Estimating Person Positions using a Camera and Wireless Devices in a Space with Temporary Shielding

Yuji Makimura[1], Daichi Hayashi[1], Teppei Kobiki[1], Shouki Sakemoto[1] and Masashi Nishiyama[1]
[1]*Graduate School of Engineering, Tottori University, Japan*
d22t5005b@edu.tottori-u.ac.jp

Keywords: Person, Position, Video sequence, Radio wave strength, Temporary shielding

Abstract: Camera sensors can generally estimate a person's position accurately. Here, we consider the case where the camera cannot estimate a person's position when its field of view is unavoidably blocked by temporary shielding. In recent years, several person position estimation methods have been proposed based on use of wireless devices that emit radio waves that can penetrate the shielding. However, to estimate a person's position accurately using wireless devices, we must also consider how we collect the training samples, which consist of pairs of the radio wave strength and the person's position data, because the radio signals acquired from wireless devices cannot be annotated intuitively by humans in the same way as camera images. Our method collects training samples automatically for a regression model that estimates a person's position when there is no shielding via a collaboration between the camera and the wireless devices. Then, our method can estimate person positions using the model, even when temporary shielding occurs. The experimental results obtained when temporary shielding occurred show that the person position estimation error was 15.9 cm when one person was walking in a poster space and 18.7 cm when two people were walking in the space simultaneously.

1 INTRODUCTION

There is a need for technology that can estimate the position of a person in a public space using various sensors. Cameras that can image large spaces with high spatial resolution are commonly used as sensors to estimate an individual person's position. Object detection methods based on deep learning for use with these camera images, e.g., those presented in (He et al., 2017; Chen et al., 2018; Wang et al., 2023a; Ren et al., 2015), have been used widely in recent years. The development of these methods has made it possible to estimate a person's position easily and accurately from camera images.

It is common practice to locate the camera such that the person is not hidden by shielding within the camera's field of view. However, the camera's field of view may sometimes be blocked unavoidably by objects that have been placed temporarily. For example, in poster spaces at conferences and exhibitions, the camera view may be obstructed temporarily when the organizers set up panels to add information boards quickly. Figure 1(a) shows an example of such a case. It is thus essential to consider cases where camera sensors cannot estimate a person's position in such shielded spaces.

In this work, we consider whether there is a sensor configuration that can estimate a person's position stably even when temporary shielding is placed within the camera's field of view. In recent years, research has been actively conducted into use of wireless sensing devices that emit radio waves with longer wavelengths than visible light that can penetrate shielding. For example, methods (Youssef et al., 2007; Booranawong et al., 2019) for person position estimation using the received signal strength indicator (RSSI), which is a time-series signal acquired from wireless devices, has been proposed. More recently, methods (Xue et al., 2021; Wang et al., 2023b) have been proposed to estimate a person's body shape and posture using wireless devices alone. We believe that the results from these existing methods can lead to successful estimation of a person's position using wireless devices, even when temporary shielding is place between the transmitter and receiver.

When using wireless devices as sensors, it is necessary to consider how to collect sufficiently large numbers of training samples because the existing methods are based on machine learning and deep learning techniques. For example, one existing

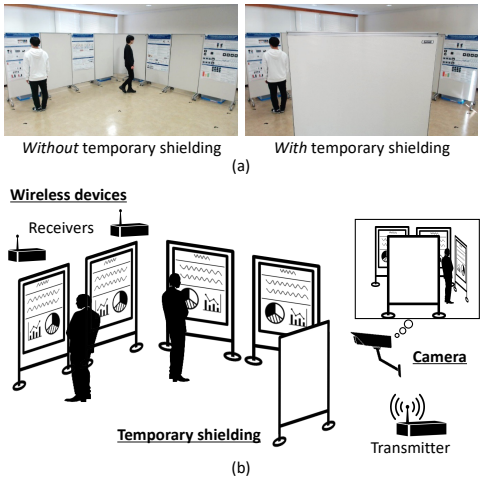

Figure 1: We develop a method to estimate a person's position, even when they are temporarily shielded within the camera's field of view. In the right image in (a), the camera cannot estimate one person's position because of the temporary shielding. As illustrated in (b), our method estimates the person's position via a collaboration between the camera and wireless devices.

method (Youssef et al., 2007) for person position estimation requires annotation by providing a teacher signal that indicates where the person is located within the space corresponding to the time-series RSSI signal. However, radio signals such as RSSI that are acquired from wireless devices cannot be understood intuitively by humans because the space in this case is invisible. This makes it difficult to determine how to perform the required annotations for wireless devices. We must therefore consider how the training samples, which consist of pairs of radio wave strength and person position data, are annotated.

In this paper, we propose a method using a camera and wireless devices that can estimate a person's position within a space even when temporary shielding is present. When there is no shielding, our method collects training samples consisting of a pair of RSSI time-series signals and a person's position automatically because the person's position can be estimated accurately using video sequences acquired from the camera. When the camera's field of view is temporarily shielded, our method applies a regression model that has been trained using the training samples to estimate the person's position based on the RSSI time-series signals acquired from the wireless devices alone. We set up a radio transmitter and radio receivers in the space in addition to the camera, as shown in Figure 1(b). The person does not need to be holding a smartphone or any other device that transmits or receives radio waves. We use Wi-Fi, a type of wireless technology that offers the advantage of easy accessibility. In the experiments, we assumed that when the temporary shielding is inserted, a person's position can be acquired at the edge of the camera's field of view, but that the shielding object is placed near the center of that field of view and that the person's position cannot be acquired using the camera alone. The experimental results show that, based on the assumption above, the error in the person's position when estimated from the RSSI time-series signals was 15.9 cm when one person was walking in the space and 18.7 cm when two people were walking in the space simultaneously.

2 POSTER PANEL AREA (PPA) DATASET

2.1 Setting

We originally collected the poster panel area (PPA) dataset. We considered the space in a poster exhibition room. Visitors to this space may stop in front of a poster of interest and read its contents, or they may simply glance at the poster and leave if they are uninterested. In a typical poster space, each poster has a staff member present, but for the PPA dataset, we assumed a situation in which posters are displayed but no staff member is present, e.g., during breaks.

Figure 2 shows the setup used to construct the PPA dataset. We prepared four posters and set up nine panels on which to hang these posters. In the figure, we hung each poster on one of the panels located at P1, P2, P3, and P4. Each panel was 180 cm high, 90 cm wide, and 4 cm thick. We placed a panel in front of the camera to act as a temporary shield. The camera (C920n, Logicool) was positioned in the upper right corner, as shown in Figure 2, to view the entire space, and its height was set at 180 cm. We placed the Wi-Fi transmitter (WXR-5700AX7S, Buffalo) below the camera at a height of 70 cm. Four Wi-Fi receivers (WI-U2-433DHP, Buffalo) were located 70 cm behind the panels at positions R1, R2, R3, and R4 in the figure. At these receivers, we recorded RSSI time-series signals that had been emitted using beacons at 102 ms intervals from the transmitter. The Wi-Fi frequency band was set to the 5 GHz band of IEEE 802.11n.

To evaluate the accuracy of a person's estimated position, we must give the person's correct position in the presence of temporary shielding. We therefore set up an auxiliary camera to provide the correct position. This auxiliary camera was located such that all persons in the space could be seen and their positions could be estimated, regardless of whether they were

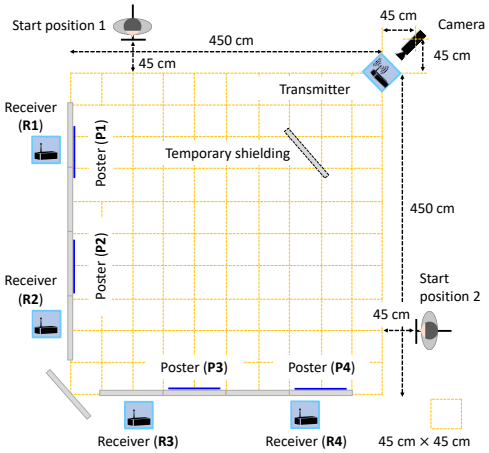

Figure 2: Experimental setting for the PPA dataset.

temporarily shielded. We treated each person's position as acquired by the auxiliary camera only as the correct signals given for the test sample used to evaluate accuracy during the prediction phase, and these signals were not included in the training samples during the training phase.

2.2 Sequences acquired from the wireless devices and the camera

Figure 3 shows examples of sequences that were contained in the PPA dataset. From left to right, these sequences show the image acquired from the camera, a bird's-eye view showing each person's position, and a graph of the Wi-Fi receiver's RSSI time-series signal. The color of each person's bounding box in the image corresponds to the color of that person's position in the overhead view. The bird's-eye view shows each person's position within the floor coordinate system of the poster space. In the RSSI time-series signal graph, the horizontal axis represents time and the vertical axis represents the RSSI values acquired from the Wi-Fi receivers. The RSSI values for each Wi-Fi receiver correspond to receivers R1, R2, R3, and R4 in order from top to bottom.

For the PPA dataset, we collected sequences in which one person was walking and where two people were walking simultaneously within the poster space. When one person was walking, we collected 15 people $\times$ 12 walking patterns $\times$ 2 with/without shielding $=$ 360 sequences. When the two people were walking simultaneously, we collected 9 person pairs $\times$ 12 walking patterns $\times$ 2 with/without shielding $=$ 216 sequences. We prepared two walking paths: one that entered from start position 1 shown in the upper left corner of Figure 2 and exited after looking at posters P1 to P4, and another that entered from start position 2 shown in the lower right corner and exited after looking at posters P4 to P1. We also prepared six different ways of looking at the posters such that the person perused two posters selected at random from among the four available and only glanced at the remaining two posters. When perusing the posters, we asked the participants to read the poster carefully to understand its title and purpose; when glancing at them, we asked the participants to look at the posters in passing after reading their titles. When one person was walking, we used 2 walking paths $\times$ 6 poster view types $=$ 12 walking patterns. When two people were walking simultaneously, there

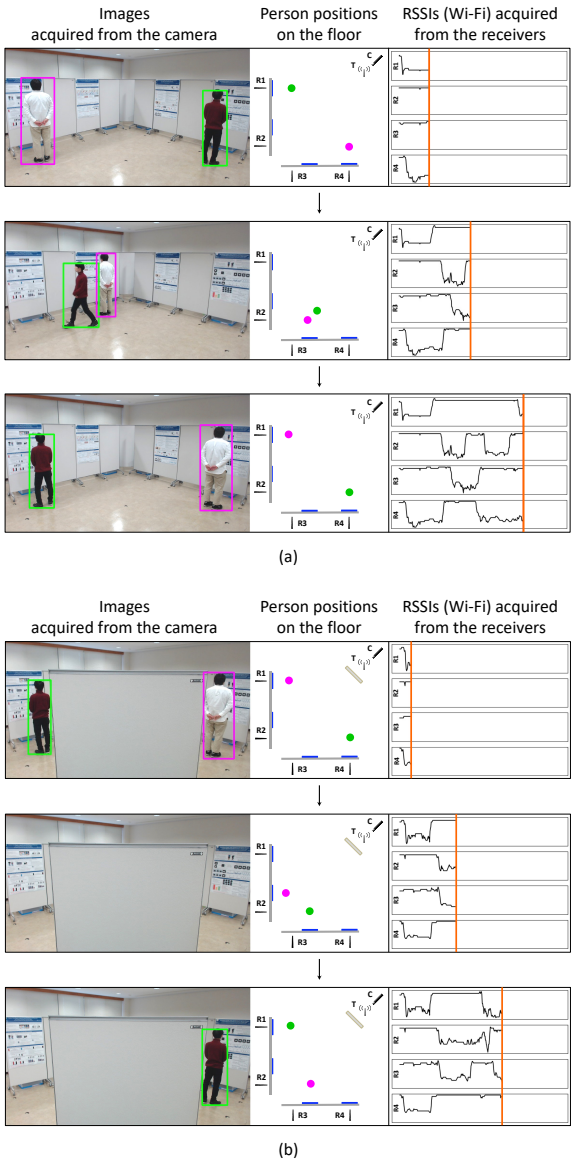

Figure 3: Examples of sequences contained within the PPA dataset.

were $6^2 = 36$ poster view types. We used 12 different walking patterns selected at random from the 36 poster view types.

3 OUR METHOD FOR PERSON POSITION ESTIMATION

3.1 Overview

As described in Section 1, our method estimates each person's position using a regression model that only uses the RSSI time-series signal acquired from the wireless devices during temporary shielding. This regression model requires annotation of training samples that comprise pairs of an RSSI time-series signal and a person's position. Specifically, when no shielding is present, we use the camera to acquire the person's position to act as a teacher signal and use the wireless devices to acquire the RSSI time-series signal. This allows automatic annotation of the training sample collection of RSSI time-series signals when it is difficult for humans to provide a visual teacher signal for the person's position.

In the presence of shielding, our method preprocesses the RSSI time-series signals and then applies the regression model using a general machine learning technique. This enables estimation of a person's position using the RSSI time-series signal acquired from the wireless devices, even if the person's position cannot be acquired from the camera because of unavoidable temporary shielding.

3.2 Procedure for the proposed method

When no shielding is present, our method collects training samples composed of pairs of person positions and RSSI time-series signals in the training phase according to the procedure shown in the upper part of Figure 4. RSSI time-series signals have lower spatial resolution than camera images, which makes it difficult to acquire dense changes caused by a person's presence or absence. Our method aims to improve the accuracy of person position estimation by sampling a short time-series signal from the RSSI time-series signal over the period from current time point t to a past time point $t-T$. We refer to these short time-series signals as training samples. Let $r_{1,t}$ to $r_{n,t}$ denote the RSSI values acquired by the N Wi-Fi receivers R1 to Rn at a time point t, respectively. The short RSSI time-series signal $\mathbf{R}_t$ used in the training

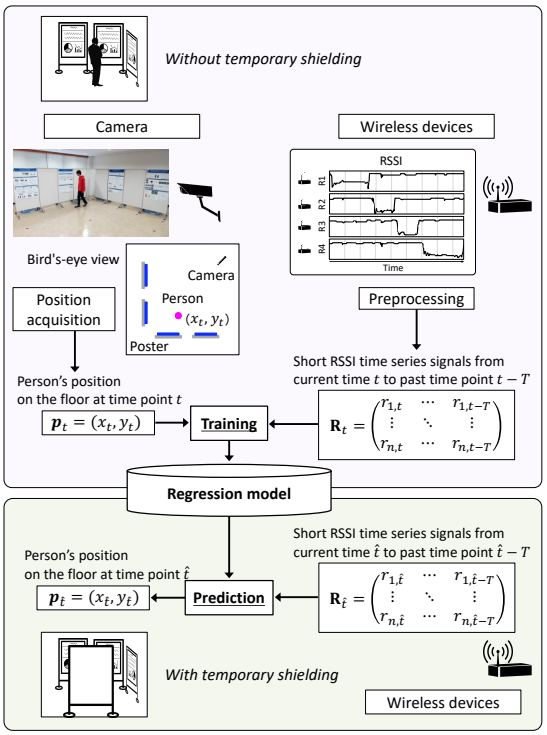

Figure 4: Overview of our person position estimation method.

phase is expressed as:

$$\mathbf{R}_t = \begin{pmatrix} r_{1,t} & \cdots & r_{1,t-T} \\ \vdots & \ddots & \vdots \\ r_{n,t} & \cdots & r_{n,t-T} \end{pmatrix}. \quad (1)$$

Let the $\boldsymbol{p}_t = (x_t, y_t)$ in the floor coordinate system denote the person's position as acquired from the camera image. We represent one training sample as $< \mathbf{R}_t, \boldsymbol{p}_t >$. Because the RSSI time-series signals contain noise, e.g., differences among the characteristics of the different receivers, we perform preprocessing to remove this noise. After collecting several training samples, our method trained a machine-learning regression model. During the training phase, a matrix $\mathbf{R}_t$ of size $N \times (T+1)$ is transformed into a $N \cdot (T+1)$-dimensional vector. We describe the process of estimating a person's position in detail in Section 3.3, and the process of the preprocessing in Section 3.4.

Next, when temporary shielding occurs, our method estimates the person's position using the RSSI time-series signals alone via the procedure shown in the lower part of Figure 4. At time point $\hat{t}$, the RSSI short time-series signal $\mathbf{R}_{\hat{t}}$ is input into the regression model at the prediction phase, and is represented as

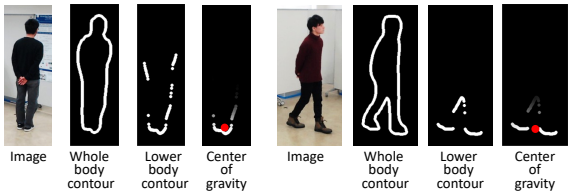

Figure 5: Examples of center of gravity estimation for the feet.

follows:

$$\mathbf{R}_{\hat{t}} = \begin{pmatrix} r_{1,\hat{t}} & \cdots & r_{1,\hat{t}-T} \\ \vdots & \ddots & \vdots \\ r_{n,\hat{t}} & \cdots & r_{n,\hat{t}-T} \end{pmatrix}. \quad (2)$$

During the prediction phase, our method also preprocesses the short time-series signal $\mathbf{R}_{\hat{t}}$. We represent the person's position output by the regression model as $\boldsymbol{p}_{\hat{t}} = (x_{\hat{t}}, y_{\hat{t}})$. We describe the regression model in Section 4.3.

The following provides an explanation of the case in which two or more people are walking within the poster space simultaneously. Our method adds each person's position to the elements of the person position $\boldsymbol{p}_t$ at a time point t. For example, $\boldsymbol{p}_t = (x1_t, y1_t, x2_t, y2_t)$ when two people are walking simultaneously within the space. When using this $\boldsymbol{p}_t$, it is necessary to switch the regression model based on the number of people present in the shielding scenario. As described in Section 1, our method assumes that the shielding occurs near the center of the camera's field of view. Our method counts the number of people present before they enter the temporary shielding area using the camera image.

3.3 Acquisition of person positions for training samples

Our method uses the camera to obtain the person's position when no temporary shielding is present to collect training samples. We estimate a person's region mask by applying a region-based convolutional neural network, Mask R-CNN (He et al., 2017), to an image acquired from a camera. Our method obtains the person's position within the floor coordinate system from the person region mask by estimating the center of gravity at their feet using the technique from (Ganaha et al., 2024). Specifically, our method obtains the contours of the person's entire body from their region mask and then calculates the center of gravity via weighted averaging using only the contours from the lower part of the region corresponding to their legs. Figure 5 shows examples of this center of gravity. We determine the person's position $\boldsymbol{p}_t$ at a time point t within the floor coordinate system by performing a homography transformation of the center of gravity of their feet within the image coordinate system.

The person's estimated position may include noise caused by opening and closing of their feet while walking. To remove this noise, we generate a time-series signal of the positions for each person by applying person tracking in the time direction in the camera image. We apply a weighted moving-average filter to the time-series signal. Because one person may hide part of another person, we use the color histogram calculated from each person's region to identify the person's region across the various time points and thus perform the tracking process.

3.4 Preprocessing of the RSSI time-series signal

This section describes the preprocessing of the RSSI time-series signal acquired from the Wi-Fi receivers. In the RSSI values $r_{1,t}$ to $r_{n,t}$ obtained from N receivers at a time point t, the values often differ from receiver to receiver, even when a person is not present, because of differences in receiver placement and the spatial characteristics of the transmission path. To align the RSSI values for the person's absence among the receivers, our method finds the offset for the person's absence for each receiver and then subtracts the offset for each receiver from all $r_{1,t}$ to $r_{n,t}$ signals. Furthermore, the RSSI values acquired may be significantly attenuated because of sudden noise. To remove these attenuated RSSI values, our method replaces an attenuated value with a constant value when it exceeds a threshold $\tilde{r}$.

4 EXPERIMENTS

4.1 Experimental conditions

To confirm the effectiveness of the proposed method, we evaluated the accuracy with which a person's position was estimated using the PPA dataset described in Section 2. We set $T = 90$ as the past time point to be referenced when generating the short time-series signals $\mathbf{R}_t, \mathbf{R}_{\hat{t}}$, as described in Section 3.2, and the number of Wi-Fi receivers was $N = 4$. In our method, we applied the weighted moving-average filter described in Section 3.3 with 45 samples during the training phase. In the method, $\tilde{r} = -35$ dBm was used as the threshold described in Section 3.4 for preprocessing of the RSSI time-series signal.

4.2 Error calculation

We used the error between the person's predicted position and the given correct position at each time point to evaluate the accuracy of the position estimation. Specifically, we calculated the Euclidean distance between the two positions. The error can be calculated simply when one person is walking within the space. However, when two people are walking simultaneously, the error must be calculated carefully. This is because the predicted person positions $\boldsymbol{p}_{\hat{t}}$ include positions for two people, and the given correct person positions also include positions for two people, and it is thus necessary to determine from which pairs of the predicted and given values the error is to be calculated. We calculated the error by selecting the pair with the smallest error from among all the pairs to ensure that the people do not overlap.

4.3 Regression models

We compared the errors in a person's estimated position using the following regression models, which are based on machine learning techniques.

- **LR**: We used ridge regression, which is a form of linear regression. The coefficient of the regularization term was $\alpha = 1$.
- **PLS**: We used partial least squares regression. The number of components used was 10.
- **MLP**: We used multilayer perceptron regression. Three linear layers and ReLU activation functions were applied. The learning rate was 0.001 and the optimizer was the adaptive moment.
- **RF**: We used random forest regression. The number of decision trees was 100 and the maximum depth of the decision trees was 13.
- **XGB**: We used XGBoost regression (Chen and Guestrin, 2016). The number of decision trees was 100, the maximum depth of the decision trees was 13, and the learning rate was 0.1.
- ***k*NN**: We used *k*-nearest neighbor regression. The number of nearest neighbor samples was $k = 5$. A kd-tree structure was used, with a leaf size of 75. The distance measure was the Manhattan distance.

We applied a leave-one-person-out process when only one person was walking in the space. Specifically, of the 15 people in the PPA dataset described in Section 2, the sequences without temporary shielding of 14 people were used as the training samples, and the sequence with temporary shielding for the remaining person was used as the test sample. This process was repeated for all 15 subjects and the average error was then calculated. Furthermore, we applied a leave-one-pair-out process when two people were walking simultaneously. Specifically, of the nine pairs in the PPA dataset, the sequences without temporary shielding of eight pairs were used as the training samples, and the sequence with temporary shielding of the remaining pair was used as the test sample. This procedure was repeated for all nine pairs and the average error was subsequently calculated.

Table 1: Error (in cm) in Person Position Estimation when One Person was Walking in the Space

Regression	W/O Preprocessing	W/ Preprocessing
LR	27.0 ± 7.8	26.6 ± 7.6
PLS	27.2 ± 7.9	26.8 ± 7.6
MLP	18.5 ± 6.7	16.2 ± 4.7
RF	20.4 ± 10.2	12.4 ± 3.8
XGB	20.5 ± 9.2	13.1 ± 3.8
*k*NN	17.9 ± 5.9	15.9 ± 5.2

Table 2: Error (in cm) in Person Position Estimation when Two People were Walking Simultaneously in the Space

Regression	W/O Preprocessing	W/ Preprocessing
LR	61.7 ± 11.2	64.1 ± 10.1
PLS	61.6 ± 11.0	64.0 ± 10.0
MLP	43.0 ± 11.3	38.1 ± 8.6
RF	48.2 ± 14.5	36.6 ± 11.9
XGB	51.1 ± 18.3	40.5 ± 11.8
*k*NN	22.0 ± 9.6	18.7 ± 6.6

4.4 Results for the position accuracy

Table 1 shows the error in the person position estimation when only one person was walking within the space. The numbers given in the table represent the average and the standard deviation of the error for each regression model. Among the regression models, MLP, RF, XGB, and *k*NN had reduced errors when compared with LR and PLS. The error was improved in each case when our preprocessing step was applied when compared with the case where no preprocessing was performed. The error was 12.4 ± 3.8 cm for our method using RF and preprocessing when only one person was walking within the space. The average foot length ranges were reported to be approximately 25 to 27 cm for men and 22 to 25 cm for women. We believe that the error for our method when using MLP, RF, XGB, or *k*NN is small when compared with the average foot length, which indicates that it is possible to estimate a person's position when temporary shielding occurs.

Next, Table 2 shows the error for person position estimation when two people were walking simultane-

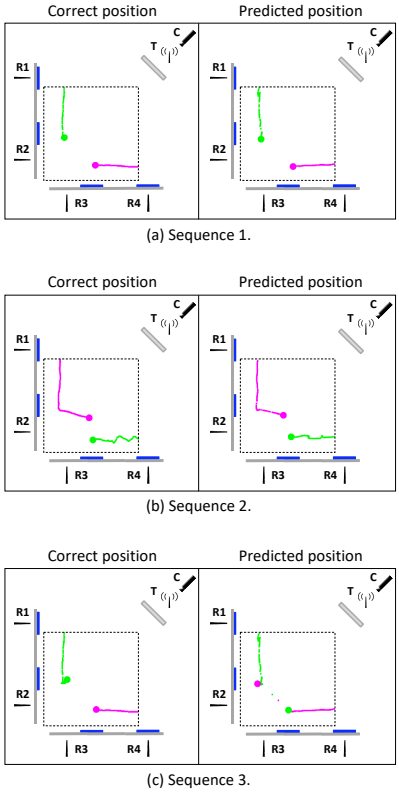

Figure 6: Visualization of the person positions estimated using our method.

ously. For all regression models, the error increased when compared with the case where only one person was walking, as described earlier. However, only *k*NN showed a small increase in the error, whereas LR, PLS, MLP, RF, and XGB all showed significant increases. Similar to the case with one person, the error was improved for MLP, RF, XGB, and *k*NN when preprocessing was applied in the case with two people. The error was 18.7 ± 6.6 cm with our method when we used *k*NN and our preprocessing step when two people were walking simultaneously. We again believe that the error from our method is smaller than the average foot length, thus indicating that it is possible to estimate a person's position in this way.

4.5 VISUALIZATION

Figure 6 shows the visualization results for the estimated person positions in the presence of temporary shielding. The left side of the figure shows the given correct person position and the right side shows the person position when estimated using our method. We plotted each person's position in temporary order on the birds-eye view shown in the center of Figure 3. The circles represent each person's current time position and the thin lines accompanying each circle represent their past time positions. The dotted lines indicate the spaces where the camera cannot acquire images of the person because of the temporary shielding. We also visualized the case where two people were walking simultaneously. The sequences in Figure 6(a) and (b) show that our method can estimate a person's position with reasonable accuracy, even if temporary shielding occurs. However, our method does not track people in the time direction during the prediction phase. For this reason, there were cases where the people's positions were swapped in the time direction, as illustrated in Figure 6(c). In future work, we intend to develop a function that can perform tracking stably for each person when using only the RSSI time-series signal.

5 CONCLUSIONS

We have proposed a method that uses a camera and wireless devices to collect training samples of the RSSI time-series signal and a person's position automatically when no temporary shielding is present and to estimate a person's position accurately using a regression model from the RSSI time-series signals when temporary shielding exists. In future work, we intend to develop a method to estimate a person's position that is independent of the number of people walking within the shielded space simultaneously. We also intend to expand the evaluation when increased numbers of people are present and when the way that these people walk in the space changes. We thank Professor Y. Iwai and Dr. M. Inoue for their valuable advice and suggestions during this research.

REFERENCES

Booranawong, A., Jindapetch, N., and Saito, H. (2019). Adaptive filtering methods for rssi signals in a device-free human detection and tracking system. *IEEE Systems Journal*, 13(3):2998–3009.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of 15th European Conference on Computer Vision, Part VII*, pages 833–851.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Ganaha, W., Ozaki, T., Inoue, M., and Nishiyama, M. (2024). Conversation activity recognition using interaction video sequences in pedestrian groups. In *Pro-*

ceedings of 27th International Conference on Pattern Recognition, pages 359–374.

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

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Advances in Neural Information Processing Systems*, volume 28, pages 91–99.

Wang, C., Bochkovskiy, A., and Liao, H. (2023a). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7464–7475.

Wang, Y., Ren, Y., Chen, Y., and Yang, J. (2023b). Wi-Mesh: A WiFi vision-based approach for 3D human mesh construction. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pages 362–376.

Xue, H., Ju, Y., Miao, C., Wang, Y., Wang, S., Zhang, A., and Su, L. (2021). mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*, pages 269–282.

Youssef, M., Mah, M., and Agrawala, A. (2007). Challenges: device-free passive localization for wireless environments. In *Proceedings of the 13th annual ACM International Conference on Mobile Computing and Networking*, pages 222–229.