Extracting Temporal Features Robust to Headwear Variation from Video Sequences of Body Sway for Person Identification

Takuya Kamitani, Haruki Nakayama, and
Masashi Nishiyama[0000−0002−5964−3209]

Tottori University, Tottori 680-0945, Japan
{takuya.kamitani,nishiyama}@tottori-u.ac.jp

Abstract. Person identification can be performed using temporal features extracted from video sequences of body sway captured by an overhead camera. We propose a method for extracting such temporal features in a manner that is robust to headwear variation. When people are wearing headwear, such as a cap or helmet, their head shapes observed by the camera change significantly according to the type of headwear. Existing methods for person identification cannot achieve high accuracy in the presence of headwear variation because the features used by existing methods are strongly affected by the changes in head shapes. To extract temporal features that are not influenced by headwear variation, we measure the time-series signals representing body sway by estimating the center positions from head shapes. Moreover, we propose a learning-based low-pass filter to remove the components that are uninformative from the frequency components of the time-series signals, while retaining the informative components. Experimental results show that our temporal features significantly enhance the accuracy of person identification in the presence of headwear variation, compared with the use of existing features.

Keywords: Person identification · Headwear variation · Body sway · Temporal feature extraction · Learning-based low-pass filter.

1 Introduction

Person identification is a technique for determining whether the same people appear in video sequences captured by cameras installed at different locations [1, 11, 12]. Person identification is often used to monitor people entering and exiting specific areas, such as factories and event venues. To identify people accurately, it is important to select cues that represent personal identities. In recent years, body sway has attracted attention as a suitable cue for person identification. Body sway refers to the subtle swaying movement of the entire body of a person while standing. By using body sway, it is possible to accurately identify people who are waiting for traffic lights to change or elevators to arrive.

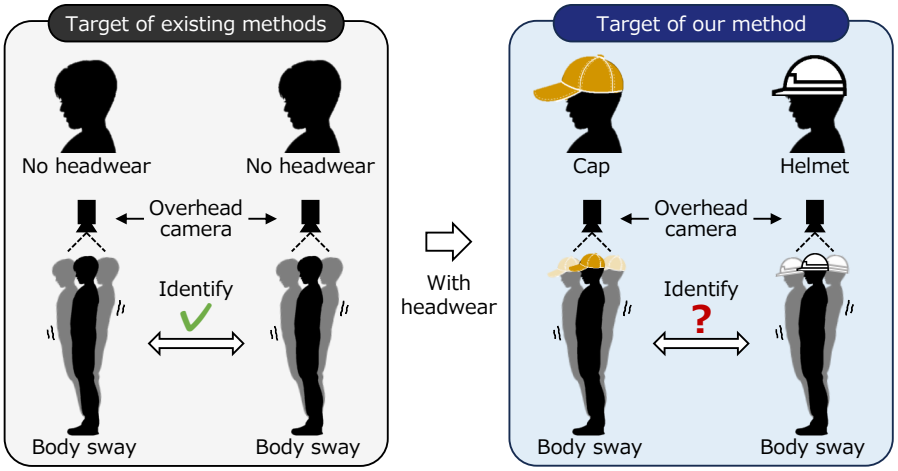

Fig. 1. Target of our method. Our method is intended to identify people with headwear in video sequences of body sway.

Several existing methods [6, 5] have been proposed to identify people by their body sway from video sequences captured by an overhead camera. These existing methods determine whether two video sequences show the same person by extracting spatiotemporal features of head shape and head movement that uniquely identify each person. Existing methods have the constraint that people must not be wearing any headwear, as shown in Fig. 1. However, people sometimes do wear headwear that is suitable and appropriate for the situation. For example, factory workers may wear caps or helmets while working on production lines, and participants in costume events may wear long-hair wigs or Afro wigs to dress up.

When people are wearing headwear, their head shapes observed by an overhead camera change according to the type of headwear worn. Figure 2 shows examples of the head shapes of the same person without headwear and with a cap, a helmet, a long-hair wig, and an Afro wig. The figure compares the head shapes with and without headwear. The head shape with a cap is similar to an ellipse because of the hat brim. In addition, the head shape with a helmet is similar to a circle, and the shape with a long-hair wig or an Afro wig is similar to that of the wig itself. In summary, the head shape varies according to the type of headwear worn.

Problems arise if existing methods [6, 5] are used directly to identify people with various types of headwear. The spatiotemporal features used by these existing methods capture personal identity by the spatial representation of shape and the temporal representation of movement. The spatial shape representation focuses on the head shape itself, whereas the temporal movement representation focuses on the slight movement of the head shape. Because both the spatial

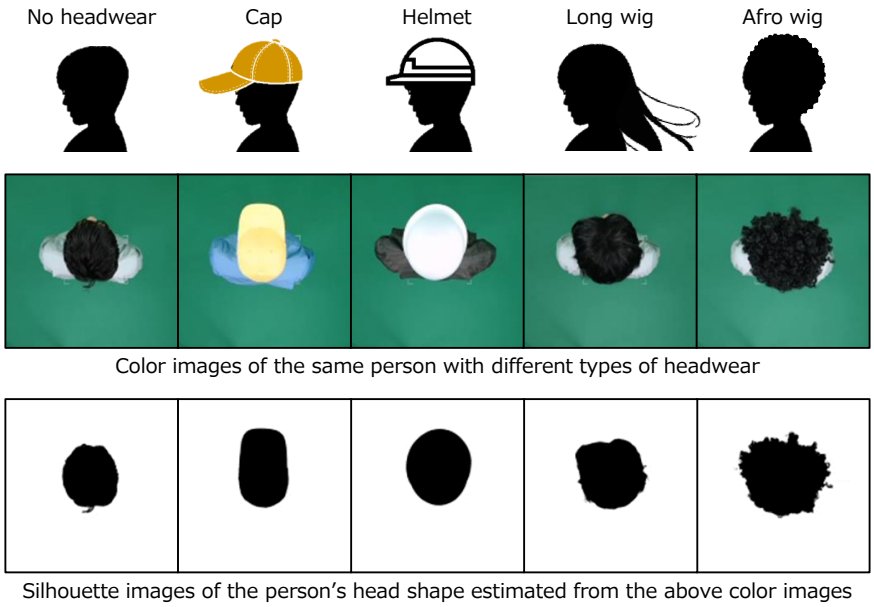

Fig. 2. Examples of variation in headwear types. Color images were acquired by an overhead camera, showing the same person wearing different types of headwear. Silhouette images, representing the head shape of the person with each type of headwear, were estimated from the color images.

shape representation and the temporal movement representation rely on the head shape, the spatiotemporal features vary when the head shape varies. Therefore, person identification becomes less accurate if the existing spatiotemporal features are used to identify people wearing various types of headwear.

In this paper, we propose a method for extracting temporal features that are robust to headwear variation, in person identification using body sway, by avoiding representations that rely on head shape. We measure the time-series signals of the center positions estimated by spatially averaging the pixel coordinates of each head shape from a video sequence captured by an overhead camera. To enhance the accuracy of person identification, we then extract temporal features with only the low-frequency components of the time-series signal, using a newly designed learning-based low-pass filter.
Our contributions are summarized as follows.

- We clarify the existence of informative components and uninformative components contained in temporal features (Section 2.4).
- We present a learning-based low-pass filter that removes the uninformative components and retains the informative components to extract temporal features that are robust to headwear variation (Section 3.1).

– Using an original dataset of video sequences of people with no headwear and with four different types of headwear, we demonstrate significantly enhanced accuracy of person identification in the presence of headwear variation, compared with existing methods (Section 3.2 and Section 3.4).

The experimental results show that the person identification accuracy of our temporal features is higher than that of existing spatiotemporal features in the presence of headwear variation.

2 Influence of Headwear Variation on Temporal Features Extracted from a Video Sequence of Body Sway

2.1 Design of Temporal Features

As noted in Section 1, the spatiotemporal features of existing methods [6, 5] are strongly affected by head shape variation because these features have representations that rely on the head shape. Using these representations, it is hard to capture personal identity in the presence of headwear variation. Therefore, we consider extracting temporal features that focus only on the head movement, ignoring the head shape. To obtain such features, we improve the procedure for temporal feature extraction used by the existing methods. Specifically, we estimate the center positions of the head shape by averaging, and measure the time-series signals of head movement.

Figure 3 shows the procedure for temporal feature extraction that improves on that of existing methods [6, 5]. First, in step P1, we acquire a color video sequence using an overhead camera and estimate a silhouette sequence of the head shape from it. We apply the deep image matting technique [10] to perform this silhouette estimation. Second, in step P2, we estimate the center positions of the head shapes in the silhouette sequence. Specifically, in each silhouette frame, we calculate the center position as a two-dimensional vector by spatially averaging the pixel coordinates of the head shape. We then set the origin to the average of the center positions. Third, in step P3, we measure two time-series signals by treating each component of the two-dimensional vectors separately in the left–right and front–back directions. Fourth, in step P4, we estimate power spectral densities [9] from each time-series signal using the Fourier transform. Finally, we extract temporal features by combining these power spectral densities into one. In this paper, we refer to this feature extraction procedure as the improved existing method. Despite our improvements, we found that this method achieved only small improvements in person identification accuracy in the presence of headwear variation. In the following sections, we experimentally investigate why the accuracy improvements were small when the improved existing method was used.

2.2 Evaluation Dataset

We created an original dataset to evaluate person identification accuracy in the presence of headwear variation. We acquired color video sequences of the body

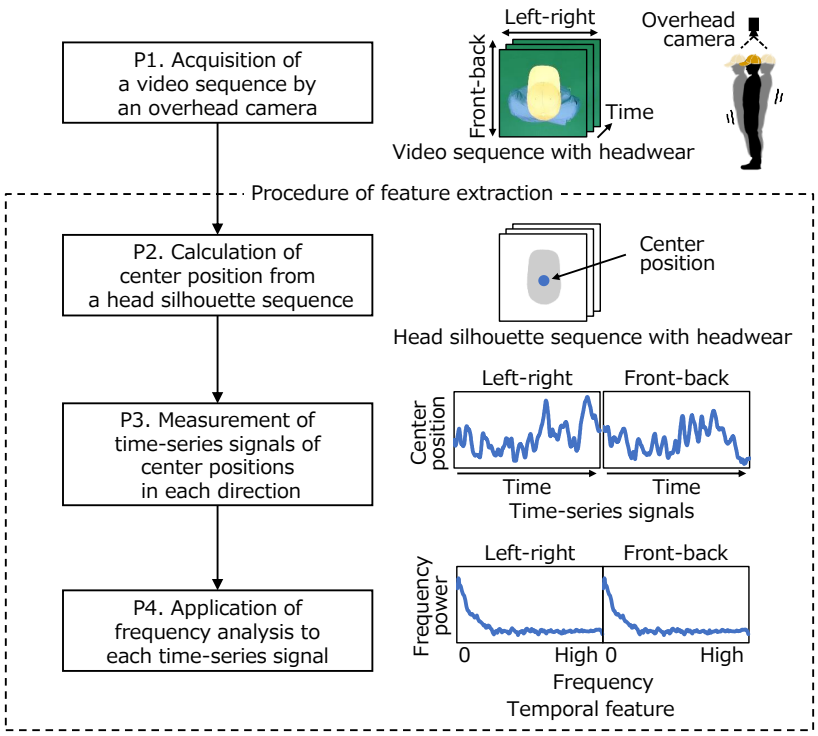

Fig. 3. Procedure used by the improved existing method of Section 2.1 for extracting temporal features from a video sequence of body sway by estimating center positions from head shapes.

sway of participants with no headwear and with the following four different types of headwear.

No headwear: Participants had no headwear, representing a situation in which people are commuters. There were no restrictions on the hairstyle of each participant. Figures 4(a) and 5(a) show examples of participants with no headwear observed from above and from the side, respectively.

Cap: Participants wore yellow caps with brims, representing a situation in which people work on a production line of small consumer goods. We instructed all participants to wear these caps so that the brims were in the same direction as their faces. Figures 4(b) and 5(b) show examples of participants wearing caps observed from above and from the side.

Helmet: Participants wore white helmets with short brims, representing a situation in which people are employed in a large industrial factory. We instructed all participants to wear these helmets so that the short brims were in the same

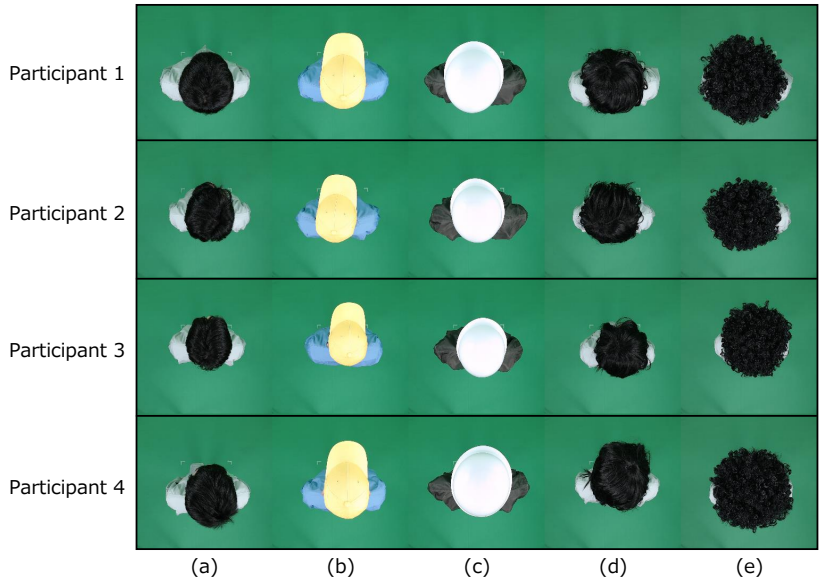

Fig. 4. Examples of color images of participants when wearing different types of headwear. The participants were observed from an overhead camera. The images in each row show the same person and those in each column depict the same type of headwear. Columns (a), (b), (c), (d), and (e) show images with no headwear, a cap, a helmet, a long wig, and an Afro wig, respectively.

direction as their faces. Figures 4(c) and 5(c) show examples of participants wearing helmets observed from above and from the side.

Long wig: Participants wore long wigs, representing a situation in which people appear on a theatre stage. We instructed all participants to wear these wigs so that the long-hair parts of the wigs remained behind their heads. Figures 4(d) and 5(d) show examples of participants wearing long wigs observed from above and from the side.

Afro wig: Participants wore Afro wigs, representing a situation in which people attend a costume event. We instructed all participants to wear these wigs in an arbitrary direction. Figures 4(e) and 5(e) show examples of participants wearing Afro wigs observed from above and from the side.

We observed 50 participants with mean age of 22.0 (standard deviation 1.9) years. Figure 6 shows the environment in which the body sway of the participants was observed. Participants were instructed to maintain an upright posture and focus on the designated marker during observation. An overhead camera was set at a height of 2.4 m above the floor. A mark was placed on the floor under the overhead camera and participants were instructed to stand on it. The camera resolution was set to 1920×1080 pixels and the sampling rate to 30 Hz. Each participant was observed twice for each type of headwear, and each par-

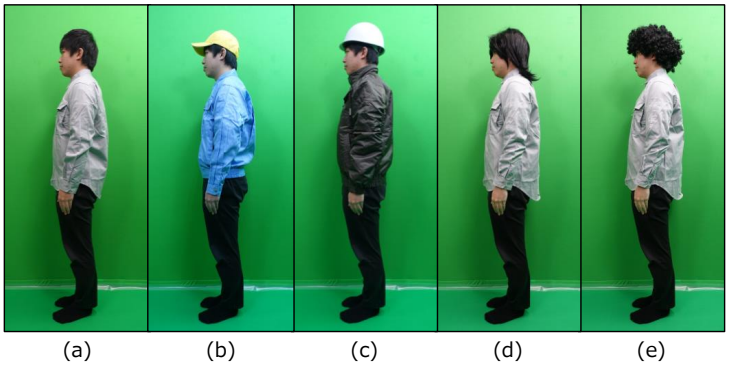

Fig. 5. Examples of color images of a participant wearing different types of headwear. The participant was observed by a side-view camera.

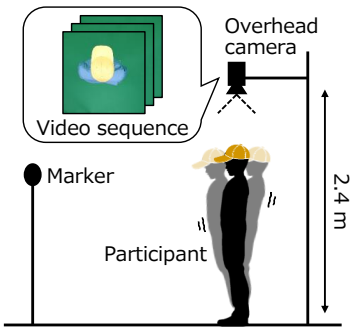

Fig. 6. Experimental setting for capturing a video sequence using an overhead camera.

ticipant's headwear was replaced in a random order. Hence, the total number of observations per participant was ten. The length of each observation was 120 s.

2.3 Accuracy of Person Identification Using Temporal Features of Improved Existing Method

We evaluated the accuracy of person identification in the presence of headwear variation. The purpose of this experiment was to confirm that the accuracy enhancement was slight even if we used the temporal features of the improved existing method, as noted in Section 2.1. The person identification task performed in this experiment was a search for the person appearing in a query video sequence among preregistered galleries of video sequences. We selected one of the five types of headwear for the query and a different type for the gallery in each trial of person identification. Hence, there were ${}_5\mathrm{P}_2 = 20$ permutations of headwear types for person identification. We applied the nearest neighbor algorithm

Table 1. Comparison of the accuracy (%) of person identification using the existing methods [6, 5] and the improved existing method (Section 2.1).

Method	n=1	n=5	n=10	n=15	nAUC
Existing methods [6, 5]	3.7	14.2	26.4	36.2	56.2
Improved existing method (Section 2.1)	5.2	18.8	31.3	43.3	59.7

with Euclidean distance for identifying people by using temporal features of the improved existing method. The number of participants for person identification was 50. We used the n-th matching rate and the normalized area under the curve (nAUC) of the cumulative match characteristics to evaluate the accuracy of person identification. The n-th matching rate is the probability that the person appearing in a query also appears in one of the top n galleries when the galleries are ranked in order of similarity (to the query). A larger value of the n-th matching rate indicates better accuracy. The nAUC is the area under the curve plotting the n-th rank (on the horizontal axis) against the n-th matching rate (on the vertical axis). A larger value of nAUC indicates better accuracy.

Table 1 shows the n-th matching rates and nAUC of the improved existing method and of the existing methods. From the experimental results, we confirmed that the improved existing method is only a slight enhancement of the existing methods. In the next section, we investigate the reason for this and explore the possibility of a more significant enhancement of the person identification accuracy.

2.4 Influence of Headwear Variation on Temporal Features

In this section, we investigate why the improved existing method could not enhance the accuracy of person identification. We assume the reason to be that its temporal features contain components that are uninformative for person identification in the presence of headwear variation. When these uninformative components are present, even if the query and gallery show the same person, the distance between the temporal features of the query and those of the gallery may be large. Furthermore, if the query and gallery show different people, the distance between the temporal features of the query and those of the gallery may be small. Consequently, we believe that these uninformative components may lead to incorrect identification.

To investigate the existence of uninformative components contained in the temporal features of the improved existing method, we calculated the within-person and between-person distances from the temporal features. The within-person distance was calculated when the query and gallery showed the same person. Conversely, the between-person distance was calculated when the query and gallery showed different people. To obtain these distances, we subtracted the frequency powers of the temporal features of the query from those of the temporal features of the gallery, squared the differences, and summed them for

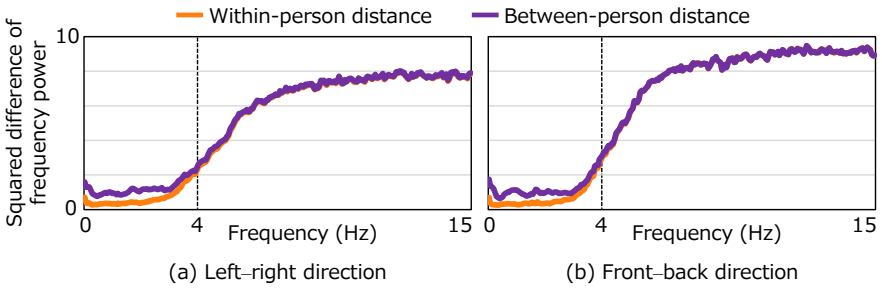

Fig. 7. Within-person and between-person distances in temporal features of the improved existing method. The orange and purple curves plot the within-person and between-person distances, respectively. (a) Left–right direction. (b) Front–back direction.

all frequency powers. We believe that the frequency bands in which the within-person and between-person distances are similar contain uninformative components that cause errors in person identification.

Figure 7 shows the within-person and between-person distances in temporal features of the improved existing method. The figure shows that the within-person and between-person distances are similar for frequencies greater than 4.0 Hz (the high-frequency band) in both the left–right (a) and front–back (b) directions. Conversely, the within-person and between-person distances are different for frequencies less than 4.0 Hz (the low-frequency band). These results suggest that the temporal feature components in the high-frequency band are uninformative and should be removed to enhance the person identification accuracy in the presence of headwear variation. Conversely, the temporal feature components in the low-frequency band are informative and should be retained. In the next section, we describe how to extract temporal feature components only in the low-frequency band and remove those in the high-frequency band.

3 Temporal Features of the Proposed Method

3.1 Design of a Learning-based Low-Pass Filter

To enhance the accuracy of person identification in the presence of headwear variation, we need to extract temporal feature components in the low-frequency band, as explained in Section 2.4. To achieve this, we propose a learning-based low-pass filter that determines the range of the low-frequency band containing only components that are informative for person identification. To design this low-pass filter, we need to select a suitable threshold. The threshold is a parameter that determines the range of the low-frequency band. We consider selecting this threshold automatically, according to the distances between temporal features. The ideal threshold should decrease the within-person distance and

increase the between-person distance. The proposed method uses a separation metric [7] to automatically find such a threshold. The separation metric should have a high value if the within-person distances are small and the between-person distances are large. We use this method to select thresholds separately in the left-right and front-back directions of body sway.

To obtain a threshold that is robust to headwear variation, we prepare a training dataset of temporal features of people wearing multiple types of headwear. The separation metric chosen is the ratio of the between-person distance to the within-person distance. We calculate the values of the separation metric by searching frequencies of temporal features at fixed intervals in the training dataset from zero to the maximum obtainable frequency. We then select the threshold that maximizes the value of the separation metric.

The proposed method is intended to perform person identification in a manner that is robust to headwear variation by extracting temporal features with a learning-based low-pass filter. Figure 8 shows the procedure used by the proposed method. The steps from P1 to P4 are the same as in Fig. 3. Step P5 of the proposed method uses a learning-based low-pass filter to extract the frequency powers of the low-frequency band only.

3.2 Accuracy of Person Identification Using Temporal Features of the Proposed Method

We investigated whether temporal features of the proposed method with a learning-based low-pass filter could enhance the accuracy of person identification in the presence of headwear variation. We created two sets (training set and evaluation set) of video sequences from the 50 participants described in Section 2.2. The training set contained ten randomly selected participants and the evaluation set contained the remaining 40 participants. The training set always included the five types of headwear described in Section 2.2. For each training set, we searched for the optimal threshold of the low-pass filter in 0.05-Hz increments. We repeated each trial, in which ten participants were randomly selected as the training set, five times and calculated the accuracy of person identification for each trial. We set the threshold of the low-pass filter to a constant value for each evaluation set. The other experimental conditions were the same as those described in Section 2.3. We again used the n-th matching rate and nAUC to evaluate the accuracy of person identification.

Table 2 shows the n-th matching rate and nAUC of person identification using temporal features of the proposed method and the improved existing method. The proposed method and the improved existing method were described in Sections 3.1 and 2.1, respectively. The results show that the temporal features of the proposed method are superior to those of the improved existing method with respect to person identification accuracy. Therefore, we believe that the proposed method can enhance the accuracy of person identification in the presence of headwear variation.

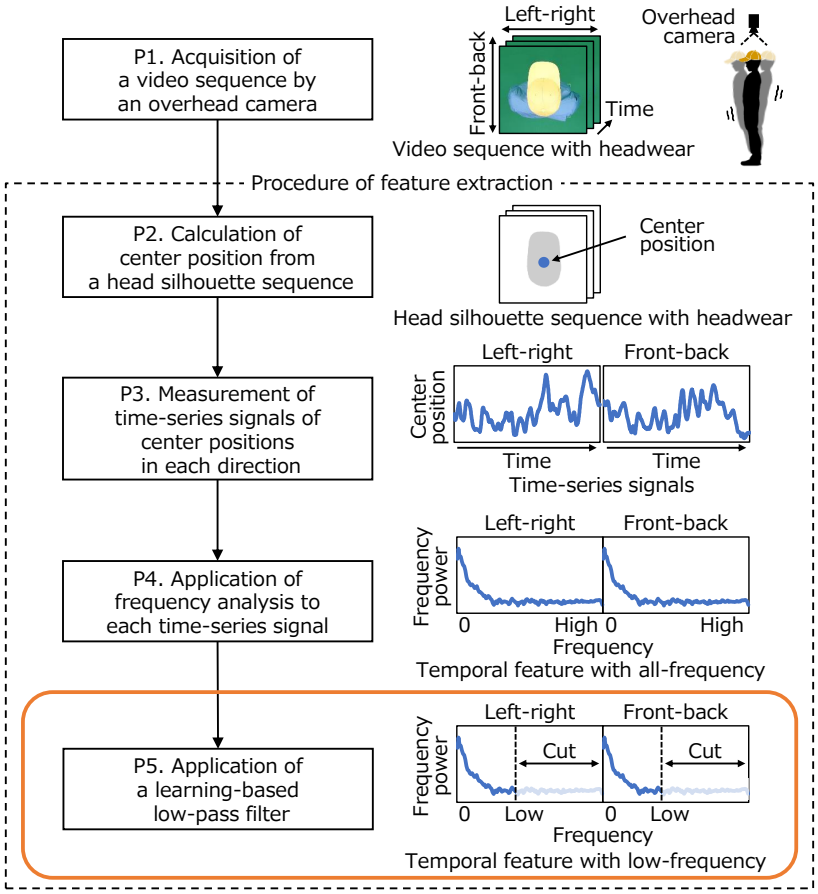

Fig. 8. Procedure used by the proposed method with a learning-based low-pass filter for extracting temporal features in the low-frequency band from a video sequence of body sway.

3.3 Comparison with the Use of Temporal Features in the High-Frequency Band

In the previous section, we demonstrated that the proposed method could enhance the accuracy of person identification by extracting temporal features from only the low-frequency band, using a learning-based low-pass filter. In this section, we discuss the accuracy achieved when temporal features are extracted from the high-frequency band. We compared temporal features extracted from the low-frequency band with those extracted from the high-frequency band and those obtained without filtering. The other experimental conditions were the same as those described in Section 3.2.

Table 2. Comparison of the accuracy (%) of person identification using our proposed method (Section 3.1) and the improved existing method (Section 2.1).

Method	n=1	n=5	n=10	n=15	nAUC
Proposed method (Section 3.1)	**44.2**	**78.6**	**89.2**	**94.2**	**92.1**
Improved existing method (Section 2.1)	6.3	21.9	38.1	51.0	60.1

Table 3. Comparison of the accuracy (%) of person identification using temporal features extracted from the low-frequency band, the high-frequency band, and without filtering.

Frequency band	n=1	n=5	n=10	n=15	nAUC
Low-frequency band (proposed)	**44.2**	**78.6**	**89.2**	**94.2**	**92.1**
High-frequency band	4.5	17.1	31.5	44.6	55.6
Without filtering (low+high frequencies)	6.3	21.9	38.1	51.0	60.1

Table 3 shows the n-th matching rate and nAUC of person identification using temporal features extracted from the low-frequency band, the high-frequency band, and without filtering. The results indicate that the use of temporal features from only the low-frequency band enhances the accuracy of person identification compared with the use of features from only the high-frequency band or without filtering. Therefore, we believe that using only a low-frequency band is beneficial for temporal feature extraction for identifying people accurately from body sway in the presence of headwear variation.

3.4 Comparison with Features of Existing Methods Based on Human Movement

The proposed method extracts temporal features from body sway, which is a type of human movement. Existing methods have been proposed that extract features from human movement other than body sway, for example, gait recognition [4] and action recognition [3, 8, 2]. Because the features extracted by these existing methods are based on human movement, they may achieve high accuracy in person identification when head shape varies because of headwear variation. To test this hypothesis, we evaluated the person identification accuracy achieved by the use of features extracted with GEI [4], DI [3], C3D [8], and TimeSformer [2]. It should be noted that we used convolutional neural networks to reduce the dimensionality of GEI and DI features and extracted only factors that are significant for enhancing the accuracy of person identification. We applied the nearest neighbor algorithm when identifying people using each feature, in the same manner as in Section 3.2. The experimental conditions other than the feature extraction methods were the same as those described in Section 3.2. The details of the feature extraction of existing methods are as follows.

Gait energy image (GEI) [4]: We extracted features by using GEI and ResNet101. We created 15 GEIs by randomly cutting out 15 short video se-

Table 4. Comparison of the accuracy (%) of person identification using the proposed method, the existing method using gait recognition [4], and the existing methods using action recognition [3, 8, 2].

Method	n=1	n=5	n=10	n=15	nAUC
Proposed	**44.2**	**78.6**	**89.2**	**94.2**	**92.1**
GEI [4]	8.0	30.2	53.3	69.9	72.4
DI [3]	15.8	45.9	67.5	81.4	80.4
C3D [8]	16.2	47.8	68.1	79.7	79.8
TimeSformer [2]	10.1	33.4	52.1	64.7	70.1

quences of length 8.5 s from the silhouette sequence described in Section 2.1 and averaging the short video sequence as a single image. To enhance the person identification accuracy, we used a pretrained ResNet101 model with ImageNet-1k and fine-tuned it with 1500 GEIs created from the training set described in Section 3.2. Finally, we obtained feature maps by inputting the GEIs into ResNet101.

Dynamic image (DI) [3]: We extracted features by using DI and ResNet101. We created 15 DIs by randomly cutting out 15 short video sequences from a color video sequence, in the same manner as described above for GEI, and applying RankSVM to each short video sequence. To enhance the person identification accuracy, we used a pretrained ResNet101 model with ImageNet-1k and fine-tuned it with 1500 DIs from the training set. Finally, we obtained feature maps by inputting the DIs into ResNet101.

C3D [8]: We extracted features by using C3D. We acquired 15 short video sequences in the same manner as described above for DI. We selected 16 frames from each short video sequence at equal intervals to reduce memory usage. We then trained the C3D model with 1500 short video sequences created from the training set. Finally, we obtained feature maps by inputting the short video sequences into C3D.

TimeSformer [2]: We extracted features by using TimeSformer. We acquired 15 short video sequences in the same manner as described above for DI. We selected eight frames from each short video sequence at equal intervals to reduce memory usage. We then used a pretrained TimeSformer model with Kinetics-400 and fine-tuned it with 1500 short video sequences created from the training set. We used divided space-time attention as a self-attention scheme. Finally, we obtained feature maps by inputting the short video sequences into TimeSformer.

Table 4 shows the n-th matching rate and nAUC of person identification using the features of our method of Section 3.1 and the existing methods [4, 3, 8, 2] that are based on human movement. The results indicate that the features of the proposed method are superior to those of the existing methods with respect to person identification accuracy. This experiment provided further confirmation that the proposed method can enhance the accuracy of person identification in the presence of headwear variation.

4 Conclusions

In this paper, we proposed a method of extracting temporal features that are robust to headwear variation for person identification using a video sequence of body sway observed by an overhead camera. When people wear headwear, their head shapes in a video sequence become different even though they have the same identity. To alleviate the influence of head shape changes caused by headwear variation, we measured time-series signals by estimating center positions from each person's head shapes. Furthermore, we discovered that frequency powers of the time-series signal in the low-frequency band contained components that were more informative for person identification in the presence of headwear variation. Therefore, we designed a learning-based low-pass filter using separation metrics to automatically extract the frequency powers in the low-frequency band from the time-series signal. The experimental results show that the proposed method using the learning-based low-pass filter can extract temporal features that are informative for person identification in the presence of headwear variation. Moreover, we confirmed that the proposed method achieves significantly higher accuracy of person identification than existing methods that were designed for extracting features from human movement.

In future work, we will further enhance the accuracy of person identification in the presence of headwear variation by examining the underlying mechanism of body sway. Furthermore, we plan to identify people wearing other types of headwear.

Acknowledgments We thank Professor Yoshio Iwai for his valuable advice and suggestions during this research.

References

1. Bedagkar-Gala, A., Shah, S.K.: A survey of approaches and trends in person re-identification. Image and vision computing **32**(4), 270–286 (2014)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (2021)
3. Bilen, H., Fernando, B., Gavves, E., Vedaldi, A.: Action recognition with dynamic image networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 2799–2813 (2017)
4. Han, J., Bhanu, B.: Individual recognition using gait energy image. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(2), 316–322 (2006)
5. Kamitani, T., Yamaguchi, Y., Nishiyama, M., Iwai, Y.: Identifying people using body sway in case of self-occlusion. In Proceedings of International Workshop on Frontiers of Computer Vision pp. 1–13 (2020)
6. 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)
7. Otsu, N.: A threshold selection method from gray-level histograms. IEEE transactions on systems, man, and cybernetics **9**(1), 62–66 (1979)

8. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision pp. 4489–4497 (2015)
9. 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)
10. Xu, N., Price, B., Cohen, S., Huang, T.: Deep image matting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 2970–2979 (2017)
11. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision pp. 1116–1124 (2015)
12. Zheng, L., Zhang, H., Sun, S., Chandraker, M., Yang, Y., Tian, Q.: Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 1367–1376 (2017)