Identifying People Using Body Sway in Case of Self-Occlusion

Takuya Kamitani, Yuta Yamaguchi, Masashi Nishiyama, and Yoshio Iwai

Tottori University, Tottori 680-8550, Japan

Abstract. We propose a method of identifying people in case of self-occlusion by using body sway measured at the head using a top-view camera. To accurately represent the identities of people as reflected in body sway, it is important to acquire accurate appearances in images. However, such images frequently contain defects, especially self-occlusion, that degrade the performance of a prevalent method of identification because they use whole-body regions to identify people. To solve the problem of self-occlusion in this context, our method computes silhouette images of regions at the head by applying a segmentation technique. To reflect people's identities using body sway, we spatially divide the head region into local blocks and temporally measure movements in them. The results of experiments show that the proposed method can improve the performance of the prevalent method of identification from 17.3% to 57.9%.

Keywords: Body sway · Self-occlusion · Identification.

1 Introduction

The widespread use of surveillance cameras is expected to help further develop biometric authentication systems [21, 14]. To identify people accurately from images captured through such cameras, behavioral characteristics [7, 13, 9] have been considered in research on biometrics as they can be used to identify people based on their movements. Features of the gait [7, 13] 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 [9] 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 [20, 16, 1] or to track people's walking routes [11, 15]. However, regions around the head have not been used to aim to identify people. Another such method [9] 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 for the features of identification of individuals.

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.

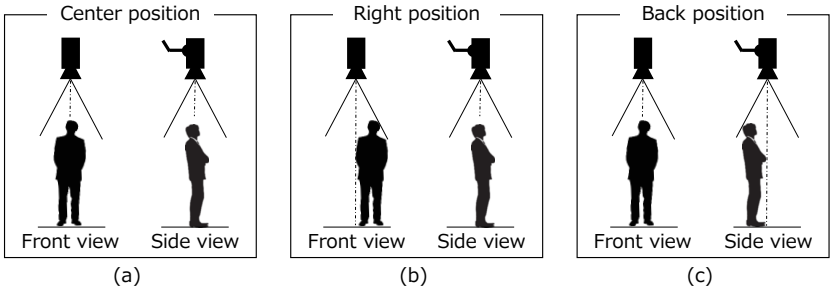

Fig. 1. The standing positions of a person used to investigate the influence of self-occlusion.

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 1. We defined the point where the optical axis of the camera was orthogonal to the floor as the center. Figure 1 (a) shows the condition when the person standing at the center was observed, and Figures 1 (b) and (c) show conditions of observation of people standing to the right and behind the center, respectively.

Figure 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 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 shows examples of head regions acquired under the same observation conditions as in Figure 2. The green pixels in the images represent the head regions. A comparison of Figures 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. Regions of the left and right shoulders were symmetrically at the center as shown in Figure 3 (a). However, in Figure 3 (b), part of regions of the right shoulder acquired at the center are hidden by the head regions. And part of regions of the left shoulder what a camera did not observe at center position appear. The same tendency can be observed in Figure 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.

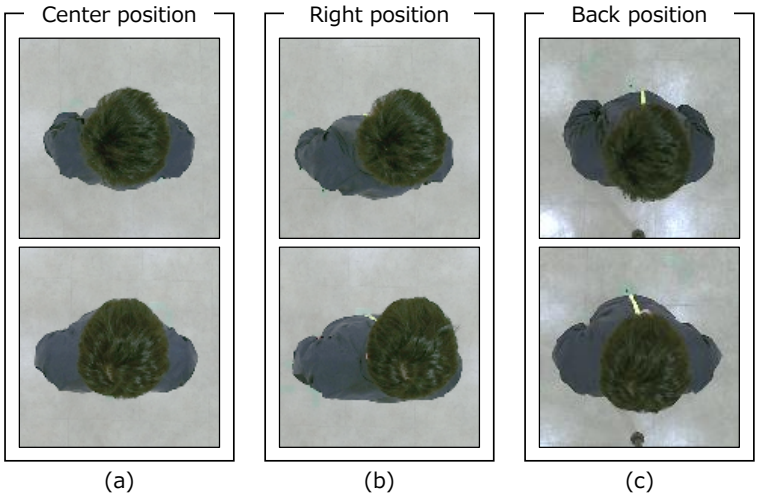

Fig. 2. Examples of the appearance of entire body acquired from two people standing in three different positions.

3 Our method

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 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.

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 [19, 4, 5, 3, 18, 10]. Segmentation methods that use deep learning [12, 17, 2] 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 5 (a) shows examples of the annotation labels of the head regions, and Figure 5 (b) shows examples of the images of people that were used. The trained network model output candidate head regions.

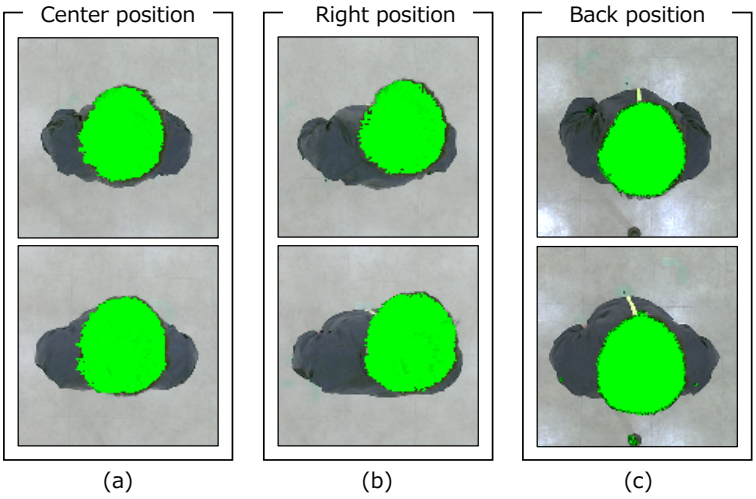

Fig. 3. Examples of head regions acquired from two people in three different standing positions. The green pixels represent the head regions.

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 6 (a) shows examples of candidate head regions containing noise. To 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 6 (b) shows examples of silhouette images of the head regions after noise had been reduced.

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 [9]. 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) [23] of each movement.

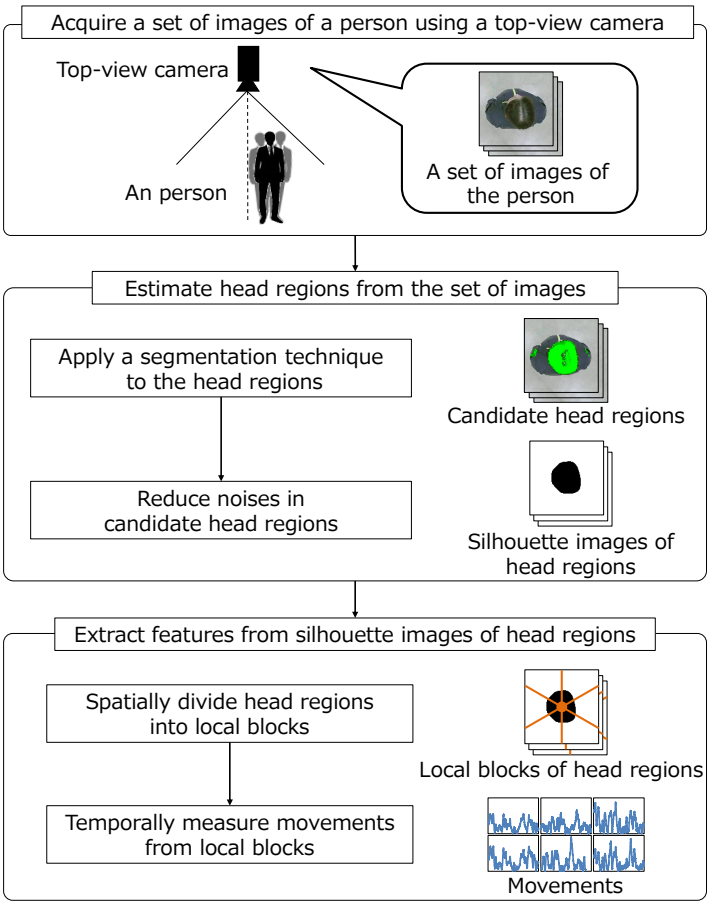

Fig. 4. Overview of our method.

4 Experiments

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 ± 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 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.

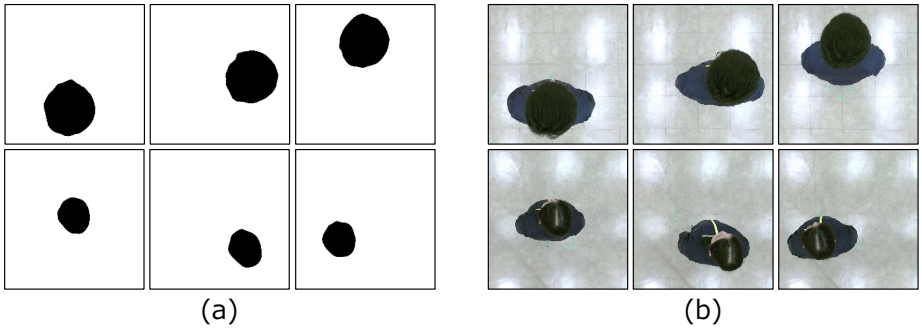

Fig. 5. Examples of annotation labels of head regions and images of people used to train a network model for head segmentation.

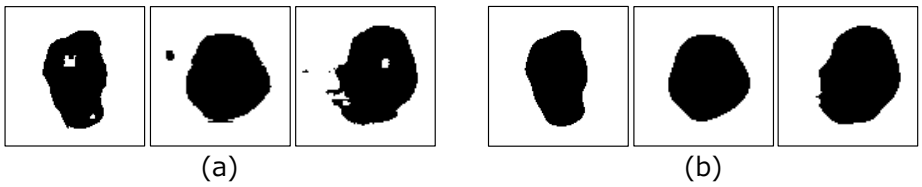

Fig. 6. Examples of candidate head regions containing noise, and silhouette images of these regions having reduced reducing noise.

Figure 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 7 (c). Figure 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. We cropped the 1920×1080-pixels images of all participants to 1080×1080 pixels, and resized them to 256×256 pixels.

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 [17], which was used in research [11], to estimate

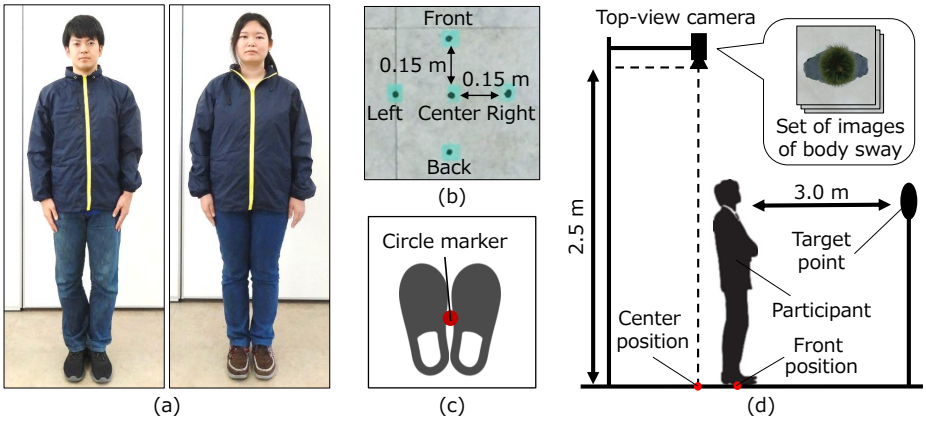

Fig. 7. The conditions under which each participant was observed. (a) shows their poses and clothes, and (b) shows their standing positions set on the floor. (c) shows the circle marker to align the position of the feet of the participants, and (d) shows the setup used to acquire a set of images of the body sway.

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 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 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 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 8 (a) and (b) were estimated with high accuracy in all positions. The appearances of the head regions in Figure 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.

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 9 (a) shows examples of them.

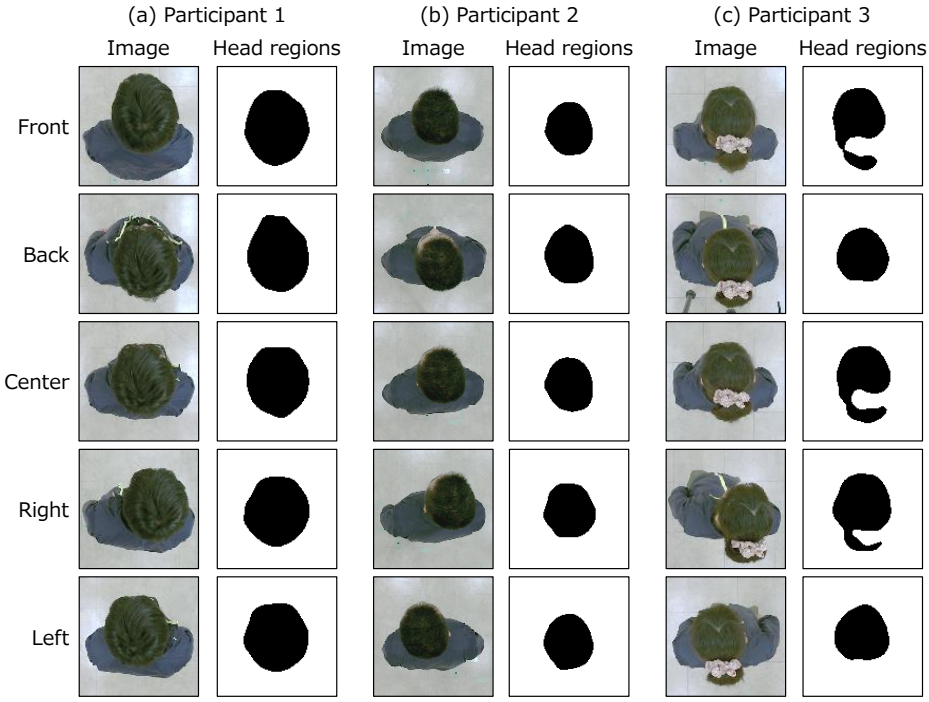

Fig. 8. Examples of head regions estimated from images of three participants in five standing positions using U-net.

Entire body: We used entire-body regions as used in a prevalent method [9]. Figure 9 (b) shows examples of them.

Shoulder: We used shoulder regions excluding the head regions from entire-body regions. Figure 9 (c) shows examples of them.

We estimated the whole-body regions and shoulder regions by applying the method described in Section 3.2. We extracted features to identify people from silhouette images of each body part by applying the method described in Section 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 performance. The proposed applied a metric learning technique, the large-margin nearest-neighbor (LMNN) method [22]. We randomly selected 25 participants not from the target and the query 4.1, and used them for LMNN. Data for the remaining 25 participants were used for identification. We repeated the random selection five times.

Table 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 per-

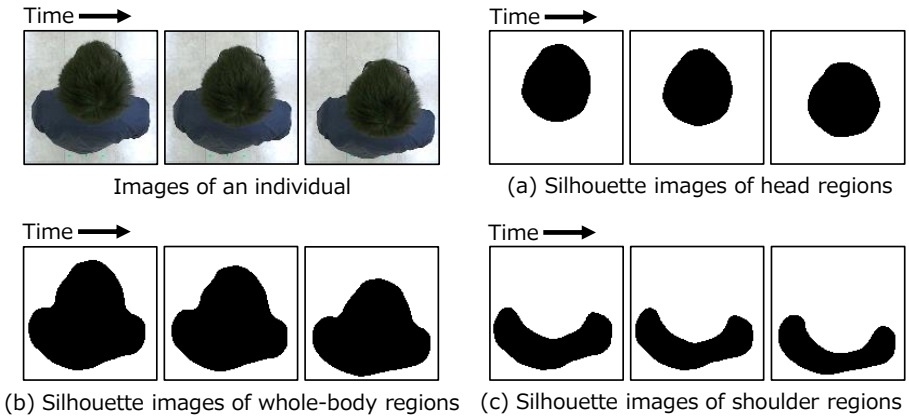

Fig. 9. Examples of the silhouette images of each body part.

Table 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

formance 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.

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: 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.

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

Table 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

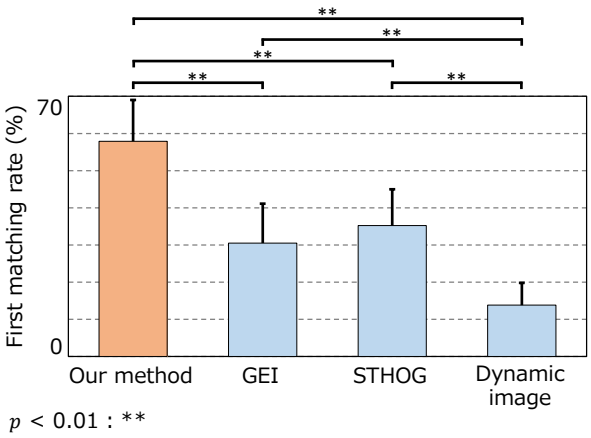

Fig. 10. Comparison of the identification performance of our method, which uses spatio-temporal features, with prevalent methods.

images. We used the temporal change of the central positions as the feature for identification.

The experimental conditions except for the features used were the same as described in Section 4.3.

Table 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.

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 [7] and STHOG [8] methods are widely used to authenticate gait, and were chosen along with the dynamic image method [6], which

is used in action recognition, for comparison with the proposed method. To extract the GEI, we computed the average image of the silhouette images for 60 seconds. To extract the STHOG, we set 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 4.3.

Figure 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. We see that there was a significant difference between our method and GEI. The same tendencies are observed for STHOG, and Dynamic image.

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 movement 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.

Acknowledgments This work was partially supported by JSPS KAKENHI under grant number JP17K00238 and MIC SCOPE under grant number 172308003.

References

1. Agusta, B.A.Y., Mittrapiyanuruk, P., Kaewtrakulpong, P.: Field seeding algorithm for people counting using kinect depth image. Indian Journal of Science and Technology **9**, 48 (2016)
2. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(12), 2481–2495 (2017)
3. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**, 122–1239 (2001)
4. Brox, T., Bourdev, L., Maji, S., Malik, J.: 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)
5. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient graph-based image segmentation. International Journal of Computer Vision **59**(2), 167–181 (2004)

6. Fernando, B., Gavves, E., Oramas, J., Ghodrati, A., Tuytelaars, T.: Rank pooling for action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **39**(4), 773–787 (2016)
7. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316–322 (2006)
8. Hua, C., Makihara, Y., Yagi, Y.: Pedestrian detection by using a spatio-temporal histogram of oriented gradients. IEICE Transactions on Information and Systems **96**(6), 1376–1386 (2013)
9. Kamitani, T., Yoshimura, H., Nishiyama, M., Iwai, Y.: Temporal and spatial analysis of local body sway movements for the identification of people. IEICE Transactions on Information and Systems **102**(1), 165–174 (2019)
10. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. Advances in Neural Information Processing Systems pp. 109–117 (2011)
11. Liciotti, D., Paolanti, M., Pietrini, R., Frontoni, E., Zingaretti, P.: 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)
12. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 3431–3440 (2015)
13. Makihara, Y., Sagawa, R., Mukaigawa, Y., Echigo, T., Yagi, Y.: Gait recognition using a view transformation model in the frequency domain. Proceedings of 9th European Conference on Computer Vision pp. 151–163 (2006)
14. Min, R., Choi, J., Medioni, G., Dugelay, J.L.: Real-time 3d face identification from a depth camera. Proceedings of the 21st International Conference on Pattern Recognition pp. 1739–1742 (2012)
15. Mukherjee, S., Saha, B., Jamal, I., Leclerc, R., Ray, N.: Anovel framework for automatic passenger counting. Proceedings of 18th IEEE International Conference on Image Processing pp. 2969–2972 (2011)
16. Munir, S., Arora, R.S., Hesling, C., Li, J., Francis, J., Shelton, C., Martin, C., Rowe, A., Berges, M.: 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)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention pp. 234–241 (2015)
18. Shotton, J., Winn, J., Rother, C., Criminisi, A.: Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. International Journal of Computer Vision **81**(1), 2–23 (2009)
19. Tighe, J., Niethammer, M., Lazebnik, S.: Scene parsing with object instances and occlusion ordering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
20. Vera, P., Monjaraz, S., Salas, J.: Counting pedestrians with a zenithal arrangement of depth cameras. Machine Vision and Applications **27**(2), 303–315 (2016)
21. Wang, X.: Intelligent multi-camera video surveillance: A review. Pattern Recognition Letters **34**(1), 3 – 19 (2013)
22. Weinberger, K.Q., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. Journal of Machine Learning Research **10**(Feb), 207–244 (2009)
23. Welch, P.: 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 **15**(2), 70–73 (1967)