of the time-series signals. These studies focused on obtaining medical data on body sway and did not consider practical applications. To apply such medical data for gender classification, a contact-type sensor must be placed on the floor.

Here, we observe body sway using a camera, which is non-contact-type sensor. Previous studies [93, 94, 95, 96, 70] have measured body sway using a camera instead of a force plate for applications such as fall prevention assessment, avatar video generation, and person re-identification. However, the features of body sway were not used to distinguish between females and males.

Here, we investigate whether body sway can be used to classify the gender of a standing person by extracting a feature from a video sequence. We used an overhead camera attached to the ceiling in our experimental setting. We assumed that the head of a standing person makes larger movements than those of the legs and waist. An overhead camera can observe upper body sway, including that of the head. In our method, we estimate the upper-body region in a video sequence to obtain a silhouette sequence. We measure the time-series signals of body sway from the silhouette sequence and extract a feature for gender classification. We created a dataset of video sequences of the body sway of 60 participants to evaluate gender classification accuracy.

We found that our method obtained $90.3 \pm 1.3\%$ accuracy for gender classification on our dataset. We also compared the accuracy of our method with that of features derived from medical data and found that our method has superior accuracy. To the best of our knowledge, the use of body sway in video sequences for gender classification has not been previously reported. Our main contribution is the development of a method for gender classification based on body sway. The remainder of this paper is organized as follows. Section 4.2 reviews related work. Section 4.3 describes our method and Section 4.4 shows the experimental results of gender classification. Finally, Section 4.5 presents the conclusions.

4.2 Related Work

4.2.1 Video Sequences of Walking People for Gender Classification

To classify the gender of a walking person in a video sequence, some methods [56, 57, 58] use GEI features extracted from gait. A GEI feature [54] is represented by an average image calculated from a silhouette sequence containing the movements of arms and legs during one gait cycle. Shan et al. [56] applied GEI features directly to gender classification. Martín-Félez et al. [57] temporally divided one gait cycle into four intervals and extracted a GEI feature from each interval for gender classification. Yu et al. [58] assumed that the movements of arms and legs affect gender classification accuracy. Their method extracts a GEI feature that represents the movement of each body part and assigns an adaptive weight to each feature. Various methods [56, 57, 58] assume that arms and legs provide the most information. However, the body sway of a standing person rarely includes large movements of the arms and legs. In this paper, we extract a feature from

body sway for gender classification.

4.2.2 Use of Single Images for Gender Classification

Some methods extract features from a single image for gender classification. Studies [97, 98, 99, 100] have proposed the use of low-level features derived from the colors and gradients in a single image. Other studies [12, 13, 101, 102] applied a convolutional neural network (CNN) to extract features from a single image in an end-to-end framework. These methods achieve high accuracy in gender classification when trained using a large number of images. Here, we increase gender classification accuracy by incorporating single images with temporal movements. Convolutional three-dimensional (C3D) [103] features are well-known spatio-temporal features. Xu et al. [104] and Liu et al. [105] reported that C3D features are useful for action recognition for classifying large movements, such as soccer shots, table tennis shots, and swimming strokes. However, C3D features are not designed for gender classification. We thus extract a spatio-temporal feature from body sway that are suitable for gender classification.

4.2.3 Analytical Research on Differences between Female and Male in Terms of Body Sway

Analytical studies [90, 91, 92] have been conducted to determine the differences between females and males in terms of body sway. These studies obtained time-series signals of body sway using a force plate placed on the floor and reported that there are significant differences in these signals between females and males in terms of frequency characteristics and trajectories [90], the elliptic approximated from trajectories [91], and the specific band of frequency characteristics [92]. However, they did not apply these parameters to gender classification. In preliminary experiments, we found

that we could not achieve high gender classification accuracy using medical data. We thus extract a feature from body sway to accurately classify the gender of a standing person.

4.2.4 Applications of Body Sway in Video Sequences

Using a camera instead of a force plate to measure the body sway of a standing person has various applications [93, 94, 95, 96, 70]. Wang et al. [93] used body sway to evaluate the risk of falling. They observed a person from various directions using multiple cameras and obtained the time-series signals of three-dimensional centers. Nishiyama et al. [94] used body sway to generate a video sequence of an avatar of a person. They observed a person using a camera placed in front of the person and estimated the center position from time-series signals. Yeung et al. [95] and Lv et al. [96] applied body sway to evaluate a person's balance in the clinical field. They analyzed the time-series signals of body joints obtained from Microsoft Kinect. Kamitani et al. [70] applied body sway to identify people. They obtained the time-series signals of body sway recorded by an overhead camera and extracted the feature representing the identity of an individual. The above examples demonstrate that body sway can be used for various applications. However, body sway has not been applied to gender classification. In this paper, we investigate whether body sway can be used to classify the gender of a standing person.

4.3 Proposed Gender Classification Method

4.3.1 Overview

The proposed method can classify the gender of a standing person using a video sequence of body sway. We acquire a video sequence of a standing

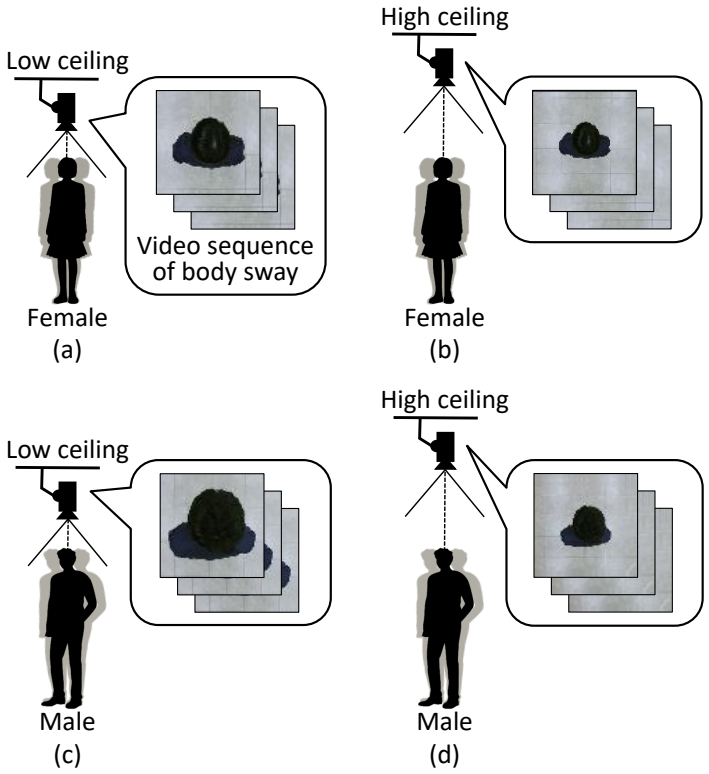


Figure 4.1: Examples of the variation of the apparent size of the upper body in our experimental setting, where the camera height was randomly changed. Female recorded by (a) low and (b) high camera. Male recorded by (c) low and (d) high camera.

person using an overhead camera attached to the ceiling. The overhead camera is used to observe the upper body of a standing person, where the amount of movement is larger than that of the lower body. For the upper body, the head has the largest movement. We extract an informative feature from the upper body for gender classification by acquiring a video sequence of body sway.

Here, we discuss the camera setting used to view the upper body of a standing person. Ceiling height, which varies in real-world scenarios, affects the apparent size of a person and thus the amount of the movement observed from body sway. Fig. 4.1 shows examples of the apparent size of the upper

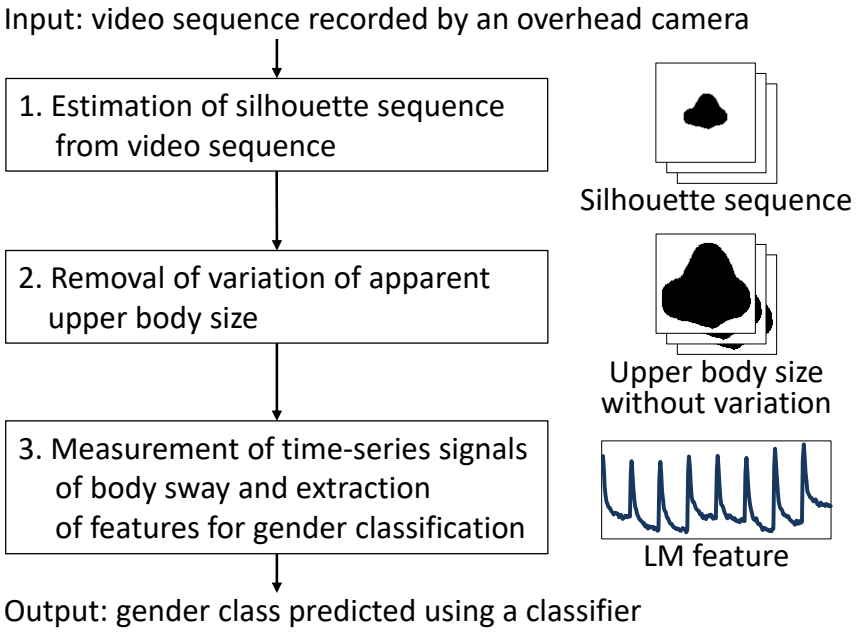


Figure 4.2: Overview of proposed three-step method for gender classification. The input is a video sequence containing body sway recorded by an overhead camera and the output is a gender class predicted using a classifier and an extracted feature.

body. Although Figs. 4.1 (a) and (b) show the upper body of the same female, the apparent sizes are completely different because the ceiling heights are different. The same tendency for males is shown in Figs. 4.1 (c) and (d). We thus develop a method for body sway measurement that does not depend on ceiling (camera) height.

The proposed method consists of the following three steps. We assume that a person is standing below the overhead camera and maintains the same posture. Fig. 4.2 shows an overview of the proposed method. In the first step, we acquire a video sequence of the standing person using the overhead camera and use it to estimate a silhouette sequence that represents the upper body. In the second step, we remove the variation of the apparent size of

Figure 4.3: Examples of silhouette sequence frames estimated from video sequences. The overhead camera was set at different heights. Female recorded by (a) low and (b) high camera. Male recorded by (c) low and (d) high camera. Black and white pixels respectively represent the upper body and background.

the upper body in the silhouette sequence due to the height of the overhead camera. In the third step, we measure the time-series signals of body sway from the silhouette sequence and extract a feature for gender classification. We determine the gender class using the extracted feature and a pre-trained classifier. We describe the removal of the variation in the apparent size of the body region in Section 4.3.2 and the extraction of a feature from body sway for gender classification in Section 4.3.3.

4.3.2 Removal of Variation in Apparent Size of Person in Silhouette Sequence

To accurately classify gender, the intra-class variation of appearance should be small. However, the apparent size of the upper body can increase this variation when the height of the overhead camera varies, as described in Section 4.3.1. A silhouette sequence is also affected by the variation of apparent size. Fig. 4.3 shows examples of the apparent size of silhouette sequences of the upper body. The frames of the silhouette sequences in Figs. 4.3 (a)

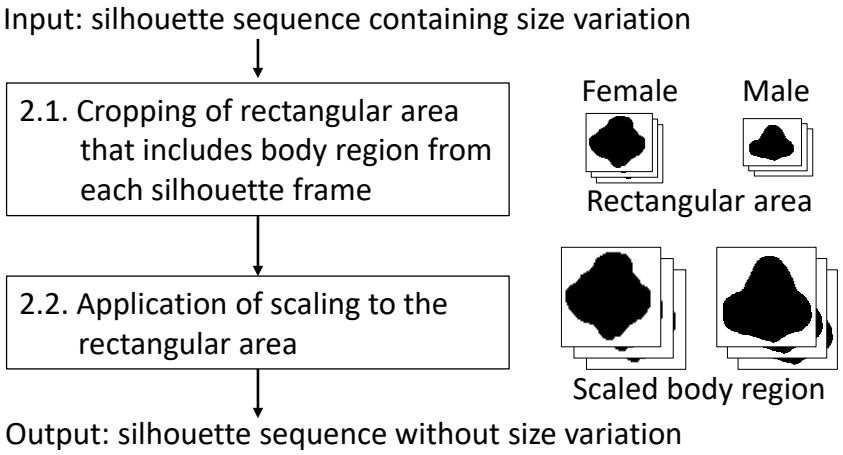


Figure 4.4: Removal of the variation of the apparent size of the upper body in a silhouette sequence.

and (b) are estimated from the same female but different camera heights. Although the silhouette sequences both belong to the female class, their apparent sizes of the upper body are different. The same tendency can be seen for the male class in Figs. 4.3 (c) and (d). If we do not consider the variation of the apparent size, gender classification accuracy will be low because the intra-class variation of appearance will be large.

We thus remove the variation of the apparent size in our method, as shown in Fig. 4.4. In this step, we crop a rectangular area from each input silhouette frame to remove the background region. The rectangular area includes the upper body and has a margin to prevent cutting off the upper body. We set this margin based on the maximum amount of movement. We determine the margin for each silhouette sequence. Finally, we apply a scaling technique so that the height and width of the rectangular area are equal to the reference values H and W, respectively.

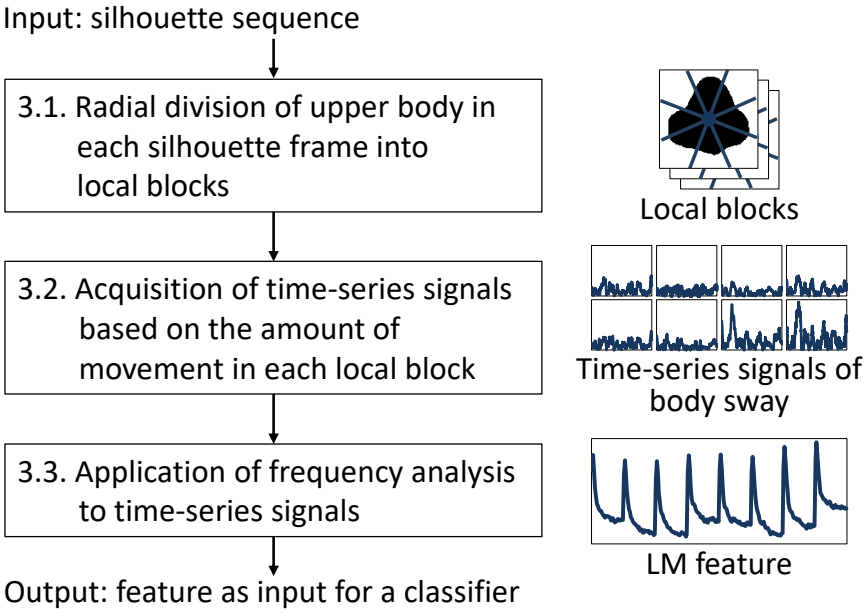


Figure 4.5: Feature extraction step. An LM feature is extracted from a silhouette sequence of body sway.

4.3.3 Extraction of a Feature from Body Sway for Gender Classification

We now describe the extraction of a spatio-temporal feature from body sway for gender classification. Our method is inspired by the framework of an existing method [70] for person re-identification. Fig. 4.5 shows the feature extraction step. The upper body in each silhouette frame is radially divided into I local blocks to extract a spatial feature. Subtractions between a reference silhouette frame and silhouette frames are calculated to extract a temporal feature. We now describe the determination of the reference silhouette frame. The distances between all silhouette frames are calculated. The reference silhouette frame with the smallest distance is selected. In each local block, the amount of movement is calculated by summing the absolute

values of all subtractions to obtain the time-series signals. Then, a window function of length L is convoluted into the time-series signals of the amount of movement in each local block. The power spectral density (PSD) [68] is estimated from the time-series signals. PSD consists of the component of the power value corresponding to each frequency. The number of components of PSD in each local block is $L/2$. Finally, the PSDs of all local blocks are combined into a feature vector for gender classification. The dimension of the feature is $IL/2$. The vector of PSDs is denoted as a local movement (LM) feature.

4.4 Experiment

4.4.1 Dataset

We evaluated whether the gender of a standing person can be classified based on a video sequence of body sway recorded by an overhead camera. We acquired video sequences of the body sway of 60 participants (30 females and 30 males). Table 4.1 shows the details of the participants. The same instructions were given to all participants. We asked the participants to maintain an upright posture (Romberg's pose), shown in Fig. 4.6 (a), during the acquisition of their video sequence. We assumed a scenario where people wear the same work clothes in a factory. To reduce the changes in face orientation during the acquisition of a video sequence, we asked all participants to keep looking at a timer placed 3.0 m away. We set the height of the timer at 1.4 m. Fig. 4.6 (b) shows the experimental setting for the acquisition of video sequences of body sway. We randomly set the height of the overhead camera to between 2.0 and 4.0 m from the floor. The resolution and sampling rate of the overhead camera were 1920×1080 pixels and 30.0 Hz, respectively. We calibrated the overhead camera such that the optical axis

Table 4.1: Details of the participants in our dataset of video sequences containing body sway.

	Female	Male
Number of participants	30	30
Average age (years)	22.4 ± 6.3	21.6 ± 1.3
Average height (cm)	158.7 ± 4.7	170.2 ± 6.4

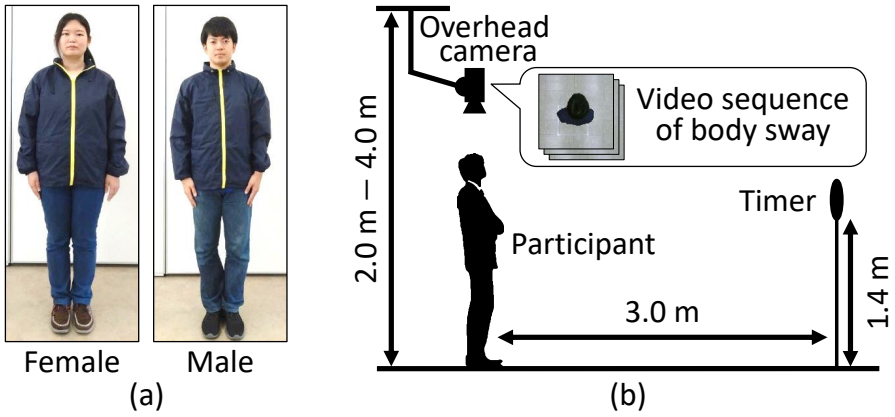
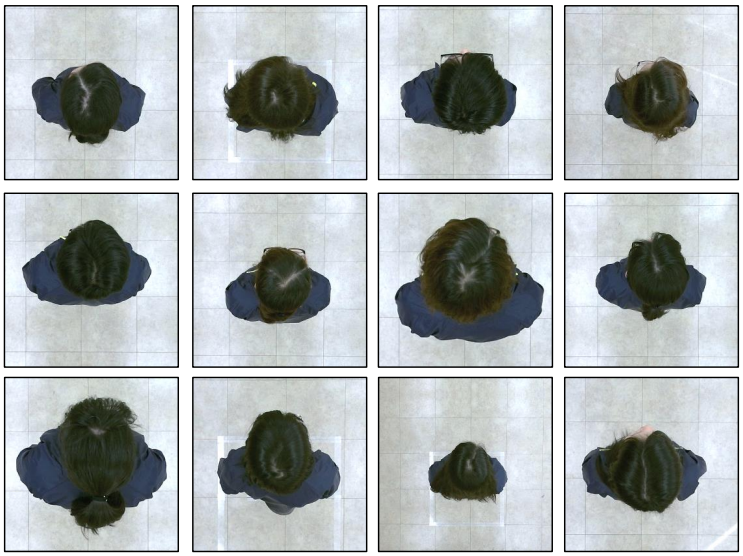


Figure 4.6: Experimental setting for observing participants using an overhead camera. (a) Examples of a female and a male standing with an upright posture. (b) Camera setting for acquiring a video sequence of body sway.

coincided with the direction normal to the floor. The internal parameters of the overhead camera were fixed. We set the time length of each video sequence to 60 s. Fig. 4.7 shows examples of color images of females and males in video sequences acquired in our experimental setting. The apparent size of the upper body in the color images varies because of differences in camera height. The inter-class variation between females and males is small even though the intra-class variation of the apparent sizes is large.



Females

Males

Figure 4.7: Examples of color images of females and males in video sequences acquired by an overhead camera. These images show variation in the apparent size of the upper body due to camera height. The inter-class variation between females and males is small even though the intra-class variation of the apparent size is large.

4.4.2 Evaluation of Gender Classification Accuracy

We compared the accuracy of our method with those of three other methods in our experiment. The details of the methods are as follows.

Proposed method (LM): We used the LM features described in Section 4.3 to represent the body sway of a standing person. To extract an LM feature, we set the number of local blocks to $I = 8$ and the length of a window function to $L = 64$ (2.1 s). We estimated a silhouette sequence from a video sequence using a conventional background subtraction technique. We set $H = 100$ pixels and $W = 100$ pixels. We applied a linear support vector machine (SVM) as the classifier and set its regularization parameter to $C = 1.0$.

Alternative method 1 (GEI): We used the GEI [54] features reported in previous studies on the gender classification of a walking person. To extract a GEI feature, we calculated a temporal average image of all frames in a 60-s silhouette sequence. We used the same silhouette sequences as those in our method. We applied a linear SVM as the classifier and set its regularization parameter to $C = 1.0$.

Alternative method 2 (CNN): We used a CNN [101] with single images as a representative of conventional classification techniques. The structure of the CNN consisted of four two-dimensional convolutional layers and four two-dimensional pooling layers. We used 45000 images of females and 45000 images of males as training samples for the CNN. The size of the sample images was set to 100×100 pixels. Each pixel had RGB color values. Binary cross-entropy with the stochastic gradient descent was used.

Alternative method 3 (C3D): We used C3D [103] with short video sequences as a representative of spatio-temporal feature extraction. The structure of C3D consisted of four three-dimensional convolutional layers and four three-dimensional pooling layers. We used 2800 short video sequences of fe-

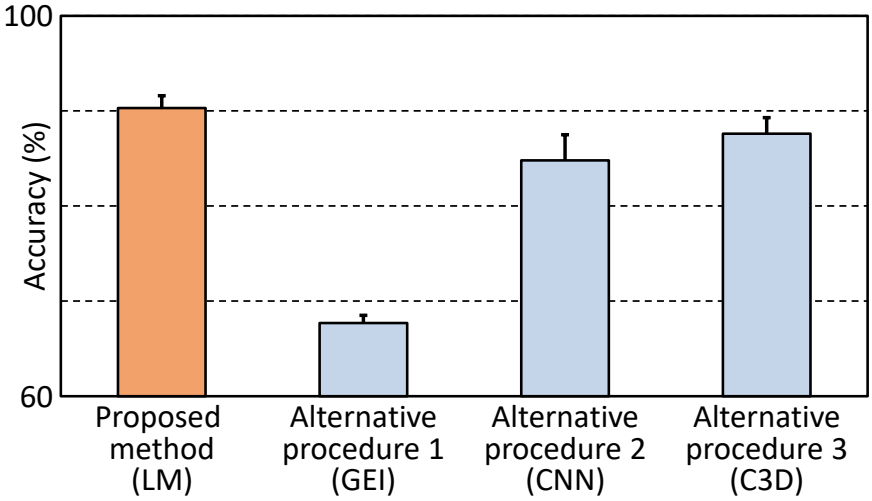


Figure 4.8: Comparison of gender classification accuracy obtained using proposed LM, GEI, CNN, and C3D features.

males and 2800 short video sequences of males as training samples. Each sample of a video sequence consisted of 16 frames. The size of each frame was set to 100×100 pixels. Each pixel of a frame had RGB color values. Binary cross-entropy with the stochastic gradient descent was used.

We randomly shuffled 60 participants and selected 50 participants as training samples and 10 participants as test samples. We completely separated the participants between the training samples and the test samples. We conducted the random shuffling 30 times and calculated the average and standard deviation of the gender classification accuracy for each method.

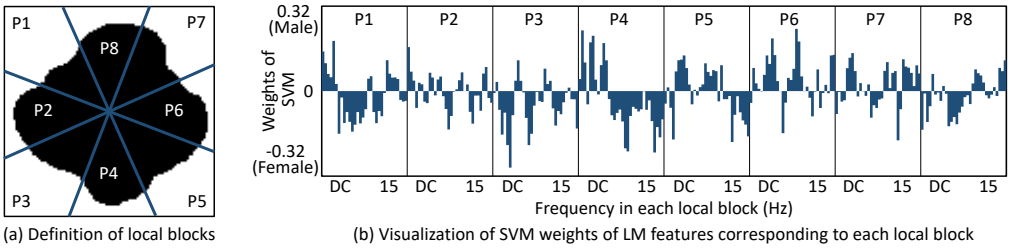
Fig. 4.8 shows the gender classification accuracy for each method. The accuracy of our method is much higher than that of GEI features. GEI features were designed to represent the large movements of the arms and legs during walking and thus cannot accurately represent the slight movements of a standing person. Conversely, the proposed LM features were designed to represent body sway. They thus have higher gender classification accu-

racy compared with that of GEI features. Furthermore, the accuracy of our method was superior to that of the CNN. The CNN extracted only spatial features from single images. It was not designed to extract temporal features from movement. We used C3D features to extract spatio-temporal characteristics. The proposed LM features outperform these features. C3D features were not designed to represent body sway (their target is large movements during large movements) and thus have relatively low gender classification accuracy. The proposed LM features include better spatio-temporal characteristics for representing body sway.

4.4.3 Visualization of SVM Weights Calculated from LM Features

We visualized the SVM weights calculated when training a gender classifier to determine the most informative component of the proposed LM features for gender classification. We used $I = 8$ local blocks to extract an LM feature (see Section 4.4.2). The local blocks are labeled P1 to P8 as shown in Fig. 4.9 (a). Local blocks P2, P4, P6, and P8 correspond to the left hand, back, right hand, and face, respectively. Fig. 4.9 (b) shows the SVM weights of the LM features corresponding to the local blocks. The horizontal axis represents the frequency in each local block. The extreme left and right on the horizontal axis of each local block represent DC and 15 Hz, respectively. The vertical axis represents the weight of each component. A component with a negative (positive) weight contributes to the classification of females (males). The number of components of an LM feature was $I \times L/2 = 8 \times 64/2 = 256$.

First, we identified the most informative local block for gender classification. We calculated the sum of the absolute SVM weights in each local block. The sum for local blocks P1 to P8 was 2.65, 1.72, 2.51, 3.44, 2.36, 2.36, 2.29, 2.11, and 1.81, respectively. A local block was more informative for gender



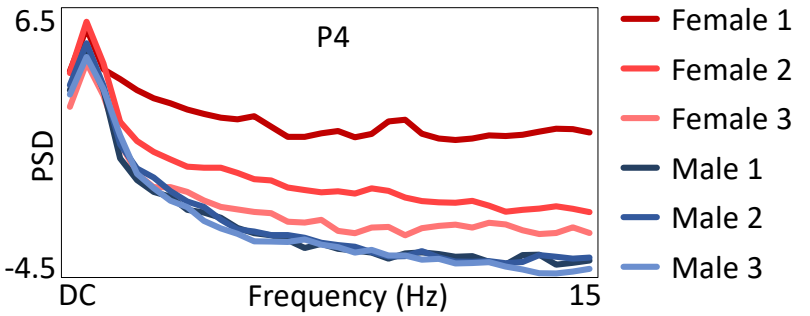
(a) Definition of local blocks

(b) Visualization of SVM weights of LM features corresponding to each local block

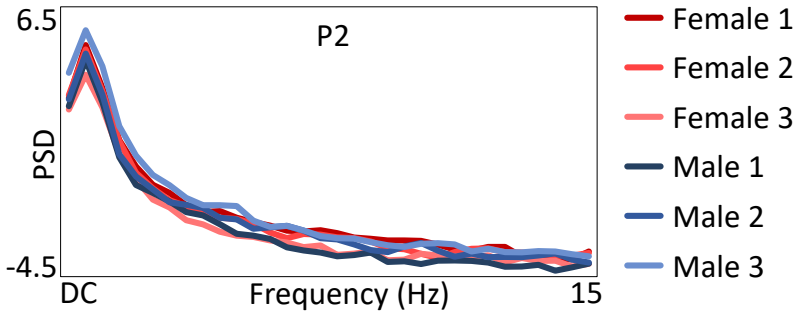
Figure 4.9: Visualization of SVM weights for determining the most informative component of proposed LM features for gender classification. (a) Definition of local blocks P1 to P8. (b) SVM weights of LM features corresponding to the local blocks.

classification when its sum was higher. Local block P4 (corresponding to a person's back) had the highest sum and was thus important for gender classification.

Next, we identified the most informative frequency band in local block P4 for gender classification. The high-frequency band is more informative than the low-frequency band for classifying the female class using the SVM weights corresponding to P4 in Fig. 4.9 (b). To determine the differences in LM features between females and males, some examples of the features corresponding to P4 are shown in Fig. 4.10 (a). Examples of those corresponding to P2 are shown in Fig. 4.10 (b). Local block P2 was not discriminative because it had the lowest sum of absolute SVM weights. We compared the features of P2 with those of P4. In P4, there are large differences in the high-frequency band (3.0 to 15.0 Hz), in which LM features of females are higher than those of males. Conversely, there are no remarkable differences between females and males in P2. We now discuss the reasons for the differences in P4. Local block P4 contained the back of the head, where females often have longer hair. We believe that the differences in the high-frequency band appear because the long hair moved during body sway.



(a) Component of LM feature corresponding to local block P4



(b) Component of LM feature corresponding to local block P2

Figure 4.10: Examples of LM features corresponding to local block P2 or P4 for three females and three males.

4.4.4 Gender Classification Accuracy Obtained using Medical Data

We evaluated the gender classification accuracy of parameters derived from medical data. As described in Section 4.2.3, analytical studies [90, 91, 92] have reported that the differences between females and males can be observed in the frequency characteristics and trajectories of time-series signals. Note that these studies used a force plate to acquire the parameters. In our experiment, we used an overhead camera instead of a force plate to acquire the time-series signals of the center positions of the upper body using silhouette sequences.

We used ten parameters, namely F1 to F6 for frequency characteristics and T1 to T4 for trajectories, reported in previous studies [90, 91, 92]. The parameters were set as follows:

F1 [90]: DC component of the power spectrum of center positions.

F2 [90]: Top accumulated frequency component of the power spectrum of center positions.

F3 [91]: DC component of the power spectrum of velocities.

F4 [92]: Sum of the vertical power spectrum at frequencies lower than 0.2 Hz.

F5 [92]: Sum of the horizontal power spectrum at frequencies lower than 0.2 Hz.

F6 [92]: Sum of the vertical power spectrum at frequencies higher than 2.0 Hz.

T1 [91]: Area of an ellipse approximated to the trajectory of center positions.

T2 [91]: Length of the major axis of the ellipse.

T3 [91]: Length of the minor axis of the ellipse.

T4 [91]: Length of the trajectory of center positions.

We tested each parameter as a one-dimensional feature vector for evaluating gender classification accuracy. We also combined the parameters into a ten-dimensional feature vector (**All**) to improve accuracy. The experimental conditions, except for the features, were the same as those for our method in Section 4.4.2.

Fig. 4.11 shows a comparison between the accuracy of each parameter derived from medical data and that of the proposed LM feature. The proposed LM feature had higher accuracy. The accuracy of the combined parameters was higher than that of individual parameters. The results show that the proposed LM feature has higher gender classification accuracy than that of parameters derived from medical data.

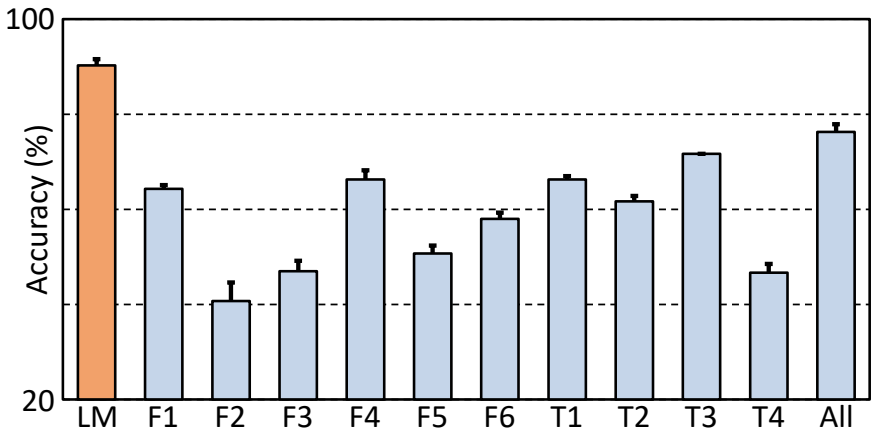


Figure 4.11: Accuracy of parameters F1-F6 and T1-T4 derived from medical data and proposed LM feature. 'All' represents a feature that combines all parameters.

4.5 Conclusions

We investigated whether the gender of a standing person can be classified by extracting a feature that represent body sway in a video sequence recorded by an overhead camera. Our method normalizes the apparent size of silhouette sequences of the upper body to remove variation. We divided the upper body into local blocks to represent spatial features and measured the time-series signals of body sway from each local block to represent temporal features. We acquired video sequences containing body sway for 60 participants to evaluate gender classification accuracy. The gender classification accuracy was $90.3 \pm 1.3\%$ for our spatio-temporal feature. We confirmed that body sway in a video sequence improves gender classification accuracy compared with that for parameters derived from medical data.

In future work, we intend to develop a method for extracting features that represent essential gender differences and are robust against posture changes. We will also investigate whether body sway can be used to classify attributes other than gender, such as age and clothing.

Chapter 5

Conclusions

5.1 Summary

In this thesis, the author developed techniques for identifying people and classifying their gender. To extract features for identification and classification from body sway, the author divided the body into multiple regions and observed local body sway movements in each region. the author then applied frequency analysis to local body sway movements. To evaluate the proposed methods, the author collected original datasets of video sequences of body sway using an overhead camera.

Chapter 2 presented a technique that identifies standing people by extracting features from local body sway movements. Chapter 3 presented a technique that improves the performance of identification in the case of self-occlusion by measuring body sway from the head regions of a person. Chapter 4 presented a technique that determines the gender of a person using body sway observed from an overhead camera by removing the variation in the apparent size of the body regions in a silhouette sequence.

5.2 Contributions

Temporal and spatial analysis of local body sway movements for the identification of people

The author designed features for identifying people accurately using body sway observed from an overhead camera. The author acquired a silhouette sequence that represents body regions from a video sequence using background subtraction. To extract features that are informative for person re-identification, the author measured local body sway movements in local regions. The author acquired the local regions by dividing the silhouette sequence of the body regions. To evaluate this technique, The author collected original datasets of video sequences of body sway using an overhead camera. The author confirmed that the identification performance of the proposed technique is superior to that of existing methods.

Identifying people using body sway in the presence of self-occlusion

The author developed a technique that identifies people in the presence of self-occlusion. To overcome the variation in appearance of a person by self-occlusion, the author measured body sway from head regions. The author acquired the silhouette sequence representing the head region from a video sequence by applying U-Net. To evaluate this technique, the author collected video sequences of body sway from 50 participants in the presence of self-occlusion. The author confirmed that the proposed technique using head regions can improve the identification performance of the existing method.

Gender classification using video sequences of body sway recorded by overhead camera

The author developed a technique that classifies the gender of a person using body sway observed from an overhead camera. To apply this technique in the real world, the author considered the variation in the height of the ceiling. The author removed the variation in the apparent size of body regions, and extracted features of gender classification from these body regions. To evaluate the proposed technique, the author collected video sequences of body sway featuring 30 females and 30 males using an overhead camera. The author showed that the proposed technique significantly improves the accuracy of gender classification, compared with existing methods in the medical area.

5.3 Future Directions

Extending the proposed methods of person re-identification and gender classification using body sway

The proposed methods of person re-identification and gender classification using body sway suffer from some limitations. In the future, the author intends to apply the proposed methods to the cases in which a person reads an advertisement displayed on a station platform, waits for an elevator in a building, or wonders whether to buy a product in a store. To achieve high performance in these use cases, the author needs to extend the proposed methods. In various use cases, it is difficult to make people maintain an upright posture for as long as several tens of seconds. People often wear different clothes from each other. Furthermore, many people may be standing around a target person.

To apply the proposed methods in the above use cases, the author will design a new feature that can identify people and classify their gender using a shorter video sequence of body sway. To ensure that the proposed methods are robust to variations in posture, the author needs to determine the period in which a person does not change their posture, and measure body sway during this period. To decrease the influence of variations in clothing, the author will consider extracting the specific frequency components of body sway in future work. When multiple people stand densely, the author will attempt to detect each person and measure the body sway of that person. For some tasks in person re-identification and gender classification, the author must analyze each task in detail and solve the problems in applying the proposed methods to the use cases.

Application to classification of attributes other than gender

In this thesis, the author investigated whether it is possible to classify the gender of a person using body sway. However, gender is not the only attribute of a person. People have other attributes, such as age, height, and weight. In addition, there are attributes of the state of a person, such as whether the person is drunk, carrying luggage, or tired. If it is possible to classify other attributes of a person using body sway, the proposed method can be applied to various other applications. In future work, the author would like to clarify whether it is possible to classify various attributes, other than the gender, of a person using body sway. Furthermore, instead of an attribute label, the author plans to predict a numerical value for each attribute, such as the age, height, or weight of a person, using body sway.

Bibliography

[1] X. Tan, S. Chen, Z.H. Zhou, and F. Zhang, "Face recognition from a single image per person: A survey," Pattern Recognition, vol.39, no.9, pp.1725–1745, 2006.

[2] R. Jafri and H.R. Arabnia, "A survey of face recognition techniques," Journal of Information Processing Systems, vol.5, no.2, pp.41–68, 2009.

[3] V. Kumar, A. Namboodiri, M. Paluri, and C. Jawahar, "Pose-aware person recognition," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.6223–6232, 2017.

[4] E. Gonzalez-Sosa, R. Vera-Rodriguez, J. Fierrez, and V.M. Patel, "Exploring body shape from mmw images for person recognition," IEEE Transactions on Information Forensics and Security, vol.12, no.9, pp.2078–2089, 2017.

[5] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," Proceedings of the IEEE International Conference on Computer Vision, pp.1116–1124, 2015.

[6] A. Bedagkar-Gala and S.K. Shah, "A survey of approaches and trends in person re-identification," Image and Vision Computing, vol.32, no.4, pp.270–286, 2014.

[7] C. Liu, S. Gong, C.C. Loy, and X. Lin, "Person re-identification: What features are important?," Proceedings of the European Conference on Computer Vision, pp.391–401, 2012.

[8] R. Layne, T.M. Hospedales, S. Gong, and Q. Mary, "Person re-identification by attributes.," Proceedings of the British Machine Vision Conference, vol.2, no.3, p.8, 2012.

[9] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.2285–2294, 2018.

[10] E. Ahmed, M. Jones, and T.K. Marks, "An improved deep learning architecture for person re-identification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3908–3916, 2015.

[11] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.1367–1376, 2017.

[12] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.34–42, 2015.

[13] G. Ozbulak, Y. Aytar, and H. Ekenel, "How transferable are cnn-based features for age and gender classification?," Proceedings of the International Conference of the Biometrics Special Interest Group, pp.1–6, 2016.

[14] E. Mäkinen and R. Raisamo, "An experimental comparison of gender classification methods," Pattern Recognition Letters, vol.29, no.10, pp.1544–1556, 2008.

[15] E. Makinen and R. Raisamo, "Evaluation of gender classification methods with automatically detected and aligned faces," IEEE Transactions

on Pattern Analysis and Machine Intelligence, vol.30, no.3, pp.541–547, 2008.

[16] Z. Yang, M. Li, and H. Ai, "An experimental study on automatic face gender classification," Proceedings of 18th International Conference on Pattern Recognition, vol.3, pp.1099–1102, 2006.

[17] S. Lapuschkin, A. Binder, K.R. Muller, and W. Samek, "Understanding and comparing deep neural networks for age and gender classification," Proceedings of the IEEE International Conference on Computer Vision Workshops, pp.1629–1638, 2017.

[18] S.S. Liew, M.K. Hani, S.A. Radzi, and R. Bakhteri, "Gender classification: a convolutional neural network approach," Turkish Journal of Electrical Engineering & Computer Sciences, vol.24, no.3, pp.1248–1264, 2016.

[19] S. Shigematsu, H. Morimura, Y. Tanabe, T. Adachi, and K. Machida, "A single-chip fingerprint sensor and identifier," IEEE Journal of Solid-State Circuits, vol.34, no.12, pp.1852–1859, 1999.

[20] M. Lourde and D. Khosla, "Fingerprint identification in biometric securitysystems," International Journal of Computer and Electrical Engineering, vol.2, no.5, pp.852–855, 2010.

[21] A. Jain, L. Hong, and S. Pankanti, "Biometric identification," Communications of the ACM, vol.43, no.2, pp.90–98, 2000.

[22] R. Kaur and S.G. Mazumdar, "Fingerprint based gender identification using frequency domain analysis," International Journal of Advances in Engineering & Technology, vol.3, no.1, pp.295–299, 2012.

[23] A. Falohun, O. Fenwa, and F. Ajala, "A fingerprint-based age and gender detector system using fingerprint pattern analysis," International Journal of Computer Applications, vol.136, no.4, pp.43–48, 2016.

[24] J. Yang, Y. Shi, and J. Yang, "Personal identification based on finger-vein features," Computers in Human Behavior, vol.27, no.5, pp.1565–1570, 2011.

[25] S.K. Im, H.M. Park, Y.W. Kim, S.C. Han, S.W. Kim, C.H. Kang, and C.K. Chung, "An biometric identification system by extracting hand vein patterns," Journal-Korean Physical Society, vol.38, no.3, pp.268–272, 2001.

[26] J.D. Wu and C.T. Liu, "Finger-vein pattern identification using svm and neural network technique," Expert Systems with Applications, vol.38, no.11, pp.14284–14289, 2011.

[27] R. Raghavendra and C. Busch, "A low cost wrist vein sensor for biometric authentication," Proceedings of IEEE International Conference on Imaging Systems and Techniques, pp.201–205, 2016.

[28] F. Marra, G. Poggi, C. Sansone, and L. Verdoliva, "A deep learning approach for iris sensor model identification," Pattern Recognition Letters, vol.113, pp.46–53, 2018.

[29] A. Agarwal, R. Keshari, M. Wadhwa, M. Vijh, C. Parmar, R. Singh, and M. Vatsa, "Iris sensor identification in multi-camera environment," Information Fusion, vol.45, pp.333–345, 2019.

[30] S. Lagree and K.W. Bowyer, "Predicting ethnicity and gender from iris texture," Proceedings of IEEE International Conference on Technologies for Homeland Security, pp.440–445, 2011.

[31] A. Kuehlkamp, B. Becker, and K. Bowyer, "Gender-from-iris or gender-from-mascara?," Proceedings of IEEE Winter Conference on Applications of Computer Vision, pp.1151–1159, 2017.

[32] J.E. Tapia, C.A. Perez, and K.W. Bowyer, "Gender classification from the same iris code used for recognition," IEEE Transactions on Information Forensics and Security, vol.11, no.8, pp.1760–1770, 2016.

[33] I.B. Barbosa, M. Cristani, A. Del Bue, L. Bazzani, and V. Murino, "Re-identification with rgb-d sensors," Proceedings of European Conference on Computer Vision, pp.433–442, 2012.

[34] A. Mogelmose, C. Bahnsen, T. Moeslund, A. Clapés, and S. Escalera, "Tri-modal person re-identification with rgb, depth and thermal features," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp.301–307, 2013.

[35] K. Koide, E. Menegatti, M. Carraro, M. Munaro, and J. Miura, "People tracking and re-identification by face recognition for rgb-d camera networks," Proceedings of European Conference on Mobile Robots, pp.1–7, 2017.

[36] F.M. Castro, M.J. Marín-Jiménez, and N. Guil, "Multimodal features fusion for gait, gender and shoes recognition," Machine Vision and Applications, vol.27, no.8, pp.1213–1228, 2016.

[37] A.K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," IEEE Transactions on Circuits and Systems for Video Technology, vol.14, no.1, pp.4–20, 2004.

[38] X. Wang, "Intelligent multi-camera video surveillance: A review," Pattern Recognition Letters, vol.34, no.1, pp.3 – 19, 2013.

[39] J. Ashbourn, "Biometrics: Advanced identity verification," 2000.

[40] A. Dantcheva, C. Velardo, A. D'Angelo, and J.L. Dugelay, "Bag of soft biometrics for person identification," Multimedia Tools and Applications, vol.51, no.2, pp.739–777, 2011.

[41] M.S. Nixon, P.L. Correia, K. Nasrollahi, T.B. Moeslund, A. Hadid, and M. Tistarelli, "On soft biometrics," Pattern Recognition Letters, vol.68, pp.218 – 230, 2015.

[42] P. Tome, J. Fierrez, R. Vera-Rodriguez, and M.S. Nixon, "Soft biometrics and their application in person recognition at a distance," IEEE Transactions on Information Forensics and Security, vol.9, no.3, pp.464–475, 2014.

[43] A.K. Jain, S.C. Dass, and K. Nandakumar, "Soft biometric traits for personal recognition systems," Proceedings of International Conference on Biometric Authentication, pp.731–738, 2004.

[44] A.K. Jain, S.C. Dass, and K. Nandakumar, "Can soft biometric traits assist user recognition?," Biometric Technology for Human Identification, vol.5404, pp.561–572, 2004.

[45] A.K. Jain, K. Nandakumar, X. Lu, and U. Park, "Integrating faces, fingerprints, and soft biometric traits for user recognition," Proceedings of International Workshop on Biometric Authentication, pp.259–269, 2004.

[46] D.S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino, "Custom pictorial structures for re-identification.," Proceedings of the British Machine Vision Conference, vol.1, no.2, pp.68.1–68.11, 2011.

[47] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Person re-identification using spatial covariance regions of human body parts," Proceedings of 7th IEEE International Conference on Advanced Video and Signal Based Surveillance, pp.435–440, 2010.

[48] W.S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, and S. Gong, "Partial person re-identification," Proceedings of the IEEE International Conference on Computer Vision, pp.4678–4686, 2015.

[49] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.384–393, 2017.

[50] A. Li, L. Liu, K. Wang, S. Liu, and S. Yan, "Clothing attributes assisted person reidentification," IEEE Transactions on Circuits and Systems for Video Technology, pp.134–146, 2015.

[51] C.H. Kuo, S. Khamis, and V. Shet, "Person re-identification using semantic color names and rankboost," Proceedings of IEEE Workshop on Applications of Computer Vision, pp.281–287, 2013.

[52] J.E. Lee, A.K. Jain, and R. Jin, "Scars, marks and tattoos (smt): Soft biometric for suspect and victim identification," Proceedings of Biometrics Symposium, pp.1–8, 2008.

[53] Y. Chen, S. Duffner, A. Stoian, J.Y. Dufour, and A. Baskurt, "Deep and low-level feature based attribute learning for person re-identification," Image and Vision Computing, vol.79, pp.25–34, 2018.

[54] 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, Feb. 2006.

[55] Y. Makihara, R. Sagawa, Y. Mukaigawa, T. Echigo, and Y. Yagi, "Gait recognition using a view transformation model in the frequency domain," pp.151–163, 2006.

[56] C. Shan, S. Gong, and P. McOwan, "Fusing gait and face cues for human gender recognition," Neurocomputing, vol.71, no.10-12, pp.1931–1938, 2008.

[57] R. Martín-Félez, R. Mollineda, and J. Sánchez, "Gender classification from pose-based geis," Proceedings of the International Conference on Computer Vision and Graphics, pp.501–508, 2012.

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

[59] N. Nies and P. Sinnott, "Variations in balance and body sway in middle-aged adults. subjects with healthy backs compared with subjects with low-back dysfunction," Spine, vol.16, no.3, pp.325–330, 1991.

[60] Y. Yu, H.C. Chung, L. Hemingway, and T.A. Stoffregen, "Standing body sway in women with and without morning sickness in pregnancy," Gait & Posture, vol.37, no.1, pp.103–107, 2013.

[61] P. Bergin, A. Bronstein, N. Murray, S. Sancovic, and D. Zeppenfeld, "Body sway and vibration perception thresholds in normal aging and in patients with polyneuropathy," Neurol Neurosurg Psychiatry, vol.58, no.3, pp.335–340, 1995.

[62] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23, no.3, pp.257–267, 2001.

[63] P. Kasprowski and J. Ober, "Eye movements in biometrics," Proceedings of International Workshop on Biometric Authentication, pp.248–258, 2004.

[64] N. Cuong, V. Dinh, and L. Ho, "Mel-frequency cepstral coefficients for eye movement identification," Proceedings of 24th International Conference on Tools with Artificial Intelligence, vol.1, pp.253–260, 2012.

[65] K.Q. Weinberger and L.K. Saul, "Distance metric learning for large margin nearest neighbor classification," Journal of Machine Learning Research, vol.10, no.Feb, pp.207–244, 2009.

[66] H.T. Lin, C.J. Lin, and R.C. Weng, "A note on plattfs probabilistic outputs for support vector machines," Machine Learning, vol.68, no.3, pp.267–276, 2007.

[67] M. Nishiyama, T. Miyauchi, H. Yoshimura, and Y. Iwai, "Synthesizing realistic image-based avatars by body sway analysis," Proceedings of the Fourth International Conference on Human Agent Interaction, pp.155–162, 2016.

[68] P. Welch, "The use of fast fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," IEEE Transactions on Audio and Electroacoustics, vol.15, no.2, pp.70–73, 1967.

[69] R. Min, J. Choi, G. Medioni, and J.L. Dugelay, "Real-time 3d face identification from a depth camera," Proceedings of the 21st International Conference on Pattern Recognition, pp.1739–1742, 2012.

[70] T. Kamitani, H. Yoshimura, M. Nishiyama, and Y. Iwai, "Temporal and spatial analysis of local body sway movements for the identification

of people," IEICE Transactions on Information and Systems, vol.102, no.1, pp.165–174, 2019.

[71] P. Vera, S. Monjaraz, and J. Salas, "Counting pedestrians with a zenithal arrangement of depth cameras," Machine Vision and Applications, vol.27, no.2, pp.303–315, 2016.

[72] S. Munir, R.S. Arora, C. Hesling, J. Li, J. Francis, C. Shelton, C. Martin, A. Rowe, and M. Berges, "Real-time fine grained occupancy estimation using depth sensors on arm embedded platforms," Proceedings of 2017 IEEE Real-Time and Embedded Technology and Applications Symposium, pp.295–306, 2017.

[73] B.A.Y. Agusta, P. Mittrapiyanuruk, and P. Kaewtrakulpong, "Field seeding algorithm for people counting using kinect depth image," Indian Journal of Science and Technology, vol.9, p.48, 2016.

[74] D. Liciotti, M. Paolanti, R. Pietrini, E. Frontoni, and P. Zingaretti, "Convolutional networks for semantic heads segmentation using top-view depth data in crowded environment," Proceedings of 24th International Conference on Pattern Recognition, pp.1384–1389, 2018.

[75] S. Mukherjee, B. Saha, I. Jamal, R. Leclerc, and N. Ray, "Anovel framework for automatic passenger counting," Proceedings of 18th IEEE International Conference on Image Processing, pp.2969–2972, 2011.

[76] J. Tighe, M. Niethammer, and S. Lazebnik, "Scene parsing with object instances and occlusion ordering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.

[77] T. Brox, L. Bourdev, S. Maji, and J. Malik, "Object segmentation by alignment of poselet activations to image contours," Proceedings of

the IEEE Conference on Computer Vision and Pattern Recognition, pp.2225–2232, 2011.

[78] P.F. Felzenszwalb and D.P. Huttenlocher, "Efficient graph-based image segmentation," International Journal of Computer Vision, vol.59, no.2, pp.167–181, 2004.

[79] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.23, pp.122–1239, 2001.

[80] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," International Journal of Computer Vision, vol.81, no.1, pp.2–23, 2009.

[81] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," Advances in Neural Information Processing Systems, pp.109–117, 2011.

[82] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.3431–3440, 2015.

[83] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, pp.234–241, 2015.

[84] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.12, pp.2481–2495, 2017.

[85] C. Hua, Y. Makihara, and Y. Yagi, "Pedestrian detection by using a spatio-temporal histogram of oriented gradients," IEICE Transactions on Information and Systems, vol.96, no.6, pp.1376–1386, 2013.

[86] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.39, no.4, pp.773–787, 2016.

[87] X. Wang, S. Zheng, R. Yang, B. Luo, and J. Tang, "Pedestrian attribute recognition: A survey," CoRR, 2019.

[88] C.B. Ng, Y.H. Tay, and B. Goi, "Recognizing human gender in computer vision: a survey," Proceedings of the Pacific Rim International Conference on Artificial Intelligence, pp.335–346, 2012.

[89] S. Khan, M. Nazir, S. Akram, and N. Riaz, "Gender classification using image processing techniques: A survey," Proceedings of the IEEE 14th International Multitopic Conference, pp.25–30, 2011.

[90] J. Kim, G. Eom, C. Kim, D. Kim, J. Lee, B. Park, and J. Hong, "Sex differences in the postural sway characteristics of young and elderly subjects during quiet natural standing," Geriatrics & Gerontology International, vol.10, no.2, pp.191–198, 2010.

[91] M. Plandowska, M. Lichota, and K. Górniak, "Postural stability of 5-year-old girls and boys with different body heights," PLoS ONE, vol.14, no.12, pp.1–10, 2020.

[92] T. Kitabayashi, S. Demura, M. Noda, and T. Yamada, "Gender differences in body-sway factors of center of foot pressure in a static upright posture and under the influence of alcohol intake," Journal of Physiological Anthropology and Applied Human Science, vol.23, no.4, pp.111–118, 2004.

[93] F. Wang, M. Skubic, C. Abbott, and J. Keller, "Body sway measurement for fall risk assessment using inexpensive webcams," Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology, pp.2225–2229, 2010.

[94] M. Nishiyama, T. Miyauchi, H. Yoshimura, and Y. Iwai, "Synthesizing realistic image-based avatars by body sway analysis," Proceedings of the Fourth International Conference on Human Agent Interaction, pp.155–162, 2016.

[95] L. Yeung, K. Cheng, C. Fong, W. Lee, and K. Tong, "Evaluation of the microsoft kinect as a clinical assessment tool of body sway," Gait & Posture, vol.40, no.4, pp.532–538, 2014.

[96] Z. Lv, V. Penades, S. Blasco, J. Chirivella, and P. Gagliardo, "Evaluation of kinect2 based balance measurement," Neurocomputing, vol.208, pp.290–298, 2016.

[97] L. Cao, M. Dikmen, Y. Fu, and T. Huang, "Gender recognition from body," Proceedings of the 16th ACM International Conference on Multimedia, pp.725–728, 2008.

[98] J. Tapia and C. Perez, "Gender classification based on fusion of different spatial scale features selected by mutual information from histogram of lbp, intensity, and shape," IEEE Transactions on Information Forensics and Security, vol.8, no.3, pp.488–499, 2013.

[99] M. Collins, J. Zhang, P. Miller, and H. Wang, "Full body image feature representations for gender profiling," Proceedings of the 12th International Conference on Computer Vision Workshops, pp.1235–1242, 2009.

[100] M. Yildirim, O. Ince, Y. Salman, J. Song, J. Park, and B. Yoon, "Gender recognition using hog with maximized inter-class difference," Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, pp.106–109, 2016.

[101] G. Antipov, S. Berrani, N. Ruchaud, and J. Dugelay, "Learned vs. hand-crafted features for pedestrian gender recognition," Proceedings of the 23rd ACM International Conference on Multimedia, pp.1263–1266, 2015.

[102] M. Nishiyama, R. Matsumoto, H. Yoshimura, and Y. Iwai, "Extracting discriminative features using task-oriented gaze maps measured from observers for personal attribute classification," Pattern Recognition Letters, vol.112, pp.241–248, 2018.

[103] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," Proceedings of the IEEE International Conference on Computer Vision, pp.4489–4497, 2015.

[104] H. Xu, A. Das, and K. Saenko, "R-c3d: Region convolutional 3d network for temporal activity detection," Proceedings of the IEEE International Conference on Computer Vision, pp.5783–5792, 2017.

[105] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-c3d: Temporal convolutional 3d network for real-time action recognition," Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp.7138–7145, 2018.

Publications

Journals

1. 神谷 卓也, 山口 優太, 中谷真太朗, 西山 正志, 岩井 儀雄. 頭上カメラから観測された身体動揺を用いた性別認識の精度評価. 映像メディア学会誌, (条件付き採録).

2. 神谷 卓也, 山口 優太, 西山 正志, 岩井 儀雄. 天井カメラを用いて観測された身体動揺における自己遮蔽の影響を考慮した人物対応付け. 電気学会論文誌 C, Vol.140, No.6, pp.629–637, June, 2020.

3. Takuya Kamitani, Hiroki Yoshimura, Masashi Nishiyama, and Yoshio Iwai. Temporal and spatial analysis of local body sway movements for the identification of people. IEICE Transactions on Information and Systems, Vol. 102, No.1, pp.165–174, January, 2019.

Refereed conferences

4. Takuya Kamitani, Yuta Yamaguchi, Shintaro Nakatani, Masashi Nishiyama, Yoshio Iwai. Gender Classification Using Video Sequences of Body Sway Recorded by Overhead Camera. Proceedings of 25th International Conference on Pattern Recognition, pp.9196–9202, January, 2021.

5. Takuya Kamitani, Yuta Yamaguchi, Masashi Nishiyama, Yoshio Iwai. Identifying People Using Body Sway in Case of Self-Occlusion. Proceedings of International Workshop on Frontiers of Computer Vision. Vol.1212, pp.136–149, April, 2020.

6. Takuya Kamitani, Hiroki Yoshimura, Masashi Nishiyama, and Yoshio Iwai. Identifying People using Temporal and Spatial Changes in Local Movements Measured from Body Sway. Proceedings of 4th Asian Conference on Pattern Recognition, pp.828–833, 2017.

Others

7. Takuya Kamitani, Syoji Fujimoto, Hiroki Yoshimura, Masashi Nishiyama, Yoshio Iwai. Anomaly detection using local regions in road images acquired from a hand-held camera, Proceedings of 7th Global Conference on Consumer Electronics, pp. 347–350, October, 2018.

8. 神谷 卓也, 西山 正志, 岩井 儀雄. カメラの身体動揺を用いた人物対応付けにおける服装変動に頑健な特徴量の設計, ビジョン技術の実利用ワークショップ (ViEW), pp. 240–245, December, 2020.

9. 神谷 卓也, 山口 優太, 西山 正志, 岩井 儀雄. カメラの高さと向きの変動に頑健な身体動揺を用いた人物対応付けの検討, 画像の認識・理解シンポジウム (MIRU), PS2-73, August 2019.

10. 神谷 卓也, 安形 俊輝, 吉村 宏紀, 西山 正志, 岩井 儀雄. 身体動揺を用いた人物対応付けにおける立ち位置変動の影響調査. バイオメトリクス研究会 (BioX), pp. 5–10, July. 2018.

11. 神谷 卓也, 吉村 宏紀, 西山 正志, 岩井 儀雄. 身体動揺を用いた人物対応付けにおける姿勢変化の調査〜足開閉が認識性能に与える影響〜. バイオメトリクスと認識・認証シンポジウム (SBRA), pp. 54–55, November 2017.

12. 神谷 卓也, 吉村 宏紀, 西山 正志, 岩井 儀雄. 身体動揺から計測した局所振動量を用いた人物対応付け. 画像の認識・理解シンポジウム (MIRU),

PS3–57, August 2017.

13. 神谷 卓也, 吉村 宏紀, 西山 正志, 岩井 儀雄. カメラ映像を用いた身体動揺の計測による人物対応付けの検討. バイオメトリクスと認識・認証シンポジウム (SBRA), pp. 94–95, November 2016.

14. Yuta Yamaguchi, Takuya Kamitani, Masashi Nishiyama, Yoshio Iwai, Daisuke Kushida. Extracting features of body sway for baggage weight classification, Proceedings of 9th Global Conference on Consumer Electronics, pp.443–446, October, 2020.

15. 中山晴貴, 山口優太, 神谷卓也, 西山正志, 岩井儀雄. 身体動揺を用いた人物対応付けにおける履物の影響調査　履物の高さが認識精度に与える影響 , ビジョン技術の実利用ワークショップ (ViEW), pp. 236–239, December, 2020.

16. 山口優太, 神谷卓也, 西山正志, 岩井儀雄. 身体動揺を用いた重量物所持の認識可能性の検証. ビジョン技術の実利用ワークショップ (ViEW), IS2-A1, December, 2019.

17. 宮城 裕也, 山口優太, 神谷卓也, 西山正志, 岩井儀雄. カメラ画像列を用いた性別認識における身体動揺の有効性の検討, 電気・情報関連学会中国支部連合大会, October, 2019.

18. 山口優太, 神谷卓也, 吉村宏紀, 西山正志, 岩井儀雄. 身体動揺を用いた人物対応付けにおける待ち姿勢の影響調査, 情報科学技術フォーラム (FIT), pp.93–94, September, 2018.

Awards

1. 平成 29 年度 MIRU 学生奨励賞. 第 20 回 画像の認識・理解シンポジウム (MIRU). 身体動揺から計測した局所振動量を用いた人物対応付け.