2026 IEEE/SICE International Symposium on System Integration (SII)
January 11-14, 2026. Cancún, México

Action Transition Recognition using ST-GCN for Worker Following in Agricultural Support Robots

Go Oya[1], Akihisa Ohya[2], Takashi Tsubouchi[2], Rui Fukui[1] and Ayanori Yorozu[2]

***Abstract*— In recent years, the increasing labor burden in Japanese agriculture has become a serious issue, driving the development of robots to assist in transporting harvested crops. This study proposes a method that recognizes the action transitions of agricultural workers and enables smooth transport assistance by utilizing three-dimensional skeletal information obtained from RGB-D images. Specifically, we employ Spatial Temporal Graph Convolutional Networks (ST-GCN) to detect the transition from "harvesting" to "loading." The recognition results are used to control the robot so that it approaches the worker before the loading action begins. The proposed method introduces a new labeling scheme tailored to harvesting and crop-loading motions, thereby improving recognition performance with ST-GCN. Evaluation experiments verified its generalization capability to different harvesting postures and workers, demonstrating an 18.7% improvement in action transition recognition accuracy compared with conventional methods. Furthermore, in robot-following experiments with the proposed method implemented, we confirmed that the system could both recognize action transitions and adjust the target following distance before the worker started loading. These results show that an ST-GCN specialized for agricultural tasks can effectively recognize harvesting action transitions, contributing to reducing the burden on workers during crop transportation.**

I. INTRODUCTION

In recent years, Japanese agriculture has been facing challenges such as an aging workforce and an increased labor burden caused by the expansion of farm management scales due to farmland consolidation [1]. To alleviate these burdens, robots that assist in transporting harvested crops have been proposed. Fig. 1 shows the agricultural robot targeted in this study performing transport assistance. A farm worker harvests crops, and when they can no longer carry them, loads them onto the robot. This study aims to develop a robot that maintains a safe distance behind the worker during harvesting and leaf/fruit thinning, and rapidly approaches within reach during loading. The target applications include leaf removal and fruit thinning for ridge-cultivated crops such as tomatoes, eggplants, and watermelons. In these tasks, depending on crop height and the specific operation, workers adopt a variety of postures, including squatting, bending at the waist, and standing while working overhead. The key challenge is to reliably recognize loading actions from multiple harvesting postures—despite differences in worker height and movement patterns—and to detect transitions from harvesting to loading promptly, enabling the robot to approach the worker at the right moment.

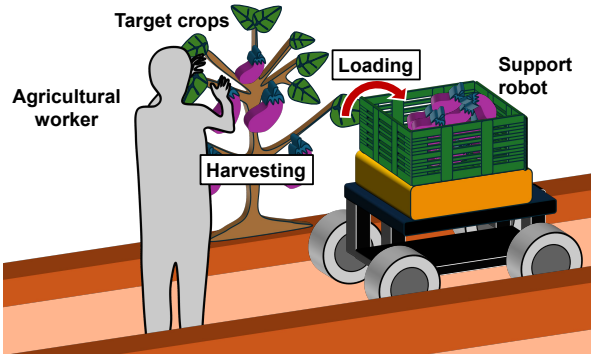


Fig. 1: An example of harvesting eggplants in a standing position using the transport support robot targeted in this study. The worker harvests crops growing at various heights in a variety of postures and loads them onto the robot.

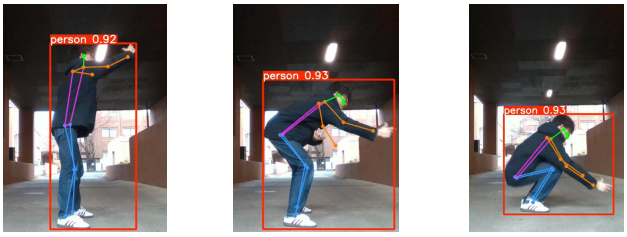
In prior work, a method was proposed in which an RGB-D camera mounted on the robot was used to maintain a constant distance from the farm worker and follow safely in narrow crop rows, as illustrated in Fig. 1 [2]. However, rather than simply keeping a constant distance, it is desirable for the robot to maintain an appropriate distance during harvesting so as not to interfere with the work, and to approach the worker during loading so that transportation assistance can be provided without interrupting the task. Furthermore, when the worker transitions from harvesting to loading, the robot should approach before the transition is complete, enabling the worker to continue tasks seamlessly without having to wait for the robot to arrive.

In the context of cooperative work between robots and workers in agriculture, a method has been proposed that adjusts the following distance based on the worker's center of gravity [3]. However, this method is applicable only when the worker's center of gravity differs between the "harvesting" and "loading" phases, and thus its application to a variety of crop harvesting scenarios is desirable. In addition, research has been conducted in which workers are equipped with IMUs and an LSTM model is used to achieve highly accurate action recognition [4]. Nevertheless, considering the aim of this study to reduce workload, a non-contact recognition method using sensors mounted on the robot is preferred.

In this study, we propose a method for recognizing the transition of agricultural worker's actions from harvesting to

[1]Go Oya and Rui Fukui are with the Department of Human and Engineered Environment Studies, The University of Tokyo, Chiba, 277-0882, Japan oya_go@lelab.t.u-tokyo.ac.jp

[2]Akihisa Ohya, Takashi Tsubouchi, and Ayanori Yorozu are with the Institute of Systems and Information Engineering, University of Tsukuba, Ibaraki, 305-8573, Japan yorozu@cs.tsukuba.ac.jp



(a) Standing (b) Semi-Squatting (c) Crouching

Fig. 2: Harvesting postures

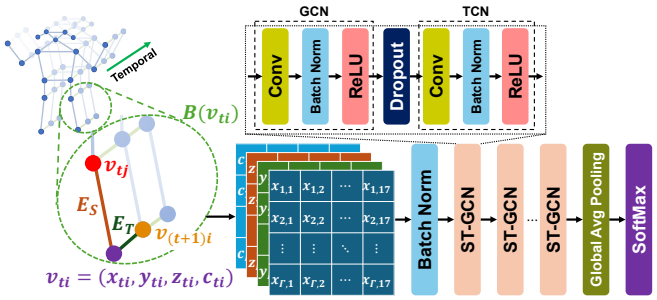


Fig. 3: The model structure of the ST-GCN used in this study

loading by applying the Spatial Temporal Graph Convolutional Network (ST-GCN) [5] to three-dimensional skeletal information obtained from an RGB-D camera. ST-GCN models each joint as a node and constructs a graph structure by connecting edges not only between joints within a frame but also between the same joints across adjacent frames along the temporal axis, achieving high recognition accuracy. However, although the datasets used in prior evaluations include various movements, they do not evaluate harvesting actions in agricultural work. Furthermore, ST-GCN uses a partitioning strategy dividing joint nodes into subsets to capture local motion features during graph convolution, with ongoing research to improve it [6] [7] [8]. However, existing strategies do not consider harvesting work characteristics, leaving the effect of a harvesting-specific partitioning on recognition accuracy unclear.

To address this, we propose a new partitioning strategy tailored to harvesting–loading work and evaluate its impact on recognition accuracy and the recognition delay time for the harvesting-to-loading transition. We also verify whether the following distance of the robot can be adjusted before the loading motion begins, based on the recognition results.

The major contributions of this study are as follows: 1)We propose a method to recognize action transitions of agricultural workers, whose harvesting postures vary depending on the type of crop, by applying ST-GCN. 2)Harvesting actions are characterized by the independent and consistent motion patterns of the arms and torso. We propose a new partitioning strategy based on this characteristic, termed Arm–Body Partitioning, and demonstrate that adopting a task-specific partitioning strategy contributes to improved recognition accuracy. 3)We implement the proposed method in a transport-assist robot and demonstrate that, with sufficient recognition accuracy and acceptable recognition delay between harvesting and loading, the robot can follow without interfering with the worker's task.

II. Action Transition Recognition and Following Using ST-GCN with a Novel Partitioning Strategy

A. Definition of harvesting actions in agricultural work

In this study, RGB images and depth information are acquired at 30 frames per second using an RGB-D camera. Based on manually classified input data, ST-GCN is trained to recognize harvesting and loading actions. The dataset used for training, validation, and evaluation includes a series of harvesting tasks involving transitions from each harvesting posture to loading. To enhance the generalization performance of recognition, harvesting tasks were performed from both left and right directions relative to the camera viewpoint for all harvesting postures.

We classify the actions into the following five categories. Harvesting actions are divided into three classes based on posture: "standing," "semi-squatting," and "crouching" harvesting, examples of which are shown in Fig. 2. The transition from harvesting to loading and the loading actions themselves are classified into two classes, "transition" and "loading," regardless of the harvesting posture. The reason for dividing them into two classes is to improve the accuracy of action transition recognition by enabling "transition" to be learned as an individual class.

B. Model Architecture Definition of ST-GCN

Spatial Temporal Graph Convolutional Networks (ST-GCN) [5] realize action recognition based on a spatio-temporal graph structure constructed from human joint node data. Discrete joint node coordinates are connected spatially by linking physically adjacent joint nodes, forming a spatial skeletal graph at a given time.

First, an undirected graph $G = (V, E)$ with N nodes and T frames is constructed. Here, the node set V is defined as follows, allowing the construction of a graph for a single frame:

$$V = \{v_{ti} \mid t = 1, \ldots, T,\ i = 1, \ldots, N\} \tag{1}$$

v_{ti} represents the feature of joint i at frame t. Each node v_{ti} has the following dimensions, including the three-dimensional coordinates x_{ti}, y_{ti}, and z_{ti}, as well as the confidence score c_{ti} from the skeleton estimation algorithm.

$$v_{ti} = (x_{ti}, y_{ti}, z_{ti}, c_{ti}) \tag{2}$$

The edge set E is divided into spatial edges E_{S} and temporal edges E_{T}, constructing the spatio-temporal graph. The spatial edges E_{S} represent connections between different joints within the same frame, and are defined as

$$E_{\mathrm{S}} = \{v_{ti}, v_{tj} \mid (i, j) \in H\} \tag{3}$$

where H is the set of human joint connections. The temporal edges E_{T} represent connections between the same joints

across different frames, and are defined as

$$E_{\mathrm{T}} = \{v_{ti}, v_{(t+1)i}\} \tag{4}$$

By combining these two types of edge sets, the spatio-temporal graph of the skeleton is constructed, allowing simultaneous acquisition of spatial and temporal information. These are illustrated as shown in Fig. 3.

As an action recognition network, ST-GCN enables the recognition of temporal sequences of actions by processing spatial and temporal information through two components: spatial graph convolution (SGCN) and temporal convolution (TCN). Based on the constructed human skeletal graph, the spatial graph convolution is defined as follows:

$$f_{\text{out}}(v_{ti}) = \sum_{v_{tj} \in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{\text{in}}(v_{tj}) \cdot W(l(v_{ti}, v_{tj})) \tag{5}$$

Here, $B(v_{ti})$ denotes the set of neighboring nodes for the sampling center v_{ti}, including nodes whose shortest distance (hop count) to v_{ti} is at most a constant D, where in this study $D = 1$. Z_{ti} is the number of elements in the neighboring node set, f_{in} represents the input features, which are the three-dimensional skeletal coordinates, and W is the weighting function. $l(v_{ti}, v_{tj})$ indicates the label mapping between node v_{ti} and its neighboring node v_{tj} based on the partitioning strategy. The partitioning strategy is described later in Section II-D. After processing with spatial graph convolution, nodes corresponding to the same joint across different frames are connected along the temporal direction, linking the spatial graph structures to form a temporal graph structure. As a result, a set of nodes with frame indices q satisfying the following condition is obtained, and based on this node set, actions are recognized:

$$B(v_{ti}) = \left\{ v_{qj} \;\middle|\; d(v_{tj}, v_{ti}) \leq D,\ |q - t| \leq \frac{\Gamma}{2} \right\} \tag{6}$$

Here, Γ denotes the temporal kernel size, q and t are frame index, and $d(v_{tj}, v_{ti})$ represents the shortest distance (number of hops) between nodes.

Based on these, the label mapping is modified as follows:

$$l_{ST}(v_{qj}) = l(v_{ti}, v_{tj}) + (q - t + \lfloor \Gamma/2 \rfloor) \times K \tag{7}$$

Here, K denotes the number of subsets into which the neighboring node set is divided according to the partitioning strategy.

C. Model Architecture for Action Recognition in Agricultural Work

The overall structure of the proposed network is shown in Fig. 3. As described in Section II-B, skeleton information spanning multiple frames is used as input, where human joints are represented as nodes and the connections of the skeleton as edges to construct the graph. With the number of joints J, the number of input frames T, and the feature dimension per node C (three-dimensional coordinates and joint detection confidence), the graph-structured data has dimensions of $J \times T \times C$ and serves as input to the spatio-temporal graph convolutional network.

Each basic unit of the proposed method consists of two modules: a spatial graph convolution module (GCN) and a temporal convolution module (TCN). The TCN follows the GCN, and each contains convolutional layers, batch normalization layers, and ReLU activation layers that extract temporal and spatial features, respectively. The entire network is composed of multiple basic units with residual connections applied. Before inputting to the network, normalization is performed via a batch normalization layer to accelerate convergence. The output feature maps are converted into fixed-size feature vectors by global average pooling. At the end of the network, a Softmax function is applied to classify the action classes and produce the final prediction.

D. Proposal of a new partitioning strategy

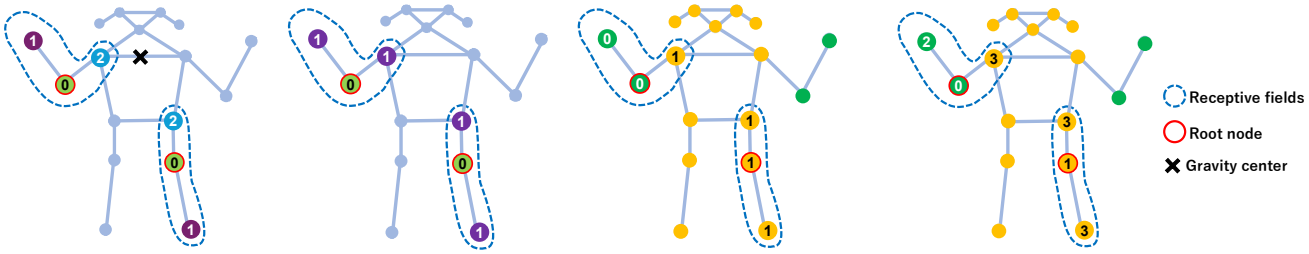
Fig. 4 illustrates the conventional partitioning strategy and the new partitioning strategy proposed in this study. In 2D convolution, adjacent pixels have a fixed spatial order, allowing the implementation of weight functions by indexing tensors according to this order. However, such an order does not exist in general graphs. Therefore, instead of assigning a unique index to each neighboring node, ST-GCN divides the set of neighboring nodes into predefined subsets and assigns different indices to each subset. This method is called the partitioning strategy. The area indicated by the blue dashed line shows an example of the filter's receptive field, and the node enclosed by the red box corresponds to the root node of the receptive field for the filter.

Here, Fig. 5 illustrates the difference in adjacency matrices with and without the partitioning strategy in a given graph structure. In the adjacency matrix, 1 is assigned if an edge exists between nodes, and 0 otherwise. By introducing the partitioning strategy, the adjacency matrix is formed separately for each divided subset, and the weight matrices are also learned individually for each subset, enabling the recognition of local features.

1) Spatial configuration: ST-GCN [5] introduces a partitioning strategy to improve recognition accuracy and, based on validation with large-scale datasets [9] [10], proposes a new partitioning strategy called "Spatial configuration partitioning," as shown in Fig. 4(a). This strategy divides the neighboring set into three subsets: 1) the root node itself, 2) the centripetal group: neighboring nodes closer to the skeleton's center of gravity than the root node, and 3) the centrifugal group: neighboring nodes farther from the skeleton's center of gravity than the root node.

Here, the center of gravity is defined as the average coordinate of all joints within the skeleton in a frame. Formally, let r_j be the distance from the root node to the center of gravity, and r_i be the distance from the center of gravity to joint i, represented by the following:

$$l(v_{ti}, v_{tj}) = \begin{cases} 0 & \text{if } r_j = r_i \\ 1 & \text{if } r_j < r_i \\ 2 & \text{if } r_j > r_i \end{cases} \tag{8}$$



(a) Spatial configuration (b) Distance (c) Arm-Body (d) Arm-Body & Distance

Fig. 4: Partitioning strategies used in the conventional method and the proposed method in this study. The numbers inside the nodes represent the label values for each strategy. **(a) Spatial configuration** partitioning. Nodes are labeled by their distance to the skeleton's center of gravity (black cross) relative to the root node (green): longer (purple, centrifugal) or shorter (blue, centripetal). **(b) Distance** partitioning. The two subsets are the root node itself (distance 0, green) and the neighboring nodes (distance 1, purple). **(c) Arm-Body** partitioning. The partition is made between the arms (green) and the body (yellow). **(d) Arm-Body & Distance** partitioning. The four subsets are: root–arm (green, red circle), root–body (yellow, red circle), non-root–arm (green, no circle), and non-root–body (yellow, no circle).

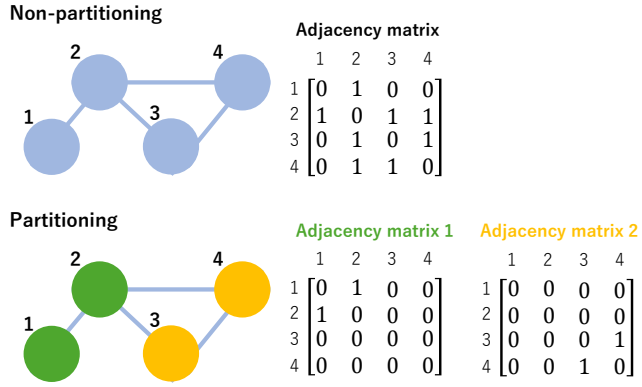


Fig. 5: Changes in the adjacency matrix due to the introduction of graph partitioning strategy

2) Distance Partitioning: A partitioning method based on the distance from the root node has also been proposed for dividing the node set [5]. In this study, since the 1-hop neighboring node set is used, the partitioning is as shown in Fig. 4(b): 1) the root node itself, and 2) all other nodes, which are represented by the following:

$$l(v_{ti}, v_{tj}) = d(v_{tj}, v_{ti}) \tag{9}$$

3) Arm-Body Partitioning: In action recognition using skeletal graphs, it has been reported that partitioning based on human body parts improves recognition accuracy and robustness [11] [12]. In harvesting tasks as well, arm movements and movements of other body parts have distinct characteristics specific to each action, and considering them separately is expected to improve recognition accuracy. Therefore, as shown in Fig. 4(c), the nodes are partitioned into 1) the arms and 2) other nodes, represented by the following:

$$l(v_{ti}, v_{tj}) = \begin{cases} 0 & \text{if arm} \\ 1 & \text{if body} \end{cases} \tag{10}$$

4) Arm-Body & Distance Partitioning: In the arm-distance partitioning, if the neighboring node set consists entirely of arm nodes or entirely of non-arm nodes, the representational capacity may decrease. Therefore, we propose a partitioning method that combines the arm partition with a distance-based partition from the root node (red circle), as shown in Fig. 4(d). The colors of the nodes indicate whether they are arms or not, and the numbers inside the nodes represent the distance from the root node. This is expressed by the following:

$$l(v_{ti}, v_{tj}) = \begin{cases} 0 & \text{if root \& arm} \\ 1 & \text{if root \& body} \\ 2 & \text{if non-root \& arm} \\ 3 & \text{if non-root \& body} \end{cases} \tag{11}$$

III. Implementation of an Action Transition Recognition Method in Harvesting Work

The flowchart of the following behavior based on action recognition using ST-GCN is shown in Figure **??**. During the training process, an action transition recognition model is created using joint position coordinates obtained by applying a pose estimation algorithm to RGB images and depth data. During the following process, actions are recognized using this model, and the robot is commanded to adjust the following distance to the worker based on the recognition results. The following sections provide detailed explanations of each process.

A. Acquisition of joint position coordinates during the training process

In this study, to create data for the training process, action data simulating actual harvesting tasks are collected using an RGB-D camera from a sufficiently distant position relative to the worker. For the RGB images, skeleton estimation is performed using YOLOv8n-pose [13], which detects 17 joint coordinates as 2D points in the image. Using the depth information and the estimated 2D joint coordinates, the three-dimensional joint positions with the camera as the origin are calculated by applying Eq. (12). The coordinate system is

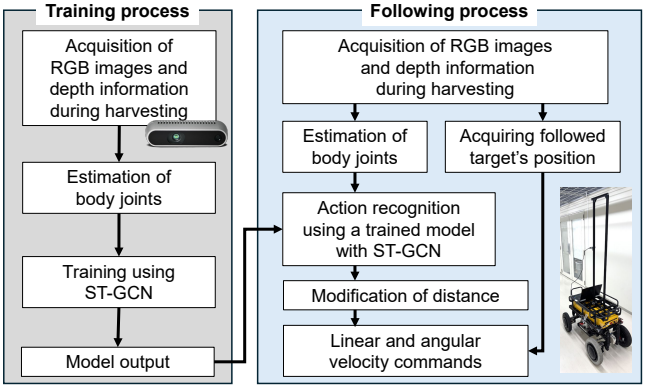
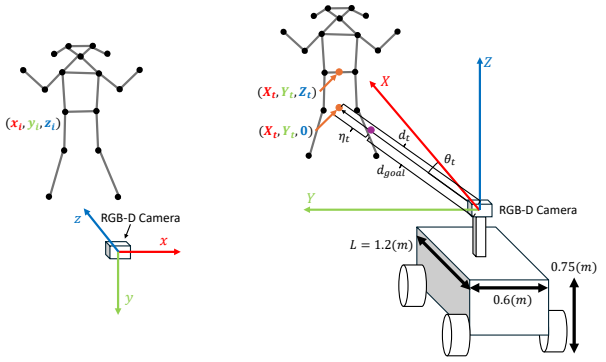


Fig. 6: Flowchart of action recognition using ST-GCN for a following robot



(a) Camera coordinate system (b) Robot coordinate system

Fig. 7: Comparison of coordinate systems used in action recognition and worker following. The camera coordinate system is used for action recognition, while the robot coordinate system is used for worker following.

defined as shown in Fig. 7(a).

$$(x_i, y_i, z_i) = ((u_i - c_x)d_i/f_x, (v_i - c_y)d_i/f_y, d_i) \quad (12)$$

Here, using the optical center (c_x, c_y) and focal lengths (f_x, f_y), the 3D coordinates of each keypoint (index i) obtained by YOLO are calculated from its 2D image coordinates (u_i, v_i) and depth information d_i. The data acquired in this way were classified by action, and the resulting dataset was used for training and model construction.

B. Acquisition of joint position coordinates during the following process

During robot following, the robot approaches the worker, which may cause the entire body of the worker to not fit within the camera's field of view. Therefore, to ensure that the whole body of the worker fits within the camera's view, the camera was installed with a pitch rotation of θ at the head height (1.7 m above the ground). Fig. 7(b) shows an example of the positional relationship between the robot and the worker. The transformation formula for joint positions is modified accordingly and, as shown in Eq. (13), a rotational transformation is applied to the x_i, y_i, z_i obtained by Eq. (12).

$$(x_i', y_i', z_i') = (x_i, y_i \cos\theta - z_i \sin\theta, y_i \sin\theta + z_i \cos\theta) \quad (13)$$

C. Calculation of the target-following position

Using the method described in Section III-B, the 3D coordinates of each joint are obtained, based on which the target position for following is determined. Specifically, using the 3D coordinates of the left hip ($i = 11$) and right hip ($i = 12$), the following point is calculated as shown in Eq. (14). Since the camera coordinate system (x, y, z) (Section II-B) and the robot coordinate system (X, Y, Z) differ, the transformed coordinates (x_i', y_i', z_i') obtained by Eq. (13) are used. Note that for distance calculation on the 2D plane, only the X and Y components of the following point are used:

$$(X_t, Y_t, Z_t) = \left(\min(z_{11}', z_{12}'), -\frac{x_{11}' + x_{12}'}{2}, -\frac{y_{11}' + y_{12}'}{2}\right) \quad (14)$$

Here, $\min(z_{11}', z_{12}')$ denotes the smaller value of the z' coordinates of the left and right hip nodes. If only one of the hip positions can be estimated, that coordinate is used; if neither can be estimated, the robot stops in place.

D. Calculation of translational and angular velocities based on recognition results

In this study, PID control is used to maintain the target following distance. In the robot coordinate system, the distance to the target is defined as $d_t = \sqrt{X_t^2 + Y_t^2}$, and the target distance that the robot should maintain is defined as d_{goal}, which is calculated as shown in Eq. (15).

$$d_{\text{goal}}\ (m) = \begin{cases} 1.4 + L/2 & \text{(if harvesting)} \\ 0.4 + L/2 & \text{(if loading)} \end{cases} \quad (15)$$

Here, L represents the total length of the robot (1.2 m), and since the camera is installed at the center of the robot, $L/2$ is the distance in the X direction from the robot's front end to the camera. These target following distances were determined based on actual measurements of comfortable distances for each task between the RGB-D camera and the worker.

Therefore, to control the velocity v_t so that d_t approaches d_{goal}, the error is defined as $\eta_t = d_t - d_{\text{goal}}$, and the control input is calculated using Eq. (16). Here, the PID control gains used in this study were experimentally determined as $K_{\text{p}} = 0.8$, $K_{\text{i}} = 10^{-5}$, and $K_{\text{d}} = 0.1$.

$$v_t = K_{\text{p}} \cdot \eta_t + K_{\text{i}} \cdot \int \eta_t \, dt + K_{\text{d}} \cdot \frac{d\eta_t}{dt} \quad (16)$$

The angular velocity ω_t is calculated according to Eq. (17). This method is based on pure pursuit [14], where ω_t is

TABLE I: Comparative evaluation of partitioning strategies on the constructed harvesting action dataset.

Partitioning Strategy	Accuracy (%)	Loading Accuracy (%)	Mean Recognition Delay Time (s)	Standard Deviation of Recognition Delay (s)
Spatial configuration	68.5	46.2	1.12	0.337
Distance	80.0	**83.3**	0.718	0.120
Arm-Body	87.1	57.1	0.597	0.307
Arm-Body & Distance	**87.2**	**83.3**	**0.535**	0.336

Bold numbers indicate the highest accuracy values or the shortest recognition delay times.

TABLE II: Action recognition performance based on NTU-RGB+D dataset. The accuracies on the cross-subject (X-Sub) and cross-view (X-View) benchmarks are reported.

Partitioning Strategy	X-Sub (%)	X-View (%)
Spatial configuration [5]	**81.5**	88.3
Distance	74.1	79.7
Arm-Body	80.5	89.9
Arm-Body & Distance	80.0	**91.9**

Bold numbers indicate the highest accuracy values.

computed using the following distance $\eta_t/(2\sin\theta_t)$ and the angle $2\theta_t$.

$$\theta_t = \arctan(Y_t/X_t), \quad \omega_t = 2(v_t/\eta_t)\sin\theta_t \tag{17}$$

IV. Experiment

A. Experimental Setup

In this study, a dataset consisting of approximately 82,000 frames simulating harvesting tasks in agricultural work was created using the Intel RealSense Depth Camera D435i. Using this dataset, an ST-GCN model tailored for harvesting tasks was trained, and recognition models were constructed under various conditions.

The computations were performed on a laptop PC equipped with an Intel Core i7-13620H CPU and an NVIDIA GeForce RTX 4070 Laptop GPU. The processing time was on average 2.4 ms per frame for image acquisition and skeleton recognition, and 93.4 ms on average for action recognition and output, totaling 95.8 ms. Thus, real-time processing for target point calculation in action recognition and robot following at 10 Hz operation is feasible.

Furthermore, following a four-wheel-drive, front and rear wheel steering robot, the robot maintains a longer distance for standing, semi-squatting, or crouching harvesting, and a shorter distance for transition or loading.

In this study, since the objective is to change the robot's following distance based on the worker's action transitions, evaluation in the experiments was conducted as a binary classification based on the target following distance.

B. Evaluation Experiments for Model Construction

1) Objective of the Verification and Evaluation Metrics: This section evaluates the performance of ST-GCN for harvesting action recognition based on skeleton information. First, using only the harvesting task data from participant A, the applicability of ST-GCN is confirmed, and a comparative evaluation is conducted among the proposed partitioning strategy and other methods.

In this experiment, the evaluation metrics were accuracy, loading accuracy, mean recognition delay time, and its standard deviation. Accuracy is the proportion of correctly recognized action frames, while loading accuracy measures correctness specifically during the loading task. High loading recognition accuracy helps prevent the robot from moving away at inappropriate times, enabling it to maintain a proper following distance for smooth operation. Mean Recognition delay time is the average duration between when a frame is classified as "transition" and when it is recognized as such. A delay shorter than the actual transition time (0.799 seconds in the evaluation dataset) indicates the system can anticipate the worker's intention before loading begins and adjust accordingly.

To verify whether the proposed partitioning strategy maintains performance in recognizing actions other than harvesting, this study uses the NTU RGB+D Action Recognition Dataset [10], provided by the ROSE Lab at Nanyang Technological University, Singapore. Since the worker varies across subjects and the body orientation of the worker changes depending on the camera viewpoint, we performed cross-subject and cross-view evaluations.

Next, using the partitioning strategy that achieved the highest evaluation, an cross-subject evaluation was conducted and verified using the same evaluation metrics described above.

2) Evaluation of Partitioning Strategies: The comparison of recognition results for each partitioning strategy is shown in Table I. The results demonstrate that the proposed "Arm-Body & Distance Partitioning" achieves the highest accuracy in recognizing worker actions compared to conventional methods. Specifically, the recognition rate improved from 68.5% to 87.2%, and the recognition delay time was reduced from 1.12 seconds to 0.535 seconds. Furthermore, since there is no significant difference in the standard deviation of the recognition delay time, it indicates that the proposed method can stably recognize action transitions. The results of cross-subject and cross-view evaluations based on the partitioning strategies using the NTU RGB+D [10] are presented in Table II. These results suggest that adopting the "Arm-Body & Distance Partitioning" maintains high recognition accuracy not only for the targeted harvesting and loading

TABLE III: Comparison of recognition accuracy in cross-subject evaluation.

Training data	Evaluation data	Accuracy (%)	Loading accuracy (%)	Mean recognition delay time (s)	Transition time (s)
Participant A	Participant A	87.2	83.3	0.535	0.799
	Participant B	67.7	45.2	0.164	0.515
	Participant C	66.0	56.7	0.330	0.601
Participants A & B	Participant A	**89.9**	**99.8**	**0.278**	0.799
	Participant B	**71.2**	**78.5**	0.288	0.515
	Participant C	**83.9**	**96.4**	0.363	0.601

Bold numbers indicate the highest accuracy or the shortest recognition delay, obtained by augmenting the data of participants used for training.

TABLE IV: Results of the robot following experiment

Harvesting Posture	Misrecognition rate during harvesting (%)	Following failure rate during harvesting (%)	Misrecognition rate during transition (%)	Following failure rate during transition (%)
Standing	15	5	10	0
Semi-Squatting	10	5	15	5
Crouching	0	0	10	0

actions during agricultural work but also for other actions performed by workers. This supports the realization of safe robot following. Based on these findings, this study adopts the "Arm-Body & Distance Partitioning" for all subsequent experiments.

3) Cross-Subject Evaluation: The recognition results of the cross-subject evaluation are shown in Table III. Since the transition time from harvesting to loading varies among participants, the average transition time for each participant is presented here. The model trained only on participant A's data showed a significant drop in accuracy for unseen participants. To address this, training included participants A and B, with their data adjusted for equal numbers of harvesting, loading, and transition instances. As a result, high accuracy was maintained for both trained and unseen participants. Additionally, recognition delay time was shorter than transition time, indicating transitions are detected before loading begins.

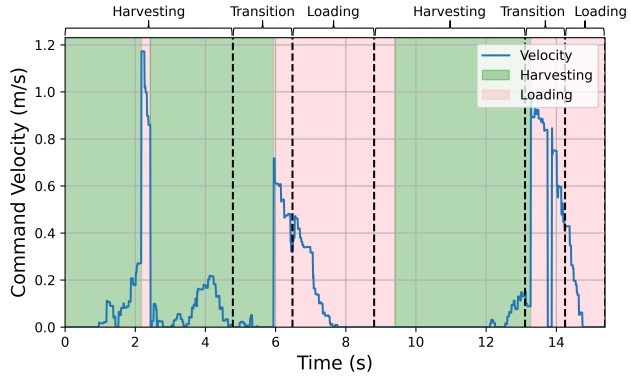
C. Evaluation Experiments with Implementation on a Following Robot

1) Verification Objective and Evaluation Metrics: In this experiment, we evaluate whether it is possible to achieve robot following based on the recognition of the behavior transition from "harvesting" to "loading" using a model capable of handling untrained participants created through cross-subject evaluation. The commands for velocity and angular velocity are issued at 10 Hz according to the recognized behavior from the acquired skeleton coordinates. Each posture (standing, semi-squatting, crouching) is performed 10 times for each left and right side, resulting in a total of 60 trials for verification. The transition from harvesting to loading within the sequence of harvesting actions is evaluated.

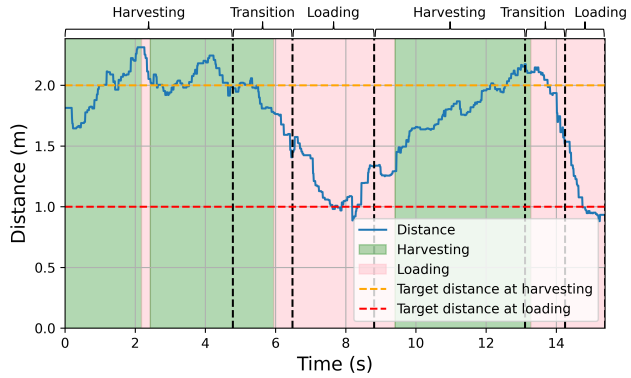
The evaluation metrics are the harvesting misrecognition rate, harvesting following failure rate, transition misrecognition rate, and transition following failure rate. The harvesting misrecognition rate indicates the proportion of frames during harvesting that were mistakenly recognized as "loading." The harvesting following failure rate indicates the proportion of frames during harvesting that were mistakenly recognized as loading and for which the following distance changed by more than 0.5 m. Considering that the difference in following target distances set for harvesting and loading is 1 m, approaching more than half of this difference could cause discomfort to the worker. Furthermore, since the average arm length of Japanese males is approximately 0.75 m [15], it was determined that the robot approaching within 1 m poses a risk of injury due to contact, and thus 0.5 m was set as the threshold. The transition misrecognition rate refers to the proportion of cases in which the system failed to recognize the transition from "harvesting" to "transition" and then to "loading" despite the worker performing it. The transition following failure rate indicates the proportion of cases where the behavior transition was not recognized even after 0.5 s had passed since the worker transitioned to loading. This is based on the fact that the average transition time obtained from the experiment was less than 0.5 s, and it was considered necessary for smooth following assistance that the worker be able to move to the loading action within 1 s after completing harvesting.

2) Experimental Results: The recognition results from the robot implementation are shown in Table IV. Fig. 8 depicts the time variation of the velocity commands sent to the robot and the distance to the following target, including two behavior transitions from semi-squatting harvesting to loading.

From Table IV, it can be seen that misrecognition occurred during both the harvesting and loading stages. However, most of the misrecognitions during harvesting were temporary, and the robot was able to correctly recognize "harvesting" again



(a) Time-series of command velocity issued to the following robot



(b) Time-series of the distance between the robot and the worker

Fig. 8: Results of the robot following experiment. Ground truth classes are shown by black dashed boundaries and captions at the top, while recognition results are shaded green for harvesting and pink for loading.

while approaching, thus maintaining an appropriate distance for harvesting. Around 2 s in Fig. 8, there is a moment where "loading" was mistakenly recognized during harvesting, but since this misrecognition was brief, it was confirmed that the distance between the robot and the worker did not become excessively close.

Additionally, when the recognition during transition was incorrect, around 7 s, the distance to the worker was not adjusted to approximately 1 m even after 0.5 s had passed since the end of the transition. However, around 14 s, the following distance was adjusted within 0.5 s after the worker transitioned to loading. These observations indicate that the system meets the requirements of an agricultural harvesting support system that does not increase the worker's burden.

Based on the above results, this experiment demonstrates that by applying ST-GCN for behavior transition recognition in the robot, the recognition delay time between harvesting and loading actions does not interfere with the work, enabling smooth following.

V. CONCLUSION

In this study, we proposed a method for recognizing behavior transitions of agricultural workers using ST-GCN with 3D skeleton data obtained from an RGB-D camera as input. To handle motions specific to harvesting and loading tasks, we introduced a new partitioning strategy for the joint graph structure called "Arm-Body & Distance partitioning." Evaluation results confirmed improved recognition accuracy and reduced delay in recognizing transitions from harvesting to loading. Consequently, the proposed method demonstrated the ability to recognize behavior transitions and to adjust the following target distance before the worker begins loading.

Future challenges include verifying the effectiveness of the proposed method more realistically by evaluating its robustness against visual occlusions caused by crops and surrounding environments in actual field conditions. Furthermore, we aim to quantitatively assess the psychological and physical burdens imposed on workers by dynamic changes in following distance, and by collecting and analyzing feedback from workers, to develop a more practical and cooperative support system.

REFERENCES

[1] Ministry of Agriculture, Forestry and Fisheries. 2020 agricultural and forestry census (in japanese), 2020 (accessed Aug. 2025). https://www.maff.go.jp/j/tokei/census/afc/2020/index.html.

[2] Ayanori Yorozu, Genya Ishigami, and Masaki Takahashi. Human-following control in furrow for agricultural support robot. In *Proceedings of the International Conference on Intelligent Autonomous Systems*, pages 155–164, 2021.

[3] Ayanori Yorozu, Genya Ishigami, and Masaki Takahashi. Ridge-tracking for strawberry harvesting support robot according to farmer's behavior. In *proceedings of Field and Service Robotics: Results of the 12th International Conference*, pages 235–245, 2021.

[4] Athanasios Anagnostis, Lefteris Benos, Dimitrios Tsaopoulos, Aristotelis Tagarakis, Naoum Tsolakis, and Dionysis Bochtis. Human activity recognition through recurrent neural networks for human–robot interaction in agriculture. *Applied Sciences*, 11(5):2188, 2021.

[5] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. *AAAI*, 2018.

[6] Shiqiang YANG, Zhuo LI, Jinhua WANG, Duo HE, Qi LI, and Dexin LI. St-gcn human action recognition based on new partition strategy. *Computer Integrated Manufacturing System*, 29(12):4040, 2023.

[7] Xiangang Cao, Chiyu Zhang, Peng Wang, Hengyang Wei, Shikai Huang, and Hu Li. Unsafe mining behavior identification method based on an improved st-gcn. *Sustainability*, 15(2):1041, 2023.

[8] Quanyu Wang, Kaixiang Zhang, and Manjotho Ali Asghar. Skeleton-based st-gcn for human action recognition with extended skeleton graph and partitioning strategy. *IEEE Access*, 10:41403–41410, 2022.

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint*, 2017.

[10] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[11] Kalpit Thakkar and PJ Narayanan. Part-based graph convolutional network for action recognition. *arXiv preprint*, 2018.

[12] Yang Qin, Lingfei Mo, Chenyang Li, and Jiayi Luo. Skeleton-based action recognition by part-aware graph convolutional networks. *The visual computer*, 36(3):621–631, 2020.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[14] Omead Amidi and Chuck E Thorpe. Integrated mobile robot control. In *Proceedings of the SPIE, Mobile Robots V*, volume 1388, pages 504–523, 1991.

[15] Makiko Kouchi and Masaaki Mochimaru. Aist anthropometric database. National Institute of Advanced Industrial Science and Technology, H16PRO 287, 2005.