International Journal of Innovative Computing, Information and Control
Volume **20**, Number **5**, October **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **1301–1317**

SKELETON-BASED INTERPOLATION APPROACH WITH MOTION GENERATION FOR ACTION RECOGNITION ON DIVERSE OCCLUSIONS

Hechen Yun[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; 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-city, Akita 010-0851, Japan
{ nobuhiko; igarashi; kawamoto }@adf.co.jp

Received December 2023; revised April 2024

Abstract. *Human action recognition (HAR) methods are pivotal in the development of safety management systems for construction sites because they involve estimating crucial information such as work actions and physical conditions. However, in challenging work environments, detecting body information and estimating models based on skeletal data may be hindered owing to occlusion by objects. This paper proposes a novel interpolation approach for skeleton-based recognition using motion generation. By combining 2D pose estimation and skeletal transformation techniques, 3D skeletal information is reconstructed to learn better spatial features from monocular video. When the action prediction accuracy is low, the proposed method generates multiple action sequences, selected by the dynamic time warping algorithm based on similarity, to interpolate missing parts of skeletons. The experimental results demonstrate the effectiveness of the proposed method in improving the recognition accuracy of 14 types of movements under nine types of missing data, with average improvements of 3.8% in the Top-1 accuracy and 1.5% in the Top-5 accuracy. These results demonstrate that our proposed approach achieved better performance than state-of-the-art methods and has the potential to enhance the performance of HAR methods in occlusion situations.*
Keywords: Skeleton, Action recognition, Interpolation, Motion generation, Occlusion

1. **Introduction.** In Japan, the population of working-age individuals has been declining in recent decades owing to the broader issue of decreasing total domestic population. Conversely, the population of middle-aged adults (aged 45-60 years) and elderly individuals (over 60 years) is growing, particularly in the construction industry [1,2]. The construction industry is a cornerstone of infrastructure development and economic progress in a country, requiring substantial investment in a large workforce to sustain its vital functions. However, jobs in the construction sector demand extended periods of mental focus and the need to sustain high productivity to meet project deadlines, while ensuring the proper functioning of equipment. These demands are challenging for elderly workers who face an increased risk of occupational accidents owing to the natural decline in physical and cognitive capabilities that accompanies aging [3]. Consequently, it is imperative to

DOI: 10.24507/ijicic.20.05.1301

find ways to maximize personal safety while maintaining the efficiency and effectiveness of workers.

Moreover, owing to the rapid advancements in computer vision and machine learning, various approaches, such as understanding human behavior, have enabled wearable devices in construction workplaces to monitor, detect, and remotely manage activities using only a few surveillance cameras throughout the day. Human action recognition (HAR) has undergone remarkable development and has been applied to these approaches; skeleton-based methods have been extensively employed and explored for analyzing human behavior, leveraging advancements in pose estimation technologies, and demonstrating good antinoise performance [4,5]. However, occlusion remains a significant factor affecting the accuracy of model detection and prediction. Therefore, it is crucial to implement methods that effectively manage occlusion and precisely monitor progress in the construction industry.

Several research groups have endeavored to address the occlusion problem in computer vision. These methods can be broadly categorized into two types: data-based and model-based. Data-based research has attempted to use a data-augmentation strategy to enhance the robustness of recognition networks by learning various representations from datasets that have many preprocessing methods for simulating occlusion problems [6,7]. However, the aforementioned methods involve high computational and resource costs for creating various masked images at the preprocessing stage and building heavyweight networks to store various representations from additional data.

Model-based research adopts intricate construction methods to construct models that enforce their ability to learn latent structural information from existing skeleton data. Cui and Sun [8] utilized a multitask graph convolutional network (GCN) with a shared spatiotemporal representation to generate incomplete parts of a skeleton for repairing missing values. However, the adaptability of recognition tasks to diverse motion features with the same action labels has not been thoroughly explored. Song et al. [9] sought to solve the occlusion problem by constructing multi-stream models using a spatiotemporal graph convolutional network (ST-GCN) with a concatenation operation to increase the accuracy of action recognition tasks. However, the occlusion situations for various body parts have not been sufficiently investigated.

Generative machine learning techniques, such as generative adversarial networks (GANs) and variant networks, which are usually constructed by generators and discriminators to implement unsupervised learning, have also received much attention and have recently been shown to generate unexcited information based on learned data [10-12].

In a previous study [13], we verified the effectiveness of restoration methods in improving the accuracy of action recognition models. This is achieved by interpolating the skeleton information using the continuity constraint of the front and rear frames of the images. However, the limited number of interpolation algorithms has proven challenging in covering diverse occlusion situations that are not included in interpolation objects, but probably occur in real-world scenarios. To address the challenges posed by occlusions and the problems in previous research, we applied generative techniques of human motion to the HAR approach and aimed to address issues related to information deficiency.

In this study, we propose a skeleton-based interpolation approach aimed at enhancing the accuracy of action recognition in diverse occlusion scenarios. This approach synthesizes similar motions based on the prediction results from the action-recognition module and attempts to fill the information gaps in incomplete skeletons. Furthermore, a multidimensional dynamic time warping (DTW) algorithm was employed to integrate the original and generated segments of motion information. To evaluate the performance of the proposed approach, nine masking processes were designed to simulate real scenarios

at a construction worksite and experiments were conducted. The contributions of this study can be summarized as follows.

- This study developed a novel skeleton-based interpolation approach that can improve the accuracy of human action recognition when limited information is obtained from occlusion problems. Specifically, we combined the off-the-shelf action recognition model with the motion synthesis method and time-series analysis algorithm to better estimate missing skeletons based on previously predicted results.
- The proposed approach supplies additional motion information from the generated samples such that it requires learning from only small datasets, both the recognition model and the generation model, which efficiently avoids unnecessary costs.
- Nine types of occlusion processes were designed to simulate real-world constriction worksite scenarios where the target persons were partially and frequently invisible, and the performances of the ST-GCN++ model were assessed for various occlusion situations.

The remainder of this paper is organized as follows. Section 2 provides an overview of related works, encompassing various methods for skeleton-based human action recognition and motion generation. Section 3 describes the data acquisition process, baseline dataset, and occlusion processing used to simulate real-world scenarios. Section 4 outlines the complete procedure of the proposed approach, including the skeleton extraction, action recognition module, motion synthesis module, and integration module with recognition processing. The experimental results and a pertinent discussion are presented in Section 5. Section 6 provides the conclusions and suggestions for future work.

2. **Related Work.**

Human action recognition technology has found widespread application in human-computer interactions and surveillance systems, enabling timely and efficient collection and analysis of human behaviors [14-16]. In this section, we introduce the related studies on skeleton-based action recognition and motion generation methods.

2.1. **Skeleton-based action recognition method.**

Many research groups have focused on skeleton-based action recognition methods owing to their compactness, efficiency, and the subsequent development of human pose estimation technology [17]. Based on our limited understanding, skeleton-based action recognition methods can be classified into three categories according to the deep learning frameworks used: RNN-based, CNN-based, and GCN-based methods.

RNN-based approaches focus on learning the dynamic features of skeleton data as a type of temporal sequence in recurrent neural networks (RNNs) and extracting features from different temporal dimensions. Despite many research groups [18-20] attempting to link features between layers within networks, RNN-based approaches still have difficulty capturing latent representations from skeleton data, such as the relative position of the human body structure, which utilizes only single joints or parts of joints as individual sequences.

CNN-based approaches normally extract spatial information from frames or videos, but they still work for 2D or 3D array structures of skeletons that resemble images in format [21]. Moreover, CNNs can be combined with RNNs to simultaneously explore spatial and temporal information at the same time [22]. However, CNN-based approaches must implement a transformation process that converts skeleton data into image-like arrays, which is computationally expensive, as in traditional classification methods.

In GCN-based methods, skeletons are transformed into graphs with generalized topological structures that use graph edges and nodes to represent the lambs and joints of the human body. This approach is based on an efficient variant of convolutional neural

networks that can deeply mine the correlation information of the body structure in skeletons [23,24]. Duan et al. [25] presented a GCN-based model called ST-GCN++ with competitive recognition performance by simplifying and optimizing complicated architectural designs in the original ST-GCN [26]. However, current GCN-based models typically assume that the node feature information of the graph data is complete, enabling the extraction of sufficient representations for recognition tasks. This makes these models easier to have deleterious effects from incomplete graph data or missing data, which often occur in real-world situations. To improve the robustness of the GCN-based model, we propose an interpolation method to maintain the completeness of the skeletal graph data for recognition models.

Considering the simplicity and support of a wide variety of skeleton action recognition methods, ST-GCN++ was used to achieve action recognition and provide conditions for generating extra features for the interpolation process in our proposed approach.

2.2. **Human motion generation method.** Human motion generation, also known as human motion synthesis, aims at generating natural, realistic, and diverse human pose sequences. A generative adversarial network (GAN) is a classical generative model that has undergone significant development and expansion owing to its stability and simplicity. Mirza and Osindero [27] introduced conditional GANs that enabled the model to be trained under limited conditions and noise, ensuring that the generated results exhibited both good convergence and diversity. Building on this work, many studies on human motion generation have achieved superior performance using a GAN training scheme [28,29]. Degardin et al. [30] proposed an architecture that combines the benefits of both GANs and GCN, referred to as a Kinetic-GAN, to synthesize different action sequences. This approach generates motions directly from the latent space via spatiotemporal graph convolutions in the model, which maintains long-term relationships between frames.

To the best of our knowledge, no related work has adapted motion generation to interpolate graph data containing missing skeletal information and to help improve the performance of action recognition. Therefore, this study is the first to adopt kinetic-GAN as the generation model of our proposed approach to generate parts of the missing data according to the conditions from the first recognition process.

3. **Data Acquisition.** Figure 1 illustrates the data acquisition environment and examples of the actions used to conduct the validity experiments of the proposed approach. A total of 5,200 videos (59.94 fps, 1920 × 1080 pixels) of ten subjects engaged in 14 types of simulated workplace actions were captured. Table 1 lists the details of the captured actions simulated with work content in the construction industry. The subjects were

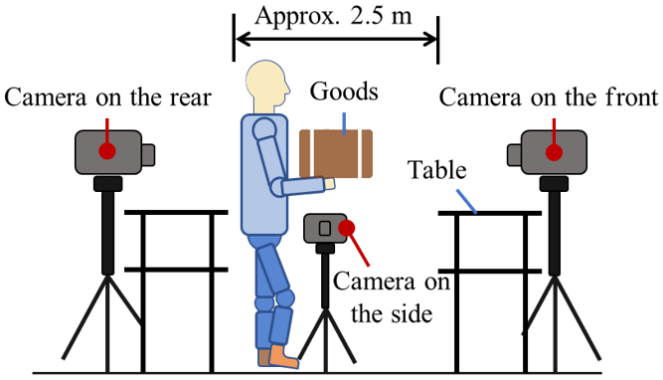


FIGURE 1. Data acquisition environment

TABLE 1. Information of captured actions

No.	Action name	Details
1	Walk	Walking at a constant speed while waving hands.
2 3	Sit down Sit up	Sitting in a chair with a backrest, placing hands on legs, and returning to standing position.
4 5	Climb down Climb up	Climbing down a stepladder step-by-step to reach the ground and climbing back up step-by-step.
6 7	Squat down Squat up	Sitting with knees bent and heels close to buttocks and returning to the standing position, where the back and legs are straightened.
8 9	Take up Throw	Lifting a large object above the head and dropping it forward or behind the body.
10 11	Tumble down Tumble up	Tumbling slowly into a mattress, relaxing the body, getting up from the mattress, and returning to the standing position.
12 13	Pick up Put down	Picking up objects from the ground or table using hands and putting them back.
14	Carry	Holding objects, turning around, and moving to opposite position.

captured primarily from three perspectives: front, rear, and side. However, the relative turn-back motion was also recorded if it was part of the action. Additionally, all videos were converted to image sequences at a 29.97 fps rate for easier processing of the experimental data. The data used in this investigation were acquired in accordance with the ethical regulations concerning human studies at Akita University, Japan.

4. **Proposed Method.** For HAR, some off-the-shelf models have already obtained sufficient research to make the prediction results from those models cover the true one within a limited range such as the Top-5. However, in real scenarios, the first result is usually adopted and other high-likelihood results are only considered for reference. This leads to the accumulation of errors and discrepancies in the results of the system analysis when occlusion and other noise problems occur. Meanwhile, 3D skeleton-based HAR models perform better than 2D skeleton-based models because the depth information provides additional feature representations for discriminating complex actions [31]. However, depth information is difficult to obtain when considering a surveillance system that uses a monocular camera at the current construction worksite.

To overcome the occlusion problems of vision limitations and enhance the reliability of the management system, we designed a recognition pipeline for HAR by introducing a skeleton-based interpolation method using a trained motion synthesis model to interpolate the original skeleton data. The interpolation method is based on a limited range of results from the off-the-shelf model to generate extra information for completing the skeleton graph data and to improve the accuracy of the first result. The proposed approach comprises four modules: 1) skeleton extraction, 2) action recognition, 3) motion generation, and 4) integration processing. An overview of the proposed approach is shown in Figure 2.

4.1. **Skeleton extraction.** In this study, 3D skeletal information was used to learn a better skeletal topology for motion representation with depth. Human pose estimation (HPE) involves identifying body representations (e.g., skeletal information) from images or videos, usually referred to as keypoints. However, to the best of our knowledge, published open-source human pose estimation techniques based on monocular videos for 3D skeletons (direct estimation approaches without intermediate 2D representations) have

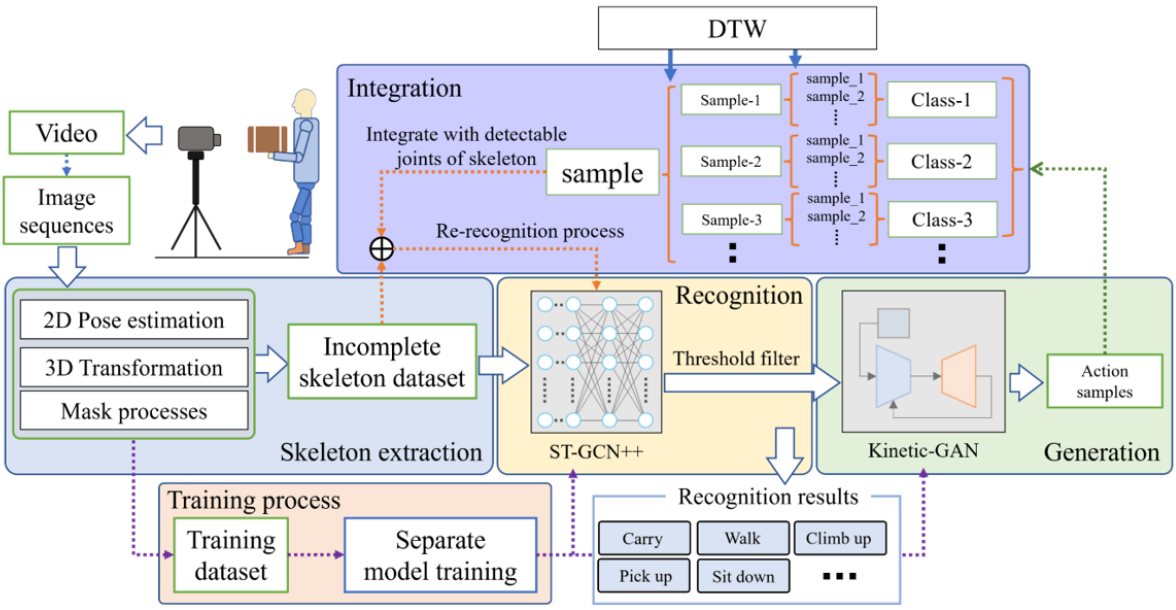


FIGURE 2. Overview of the proposed approach

had difficulty achieving both stability and practicability until recently (e.g., MediaPipe-BlazePose and AlphaPose [32,33]). By contrast, 2D pose estimation algorithms have been developed several times and have achieved good performance, even for complex scenes or multiple-person tasks with real-time implementations. Moreover, 2D-to-3D lifting human pose estimation methods have achieved significant success, owing to the advantages of self-attention mechanisms. However, the effectiveness of 2D-to-3D lifting technology in surveillance tasks has not been sufficiently discussed and has not been specifically designed for recognizing work actions during construction. Therefore, to address these problems and obtain stable skeleton data, we employed a two-stage strategy that uses a 2D human pose detector, YOLOv8-Pose [34], with a 2D-to-3D lighting approach, PoseFormerV2 [35], to estimate 3D skeleton data from monocular videos of data acquisition. A flowchart of the skeleton extraction process is shown in Figure 3.

Specifically, YOLOv8-Pose was used to extract the 2D coordinates of the human joints along with the relative confidence scores from each video frame (image of sequence). The confidence scores of the joints indicated the prediction accuracy of the pose estimation model or the visibility of the joints, with values ranging from 0.0 to 1.0. According to our observations, the HPE model is sufficiently robust to estimate the joints when the parts of

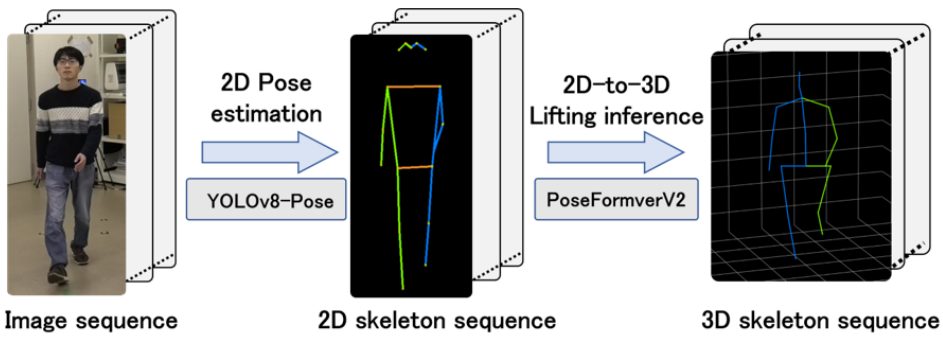


FIGURE 3. Overview of skeleton extraction

the human body are invisible; however, the results become unreliable when the confidence scores of the estimated joints are low. Therefore, to maintain both the robustness of the HPE model and avoid inaccurate predicted results in the model's estimation in this study, a threshold value of 0.30 is applied to filter out noise in the skeleton data, where it is difficult for the pose estimator to detect joints even to estimate those positions in the image. When the confidence score of a joint was below a threshold value, the coordinate value was set to zero. This adjustment was made because, in our observations, occlusion problems arise from complex motion and interactions with the environment and lead to the joints becoming invisible and undetectable; however, the estimator still attempts to provide tentative values based on previous training information. Additionally, some joints were considered missing when the coordinate values (x, y) were zero, irrespective of the confidence scores. After the filtering process, the residuals of the 2D skeleton data are passed on to the next process.

Second, 3D skeleton data (coordinates with depth information) were inferred based on the estimated 2D skeletons using PoseFormerV2. Benefiting from the self-attention mechanism of transformers and the utility of frequency-domain representation, PoseFormerV2 can efficiently and robustly predict the relative lengths of sequences. However, the 2D-to-3D lifting approach generates 3D skeletons with unnatural poses when parts of the joints are inevitably missing in 2D skeletons. This is because PoseFormerV2 learns global temporal correlations across the entire frame of joint sequences, leading to the model outputting a plausible result that includes missing joints in frames based on its pretrained information.

Therefore, to better understand situations in which information from natural movements is lacking, we set the coordinate values of joints to zero in the 3D skeleton data when relevant 2D joints are missing in the same frame. The extracted 3D skeleton data for the 17 joints were used in the experiments.

4.2. **Action recognition.** The recognition process was constructed using a PyTorch-based open-source toolkit called MMAction2 [36], and ST-GCN++ was chosen as the recognition model because it supports skeleton-based datasets and achieves state-of-the-art performance.

ST-GCN++ was proposed based on a GCN with a complicated attention mechanism from the ST-GCN and achieved significant performance on various recognition tasks, comparable to other approaches. A spatiotemporal graph was constructed to represent the human skeleton sequence hierarchically with 17 joints and various frames. The skeleton joints are used as graph nodes and are connected by human body structures, and the same positional joints across consecutive frames are the graph edges. Spatiotemporal graph convolution operations are applied when 3D skeleton data are input, and the output value can be written as Equation (1), according to ST-GCN:

$$f_{out} = \sum_{d=0}^{D_{\max}} W_d f_{in} \left(\Lambda_d^{-\frac{1}{2}} A_d \Lambda_d^{-\frac{1}{2}} \otimes M_d \right), \tag{1}$$

where $D_{\max}$ is the predefined maximum graph distance and f_{in} and f_{out} are the input and output feature maps, respectively. W and M denote the learnable weight matrices for computing the inner product using the input feature maps and for normalizing the corresponding subset to the output, respectively. $\otimes$ represents elementwise multiplication, A_d represents the d-th order adjacency matrix that marks the pairs of joints with a graph distance d, and two parts of Λ are used to normalize A_d.

When the prediction results were obtained from the trained ST-GCN++ model, the probabilities of predicting all action categories were collected and passed to the generation module in the proposed approach for interpolation. However, this approach requires a large number of computational resources to generate samples for interpolating missing information, which is unnecessary for samples that have already achieved high prediction scores and cannot significantly change the prediction results. To decrease the computational costs and improve the success rate of interpolation, a threshold value of 0.50 was set to filter out unnecessary samples. The samples that achieved or exceeded the threshold value were not processed, whereas those that did not pass to the next module were considered the final recognition results.

4.3. **Motion generation.** In this module, we selected the Kinetic-GAN framework as the generation model to synthesize the human actions of the skeleton based on the label output from the recognition module.

The kinetic-GAN utilizes graph convolutions on a generative adversarial network to ensure the stable generation of skeletal motions that conform to the human body structure. A GCN consists of correlational information between joint connections in spatial and changing temporal dimensions. Conditional synthesis of action sequences maintains long-term relationships between frames and can provide more realistic data than unsupervised learning. Additionally, kinetic-GAN applies graph pyramids with an encoder-decoder architecture, which upsamples and downsamples the input graphs at multiple resolution levels to explore the latent information inside the intermediate nodes over the spatial and temporal dimensions.

To train the generator and discriminator of the Kinetic-GAN, a graphical construction was designed to be compatible with the skeleton extraction process and recognition module. Moreover, the root nodes were set to the upper torso joints to implement graph convolution to up-sample and down-sample each sample. The exploration path of graph pyramids for a single frame is shown in Figure 4. The downsampling process aims to extract high-level information in the spatial dimension by eliminating redundant nodes, while preserving structural information. This process involves different levels of graphs containing multilevel semantic information.

Furthermore, to generate the missing motion information from the generator, we used the prediction results passed from the recognition module as label embeddings to generate

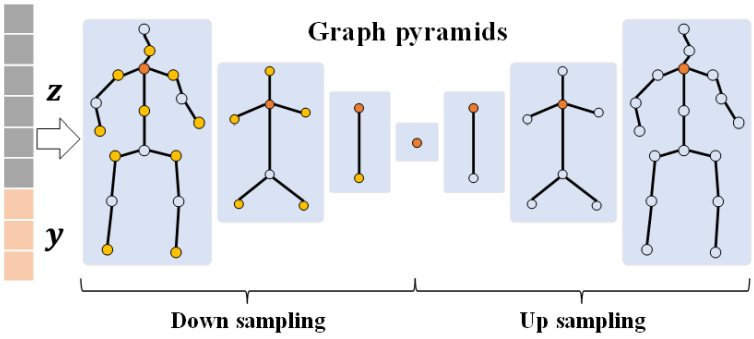


FIGURE 4. Exploring path of graph pyramids
(Root node represents the center node for graph structure; yellow node represents the node that will be down sampled at next level; embedded action label y and Gaussian random noise z are combined to train the generator.)

each action sequence. The generator was used to generate samples of 64-frame action sequences with subtle differences due to random latent noise z after the training procedures. Samples of the generated action sequences are provided to the integration module.

4.4. **Integration process.** The integration module filters generated action sequences from numerous samples to a single sample that is able to fill missing parts of joints with generated coordinate information when estimated skeletons are incomplete.

Considering that the action sequence consists of both spatial joint vectors and temporal frames, an adaptive measurement is necessary to calculate the similarity between the two action samples of the temporal sequence. In this study, a multidimensional matching algorithm with conventional DTW was employed during the sample-filtering process for multivariate time series of action sequences. DTW is an effective algorithm for measuring the similarity between two sequences with different time-series lengths using dynamic time warping for distance measurement. Inspired by the open-source Python library DTAIDistance supported by the DTAI research group [37,38], we used the Euclidean distance as the numerical assessment standard to calculate the similarity between the generated samples and original incomplete skeleton data. Coordinates were matched using a nonlinear mapping method to calculate an optimal match to two given sequences, and the joints of coordinates along the x, y, and z axes (three dimensions) were used to calculate the sum of all distances between the mapped sequence elements. Only the joints related to the existing part from the original skeleton data were calculated and summarized, and incomplete parts were ignored. The three sums of the Euclidean distances are considered a loss function, and the lower the value, the greater the similarity between the two sequences.

Using the above measurement method, all similarities between the generated sample sequences and the original incomplete skeleton data were calculated. The most similar sample (shortest distance) was selected as the final sample for integration with the original skeletal data.

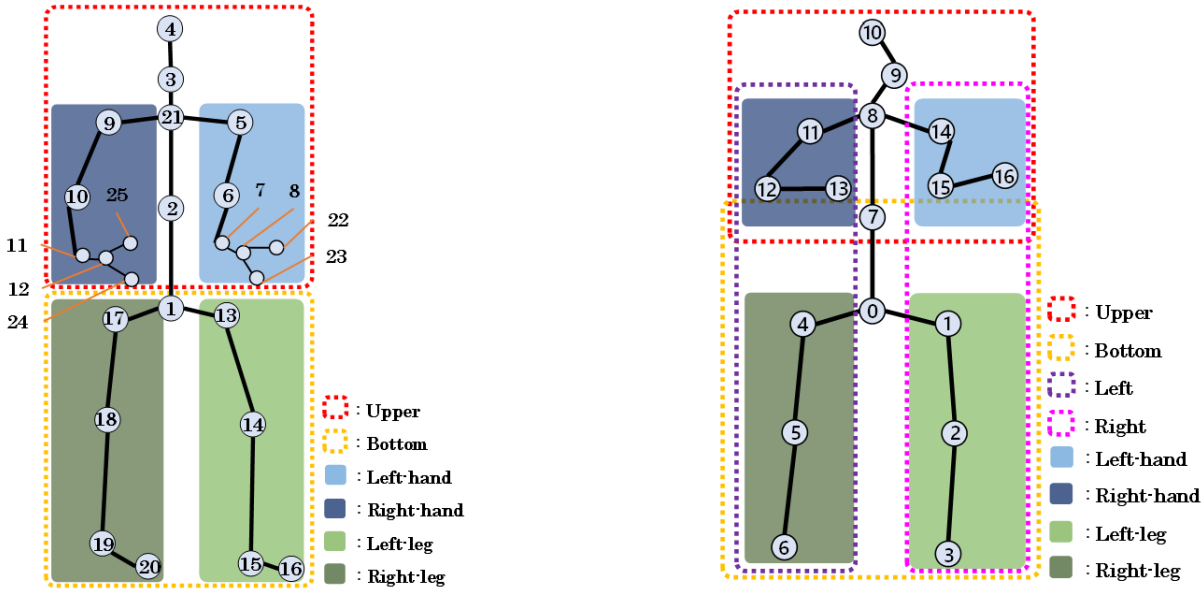
Because the generated sample contains 64 frames, which are fixed owing to the architecture of the Kinetic-GAN, the original skeleton data vary for the generated samples. Therefore, the final generated sample is transformed at the time-series level to fit the frame numbers of the original data. Specifically, if the generated sample has fewer frames than the related original sample, the generated sample will be padded with empty frames or even temporary frames based on the difference in frame numbers between the two frames. Then, the linear interpolation algorithm is used to fill the coordinate values in the empty frames. Conversely, if the generated sample was longer than the related original sample, it was clipped evenly with temporary frames to fit the original sample. Finally, the missing joints of the x, y, and z coordinates from the original data were filled with the generated sample, and the existing joints remained original in the corresponding frame after frame transformation. The integrated data were rerecognized using the same recognition model.

5. Experiments and Discussion.

5.1. **Datasets and evaluation metrics.** Our experiments were based on two types of datasets to evaluate the performance of the proposed approach: one was from the public dataset NTU RGB+D 60 [39], and the other was extracted from the data acquisition in Section 3. With respect to NTU RGB+D, 56880 video samples of 60 action classes were obtained from 40 distinct subjects captured using Microsoft Kinect V2, which can estimate the 3D skeletal data of the coordinator of 25 major body joints from time-sequencing images with high accuracy. After sample filtering with missing skeleton data based on the cross-subject set, the NTU dataset that contains 40091 and 16487 samples

was used for training and evaluation, respectively. In this study, only the 3D skeletons on the X-Sub benchmark of the NTU RGB+D 60 dataset were used.

To evaluate the performance of the proposed approach in real-world scenarios under occlusion conditions, we designed seven types of masking processes for the NTU dataset and two additional types of masking processes for the data acquisition dataset (nine mask categories). The datasets processed by masking were used individually for the evaluation experiments. Specifically, the mask categories included upper, bottom, left, right, left hand, right hand, left leg, right leg, and intermediate. With respect to the dataset from data acquisition, the 'Left' and 'Right' mask categories were added, which were based on the mask categories for the NTU dataset. Moreover, 'Intermediate' is applied for masking all joints of the middle 30 frames in each sample for both the NTU dataset and the acquired dataset. In addition, mask processing was implemented on the joints, and the x, y, and z coordinates were set to zero. The masking designs for the positional and intermediate frames of each dataset are presented in Figures 5 and 6, respectively.



(a) NTU RGB+D dataset

(b) Dataset of data acquisition

FIGURE 5. Design illustrations of masking processes for positional joints

The original and masked datasets were used to evaluate the performance of the recognition module, and the proposed approach interpolates only the masked datasets and classifies them using the same recognition model, which implements both the NTU and acquired datasets. All action classes were calculated as probabilities, and the highest prediction probability of the action class and the top five probabilities of prediction were used as evaluation standards and defined as Top-1 and Top-5, respectively.

5.2. **Implementation details.** The NTU and acquired datasets were employed separately in experiments to evaluate the validity and usability of the proposed approach. On the one hand, to evaluate the validity of the proposed approach within a limited time, only the evaluation part of the NTU dataset was used in the experiments, and the models were already pretrained with the NTU dataset provided by the research groups of MMaction2 and Kinetic-GAN, which included the ST-GCN++ recognition model and motion generation model. Moreover, we generated action sequences within the top five classes based on the prediction results because of limitations in computational resources. In addition, we employed SOTA skeleton-based HAR methods that include ST-GCN, 2s-AGCN [40],

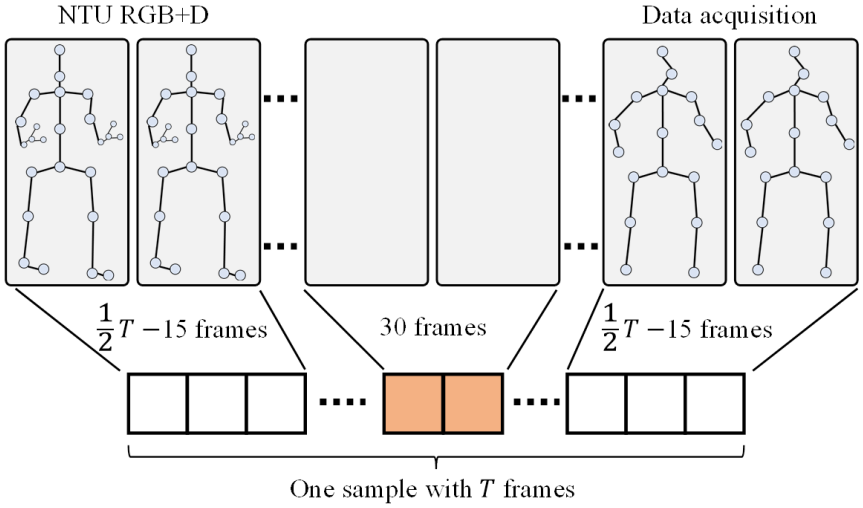


FIGURE 6. Design illustration of masking process for intermediate frames

MSG3D [41], CTRGCN [42], and the original ST-GCN++ as baselines for comparison with our approach.

On the other hand, the graph structure of the skeleton is redesigned based on 17 joints according to the results of the pose estimation technique because of the compatibility of both the recognition model and the generation model when those models are trained on the dataset used for data acquisition. The ST-GCN++ model and other comparison methods was trained with 200 epochs on the acquired dataset, and those models with the best validation accuracy were selected and employed in the experiments. The motion generation model was trained with 2,000 epochs, and the latest model was used to generate action sequences during the experiments, which achieved convergence between the generator and discriminator. Models were trained only on the original dataset to ensure that incomplete skeleton patterns were not learned during training. All the experiments were conducted on a computer with an NVIDIA RTX A4000 GPU, PyTorch 2.0.1 framework, and Ubuntu 22.04 LTS OS.

5.3. **Recognition results for the NTU dataset.** Table 2 presents the comparison results of the recognition accuracies between the original and masked NTU datasets with seven types of mask processing. According to the prediction results, the performance of the recognition models has various degrees of negative impact in occlusion situations. This

TABLE 2. Recognition results for NTU dataset [%]

Mask category	No interpolated		Interpolated		Diff.	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Normal	89.23	98.03				
Upper	14.08	31.98	15.41	33.86	1.33	1.88
Bottom	78.29	94.58	80.74	95.22	2.45	0.64
Left-hand	67.73	89.59	71.52	90.77	**3.79**	1.18
Right-hand	53.80	79.89	57.03	80.87	**3.23**	0.98
Left-leg	87.47	97.72	87.94	97.82	0.47	0.10
Right-leg	86.80	97.43	87.32	97.41	0.52	−0.02
Intermediate	68.62	91.73	62.73	86.70	*−5.89*	*−5.03*
Avg.	65.26	83.27	66.10	83.24	0.84	−0.04

effect was more pronounced for data in the upper-mask category. This is because most of the effective information for action discrimination was expressed in the upper body of the NTU dataset. In particular, the 'Upper' dataset has significantly lower accuracies: Top-1 is 14.08%, and Top-5 is 31.98%. Moreover, the performances of the 'Left-hand' and 'Right-hand' masks also present the fact that incomplete skeleton situations have more effects on the recognition model.

According to the comparison results, the interpolation approach achieved a better performance in recognizing incomplete skeletons in the NTU dataset. Compared to the recognition results without integrating the generated skeleton, the upper-related and left-leg-related occlusion situations obtained better Top-1 and Top-5 accuracies. Moreover, Top-1 benefited from our approach more than did Top-5 (3.79% and 3.23% improvement in the 'Left-hand' and 'Right-hand' datasets, respectively). In addition, the average accuracy of the Top-1 dataset increased slightly by 0.84% after interpolation. To some extent, these results validate the rationality of our assumption that motion generation, which involves previous training with known motion information, can help recognition perform better when the recognition model has a certain level of reliability. In addition, the results demonstrate that the proposed interpolation method can improve the accuracy of the HAR model with many daily action categories in occlusion situations where occluded daily actions frequently occur at real construction worksites.

However, the proposed approach obtained the same or even lower recognition accuracies than did the normal dataset for the left-leg, right-leg, and intermediate masks. One reason for this is that the motion information is not as important as that of the upper-body recognition model, which already achieves good performance even without extra processing that supports unnecessary features and extra noise in the recognition model.

5.4. **Recognition results for the datasets of data acquisition.** In Table 3, we compare the proposed approach with the SOTA methods in terms of Top-1 accuracy on the data acquisition dataset with nine types of mask categories. The results show that the accuracy gaps between the compared methods and ST-GCN++ become much smaller when using the proposed approach. The comparison methods demonstrate better robustness on an occluded dataset because they are trained with a more complex model design, more parameters, and more FLOPs than the original ST-GCN++, indicating that ST-GCN++ has better computational efficiency and is more usable for real-world implementation.

Table 3. Comparison results of Top-1 accuracy [%]

	Normal	Upper	Bottom	Left	Right	Left-hand	Right-hand	Left-leg	Right-leg	Intermediate	Avg.
ST-GCN	97.06	8.19	55.35	69.38	63.74	85.58	**88.34**	87.11	76.45	88.78	69.21
2s-AGCN	97.59	14.69	**57.81**	**75.13**	65.50	88.77	88.00	**87.28**	79.15	90.93	71.92
MSG3D	97.44	13.91	42.74	58.32	48.47	80.47	85.32	74.27	62.43	**93.57**	62.17
CTRGCN	97.40	**21.18**	55.26	73.98	**67.10**	**88.82**	87.69	87.26	80.07	91.91	**72.59**
ST-GCN++	97.24	12.82	46.34	60.70	49.68	82.51	86.00	78.17	64.21	92.37	63.64
ST-GCN++ with interpolation (Ours)		14.25	55.00	63.48	57.94	82.17	86.02	80.91	75.83	91.53	67.46

Table 4 lists the comparison of average recognition results from ten validation subjects on the dataset of data acquisition, which was processed with nine types of mask categories. We observe that the recognition model achieved better performance after the implementation of the proposed interpolation approach, in which the average accuracy of the Top-1 improved from 63.64% to 67.46%, and that of the Top-5 improved from 86.57%

TABLE 4. Recognition results for dataset of data acquisition [%]

Mask category	No interpolated		Interpolated		Diff.	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Normal	97.24	99.98				
Upper	12.82	38.30	14.25	41.26	1.43	2.96
Bottom	46.34	82.29	55.00	85.39	**8.66**	3.10
Left	60.70	93.13	63.48	92.88	2.78	−0.25
Right	49.68	83.73	57.94	88.58	**8.26**	4.84
Left-hand	82.51	98.88	82.17	97.69	−0.34	*−1.19*
Right-hand	86.00	99.20	86.02	99.39	0.03	0.19
Left-leg	78.17	95.81	80.91	95.77	2.74	−0.04
Right-leg	64.21	88.49	75.83	93.35	**11.62**	4.86
Intermediate	92.37	99.31	91.53	98.45	−0.84	−0.86
Avg.	63.64	86.57	67.46	88.08	3.81	1.51

to 88.08%. This indicates that our interpolation approach indeed effectively assists the pattern classification of the recognition model when skeleton features are not sufficient.

Specifically, the ST-GCN++ model yields a low prediction accuracy when most parts of the body, such as the upper and bottom parts, are occluded. We believe that this is due to the lack of important motion information for classification, which is the same as the performance on the NTU dataset. However, the difference between the two experimental results (Section 5.3 and Section 5.4) shows that our approach performed better on the bottom and right mask categories. Compared to those in the NTU dataset, the proportions of bottom-body relevant samples, such as those related to walking, climbing, and sitting, are greater than those in the other datasets. It is efficient to complement this motion information using generated samples, including generators trained on different subjects that probably exhibit various personal behaviors. Moreover, the participants were instructed that their ability to use their right hand only to perform motions related to single-hand interactive movements with objects during data acquisition would be restricted. Therefore, losing the skeleton of the right body parts is fatal for the recognition model when classifying this dataset, and proper feature supplements are useful for the generation and integration modules. Moreover, the interpolated prediction results also significantly improved upon the uncorrected results on the 'Right-leg' mask category. We believe that this is because the setting position of the camera, which included approximately three-quarters of the data, was captured by the camera set up on the side that only captured the right-side aspect of the subjects. Although we reconstructed the skeleton information from 2D to 3D to alleviate the limitations of monocular cameras, certain errors and jitter in the skeleton coordinates still exist based on the pose transformation technique. When the recognition model relies on reconstructed information without other sides of the skeleton that are directly observable, for which the right parts are observable and the left parts are reconstructed, it is difficult for the model to provide correct predictions with many noise and uncertain features. The main component of the dataset is also a factor affecting the recognition performance of the 'Right-leg' mask. The generation model supplies whole skeletons of actions that allow DTW to calculate similarities based on the most efficient features, which are the upper-body parts, and selects the correct sample to facilitate the recognition process. For the 'Intermediate' mask, the experimental results show high accuracy for both types of datasets; these results are different from those for the NTU dataset, and the proposed approach has also no opportunity to be

implemented effectively. We believe this is due to the frame number of videos from data acquisition being more than 30 frames (normally two times of 30 frames or more), which allows the recognition model to utilize the residual frames to extract motion patterns, which are sufficient for the classification of the model.

According to the experimental results, the proposed approach demonstrates the versatility of improving the performance of the 3D skeleton-based HAR model when the body parts of the target person are occluded in the real world, thereby boosting the surveillance system to collect correct data for safety management. They also show that the proposed approach can mitigate the problem of unstable model accuracy owing to partially missing graph data and demonstrate the potential to recognize complex work contents under occlusions.

6. **Conclusion.** In this paper, we propose a novel approach that connects a skeleton-based action recognition model with a motion generation technique to reconstruct skeleton features in occluded situations. The proposed approach outperforms the baseline method of ST-GCN++ in terms of performance on simulation datasets for nine types of occlusions based on a public dataset and data acquisition. Moreover, the recognition model and the generation model were only trained on an unmasked dataset, which was not necessary to augment the dataset at the preprocessing stage. The results led to the following conclusions.

1) The proposed approach is valid for improving the accuracy of recognition models by interpolating missing parts of skeletons when occlusion occurs and has the potential to be utilized in real construction worksites.
2) The proposed approach achieved an average improvement in accuracy of 3.81% for Top-1 and 1.51% for Top-5 compared with the ST-GCN++ baseline method.

In future, we plan to collect video data from real construction sites to train models that can recognize and generate professional actions based on construction-related expertise. Additionally, new methods for enhancing the performance of skeleton integration across different sample frame numbers were considered.

REFERENCES

[1] Ministry of Health, Labour and Welfare of Japan, *A Survey on Occupational Accidents of Older Workers in 2023*, https://www.mhlw.go.jp/content/11302000/001099505.pdf, Accessed on April 8, 2024.

[2] Japan Construction Occupational Safety and Health Association, *A Survey on Occupational Accidents in the Construction Industry*, https://www.kensaibou.or.jp/safe_tech/statistics/occupational_accidents.html, Accessed on April 8, 2024.

[3] Ministry of Land, Infrastructure, Transport and Tourism, Japan, *Current Status of the Construction Industry*, https://www.mlit.go.jp/policy/shingikai/content/001602250.pdf, Accessed on April 8, 2024.

[4] X. Zhao, Construction risk management research: Intellectual structure and emerging themes, *International Journal of Construction Management*, pp.1-11, DOI: 10.1080/15623599.2023.2167303, 2023.

[5] S. M. E. Sepasgozar, Differentiating digital twin from digital shadow: Elucidating a paradigm shift to expedite a smart, sustainable built environment, *Buildings*, vol.11, no.4, 151, DOI: 10.3390/buildings11040151, 2021.

[6] 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.

[7] L. Lin, S. Song, W. Yang and J. Liu, MS2L: Multi-task self-supervised learning for skeleton based action recognition, *Proc. of the 28th ACM International Conference on Multimedia*, pp.2490-2498, DOI: 10.1145/3394171.3413548, 2020.

[8] Q. Cui and H. Sun, Towards accurate 3D human motion prediction from incomplete observations, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, pp.4801-4810, 2021.

[9] Y. Song, Z. Zhang, C. Shan and L. Wang, Richly activated graph convolutional network for robust skeleton-based action recognition, *IEEE Transactions on Circuits and Systems for Video Technology*, vol.31, no.5, pp.1915-1925, DOI: 10.1109/TCSVT.2020.3015051, 2021.

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, Generative adversarial networks, *Communications of the ACM*, vol.63, no.11, pp.139-144, DOI: 10.1145/3422622, 2020.

[11] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta and A. A. Bharath, Generative adversarial networks: An overview, *IEEE Signal Processing Magazine*, vol.35, no.1, pp.53-65, DOI: 10.1109/MSP.2017.2765202, 2018.

[12] J. Zhu, T. Park, P. Isola and A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.2242-2251, DOI: 10.1109/ICCV.2017.244, 2017.

[13] H. Yun, E. Nakamura, Y. Kageyama, C. Ishizawa, N. Kato, K. Igarashi and K. Kawamoto, Action recognition of simulated workplace with occlusion based on interpolated skeleton data using OpenPose, *International Journal of Innovative Computing, Information and Control*, vol.20, no.1, pp.231-247, DOI: 10.24507/ijicic.20.01.231, 2024.

[14] U. A. Usmani, J. Watada, J. Jaafar, I. A. Aziz and A. Roy, Particle swarm optimization with deep learning for human action recognition, *International Journal of Innovative Computing, Information and Control*, vol.17, no.6, pp.1843-1870, DOI: 10.24507/ijicic.17.06.1843, 2021.

[15] Z. Sun, Q. Ke, H. Rahmani, M. Bennamoun, G. Wang and J. Liu, Human action recognition from various data modalities: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.45, no.3, pp.3200-3225, DOI: 10.1109/TPAMI.2022.3183112, 2023.

[16] Y. Kong and Y. Fu, Human action recognition and prediction: A survey, *International Journal of Computer Vision*, vol.130, pp.1366-1401, DOI: 10.1007/s11263-022-01594-9, 2022.

[17] C. Zheng, W. Wu, C. Chen, T. Yang, S. Zhu, J. Shen, N. Kehtarnavaz and M. Shah, Deep learning-based human pose estimation: A survey, *ACM Computing Surveys*, vol.56, no.1, pp.1-37, DOI: 10.1145/3603618, 2023.

[18] Y. Du, W. Wang and L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp.1110-1118, DOI: 10.1109/CVPR.2015.7298714, 2015.

[19] W. Li, L. Wen, M.-C. Chang, S. N. Lim and S. Lyu, Adaptive RNN tree for large-scale human action recognition, *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, pp.1444-1452, DOI: 10.1109/ICCV.2017.161, 2017.

[20] J. Liu, G. Wang, L.-Y. Duan, K. Abdiyeva and A. C. Kot, Skeleton-based human action recognition with global context-aware attention LSTM networks, *IEEE Transactions on Image Processing*, vol.27, no.4, pp.1586-1599, DOI: 10.1109/TIP.2017.2785279, 2018.

[21] H. Duan, Y. Zhao, K. Chen, D. Lin and B. Dai, Revisiting skeleton-based action recognition, *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp.2969-2978, DOI: 10.1109/CVPR52688.2022.00298, 2022.

[22] H. Zan and G. Zhao, Human action recognition research based on fusion TS-CNN and LSTM networks, *Arabian Journal for Science and Engineering*, vol.48, pp.2331-2345, DOI: 10.1007/s13369-022-07236-z, 2023.

[23] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, *arXiv Preprint*, arXiv: 2107.12213, 2021.

[24] J. Lee, M. Lee, D. Lee and S. Lee, Hierarchically decomposed graph convolutional networks for skeleton-based action recognition, *arXiv Preprint*, arXiv: 2208.10741, 2023.

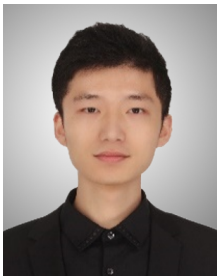
[25] H. Duan, J. Wang, K. Chen and D. Lin, PYSKL: Towards good practices for skeleton action recognition, *Proc. of the 30th ACM International Conference on Multimedia*, pp.7351-7354, DOI: 10.1145/3503161.3548546, 2022.

[26] S. Yan, Y. Xiong and D. Lin, Spatial temporal graph convolutional networks for skeleton-based action recognition, *Proc. of the AAAI Conference on Artificial Intelligence*, vol.32, no.1, DOI: 10.1609/aaai.v32i1.12328, 2018.

[27] M. Mirza and S. Osindero, Conditional generative adversarial nets, *arXiv Preprint*, arXiv: 1411.1784, 2014.

[28] L. Xu, Z. Song, D. Wang, J. Su, Z. Fang, C. Ding, W. Gan, Y. Yan, X. Jin, X. Yang, W. Zeng and W. Wu, ActFormer: A GAN-based transformer towards general action-conditioned 3D human motion generation, *arXiv Preprint*, arXiv: 2203.07706, 2023.
[29] Q. Men, H. P. H. Shum, E. S. L. Ho and H. Leung, GAN-based reactive motion synthesis with class-aware discriminators for human-human interaction, *Computers & Graphics*, vol.102, pp.634-645, DOI: 10.1016/j.cag.2021.09.014, 2022.
[30] B. Degardin, J. Neves, V. Lopes, J. Brito, E. Yaghoubi and H. Proença, Generative adversarial graph convolutional networks for human action synthesis, *arXiv Preprint*, arXiv: 2110.11191, 2022.
[31] P. Elias, J. Sedmidubsky and P. Zezula, Understanding the gap between 2D and 3D skeleton-based action recognition, *2019 IEEE International Symposium on Multimedia (ISM)*, pp.192-1923, DOI: 10.1109/ISM46123.2019.00041, 2019.
[32] Google, *MediaPipe*, https://developers.google.com/mediapipe, Accessed on April 8, 2024.
[33] H. Fang, J. Li, H. Tang, C. Xu, H. Zhu, Y. Xiu, Y. Li and C. Lu, AlphaPose: Whole-body regional multi-person pose estimation and tracking in real-time, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.45, no.6, pp.7157-7173, DOI: 10.1109/TPAMI.2022.3222784, 2023.
[34] Ultralytics, *Ultralytics YOLOv8*, https://github.com/ultralytics/ultralytics, Accessed on December 15, 2023.
[35] Q. Zhao, C. Zheng, M. Liu, P. Wang and C. Chen, PoseFormerV2: Exploring frequency domain for efficient and robust 3D human pose estimation, *arXiv Preprint*, arXiv: 2303.17472, 2023.
[36] OpenMMlab, *MMAction2*, https://github.com/open-mmlab/mmaction2, Accessed on April 8, 2024.
[37] K. Paliwal, A. Agarwal and S. Sinha, A modification over Sakoe and Chiba's dynamic time warping algorithm for isolated word recognition, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.1259-1261, DOI: 10.1109/ICASSP.1982.1171506, 1982.
[38] DTAI Research Group, *DTAIDistance*, https://github.com/wannesm/dtaidistance, Accessed on April 30, 2024.
[39] A. Shahroudy, J. Liu, T. Ng and G. Wang, NTU RGB+D: A large-scale dataset for 3D human activity analysis, *arXiv Preprint*, arXiv: 1604.02808, 2016.
[40] L. Shi, Y. Zhang, J. Cheng and H. Lu, Two-stream adaptive graph convolutional networks for skeleton-based action recognition, *arXiv Preprint*, arXiv: 1805.07694, 2019.
[41] Z. Liu, H. Zhang, Z. Chen, Z. Wang and W. Ouyang, Disentangling and unifying graph convolutions for skeleton-based action recognition, *arXiv Preprint*, arXiv: 2003.14111, 2020.
[42] Y. Chen, Z. Zhang, C. Yuan, B. Li, Y. Deng and W. Hu, Channel-wise topology refinement graph convolution for skeleton-based action recognition, *arXiv Preprint*, arXiv: 2107.12213, 2021.

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.

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. 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. 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, he 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., he 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.