International Journal of Innovative Computing, Information and Control
Volume **20**, Number **1**, February **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **231–247**

ACTION RECOGNITION OF SIMULATED WORKPLACE WITH OCCLUSION BASED ON INTERPOLATED SKELETON DATA USING OPENPOSE

Hechen Yun[1], Etsuro Nakamura[1], Yoichi Kageyama[1,*], Chikako Ishizawa[1]
Nobuhiko Kato[2], Ken Igarashi[2] and Ken Kawamoto[2]

[1]Graduate School of Engineering Science
Akita University
1-1 Tegata Gakuen-machi, Akita-shi, Akita 010-8502, Japan
d8521052@s.akita-u.ac.jp; merumeru1116@gmail.com; ishizawa@ie.akita-u.ac.jp
*Correspongding author: kageyama@ie.akita-u.ac.jp

[2]ADK Fuji System Co., Ltd.
110-3 Tegata Yamazaki-machi, Akita-shi, Akita 010-0851, Japan
{ nobuhiko; igarashi; kawamoto }@adf.co.jp

Received April 2023; revised August 2023

Abstract. *As the Japanese construction industry sees a rise in the proportion of elderly workers, it becomes crucial to establish a management system that ensures their safety and health. Several action recognition methods have been proposed as key techniques for such systems, but real-world occlusions present a significant challenge. This study introduces an action recognition approach that leverages skeleton data and linear interpolation methods to address this issue. The proposed approach identifies and categorizes missing skeletal position values into three distinct scenarios. It then employs three types of interpolation methods to impute these missing values, considering the human body structure and temporal changes. To emulate real-world occlusions, four types of occlusion processes are applied prior to extracting skeleton data. Features such as skeleton coordinates, orientation, distance, and trajectory are calculated for training and validation of an LSTM-based classification network. This network is designed to recognize 13 different types of actions. Experimental results indicate that the proposed approach can accurately identify actions with an average accuracy of 76.35% from masked datasets, demonstrating its potential for recognizing work actions in occluded situations.*
Keywords: Action recognition, Skeleton, Machine learning, Occlusion, Interpolation method, OpenPose

1. **Introduction.** Recently, with the rapidly increasing aging problem in Japan, the population of middle-aged adults (aged 45-60) and older adults is growing at work sites in the Japanese construction industry, which has the highest proportion among other industries [1]. The decline in physical and mental functions with age [2,3] has affected the judgment of workers under unexpected situations such as slippery floors, differences in concrete segments, and protrusions. With the increased attention of modern employees to the safety and comfort of their work environment, developing a management system to establish such a work environment for them is required. In particular, the applications of such a management system would effectively optimize the total project cost [4] by assessing the level of risks at a workplace to restrict workers from making unsafe decisions and avoid accidents. Moreover, integration with fatigue risk management systems would

DOI: 10.24507/ijicic.20.01.231

improve the efficiency of workers by monitoring their current state to prevent long- and short-term fatigue on their physical and mental faculties [5].

Furthermore, in recent decades, surveillance devices have been more commonly integrated into workplaces and accepted by people; more and more surveillance cameras are being installed on work sites. With the increased installation of these cameras [6], substantial video data can be easily captured from workplaces and inputted into the management system. As a key part of the vision-based system, action recognition can be achieved using discriminative models when there is enough action data on learning and worker actions can be analyzed on the vision level. Following improvements in computer vision and deep learning, some companies and research groups worked on action recognition obtained excellent results and accuracy through video data [7-9]. In previous research, action recognition methods have been proposed based on vision information or extracted skeleton key-point coordinates. They were aimed at minimizing the lack of action information available due to the occlusion problem on construction sites [10,11]. In this study, occlusion is defined as a situation where the camera does not capture the human body entirely due to the existence of other objects or the position of self-paced limb motion. A skeleton-based action recognition strategy for extracting high-level features from the human body is proposed. The proposed strategy also includes the ability to discriminate patterns to avoid the effects of unnecessary information on the recognition results, such as the movement of objects or the colors of clothes [10,12].

The occlusion problem is one of the biggest challenges in action recognition for real-world situations. This problem could be caused by object interactions, illumination variances, viewpoint changes, and even simple actions. Some research groups on face recognition improved model performances by adopting some data augmentation methods, such as cropping parts of the face area on the preprocessing step and overlapping real-like mask templates on the corresponding position of the face [13]. A method for action recognition adopted data augmentation to alleviate the occlusion problems by mixing two samples at the token level and creating a larger dataset to train models [14]. However, data augmentation increases the computational cost by processing each frame of action videos to train action recognition models. A method for encoding the feature representation of videos before classification was proposed to enhance robustness under occlusion [15]. However, this method struggles to process the temporal information between the non-occluded and occluded environments. Meanwhile, skeleton-based action recognition methods to enhance the robustness of the noisy or incomplete skeleton data in occlusion situations were proposed. On the one hand, some research groups provided transformations from 2D to 3D skeleton data to address the limitations due to insufficient spatial information. One method tried to predict missing parts of the 3D human skeleton sequence by combining the skeleton with the surrounding situation using the attention mechanism [16]. Nevertheless, training models with image features of the surrounding situation may lead to negative effects when implemented at complex work sites, such as construction sites. Moreover, 3D sensors are not typically used in practice because of the restrictions on their performance owing to the camera viewpoint and environmental conditions. On the other hand, a method to improve the accuracy of the pose estimation model was proposed [17]. This method can indirectly impact the results of action recognition; however, it requires the support of substantial resources, datasets, and computational requirements, increasing the project cost. A 2D skeleton-based method using the generation model to interpolate the missing values in the skeleton coordinates approach was proposed [18]. Unfortunately, the complicated matrix-generating process in this method makes it difficult to deploy it in an actual workplace with real-time feedback. Moreover, the validity of interpolation methods to extrapolate the missing data values has not been discussed sufficiently yet.

Due to the above factors, in this study, we propose an approach for recognizing human actions based on interpolated 2D skeleton data to be used as part of a management system at construction sites. In the proposed method, the actions of daily workers on a construction site were simulated and treated as subjects. The skeleton data were extracted from the video data using OpenPose, and three interpolation methods were used to complement the motion information of the actions. Furthermore, features such as joint-joint orientation, distance, and frame-to-frame trajectory were calculated from the interpolated skeleton data. The validity of this recognition model was discussed to improve its accuracy. For evaluating the performance of the proposed approach, some visual occlusion generation methods were used to simulate real-world situations in a complex construction workplace with occlusion. Additionally, to evaluate the feasibility of the approach, we compared the evaluated results with those from previous studies. The comparison showed that our approach provided good performance and could solve occlusion problems better in complicated work environments. Our contributions are summarized as follows.

- We present a new skeleton-based action recognition approach to deal with occlusion problems on construction sites.
- Four kinds of occlusion processes are considered to simulate the real-world situations of construction work sites where workers are only partially visible.
- The missing detection method and three kinds of interpolation methods are proposed to assign the missing position values of the skeleton estimated using OpenPose.
- Our proposed approach demonstrates the potential to recognize work-like actions even in occlusion situations.

This paper is structured as follows. Section 2 describes the data acquisition and occlusion processing with four kinds of methods to simulate real-work situations. Section 3 presents the entire process of the proposed method, including skeleton data extraction, skeleton interpolation with linear interpolation methods, feature extraction, dataset creation, and the construction of an action classification network based on LSTM. The details of the experiments are presented in Section 4, and the results from experiments are presented and discussed in Section 5. Section 6 concludes the work.

2. Materials and Preprocess.

First, we describe the data acquisition method. Next, we describe the occlusion process.

2.1. Data acquisition.

Figure 1 shows the data acquisition environment. Videos of 10 subjects with 13 types of simulated workplace actions were captured from three cameras, with each action captured at least three times. In particular, 5,515 action videos (59.94 fps, 1920 $\times$ 1080 pixels) were acquired and converted to sequence images as experimental data, with a 29.97 fps rate. The action definitions and the number of videos in each category are listed in Table 1. Three monocular video cameras (SONY: FDR-AX60, PANASONIC: HC-VX2M $\times$ 2) were used for data acquisition. The cameras were set in the front, behind, and at the side; these were the relative positions of the subjects at the preparation time. For the action categories without position changes, subjects stood at the specified center point and performed actions. Walking actions were captured when the subjects moved between the front and back points and between the two side points. The data used in this investigation were acquired in accordance with the ethical regulations concerning studies involving humans at Akita University, Japan.

2.2. Occlusion process.

Considering that the pose estimation technique was robust enough to avoid occlusion, different masks were added to the image that simulated actual skeleton extraction in real-world situations. Inspired by Fong and Vedaldi [19], in this study, three different types of mask-creation methods were adopted to simulate situations

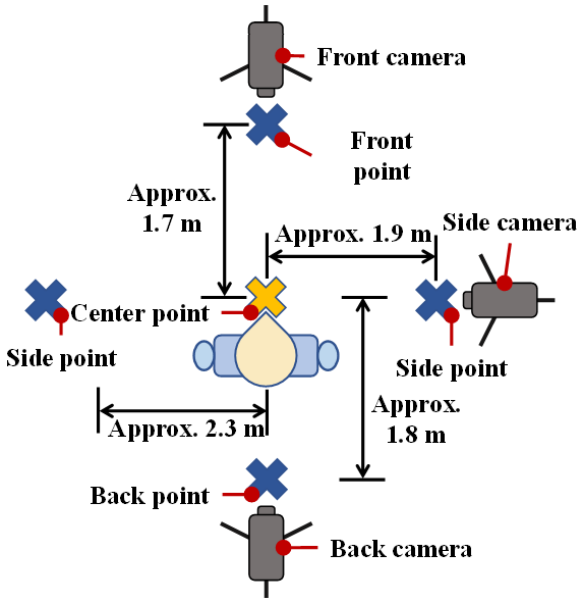


FIGURE 1. Data acquisition

TABLE 1. Information in the data

Action name	Avg. time/sec.	Avg. frames	Count	Details
Walk	5.32	159.53	504	Walking at a constant speed while waving hands.
Sit down Sit up	2.28 1.90	68.42 57.00	120 120	Sitting in a chair with a backrest, placing hands on legs, and returning to standing position.
Climb down Climb up	3.97 4.23	119.08 126.82	208 208	Climbing down a stepladder step-by-step to reach the ground and climbing back up step-by-step.
Squat down Squat up	1.81 1.66	54.25 49.86	123 123	Sitting with knees bent and heels close to buttocks and returning to the standing position, where the back and legs are straightened.
Take up Throw	2.01 1.98	60.14 59.41	254 252	Lifting a large object above the head and dropping it forward or behind the body.
Tumble down Tumble up	3.92 3.00	117.49 89.89	368 368	Tumbling slowly into a mattress, relaxing the body, getting up from the mattress, and returning to the standing position.
Pick up Put down	2.84 2.97	85.17 88.88	1438 1429	Picking up objects from the ground or table using hands and putting them back.

with occlusions. Body masks were created according to the bounding box, performing calculations from the skeleton data without occlusion. For covering entire body parts, including muscles and clothes, the edges of the top and bottom of the bounding box were extended by 10 pixels, and those of the left and right were extended by 19 pixels. Masks were added to the top, middle, and bottom of the bounding box, and the size of the masks was changed, corresponding to 30%, 50%, and 70% of these regions of the box, respectively. Furthermore, the "Hide-and-Seek" [20] and "Cutout" [21] techniques were used to introduce the mask generation methods of stochastic input-level occlusions that

could improve the robustness of deep learning models. Moreover, we considered that those methods could estimate the performance of the pose estimation technique and the action recognition method as well. Therefore, using the Hide-and-Seek method, the images were divided into 36 equal regions, and 12 regions were randomly masked for each video frame. In addition, 12 square regions of sequence images – which randomly selected 12 positions in image size for one sequence frame of actions as the left-top point of the square – were masked using the Cutout method. Subsequently, masks with a fixed size (200 × 200 pixels) were added. Additionally, horizontal masks were added in the middle of the images with 270 pixels in width. Here, the mask was a region (RGB: 0, 0, 0) in the image. Figure 2 and Figure 3 show examples of masked images.

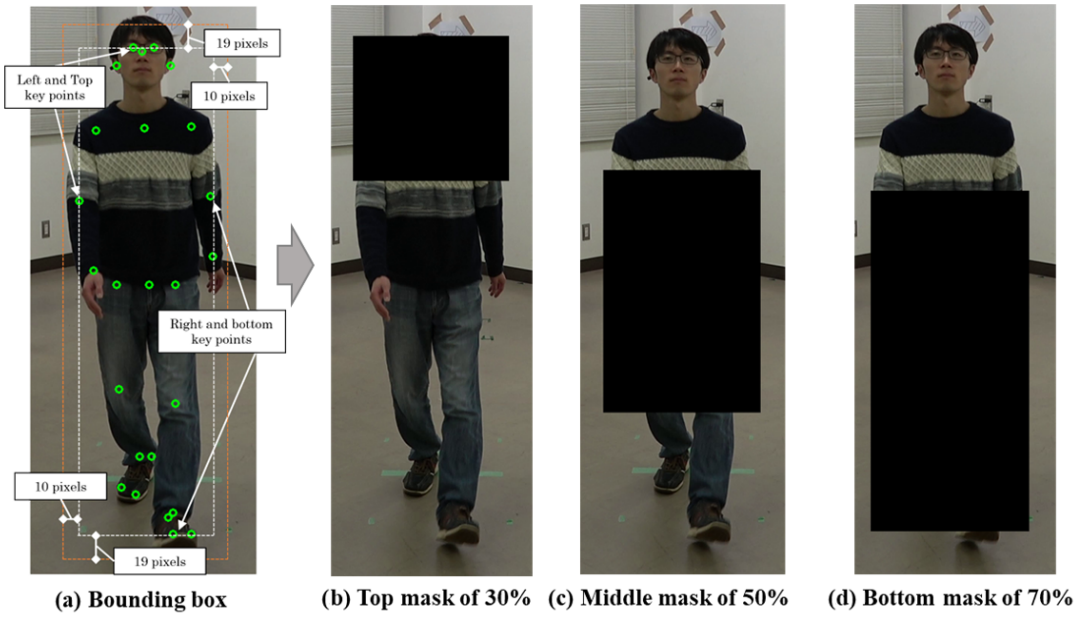


FIGURE 2. Example of top, middle, and bottom of the bounding box with 30%, 50%, and 70% proportions

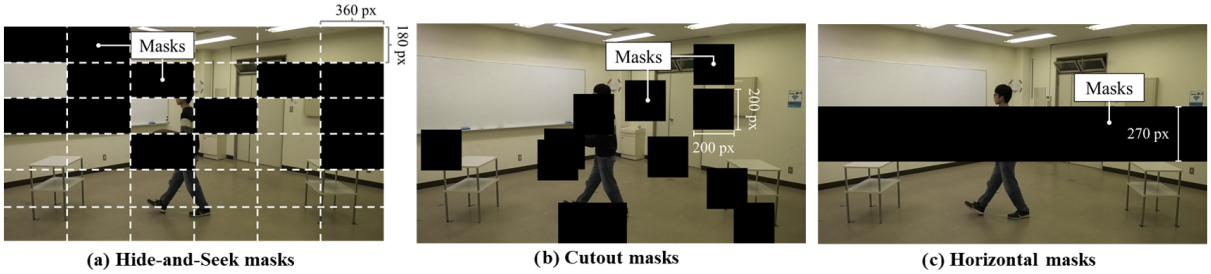


FIGURE 3. Example of Hide-and-Seek, Cutout, and horizontal masks

3. **Proposed Method.** In this study, an action recognition approach based on OpenPose was proposed. This can be used as an essential part of a management system that can automate the monitoring of construction workers to prevent accidents from happening and improve work efficiency. Figure 4 presents an overview of the proposed method. The proposed approach is divided into the following steps: 1) extracting skeleton data from the sequence images of each action; 2) interpolating the missing data of the key points based on three interpolation methods; 3) calculating the four types of action features using

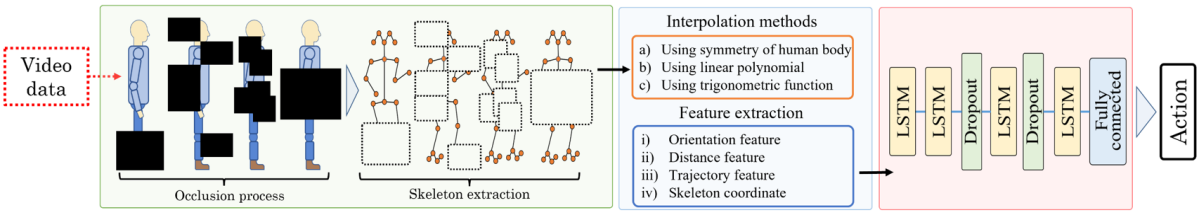


FIGURE 4. Overview of the proposed method

the imputed data; 4) establishing and training the long short-term memory (LSTM) classification model using the imputed data and action features; and 5) predicting the actions of the subjects according to the trained model.

3.1. **Skeleton data extraction.** The OpenPose library and BODY_25 [22] model were adopted as post-estimation techniques to obtain 25 skeletal key points with coordinates and the confidence scores of the estimations. However, owing to the camera's viewpoint and the occlusions, parts of the subject's body were hidden and became undetectable. In this study, the missing detection method is used to search for and record the missing key points from extracted skeleton data using OpenPose at the preprocessing step. Based on our observations, OpenPose tries to predict the key points' position of missing parts when it detects a partially invisible human body, based on the detectable existing vision information and before-learned human body structure, and then outputs the prediction probability as a confidence score from 0.0-1.0. The confidence score of the key point is higher than 0.7 when that body part is completely visible or only slightly obstructed by non-disturbing objects, such as thin lines. Conversely, the confidence score of the key point is less than 0.7 when a body part is somehow invisible, which may lead to occlusion, low resolution, and an unnormal body pose, and then the coordinate of the key point is possibly obtained by a guess value or zero. Consider that the prediction of OpenPose has certain validity when the confidence score is between 0.3-0.7, in which the coordinate of the key point is guess value, and we only want to deal with the missing parts of the skeleton; key points will be recorded when the coordinates are zeros or the confidence score is less than 0.3, which means that part of the body was completely invisible and unpredictable for guess value using OpenPose. Then, the recorded missing key points will be categorized and interpolated by the relevant interpolation method. Those key points' probabilities higher than 0.3 will be ignored from interpolation processes and directly used to calculate features.

3.2. **Skeleton interpolation.** To reduce the influence of the missing points, we proposed an interpolation procedure with three steps to process the row skeleton data. The interpolation procedure was used to detect the missing points and estimate the values at each step accordingly.

1) To interpolate the points of the symmetric joints (e.g., those of the hands and feet), the coordinate values of the missing points are interpolated by the corresponding points, which can be detected at the position of the same limb on the other half of the body.
2) To interpolate the asymmetric missing points, wherein limited frames n with missing points exist between the two frames of the detectable points P, the coordinate values of the missing points P_m are interpolated using linear interpolation, assuming that the coordinate values increase with constant velocity within the detectable first and second frames. For solving the above situation, the missing points are calculated using

Equation (1):

$$P_m(t+i) = P(t) + i\frac{P(t+n) - P(t)}{n}, \quad 0 < i < n,\ n \in T. \tag{1}$$

Moreover, linear interpolation is used to calculate the values of the missing points for which the frames before or after can only be detected by adding values (positive or negative depending on the time-series variation of motion) obtained from five detectable frames before or after the missing points. The missing points are calculated using Equations (2) and (3):

$$P_m(t+i) = P(t) + i\frac{P(t+4) - P(t)}{4}, \quad 4 < i < T, \tag{2}$$

$$P_m(t-i) = P(t) + i\frac{P(t) - P(t-4)}{4}, \quad 4 < i < T, \tag{3}$$

where T is the number of total frames in each action, t is the position number of the last detected frame before the frames with the missing points, $t+n$ is the position number of the first detected frame after the frames with the missing points, and i is the position number of the frames that are counted from frame t.

3) When a situation arises wherein the subject's back is turned to the camera, the facial information of the subject cannot be detected. According to practical observations of the human face structure, the nose, ears, and eyes positions form a right-angled triangle on the left and right sides of the face. Therefore, this structure relationship is used to interpolate when the missing points are both the eyes and nose. In this study, the positions of the eyes are assumed to always be at the upper of the ears and nose in two-dimensional space according to common sense of work actions. Figure 5 shows an example of the performed face interpolation process.

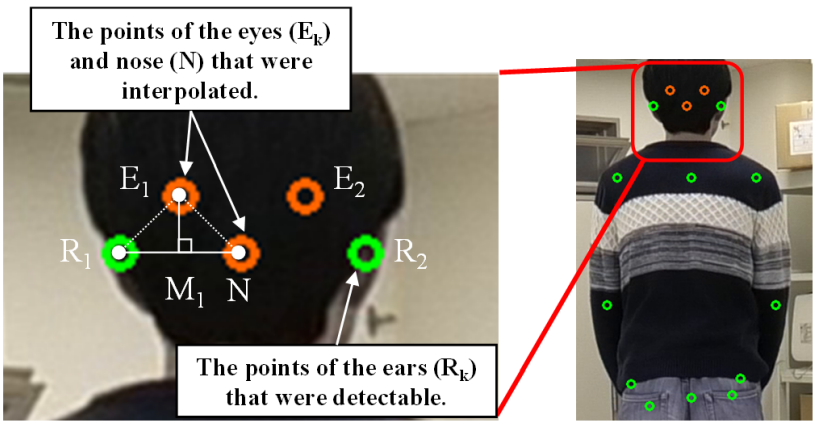


FIGURE 5. Example of the face interpolation process (Orange points: interpolated; Green points: detectable)

The distance between the two ears is assumed to have four equal lengths m, which are the lines R_1M_1, R_2M_2, M_1N, and M_2N, and the point of the nose is set in the middle, between R_1 and R_2. Lines E_1M_1 and E_2M_2 are assumed always having a vertical relationship with R_1N and R_2N, respectively. Each point of the ear and eye forms a triangle, with the nose point on one side. Points M_k were set as the connecting points of the line perpendicular to the edge from point R_k to N, and the coordinates were set as $\left(\frac{x_{R_k}+x_N}{2}, \frac{y_{R_k}+y_N}{2}\right)$. Based on the property of perpendicularity of two-dimensional vectors and the assumption that eyes E_k are always above lines R_1N and R_2N, the

vector $\overrightarrow{E_kM_k}$ can be calculated by swapping the components of the unit vector $\hat{u}_{\overrightarrow{R_kM_k}}$ and setting a minus signal at the y axis, and then multiplying the corresponding magnitude, which is of length m. Equations (4) and (5) were used to compute the unit vector $\hat{u}$ of line R_kM_k, and the Euclidean lengths of lines E_kM_k from the assumed right triangle, respectively:

$$\hat{u}_{\overrightarrow{R_kM_k}} = \left(\frac{x_{\overrightarrow{R_kM_k}}}{\sqrt{x^2_{\overrightarrow{R_kM_k}} + y^2_{\overrightarrow{R_kM_k}}}}, \frac{y_{\overrightarrow{R_kM_k}}}{\sqrt{x^2_{\overrightarrow{R_kM_k}} + y^2_{\overrightarrow{R_kM_k}}}} \right), \tag{4}$$

$$\overrightarrow{E_kM_k} = \left(my_{\hat{u}_{\overrightarrow{R_kM_k}}}, -mx_{\hat{u}_{\overrightarrow{R_kM_k}}} \right), \tag{5}$$

where k is the side id (1 means left and 2 means right) of the assumed triangle on the face, $\hat{u}$ is the unit vector of non-zero vector in the direction $\overrightarrow{EM}$ normalized by the length of $\overrightarrow{EM}$, which is computed by $\left(\frac{x}{\sqrt{x^2+y^2}}, \frac{y}{\sqrt{x^2+y^2}} \right)$ in the Cartesian coordinate system. $x_{\overrightarrow{R_kM_k}}$, $y_{\overrightarrow{R_kM_k}}$, $x_{\hat{u}_{\overrightarrow{R_kM_k}}}$ and $y_{\hat{u}_{\overrightarrow{R_kM_k}}}$ mean they are coordinate values x and y of vector $\overrightarrow{R_kM_k}$ and unit vector $\hat{u}_{\overrightarrow{R_kM_k}}$, respectively. m is one-quarter the length between the two points of ear R_1 and R_2.

3.3. **Feature extraction.** Because the proposed network could learn the pattern of actions efficiently from a wider aspect, four types of features (coordinates, orientation, distance, and trajectory) were extracted to quantify the action information using interpolated skeleton data.

3.3.1. *Coordinate features.* The coordinate values (P^x, P^y) of 25 interpolated key points were used to represent the position of the limbs and trunk to obtain the basic spatial information from the skeleton. Furthermore, orientation, distance, and trajectory features were extracted based on these coordinate values. Specifically, if the value remained at zero even after processing with the interpolation methods, the value of each feature related to it would also be set to zero.

3.3.2. *Orientation features.* To quantify the spatial information related to the relative positions of the limbs and trunk depending on the type of action, the coordinates of the skeletal key points were used to extract the radian values as orientation features. Figure 6(a) shows an example of this process where the origin lines are the sublines of the angle calculation. We decided upon the neck position (point No. 1) as the axis point, and the angles between the other points G from the key points 1, 6, and 4 of the head, hands, and legs, respectively, were calculated as follows:

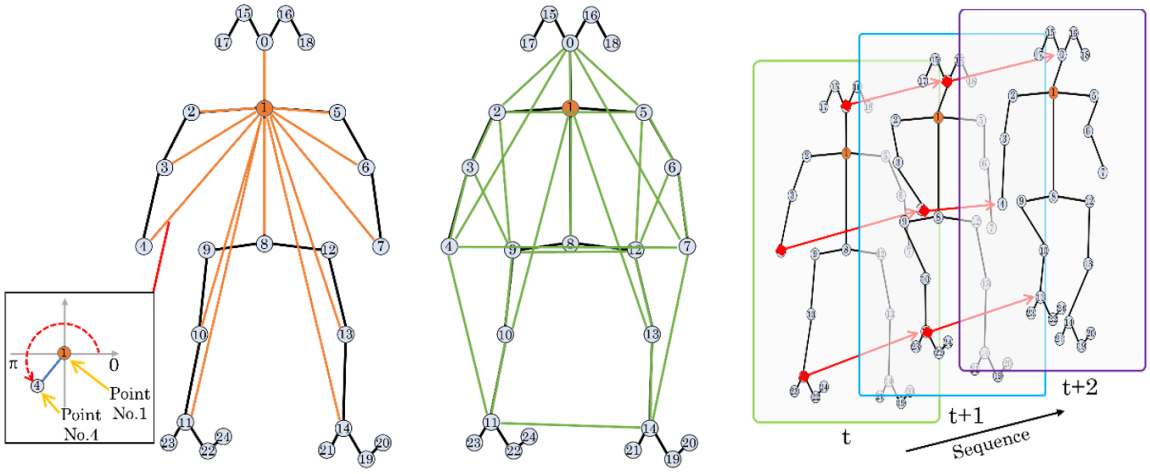
$$f^{ori} = \arctan\left(P^y_i - P^y_1 / P^x_i - P^x_1 \right), \quad i \in G, \tag{6}$$

where the orientation feature f^{ori} is calculated using the coordinate (P^x_i, P^y_i), and is used to represent the spatial information between the neck and each limb in a single frame.

3.3.3. *Distance features.* The distance changes between some parts of the body represented the features of the actions in a spatial aspect. To extract them, we selected some groups with significant changes in the time sequence between points No. 0 and No. 14. Figure 6(b) shows a representative example of this process, where the green lines are the distance edges between the selected key points. Each distance was calculated as follows:

$$f^{dis} = \left| \sqrt{\left(P^x_j - P^x_k \right)^2 + \left(P^y_j - P^y_k \right)^2} \right|, \quad j \in V,\ k \in E, \tag{7}$$

where V and E are the key points in the selected groups. Moreover, the distance of the edge f^{dis} between the two points with the coordinate values (P_j^x, P_j^y and P_k^x, P_k^y) is computed to learn the action's patterns in a single frame.



(a) Orientation feature (b) Distance feature (c) Trajectory feature

FIGURE 6. (color online) Example of features based on the skeleton data. The circles with numbers are the key points representing the features coordinates.

3.3.4. *Trajectory features.* Each action had a similar movement pattern within a short time, even if different people performed the action. The trajectory feature was extracted from the key points as temporal information on the action from two continuous frames. The coordinate change between two frames with the same key point was calculated as a distance vector according to Equation (7). Figure 6(c) shows an example of this process, where the red vectors are the distance vectors with the key points in order of the frame sequence.

3.4. **Dataset creation.** For training the model, samples of the dataset were obtained from sequence-image data by scanning 60 consecutive frames and shifting one frame in the order of the time-series stream. Additionally, each sample included 60 consecutive extracted frames. Then, four types of action features (113 features) were calculated from them. Furthermore, for evaluating the performance of the proposed method and the effects of interpolation methods and features applied in the approach, four extra datasets were created based on one unmasked dataset by removing the interpolation methods and the 113 action features one by one to investigate the contribution of each component. Finally, four unmasked datasets and 12 masked datasets were obtained and used in the experiments.

3.5. **LSTM-based classification network.** This study implemented a single-channel LSTM network to learn temporal information on actions from the variance in the coordinates and features. Structurally, the model had four LSTM layers in its middle layers, with L1 and L2 regularization for improving the performance of the model. The Adam optimizer [23] was adopted as the optimization function for processing the model. Samples of the dataset were adopted as inputs for the classification model, which were then used to recognize the actions.

4. **Experiments.** In this investigation, both masked and unmasked datasets were employed. The dataset from data collection that was processed using the combination of approaches in the suggested approach corresponded to four unmasked datasets. The 12 masked datasets created from 12 types of occlusion-processed data to simulate various possible situations on construction sites and the proportions of top, middle, and bottom type masks were defined as '0.3', '0.5', and '0.7' followed by the mask categories. Each dataset had the same 13 actions, and each sample with the extracted features had only one class. The nine subjects of the samples in the datasets were used to train the network, and one subject was used for validating the training process. The dataset sample was recognized based on the classes of the actions for each sample, using the trained network.

In this study, three different experiments were conducted to evaluate the performance of the proposed approach, including 1) estimation of interpolation methods; 2) training performance on unmasked datasets; and 3) recognition performance of masked datasets.

Based on our limited knowledge, it is challenging to find suitable research as a comparison method that targets interpolating the missing parts of the skeleton with the whole sequence at a two-dimensional level. The I3D model (Inflated 3D ConvNets) [24] is an RGB-based state-of-the-art approach that demonstrates good performance under diverse environments to recognize actions based on a new two-stream inflated 3D ConvNets that could be used to learn spatio-temporal features. We believe the I3D model has enough ability to recognize actions on complex work sites at the RGB level. Therefore, we chose the I3D model to compare with the proposed approach under masked datasets.

In the experiments, the I3D model was set up with ResNet-50 as its backbone and already pretrained by an open-action video dataset named Kinetics 400 [25]. To recognize the 13 classes of action that were obtained during data acquisition, each action sample video had 60 frames extracted as one dataset sample and resized in advance from 1920 $\times$ 1080 pixels to 224 $\times$ 224 pixels to fit the model's requirements. For samples with more than 60 frames, they were clipped evenly with temporary frames; for samples with less than 60 frames, they were filled with the nearest valid frame corresponding to the empty frame position at equal intervals.

The proposed network was built using TensorFlow 2.1.0 with Keras 2.3.1, and the comparison approach was constructed using the deep learning toolkit GluonCV [26] and PyTorch 1.6.0.

5. Results and Discussion.

5.1. Estimation of interpolation methods.

For better evaluating and analyzing the performance of the proposed interpolation methods, the situations of missing key points were quantified by comparing the original and masked datasets. Table 2 lists the estimation results of missing key points on each type of dataset that was created from occlusion processes. Each dataset had the same number of action samples (4,840) processed by interpolation methods. The key points that were not detected by the missing detection method were ignored and not counted in this estimation. The table shows that the missing rate of key points achieved a significantly low level, regardless of whether the human bodies were visible or occluded in samples, and the number of detectable key points under occlusion situations exceeded our expectations. We believe it benefited from the powerful pose estimation technique of OpenPose, which can predict the position of invisible key points by using a large action database in the pre-training stage.

In contrast, 12 types of mask processing indeed affect the performance of key point estimation. Following the increase in proportions in the top, middle, and bottom mask datasets, the missing rate of key points also got higher than before, especially in the

TABLE 2. Estimation results of interpolation methods for each dataset

Dataset category	Interpolated action samples	Number of missing points	Number of interpolated points	Missing rate [%]	Interpolation rate [%]
Unmasked	4,018	1,366,041	226,971	0.025	16.615
Bottom0.3	4,753	1,949,147	664,962	0.035	34.115
Bottom0.5	4,835	3,097,279	1,052,662	0.057	33.986
Bottom0.7	4,835	4,179,687	1,284,275	0.077	30.726
Middle0.3	4,613	1,633,358	481,946	0.030	29.506
Middle0.5	4,699	1,868,834	677,176	0.034	36.235
Middle0.7	4,822	3,187,522	1,625,595	0.059	50.999
Top0.3	4,797	1,679,493	713,375	0.031	42.476
Top0.5	4,818	2,134,969	1,131,844	0.039	53.014
Top0.7	4,833	3,675,176	2,350,305	0.068	63.951
Horizontal mask	4,779	1,927,421	751,634	0.035	38.997
Hide-and-Seek mask	4,840	2,328,480	1,403,359	0.043	60.269
Cutout mask	4,619	1,694,945	564,594	0.031	33.310

'Bottom0.7' dataset, which achieved the highest value of 0.077% compared to others. Moreover, the interpolated number of key points also increased when much more areas of the body became invisible. Especially, the 'Top0.7' mask dataset got the highest interpolation rate by 63.951%. We think this is because the pose estimation model is good at detecting the leg parts of the body and recognizing them as humans, then providing the relevant key points of the leg parts. It is also giving enough information to interpolation methods that can estimate the position values of missing parts based on that and interpolate facial and temporal information. However, comparatively, in bottom mask datasets, the interpolation rate did not increase with the rise in the missing rate because it was hard for OpenPose to detect the object as human in samples and provide skeleton information to interpolation methods.

5.2. **Training results for unmasked datasets.** Table 3 lists the training results of recognition models on four types of unmasked datasets to investigate the effects of interpolation methods and calculated features from the proposed approach. The I3D model was also trained on the same types of unmasked datasets for 200 epochs to compare with the proposed approach. Each target subject was used as validation data, and residual subjects were used to train the recognition models, then shifted to the next subject as validation data. Each training procedure was individual. The table gives the highest validation accuracy of one epoch step of the training processing.

The results show that the recognition model of the proposed approach had good training performances on each type of unmasked dataset, with an average accuracy of over 90% for each model, compared with the I3D approach, which had an average accuracy of 82.53%. We think this is because the proposed approach can get essential information about motion patterns from skeleton data, which makes the model stable and easy to achieve convergence. Compared with that, I3D is a large model that needs more training steps and epochs to extract important features from RGB-based information and is constrained by multiple factors, like the color of cloths and scale of subjects.

In contrast, models trained on datasets that only included interpolation methods or calculated features, or neither, performed better than models trained on datasets that included the proposed approach, especially when subject No. 1 was used as validation data

TABLE 3. Training results for unmasked datasets [%]

Validation of subject No.	Coordinate	Coordinate + Interpolation	Coordinate + Features	Coordinate + Interpolation + Features	I3D (RGB-based)
1	97.12	96.39	96.88	88.70	81.84
2	84.60	85.49	83.04	78.79	83.05
3	99.33	99.11	99.11	98.44	82.01
4	99.79	99.37	98.96	99.17	85.69
5	98.63	98.24	99.02	97.07	85.14
6	99.17	98.54	99.37	95.83	81.07
7	99.79	99.58	99.79	98.96	80.28
8	98.12	98.96	98.96	94.58	82.44
9	99.79	98.96	98.75	94.17	82.32
10	97.71	98.96	98.12	98.33	81.50
Average	**97.41**	97.36	97.20	*94.40*	82.53

their performances decreased by about 8%. We believe that the suggested strategy will offer unnatural motion information for model training when the missing rate of important points is at low levels, and that having more features will result in a model that is heavier in scale and more difficult to converge within short epochs than having fewer features. Moreover, according to our observations, the personal habits of subject No. 1 in captured action videos are different from those of other subjects, such as hunchback motions, which make those samples hard to recognize when the model cannot learn from the dataset.

5.3. **Recognition results for masked datasets.** Occlusion processing was conducted in this study to assess the performance of the proposed approach and the comparison method (I3D) when faced with various situations on construction sites. All recognitions in this experiment were trained on an unmasked dataset to evaluate the robustness of occlusion situations.

For better comparison of the differences between each approach, the basic approach was also trained with the same training procedures but only with the coordinates of the skeleton key points. For selecting a model that has both versatility and usability, two models were used in this experiment based on the training performance. One is the highest validation accuracy within 200 epochs (Max of validation), using subject No. 4 as validation data. Another is the highest average validation accuracy between 11-200 epochs (Max average of validation), with subject No. 10 as validation data, which means it is more stable for recognition. Table 4 lists the recognition results of the proposed approaches and the comparison method for 12 types of masked datasets.

The results show that skeleton-based approaches have better recognition accuracy compared with RGB-based I3D in this experiment, which includes a basic approach and proposed approaches. Furthermore, the Max of validation of the proposed approach increased by 1.77% compared with the basic approach, which only trained on and recognized coordinate values of key points. The basic approach obtained higher accuracy when the proportions of 30% of body positions were higher than the proposed approach, but got decreased significantly when the proportion of masks became higher. Compared with that, the proposed approach has the ability to alleviate the effects of masks and mitigate the trend of decreasing accuracies, such as 'Bottom0.3' to 'Bottom0.5' and 'Middle0.3' to 'Middle0.5'. Moreover, the model from Max of validation has better recognition performances on each mask dataset without 'Bottom' types of masks compared with the Max

TABLE 4. Recognition results for masked datasets [%]

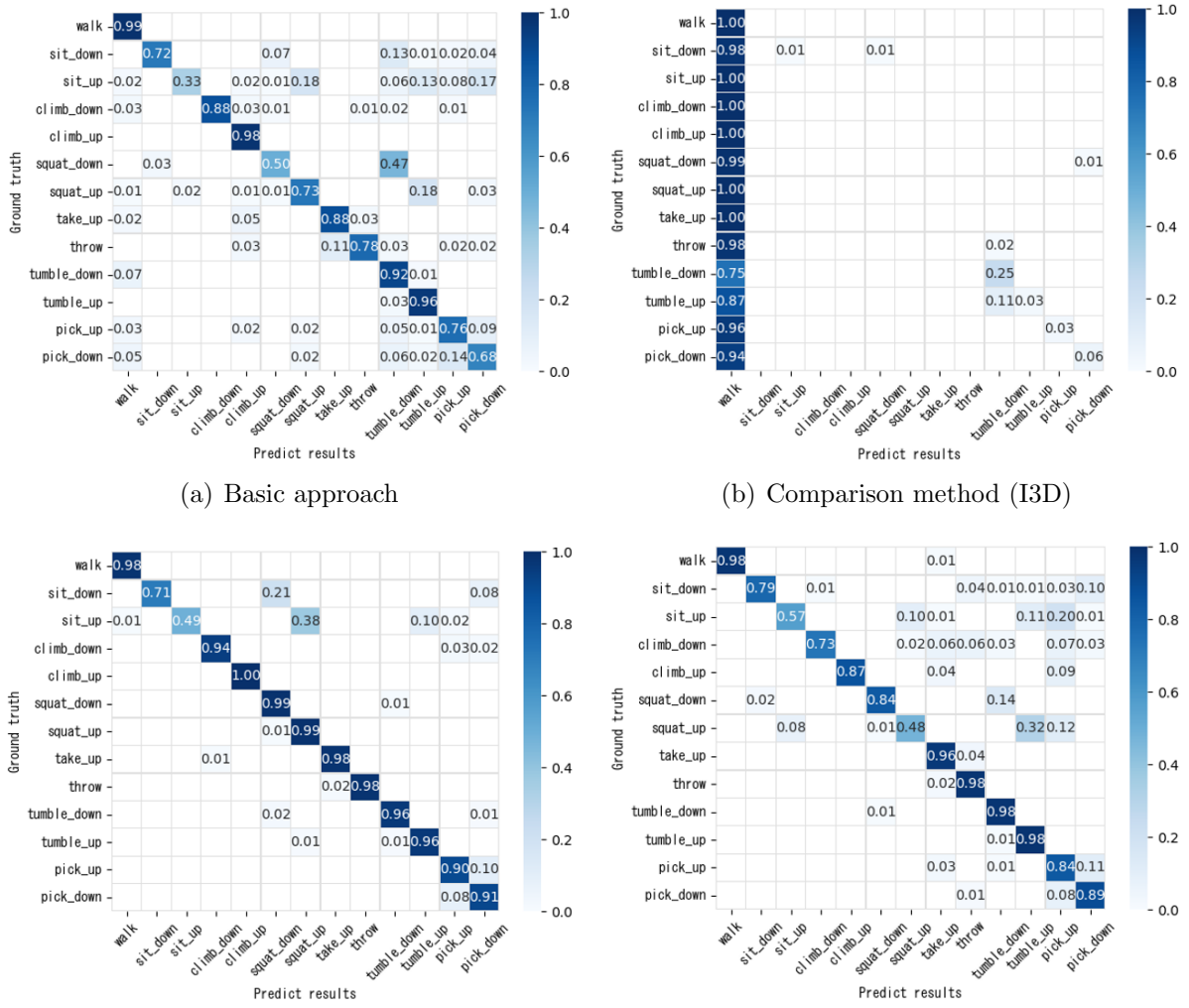
Dataset category	Basic approach (Coordinate only)	Proposed approach (Max of validation)	Proposed approach (Max average of validation)	Comparison method (RGB-based I3D)
Unmasked	**96.42**	94.88	93.43	82.53
Bottom0.3	95.14	86.53	90.50	80.85
Bottom0.5	90.90	84.62	85.94	77.23
Bottom0.7	61.60	60.00	60.78	72.64
Middle0.3	87.17	82.19	80.40	78.31
Middle0.5	73.22	67.86	67.33	68.72
Middle0.7	49.94	59.11	52.83	54.83
Top0.3	90.54	91.10	85.10	53.00
Top0.5	69.61	74.32	71.05	46.28
Top0.7	*31.00*	*55.72*	45.99	42.23
Horizontal mask	74.88	73.09	73.33	12.05
Hide-and-Seek mask	77.79	92.01	87.77	51.65
Cutout mask	93.18	89.71	88.47	75.64
Average of masks	74.58	**76.35**	74.12	59.45

*All models were trained on unmasked dataset.

average of the validation model. Therefore, those results show that the proposed approach has the potential to alleviate the effects of occlusion problems and maintain recognition performance at the construction worksite.

Figure 7 shows the normalized confusion matrix of the recognition results from the basic approach, the proposed approach, and the comparison approach when recognizing the masked dataset of the Hide-and-Seek mask. The comparison between Figure 7(a) and Figure 7(b) indicated that the comparison approach barely classified 'Walk' actions that only detected the position changes of the subjects. The reason might be that extracting sufficient action information from time-series images without coherent changes using the RGB-based I3D model was difficult. Compared with that, skeleton-based approaches, including the basic and the proposed approaches, can achieve high accuracy even under stochastic occlusions. Figure 7(c) indicated that the proposed approach with Max of validation as selection measure obtained stable recognition results over all action classes. It benefits from linear interpolations in the proposed approach that make the model possible to be trained on time-sequential features with a stable variance of action patterns. Nevertheless, samples of 'sit down' and 'sit up' were hard to recognize compared with Figure 7(d), which is the model with Max average of validation as the selection measure. We consider that it is because the model of Max of validation cannot recognize the differences between 'sit', 'climb', and 'squat' when subject No. 4 is used as validation data, which includes some personal habits as hands-up before each action. It seems like the proposed approach can be trained with higher versatility when selecting the model using Max average of validation as a measure.

However, the proposed approach is hard to recognize the samples under 'Bottom' masks. We think it is due to the fact that the proposed interpolation methods are only focused on time series and facial parts but do not have optimization algorithms for bottom parts of human bodies. Moreover, the extra features calculated from missing key points also become a kind of noise for models during the training and recognition procedures.



(a) Basic approach

(b) Comparison method (I3D)

(c) Proposed approach (Max of validation)

(d) Proposed approach (Max average of validation)

FIGURE 7. Normalized confusion matrix of basic approach, the proposed approaches and comparison approach on the Hide-and-Seek mask dataset

Therefore, to improve the recognition accuracy when the bottom parts of the body are occluded, it is necessary to add new interpolation algorithms for estimating the missing key points.

6. **Conclusion.** In this study, we proposed a skeleton-based action recognition approach to classify the working actions simulating situations on construction sites. The proposed approach estimates the missing information of occluded human bodies by three types of interpolation methods and learns action patterns from the coordinates of the skeleton with four calculated features. Furthermore, we designed 12 types of masks to simulate the occlusion situation that could potentially happen at a real construction worksite. The recognition model was trained on an unmasked dataset, and masked datasets were recognized to estimate the performance of the proposed approach. The results provided the following conclusions.

1) The proposed approach shows the potential to recognize working actions by interpolating missing skeleton information at a real construction worksite when occlusion happened.

2) The proposed approach achieved the average accuracy of 76.35% for all masked datasets. These performances are better than the comparison method I3D by 16.90% and the coordinate-only basic approach by 1.77%.

In the future, new interpolation algorithms for the bottom parts of the human body will be considered for specific occlusions. Moreover, the usability and versability of the proposed approach will be investigated by acquiring working videos from actual construction worksites.

REFERENCES

[1] Ministry of Land, Infrastructure, Transport and Tourism, Japan, *Current Status of the Construction Industry and Construction Workers*, https://www.mlit.go.jp/common/001180947.pdf, Accessed on 04 April, 2023.

[2] I. J. Deary, J. Corley, A. J. Gow, S. E. Harris, L. M. Houlihan, R. E. Marioni, L. Penke, S. B. Rafnsson and J. M. Starr, Age-associated cognitive decline, *British Medical Bulletin*, vol.92, no.1, pp.135-152, DOI: 10.1093/bmb/ldp033, 2009.

[3] J. F. Sallis, Age-related decline in physical activity: A synthesis of human and animal studies, *Medicine & Science in Sports & Exercise*, vol.32, no.9, pp.1598-1600, DOI: 10.1097/00005768-200009000-00012, 2000.

[4] J. Yang, M. W. Park, P. A. Vela and M. Golparvar-Fard, Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future, *Advanced Engineering Informatics*, vol.29, no.2, pp.211-224, DOI: 10.1016/j.aei.2015.01.011, 2015.

[5] M. Sprajcer, M. J. W. Thomas, C. Sargent, M. E. Crowther, D. B. Boivin, I. S. Wong, A. Smiley and D. Dawson, How effective are Fatigue Risk Management Systems (FRMS)? A review, *Accident Analysis & Prevention*, vol.165, 106398, DOI: 10.1016/j.aap.2021.106398, 2022.

[6] Yano Research Institute Ltd., *Global Surveillance Camera Systems Market: Key Research Findings 2021*, https://www.yanoresearch.com/en/press-release/show/press_id/2813, Accessed on 04 April, 2023.

[7] K. Muhammad, A. Ullah, A. S. Imran, M. Sajjad, M. S. Kiran, G. Sannino and V. H. C. de Albuquerque, Human action recognition using attention based LSTM network with dilated CNN features, *Future Generation Computer Systems*, vol.125, pp.820-830, DOI: 10.1016/j.future.2021.06.045, 2021.

[8] S.-H. Kim and D. Cho, Viewpoint-aware action recognition using skeleton-based features from still images, *Electronics*, vol.10, no.9, 1118, DOI: 10.3390/electronics10091118, 2021.

[9] C. Leng, Q. Ding, C. Wu and A. Chen, Augmented two stream network for robust action recognition adaptive to various action videos, *Journal of Visual Communication and Image Representation*, vol.81, 103344, DOI: 10.1016/j.jvcir.2021.103344, 2020.

[10] Z. Li and D. Li, Action recognition of construction workers under occlusion, *Journal of Building Engineering*, vol.45, 103352, DOI: 10.1016/j.jobe.2021.103352, 2022.

[11] S. Yamanaka, C. Lee and S. Date, A parallel LSTM-based missing body feature point completion in video frames, *2019 International Conference on Computational Science and Computational Intelligence*, pp.646-651, DOI: 10.1109/CSCI49370.2019.00121, 2019.

[12] F. Angelini, Z. Fu, Y. Long, L. Shao and S. M. Naqvi, 2D pose-based real-time human action recognition with occlusion-handling, *IEEE Transactions on Multimedia*, vol.22, no.6, pp.1433-1446, DOI: 10.1109/TMM.2019.2944745, 2020.

[13] R. J. Salim and N. Surantha, Masked face recognition by zeroing the masked region without model retraining, *International Journal of Innovative Computing, Information and Control*, vol.19, no.4, pp.1087-1101, DOI: 10.24507/ijicic.19.04.1087, 2023.

[14] Z. Xing, Q. Dai, H. Hu, J. Chen, Z. Wu and Y. Jiang, SVFormer: Semi-supervised video transformer for action recognition, *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.18816-18826, DOI: 10.48550/arXiv.2211.13222, 2023.

[15] S. Nazir, M. H. Yousaf, J. C. Nebel and S. A. Velastin, A bag of expression framework for improved human action recognition, *Pattern Recognition Letters*, vol.103, pp.39-45, DOI: 10.1016/j.patrec.2017.12.024, 2018.

[16] T. Fujita and Y. Kawanishi, Future pose prediction from 3D human skeleton sequence with surrounding situation, *Sensors*, vol.23, no.2, 876, DOI: 10.3390/s23020876, 2023.

[17] Y. Wang, Z. Xu, L. Li and J. Yao, Robust multi-feature learning for skeleton-based action recognition, *IEEE Access*, vol.7, pp.148658-148671, DOI: 10.1109/ACCESS.2019.2945632, 2019.

[18] D. Yang, Y. Wang, A. Dantcheva, L. Garattoni, G. Francesca and F. Brémond, Self-supervised video pose representation learning for occlusion-robust action recognition, *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition*, pp.1-5, DOI: 10.1109/FG52635.2021.9667032, 2021.
[19] R. Fong and A. Vedaldi, Occlusions for effective data augmentation in image classification, *2019 IEEE/CVF International Conference on Computer Vision Workshop*, pp.4158-4166, DOI: 10.1109/ICCVW.2019.00511, 2019.
[20] K. K. Singh and Y. J. Lee, Hide-and-Seek: Forcing a network to be meticulous for weakly-supervised object and action localization, *2017 IEEE International Conference on Computer Vision*, pp.3544-3553, DOI: 10.1109/ICCV.2017.381, 2017.
[21] T. DeVries and G. W. Taylor, Improved regularization of convolutional neural networks with Cutout, *arXiv Preprint*, arXiv: 1708.04552, 2017.
[22] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Sheikh, OpenPose: Realtime multi-person 2D pose estimation using part affinity fields, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.43, no.1, pp.172-186, DOI: 10.1109/TPAMI.2019.2929257, 2021.
[23] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv Preprint*, arXiv: 1412.6980, 2015.
[24] J. Carreira and A. Zisserman, Quo Vadis, action recognition? A new model and the kinetics dataset, *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp.4724-4733, DOI: 10.1109/CVPR.2017.502, 2017.
[25] W. Kay, J. Carreira, K. Simonyan et al., The kinetics human action video dataset, *arXiv Preprint*, arXiv: 1705.06950, 2017.
[26] J. Guo, H. He, T. He et al., GluonCV and GluonNLP: Deep learning in computer vision and natural language processing, *The Journal of Machine Learning Research*, vol.21, no.23, pp.1-7, 2020.

Author Biography

Hechen Yun received his B.E. degree in Computer Science and Technology from Hangzhou Normal University, China, in 2017, and M.E. degree in Computer Science and Engineering from Akita University, Japan, in 2021. He is now enrolled in a doctoral program with the Graduate School of Engineering Science in Akita University. His research interests include computer vision, machine learning, and pattern recognition.

Etsuro Nakamura received the B.E. and M.E. degrees in Computer Science and Engineering, as well as a Dr. Eng. degree from Akita University, Japan, in 2017, 2019, and 2022, respectively. His research interests include human sensing, image processing, and speech content estimation. During his doctoral program, he developed a speaker-identification system for Japanese speech content and has authored over 30 papers in journals and conferences. At present, Dr. Nakamura has joined Japan Business Systems, Inc., where he currently serves as a development engineer in the Cloud Management Services Department.

Yoichi Kageyama received the B.E. and M.E. degrees in Computer Science and Engineering and the Dr. Eng. degree from Akita University, Japan, in 1995, 1997, and 2001, respectively. He joined Akita University as a Research Associate in 1997. He became an Assistant Professor in 2001 and an Associate Professor in 2004. He is now a Professor with the Department of Mathematical Science and Electrical Electronic Computer Engineering, Graduate School of Engineering Science, Akita University. His research interests include human sensing, remote sensing, and image processing.

Chikako Ishizawa received the B.E. degree in Chemical Engineering for Resources from Akita University, Japan, in 1992, and joined FUJIFILM Software Co., Ltd. She joined Akita University in 1995. She received a Dr. Eng. degree from Akita University in 2012. She is now a Professor with the Department of Mathematical Science and Electrical Electronic Computer Engineering, Graduate School of Engineering Science, Akita University. Her research interests include visual information processing and log analysis.

Nobuhiko Kato received the B.E. degree in electrical and electronic engineering from Akita University, Japan, in 1997, and joined ADK Fuji System Co., Ltd. Currently, he works in the Human Resources and Development Department. In recent years, his efforts have extended to designing and implementing internship curricula for students based on IoT and AI technologies. Since 2023, he is serving as a part-time lecturer to teach DX introduction lectures at National Institute of Technology, Akita College, Japan. Alongside designing and implementing curricula that focus on training new employees within and outside the company, he conducts collaborative lectures with universities, technical colleges, and vocational schools. His research interests include software engineering and computer programming education method.

Ken Igarashi received the B.E. degree in the Department of Information Engineering at the Faculty of Engineering from Toyo University, Japan, in 1995, and joined ADK Fuji System Co., Ltd. Throughout his tenure with the company, he has played a pivotal role in the development of diverse business systems for both public and private sectors. His contributions extend to the creation of elderly care systems utilizing human and environmental sensors, as well as agricultural IoT services, showcasing his proficiency as a systems engineer and project manager. Presently, Igarashi directs his focus towards the development of solutions that harness the power of AI-based technology. In his role as the head of the DX Solution Business Division within the company, he actively supports customers in their endeavors to achieve Digital Transformation (DX).

Ken Kawamoto received the B.E. degree in the Department of Applied Physics, Faculty of Engineering, Tohoku Gakuin University in 2000, and subsequently joined ADK Fuji System Co., Ltd. Throughout his career, he has contributed significantly to enhancing business efficiency, particularly through the development of web applications related to business processes. In his current role as the Manager of the company's DX Solution Department, he actively supports customers in their endeavors to achieve Digital Transformation (DX).

Beyond his responsibilities at ADK Fuji System Co., Ltd., Kawamoto operates his own services catering to small and medium-sized construction companies. Additionally, he engages in research and development to create new services aimed at enhancing the safety and security of the construction industry.