Action Recognition of Excavators Using Physical Simulator and Real Image Data with Class-Dependent Data Augmentation

Tomohiro Katsura[1], Ren Komatsu[1], Keiji Nagatani[1], Takumi Chiba[2], Kazuhiro Chayama[2], Atsushi Yamashita[3] and Hajime Asama[1]

***Abstract*— In this study, we proposed a training method using joint points obtained from physical simulations and real image data for the action recognition of excavators. The proposed method classifies the action recognition of excavators into position detection, skeleton detection, and action recognition models. The first two models are trained using the real image data, whereas the action recognition model is trained using the joint point data obtained from the physical simulation. For the action recognition model, we proposed a data augmentation method based on the features of the actions of the excavator. Experimental results indicate that the proposed method can achieve better accuracy than the conventional method that uses real video data, though the proposed method does not use any real video data for training.**

I. INTRODUCTION

The construction industry, which is considered to have lower productivity compared to other industries, relies on construction machinery for the majority of its operation [1]. Excavators are a major part of construction machinery. Therefore, it is important to improve the productivity of excavators based on evaluation and analyses. The excavators repeat the following operations: "Dig," "Swing," "Load," and "Swing" (Fig. 1). This cycle time is considered an important indicator in evaluating and analyzing the productivity of excavators [2]. To measure and analyze the cycle time, it is necessary to record action time. The manual recording of action time is time-consuming, expensive, and prone to errors [3]. To reduce time and cost, an automated system for the action recognition of excavators is highly required. Automated systems for the action recognition of excavators can be classified into two categories: internal-sensor- and external-sensor-based methods. The method with internal sensors recognizes the action of excavators based on the information obtained from sensors mounted on the excavators [4], [5], [6], whereas the method with external sensors recognizes the action of excavators by externally capturing videos using cameras.

[1]T. Katsura, R. Komatsu, K. Nagatani and H. Asama are with the Department of Precision Engineering, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan, {katsura, komatsu, asama}@robot.t.u-tokyo.ac.jp, keiji@ieee.org

[2]T. Chiba and K. Chayama are with Fujita Corporation, 4-25-2 Sendagaya, Shibuya-ku, Tokyo, 151-8570, Japan, takumi.chiba@fujita.co.jp, fujita.chayama@gmail.com

[3]A. Yamashita is with the Department of Human and Engineered Environmental Studies, Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba, 277-8563, Japan, yamashita@robot.t.u-tokyo.ac.jp

(a) Dig (b) Swing (c) Load (d) Swing

Fig. 1: Excavators repeat "Dig," "Swing," "Load," and "Swing."

The internal-sensor-based method consumes a lot of time for attaching and detaching sensors, thereby incurring high cost [1]. Moreover, in several cases, sensors cannot be installed in rented or traditional excavators [7]. Therefore, a low-cost automated system is required for the action recognition of excavators using camera images and does not involve the attachment or detachment of sensors.

Herein, the images of real environments with bounding boxes (BBox), keypoints, and labels are referred to as real image data and the videos of real environments with BBoxes, keypoints, and labels are referred to as real video data. The method that uses machine learning to perform action recognition based on camera videos first detects the positions of excavators and then recognizes the action of the excavators [8]. However, there is a scarcity of real video data for training excavators to achieve sufficient action recognition accuracy, and the system is susceptible to differences in camera viewpoints [8]. Moreover, the action recognition of an excavator is different from general action recognition, in which the same movement is made only once during a single action event, making action recognition difficult [9]. For instance, when humans walk or run, their arms and legs repeat the same movements. However, when an excavator swings, it only rotates once and does not repeat the same movement during a single action event. In this case, recognizing the action of an excavator becomes more difficult than that of a human.

A large volume of the real image data of excavators that

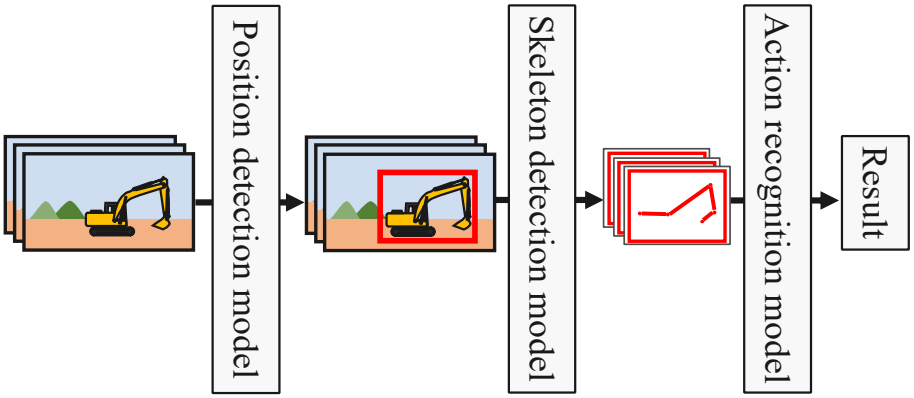


Fig. 2: Proposed method, where the action recognition of excavators is divided into three models: position detection, skeleton detection, and action recognition.

can be used to train object detection models are publicly available [10], [11]. In these datasets, there are variations in viewpoints, the postures of the excavator, and backgrounds. Meanwhile, a few hundred real video datasets are available for training the action recognition model [12], which are insufficient for varying viewpoints, excavator actions, and backgrounds. Creating a real video dataset with varying viewpoints, excavation actions, and backgrounds is time-consuming and expensive. Therefore, a method that can learn from other types of data is desirable.

To perform training without using real video data, Kasahara et al. inserted the real images of actual shooting sites into the background of video data obtained from physical simulation and made it possible to learn for action recognition [13]. However, the real images in the background were not used for simulation. To obtain real images for the background, the location of the camera for action recognition must be fixed before learning, and the work site must not be significantly different between the time of training and time of action recognition. As work progresses, the work site changes day by day, and the appearance of the work site during the time of training is different from that during the time of action recognition. Therefore, a method that can respond to daily changes in the work site is required.

The development of a method that can address the following two situations is desirable: (a) There are only real video datasets with few variations in the viewpoint and the excavating motion itself; (b) it is not possible to use information specific to particular surrounding environments. Therefore, this study aims to develop a method that can learn from data other than real video data and cope with daily changes in the work site. In particular, the first goal is to develop a method that can learn from physical simulation and real image data without using any worksite-specific data, and the second goal is to demonstrate an approach for efficiently augmenting data from physical simulation.

II. PROPOSED METHOD

A. Concept

It is important to accurately recognize actions based on camera images captured from arbitrary directions. Therefore, to achieve highly accurate action recognition, data considering different viewpoints are required during training. In the case of excavators, as it is difficult to obtain real video datasets viewed from various directions, we consider collecting data viewed from various directions via physical simulation and using them as training data.

Only the skeletal data of the excavator are needed to recognize the action of the excavator. Therefore, we focus on the movement of the skeleton of the excavator. We extract the time-series skeletal transition data of the excavator from the video data for action recognition and perform action recognition using an action recognition model. The action recognition model is trained by obtaining the time-series skeletal data of the excavator from the simulation data.

The range of variation in the way an excavator moves differs for each movement. Therefore, the accuracy of action recognition can be effectively improved by expanding the data based on the characteristics of each action event.

B. Proposed Method

As shown in Fig. 2, the proposed method classifies the action recognition of excavators into three models: excavator position detection, skeleton detection, and action recognition. Each model is explained in detail.

Compared to the real video data, the real image data of excavators possess more variations in the directions of shooting and excavation action. Therefore, the position detection model is trained using the real image data. Position detection is performed prior to excavator skeleton detection to remove unnecessary information and increase the accuracy of skeleton detection.

Moreover, the skeleton detection model is trained using real image data. The skeleton of an excavator is detected, and it is assumed that the action recognition model can be trained via physical simulation.

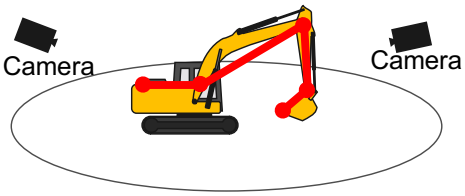


Fig. 3: Data acquisition. The excavator is moved via physical simulation, and the skeleton model of the excavator is projected on multiple 2D planes, as if it were being photographed from various viewpoints.

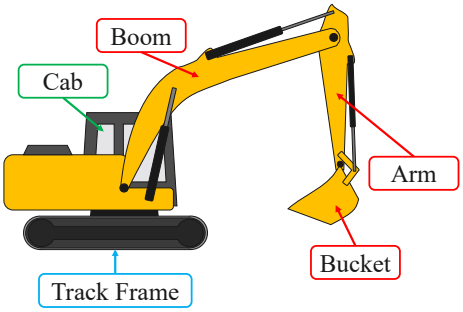


Fig. 4: Name of each part of an excavator.

.

There is a scarcity of real video data in terms of variation in the viewpoint and excavating motion. Therefore, the action recognition model is learned based on the time-series skeletal transition data obtained from physical simulation.

C. Acquiring Training Data for Action Recognition Models

The training data for the action recognition model are obtained via physical simulation. The simulation process is shown in Fig. 3. In the simulation process, an excavator is moved, and the time-series 3D coordinates of the skeleton are obtained. Next, we estimate the range where a camera for action recognition can be installed. At arbitrary locations within the range, the time-series 3D coordinates of the skeleton are simultaneously projected on multiple 2D planes as if they were captured using a camera. This process helps in obtaining time-series skeletal transition data viewed from various directions at a time. A more detailed explanation is provided hereafter. To obtain training data, the motion of the excavator is reproduced M_{train} times in the simulation process. The obtained 3D time-series coordinates of the skeleton are projected on M_{view} 2D planes, which helps extend the data to N_{train}. Similarly, to obtain validation data, in the simulation process, the motion is reproduced M_{val} times and projected on M_{view} 2D planes to extend the data to N_{val}. $N_{\mathrm{train}} = M_{\mathrm{train}} M_{\mathrm{view}}$ and $N_{\mathrm{val}} = M_{\mathrm{val}} M_{\mathrm{view}}$. By increasing the number of projection directions, it becomes easier to obtain time-series skeletal transition data viewed from multiple directions.

D. Data Augmentation

To learn sufficiently, we augment the data based on the characteristics of the excavator's movements. In addition, general data augmentation methods are performed.

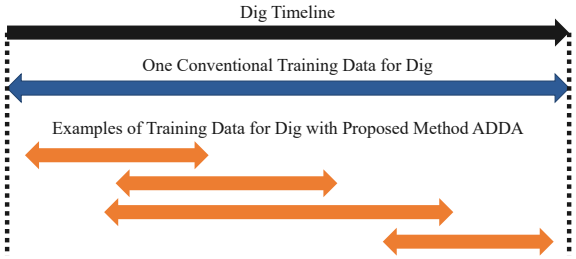


Fig. 5: Data augmentation method (ADDA). Compared to other actions, each dig exhibits a considerable difference owing to the influence of changes in the ground surface and the amount of excavation. Therefore, we propose a data augmentation method, ADDA, which extracts a random part of the data of each dig with a random length and uses it as training data. This method improves robustness.

1) Action-Driven Data Augmentation: Fig. 4 shows the names of each part of the excavator. Excavators repeatedly dig, swing, and load. Dig has more diverse movements than swing and load. In swing, the track frame is fixed, and the cab, boom, arm, and bucket rotate at an arbitrary angle in unison. The swing is not affected by differences in the shape of the ground surface or the amount of excavation, and there is no significant difference in each swing. Load is an action in which the track frame and cab are fixed and the bucket position itself is not moved much, and the boom, arm, and bucket are moved so that the bucket is inverted. Load is not affected by differences in the shape of the ground surface or the amount of excavation, and thus each load does not differ significantly. Dig, on the other hand, is a large movement of the boom, arm, and bucket, while the track frame and cab remain fixed. Each dig differs considerably due to differences in the shape of the ground surface and the amount of excavation. Therefore, for data augmentation, a random portion of each dig of random length was extracted and used as training data (Fig. 5). The data augmentation method, i.e., action-driven data augmentation, is hereafter referred to as "ADDA."

2) General Data Augmentation: Excavators may be placed in a tilted position. Therefore, we rotate the obtained time-series skeletal transition data and perform data augmentation. In addition, to make the data robust to noise, which is generated when the excavator's skeleton is detected, we mix noise with the time-series skeletal transition data and perform data augmentation.

III. EXPERIMENTS

The first objective of this experiment is to confirm that action recognition models can be trained using time-series skeletal transition data obtained via physical simulation. The second objective is to confirm that ADDA improves action recognition accuracy. The third objective is to provide an efficient way to collect simulation data optimal for training.

We used the proposed method for the action recognition of excavators, specifically "Dig," "Swing," and "Load," and compared its action recognition accuracy with that of the

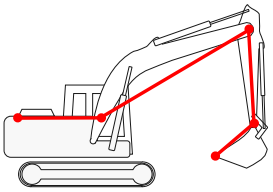


Fig. 6: Skeleton model of an excavator. Five key points are connected by straight lines: the tip of the bucket, axis of rotation of the bucket, axis of rotation of the arm, root of the boom, and rear of the excavator body to prepare the framework of the excavator.

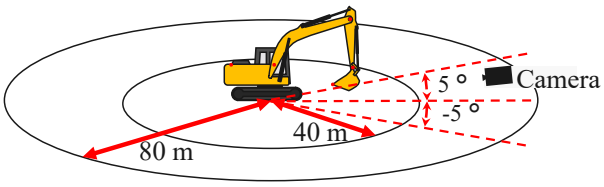


Fig. 7: Data augmentation via physical simulation. Data are extended by moving an excavator in the simulation process, and skeletal transition data are collected considering multiple viewpoints.

conventional method. The proposed method comprised position detection, skeleton detection, and action recognition models. Each training model is explained first. Then, conventional methods and experimental results are described and discussed.

A. Position Detection Model

To detect the position of an excavator and obtain its BBox, we used Mask R-CNN [14]. ResNet-101 was used as the backbone to convert the input images into features.

The Moving Objects in Construction Site (MOCS) [15] dataset was used as the training data for the position detection model. A total of 23,404 real image data in the MOCS dataset were annotated with BBox information for 12 types of construction machinery, including excavators. The real image data of MOCS included 12,057 excavators.

The Alberta Construction Image Dataset (ACID) [11] was used as the validation data for the position detection model. A total of 2850 real image data in ACID were annotated with BBox information for 3 types of construction machinery, including excavators. The real image data of the ACID included 2388 excavators.

Three methods were used to enhance the data. The first method is left–right inversion. Excavators are not always flipped upside down; however, they are sometimes flipped left and right. Left-right flipping was considered with a probability of 50%. The second method is rotation. As the ground surface may be tilted where excavators work, the images were randomly rotated at a maximum rotation angle of 20°. The third method is cropping. The areas in the image were randomly cropped while maintaining the width and height of the image at 70% or more.

The input image size was set to 1333 × 640, and the number of training epochs was set to 40. Among the results obtained, the weights of the most recent position detection model with the highest mAP were used for evaluation.

B. Skeleton Detection Model

As shown in Fig. 6, the excavator's skeleton was defined as a straight line connecting five key points: the tip of the bucket, axis of rotation of the bucket, axis of rotation of the arm, root of the boom, and rear of the body. A typical skeleton detection model, HRNet [16], was used as the skeleton detection model.

To train the skeleton detection model, we used the Luo dataset [17]. The dataset included 1281 real image data of excavators annotated with skeletal information; 1000 images were used as training data, and the remaining 281 images were used as validation data.

For data augmentation, the data were flipped left to right with a probability of 50% and randomly rotated at a maximum rotation angle of 20°.

The input image size was set to 256 × 192, and the number of training epochs was set to 350. Among the results obtained, the weights of the most recent skeleton detection model with the lowest loss function were used for evaluation.

C. Action Recognition Model

Pose-SlowOnly [18] was used as the action recognition model. The action recognition model was trained using the time-series skeletal transition data obtained from the simulation process.

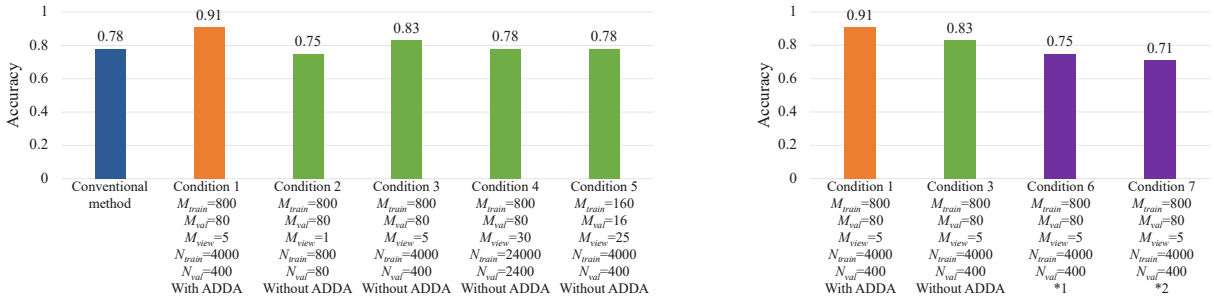
To train the action recognition model, we used the physical simulation process, PyBullet, to obtain the 3D coordinates of each key point of the excavator during each action. As shown in Fig. 7, it was assumed that the camera for action recognition was located at a horizontal distance of 40–80 m from the excavator and an elevation angle of -5° to 5°, as viewed from the excavator. The 3D coordinates of each key point are projected on M_{view} 2D planes within the assumed range.

The excavator repeated the following actions: "Dig," "Swing," "Load," and "Swing." To obtain training data, "Dig," "Swing," and "Load" were repeated M_{train} times in the simulator and projected to M_{view} 2D planes, expanding the data to N_{train}. Similarly, to obtain validation data, the actions were reproduced M_{val} times, projected to M_{view} 2d planes, and extended to N_{val} data. The relationships among M_{train}, M_{val}, M_{view}, N_{train}, N_{val}, and ADDA in experimental conditions 1–7 are summarized in Table I. The action cycle comprised "Dig," "Swing," "Load," and "Swing," and the number of "Swing" events was larger than that of "Dig" and "Load" events. Therefore, the ratio of "Dig," "Swing," and "Load" in training data was set to 1:2:1.

We randomly rotated the time-series skeletal transition data at a maximum rotation angle of 10°. Noise following a Gaussian distribution was added to each key point of the excavator to make it robust against errors in skeleton detection. The number of training epochs was set to 400,

TABLE I: Conditions of the experiment. *1: The dig and load data are maintained as they were, and only swing data are extracted from random positions of random lengths. *2: The dig and swing data are maintained as they are, and only load data are extracted from random positions of random lengths.

	Condition 1	Condition 2	Condition 3	Condition 4	Condition 5	Condition 6	Condition 7
M_{train}	800	800	800	800	160	800	800
M_{val}	80	80	80	80	16	80	80
M_{view}	5	1	5	30	25	5	5
N_{train}	4000	800	4000	24000	4000	4000	4000
N_{val}	400	80	400	2400	400	400	400
ADDA	With	Without	Without	Without	Without	$*1$	$*2$



(a) Effectiveness of the proposed method. The proposed method represents Condition 1.

(b) Effect of ADDA. *1: Only swing data are extracted from random positions of random lengths. *2: Only load data are extracted from random positions of random lengths.

Fig. 8: Action Recognition Accuracy.

and the weights of the most recent action recognition model with the highest top-1 accuracy among the results obtained were used for evaluation.

D. Conventional Methods for Comparison

The conventional method comprised a position detection model, skeleton detection model, and action recognition model. The position of the excavator (BBox) in the camera image was detected using the position detection model, the skeleton was detected from the RGB image in the detected BBox using the skeleton detection model, and action recognition was performed based on the data obtained from the skeleton model using the action recognition model.

The method described in Section III-A was used as the position detection model. The method described in Section III-B was used for the skeleton detection model. PoseSlowOnly [18] was used as the action recognition model. For training the action recognition model, we used a part of the dataset [12]. We used 222 clips as training data and 47 clips as validation data.

E. Test Data Used to Determine Action Recognition Accuracy

We used 179 clips from the Roberts dataset [12], which were not used in the previous section as test data. This test dataset contained 47 clips of "Dig," 90 clips of "Swing," and 42 clips of "Load."

F. Results and Discussion

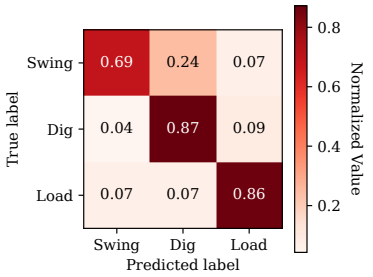
Fig. 8 shows the accuracy of action recognition. Equation (1) is used to calculate the action recognition accuracy.

$$r = \frac{N_{\mathrm{correct}}}{N_{\mathrm{video}}}. \tag{1}$$

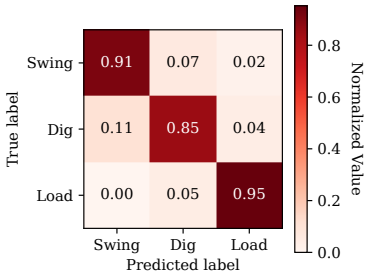
Here, r represents the action recognition accuracy, N_{correct} denotes the number of correctly judged real videos, and N_{video} denotes the total number of judged real videos. The normalized confusion matrix of each result is shown in Fig. 9.

Fig. 8(a) shows that the action recognition accuracy of the conventional method is 78%. The proposed method based on Condition 1 possesses an action recognition accuracy of 91%. Although the proposed method does not use real video data for training, it can achieve a much higher action recognition accuracy than the conventional method that uses real video data for training.

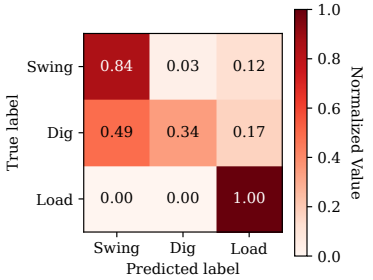
Next, we consider the approach for efficiently collecting training data from simulations. Comparing Conditions 2, 3, and 4 with different M_{view}, Condition 3 exhibits the highest action recognition accuracy. When M_{view} is extremely small, as in Condition 2, the total amount of data is extremely small and the action recognition accuracy does not improve. Meanwhile, if M_{view} is considerably large, as in Condition 4, the action recognition accuracy does not improve. In other words, for proper training, the variation in the 2D plane on which the dig is projected (M_{view}) and the variation in the excavation action (M_{train}, M_{val}) must possess an appropriate relationship. It was found that large or small



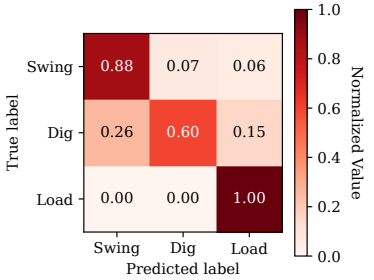
(a) Conventional method



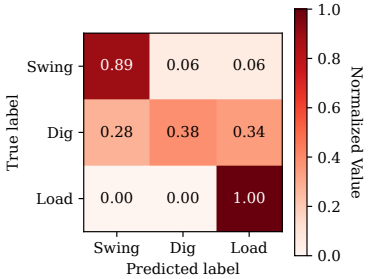
(b) Condition 1 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 5$, $N_{\mathrm{train}} = 4000$, $N_{\mathrm{val}} = 400$, With ADDA)



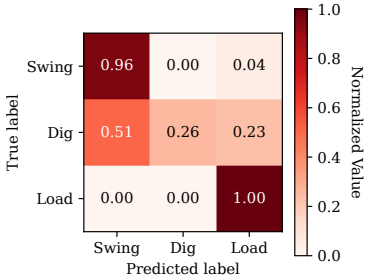
(c) Condition 2 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 1$, $N_{\mathrm{train}} = 800$, $N_{\mathrm{val}} = 80$, Without ADDA)



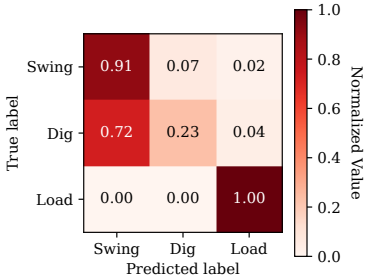
(d) Condition 3 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 5$, $N_{\mathrm{train}} = 4000$, $N_{\mathrm{val}} = 400$, Without ADDA)



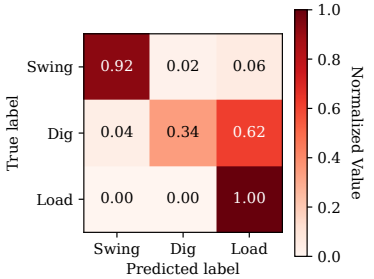
(e) Condition 4 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 30$, $N_{\mathrm{train}} = 24000$, $N_{\mathrm{val}} = 2400$, Without ADDA)



(f) Condition 5 ($M_{\mathrm{train}} = 160$, $M_{\mathrm{val}} = 16$, $N_{\mathrm{view}} = 25$, $N_{\mathrm{train}} = 4000$, $N_{\mathrm{val}} = 400$, Without ADDA)



(g) Condition 6 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 5$, $N_{\mathrm{train}} = 4000$, $N_{\mathrm{val}} = 400$, *1)



(h) Condition 7 ($M_{\mathrm{train}} = 800$, $M_{\mathrm{val}} = 80$, $N_{\mathrm{view}} = 5$, $N_{\mathrm{train}} = 4000$, $N_{\mathrm{val}} = 400$, *2)

Fig. 9: Normalized confusion matrix. *1: Only swing data were extracted from random positions of random lengths. *2: Only load data were extracted from random positions of random lengths.

variations (N_{view}) in the 2D plane on which excavation was projected (M_{train} and M_{val}) did not influence the reduction in the action recognition accuracy.

Comparing Conditions 3 and 5, where N_{train} and N_{val} are the same value, Condition 3 has higher accuracy in recognizing the actions. Condition 3 projects more excavation movements onto fewer 2D planes than Condition 5. Therefore, compared to Condition 5, Condition 3 has more variations of the original action, and the accuracy of action recognition is higher.

The effect of the data augmentation method ADDA is described. Fig. 8(b) shows that Condition 1 is more accurate for action recognition than Condition 3. Figs. 9(b) and 9(d) show that the rate of correctly recognized dig videos is particularly increased. Therefore, it is clear that the accuracy of action recognition was improved by ADDA.

Next, we compare Condition 1, Condition 6, and Condition 7. Compared to the cases in which the time-series

skeletal transition data of swing and load are extracted from random positions for random lengths (Condition 6 and Condition 7), Condition 1 is much more accurate in action recognition. Therefore, it was found that the accuracy of action recognition can be improved by applying data augmentation to the dig, as in ADDA. In addition, Fig. 9(d), 9(g) and 9(h) show that applying data augmentation only to swing and load does not significantly improve the recognition accuracy, indicating that data augmentation based on the characteristics of each action is necessary, as in ADDA.

IV. CONCLUSION

Herein, we proposed a method that can learn from simulated and real image data. Moreover, we demonstrate an approach to efficiently augment data via physical simulation. Although the proposed method does not use real video data for training, it was found to be more accurate than conventional methods that use real video data for training. The proposals are as follows:

(1) The proposed method divided the action recognition of excavators into the following models: position detection, skeleton detection, and action recognition. We proposed the method to learn the first two models from real image data and the action recognition model from time-series skeleton transition data obtained via physical simulations.

(2) We proposed an efficient method to collect time-series skeletal transition data via physical simulation. The time-series skeletal transition data were obtained by obtaining the 3D coordinates of an excavator's skeleton via physical simulation and projecting them on a 2D plane as if it were photographed from various directions. Moreover, we proposed ADDA to augment the data by extracting only the data of the excavator's dig from a random position for a random time length, thereby taking advantage of the characteristics of the excavator's action. We found that this method significantly improves the accuracy of action recognition.

As the proposed method does not require annotated real video data, it takes less time and is less expensive than other conventional methods. Therefore, using the proposed method, it becomes easy to collect training data for "Dig," "Swing," and "Load," and for new action recognition classes, e.g., stop and move. In the future, we will try to develop a method to recognize action classes other than "Dig," "Swing," and "Load." Moreover, we will aim to realize a system that can recognize the movements of an excavator by placing a camera, and can record the movements of excavators for a day.

REFERENCES

[1] E. R. Azar and B. McCabe, "Part based model and spatial–temporal reasoning to recognize hydraulic excavators in construction images and videos," *Automation in construction*, vol. 24, pp. 194–202, 2012.

[2] T. Cheng, J. Teizer, G. C. Migliaccio, and U. C. Gatti, "Automated task-level activity analysis through fusion of real time location sensors and worker's thoracic posture data," *Automation in Construction*, vol. 29, pp. 24–39, 2013.

[3] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771–782, 2011.

[4] N. Pradhananga and J. Teizer, "Automatic spatio-temporal analysis of construction site equipment operations using gps data," *Automation in Construction*, vol. 29, pp. 107–122, 2013.

[5] R. Akhavian and A. H. Behzadan, "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers," *Advanced Engineering Informatics*, vol. 29, no. 4, pp. 867–877, 2015.

[6] H. Kim, C. R. Ahn, D. Engelhaupt, and S. Lee, "Application of dynamic time warping to the recognition of mixed equipment activities in cycle time measurement," *Automation in Construction*, vol. 87, pp. 225–234, 2018.

[7] E. Rezazadeh Azar, S. Dickinson, and B. McCabe, "Server-customer interaction tracker: computer vision–based system to estimate dirt-loading cycles," *Journal of Construction Engineering and Management*, vol. 139, no. 7, pp. 785–794, 2013.

[8] C. Chen, Z. Zhu, and A. Hammad, "Automated excavators activity recognition and productivity analysis from construction site surveillance videos," *Automation in Construction*, vol. 110, p. 103045, 2020.

[9] M. Golparvar-Fard, A. Heydarian, and J. C. Niebles, "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers," *Advanced Engineering Informatics*, vol. 27, no. 4, pp. 652–663, 2013.

[10] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Automation in Construction*, vol. 122, p. 103482, 2021.

[11] B. Xiao and S.-C. Kang, "Development of an image data set of construction machines for deep learning object detection," *Journal of Computing in Civil Engineering*, vol. 35, no. 2, p. 05020005, 2021.

[12] D. Roberts and M. Golparvar-Fard, "End-to-end vision-based detection, tracking and activity analysis of earthmoving equipment filmed at ground level," *Automation in Construction*, vol. 105, p. 102811, 2019.

[13] J. Y. L. Kasahara, S. Jinhyeok, K. Ren, C. Shota, N. Keiji, C. Takumi, Y. Shingo, C. Kazuhiro, Y. Atsushi, and A. Hajime, "Action recognition of excavator with data augmentation of simulator-generated training data," *Journal of the Japan Society of Precision Engineering*, vol. 88, no. 2, pp. 162–167, 2022.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, 2017.

[15] A. Xuehui, Z. Li, L. Zuguang, W. Chengzhi, L. Pengfei, and L. Zhiwei, "Dataset and benchmark for detecting moving objects in construction sites," *Automation in Construction*, vol. 122, p. 103482, 2021.

[16] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5693–5703, 2019.

[17] H. Luo, M. Wang, P. K.-Y. Wong, and J. C. Cheng, "Full body pose estimation of construction equipment using computer vision and deep learning techniques," *Automation in Construction*, vol. 110, p. 103016, 2020.

[18] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2969–2978, 2022.