IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 13 December 2023, accepted 3 March 2024, date of publication 13 March 2024, date of current version 19 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3376656

RESEARCH ARTICLE

Subjective Baggage-Weight Estimation Based on Human Walking Behavior

MASAYA MIZUNO[1], **TOMOHIRO FUJITA**[2], **YASUTOMO KAWANISHI**[1,2], **(Member, IEEE), DAISUKE DEGUCHI**[1], **(Member, IEEE), AND HIROSHI MURASE**[1], **(Life Fellow, IEEE)**
[1]Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan
[2]Guardian Robot Project, RIKEN Information Research and Development and Strategy Headquarters, RIKEN, Souraku, Kyoto 619-0288, Japan

Corresponding author: Yasutomo Kawanishi (yasutomo.kawanishi@riken.jp)

This work was supported in part by the Japan Society for the Promotion of Science (JSPS) Grant-in-Aid for Scientific Research (KAKENHI) under Grant JP17H00745.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee at Nagoya University.

ABSTRACT We address a new computer vision problem of subjective baggage-weight estimation, where the term *subjective weight* is defined as how heavy the person feels. In this paper, we propose a method named G2SW+ (Gait to Subjective Weight plus), which is an extension of our previous method, G2SW. The method uses human walking behavior, including 3D locations and velocities of body joints and silhouettes, as input. It estimates the subjective weight using a combination of a Convolutional Neural Network and a Graph Convolutional Network. It also estimates human body weight and recognizes the type of baggage as subtasks based on the assumption that body weight and type of baggage affect human gait. For the evaluation, we built a dataset for subjective baggage-weight estimation, consisting of pairs of 3D skeleton and human silhouette sequences with subjective weight, body weight, and baggage-type annotations. We confirmed that the proposed method can accurately estimate the subjective baggage weight. Moreover, we confirmed that training with the subtasks and utilizing the human silhouette sequence as an additional input improves the performance of the subjective weight estimation.

INDEX TERMS Subjective baggage-weight, gait to subjective weight plus (G2SW+), human silhouette image, graph convolution, multi-task learning.

I. INTRODUCTION

In recent years, considerable attention has been paid to the development of assistive robots [1], [2]. To realize this assistance, the system should be environmentally aware and provide proactive support. In this study, we focus on the support provided by a robot to a person carrying heavy baggage. To clarify the problem setting, we assumed a situation in which one person walks with one piece of baggage, as shown in Fig. 1.

To assist people in carrying heavy baggage, it is necessary to develop robots capable of carrying such baggage; however, it is also important to develop a function that determines whether or not to support a person. How heavy a person feels when he/she has a piece of baggage would be valuable information to determine. We named this subjective baggage weight, which quantifies that sensation. The robot can decide whether or not to provide support based on the person's estimated subjective weight. In this research, we focus on the fact that subjective weight changes the human gait. Based on the assumption, we have proposed G2SW (Gait to Subjective baggage-weight), a network to estimate subjective weight by human gait characteristics [3].

However, the G2SW still has two limitations. First, the physique information is not included in the human skeleton sequences. Although this method estimates body weight as well as subjective weight, the input only contains information on the human skeleton. Human skeleton information alone is insufficient for weight estimation because it lacks body size and strength information related to body weight.

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Liu.

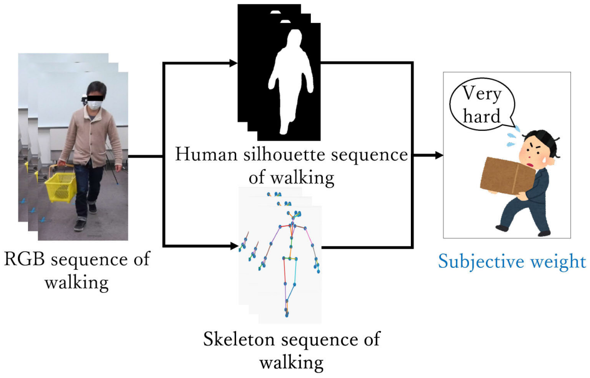

RGB sequence of walking
Human silhouette sequence of walking
Skeleton sequence of walking
Very hard
Subjective weight

FIGURE 1. Estimation of the subjective baggage-weights from human gait.

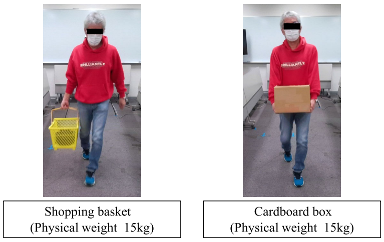

Shopping basket (Physical weight 15kg)
Cardboard box (Physical weight 15kg)

FIGURE 2. Example of the gait with the different types of baggage.

Second, the type of baggage is not considered, and the information was assumed to be given. In practical cases, recognizing the type of baggage is not trivial. As shown in Fig. 2, human gait varies not only in body weight but also in the type of baggage.

To address these limitations, in this paper, we propose G2SW+ (Gait to Subjective Weight plus), which is a method for estimating subjective weight from the human gait (Fig. 1), by further extending the previous G2SW [3].

As noted above, physique information should be considered when estimating subjective weight. For this reason, we also use a human silhouette image sequence for one walking cycle as the input. A human silhouette image has the advantage of representing physique information while eliminating the background, clothing, and other information that is not relevant to subjective weight estimation.

Subjective baggage-weight is also affected by not only the body weight of the person but also the type of baggage. To consider this in the estimation, the proposed method simultaneously recognizes the type of baggage and estimates the body-weight of the person as subtasks in the training phase. Using the subtasks, the network is trained to consider not only body weight but also the type of baggage in the subjective baggage-weight estimation.

Note that this paper is an extended version of our previous paper [3] with further improvements as follows:

- We propose G2SW+, an estimation method for subjective baggage-weight from the human gait. The method improves the previous method, G2SW, by using physique information as additional input, and not only body-weight estimation but also the type of baggage recognition are added as subtasks to focus on the differences in the types of baggage and persons.

In the rest of the paper, Section II presents a literature review. Section III presents the proposed method, which estimates subjective weights. Section IV presents the experimental results and discussions. Finally, Section V summarizes and discusses future issues.

II. LITERATURE REVIEW

A. ESTIMATION OF BAGGAGE WEIGHT

Yamaguchi et al. presented a technique for estimating the weight of a piece of baggage based on body sway [4]. Body sway is the natural movement of a person's body, even when stationary and upright. This approach estimates the weight of baggage by exploiting the characteristic that heavier weight leads to greater body sway. However, this technique is unsuitable for direct use in robotic applications because it requires a bird's-eye view of a stationary person.

Oji et al. presented a weight estimation method from lifting motion [5]. This method estimates the weight of an object from a hand motion by focusing on the fact that the hand motion changes depending on the weight of the object when lifting it up. However, because it requires a specific motion, that is, object lifting, its applicability is limited.

We have proposed a method for estimating subjective baggage weight, named G2SW [3]. This method uses a human skeleton sequence, which represents human gait, as the input. The method extracts the feature using a Graph Convolutional Network from a sequence of one walking cycle. To fix the number of frames while preserving speed information, the method samples a fixed number of frames from a walking cycle and builds a location-and-velocity graph that represents the body joint locations and their moving speeds. The method also considers the person's body weight as a subtask and achieves acceptable accuracy. However, there is room for improvement because of the limitations mentioned in Section I.

B. ACTION RECOGNITION BY A BODY SKELETON SEQUENCE

A network architecture called Long Short-Term Memory (LSTM), which captures temporal information, is often used for action recognition from a skeleton sequence [6], [7], [8], [9]. LSTM models are often used in skeleton sequence recognition because they can effectively capture temporal features. Katoh et al. [10] have proposed a gait style recognition method based on the skeleton sequence. In this study, onomatopoeia were used to describe motion styles. The method employs an LSTM model to estimate onomatopoeia from a skeleton sequence of walking. Nishida et al. [11] have proposed a method that uses LSTMs to recognize whether a person is using a white cane or not from gait. Multiple LSTMs are used to tackle the problem of orientation variations of people.

In recent years, graph convolutional networks (GCN) consisting of graph convolution layers have been widely used in action recognition. It regards the skeleton as a graph. Each body joint and each limb are represented as a vertex and an edge in a graph, respectively. Yan et al. [12] introduced STGCN, a technique for recognizing actions in skeleton sequences by treating the sequence as a graph with temporal connections. In this approach, the spatial characteristics are obtained by employing graph convolutions on individual frames. Subsequently, temporal convolutions are applied to each temporal sequence of a body joint to capture temporal traits. This approach allows the incorporation of skeletal structure and motion information, thereby enhancing the effectiveness of action recognition tasks.

Several methods have been developed based on the foundation of the ST-GCN. Among these, one notable extension is the multi-scale direction represented by MS-G3D [13]. The primary component of this method is the G3D module, which is a variant of the I3D module [14] adapted for graph convolutions. This module implements graph convolutions over a spatio-temporal graph corresponding to a skeleton sequence. The module is further extended using multiple graphs of different multi-hop connections to introduce a multi-scale aspect. Incorporating multi-hop connections facilitates direct linkage between body joints that are skeletally distant from each other but are relevant for recognition tasks. The efficacy of the multi-scale strategy has been demonstrated in diverse studies [15], [16], [17]. Another direction of ST-GCN extension involves the integration of an attention mechanism represented by STA-GCN [18]. This method introduces attention nodes that gauge the significance of body joints on a per-frame basis. Additionally, attention edges are introduced to assess the pairwise significance of joints, thereby capturing the relationships among joints that carry varying degrees of significance for distinct movements.

III. PROPOSED METHOD

A. OVERVIEW

This paper proposes a method for estimating the subjective baggage-weight from the human gait, named G2SW+ (Gait to Subjective Weight). In this study, we basically use a 3D human skeleton sequence to represent human gait. A 3D human skeleton sequence is a set of (X, Y, Z) coordinates of joint locations in the global coordinate system. Here, $(X_t^j, Y_t^j, Z_t^j)^\top$ denotes the location of the j-th body joint in t-th frame. Additionally, noting that walking is a repetition of two steps, we define two steps as one cycle of walking and use the 3D human skeleton sequence $\mathcal{S}_i$ for one cycle of walking as the input. We also utilize a human silhouette sequence, $\mathcal{U}_i = \{\mathbf{u}_i^j\}$, as the input for G2SW+.

Figure 3 shows the flowchart of the training and estimation steps of the proposed method. As a preprocessing, the i-th 3D human skeleton sequence $\mathcal{S}_i$ is converted into a location and velocity graph $\widehat{\mathcal{S}_i}$ as previous G2SW [3], and the i-th human silhouette sequence $\mathcal{U}_i$ is converted into a mean silhouette image $\widehat{\mathbf{u}}_i$. Then the location and velocity graph $\widehat{\mathcal{S}_i}$ and the mean silhouette image $\widehat{\mathbf{u}}_i$ are input to the subjective weight estimator (G2SW+) to estimate the subjective weight. To train the G2SW+, we employ multi-task learning that handles baggage-type recognition and body weight estimation as subtasks.

In the following sections, we first define subjective weights in Section III-B. Then, the preprocessing for the input is explained in Section III-C. The network architecture and multi-task learning are explained in Section III-D.

TABLE 1. New borg scale.

Scale	Description	Scale	Description
10	Very, very Hard	4	Somewhat Hard
9		3	Moderate
8		2	Light
7	Very Hard	1	Very Light
6		0.5	Ver, very Light
5	Hard	0	Nothing at all

B. DEFINITION

In this study, we use the definition from our previous study [3]. The subjective weight is defined as how heavy a person feels. We employ the New Borg Scale [19] to quantify subjective weight. Originally, the New Borg Scale quantifies how hard the activity is, as shown in Table 1. Since we considered that carrying a heavy baggage can be considered as a hard activity, we employ the scale to quantify how heavy a person feels, and the proposed method, G2SW+, estimates the value of the New Borg Scale.

C. PREPROCESSING

In the proposed method, a 3D human skeleton sequence of one walking cycle and a human silhouette sequence of one walking cycle are assumed to be cropped beforehand based on the frame in which the positions of the left and right legs are the most distant. The cropped 3D human skeleton and silhouette sequences are then preprocessed separately.

For each 3D human skeleton sequence, following our previous work [3], we applied the following three preprocessing steps:

1) location and orientation normalization,
2) velocity calculation,
3) frame sampling for a fixed length.

This preprocessing normalizes and enhances the input 3D skeleton sequence. The output of the preprocessing is named *location-and-velocity graph* $\overline{S}_i$. For the details, please refer to the previous paper [3].

A body silhouette sequence is provided for a walking cycle. We compute a mean silhouette image, which can represent important gait features in a single image, known as Motion History Image [20]. The calculation is as follows:

$$\bar{\mathbf{u}}_i = \sum_{\mathbf{u}_i^j \in \mathcal{U}_i} \mathbf{u}_i^j. \tag{1}$$

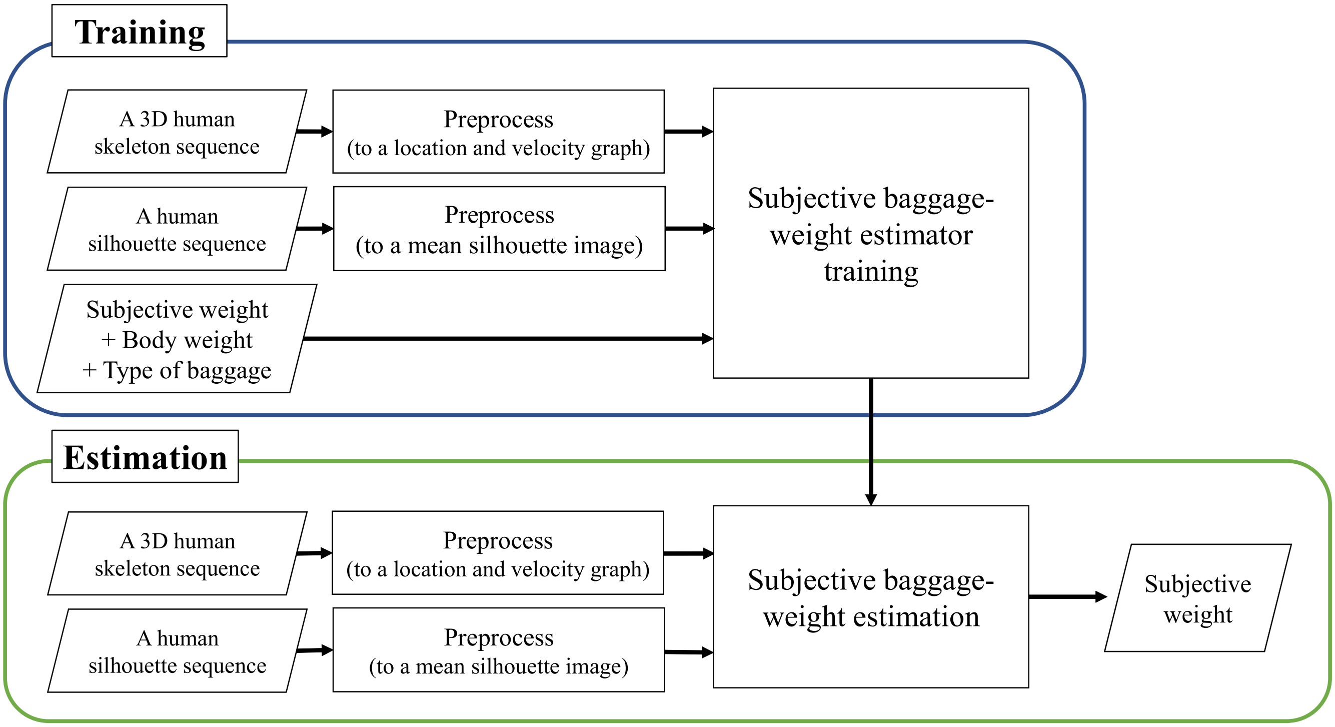

Training
A 3D human skeleton sequence
Preprocess (to a location and velocity graph)
A human silhouette sequence
Preprocess (to a mean silhouette image)
Subjective weight + Body weight + Type of baggage
Subjective baggage-weight estimator training
Estimation
A 3D human skeleton sequence
Preprocess (to a location and velocity graph)
A human silhouette sequence
Preprocess (to a mean silhouette image)
Subjective baggage-weight estimation
Subjective weight

FIGURE 3. The training and estimation processes of the proposed method.

D. THE PROPOSED G2SW+ AND ITS MULTI-TASK TRAINING

In the proposed G2SW+, the subjective weight is estimated from the location and velocity graph $\overline{\mathcal{S}}_i$, and the mean silhouette image $\bar{\mathbf{u}}_i$.

The architecture of the proposed G2SW+ is shown in Fig. 4. In the proposed G2SW+, a middle skeleton feature representation in graph shape $\mathbf{M}_i$ is calculated using a GCN-based feature extractor f_m, and a skeleton feature representation in graph shape $\mathbf{P}_i$ is calculated using a GCN-based feature refinement f_p as

$$\mathbf{M}_i = f_m(\overline{\mathcal{S}}_i;\ \theta_m,\ A), \tag{2}$$

$$\mathbf{P}_i = f_p(\mathbf{M}_i;\ \theta_p,\ A), \tag{3}$$

where A denotes an adjacent matrix that defines the adjacency of human body joints. These functions f_m and f_p consist of multiple graph convolution layers. The proposed method uses consecutive blocks of an MS-G3D module [13] for these functions. Here, the sets of parameters in these networks are represented by θ_m and θ_p. After the MS-G3D blocks, the graph-shaped outputs $\mathbf{M}_i, \mathbf{P}_i$ are reshaped into a 1-dimensional vector $\mathbf{m}_i$, $\mathbf{p}_i$.

A physique feature representation $\mathbf{a}_i$ is also calculated using a CNN-based feature extractor f_a as

$$\mathbf{a}_i = f_a(\widehat{\mathbf{u}}_i; \theta_a), \tag{4}$$

where θ_a is a set of parameters of f_a.

The subjective weight w_i^s is then calculated using the fully-connected layers g_s. At the same time, the body weight w_i^b and type of baggage k_i are also calculated using fully-connected layers g_b and g_k. We use the concatenated feature $(\mathbf{p}_i, \mathbf{a}_i)$ as the input of g_s and g_b, and concatenated feature $(\mathbf{m}_i, \mathbf{a}_i)$ as the input of g_k, respectively.

$$w_i^s = g_s((\mathbf{p}_i, \mathbf{a}_i); \theta_s), \tag{5}$$

$$w_i^b = g_b((\mathbf{p}_i, \mathbf{a}_i); \theta_b), \tag{6}$$

$$k_i = g_k((\mathbf{m}_i, \mathbf{a}_i); \theta_k). \tag{7}$$

Here, because the recognition of the baggage type is quite a different task from the others, it uses a shallower feature than other tasks. These three functions g_s, g_b, and g_k consist of four fully-connected layers, whose parameters are θ_s, θ_b and θ_k, respectively. Leaky ReLU [21] is used as the activation function for the hidden layers. The final layer uses a sigmoid function as the activation function.

Given a batch of $\overline{\mathcal{S}}_i$, $\bar{\mathbf{u}}_i$, and corresponding sets of ground truth of subjective weight, body weight, and type of baggage $(\widehat{w}_i^s, \widehat{w}_i^b, and\widehat{k}_i)$, the network is trained in multi-task learning manner. The parameters θ_m, θ_p, θ_a, θ_s, θ_b, and θ_k are updated using backpropagation to minimize the total loss L consisting of subjective weight loss L_s, body weight loss L_b, and baggage-type loss L_k. L_s and L_b are the mean squared errors, whereas L_k is a categorical cross-entropy loss represented by h.

$$L = \lambda_s L_s + \lambda_b L_b + \lambda_k L_k, \tag{8}$$

$$L_s = \sum_i (w_i^s - \widehat{w}_i^s)^2, \tag{9}$$

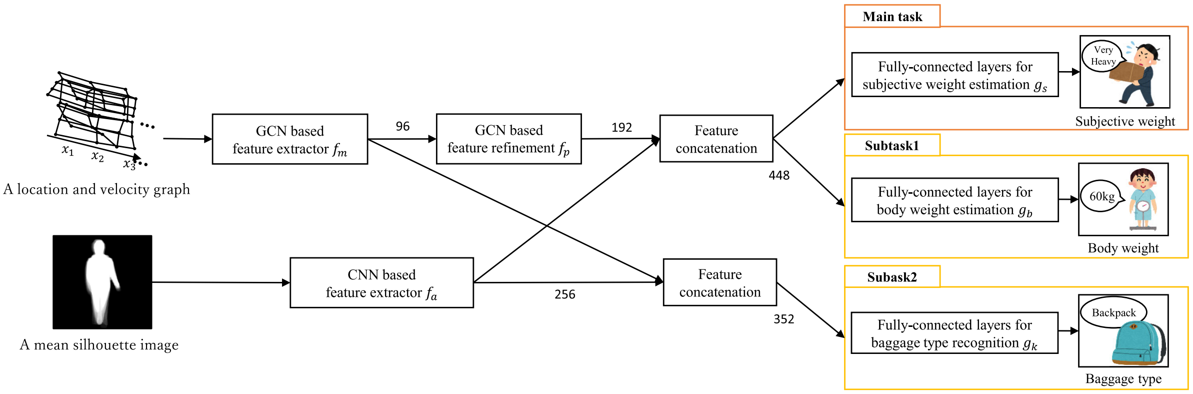

x_1
x_2
x_3
A location and velocity graph
A mean silhouette image
GCN based feature extractor f_m
96
GCN based feature refinement f_p
192
Feature concatenation
448
CNN based feature extractor f_a
256
Feature concatenation
352
Main task
Fully-connected layers for subjective weight estimation g_s
Very Heavy
Subjective weight
Subtask1
Fully-connected layers for body weight estimation g_b
60kg
Body weight
Subask2
Fully-connected layers for baggage type recognition g_k
Backpack
Baggage type

FIGURE 4. **Network architecture of the proposed G2SW+. This network accepts a location and velocity graph and a mean silhouette image, and outputs the subjective baggage weight, body weight, and baggage type. The numbers on the arrows indicate the output vector dimensions.**

$$L_b = \sum_i (w_i^b - \widehat{w}_i^b)^2, \tag{10}$$

$$L_k = \sum_i h(\widehat{k}_i, k_i), \tag{11}$$

where λ_s, λ_b, and λ_k are the weights of the losses. These weights are selected as the best of several combinations. Here, the ranges of the subjective weights are normalized in the range [0, 1]. The range of body weight is also normalized in the range [0, 1], such that the maximum value in the training data is 1 and the minimum value is 0.

IV. EVALUATION

A. DATASET

Because there are no publicly available that consist of 3D skeleton sequences and silhouette sequences with annotations of subjective baggage weights, we originally captured a dataset for evaluation. Note that this dataset is an extension of our previous dataset [3] by adding human silhouette sequences. For the details, please refer to the previous paper. Only the differences are described here.

In this study, we assume a situation in which one person walks with a piece of baggage. The 3D human skeleton sequences were collected by observing each participant walking with a piece of baggage using a Microsoft Azure Kinect sensor installed at a height of 2 m. A skeleton consists of 32 body joints. The frame rate of the captured images is 30 fps. Fig. 5 shows the captured images, 3D human skeletons, and human silhouette images for each type of baggage. Detailed information, including the statistics of the dataset is described in a previous paper [3]. To obtain the silhouette images, we applied Mask R-CNN [22] to every captured image. The silhouette images are resized to 128 × 128 pixels to input to the CNN.

In each session, each participant walked with a piece of the prepared baggage. A short break was inserted after each session to prevent the effects of the previous session. In this experiment, 30 patterns (five baggage types × six weights (0–15 kg)) of 3D human skeleton sequences were captured for each subject.

All the participants consented to using and disclosing their data for research purposes. It should be noted that the Ethics Committee at Nagoya University has approved this experiment.

B. EVALUATION PROTOCOL AND METRICS

In this experiment, following a previous paper [3], we performed 5-fold cross-validation that split five people for evaluation and the remaining 30 people for training from the dataset.

Because of the limited number of preprocessed 3D human skeletons, data augmentation was applied for training. From the input 3D skeleton sequence of one walking cycle, three frames were randomly dropped. We performed this ten times for each walking cycle, thus increasing the data volume to 240,150 walking cycles. In the experiment, the frame length of the location and velocity graph was set to $M = 50$ after data augmentation.

We evaluated G2SW+ performance for subjective weight estimation for each type of baggage. The mean absolute error (MAE) of the estimation results was used as an evaluation metric. It is defined as

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |w_i^s - \widehat{w}_i^s|, \tag{12}$$

where N denotes the number of 3D skeleton sequences.

For practical application of deciding whether to support a person, the performance of estimation within a tolerance error threshold, named Tolerance Accuracy (TA), are also introduced. It is defined as

$$\mathrm{TA}_\tau = 100 \frac{NW_\tau}{N}, \tag{13}$$

where τ is the tolerance error threshold, and NW_τ represents the number of data within the estimation error τ.

TABLE 2. Comparison of subjective weight estimation accuracy between G2SW+ and G2SW.

	MSE↓	TA_1 ↑	TA_2 ↑	TA_3 ↑
G2SW+	**1.41**	**46.8%**	**72.3%**	**87.5%**
G2SW	1.44	46.2%	71.4%	86.9%

TABLE 3. Evaluation scores of G2SW+ for different baggage types (subjective weight: 0–10).

Type of Baggage	MSE↓	TA_1 ↑	TA_2 ↑	TA_3 ↑
Handbag	1.47	44.2%	69.8%	86.4%
Shoulder bag	1.14	53.9%	80.9%	93.8%
Backpack	1.54	44.1%	69.5%	84.4%
Cardboard box	1.56	42.8%	68.6%	85.1%
Shopping basket	1.35	49.0%	72.6%	87.8%
Total	1.41	46.8%	72.3%	87.5%

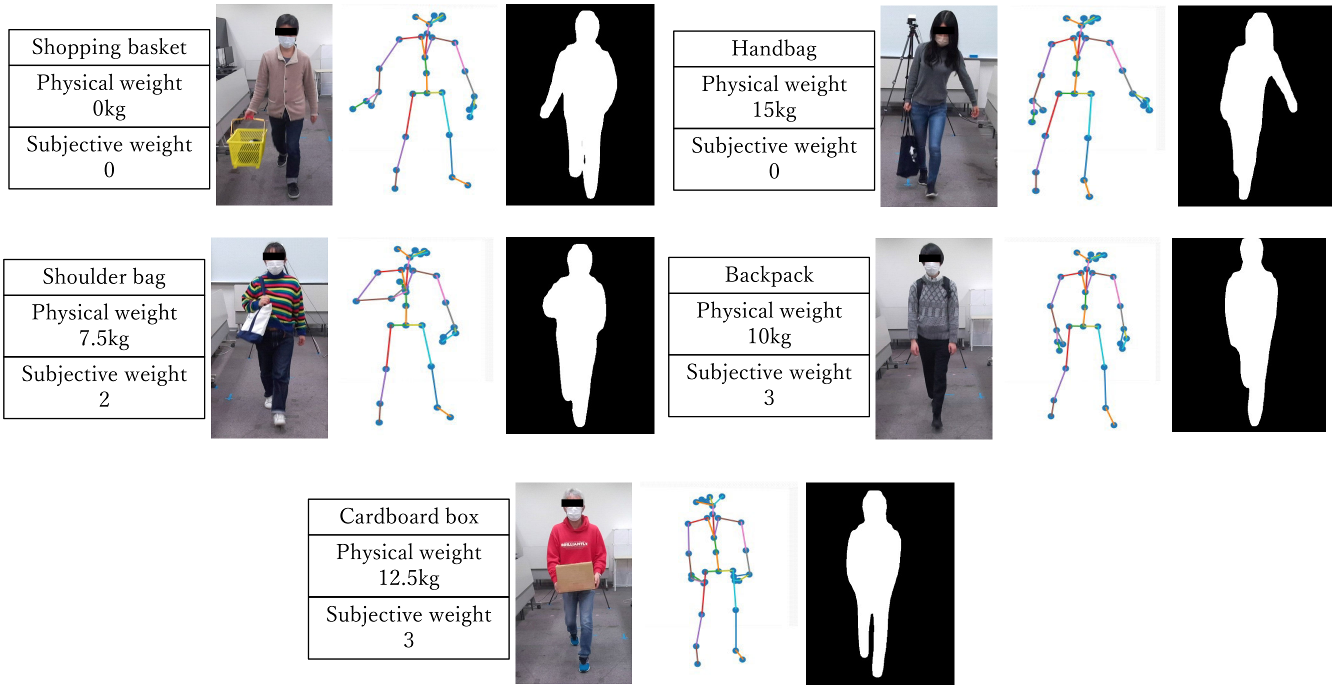


FIGURE 5. Examples of captured images and 3D human skeletons of each type of baggage.

C. MAIN TASK: SUBJECTIVE BAGGAGE-WEIGHT ESTIMATION

Table 3 presents a comparison of the mean absolute errors of the subjective weight estimation and Tolerance Accuracy of $\tau = 1, 2$, and 3 between G2SW+ and G2SW. Figure 6 presents an example of the estimation results. From Table 2, we confirmed that G2SW+ could estimate subjective weights with a mean absolute error of 1.41 in the New Borg Scale as the average of the entire baggage. G2SW+ achieved a Tolerance Accuracy of 46.8% with $\tau = 1(TA_1)$, 72.3% with $\tau = 2(TA_2)$, and 87.5% with $\tau = 3(TA_3)$ as the average of the entire baggage. These accuracies were better than those of G2SW.

D. DISCUSSION

Through the experiments, we confirmed that G2SW+ can estimate the subjective weight better than G2SW.

1) EFFECTIVENESS OF SUBTASKS

In G2SW+, we used baggage-type recognition as a subtask, in addition to weight estimation, which was used in G2SW. To confirm the effectiveness of the subtasks, we compared the proposed method with a method that did not use these subtasks. Table 4 presents a comparison of the subjective weight estimation accuracies with and without each subtask. From the table, it was confirmed that the accuracy of subjective weight estimation was improved by using not only the body weight estimation but also the type of baggage recognition as subtasks.

2) EFFECTIVENESS OF THE VELOCITY FEATURE

In the proposed method, we propose a location and velocity graph to preserve velocity information in a fixed sequence length for one cycle of walking. To confirm that the velocity information is also valid when using the physique feature,

TABLE 4. Comparison of subjective weight estimation accuracy with and without subtasks.

Body weight estimation	Type of baggage recognition	MSE↓	TA_1 ↑	TA_2 ↑	TA_3 ↑
✓	✓	**1.41**	**46.8%**	**72.3%**	**87.5%**
✓		1.43	46.5%	71.5%	86.8%
	✓	1.43	46.2%	71.4%	86.9%
		1.43	46.1%	71.5%	87.2%

TABLE 5. Comparison of subjective weight estimation accuracy with and without velocity information.

	MSE↓	TA_1 ↑	TA_2 ↑	TA_3 ↑
with velocity	**1.41**	**46.8%**	**72.3%**	**87.5%**
without velocity	1.46	45.8%	70.7%	86.1%

TABLE 6. Comparison of subjective weight estimation accuracy with and without appearance information.

	MSE↓	TA_1 ↑	TA_2 ↑	TA_3 ↑
with physique feature	**1.41**	**46.8%**	**72.3%**	**87.5%**
without physique feature	1.46	45.8%	70.7%	86.1%

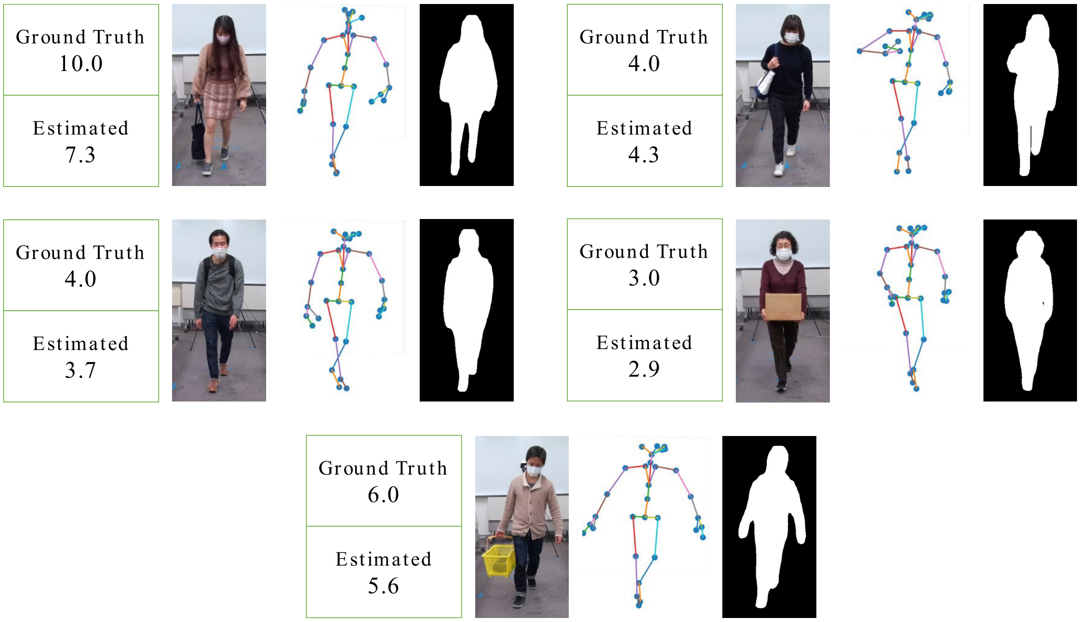


FIGURE 6. Example of estimation results of subjective baggage-weight estimation.

we compared our method with a method that did not use velocity information. Table 5 shows a comparison of the subjective weight estimation accuracies with and without the velocity information. From the table, it was confirmed that the accuracy of subjective weight estimation was improved by using velocity information as additional information when using the physique feature.

3) EFFECTIVENESS OF THE PHYSIQUE FEATURE

In this method, we use physique information, which is the human silhouette sequence, as the input to consider body weight. To confirm the effectiveness of the additional physique features as inputs, we compared the proposed method with a method that does not use physique information. Table 6 presents a comparison of the subjective weight estimation accuracy with and without physique information. The table confirms that the accuracy of the subjective weight estimation was improved by using physique information as additional information.

4) CHALLENGES OF THE PRACTICAL USE

In the proposed method, G2SW+, the estimation is performed on the 3D skeleton and silhouette sequences for one cycle of walking cropped from the sequence during walking.

In contrast, several walking cycles can be obtained from a captured walking sequence in practical applications. Thus, multiple estimation results are obtained for a single sequence during walking. In the future, an integration method for multiple estimation results should be considered.

The 3D human skeleton sequences used in this study were estimated using the Azure Kinect SDK; however, the estimation failed in some cases. This was often the case when the participants wore oversized clothes or black facial masks. In practical applications, a system that discards such estimations should be introduced to maintain accuracy.

A further limitation is that the 3D human skeleton and silhouette may not be accurately estimated in situations where a person is carrying a very large load.

V. CONCLUSION

In this study, we proposed a subjective baggage-weight estimation method named G2SW+, which is an extension of G2SW, from human gait when a person is walking with a piece of baggage. Because subjective weights affect human gait, we employed a 3D human skeleton sequence in the subjective weight estimation method to represent human gait. In addition to the locations and velocities of body joints, we also employed human silhouette sequences, which are expected to represent the physique feature, as an additional input. The estimation method G2SW+ was trained with an additional subtask, baggage-type estimation, while the previous G2SW was trained with a subtask, human body weight estimation.

Future work includes a further update of the gait representation to more effectively describe the motion of skeletons by analyzing the contribution of each body joint. Developing a human support robot that uses this estimation method so that the robot can decide whether or not to assist a person is also future work.

REFERENCES

[1] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of human support robot as the research platform of a domestic mobile manipulator," *ROBOMECH J.*, vol. 6, no. 1, pp. 1–15, Apr. 2019.

[2] A. Yuguchi, S. Kawano, K. Yoshino, C. T. Ishi, Y. Kawanishi, Y. Nakamura, T. Minato, Y. Saito, and M. Minoh, "Butsukusa: A conversational mobile robot describing its own observations and internal states," in *Proc. 17th ACM/IEEE Int. Conf. Human-Robot Interact. (HRI)*, Mar. 2022, pp. 1114–1118.

[3] M. Mizuno, Y. Kawanishi, T. Fujita, D. Deguchi, and H. Murase, "Subjective baggage-weight estimation from gait: Can you estimate how heavy the person feels?" in *Proc. 18th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2023, pp. 567–574.

[4] Y. Yamaguchi, T. Kamitani, M. Nishiyama, Y. Iwai, and D. Kushida, "Extracting features of body sway for baggage weight classification," in *Proc. IEEE 9th Global Conf. Consum. Electron. (GCCE)*, Oct. 2020, pp. 345–348.

[5] T. Oji, Y. Makino, and H. Shinoda, "Weight estimation of lifted object from body motions using neural network," in *Haptics: Science, Technology, and Applications*, vol. 10894. Cham, Switzerland: Springer, Jun. 2018, pp. 3–13.

[6] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Computer Vision—ECCV 2016*. Cham, Switzerland: Springer, 2016, pp. 816–833.

[7] J. Liu, G. Wang, P. Hu, L.-Y. Duan, and A. C. Kot, "Global context-aware attention LSTM networks for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3671–3680.

[8] A. Ullah, J. Ahmad, K. Muhammad, M. Sajjad, and S. W. Baik, "Action recognition in video sequences using deep bi-directional LSTM with CNN features," *IEEE Access*, vol. 6, pp. 1155–1166, 2018.

[9] M. Majd and R. Safabakhsh, "Correlational convolutional LSTM for human action recognition," *Neurocomputing*, vol. 396, pp. 224–229, Jul. 2020.

[10] H. Kato, T. Hirayama, Y. Kawanishi, K. Doman, I. Ide, D. Deguchi, and H. Murase, "Toward describing human gaits by onomatopoeias," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2017, pp. 1573–1580.

[11] N. Nishida, Y. Kawanishi, D. Deguchi, I. Ide, H. Murase, and J. Piao, "SOANets: Encoder–decoder based skeleton orientation alignment network for white cane user recognition from 2D human skeleton sequence," in *Proc. 15th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl.*, 2020, pp. 435–443.

[12] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell. (AAAI)*, 2018, pp. 7444–7452.

[13] Z. Liu, H. Zhang, Z. Chen, Z. Wang, and W. Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 140–149.

[14] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[15] S. Abu-El-Haija, B. Perozzi, A. Kapoor, N. Alipourfard, K. Lerman, H. Harutyunyan, G. V. Steeg, and A. Galstyan, "MixHop: Higher-order graph convolutional architectures via sparsified neighborhood mixing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3595–3603.

[16] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3590–3598.

[17] F. Wu, A. H. Souza Jr., T. Zhang, C. Fifty, T. Yu, and K. Q. Weinberger, "Simplifying graph convolutional networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6861–6871.

[18] K. Shiraki, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Spatial temporal attention graph convolutional networks with mechanics-stream for skeleton-based action recognition," in *Computer Vision—ACCV 2020*, vol. 12626. Cham, Switzerland: Springer, Feb. 2021, pp. 341–357.

[19] G. A. V. Borg, "Psychophysical bases of perceived exertion," *Med. Sci. Sports Exerc.*, vol. 14, no. 5, May 1982, Art. no. 377381.

[20] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[21] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. 30th Int. Conf. Mach. Learn. (ICML)*, Jun. 2013, vol. 30, no. 1, pp. 1–6.

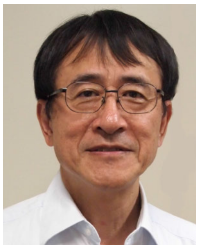
[22] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

MASAYA MIZUNO received the B.Inf. and M.Inf. degrees from Nagoya University, Japan, in 2021 and 2023, respectively. He joined SoftBank Corporation, in 2023. His research interest includes human behavior analysis from an image sequence.

TOMOHIRO FUJITA received the B.Eng., M.Syst.Inf., and Dr.Syst.Inf. degrees from Kobe University, in 2017, 2019, and 2021, respectively. He is currently a Postdoctoral Researcher with the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His research interests include deep learning in artificial intelligence, natural language processing, time series analysis, and human pose prediction.

YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Postdoctoral Fellow with Kyoto University, in 2012. He moved to Nagoya University, Japan, as a Designated Assistant Professor, in 2014, where he became an Assistant Professor and a Lecturer, in 2015 and 2020, respectively. Since 2021, he has been the Team Leader of the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and computer vision for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He received the Best Paper Award from SPC2009 and the Young Researcher Award from IEEE ITS Society Nagoya Chapter. He is a member of IIEEJ and a Senior Member of IEICE.

DAISUKE DEGUCHI (Member, IEEE) received the B.Eng. and M.Eng. degrees in engineering and the Ph.D. degree in information science from Nagoya University, Japan, in 2001, 2003, and 2006, respectively. He became a Postdoctoral Fellow with Nagoya University, in 2006. From 2008 to 2012, he was an Assistant Professor with the Graduate School of Information Science. From 2012 to 2019, he was an Associate Professor with Information Strategy Office. Since 2020, he has been an Associate Professor with the Graduate School of Informatics. His research interests include the object detection, segmentation, recognition from videos, and their applications to ITS technologies, such as detection and recognition of traffic signs. He is a member of IEICE and IPS Japan.



HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York City. He has been a Professor with Nagoya University, since 2003 and an Emeritus Professor, since 2021. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of IAPR, IPSJ, and IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He got a Medal with Purple Ribbon from the Government of Japan, in 2012.