SOANets: Encoder-decoder based Skeleton Orientation Alignment Network for White Cane User Recognition from 2D Human Skeleton Sequence

Naoki Nishida[1], Yasutomo Kawanishi[1], Daisuke Deguchi[2], Ichiro Ide[1], Hiroshi Murase[1] and Jun Piao[3]
[1]*Graduate School of Informatics, Nagoya University, Aichi, Japan*
[2]*Information Strategy Office, Nagoya University, Aichi, Japan*
[3]*Data Science Research Laboratories, NEC Corporation, Kanagawa, Japan*

Keywords: Skeleton Orientation Alignment, Skeleton Representation Sequence, White Cane User Recognition.

Abstract: Recently, various facilities have been deployed to support visually impaired people. However, accidents caused by visual disabilities still occur. In this paper, to support the visually impaired people in public areas, we aim to identify the presence of a white cane user from a surveillance camera by analyzing the temporal transition of a human skeleton in a pedestrian image sequence represented as 2D coordinates. Our previously proposed method aligns the orientation of the skeletons to various orientations and identifies a white cane user from the corresponding sequences, relying on multiple classifiers related to each orientation. The method employs an exemplar-based approach to perform the alignment, and heavily depends on the number of exemplars and consumes excessive memory. In this paper, we propose a method to align 2D skeleton representation sequences to various orientations using the proposed Skeleton Orientation Alignment Networks (SOANets) based on an encoder-decoder model. Using SOANets, we can obtain 2D skeleton representation sequences aligned to various orientations, extract richer skeleton features, and recognize white cane users accurately. Results of an evaluation experiment shows that the proposed method improves the recognition rate by 16%, compared to the previous exemplar-based method.

1 INTRODUCTION

In recent years, various kinds of facilities have been deployed to support visually impaired people. Therefore, they have become able to go out on their own actively. For example, braille blocks can be found around cities and public facilities to guide visually impaired people (*e.g.* low-vision and partially sighted). However, accidents involving them still occur, such as, falling on a railway track from a station platform. To prevent such accidents, platform screen doors are installed at stations. However, as their installation is limited to major stations, social support is still necessary to prevent accidents.

Therefore, necessity to identify visually impaired people in the public space is increasing. Information from surveillance cameras installed in public places can be used for this purpose. For example, Tanikawa et al. proposed a method to automatically recognize and track wheelchair users in security camera footages (Tanikawa et al., 2017), and then to notify personnel to support them promptly.

Usually, it is not necessary to provide sighted people with notifications intended for visually impaired people. Therefore, it is essential to distinguish sighted and visually impaired people to provide support only for the latter.

Visually impaired people usually employ a white cane to search for obstacles. It also serves as a medium that helps other people recognize their impairment. Therefore, a white cane detector can be used to identify visually impaired people in images from surveillance cameras. However, even state-of-the-art object detectors (He et al., 2017; Redmon and Farhadi, 2018; Cai and Vasconcelos, 2018) may mis-detect objects with appearances similar to a white cane, such as a white umbrella.

To overcome this issue, it is preferable to recognize them not only by the presence of a white cane, but also by unique actions peculiar to persons walk-

Nishida, N., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H. and Piao, J.
SOANets: Encoder-decoder based Skeleton Orientation Alignment Network for White Cane User Recognition from 2D Human Skeleton Sequence.
DOI: 10.5220/0008988204350443
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 435-443
ISBN: 978-989-758-402-2; ISSN: 2184-4321

(a) White-cane user (b) Sighted person

Figure 1: Walking actions of a white cane user and a sighted person. The difference is mainly seen in the movement of their arms. The sighted person swings his arms back and forth. In contrast, the arm of the white cane user holding his cane is fixed forward.

ing with a white cane. It is evident that there are differences between the actions of a white cane user and a sighted person as shown in Figure 1. Therefore, in this paper, we focus on recognizing a white cane user analyzing only the motions typical to pedestrians.

Various studies have been performed to propose an efficient way to recognize actions of human skeletons (Le et al., 2018; Baptista et al., 2019; Wang et al., 2016). As it is difficult to create a 3D representation of a skeleton using the data from a monocular camera, a sequence of 2D skeleton representations is usually considered for estimating human actions. However, as the appearance of a 2D skeleton varies widely according to the skeleton's orientation as shown in Figure 2, the performance of action recognition may be deteriorated depending on the orientation. Therefore, it is desirable to align a skeleton orientation in the input for a white cane user recognition to a specific orientation. However, since there may be several orientations that work effectively for user recognition, not only a single orientation but also several orientations combined together can be effective. Hence, to achieve more accurate recognition, it is required to obtain several human skeleton representation sequences aligned to different orientations.

We have previously proposed a skeleton orientation alignment method based on exemplar-based 2D skeletons for white cane user recognition (Nishida et al., 2019). We prepared a database with 2D skeletons in various orientations and aligned the input 2D skeletons to various orientations by finding suitable skeletons in this database. The performance of the method depended on the number of skeleton exemplars contained in the database. Using this method, slightly different skeletons could be converted to the same skeleton, as the dictionary is discrete. Therefore, unnatural skeleton representation sequences could be generated.

To tackle this problem, we propose an encoder-decoder model named Skeleton Orientation Alignment Networks (SOANets) for skeleton orientation alignment, as an alternative to the above exemplar-based approach. Each SOANet corresponding to a target orientation aligns an arbitrary orientation of an

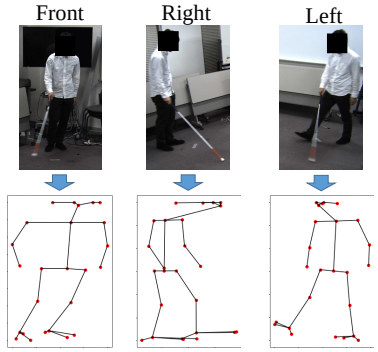

Figure 2: Difference in skeletons depending on their orientations.

input 2D skeleton representation to the target orientation. As a result, we can obtain skeletons in multiple orientations corresponding to a single input 2D skeleton. Unlike in the exemplar-based approach, we can obtain a continuous skeleton representation and achieve natural skeleton representation sequences owing to the fact that a skeleton representation is regressed directly by SOANets.

The process of recognizing a white cane user is similar to that of the exemplar-based approach. By using aligned skeleton representation sequences obtained through SOANets, identification of whether each skeleton representation sequence corresponds to a white cane user or not is performed for all orientations independently. Finally, a classification result is obtained by applying weight to the classification results corresponding to each orientation of the considered input skeleton representation sequences and aggregating the classification results.

Contributions of the present paper are summarized as follows:

- We introduce SOANets based on an encoder-decoder model to align an orientation of a 2D skeleton representation to multiple orientations. As the proposed method can output a skeleton representation in a form of continuous sequence, the output will be much natural than that of our previously proposed exemplar-based approach.
- We achieve more accurate classification of white cane users by replacing the skeleton orientation alignment of the exemplar-based approach with the proposed SOANets.
- Through performance evaluation using the images collected in several real environments, we demonstrate that the proposed method achieves the highest accuracy for white cane user recognition.

The rest of this paper is organized as follows. In section 2, we describe related research works. In section

3, we present the proposed method to classify pedestrian's 2D skeleton representation sequences by the orientation alignment of a 2D skeleton representation sequence. In section 4, we describe the conducted experiments and discuss their results. In section 5, we conclude this paper and discuss the future directions.

2 RELATED WORK

Considering the task of human action recognition, it is necessary to extract the temporal transition of human features. To address this, Recurrent Neural Network (RNN) is often used to recognize continuous sequences such as sentences and videos. In particular, Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997), a type of RNN, is employed to handle long-term sequences, and methods based on this approach have achieved high accuracy for action recognition (Sun et al., 2017; Si et al., 2019).

Recently, representation of human skeletons is often used as a feature to recognize human actions. Convolutional pose machine (Wei et al., 2016) and OpenPose (Cao et al., 2017) are mentioned as well-known methods for human skeleton estimation. These methods are used to estimate 2D coordinates of each body joint that composes a human body by the Convolutional Neural Network (CNN).

It is desirable that a skeleton is represented in 3D rather than in 2D, as a 2D skeleton greatly differs depending on the skeleton's orientation. Therefore, some researchers have used a 3D skeleton for action recognition (Le et al., 2018; Baptista et al., 2019; Wang et al., 2016). However, in these research works, 3D skeletons are prepared in advance or are estimated based on images captured from multiple cameras. In a real scene, as it is difficult to install multiple cameras for capturing people simultaneously everywhere, it is more efficient if a skeleton is estimated from a single image. Recently, a method to estimate a 3D skeleton from a 2D skeleton was proposed (Martinez et al., 2016), but it is still inaccurate.

To realize recognition of human actions, we have proposed a 2D skeleton orientation alignment method by an exemplar-based method (Nishida et al., 2019). The alignment is performed by obtaining skeleton representations from the skeleton database prepared in advance. Using the skeleton orientation alignment, richer features of a 2D skeleton with various orientations can be obtained compared to methods based on a single skeleton orientation.

3 WHITE CANE USER RECOGNITION FROM VARIOUS ORIENTATIONS BY SOANets

When recognizing a white cane user based on a skeleton representation sequence, there is a problem that the appearance of a skeleton greatly varies depending on the skeleton's orientation. To address this problem, we basically follow the framework proposed in our previously proposed exemplar-based method (Nishida et al., 2019). However, this method has a deficiency in the skeleton alignment process, as described in the previous section. To overcome this deficiency, we propose the SOANets method regressing the orientation-aligned 2D skeleton instead of performing exemplar-based skeleton orientation alignment.

The procedure of the white cane user recognition framework is shown in Figure 3 (The part proposed in this paper is indicated by a double-lined box). First, a pedestrian image sequence is used as an input, and 1) the skeleton of the pedestrian in each frame is estimated. Then, 2) for each frame, 2D skeleton representation sequences with various orientations are obtained from the 2D skeletons by the proposed SOANets. Finally, 3) the aligned 2D skeleton representation sequences are classified by a classifier corresponding to their orientation, and the results are integrated to output the final decision for the input sequence.

Here, we propose the SOANets procedure discussed in point 2). In this method, we align the orientation of a 2D skeleton to various orientations. We use an encoder-decoder model for a target orientation to align an input 2D skeleton. Each SOANet regresses the 2D skeleton aligned to a certain orientation. We obtain 2D skeletons of various orientations by encoder-decoder models corresponding to each orientation. The collection of these encoder-decoder models are named as SOANets.

Details of each procedure is presented in the rest of this section.

3.1 Skeleton Estimation of a Pedestrian

We define a human skeleton using a set of 2D coordinates of body joints such as wrists, elbows, knees, etc. Assuming that the number of body joint points is J, a 2D skeleton with a certain orientation can be represented as $\mathbf{p} \in \mathbb{R}^{2J}$. Here, a 2D skeleton of the n-th frame in a pedestrian image sequence is estimated as $\mathbf{p}_n = (x_n^1, y_n^1, ..., x_n^j, y_n^j, ..., x_n^J, y_n^J)^T$, $x_n^j, y_n^j \in \mathbb{R}$. The

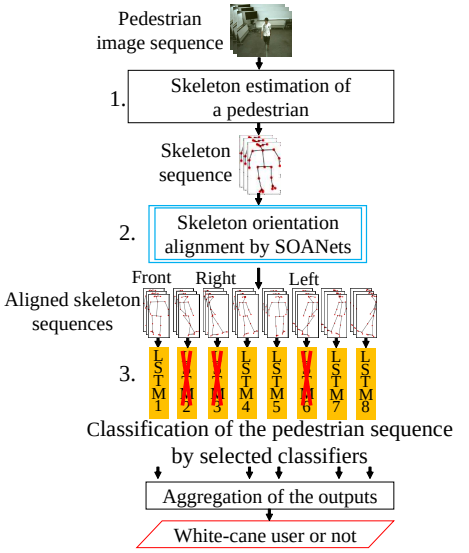

Figure 3: Procedure of the white cane user recognition following the exemplar-based approach. The double-lined box indicates the part proposed in the present paper.

sequence $I = \{\mathbf{I}_1, ..., \mathbf{I}_n, ..., \mathbf{I}_N\}$ consists of N color images obtained by tracking pedestrians, whose size is $w \times h$ [pixels].

We use OpenPose (Cao et al., 2017) for 2D skeleton estimation. For each frame $\mathbf{I}_n$, the method estimates heat maps indicating probabilities of all body joints, and part affinity fields indicating the connection between each body joint pair. These maps and image features are used as input, and a 2D skeleton representation $\mathbf{p}_n$ and its probability o_n are the output.

From the estimated 2D skeleton representation sequence $\mathcal{P} = \{\mathbf{p}_n\}_{n=1}^{N}$, we construct a 2D skeleton representation sequence $\mathcal{S}$ as the input for the next process (Skeleton orientation alignment). Here, we eliminate the frames of low-confident estimations, and construct $\mathcal{S}$ from $\mathcal{P}$ as follows:

$$\mathcal{S} = \{\mathbf{p}_n | \forall n, o_n \leq \tau\}, \tag{1}$$

where o_n is the probability of $\mathbf{p}_n$. For the estimated coordinates (x_q^j, y_q^j) of each body joint, their value range is normalized as follows:

$$x_q^{'j} = \frac{x_q^j - \min_i(x_q^i)}{\max_i(x_q^i) - \min_i(x_q^i)}\theta_x, \tag{2}$$

$$y_q^{'j} = \frac{y_q^j - \min_i(y_q^i)}{\max_i(y_q^i) - \min_i(y_q^i)}\theta_y, \tag{3}$$

where θ_x and θ_y are the constants to adjust the width and height of a skeleton. Examples of the estimated 2D skeletons are shown in Figure 4.

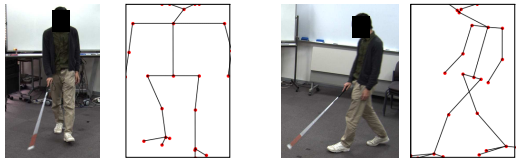

Figure 4: Examples of estimated 2D skeletons.

In addition, we expand body joint coordinates $\mathbf{p}_n$ and skeleton representation sequence $\mathcal{S}$ to clarify whether the coordinates of body joints can be estimated:

$$\mathbf{p}_n^{'} = (x_n^{'1}, y_n^{'1}, f_n^1, ..., x_n^{'j}, y_n^{'j}, f_n^j, ..., x_n^{'J}, y_n^{'J}, f_n^J), \tag{4}$$

$$\mathcal{S}^{'} = \{\mathbf{p}_n^{'} | \forall n, o_n \leq \tau\}, \tag{5}$$

$$f_n^j = \begin{cases} 1 & (Successfully\ detected) \\ 0 & (Misdetected) \end{cases} \tag{6}$$

3.2 Skeleton Orientation Alignment using SOANets

Using SOANets, a 2D skeleton representation $\mathbf{p}_n^{'}$ is transformed to a set of 2D skeleton representations $\{\tilde{\mathbf{p}}_{nd}\}_{d=1}^{D}$ viewed from D different orientations. Here, the function $T(\mathbf{p}_n^{'})$ performing this transformation is defined as follows:

$$T(\mathbf{p}_n^{'}) = \{\tilde{\mathbf{p}}_{n1}, ..., \tilde{\mathbf{p}}_{nd}, ..., \tilde{\mathbf{p}}_{nD}\}, \tag{7}$$

We perform the skeleton orientation alignment by encoder-decoder models. Each encoder-decoder model transforms an input 2D skeleton into an arbitrary orientation to the aligned 2D skeleton representation in a specific orientation as shown in Figure 5. This set of encoder-decoder networks is used to regress the 2D skeleton that is aligned to the input skeleton orientation. We can obtain skeletons in various orientations from a single input skeleton as shown in Figure 6. All encoder-decoder models have the same architecture as presented in Figure 7. The collection of these the encoder-decoder models are named as SOANets. The function $T_d(\mathbf{p}_n^{'})$ consists of several SOANets, and each SOANet is defined as follows:

$$T_d(\mathbf{p}_n^{'}) = \text{Decoder}_d(\text{Encoder}_d(\mathbf{p}_n^{'})) = \tilde{\mathbf{p}}_{nd}. \tag{8}$$

Therefore, $T(\mathbf{p}_n^{'})$ also can be defined as follows:

$$T(\mathbf{p}_n^{'}) = \{T_1(\mathbf{p}_n^{'}), ..., T_d(\mathbf{p}_n^{'}), ..., T_D(\mathbf{p}_n^{'})\}. \tag{9}$$

Next, we describe the details on inputs and outputs of SOANets. To realize the skeleton orientation alignment from input skeletons in various orientations, we prepare 2D skeletons in D orientations as the training input.

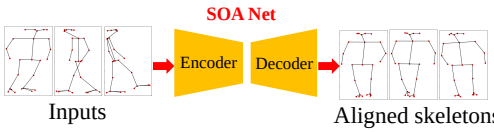

Figure 5: Skeletons in any orientation is aligned to a single orientation using SOANet.

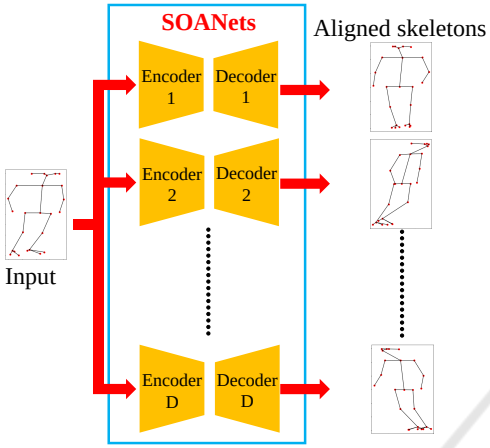

Figure 6: Skeleton orientation alignment using SOANets.

It is difficult to obtain a representation of the body joints of a skeleton due to occlusion, distance from cameras, etc. Therefore, misdetection of some body joints affects not only the skeleton orientation alignment, but also classification using aligned skeleton representation sequences. To solve this issue, we use skeleton representations with missing values of body joint coordinates as the input, and the complete skeleton representations as the output.

Using the training data, SOANets are trained to regress the complete 2D skeleton representation $\tilde{\mathbf{p}}_{nd} = (\tilde{x}^1_{nd}, \tilde{y}^1_{nd}, ..., \tilde{x}^J_{nd}, \tilde{y}^J_{nd})$ in the aligned orientation d from an input skeleton representation sequence $\mathbf{p}'_n \in \mathbb{R}^{3J}$. As all missing values corresponding to body joints are supposed to be restored, f^j_n indicating the existence of the missing value of body joints is excluded from $\tilde{\mathbf{p}}_n \in \mathbb{R}^{2J}$.

Finally, orientations of all 2D skeletons in the input skeleton representation sequence $\mathcal{S}' \in \mathbb{R}^{3JN}$ are aligned by SOANets. The aligned skeleton representation sequences $\{\widetilde{\mathcal{S}}_d\}_{d=1}^D$ to D orientations are denoted as follows:

$$\{\widetilde{\mathcal{S}}_d\}_{d=1}^D = \{(\tilde{\mathbf{p}}_{1d}, ..., \tilde{\mathbf{p}}_{nd}, ..., \tilde{\mathbf{p}}_{Nd})\}_{d=1}^D. \quad (10)$$

Each $\widetilde{\mathcal{S}}_d$ is used as the input for classification to identify, whether a pedestrian is a white cane user or not. An example of aligned 2D skeletons obtained by the proposed SOANets from an input 2D skeleton is shown in Figure 8.

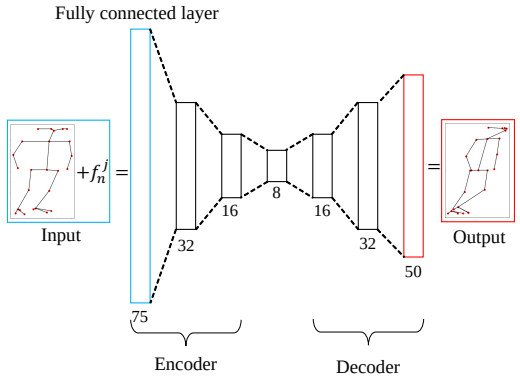

Figure 7: Encoder-decoder architecture of each SOANet.

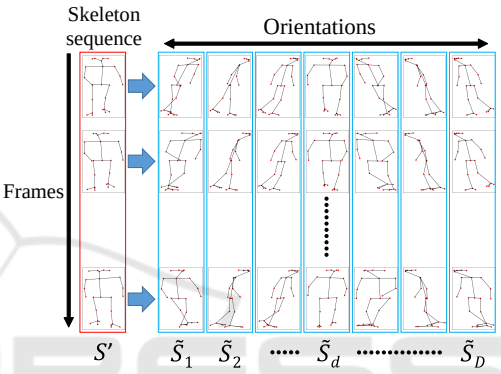

Figure 8: Examples of aligned 2D pseudo-skeletons.

3.3 Classification of a Pedestrian Sequence

Each 2D skeleton representation sequence $\widetilde{\mathcal{S}}_d$ obtained by SOANets is classified, whether it corresponds to a white cane user or not according to the classifier corresponding to each skeleton orientation. We prepare D individual classifiers $\mathcal{C} = \{C_1, ..., C_d, ..., C_D\}$ for D different skeleton orientations obtained using the methods based on LSTM network and the exemplar-based approach.

Following the exemplar-based approach, all classification scores of each class (a white cane user or a sighted person) are weighted and integrated as follows:

$$g^\ell(\widetilde{\mathcal{S}}) = \frac{1}{D}\sum_{d=1}^{D} w_d g^\ell_d(\widetilde{\mathcal{S}}), \quad (11)$$

$$w_d = \begin{cases} 1 & a(C_d) \geq \delta, \\ 0 & a(C_d) < \delta, \end{cases}, \quad (12)$$

where $g^\ell(\widetilde{\mathcal{S}})$ is the integrated classification score of class ℓ; $g^\ell_d(\widetilde{\mathcal{S}})$ is the classification score of C_d; w_d is

the weight for $g_d^\ell(\widetilde{\mathcal{S}})$ of all classes; $a(C_d)$ is the accuracy of the classifier C_d for the training dataset; and δ is a threshold. Finally, the class $\tilde{\ell} = \text{argmax}_\ell\, g^\ell(\widetilde{\mathcal{S}})$ with the highest score is output as the classification result.

4 EVALUATION

In this section, we introduce two experiments: 1) evaluation of the skeleton orientation alignment performed by SOANets, and 2) confirmation of the effectiveness of the proposed method in terms of pedestrian classification.

4.1 Dataset for Training SOANets

For training SOANets, we captured images of pedestrians using three calibrated cameras. We captured the data at one specific location, and the same participant performed both the roles of a sighted and a visually impaired person. We estimated their 3D poses by OpenPose and obtained M $(= 4,616)$ complete 3D skeleton representations. The skeleton's orientation that faces the front is labeled as 0°, and the other orientations are set by rotating the skeleton counterclockwise with the step of 10° around the vertical axis. As a result, using a 3D skeleton representation, D $(= 36)$ sets of 2D complete skeletons in D orientations are obtained. Finally, training data for SOANets is composed of MD $(= 166,176)$ 2D complete skeleton representations.

4.2 Dataset for Classification

For training and testing data, we prepared a dataset by capturing sequences of several walking white cane users and sighted pedestrians in the same manner as in our previous study (Nishida et al., 2019). In the experiment, seventeen sighted participants played both roles, and five visually impaired people also participated as white cane users. The sequences were captured at five different locations including both indoors and outdoors. We composed pedestrian sequences by selecting frames where pedestrians existed, resulting in 266 pedestrian sequences. The details of the prepared dataset are summarized in Table 1, and examples of the images at each location are presented in Figure 9.

Table 1: Number of pedestrian image sequences in the dataset.

Location	1	2	3	4	5	All
#White-cane user	23	12	12	10	76	133
#Non user	26	6	25	0	76	133
#All sequences	49	18	37	10	152	266

Figure 9: Examples of images at each location.

4.3 Experiment 1: Evaluation of the Skeleton Orientation Alignment

4.3.1 Settings

In this experiment, we evaluate the performance of the skeleton orientation alignment by SOANets. As a metric for evaluation we use Root Mean Squared Error (RMSE) computed from sets of E aligned skeletons $\{\mathbf{p}_e^{a\ell} = (x_{e1}^{a\ell}, y_{e1}^{a\ell}, ..., x_{eJ}^{a\ell}, y_{eJ}^{a\ell})\}$ and E corresponding ground-truth skeletons $\{\mathbf{p}_e^{GT} = (x_{e_1}^{\text{GT}}, y_{e_1}^{\text{GT}}, ..., x_{e_J}^{\text{GT}}, y_{e_J}^{\text{GT}})\}$. Here, RMSE is defined as follows:

$$\text{RMSE} = \frac{1}{E}\sum_{e=1}^{E}\sqrt{\frac{1}{J}\sum_{j=1}^{J}\{(x_{e_j}^{\text{GT}} - x_{e_j}^{\text{a}\ell})^2 + (y_{e_j}^{\text{GT}} - y_{e_j}^{\text{a}\ell})^2\}} \tag{13}$$

We considered two to five random body joints from the input skeleton as the misdetected body joints. This pseudo-misdetection is applied iteratively ten times for each input to augment the data. Therefore, the number of skeleton data for SOANets is $10MD$ $(= 1,661,760)$, where 1,600,000 skeletons are used for training, and 61,760 skeletons are used for testing.

4.3.2 Result

The results of applying the proposed method based on SOANets for all considered aligned skeleton orientations are shown in Figure 10. Examples of the skeleton orientation alignment using SOANets are

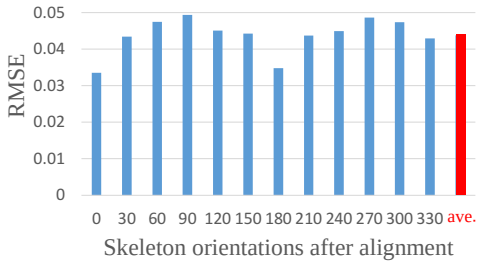

Figure 10: Results of the skeleton orientation alignment evaluation.

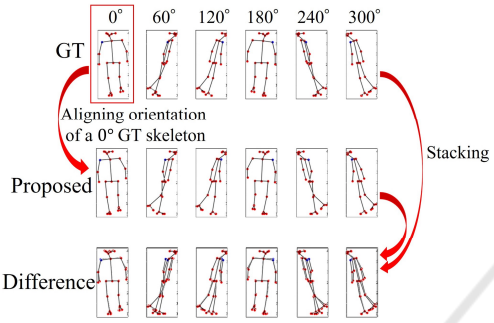

Figure 11: Examples of the skeleton orientation alignment.

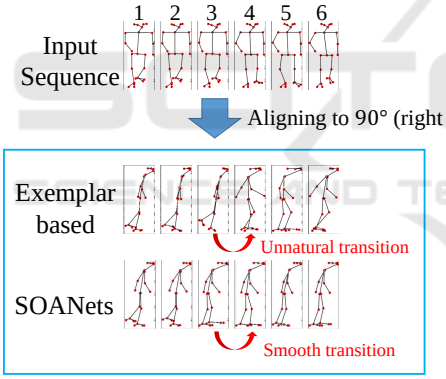

Figure 12: Examples of the skeleton orientation alignment by the exemplar-based approach and SOANets (proposed).

presented in Figure 11, and comparison of the skeleton orientation alignment results by SOANets and the exemplar-based approach is shown in Figure 12.

4.3.3 Discussion

In this section, we discuss the results of the skeleton orientation alignment. From Figure 10, we can see that RMSE is the lowest when the aligned orientations are 0° and 180° corresponding to the front and back orientations, respectively. In contrast, when the aligned orientations are 90° and 270° corresponding to right and left orientations, respectively, RMSE is the highest. The depth of arm and leg body joints is considered to be the cause of this difference of RMSE.

For a 2D skeleton in the front or back orientations, it is difficult to estimate the depth of limbs. However, for a 2D skeleton in the side orientation, the limbs appear clearly. Therefore, we consider that the alignment error becomes larger due to the lack of information on the limb depth, when a front or back skeleton representation sequence is aligned to the side orientation skeleton. As we can see in the bottom figure of Figure 11, the error related to the limb joints is larger than that of other joints. To address this problem, it is necessary to obtain information on the limb depth from a front or back skeleton representation sequence. For example, we can consider the use of multiple frames for input of SOANets and to obtain more features than a single frame.

From Figure 12, we can see that the aligned skeleton representations obtained using the previously proposed exemplar-based method change greatly from the third to the fourth frame. As it aligns the skeleton orientation by obtaining skeleton representations from the database, this alignment is limited to the number of skeleton patterns available in the database. Therefore, in Figure 12, the exemplar-based approach aligns the skeleton representation sequence with unnatural transition. On the other hand, since the proposed SOANets regress the aligned skeleton directly from an input skeleton, it has no such limitation and was able to align the skeleton orientation more naturally.

4.4 Experiment 2: Evaluation of Pedestrian Classification

4.4.1 Settings

In this experiment, we evaluate the accuracy of classifying pedestrian sequences.

For evaluation, the length of each 2D skeleton representation sequence is set to 64 frames, and each sequence is divided into overlapping five sequences, which is re-composed of 32 frames of the input skeleton representation sequence with the step of eight. The total number of skeleton representation sequences (before applying the skeleton orientation alignment) is $266 \times 5 = 1{,}330$. The number of detected body joints for each 2D skeleton is $J = 25$, and the value range of coordinate values of body joints is $[-1.0, 1.0]$ in the horizontal and vertical directions with $\theta_x = 1.0$ and $\theta_y = 1.0$. For evaluation, five-fold cross-validation is performed on the dataset obtained at each of five locations considered for evaluation, and the dataset of other four locations is used for training the SOANets.

To estimate the performance of the proposed

Table 2: Classification results.

Location	1	2	3	4	5	All
No alignment	0.82	0.70	0.55	0.53	0.64	0.66
Exemplar-based						
No weighting	0.77	0.70	0.76	0.52	0.80	0.77
Weighting	0.80	0.69	**0.75**	0.50	0.81	0.78
SOANets						
No weighting	0.85	0.84	0.68	0.64	**0.82**	0.80
Weighting(Ours)	**0.87**	**0.90**	0.71	**0.70**	**0.82**	**0.82**

method, we compare the accuracy of the following five methods:

- No alignment of the skeleton's orientation for an input skeleton representation and using a single classifier (No alignment).
- Aligning the skeleton's orientation by the exemplar-based approach and integrating all results of classifiers without weighting (Exemplar-based, No weighting).
- Aligning the skeleton's orientation by the exemplar-based approach and integrating all results with weighting (Exemplar-based, Weighting).
- Aligning the skeleton's orientation by the SOANets method and integrating all results without weighting (SOANets, No weighting).
- Aligning the skeleton's orientation by the SOANets method and integrating all results with weighting (Proposed: SOANets, Weighting).

4.4.2 Results

The results are summarized in Table 2. In this table, "All" indicates the average of the results in all considered locations weighted by the number of samples shown in Table 1. The accuracy of the proposed method improved by 16% compared with the method without the skeleton orientation alignment. Moreover, the accuracy of the proposed method was improved compared with the previously proposed exemplar-based approach with and without weighting considering each classifier. As a result, the effectiveness of the proposed method was confirmed.

4.4.3 Discussion

Here, we discuss the experimental results. We focused on following two points : 1) locations, where the data were captured, and 2) weighting of classifiers.

First, we discuss the difference in the results in terms of the locations, where the data were captured. As shown in Table 2, for all the considered methods, the accuracy at location 4 is relatively low. There

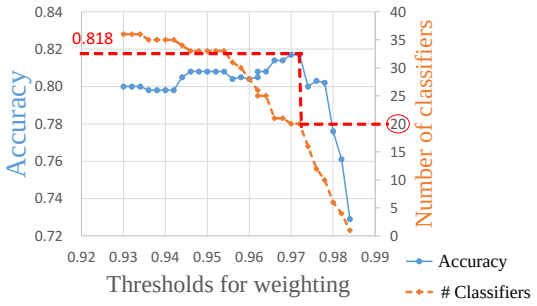

Figure 13: Relation of the number of classifiers and the accuracy.

are two possible reasons for this. One is that unlike other locations, it contains only white cane users and no pedestrians without white canes. The other is that while capturing the data at location 4, the camera position was relatively higher than that at other locations, and thereby, the tilt angle was different from the others. Therefore, we consider that the classification accuracy deteriorated due to the fact that skeleton patterns were different from those at other locations. To mitigate this problem, it is necessary to capture data by changing the camera position, target location, and subjects multiple times.

Second, we discuss applying different weights to classifiers. Let us investigate changes in the number of classifiers used for evaluation ($w_d = 1$) and the corresponding accuracy based on the results provided in Figure 13. The graph is drawn by changing the threshold parameter δ, which controls the number of classifiers. The accuracy improved when the number of classifiers used for evaluation decreased observing that the highest accuracy was obtained with the application of twenty classifiers, and $\delta = 0.972$. However, the accuracy rapidly decreased when the number of classifiers were less than twenty. Based on this observation, we can conclude that it is necessary to define the required number of classifiers corresponding to each skeleton's orientation to maintain high accuracy. Classifiers that were not used in the evaluation mainly correspond to the orientations of the front and the back. The reason for this is that a 2D skeleton representation in the front or back orientations provides less information on the depth of limbs, as described in 4.3.3. However, if a white cane user employs a white cane by swinging his/her arm left and right, the skeleton representations in the front or back orientations is important to recognize such action. Therefore, we plan to introduce a mechanism to select important classifiers according to the action presented in the input skeleton representation sequence.

5 CONCLUSION

In this paper, we aimed to solve the problem of recognizing white cane users by classifying pedestrians from the temporal transition of their skeletons. We proposed a 2D skeleton orientation alignment method named SOANets. Through experiments, we demonstrated that the accuracy of analyzing pedestrian image sequences improves by incorporating the skeleton orientation alignment of an input 2D skeleton, and the effectiveness of the proposed method was confirmed.

In the future, we plan to train SOANets with more action patterns. We will also integrate object recognition methods that can directly detect a white cane.

ACKNOWLEDGMENTS

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research

REFERENCES

Baptista, R., Ghorbel, E., Papadopoulos, K., Demisse, G. G., Aouada, D., and Ottersten, B. (2019). View-invariant action recognition from RGB data via 3D pose estimation. In *Proceeding of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2542–2546.

Cai, Z. and Vasconcelos, N. (2018). Cascade R-CNN: Delving into High Quality Object Detection. In *Proceeding of the 2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162.

Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). Real-time multi-person 2D pose estimation using part affinity field. In *Proceeding of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. (2017). Mask R-CNN. In *Proceeding of the 2017 IEEE International Conference on Computer Vision*, pages 2961–2969.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Le, T. M., Inoue, N., and Shinoda, K. (2018). A fine-to-coarse convolutional neural network for 3D human action recognition. In *Proceeding of the 29th British Machine Vision Conf*, pages 184–1–184–13.

Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2016). A simple yet effective baseline for 3D human pose estimation. In *Proceeding of the 2017 IEEE International Conference on Computer Vision*, pages 2640–2649.

Nishida, N., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., and Piao, J. (2019). Exemplar-based Pseudo-Viewpoint Rotation for White-Cane User Recognition from a 2D Human Pose Sequence. In *Proceeding of the 16th IEEE International Conference on Advanced Video and Signal-based Surveillance*, number Paper ID 29.

Redmon, J. and Farhadi, A. (2018). YOLOv3: An Incremental Improvement. *Computing Research Repository*, (arXiv:1804.02767).

Si, C., Chen, W., Wang, W., Wang, L., and Tan, T. (2019). An attention enhanced graph convolutional LSTM network for skeleton-based action recognition. In *Proceeding of the 2019 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1227–1236.

Sun, L., Jia, K., Chen, K., Yeung, D. Y., Shi, B. E., and Savarese, S. (2017). Lattice long short-term memory for human action recognition. In *Proceeding of the 2017 IEEE International Conference on Computer Vision*, pages 2147–2156.

Tanikawa, U., Kawanishi, Y., Deguchi, D., Ide, I., Murase, H., and Kawai, R. (2017). Wheelchair-user Detection Combined with Parts-based Tracking. In *Proceeding of the 12th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, volume 5, pages 165–172.

Wang, C., Wang, Y., and Yuille, A. L. (2016). Mining 3D key-pose-motifs for action recognition. In *Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2639–2647.

Wei, S. E., Ramakrishna, V., Kanede, T., and Sheikh, Y. (2016). Convolutional pose machines. In *Proceeding of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4732.