Exemplar-based Pseudo-Viewpoint Rotation for White-Cane User Recognition from a 2D Human Pose Sequence

Naoki Nishida, Yasutomo Kawanishi
Graduate School of Informatics, Nagoya University, Aichi, Japan
nishidan@murase.is.i.nagoya-u.ac.jp, kawanishi@i.nagoya-u.ac.jp
Daisuke Deguchi
Information Strategy Office, Nagoya University, Aichi, Japan
ddeguchi@nagoya-u.jp
Ichiro Ide, Hiroshi Murase
Graduate School of Informatics, Nagoya University, Aichi, Japan
ide, murase@i.nagoya-u.ac.jp
Jun Piao
Data Science Research Laboratories, NEC Corporation, Kanagawa, Japan
j-piao@cw.jp.nec.com

Abstract

In recent years, various facilities are equipped to support visually impaired people, but accidents caused by visual disabilities still occur. In this paper, to support the visually-impaired people in a public space, we aim to classify whether a pedestrian image sequence obtained by a surveillance camera is a white-cane user or not from the temporal transition of a human pose represented as 2D coordinates. However, since the appearance of the 2D pose varies largely depending on the viewpoint of the pose, it is difficult to classify them. So, in this paper, we propose a method to rotate the viewpoint of a pose from various pseudo-viewpoints based on a pair of 2D poses simultaneously observed and classify the sequence by multiple classifiers corresponding to each viewpoint. Viewpoint rotation makes it possible to obtain pseudo-poses seen from various pseudo-viewpoints, extract richer pose features, and recognize white-cane users more accurately. Through an experiment, we confirmed that the proposed method improves the recognition rate by 12% compared to the method not employing viewpoint rotation.

1. Introduction

In recent years, since various facilities are equipped to support visually impaired people, it is becoming ready that they can go out actively. For example, braille blocks can be found throughout cities and public facilities to guide visually impaired people. However, accidents caused by visual disabilities still occur, such as falling to a track from a station platform.

To prevent such accidents, platform screen doors are being installed at stations. However, since their installations are limited to major stations, human assistance is still necessary to prevent accidents.

This leads to the necessary of means to find visually impaired people from the public space, and surveillance cameras installed in public places are expected to serve this purpose. For example, Tanikawa et al. proposed a method to automatically recognize and track wheelchair users in security camera images [1], in order to automatically notify personnels to support them as soon as possible.

Usually, it is not necessary to provide sighted people with notifications that are very important for visually impaired people. Therefore, it is necessary to distinguish sighted and visually impaired people to provide support only for the latter.

Visually impaired people usually possess a white-cane to search for obstacles. It also serves as informing other people about their existence. Therefore, the existence of a white-cane can be used for finding visually impaired people from an image, and thus we aim to recognize visually impaired people who possess a white-cane (white-cane users) in video sequences.

Although appearance-based object detectors such as YOLO [2] can be used to detect a white-cane around a pedestrian, it may mis-detect objects whose appearances are

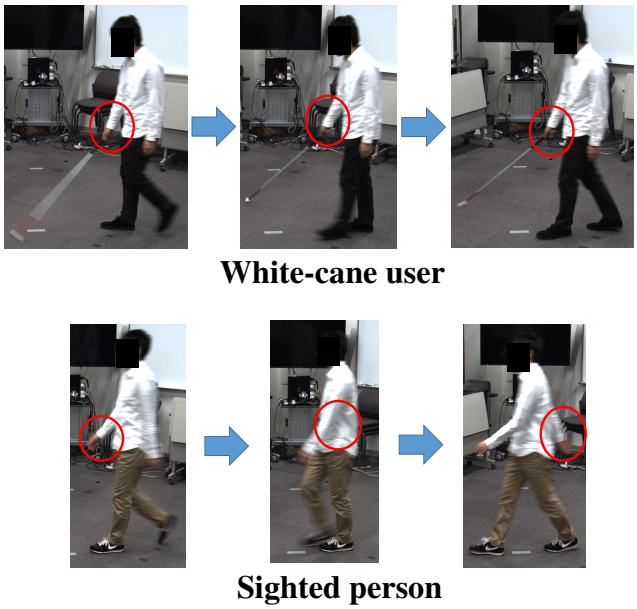

Figure 1. Walking actions of a white-cane user and a sighted person. The difference is mainly seen in the movement of their arm.

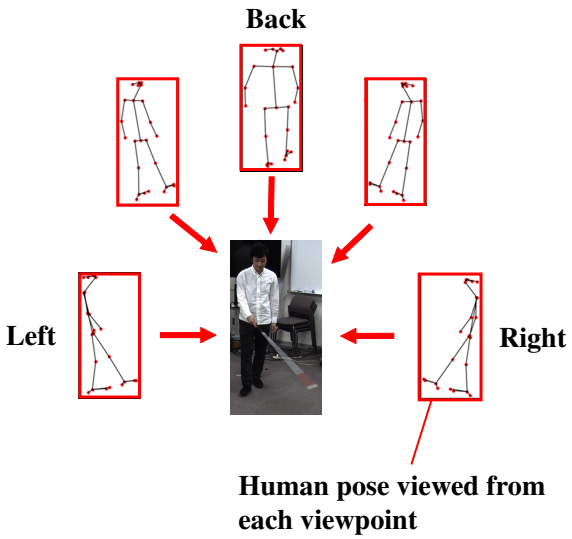

Figure 2. Pose difference depending on the viewpoint of a pose.

similar to a white-cane, such as a white umbrella.

To address this issue, it would be better to recognize them not only by the existence of a white-cane but also by unique actions when white-cane users search for obstacles. In fact, there are differences between actions of a white-cane user and a sighted person as shown in Figure 1. In this paper, we focus on recognizing a white-cane user from only a pedestrian's actions.

Various studies had been performed to recognize actions from human poses [9, 10, 11]. Since it is difficult to estimate the 3D pose of a human from a monocular camera, a sequence of 2D poses are usually used for estimating his/her actions. However, since the appearance of a 2D pose varies widely according to the viewpoint as shown in Figure 2, the performance of action recognition may be degraded according to the viewpoint.

To tackle this problem, this paper proposes a framework for recognizing a white-cane user based on multiple classifiers specialized for each viewpoint. In addition, this paper presents the novel idea to obtain poses from multiple viewpoints by rotating the 2D pose representations in each frame. To realize this viewpoint rotation, an exemplar-based approach is employed to rotate the 2D pose representation.

After obtaining pose sequences viewed from multiple viewpoints by viewpoint rotation, we classify whether each pose sequence is a white-cane user or not by classifiers corresponding to each viewpoint. Finally, we integrate the classification results weighted by the accuracy of each classifier to classify whether the pedestrian is a white-cane user or not.

The exemplar-based viewpoint rotation of a pose is realized by preparing sets of 2D pose representation pairs simultaneously observed from several viewpoints and outputting one of the sets whose pose is similar to the input.

Contributions of this paper are summarized as follows:

- We propose a framework for recognizing a white-cane user from a sequence of estimated 2D pose representations. By focusing on the pose transition of a white-cane user, we distinguish them from pedestrians who possess something similar to a white-cane such as a white umbrella.

- We also propose a method for generating a pedestrian's 2D pose representation sequence observed from various pseudo-viewpoints by rotating viewpoints for the 2D poses based on exemplars. By obtaining a 2D pose representation sequence observed from various viewpoints, we can extract pose features richer than that of pose representations viewed from a single viewpoint.

- Through evaluation using images collected in several real environments, we show 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 work. In section 3, we present the proposed method to classify a pedestrian's 2D pose representation sequence by viewpoint rotation of a human 2D pose representation sequence. In section 4, we report our experiments and discuss their results. In section 5, we conclude this paper and discuss the future work of this research.

2. Related work

For human action recognition, it is necessary to extract the temporal transition of human features. To address this, Recurrent Neural Network (RNN) is often used, which aims to recognize continuous sequences such as sentences and videos. In particular, Long Short-Term Memory (LSTM) [7], a type of RNN, could handle long-term sequences, and methods using it have achieved high accuracies in action recognition [8].

In recent years, estimated human poses are often used as a feature to recognize human actions. Convolutional Pose Machine [5] and OpenPose [6] are mentioned as well-known methods for human pose estimation. These methods estimate the 2D coordinate values of each joint that composes the human body by the Convolutional Neural Network (CNN).

It is desirable that a human pose is represented in 3D rather than in 2D, since a 2D pose greatly differs depending on the viewpoint. Therefore, some researchers have used a 3D pose for action recognition [9, 10, 11]. However, in these researches, 3D poses were prepared in advance or estimated by images captured from multiple cameras. In a real scene, since it is difficult to install multiple cameras for capturing people simultaneously everywhere, it is better than a human pose is estimated from a single image.

3. White-cane users recognition via viewpoint rotation of a 2D human pose

When recognizing a white-cane user from poses, there is a problem that the appearance of a pose greatly varies depending on the viewpoint. To cope with this problem, we propose an exemplar-based viewpoint rotation method of a human pose. By rotating the viewpoint of a 2D pose, richer features of a pose viewed from various viewpoints than viewing from one viewpoint would be obtained. When performing pedestrian classification, to extract features of white-cane users from pose sequences, it is necessary to consider not only a single pose but also its temporal transition. We utilize LSTM as a network that can take into account the temporal transition of a pose sequence.

The procedure of the proposed method is shown in Figure 3. First, a pedestrian image sequence is input, and 1) the pose of the pedestrian in each frame is estimated. Then, 2) for each frame, poses from various pseudo-viewpoints of the estimated pose are obtained by exemplar-based pseudo-viewpoint rotation. Finally, 3) each of the pseudo-rotated pose sequences is classified by a classifier corresponding to its viewpoint and the results are integrated to output a final decision for the input sequence. We present the details of each process in the rest of this section.

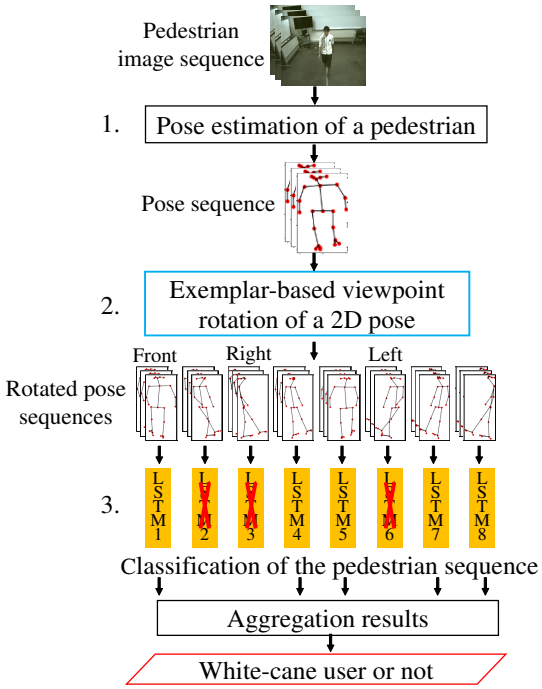

Figure 3. Procedure of the proposed method.

3.1. Pose estimation of a pedestrian

We define a human pose by a set of 2D coordinates of joints such as wrists, elbows, knees, etc. Assuming that the number of joint points is J, a 2D pose which is viewed from a certain viewpoint can be represented by $\mathbf{p} \in \mathbb{R}^{2J}$. Here, a 2D pose 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 sequence $\mathcal{I} = \{\mathbf{I}_1, ..., \mathbf{I}_n, ..., \mathbf{I}_N\}$ consists of N color images whose size are $w \times h$ [pixels] obtained by human tracking.

We use OpenPose [6] for 2D pose estimation. For each frame $\mathbf{I}_n$, the method estimates heat maps indicating probabilities of all joints and part affinity fields indicating the connection between each joint pair. These maps and image features are input and a 2D pose $\mathbf{p}_n$ and its probability o_n are output.

From the estimated 2D pose sequence $\mathcal{P} = \{\mathbf{p}_n\}_{n=1}^N$, we construct a 2D pose sequence $\mathcal{S}$ as the input of the next process (Exemplar-based viewpoint rotation for a 2D pose). Here, we removed frames of low-confident estimations, and constructed $\mathcal{S}$ 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 coordinate values (x_q^j, y_q^j) of each joint, their value range is

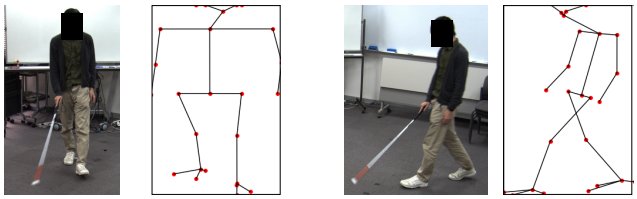

Figure 4. Examples of estimated 2D poses.

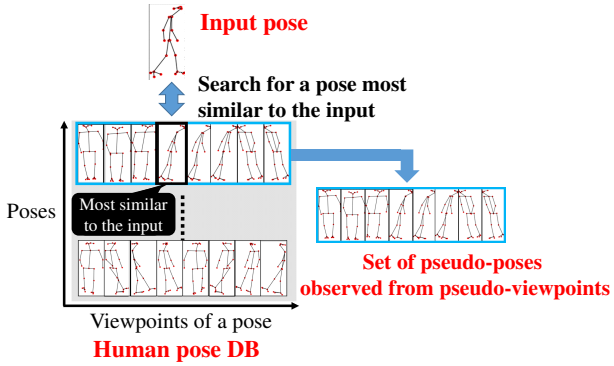

Figure 5. Viewpoint rotation of a 2D pose.

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 constants to adjust the width and height of a pose. Examples of the estimated 2D poses are shown in Figure 4.

3.2. Exemplar-based pseudo-viewpoint rotation of a 2D pose

By rotating the pseudo-viewpoint of a 2D pose, a 2D human pose sequence $\mathcal{S}$ is transformed to a set of 2D pose sequences $\{\tilde{\mathcal{S}}_d\}_{d=1}^D$. Here, $\tilde{\mathcal{S}}_d$ is defined as:

$$\tilde{\mathcal{S}}_d = T_d(\mathcal{S}), \tag{4}$$

where $T_d(\mathcal{S})$ is the function to rotate poses in $\mathcal{S}$ to ones viewed from a d-th pseudo-viewpoint. The procedure of pseudo-viewpoint rotation for a 2D human pose is shown in Figure 5.

The rotated pose can be estimated by multivariate regression. However, it is difficult to constrain the output to a realistic pose. Therefore, we propose to rotate the viewpoint in an exemplar-based approach, which outputs 2D poses in the pose database prepared in advance.

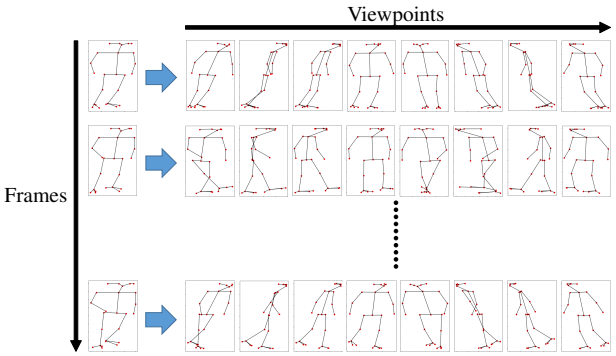

Figure 6. Examples of rotated 2D pseudo-poses.

Therefore, the proposed method consists of the following two stages; 1) construct a database of 2D poses (pose DB), and 2) rotate the viewpoint referring to the pose DB. Here, $\mathcal{B}$ is the pose DB tht consists of 2D pose sets $\mathcal{B}_m$, defined as:

$$\mathcal{B} = \{\mathcal{B}_1, ..., \mathcal{B}_m, ..., \mathcal{B}_M\}, \tag{5}$$

$$\mathcal{B}_m = \{\mathbf{b}_{1m}, ..., \mathbf{b}_{dm}, ..., \mathbf{b}_{Dm}\}, \tag{6}$$

where $\mathbf{b}_{dm}$ is one of the 2D pseudo-poses generated by viewing a 3D pose from D pseudo-viewpoints. The 3D pose composed of the 3D coordinates of each human joint is estimated using multiple calibrated cameras. By changing human poses, we construct various 3D human poses and generate various 2D poses from pseudo-viewpoints.

In the pose rotation stage, given an input 2D pose $\mathbf{p}_q$, a 2D pose $\mathbf{b}_{d'}^q$ which is the most similar to the input 2D pose $\mathbf{p}_q$ is searched. By using the Euclidean distance as the metric of measuring the similarity of 2D poses, $\mathbf{b}_{d'}^q$ can be obtained by the following equation:

$$\mathbf{b}_{d'}^q = \underset{\mathbf{b}_{dm} \in \mathcal{B}_m, \mathcal{B}_m \in \mathcal{B}}{\arg\min} (||\mathbf{p}_q - \mathbf{b}_{dm}||_2^2). \tag{7}$$

Finally, $\mathcal{B}^q = \{\mathbf{b}_1^q, ..., \mathbf{b}_{d'}^q, ..., \mathbf{b}_d^q, ..., \mathbf{b}_D^q\} \in \mathcal{B}$, which is the 2D pose set containing $\mathbf{b}_{d'}^q$, is retrieved as 2D pseudo-poses viewed from each rotated viewpoint. Therefore, by repeating this process along the input pose sequence $\mathcal{S}$, the 2D rotated pose sequence $\tilde{S}_d$ for a pseudo-viewpoint d composed of $\mathbf{b}_d^q \in \mathcal{B}^q$ is obtained as follows.

$$\tilde{S}_d = \{\mathbf{b}_d^1, ..., \mathbf{b}_d^q, ..., \mathbf{b}_d^Q\}. \tag{8}$$

An example of rotated 2D pseudo-poses obtained by the proposed pseudo-viewpoint rotation of an input 2D pose is shown in Figure 6.

3.3. Classification of a pedestrian sequence

Each 2D pose sequence generated by pseudo-viewpoint rotation is classified whether it is a white-cane user or not

by the classifier corresponding to each viewpoint. We prepare D independent classifiers $\mathcal{C} = \{C_1, ..., C_d, ..., C_D\}$ for D different viewpoints, of which, each of them is a neural network. The network structure of the classifier consists of three fully connected layers, one LSTM layer, and three fully connected layers arranged in this order. Each of the classifiers is trained with the rotated 2D pose sequences observed from the corresponding pseudo-viewpoint.

Each classifier outputs a classification score for the input. Then, the classification scores are integrated into the final result. The integration is performed by a weighted sum of the scores. In this paper, the weight for each classifier is either 0 or 1, which is determined by the accuracy of the training data. Given a set of classifiers $\mathcal{C}$, a subset of $\mathcal{C}$ is selected as $\mathcal{C}'$ as follows:

$$\mathcal{C}' = \{C_d | a(C_d) > \delta, \forall C_d \in \mathcal{C}\}, \tag{9}$$

where $a(C_d)$ is the accuracy of the classifier C_d for all training image sequences, and δ is a threshold. Finally, the classification results are integrated by summing the classification scores of each classifier in $\mathcal{C}'$, and the label with the highest score is output as the classification results.

4. Evaluation

In this section, we describe the experiment to confirm the effectiveness of the proposed method and discuss the results.

4.1. Pose DB for viewpoint rotation

For constructing the pose DB, we capture a pedestrian from three calibrated cameras as shown in Figure 7. We capture the data at one specific location, and both the sighted pedestrian role and the white-cane user role are acted by a single sighted person. Then, we estimate their 3D poses by OpenPose. In total, we obtained $M\ (= 4,616)$ 3D poses. These 3D poses are virtually observed from various viewpoints and 2D poses are created by projecting each of the 3D poses onto 2D planes in various directions. The viewpoint, which observes a pedestrian from the front, is labeled as $0°$, and a pseudo-viewpoints are set by rotating the viewpoint counterclockwise in $10°$ steps around the vertical axis. As a result, for a 3D pose, $D\ (= 36)$ sets of 2D poses virtually observed from pseudo-viewpoints are obtained. In the end, the pose DB is composed of $MD\ (= 166,176)$ 2D poses.

4.2. Dataset for evaluation

As training and testing data, we construct a dataset by capturing videos of several walking white-cane users and sighted pedestrians. 17 sighted participants played both roles and five visually impaired people also participated as white-cane users. The videos were captured at five different locations including both indoors and outdoors. We composed pedestrian sequences by selecting frames where pedestrians exist, resulting in 266 pedestrian sequences. The details of the dataset are summarized in Table 1 and examples of the images at each location are shown in Figure 8.

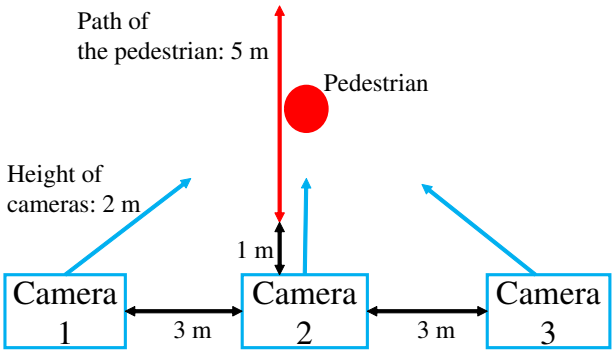

Figure 7. Positions of cameras for constructing the pose DB.

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
Sighted pedestrian without a white-cane	26	6	25	0	76	133
All sequences	49	18	37	10	152	266

4.3. Experimental settings

In the experiment, the classification accuracy for pedestrian sequences is compared by changing the conditions such as the number of viewpoints for the pose, with or without viewpoint rotation and weighting for classifier integration.

In the evaluation, the length of each 2D pose sequence was set to $Q = 64$ frames, and each sequence is divided into overlapping 5 sequences composed of 32 frames for data augmentation with stride 8. The total number of pose sequences (before viewpoint rotation), is $266 \times 5 = 1,330$. The number of estimated joints of each 2D pose is $J = 25$, the value range of coordinate values of joints is [0.0, 1.0] in the horizontal direction, and [0.0, 1.5] in the vertical direction with $\theta_x = 1$ and $\theta_y = 1.5$. For the evaluation, five-fold cross-validation was performed with the data taken at each location among the five locations as evaluation and

Figure 8. Examples of an image at each location.

the other as training.

We compare the accuracy of the following five methods:

- No rotation of the viewpoint for an input pose and use one classifier (No rotation).
- Rotate the viewpoint only to 0° (front) and use one classifier corresponding to the 0° viewpoint (0° viewpoint).
- Rotate the viewpoint only to 90° (right) and use one classifier corresponding to the 90° viewpoint (90° viewpoint).
- Rotate the viewpoint to all 36 viewpoints and integrating all results without weighting (All viewpoints, without weighting).
- Rotate the viewpoint to all 36 viewpoints and integrate all results with weighting (Proposed method).

4.4. Results

The results are summarized in Table 2. Here, "All" in the table indicates the average of all location's results weighted by the amount of data shown in Table 1. The accuracy of the proposed method achieved the highest in the overall results and is improved by 0.12 compared to the method without viewpoint rotation. Moreover, the accuracy of the proposed method is improved compared with also rotating to a specific viewpoint and without weighting for each classifier. As a result, the effectiveness of the proposed method was confirmed.

4.5. Discussion

Here, we discuss the experimental results. We focused on three points as follows: (1) the locations where the data were captured, (2) the effects of pseudo-viewpoint rotation, (3) the weighting of classifiers.

Table 2. Classification results.

Location	1	2	3	4	5	All
No rotation	**0.82**	0.70	**0.55**	0.53	0.64	0.66
0° viewpoint	0.49	0.60	**0.71**	**0.84**	0.50	0.55
90° viewpoint	0.80	0.66	0.70	**0.55**	0.71	0.71
All viewpoints without weighting	0.77	0.70	**0.76**	0.52	0.80	0.77
Proposed method	0.80	0.69	0.75	0.50	**0.81**	**0.78**

First, we discuss the difference of results by locations where the data were captured. As shown in Table 2, among all the methods, the accuracy at location 4 was relatively low. There are two possible reasons for this. One is that different from other locations, it contains only white-cane users and no sighted pedestrians. The other is that when capturing at location 4, the camera position was higher than that at the other locations, and thus the tilt angle was different from the others. Therefore, we considered that since pose patterns were different from those at other locations, the classification accuracy was degraded. To address this problem, we should capture data in many patterns of camera position, capturing location, subjects, and so on.

Second, we discuss the effects of the pseudo-viewpoint rotation. When rotating the viewpoint to the specific viewpoint 90°, the accuracy improved compared to that without viewpoint rotation. This is considered to be the result of simplifying the classification by unifying the viewpoint into one and suppressing differences between features of poses depending on the viewpoint. However, when rotating to the specific viewpoint 0°, the accuracy at locations 1 and 5 greatly decreased, and as a result, the overall accuracy also decreased. A common characteristic of the data at these two locations is that they include many pedestrian images facing to the left or the right. From these results, we considered that the accuracy decreased by rotating the viewpoint from the left and the right to the front. On the other hand, the accuracy for locations 3 and 4 improved compared with rotating to the viewpoint 90°. From this, we can see that effective viewpoints for classification vary depending on pose characteristics.

Finally, we discuss the weighting to classifiers. First, let's see the changes in the number of classifiers used for evaluation (weighted by 1) and the accuracy due to the change of the threshold value δ which determines the weights of classifiers, in Figure 9. The accuracy improved as the number of classifiers used for evaluation decreased, where the highest accuracy was obtained when 14 classifiers were used and $\delta = 0.965$. However, the accuracy rapidly decreased by less than ten classifiers. From this, we can see that in order to maintain high accuracy, it is necessary to prepare a certain number of classifiers corresponding to the viewpoints. Classifiers that were not used at the eval-

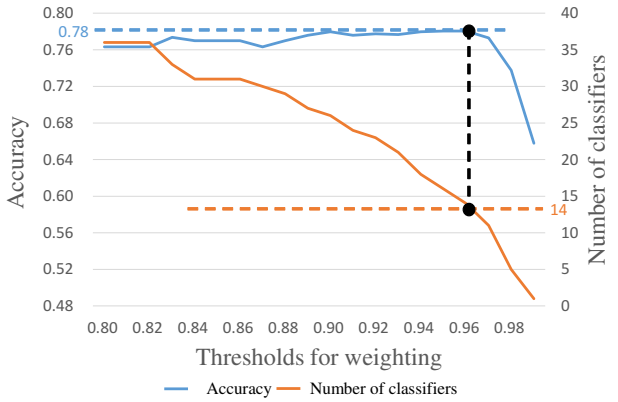

Figure 9. Changes of the number of classifiers and the accuracy.

uation mainly corresponded to viewpoints from the front and the back. The reason is, as shown in the experimental results, the classification using the viewpoints from the front have low accuracy. However, as described in the previous discussion, there are also scenes where front-facing viewpoints are effective, but the current weighting for the classifiers cannot consider the effectiveness of such specific data. Therefore, it is necessary to consider not only the accuracy for the whole data but also the accuracy for specific data for the weighting, for example, pedestrians facing originally backward, etc.

5. Conclusion

In this paper, we proposed a method to recognize white-cane users by classifying pedestrians from the temporal transition of their poses. In the proposed method, we tried to realize an accurate classification based on 2D poses of pseudo-viewpoints which are generated by various rotating viewpoints of a 2D pose. For classification, we prepared classifiers corresponding to each viewpoint and integrated all the results. In addition, we weighted each classifier output based on the training accuracy for a more accurate classification. Through an experiment, the classification accuracy of a pedestrian image sequence was improved by incorporating the viewpoint rotation of an input 2D pose and the effectiveness of the proposed method was confirmed.

As future work, we will consider viewpoint rotation methods other than the exemplar-based method. We will also integrate object recognition methods that directly detect a white-cane.

References

[1] U. Tanikawa, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, and R. Kawai, “Wheelchair-user detection combined with parts-based tracking”, Proc. 12th Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications, vol.5, pp.165–172, Feb. 2017.

[2] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, real-time object detection”, Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.779–788, June 2016.

[3] T. M. Le, N. Inoue, and K. Shinoda, “A fine-to-coarse convolutional neural network for 3D human action recognition”, Proc. 29th British Machine Vision Conf., pp184-1–184-13, Sept. 2018.

[4] M. Liu and J. Yuan, “Recognizing human actions as the evolution of pose estimation maps”, Proc. 2018 IEEE Conf. on Computer Vision and Pattern Recognition, pp.1159–1168, June 2018.

[5] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, “Convolutional pose machines”, Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp. 4724–4732, June 2016.

[6] Z. Cao, T. Simon, S.E. Wei, and Y. Sheikh, “Real-time multi-person 2D pose estimation using part affinity field”, Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition, pp.7291–7299, July 2017.

[7] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, Neural Computation, vol.9, no.8, pp.1735–1780, Nov. 1997.

[8] L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi, and S. Savarese, “Lattice long short-term memory for human action recognition”, Proc. 2017 IEEE Int. Conf. on Computer Vision, pp.2147–2156, Oct. 2017.

[9] T. M. Le, N. Inoue, and K. Shinoda, “A fine-to-coarse convolutional neuralnetwork for 3D human action recognition”, Proc. 29th British Machine Vision Conf., pp.184-1–184-13, Sept. 2018.

[10] R. Baptista, E. Ghorbel, K. Papadopoulos, G. G. Demisse, D. Aouada, and B. Ottersten, “View-invariant action recognition from RGB data via 3D pose estimation”, Proc. 2019 IEEE International Conf. on Acoustics, Speech and Signal Processing, pp.2542–2546, May 2019.

[11] C. Wang, Y. Wang, and A. L. Yuille, “Mining 3D key-pose-motifs for action recognition”, Proc. 2016 IEEE Conf. on Computer Vision and Pattern Recognition, pp.2639–2647, June 2016.