IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 28 November 2023, accepted 31 December 2023, date of publication 9 January 2024, date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3351888

RESEARCH ARTICLE

DGU-HAO: A Dataset With Daily Life Objects for Comprehensive 3D Human Action Analysis

JIHO PARK[1], **JUNGHYE KIM**[2], **YUJUNG GIL**[3], **AND DONGHO KIM**[4]
[1]Department of Artificial Intelligence, Dongguk University, Seoul 04620, South Korea
[2]Department of Information and Communication Engineering, Dongguk University, Seoul 04620, South Korea
[3]Department of Computer Science and Engineering, Dongguk University, Seoul 04620, South Korea
[4]Software Education Institute, Dongguk University, Seoul 04620, South Korea

Corresponding author: Dongho Kim (dongho.kim@dgu.edu)

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant S-2021-A0496-00167; in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program under Grant IITP-2024-2020-0-01789; and in part by the Artificial Intelligence Convergence Innovation Human Resources Development supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2024-RS-2023-00254592.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT The importance of a high-quality dataset availability in 3D human action analysis research cannot be overstated. This paper introduces DGU-HAO (**H**uman **A**ction analysis dataset with daily life **O**bjects). This novel 3D human action multi-modality dataset encompasses four distinct data modalities accompanied by annotation data, including motion capture, RGB video, image, and 3D object modeling data. It features 63 action classes involving interactions with 60 common furniture and electronic devices. Each action class comprises approximately 1, 000 motion capture data representing 3D skeleton data and corresponding RGB video and 3D object modeling data, resulting in 67, 505 motion capture data samples. It offers comprehensive 3D structural information of the human, RGB images and videos, and point cloud data for 60 objects, collected through the participation of 126 subjects to ensure inclusivity and account for diverse human body types. To validate our dataset, we leveraged MMNet, a 3D human action recognition model, achieving Top-1 accuracy of 91.51% and 92.29% using the skeleton joint and bone methods, respectively. Beyond human action recognition, our versatile dataset is valuable for various 3D human action analysis research endeavors.

INDEX TERMS 3D human action analysis, human action recognition, human activity understanding, motion capture, multi-modal dataset.

I. INTRODUCTION

Securing high-quality human action datasets has become increasingly important due to the recent surge in research activities on human action analysis, which plays a vital role in computer vision, machine learning, and artificial intelligence. These datasets mainly consist of RGB videos, sequences of images, 3D structural information of humans, etc., and annotation of each action class. The annotation data represents the label and description of each action class. They are used to train and validate computer vision and machine learning models to analyze human actions. Various applications use human action analysis datasets, such as autonomous driving, security surveillance, user behavior detection, sports analysis, and medical diagnosis.

The analysis of human actions has traditionally been based on RGB videos [1]. However, the recent availability of depth cameras, such as Microsoft Kinect [2], [3], has enabled the tracking of motion sequences in 3D. This development has led to the emergence of multi-modality datasets, which include 3D skeleton information on human actions. These datasets have garnered attention in the research field and have been shown to achieve higher accuracy [4]. Several multi-modality

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

datasets exist for human action analysis, such as Toyota Smarthome [5], NTU RGB+D [6], [7], Northwestern-UCLA [8], and PKU-MMD [9]. These datasets include 3D structural data, representing human skeleton information for each frame, as well as RGB images and videos. Human action frequently involves interactions with objects, yet current datasets face constraints in furnishing object-related information and necessitate separate pre-processing for each modality for model training. To overcome these challenges, we introduce a novel dataset, **DGU-HAO**: **H**uman **A**ction analysis dataset with daily life **O**bjects, designed to address these limitations.

DGU-HAO is a versatile dataset in human action analysis research, encompassing tasks such as human action recognition, human action generation, human pose estimation, real-time detection, and more. In this paper, our dataset is specifically validated with a focus on human action recognition. To address the constraints seen in prior datasets, DGU-HAO meticulously gathered data from various subjects, considering variations in age, body types, and movement patterns, achieving a more balanced representation of these characteristics. Moreover, it offers information on objects, encompassing everyday furniture and electronic devices, highlighting their interactions with humans in point cloud data (PCD) format. Additionally, our dataset simplifies the data pre-processing task by consolidating video, 3D human structural, and label information into a single JSON file. For an in-depth description of our dataset, please refer to Section III.

We summarize our **key contributions** in this paper as follows:

- DGU-HAO explores a new realm by gathering data that includes common human actions related to the use of furniture and electronic devices in home and office environments. Our dataset stands out for offering 3D modeling information in PCD format for objects engaged in human interactions, providing a distinctive resource. This facilitates the exploration of the human-object interaction domain.
- DGU-HAO is appealing due to its multi-modality, featuring a total of four data modalities with sufficient data samples. Moreover, its excellence lies in the provision of annotation data in JSON format, enhancing user convenience in utilizing the dataset. This annotation data comprises comprehensive information about each action class, including details on objects, action classes, subjects, section tagging for RGB videos, and 3D coordinate information for all joints.
- DGU-HAO encompasses 126 subjects, considering variations in age, body shapes, movement patterns, and supplementary multi-modal information. This enriched diversity enhances the model's generalization capacity and facilitates action recognition in various environments.

The structure of this paper is as follows: Section II reviews previous research on 3D-based human action analysis datasets and deep learning algorithms. Section III describes the structure of the proposed dataset and how we built and pre-processed our dataset. Section IV explains the dataset evaluation results with the human action recognition algorithm and performance analysis. Finally, section V summarizes the paper, provides conclusions, and discusses future work.

II. LITERATURE REVIEW

A. 2D HUMAN ACTION DATASETS

The development of computer vision and pattern recognition technologies is significantly influenced by 2D human action datasets. Notable examples such as UCF101 [10], Kinetics [1], [11], [12], HMDB51 [13], and NTU RGB+D [6], [7] encompass a diverse range of activities, providing a comprehensive platform for evaluating algorithm performance across various scenarios. These datasets play a crucial role in advancing the understanding and capabilities of action recognition algorithms.

UCF101 [10] is a widely used benchmark dataset for human action recognition in videos, comprising 13, 320 video clips across 101 action categories. The dataset encompasses diverse activities such as sports, daily life, and various human interactions. Each video clip is captured under realistic conditions, providing a rich and challenging resource for evaluating the performance of action recognition algorithms.

The Kinetics [1], [11], [12] serves as an extensive benchmark for video-based action recognition, featuring approximately 650, 000 video clips that encompass 700 distinct human action classes. Encompassing a broad spectrum of activities, such as sports, routine actions, and intricate interactions, each video clip has a duration of about 10 seconds, and there are at least 700 video clips for each action class. The dataset is compiled from videos sourced from YouTube.

The HMDB51 [13] is a widely utilized benchmark dataset for human action recognition in videos, comprising 51 action classes. It consists of 6, 766 high-quality clips extracted from various sources, including movies and YouTube, covering a diverse range of actions such as sports, dancing, and everyday activities. Each action class contains at least 101 clips.

Nevertheless, as the majority of 2D datasets heavily rely on RGB images or videos, there exists a constraint in adequately conveying information about the depth of motion and spatial location. This limitation poses challenges in accurately discerning lateral shifts, obscured sections, and interactions with objects that occur during movement. Consequently, the introduction of a 3D human action dataset aimed to address these constraints. Leveraging advancements in motion capture sensors and depth cameras like Microsoft Kinect and the Optical Motion Capture System, the shortcomings of 2D datasets were mitigated by more effectively capturing the

TABLE 1. Comparison of the proposed DGU-HAO dataset and some other datasets for 3D action recognition. Our dataset provides point cloud data of 60 objects used in daily life. The JSON within the data modalities section tags annotation data for each data sample with metadata, including actor information, motion scenario details, object code, action class, and its code.

Datasets	# Samples	# Videos	# Classes	# Subjects	# Objects	# View	Data Modalities					Year
							RGB+D	IR	3D Joints	PCD	JSON	
Northwestern-UCLA [8]	1,475	1,475	10	10	-	3	✓	-	✓	-	-	2014
NTU RGB+D 60 [6]	56,880	56,880	60	40	-	80	✓	✓	✓	-	-	2016
PKU-MMD [9]	21,545	1,076	51	66	-	3	✓	✓	✓	-	-	2017
Toyota Smarthome [5]	16,115	16,115	31	18	-	7	✓	-	✓	-	-	2019
NTU RGB+D 120 [7]	114,480	114,480	120	106	-	155	✓	✓	✓	-	-	2019
Ours (DGU-HAO)	**67,505**	**208,875**	**63**	**126**	**60**	**15**	✓	**-**	✓	✓	✓	**2022**

spatial dimension and temporal characteristics of motion, incorporating depth information alongside RGB frames.

B. 3D HUMAN ACTION DATASETS

Table 1 compares five existing 3D human action datasets and the specifications of our DGU-HAO dataset.

One of these datasets, Northwestern-UCLA [8], encompasses RGB, depth, and 3D human skeleton data concurrently captured by three Kinect cameras. It comprises ten distinct action categories, including picking up with one hand, picking up with two hands, dropping trash, walking around, sitting down, standing up, donning, doffing, throwing, and carrying. Ten actors performed each of these actions, resulting in a dataset containing $1,475$ video and data samples. It's important to note that this dataset has limitations due to the small number of human action classes and actors included.

NTU RGB+D [6] is an expansive dataset, comprising a total of $56,880$ samples derived from 40 subjects engaged in 60 diverse daily life action classes. The dataset was recorded using three different views of RGB cameras, offering a rich array of modalities, including depth maps, 3D skeleton data encompassing 25 joints, RGB frames, and infrared information.

PKU-MMD [9] consists of $1,076$ untrimmed video sequences featuring 66 subjects captured from three different camera views. This dataset contains 5.4 distinct action categories annotated, yielding nearly $20,000$ action instances and a staggering 5.4 million frames.

As described in [5], the Toyota Smarthome dataset comprises 31 action motion classes and $16,115$ RGB+D videos executed by 18 subjects. Nonetheless, this dataset has limitations, including intra-class variations, class imbalances, similarities among different action classes, unequal video lengths, and fewer actors.

The NTU RGB+D 120 dataset, as introduced in [7], involves data contributed by 106 distinct subjects and encompasses over $114,000$ video samples captured across 155 different views, comprising 8 million frames. This extensive dataset covers 120 unique action classes, encompassing daily routines, interactive activities, and health-related actions. In [7], the authors have introduced an innovative framework known as Action-Part Semantic Relevance-aware (APSR) to enhance the reliability of one-shot 3D action recognition.

As depicted in Table 1, the DGU-HAO dataset showcases a remarkable range in terms of data volume compared to existing datasets [5], [6], [8], [9], with a maximum that is approximately 46 times larger and a minimum that is 1.18 times larger, excluding the dataset in [7]. When focusing on video data, our dataset stands out in both size and resolution, surpassing all other datasets [5], [6], [7], [8], [9]. It offers roughly twice the volume of video data compared to [7], which boasts the most substantial video data among existing datasets and an astonishing 194 times more video data than the dataset with the least video content [9]. While our dataset offers fewer action classes than [7], it outnumbers all other datasets [5], [6], [8], [9]. In terms of subjects, we have gathered data from a diverse pool of individuals, including various genders, heights, weights, and ages, thus enhancing the overall robustness of our dataset.

Furthermore, our dataset includes point cloud data for 60 objects interacting with humans. Additionally, each data sample is accompanied by an annotation file containing meticulously refined motion capture information, presented in a straightforward 1:1 correspondence in JSON format. This structure greatly simplifies data processing and utilization.

C. HUMAN ACTION RECOGNITION NETWORKS

The Video-Pose Network (VPN) in [14] is integrated into the top layer of a 3D convolutional network, comprising two main components: the attention network and spatial embedding processes. The attention network involves the pose backbone and spatio-temporal coupler, which transforms 3D pose input into a graph format using Graph Convolutional Network (GCN) [15] to derive features for each 3D pose, including attention weights capturing spatial-temporal characteristics. The spatial embedding process improves alignment between RGB images and 3D poses by measuring distance mapping in the embedding space, enabling accurate identification of similar 3D pose operations using RGB images. This research is significant for predicting human behavior by combining RGB and 3D Pose skeletons, although it comes at the cost of slower processing speed.

VPN++ [16] is an advanced network that addresses the shortcomings of VPN [14]. It transforms existing VPN into VPN-F (VPN-Feature) and VPN-A (VPN-Attention) and combines them to form VPN++. Both VPN-F and VPN-A are teacher-student networks. The difference from VPN is that VPN++ uses only RGB images to reduce the time required for testing. Although the time required for action recognition has been reduced, it still exhibits a slower speed. It faces challenges related to lower accuracy when compared to recently developed action recognition networks like PoseC3D [17] or MMNet [18].

PoseC3D [17] is a 3D-CNN-based approach for skeleton-based action recognition, which takes 3D heatmap volumes as input. 3D-CNN-based approaches first extract 2D poses as coordinates from frames of a video. After extracting 2D poses to be input into PoseC3D, the model generates pseudo heatmaps for joints and limbs by stacking 2D heatmaps along the temporal dimension, creating a 3D representation. PoseC3D outperforms the GCN-based approach for robustness, interoperability, and generalization. Our dataset is specifically structured to utilize 3D coordinate values of individual keypoints as the model input. However, PoseC3D operates by utilizing the 2D coordinate values of individual keypoints in each frame. Consequently, PoseC3D may not be the most suitable model for validating our dataset due to this inherent mismatch in the input data format.

Model-based Multi-modal Network (MMNet) [18] is an ensemble model based on GCN [15] and CNN models. The input data of the MMNet is 3D skeleton data and RGB videos; the model's output is action class. The pre-processing consists of mainly 2 phases to train this model. First, extract joints and bones from the 3D skeleton data, respectively. Secondly, extract the spatio-temporal region of interest (ST-ROI) from the RGB videos. The framework of the MMNet model is constructed with 3 individual networks. First, a GCN network for training joints from 3D skeleton data. Second, a GCN network for training bones from 3D skeleton data. Lastly, a CNN-based ResNet [19] network for training ST-ROI images from RGB videos. Finally, MMNet recognizes action class by the ensemble of those 3 individual networks.

Hence, we decided that the MMNet [18] model was more appropriate for validating our dataset, so we used MMNet to validate the data. Our dataset comprises 3D motion capture data, RGB videos and images, and 3D object modeling data. The MMNet model extracts and uses 3D skeleton data from 3D motion capture data. The PoseC3D [17] model extracts 2D skeleton data from a 2D RGB image, stacks the 2D skeleton data according to the time dimension, and uses it as 3-channel data. In other words, human action data itself is not 3D data consisting of x, y, and z axes. We tracked human action using a motion capture sensor and obtained 3D coordinate values of human action for the x, y, and z axes. When using PoseC3D, it extracts its own 2D coordinate value from RGB, making it challenging to properly verify the 3D motion capture data we built. Therefore, we decided that the MMNet model, which extracts 3D skeleton joints and bones from 3D motion capture data and uses them as input data, is more suitable for verifying our data.

III. DATASET STRUCTURE

A. DATA COLLECTION

The motion capture procedure occurred in a controlled environment, where video recording, 3D motion capture, and

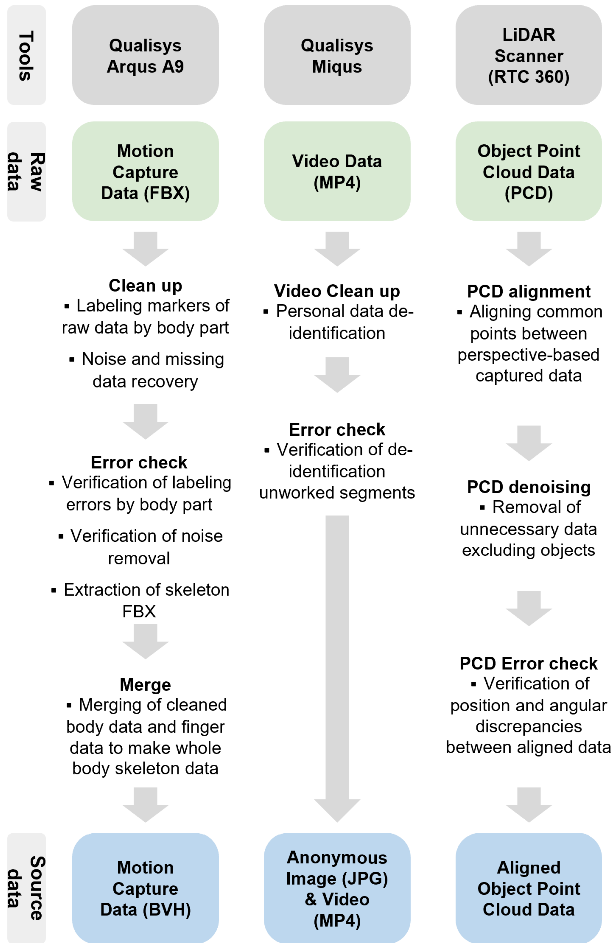
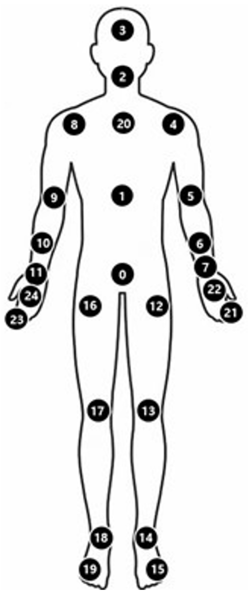


FIGURE 1. All data types were collected and built simultaneously. The motion capture data coordinates of the finger were collected separately from the motion capture data of the body part using MoCap Pro Super Splay, a hand motion data collection device. The finger motion capture data coordinates were merged with the body motion capture data coordinates according to the human skeleton's hierarchical structure.

object point cloud data collection occurred simultaneously. This environment was carefully set up to avoid light reflections and covered a range of 6 to 15 meters. The setup utilized twelve Qualisys Arqus A9 cameras (Qualisys), three Qualisys Miqus cameras (Qualisys), and a LiDar Scanner (RTC 360, Leica), as illustrated in Fig. 1. To ensure precision, calibration rods were employed to measure the capture area and fine-tune camera settings, including lens distortion, angles, and positions. Once calibrated, the cameras defined the designated capture area where participants were instructed to perform their actions. These participants wore specialized suits with markers attached to key joint reference points, enabling the capture of 3D spatial information. Before the actual motion capture, basic motions were recorded to assess capture quality and optimize settings through software and hardware adjustments. The data acquired from this optical motion capture setup served as the initial raw data, which was subsequently processed and refined into a pre-processed format.



No.	Label	No.	Label
0	Hips	13	Left Leg
1	Spine1	14	Left Foot
2	Neck	15	Left Toe
3	Head	16	Right Upper Thigh
4	Left Arm	17	Right Leg
5	Left Forearm	18	Right Foot
6	Left Wrist	19	Right Toe
7	Left Hand	20	Spine2
8	Right Arm	21	Right Finger
9	Right Forearm	22	Right Thumb
10	Right Wrist	23	Left Finger
11	Right Hand	24	Left Thumb
12	Left Upper Thigh		

FIGURE 2. **Configuration of the body joints and label in our dataset.**

In the conversion to BVH format, the initial raw data underwent processing with QTM (Qualisys Track Manager) and underwent noise reduction procedures. Simultaneously, the video data in MP4 format anonymizes sensitive information, safeguarding the privacy of individuals.

The dataset includes 3D skeleton data and RGB format data, which was captured via a camera. Using a camera, the RGB dataset offers visual appearance information in images photographed from multiple angles. Specifically, three different RGB camera views were employed, positioned at 0, -30 degrees, and 30 degrees with consistent height. This configuration enables the capture of actions from the front, left, and right perspectives. When utilizing the RGB data for training purposes, it can be segmented and applied per frame. This flexibility enhances performance by training multi-modal data, using the corresponding visual data beyond the skeletal information.

B. DATASET STRUCTURE

Our dataset consists of 67, 505 motion capture data samples involving 126 subjects interacting with 60 different objects across 63 unique action classes. It was meticulously designed to emphasize clear differentiations between actions of similar nature. To achieve this, we harnessed the power of multiple modalities, incorporating both RGB frames and 3D skeletal joint positions. This multi-modal approach equips the model with a comprehensive understanding of various facets of the data, enabling it to deduce contextual nuances and ultimately enhancing its performance. Furthermore, our dataset includes detailed labeling of 25 body joints, and you can observe the configuration of these joints in Fig. 2.

1) DATA MODALITIES

In this research, we introduce a dataset gathered through the utilization of the Qualisys Arqus A9 optical motion capture system. This comprehensive dataset encompasses diverse data types, including motion capture, video recordings, images, and 3D modeling information, amounting to 67, 505 individual samples.

TABLE 2. **Data modality and description of each modality.**

Data modality	File Format	Description
Motion Capture Data (MCD)	BVH	- 3D coordinate of joint - # of joint: 25
RGB Video	MP4	- Resolution: 1920 * 1080 - # of view: 3
Thumnail Images	JPG	- Still photo of the video - Five per video
3D Object Modeling	FBX	- 3D modeling of point cloud data - 40 furniture & 20 electronic devices
Annotation Data	JSON	- Metadata of the other data modalities - Configuration: pre-processed MCD, action code, action class, object name, object code, actor id, video section tagging, etc.

TABLE 3. **Configuration of annotation data name format.**

Field Name		Meaning
Object Code	Object Type (T)	Furniture (FN), Electronic Devices (EL)
	Object Name (N)	Chair (CH), Sofa (SO), Desk (DE), Table (TB), Other Furniture (oF), Home Equipment (HE), Office Equipment (OE)
Action Class		A01 ∼ A63
Gender (G)		Female (F), Male (M)
Age Group (A)		Young (Y), Middle Age (M), Old Age (O)
Body Shape Info (B)		H01, H02, H03, L01, L02, L03, M01, M02, M03
Actor ID		P001, P002, ...
Data ID		001, 002, ...
File Name Regular Expression		T_N_ActionClass_G+A+B_ActorID_DataID (ex. FN_SO05_A01_FMH01_P109_001)

Our dataset is structured in various formats to accommodate different aspects of the data. The motion capture data is provided in the BVH file format, with each frame containing the 3D coordinates of the joints. We offer video data stored in MP4 files for visual representation, capturing the actions dynamically. Additionally, image data is presented in JPG format, showcasing static frames of the actions, as depicted in Table 2. Furthermore, the 3D object modeling data is available in the FBX format, encompassing the hierarchical skeleton structure and joint details. The dataset includes annotation (JSON) data, encompassing segment tagging, meta-information, action scenarios, and labeling information for video segments. The Annotation data name format configuration is shown in Table 3. The annotation data contains skeleton coordinate values from the motion capture data converted to JSON format per joint to facilitate the use of the data.

2) ACTION CLASSES

We first selected 60 objects frequently used in daily life. Then, we selected 63 action classes by considering how people interact with those 60 objects and referring to the

TABLE 4. Configuration of the subject body types and age groups.

Gender	Age Group	Height (cm)	Weight (kg)
Male	$10 \sim 29$	$157 \sim 174$,	$46 \sim 61$
		$175 \sim 179$,	$62 \sim 77$
		$180 \sim 191$	$78 \sim 143$
	$30 \sim 49$	$154 \sim 173$,	$50 \sim 62$
		$174 \sim 179$,	$63 \sim 75$
		$180 \sim 191$	$76 \sim 130$
	$50 \sim 69$	$150 \sim 162$,	$49 \sim 59$
		$163 \sim 169$,	$60 \sim 70$
		$170 \sim 187$	$71 \sim 114$
Female	$10 \sim 29$	$143 \sim 160$,	$36 \sim 44$
		$161 \sim 166$,	$45 \sim 54$
		$167 \sim 178$	$55 \sim 93$
	$30 \sim 49$	$140 \sim 154$,	$38 \sim 50$
		$155 \sim 163$,	$51 \sim 62$
		$164 \sim 186$	$63 \sim 114$
	$50 \sim 69$	$140 \sim 151$,	$37 \sim 48$
		$152 \sim 157$,	$49 \sim 60$
		$158 \sim 175$	$61 \sim 109$

NTU RGB+D dataset [6], [7]. For example, in the case of a cell phone, a person's interaction with the cell phone may include pressing a button on the cell phone, answering a call, and immediately hanging up after answering the call (rejecting the call). Each action class consists of a total of three phases: the subject approaches the object, uses the object, and retreats after use. The inventory of action classes has been meticulously arranged and can be found in Table 5 and Table 6. There are 63 distinct action categories, each classified according to the furniture or electronic devices commonly employed in daily activities.

3) SUBJECTS AND OBJECTS

We selected age-specific body type standards by considering differences in movement patterns depending on the object user's age and physique and constructed a dataset by recruiting at least 126 subjects evenly from the corresponding body type standard distribution. Configuration of the body type standards by age are provided in Table 4. We evenly selected 126 subjects according to the above body type standard table considering height and weight based on the standard information on the Korean human body measured through the Korean Human Body Measurement Survey. Therefore, the male-to-female ratio is 53.5:46.5, and the age ratio is 37% in the 10s $\sim$ 20s, 35% in the 30s $\sim$ 40s, and 29% in the 50s $\sim$ 60s. Additionally, we recruited subjects from the general public so that their natural behavioral characteristics could be demonstrated. Based on scenarios and action classes in Table 5 and Table 6, we classified various subjects and constructed a dataset by configuring the approaches, uses, and retreat phases for the objects. Therefore, because we constructed the dataset using a wide subject spectrum, our dataset is effective in generalizing.

Our dataset encompasses 60 everyday objects commonly encountered in household and office settings. These objects are categorized into two main groups: furniture types and appliances. The furniture types include a variety of items such as chairs (12 types), sofas (5 types), desks

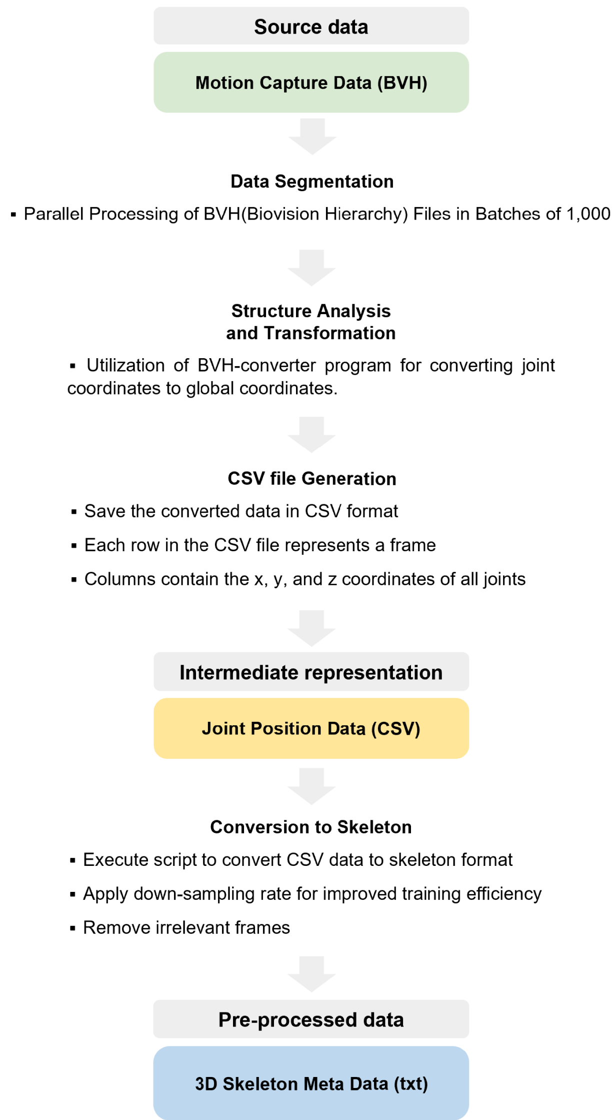


FIGURE 3. Overview of the data pre-processing to evaluate with the MMNet model. To ensure the model data loader could properly process our motion capture data obtained from Fig. 1, we converted the motion capture data in BVH format into 3D skeleton metadata in text format.

(10 types), objects placed on tables (7 types), and various other furniture pieces (e.g., wardrobes, beds, sinks, etc., totaling six types). Meanwhile, the appliances category encompasses office equipment (e.g., computers, copiers, etc., comprising 11 types) and home appliances (e.g., refrigerators, washing machines, etc., totaling nine types). A total of 63 action classes interact with 60 3D-modeled objects.

IV. DATASET PRE-PROCESSING AND EVALUATION WITH MMNeT

A. DATA PRE-PROCESSING

We used the MMNet model [18] to evaluate the DGU-HAO dataset. Therefore, we outline the pre-processing procedures involved in acquiring the skeleton data, which serves as the

TABLE 5. 15 action classes and its action code with 40 furniture objects with its object code. All action classes belong to each motion scenario, with seven motion scenarios. Each action class is interacting with one furniture object.

Object Code	Object Name	Motion	Action Class	Action Code
CH01	Back chair(leg)	Sitting down or Standing up on a chair/sofa	Sit down with a hand support	A01
CH02	Back chair(wheeled)			
CH03	Desk chair(wheeled)		Sit down without a hand support	A02
CH04	Stool(leg)			
CH05	Stool(wheeled)		Sit with crossed legs	A03
CH06	Dining chair			
CH07	Bar chair		Crossing legs	A04
CH08	Rocking chair			
CH09	Legless chair	Lying down or Standing up on the sofa	Lying down with face forward	A05
CH10	Folding chair(camping)			
CH11	Swing chair			
CH12	Sunbed		Lying on the one's side	A06
SO01	Sofa(straight)			
SO02	Sofa			
SO03	Recliner sofa		Lying down on one's stomach	A07
SO04	Folding sofa			
SO05	Bean bag			
DE01	Desk(without drawer)	Standing	Perching on the desk	A08
DE02	L-shaped desk		Retrieve an item from the shelf	A09
DE03	H-shaped desk		Placing an item	A10
DE04	Height adjustable desk		Moving an item	A11
DE05	All-in-one desk		Wiping the tabletop	A12
DE06	Sitting desk			
DE07	Meeting room desk			
DE08	Round desk			
DE09	Standing desk	Sitting	Resting the chin on one's hand	A13
DE10	Built-in desk			
TB01	Business table			
TB02	Round table		Lying face down	A14
TB03	Folding table			
TB04	Storage table			
TB05	Sitting table			
TB06	Height adjustable table			
TB07	Standing table			
OF01	Closet	Putting things into the cabinet	Putting in the item and close	A15
OF02	Double door closet			
OF03	Desk			
OF04	Drawer			
OF05	Cabinet			
OF06	Bed	Perching on the bed	Sit down with a hand support	A01
			Sit down without a hand support	A02
			Sit with crossed legs	A03
		Lying on the bed	Lying down with face forward	A04
			Lying on the one's side	A05
			Lying down on one's stomach	A06

input for the MMNet model [18] in this section. We also elaborate on the construction of the pre-processing pipeline. The sequence of pre-processing steps is shown in Fig. 3.

1) BVH TO CSV

The BVH files initially contained joint positions in a relative hierarchy, but our model necessitated global joint coordinates for each frame during training. To achieve this conversion from BVH files to CSV files containing joint positions, we employed the bvh-converter tool and utilized the BVH parser from cgkit.

The BVH parser meticulously analyzed the file structure, extracting both joint positions and rotation information. Subsequently, it converted the relative positions into global coordinates using a 'ZYX' Euler rotation sequence. This resulted in the extraction of 3D coordinates for each joint in every frame, storing the data in CSV format. In this format, each row corresponds to a frame, and the columns contain the x, y, and z coordinates of the joints. Considering the substantial size of the dataset, the conversion process was notably time-consuming. To address this, we implemented parallel processing in batches.

2) CSV TO SKELETON

To validate our dataset using the human action recognition model, we select 25 main joints from 75 keypoints refer to [6], [7] as shown in Fig. 2. We pre-processed the CSV data into 3D skeleton data, formatted identically to the NTU-RGB+D dataset [6], [7]. The information extracted from the CSV file for each frame is then written into a new skeleton file. This 3D

TABLE 6. 38 action classes and its action code with 20 electronic device objects with its object code. All action classes belong to each motion scenario, with 21 motion scenarios. Each action class is interacting with one furniture object.

Object Code	Object Name	Motion	Action Class	Action Code
HE01 HE02 HE03	Refrigerator Double door refrigerator Lid Kimchi refrigerator	Opening and Closing the refrigerator door	Putting in the item	A16
			Checking the interior	A17
		Using easy home-bar	Putting in the item	A18
			Checking inside	A19
HE04 HE05	Washing machine Drum washing machine	Opening and Closing the washing machine door	Put things in the laundry	A20
			Add detergent and Operating	A21
			Taking the things out	A22
HE06	Microwave	Operating the microwave buttons	Putting in the item	A23
			Checking inside	A24
			Pushing the buttons	A25
HE07	Coffee machine	Put ingredients into the coffee machine	Fill up with water	A26
			Fill up with coffee beans	A27
		Operating the coffee machine buttons	Brewing coffee	A28
		Steaming the milk in the coffee machine	Steaming the milk and wipe	A29
HE08	Electric kettle	Opening and closing the lid of an electric kettle	Fill with water	A30
			Pouring water	A31
		Moving the electric kettle	Bring something	A32
			Take something	A33
HE09	Vacuum cleaner	Clean up	Clean while standing	A34
			Cleaning while bending over	A35
OE01 OE02	Wireless phone Wired phone	On a call	Operating the buttons	A36
			Answer the phone and hang up	A37
			Picking up the phone and immediately hang up	A38
OE03	Desktop	Manipulating the desktop	Manipulating the keyboard	A39
			Manipulating the computer-mouse	A40
			Operating the power button	A41
			Plug in the USB	A42
			Manipulating the monitor	A43
OE04	Laptop	Manipulating the laptop	(Open the top part) Manipulating the keyboard	A44
			(Open the top part) Manipulating the mouse	A45
			(Open the top part) Manipulating the touchpad	A46
			(Open the top part) Manipulating the power button	A47
			(Open the top part) Plug in the USB	A48
			Close the top part	A49
OE05	Multi-Function Printer	Manipulating the large multifunction printer	Scan a flatbed	A50
			Copy the feeder	A51
OE06	Small printer	Manipulating the small printer	Refill the paper	A52
			Refill a toner	A53
OE07	Document shredder	Shredding documents	Put the paper inside	A54
OE08	Coating machine	Coating the documents	Coating	A55
OE09	Cash counter	Counting banknotes	Count the banknotes	A56
OE10	Office safe	Open and close the safe	Put the things inside	A57
			Take out the item	A58
OE11	Shopping cart	Moving shopping carts	Pushing the cart forward	A59
			Pull the cart backwards	A60
			Avoiding obstacles	A61
		Loading and unloading luggage	Unloading luggage	A62
			Loading luggage	A63

skeleton data encompasses essential information, including the total frame count, the number of detected individuals, details for each person, joint count, and the 3D coordinates of each joint. As we deal with the motions of a single person, the number of recognized joints remains constant at 25. The positions for the right and left hands are defined as the average of the four sets of 3D coordinates for each hand. The 3D skeleton data, sensitive to even minor positional variations, plays a pivotal role and greatly influences the accuracy of 3D human action recognition.

We applied a down-sampling rate of 10 frames per second to enhance training efficiency, removing frames unrelated to the specified action classes to eliminate extraneous data.

B. EVALUATION ENVIRONMENT

In this study, we used the MMNet model [18] to validate our dataset. The model was trained on 54, 334 samples (80.48%) and evaluated on 13, 171 samples (19.52%), with

TABLE 7. The hardware specifications.

Element	Specification
CPU	Intel Xeon Silver 4210 2.20GHz
Memory	256GB
GPUs	Nvidia GeForce RTX 3090 $\times 5$
OS	Ubuntu 18.04
Framework	Pytorch 1.7.0 + CUDA 11.2

TABLE 8. Configuration of the hyperparameters.

Hyperparameter	Value
Base learning rate	0.1
Lambda_L1	1e-05
Lambda_L2	0.0001
Optimizer	Stochastic Gradient Descent (SGD)
Drop out	0.5
Weight decay	0.0001
Batch size	64
# of Epoch	15

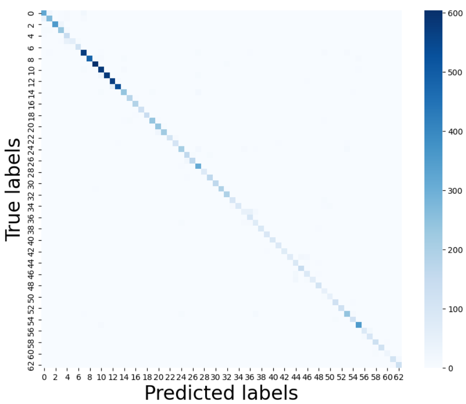


FIGURE 4. Visualizing a confusion matrix for 63 action classes based on skeleton joint data, where the x-axis represents predicted labels and the y-axis represents true labels.

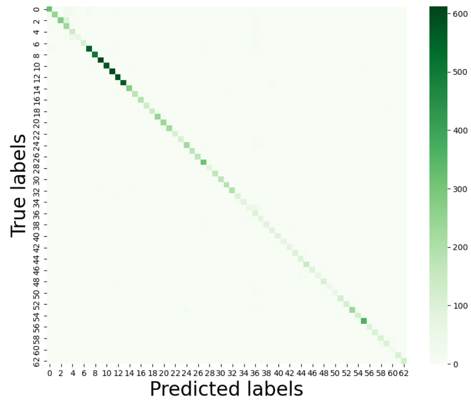


FIGURE 5. Visualizing a confusion matrix for 63 action classes based on skeleton bone data, where the x-axis represents predicted labels and the y-axis represents true labels.

TABLE 9. Top 10 accurate action classes of the different methods on our dataset. The rankings were organized according to accuracy, and additional evaluation metrics such as F1 score, precision, and recall were employed.

Method	Action Code	Rank	Accuracy	F1 Score	Precision	Recall
Skeleton Joint	1	A53	99.98%	99.19	99.19	99.19
	2	A43	99.97%	100.00	96.00	97.96
	3	A29	99.97%	96.39	98.77	97.56
	4	A39	99.97%	98.90	96.77	97.83
	5	A41	99.97%	97.85	97.85	97.85
	6	A60	99.97%	100.00	96.75	98.35
	7	A18	99.95%	99.21	96.18	97.67
	8	A40	99.95%	93.94	100.00	96.88
	9	A31	99.93%	97.63	97.06	97.35
	10	A32	99.93%	100.00	95.59	97.74
Skeleton Bone	1	A18	99.98%	98.50	100.00	99.20
	2	A29	99.97%	96.40	98.80	97.60
	3	A30	99.96%	98.80	98.20	98.50
	4	A33	99.96%	98.00	99.50	98.80
	5	A39	99.96%	97.80	96.80	97.30
	6	A41	99.96%	94.90	100.00	97.40
	7	A42	99.96%	100.00	94.00	96.90
	8	A19	99.95%	99.30	96.60	98.00
	9	A53	99.95%	100.00	95.20	97.50
	10	A43	99.94%	92.50	98.70	95.50

accuracy calculated using confusion matrices generated during training.

The hardware specifications used for evaluating the dataset are provided in Table 7. Additionally, Table 8 presents the hyperparameter configuration for the MMNet model, with most of the parameters following the established MMNet model settings [18].

C. EVAULATION RESULTS

In this study, two approaches were employed for data validation: the first method involved training the model using the 3D coordinate values of 25 joints as skeletal joints, while the second method, known as skeleton bone, entailed model training by connecting joints linked through the body's bones among the 25 joints, augmenting the dataset with real human skeleton information.

Figures 4 and 5 provide visual representations of the confusion matrices derived from the training results, where darker colors indicate higher values, with the x-axis denoting model-predicted labels and the y-axis representing ground truth labels. In Fig. 4, the diagonal matrix exhibits the highest values corresponding to the skeleton joint method, signifying effective learning in predicting action classes. Similarly, in Fig. 5, illustrating the results for the skeleton bone method, the diagonal matrix positions are the darkest, indicating successful training in action class prediction and affirming the dataset's quality and integrity.

Table 9 presents the results of sorting the ten most accurate action classes by method. In the skeleton joint method, the action 'A53: Refill a toner' achieved the highest accuracy at 99.98%. Similarly, in the skeleton bone method, the action 'A18: Putting in the item' showed the highest accuracy, also at 99.98%. Six out of the top 10 accurate action classes were common to both methods, demonstrating the robustness of the dataset.

TABLE 10. Top 10 misclassified action classes of the different methods on our dataset. The rankings were organized according to accuracy, and additional evaluation metrics such as F1 score, precision, and recall were employed.

Method	Action Code	Rank	Accuracy	F1 Score	Precision	Recall
Skeleton Joint	1	A01	99.15%	86.10	84.30	85.20
	2	A02	99.25%	87.40	82.10	84.60
	3	A05	99.32%	63.70	94.70	76.10
	4	A08	99.33%	92.60	93.30	92.90
	5	A36	99.38%	71.20	43.10	53.70
	6	A11	99.38%	91.50	95.60	93.50
	7	A13	99.43%	91.20	96.60	93.80
	8	A38	99.43%	62.30	73.10	67.30
	9	A14	99.44%	97.60	90.00	93.70
	10	A37	99.46%	60.60	92.40	73.20
Skeleton Bone	1	A04	98.67%	62.00	88.30	72.80
	2	A37	98.86%	41.00	98.10	57.90
	3	A03	99.12%	99.30	71.10	82.90
	4	A01	99.14%	86.70	83.30	85.00
	5	A05	99.18%	59.00	94.00	72.50
	6	A02	99.24%	92.90	75.40	83.20
	7	A08	99.35%	94.20	91.90	93.00
	8	A36	99.43%	73.60	48.60	58.60
	9	A06	99.56%	96.90	52.90	68.50
	10	A25	99.58%	90.00	88.20	89.10

Table 10 displays the results of sorting the ten action classes with the lowest accuracy by the method. In the skeleton joint method, the action 'A01: Sit down with a hand support' had the lowest accuracy at 99.15%. For the skeleton bone method, the action 'A04: Crossing legs' showed the lowest accuracy, but still achieved 98.67% accuracy. However, compared to the high accuracy for misclassification, it was confirmed that the performance was poor in the F1-score and other performance indicators as shown in Table 10. Similar to Table 9, six action classes out of the bottom 10 were common to both methods, reinforcing the dataset's robust construction.

Furthermore, an observation revealed that most of the top 10 accurate action classes involved interactions with office equipment rather than home appliances. This suggests slightly higher accuracy for actions related to office equipment with distinct characteristics compared to similar actions involving home appliances. However, it's noteworthy that even the lowest accuracy, 98.67% for 'A04: Crossing legs' in the skeleton bone method, is relatively high. Additionally, with only about a 1.3% difference between the lowest and highest accuracies (99.98%), it's evident that all action class data are evenly constructed.

Table 11 below compares test accuracy between our dataset and other datasets trained using the MMNet model. The accuracy results for datasets other than ours are based on the MMNet model [18]. There is a discrepancy in the experimental settings—the MMNet paper employed 80 epochs, whereas this paper utilized only 14 epochs.

For the skeleton joint method, our dataset achieved an accuracy of 91.51%, slightly surpassing that of the PKU-MMD dataset and exhibiting the highest accuracy among the compared datasets. On the other hand, the skeleton bone method recorded an accuracy approximately 1.1% lower than the PKU-MMD dataset. Despite this, it ranked second in accuracy among the seven datasets, demonstrating the robust quality of our dataset.

TABLE 11. Comparison of the test accuracy of the MMNet model in our dataset and the MMNet model in other human action recognition datasets for each skeleton joint (SJ) and skeleton bone (SB) method.

Dataset	Top1 Accuracy		Epoch
	SJ	SB	
N-UCLA [8], [18]	84.20%	83.50%	80
NTU RGB+D [6], [18]	80.40%	84.40%	
PKU-MMD [9], [18]	91.50%	93.40%	
Toyota Smarthome [5], [18]	66.60%	66.30%	
NTU RGB+D 120 [7], [18]	79.00%	81.00%	
Ours (DGU-HAO)	91.51%	92.29%	14

V. CONCLUSION

This paper introduces a novel motion capture dataset tailored for human action analysis, comprising an extensive collection of 67, 505 video samples across 63 diverse action categories. The dataset encompasses multiple data modalities, including RGB images, videos, object point cloud data, and 3D skeleton data, each associated with every action class, facilitating versatile model training. The inclusion of a wide array of human subjects has enabled the creation of a realistic benchmark for human action recognition. When compared to other 3D human action datasets (N-UCLA, NTU RGB+D, PKU-MMD, Toyota Smarthome, and NTU RGB+D 120) evaluated under the same conditions using the MMNet algorithm, our dataset demonstrated notable performance, particularly in accuracy. Specifically, the Skeleton Joint method exhibited the highest accuracy among the datasets, achieving a top-1 accuracy of 91.51%. The Skeleton Bone method produced the most favorable results, boasting a top-1 accuracy of 92.29%, surpassing even the PKU-MMD dataset, which achieved a top-1 accuracy of 93.40%. Notably, the difference in top-1 accuracy between PKU-MMD and our dataset in the Skeleton Bone method was merely 1.11%, indicating a minimal distinction. Therefore, the experimental outcomes underscore the utility of our motion capture dataset as input for human action analysis models. Our dataset is a general-purpose dataset that can be used for multiple studies that analyze 3D human actions. In this paper, our dataset was verified using a human action recognition model, MMNet [18], but it is possible to apply various models, such as human action generation and human-object interaction. Nevertheless, it is essential to acknowledge the limitation of our current evaluation, which solely focuses on human action recognition. To address this limitation, we plan to propose human action-object recognition networks, leveraging both 3D skeleton data and object point cloud data to enhance model performance.

ACKNOWLEDGMENT

The dataset was built by DTAAS consortium. Informed Consent was obtained from all the human subjects who participated in the data collection.

Dataset access: https://shorturl.at/mFOPW (accessed on 6 January 2024).

Dataset access (non-Korean): https://github.com/CSID-DGU/NIA-MoCap-1 (accessed on 6 January 2024).

REFERENCES

[1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.

[2] Z. Zhang, "Microsoft Kinect sensor and its effect," *IEEE Multimedia Mag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[3] J. Han, L. Shao, D. Xu, and J. Shotton, "Enhanced computer vision with Microsoft Kinect sensor: A review," *IEEE Trans. Cybern.*, vol. 43, no. 5, pp. 1318–1334, Oct. 2013.

[4] C. Tang, A. Tong, A. Zheng, H. Peng, and W. Li, "Using a selective ensemble support vector machine to fuse multimodal features for human action recognition," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–18, Jan. 2022.

[5] R. Dai, S. Das, S. Sharma, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2533–2550, Feb. 2023.

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[7] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.

[8] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu, "Cross-view action modeling, learning and recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2649–2656.

[9] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*.

[10] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*.

[11] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*.

[12] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," 2019, *arXiv:1907.06987*.

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.

[14] S. Das, S. Sharma, R. Dai, F. Bremond, and M. Thonnat, "VPN: Learning video-pose embedding for activities of daily living," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 72–90.

[15] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[16] S. Das, R. Dai, D. Yang, and F. Bremond, "VPN++: Rethinking video-pose embeddings for understanding activities of daily living," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9703–9717, Dec. 2022.

[17] H. Duan, Y. Zhao, K. Chen, D. Lin, and B. Dai, "Revisiting skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2959–2968.

[18] B. X. B. Yu, Y. Liu, X. Zhang, S.-H. Zhong, and K. C. C. Chan, "MMNet: A model-based multimodal network for human action recognition in RGB-D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3522–3538, Mar. 2023.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

JIHO PARK received the B.S. degree in computer science and engineering from Dongguk University, South Korea, in 2022, where she is currently pursuing the master's degree in artificial intelligence. Her research interests include artificial intelligence, machine learning, computer vision, and multi-modal learning.

JUNGHYE KIM is currently pursuing the bachelor's degree in information and communication engineering and in data science software with Dongguk University, South Korea. Her research interests include multi-modal learning and visual understanding.

YUJUNG GIL received the B.S. degree in computer science and engineering from Dongguk University, in 2023. Her current research interests include computer vision, cloud, and artificial intelligence.

DONGHO KIM received the B.S. degree in computer engineering from Seoul National University, South Korea in 1990, and the M.S. and Ph.D. degrees in computer science from the University of Southern California, Los Angeles, CA, USA, in 1992 and 2002, respectively. He is currently a Professor with the Software Education Institute, Dongguk University, South Korea. His research interests include artificial intelligence, distributed systems, networks, and security.