electronics

Article

DGU-HAU: A Dataset for 3D Human Action Analysis on Utterances

Jiho Park [1,†], Kwangryeol Park [2,†] and Dongho Kim [3,*]

[1] Department of Artificial Intelligence, Dongguk University, Seoul 04620, Republic of Korea; jiho8345@dgu.ac.kr
[2] Department of Computer Science and Engineering, Dongguk University, Seoul 04620, Republic of Korea; 2018112010@dgu.ac.kr
[3] Software Education Institute, Dongguk University, Seoul 04620, Republic of Korea
* Correspondence: dongho.kim@dgu.edu
† These authors contributed equally to this work.

Abstract: Constructing diverse and complex multi-modal datasets is crucial for advancing human action analysis research, providing ground truth annotations for training deep learning networks, and enabling the development of robust models across real-world scenarios. Generating natural and contextually appropriate nonverbal gestures is essential for enhancing immersive and effective human–computer interactions in various applications. These applications include video games, embodied virtual assistants, and conversations within a metaverse. However, existing speech-related human datasets are focused on style transfer, so they have limitations that make them unsuitable for 3D human action analysis studies, such as human action recognition and generation. Therefore, we introduce a novel multi-modal dataset, **DGU-HAU**, a dataset for 3D human action on utterances that commonly occurs during daily life. We validate the dataset using a human action generation model, Action2Motion (A2M), a state-of-the-art 3D human action generation model.

Keywords: 3D human action analysis; human activity understanding; motion capture; multi-modal dataset; utterance dataset

Citation: Park, J.; Park, K.; Kim, D. DGU-HAU: A Dataset for 3D Human Action Analysis on Utterances. *Electronics* **2023**, *12*, 4793. https://doi.org/10.3390/electronics12234793

Academic Editor: Yue Wu

Received: 11 October 2023
Revised: 21 November 2023
Accepted: 25 November 2023
Published: 27 November 2023

1. Introduction

Human action analysis research is important in understanding and interpreting human behavior from various perspectives. This field is crucial for multiple applications, including video surveillance, healthcare, robotics, sports analysis, and entertainment. Human action analysis research can enhance safety, efficiency, and automation in various industries by accurately recognizing, predicting, and modeling human actions.

Constructing datasets suitable for human action analysis is significant for advancing this research domain. The datasets provide essential ground truth annotations for training and evaluating deep learning networks of human action analysis research such as action recognition, action prediction, action generation and modeling, pose estimation, real-time action analysis, etc. As diverse and complex human actions span multiple contexts and environments, the datasets allow researchers to develop robust models that generalize well across real-world scenarios. Furthermore, well-structured datasets foster healthy competition within the research community, inspiring the development of more accurate and efficient action analysis techniques. Previous human action analysis datasets were unimodal, mainly based on RGB images or videos [1]. These datasets do not contain depth information, so they need pre-processing to reconstruct the 3D skeleton. With the introduction of depth sensors, such as Microsoft Kinect [2,3] and IR cameras [4,5], it is possible to build multi-modal human action analysis datasets containing RGB, depth, and 3D skeleton data. Therefore, we aim to introduce a general-purpose human action analysis dataset and validate the dataset with the human action generation model [6] in this paper.

The generation of human-like movements has garnered significant attention and research across disciplines such as computer vision, graphics, and animation. This field aims to develop algorithms and models that can produce realistic and natural movements resembling those of humans. By capturing the intricacies of human motion, many studies strive to enhance the quality and believability of virtual characters, avatars, and animations, thereby creating immersive experiences in various domains, including entertainment, virtual reality, and robotics.

To achieve truly immersive and effective human–computer interactions, generating nonverbal gestures that appear natural and appropriate is crucial across a range of conversational scenarios. This necessity has emerged in various applications, including communication with characters in video games, embodied virtual assistants, and avatars conversing in a metaverse. In video games, lifelike character animations convey emotions, intentions, and interactions, enabling players to engage with the virtual world more deeply. Embodied virtual assistants, such as chatbots or virtual agents, can benefit from natural gestures to enhance their expressiveness and facilitate more intuitive communication with users. Moreover, as the concept of the metaverse continues to evolve, avatars engaging in conversations within this virtual realm will require nonverbal gestures that are contextually appropriate, enabling users to connect and communicate effectively in this immersive environment. In all these scenarios, the research on generating natural nonverbal gestures aims to bridge the gap between verbal communication and nonverbal expressions, enhancing human–computer interactions' overall effectiveness and believability.

Previous studies on human action generation models include Generative Adversarial Network (GAN)-based [7,8], conditional temporal Variational Auto Encoder (VAE)-based [6], Graph Convolutional Network (GCN)-based [9], and Transformer-based models [10]. Its dataset [4,5,11,12] has focused on generating human motion for daily activities. Despite significant progress, several challenges remain in human action generation datasets; no action generation datasets are related to conversation situations. Research on gesture generation for speech involves studying the unique gesture style of an individual, replicating it, and applying it to other objects or contexts. The primary focus of this task is researching the creation of a specific speech style or style transfer rather than generating general human behavior that may occur during a conversation.

This paper introduces a novel dataset, **DGU-HAU**, a dataset for 3D **H**uman **A**ction on **U**tterances that commonly occur during daily life. Our dataset is divided into two categories: single-person presentations and conversations involving two or four people. Each category has four and ten scenarios, resulting in 142 action classes with approximately 10 classes for each scenario. The group of subjects involved in the study had an almost equal distribution of males and females, with a total of 166 participants of different age groups. Therefore, our dataset comprises 142 distinct action classes across 14 scenarios. We have collected approximately 100 motion capture data samples for each class, resulting in a total of 2408 JSON annotation files with 14,116 motion capture data samples. Each JSON annotation file contains annotation information of about six motion capture data samples. Using the Action2Motion (A2M) [6] model, we validated our dataset. The A2M model represents the latest 3D human action generation advancement as of 2023. It was validated across various datasets, employing the Fréchet Inception Distance (FID) [13] as an evaluation metric. Moreover, the Action2Motion model operates on conditional temporal VAE principles and crafts physically plausible human actions by leveraging Lie Algebra. Hence, we adopted A2M to validate our dataset by generating physically plausible human actions.

The structure of this paper is as follows. Section 2 reviews previous research on 3D-based human action analysis. Section 3 describes the structure of the proposed dataset and how we pre-processed our dataset. Section 4 explains the dataset evaluation results with the A2M model and performance analysis. Finally, Section 5 summarizes the paper and discusses future work.

2. Related Work

2.1. Generative Pre-Trained Transformer

Recent research trends in GPT (Generative Pre-trained Transformer) [14–16] have shown significant advancements in natural language processing. GPT, a state-of-the-art language model [16–19] based on the Transformer [20] architecture, has gained tremendous attention and popularity in the research community. It has demonstrated remarkable capabilities in various natural language understanding and generation tasks, including machine translation, text summarization, question answering, and conversational agents. Researchers have been actively exploring novel techniques to improve the performance, efficiency, and generalization ability of GPT models. Recent studies have addressed challenges such as model size, training efficiency, fine-tuning techniques, and ethical considerations in language generation. Additionally, efforts have been made to extend the capabilities of GPT models to handle multi-modal tasks that involve both textual and visual inputs. This paper studies building a multi-modal dataset of such a generative model. Therefore, we aim to develop a general-purpose 3D human action analysis dataset for tasks involving text and visual input to overcome the limitations of GPT studies biased toward natural language processing.

2.2. Gesture Generation

In the human action analysis study, the action generation largely consists of gesture and 3D human action generation. Firstly, gesture generation is a crucial area of research in understanding and enhancing an individual's speech pattern. The primary objective is to generate expressive gestures that align with the speech context. Several studies have explored different aspects of gesture generation, including [21]; one notable work by Ginosar et al. focused on understanding and learning the unique conversational gestures of ten celebrities. By analyzing a large dataset of their speeches, the study aimed to capture and reproduce the distinct styles of these individuals in gesture generation. Another relevant research direction is style transfer in gesture animation. The authors of [22] investigated the transfer of gesture styles between individuals. The goal was to learn consistent gesture styles from multiple individuals and apply those styles to different subjects. However, this task does not match our dataset to generate general gestures or synthesize actions during a speech to fit the speech context.

2.3. 3D Human Action Generation

In 3D human action generation, several related works have been conducted to explore various aspects of this research area. The authors of [7] generate realistic and consecutive human actions using an autoencoder and generative adversarial network. Bi-directional GAN-based [8] generates action sequences from noise. The author of [8] proposes modeling smooth and diverse transitions for action generation using a latent space of lower dimensionality. Unlike standard action prediction methods, ref. [8] can generate action sequences without any conditional action poses from pure noise. The author of [9] suggests a modified version of GCNs that selectively uses self-attention to sparsify a complete action graph in the temporal domain. The work in [10] presents a generative VAE transformer-based architecture model using SMPL [23] for 3D mesh modeling. Conditional temporal VAE-based [6] used Lie Algebra to generate physically plausible human action. This paper uses this model to validate our dataset because of the Lie Algebra.

The majority of studies on human action generation use the following datasets: Human 3.6 M [11], NTU RGB+D [4,5], HumanAct12 [6], UESTC [12]. The configuration of those datasets is described in Table 1. The study in [11] contains 3.6 million frames of motion capture data, 11 subjects performing 17 motions, four cameras, 3D skeleton motion capture data including 17 action classes (walking, running, activity, cycling, clapping, lifting, squats, etc.), and section tagging annotation. The work in [5] includes 114,480 motion capture sequences performed by 106 subjects. The dataset includes 120 motions, such as hand waving, picking up objects, sitting, standing up, walking, running, and more. The dataset

also includes RGB+D, 3D skeleton motion capture data, and section tagging annotation. The work in [6] includes motion capture data for 12 action categories, such as warm-up and lifting a dumbbell, and 34 subcategories, including warming up the elbow and lifting the dumbbell with the right hand, as well as segment tagging annotations. The work in [12] contains 40 categories of aerobic exercise with 118 subjects, as shown in Table 1. Our dataset has the largest number of action classes and subjects compared to the other datasets shown in Table 1.

Table 1. Comparison of the proposed DGU-HAU dataset and other 3D human action datasets. The Anno. (Annotation) in data is the section tagging annotation data of each data sample with the metadata, such as actor information, action code, conversation or presentation scenario information, action class, and its code. Our dataset has a text script modality: the textualization of extracted audio from each video data sample. Motion capture data (MCD) represent 3D joint information of the human body.

Dataset	# Frames	# Video	# Classes	# Subjects	# Views	# Anno.	Data Modalities				Year
							RGB	MCD	Anno.	Script	
Human3.6M [11]	3.6 M	-	17	11	4	-	✓	✓	✓	-	2013
NTU RGB+D [4]	-	56,880	60	40	80	-	✓	✓	-	-	2016
NTU RGB+D 120 [5]	-	114,480	120	106	155	-	✓	✓	-	-	2019
UESTC [12]	-	25,600	40	118	9	-	✓	✓	-	-	2019
PHSPD [24,25]	2.1 M	334	31	21	4	-	✓	✓	✓	-	2020
HumanAct12 [6]	90 K	-	12	21	4	-	✓	✓	✓	-	2020
DGU-HAU (ours)	≈144 K	1,352	142	166	15	2,408	✓	✓	✓	✓	2022

Since gesture generation is a study that learns and imitates a specific human style that can occur in conversation scenarios, the dataset used for gesture generation comprises various gestures for each particular person. On the other hand, since the 3D human action generation study is a study that learns and creates general human actions for a specific action class, the dataset for this is composed of data in which various people acted on each action class. Our dataset corresponds to 3D human action generation rather than gesture generation because it is a dataset of general human actions of specific actions that can occur in presentation and conversation scenarios. Additionally, as can be seen in Table 1, the HumanAct12 [6] dataset used in Action2Motion is a smaller dataset than our dataset. Therefore, the Action2Motion model was used to validate our dataset.

3. Dataset Structure and Building Process

This section describes the data collection environment, tools, methods, and structure. The overall data-building process for each data modality is schematized in Figure 1.

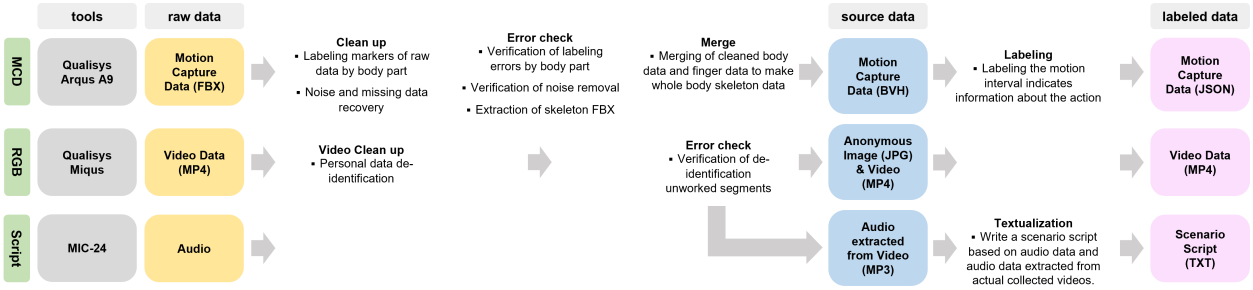


Figure 1. All data modalities were collected and built simultaneously. The finger's motion capture data were collected using MoCap Pro Super Splay, a hand motion data collection device, separate from the body part's data. They were merged with the body motion capture data coordinates according to the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

3.1. Collection Setups

We used 12 Qualisys Arqus A9 devices for motion capture data in a 6∼15 m square space and 3 Miqus video device for video data, as shown in Figure 2. Two-dimensional coordinates of each joint marker of the human body are generated from multiple motion capture cameras (Arqus A9, Qualisys, Göteborg, Sweden), and these two-dimensional coordinate data are analyzed by software (QTM) (https://www.qualisys.com/software/qualisys-track-manager/ (accessed on 10 October 2023)) to calculate coordinates in three-dimensional space. We used MoCap Pro Super Splay, a glove format with 16 sensors, to acquire the hand motion capture data. Motion capture data for the human body and hands were collected separately and then integrated using QTM to create a complete motion capture of the whole human body. There are three distinct viewpoints to collect RGB video data, and footage is shot at 60 fps or higher. The RGB video is full HD with 1920 × 1080 resolution.

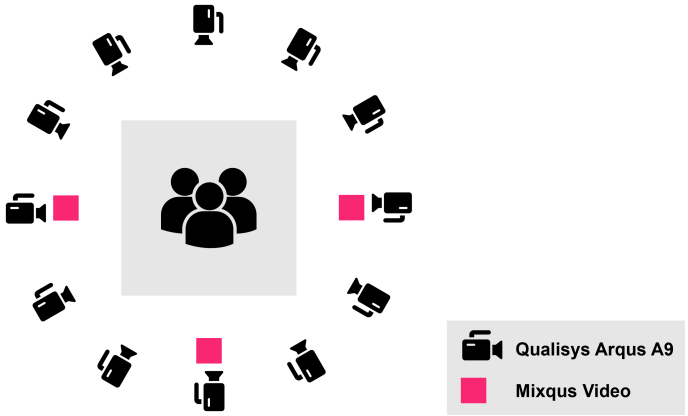
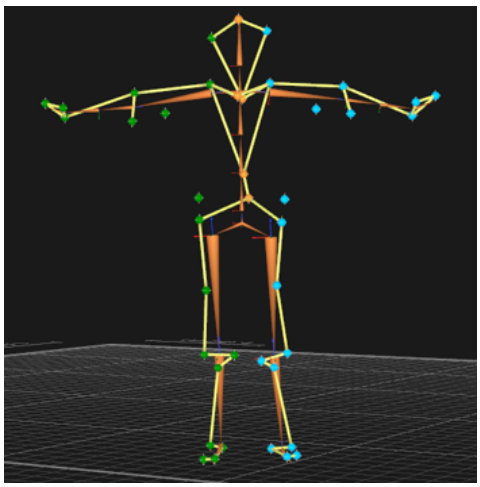


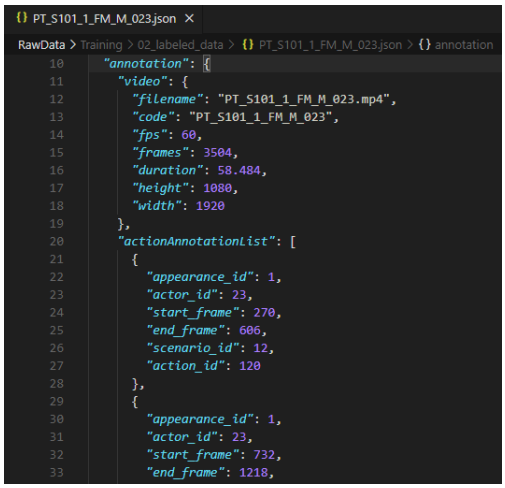
Figure 2. Setup of the data collection environment.

3.2. Data Modalities

The proposed dataset, DGU-HAU, provides 14,116 motion capture data samples with 2408 annotation data samples, and there are four data modalities: motion capture data, RGB video, scenario script, and annotation data. The overall building steps for each modality of our dataset are shown in Figure 1. The samples of each data modality are shown in Figure 3.



(**a**) Sample of the MCD



(**b**) Sample of the annotation data

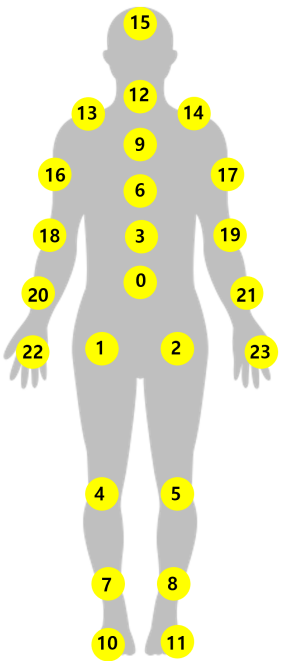
Figure 3. *Cont.*

(**c**) Sample of the RGB video data

Figure 3. Sample data of each data modality. (**a**) A sample of the 3D motion capture data (MCD), which is formatted in bvh. Green points mean right body joints, blue points mean left body joints, yellow lines mean the line connecting the connected joints, and orange means the bone.; (**b**) a sample of the annotation data, which is formatted in JSON; (**c**) a sample of the RGB video data.

3.2.1. Motion Capture Data (MCD)

Our dataset provides 14,116 motion capture data samples of the human body in BVH (BioVision Hierarchy) format. This common file format delivers motion capture data that represent 3D coordinates of the joints of humans, as shown in Figure 4. The BVH format consists of a hierarchy section and a motion section. In the hierarchy section, the information on the human skeleton joints is represented in a tree structure, and each joint has an offset and a channel list. The channel list is a transformation list for motion at that point.



No.	Label	No.	Label
0	Hips	1	LeftUpLeg
2	RightUpLeg	3	Spine
4	LeftLeg	5	RightLeg
6	Spine1	7	LeftFoot
8	RightFoot	9	Spine2
10	LeftToeBase	11	RightToeBase
12	Neck	13	LeftShoulder
14	RightShoulder	15	Head
16	LeftArm	17	RightArm
18	LeftForeArm	19	RightForeArm
20	LeftForeArmRoll	21	RightForeArmRoll
22	LeftHand	23	RightHand

Figure 4. Configuration of the body joints and label in our dataset.

First, we collected raw samples of motion capture data in FBX (Filmbox) format using 12 Qualisys Arqus A9 devices, as shown in Figures 1 and 2. Then, we cleaned up the

collected data by labeling body part markers and recovering missing data and noise. After cleaning up, we verified the collected data and extracted the skeleton information from the raw data. We collected body and finger joint data separately to create a dataset that can detect precise finger movements. Thus, we obtained 3D coordinates for 75 body joints, consisting of 27 for the human body and 24 for the left and right human fingers. We labeled the motion capture data, extracted each 3D coordinate, and converted it to JSON format for ease of use. We selected and employed 24 representative joints out of 75 for data verification. Figure 4 displays information about the position, number, and label of the selected joints in this paper.

3.2.2. RGB Video

Our dataset provides 1352 RGB videos of the various versions of each action class related to the utterance in MP4 format, of which the resolution is 1920 × 1080. Each RGB video has more than one action class, and the section tagging information of each action class is in the annotation data. As shown in Figure 2, we filmed the 14 scenarios, including about ten action classes in three different views (front, left, and right side) using Qualisys Miqus devices. After collecting the video data, we anonymized the video to protect personal information. We then verified that the data anonymization process was successful, as shown in Figure 1. Simultaneously, we extracted the audio data in MP3 format from the verified video data for the scenario script; the other modality is described in the next section.

3.2.3. Scenario Scripts

Our dataset provides 1352 scenario scripts, based on the audio extracted from RGB videos, in the text file, as shown in Table 1. We wrote a scenario script based on the audio data extracted from each collected RGB video. After that, we checked the audio-to-text scenario script for errors such as typos and profanity and whether the lines matched well with the uttered actors and times. As shown in Table 2, it consists of a total of fourteen types of scenarios; four scenario types are scenarios for one-person presentation circumstances, and ten scenario types are scenarios for two- or four-person conversation circumstances. We collected data samples for 14 types of scenarios with various combinations of actors.

Table 2. Data modality and description of each modality.

Data Modality	File Format	Description
Motion Capture Data (MCD)	BVH	- 3D coordinate of joint - Number of joints: 24
RGB Video	MP4	- Resolution: 1920 × 1080 - Number of views: 3
Scenario Scripts	TXT	- Text script of 14 action scenarios: presentation × 4 conversation × 10
Annotation Data	JSON	- Metadata of the other data modalities - Configuration: pre-processed MCD, scenario code, scenario name, action code, action class, actor ID, video section tagging, etc.

3.2.4. Annotation Data

Our dataset provides 2408 annotation data samples, and each annotation data file contains the annotation information of about 6 motion capture data samples. Therefore, we have 14,116 motion capture data samples. The annotation data are the metadata of the others. They include the information of the dataset, annotation of the video, and each action, actor, scenario, action information, corresponding video section information, and motion capture data, as shown in Table 3. In the annotation part of the motion capture data,

there are frame ID and 3D coordinates of the body joints in Figure 4. The name of the action class is tagged based on predefined start and end points and conversation content. There are a total of 75 body joints, of which 48 body joints are related to finger joints, as shown in Table 3. It represents the 3D coordinates of 24 joints for the right and left hands. The remaining 27 body joints represent important joints in the human body. When verifying the data, we referred to [6] and selected 24 body joints that appropriately expressed the human body out of 75 body joints to construct a skeleton. A more detailed body joint label annotation is described in [26].

Table 3. Configuration of annotation data in JSON format. More detailed body joint label annotation is described in [26].

Category	Type	Description
1. info	object	information on the data
1.1 name	string	name of the dataset
1.2 creator	string	name of the constructor
1.3 date_created	string	date of the deployment
2. annotation	object	information of the annotation
2.1 video	object	information of the video
2.1.1 filename	string	file name of the video
⋮	⋮	⋮
2.2 actionAnnotationList	array	action annotation
2.2.1 actor_id	number	ID of the actor
⋮	⋮	⋮
3. categories	object	information of the categories
3.1 actionCategories	array	list of the action classes
3.1.1 id	number	ID of action scenario
⋮	⋮	⋮
3.2 actionScenarioList	array	list of the action scenarios
3.2.1 id	number	ID of action scenario
⋮	⋮	⋮
4. actors	array	information of the actors
4.1 id	number	ID of the actor
⋮	⋮	⋮
5. mocap_data	array	motion capture data
5.1 frame	number	frame id
5.2 bodyJoints	object	coordinate info. of the keypoints
5.2.1 Skeletons	array	3D coordinates of the Skeleton
5.2.2 Reference	array	3D coordinates of the Reference
5.2.3 Hips	array	3D coordinates of the (0) Hips
⋮	⋮	⋮
5.2.77 HeadEnd	array	3D coordinates of the HeadEnd

3.3. Subjects

The dataset includes 166 subjects of different ages and genders, each with a unique actor ID number. The subjects were recruited with an equal number of males and females and varying ages to mitigate bias. Table 4 shows the number of data samples by age and gender groups. The subjects' age groups were divided into three groups: the young group in their teens and 20 s, the middle group in their 30 s and 40 s, and the old group in their 50 s and 60 s. Y denotes the young group, M represents the middle group, and O means the old group. There are 1128 data samples containing female subjects and 1280 data samples containing male subjects, and the data are structured in an almost 1:1 ratio in the gender of subjects. The number of data samples by group is 226, 430, and 472 in the order of old, middle, and young groups for female subjects, respectively, and 344, 509, and 427 for male subjects. In the conversation scenario, each scenario comprised more

than 84 different subjects. In the presentation scenario, each scenario included more than 170 different subjects.

Table 4. Gender and age ratio of the subjects. FM denotes female, and MA denotes male.

Gender	Age Group	Detailed	# of Data Samples	Sub Total
FM	O	Old: 50–60 s	226	1128
	M	Middle: 30–40 s	430	
	Y	Young: 10–20 s	472	
MA	O	Old: 50–60 s	344	1280
	M	Middle: 30–40 s	509	
	Y	Young: 10–20 s	427	

3.4. Action Classes

Our dataset has 142 action categories among four presentation and ten conversation scenarios. There are four types of presentations: sitting and standing presentations, as well as explanations. The ten conversations cover various scenarios that commonly occur daily, including daily life (specifically watching TV), consoling someone, celebrating a birthday, etc. Tables 5 and 6 describes the scenario code, scenario name, action class code, and action class name of the action classes and scenarios. The table only includes brief information about action classes and scenarios. The full table is in [27,28].

Table 5. Configuration of the 37 action classes and their action codes in the four presentation scenarios. The work in [27] shows the full table about the configuration of 37 action classes.

Scenario Code	Scenario	Action Code	Action Class
S101	stand up and do a presentation	A106	holding the microphone with both hands
		⋮	⋮
		A115	taking a step forward and bowing a head in greeting
S102	sit down and do a presentation	A116	counting by hand
		⋮	⋮
		A123	expressing a part of the drawn circle
S103	stand up and explain	A124	pointing to the work displayed behind with the left hand
		⋮	⋮
		A133	moving towards the work behind the audience
S104	sit down and explain	A134	waving of hand of greeting
		⋮	⋮
		A142	end greeting while sitting down

Table 6. Configuration of the 105 action classes and their action codes in the ten conversation scenarios. The work in [28] shows the full table about the configuration of the 105 action classes.

Scenario Code	Scenario	Action Code	Action Class
S201	daily life (watching TV)	A001	looking around and finding the remote control
		⋮	⋮
S202	consolation	A013	covering face with hands and sighing
		⋮	⋮
S203	celebrating birthday	A023	showing up with a cake
		⋮	⋮

Table 6. *Cont.*

Scenario Code	Scenario	Action Code	Action Class
S204	doctor's office	A033	opening the door and entering in
		⋮	⋮
S205	asking the station staff for directions	A042	taking a phone out of pocket
		⋮	⋮
S206	taking a picture of someone else	A052	taking selfie
		⋮	⋮
S207	catch up	A065	showing a phone to the opponent
		⋮	⋮
S208	shopping for clothes at a department store	A073	choosing clothes from a hanger
		⋮	⋮
S209	intercompany business meeting	A083	bowing down in greeting
		⋮	⋮
S210	grocery shopping	A096	looking at the notes while walking
		⋮	⋮
		A105	taking something and putting it in a shopping cart

4. Dataset Pre-Processing and Evaluation with Action2Motion

4.1. Data Pre-Processing

We validated our dataset using Action2Motion [6], as mentioned earlier. Therefore, to generate 3D plausible human actions, we trained the A2M model with our data and used the trained A2M model to create 3D human actions as GIF files. To do this, we pre-processed the data according to the input format of the A2M model. The overall pre-processing and generating action steps for our dataset are shown in Figure 5.

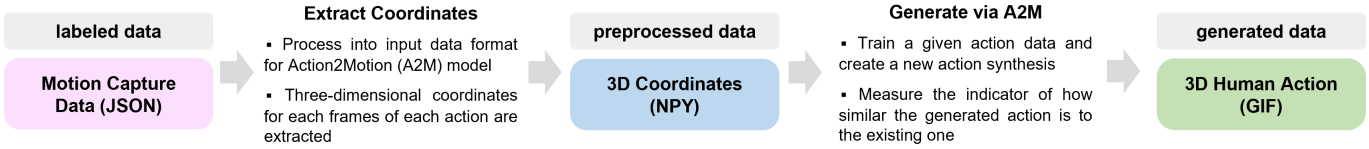


Figure 5. Overall steps of pre-processing our dataset to train the A2M [6] model. We extracted coordinates from labeled motion capture data to generate the NumPy file of 3D coordinates of human action. We trained the A2M model with pre-processed data and generated a new 3D human action in GIF format.

We obtained the first pre-processed 3D human action data in JSON file format through annotation of BVH format motion capture data, and the second pre-processed 3D coordinate data in the NumPy format through the extraction of 3D coordinates, as shown in Figure 5. The extraction of 3D coordinates involves two main steps. First, we extract the information from the 24 joint we selected among the 75 joints' information during the frame section corresponding to the action class from each data sample. Second, the 3D coordinate values of the extracted 24 joints were converted to the NumPy format to match the input format of the model. The final pre-processed dataset was trained using the A2M model to create a new action synthesis. We measured its similarity to the existing ground truth using the FID metric.

4.2. Evaluation Results for A2M Model

This paper uses the A2M model to evaluate our dataset. Following [6], four metrics, FID, accuracy, diversity, and multi-modality, are considered to validate our dataset as a metric. FID [13] is a metric used to measure the performance of generative models. It assesses the similarity between generated and ground truth data. The FID comprises two main components: Inception Score and Fréchet Distance. The *FID* is obtained by applying the following operation:

$$FID = \|\mu_{real} - \mu_{fake}\|^2 + Tr(\Sigma_{real} + \Sigma_{fake} - 2(\Sigma_{real}\Sigma_{fake})^{1/2} + FID^2_{score} \quad (1)$$

where μ_{real} and μ_{fake} represent the means of the feature vectors for real and generated data, respectively. Σ_{fake} and Σ_{real} represent the covariances of the feature vectors for generated and ground truth data, respectively. FID_{score} represents the difference in Inception Scores [13].

FID is related to human intuition in that it is based on feature vectors that capture an abstract representation between generated and real data. Therefore, FID matches well with general intuition about how similar a generated image feels to real data. Additionally, FID calculates the Frèchet Distance between two multivariate normal distributions, which can quantify and compare the similarities and differences between the two distributions through statistical methods. Therefore, we used FID as the main data evaluation metric. Diversity gauges the deviation of the generated motions across all categories of actions. Unlike diversity, multi-modality assesses the extent to which the generated motions exhibit diversity within each action category.

The hardware specifications of dataset evaluation are listed in Table 7, and the hyperparameter settings of the A2M model are shown in Table 8 below. We proceeded with training by maintaining the hyperparameter settings of the A2M model [6]. We used NVIDIA GeForce RTX 3090 GPU for training and evaluating our dataset, and the employed learning parameters are shown in Table 8 below.

Table 7. The hardware specifications.

Element	Specification
CPU	Intel Xeon Gold 6226R 2.90 GHz
Memory	256 GB
GPUs	Nvidia GeForce RTX 3090
OS	Ubuntu 18.04
Framework	Pytorch 1.8.2 + CUDA 11.1

Table 8. Configuration of the hyperparameters.

Hyperparameter	Value	Description
lambda_kld	0.001	Weight of KL Divergence
lambda_align	0.5	Weight of align loss
time_counter	true	Enable time count in generation
tf_ratio	0.6	Teacher force learning ratio
use_lie	true	Use Lie Representation
batch_size	16	Batch size of training process
iterations	50,000	Training iterations

Since the action class classifier of the A2M model can only classify up to 13 action classes, we separately trained 14 scenarios. According to Tables 5 and 6, there are no more than 13 action classes in each scenario in our dataset. Therefore, we trained the A2M model for each scenario, and the evaluation results are shown in Table 9.

Table 9. Evaluation results of our dataset for each scenario. The lower the FID and the higher the Accuracy, the better the performance.

Scenario Code	Evaluation Metrics (Real Motion)				Evaluation Metrics (Generated)			
	FID ↓	Accuracy ↑	Diversity →	Multi Modality →	FID ↓	Accuracy ↑	Diversity →	Multi Modality →
S101	0.050 ± 0.0053	0.998 ± 0.0002	6.673 ± 0.0843	1.576 ± 0.0116	0.161 ± 0.0005	0.934 ± 0.0008	6.636 ± 0.0514	1.830 ± 0.0263
S102	0.042 ± 0.0039	0.993 ± 0.0004	6.643 ± 0.0686	1.631 ± 0.0198	0.524 ± 0.0311	0.886 ± 0.0016	6.512 ± 0.0930	2.052 ± 0.0286
S103	0.045 ± 0.0030	0.983 ± 0.0005	6.599 ± 0.0969	1.674 ± 0.0185	0.342 ± 0.0113	0.864 ± 0.0016	6.569 ± 0.0832	1.871 ± 0.0177
S104	0.039 ± 0.0029	0.956 ± 0.0008	6.526 ± 0.0821	1.965 ± 0.0206	0.356 ± 0.0109	0.829 ± 0.0018	6.689 ± 0.0718	2.185 ± 0.0308
S201	0.062 ± 0.0047	0.979 ± 0.0005	6.534 ± 0.0658	2.735 ± 0.0163	0.387 ± 0.0065	0.618 ± 0.0023	6.300 ± 0.0554	3.971 ± 0.0223
S202	0.057 ± 0.0052	0.975 ± 0.0006	6.536 ± 0.0671	2.836 ± 0.0290	1.532 ± 0.0161	0.612 ± 0.0020	6.265 ± 0.0506	3.991 ± 0.0296
S203	0.051 ± 0.0035	0.974 ± 0.0007	6.501 ± 0.0648	2.357 ± 0.0205	1.082 ± 0.0261	0.552 ± 0.0018	6.154 ± 0.0557	4.198 ± 0.0282
S204	0.049 ± 0.0033	0.986 ± 0.0006	6.493 ± 0.0649	2.478 ± 0.0238	0.487 ± 0.0178	0.685 ± 0.0020	6.276 ± 0.0625	4.301 ± 0.0304
S205	0.046 ± 0.0027	0.989 ± 0.0004	6.496 ± 0.0539	3.286 ± 0.0220	1.023 ± 0.0157	0.441 ± 0.0020	6.134 ± 0.0863	4.661 ± 0.0218
S206	0.054 ± 0.0025	0.980 ± 0.0004	6.501 ± 0.0578	3.119 ± 0.0139	0.815 ± 0.0206	0.352 ± 0.0025	6.220 ± 0.0721	5.016 ± 0.0183
S207	0.051 ± 0.0036	0.981 ± 0.0007	6.463 ± 0.0587	2.985 ± 0.0483	0.841 ± 0.0150	0.594 ± 0.0013	6.188 ± 0.0703	4.628 ± 0.0339
S208	0.051 ± 0.0020	0.983 ± 0.0006	5.973 ± 0.3159	2.969 ± 0.0252	0.840 ± 0.0197	0.513 ± 0.0023	5.899 ± 0.0507	4.529 ± 0.0248
S209	0.044 ± 0.0030	0.998 ± 0.0001	6.826 ± 0.0538	2.064 ± 0.1400	0.622 ± 0.1350	0.637 ± 0.0848	6.585 ± 0.0657	2.629 ± 0.2310
S210	0.044 ± 0.0028	0.998 ± 0.0001	6.823 ± 0.0408	1.935 ± 0.1290	0.520 ± 0.1310	0.713 ± 0.0803	6.589 ± 0.0689	2.408 ± 0.2168

According to Table 9, the FID for real motion, the ground truth data for each scenario, is 0.039 to 0.062, showing that the distribution of the original data was learned very well. In addition, the FID for the generated motion is 0.342 to 1.532 for each scenario, confirming that the data were well-generated by learning the distribution of the original data well. The evaluation metric in the result table was constructed by referring to [6] for comparison. Accuracy refers to whether A2M's classifier generates the right action for real motion. For real motion, the accuracy is very high, at 97% to 99%, but for generated motion, the accuracy ranges from 35% to 93%, which shows a very large deviation. The scenario's low accuracy is due to the presence of multiple similar action classes. Diversity is an evaluation metric for whether the model generates diverse data. Table 9 shows that our data have a diversity of about 6 for all scenarios.

To compare our dataset with other datasets, we trained the A2M on our dataset under the same experimental conditions as the three datasets [4,6,29] on which the A2M was trained. As mentioned earlier, A2M can classify up to 13 action classes, so we selected 13 by applying random sampling, just as A2M applied to the other two datasets [4,29]. We sampled seven subsets, which consist of 13 action classes of our dataset, and trained using the A2M model. The results of the comparison experiment are shown in Table 10.

Table 10. Evaluation results of our dataset and comparison with other datasets based on [6]. The lower the FID and the higher the Accuracy, the better the performance.

Dataset	Evaluation Metrics (Real Motion)				Evaluation Metrics (Generated)			
	FID ↓	Accuracy ↑	Diversity →	Multi Modality →	FID ↓	Accuracy ↑	Diversity →	Multi Modality →
HumanAct12	0.092 ± 0.007	0.997 ± 0.001	6.853 ± 0.053	2.449 ± 0.038	2.458 ± 0.079	0.923 ± 0.002	7.032 ± 0.002	2.870 ± 0.037
NTU-RGB+D	0.031 ± 0.004	0.999 ± 0.001	7.108 ± 0.048	2.194 ± 0.025	0.330 ± 0.008	0.949 ± 0.001	7.065 ± 0.043	2.052 ± 0.030
CMU Mocap	0.065 ± 0.006	0.930 ± 0.002	6.130 ± 0.079	2.720 ± 0.066	2.885 ± 0.116	0.680 ± 0.003	6.500 ± 0.061	4.120 ± 0.056
DGU-HAU (Ours)	0.041 ± 0.002	0.872 ± 0.001	6.528 ± 0.047	2.129 ± 0.014	0.992 ± 0.020	0.759 ± 0.001	6.157 ± 0.069	2.291 ± 0.021

According to Table 10, the FID value for each scenario of ground truth is 0.041. This value is the second smallest among the four datasets, only surpassed by NTU RGB+D [4]

by a mere 0.01 difference. The motion generated has an FID value of 0.992, the second-best after NTU RGB+D, with a difference of 0.66.

Regarding accuracy, real motion recorded the lowest value among the four datasets at 87.2%. The value was lowered because there were other similar action classes when only 13 were randomly selected from 142. According to Table 9, the accuracy of real motion for each scenario is 97% to 99%, similar to other datasets in Table 10. The accuracy of generated motion is 75.9%, ranking third out of four datasets. The value seems to have been measured with slightly lower accuracy for the same reason as the real motion.

In the case of Diversity, both real and generated motion were about 6, which showed that the values were similar to other datasets. It has been determined that data can be produced variously.

5. Discussion

A 3D human action dataset on utterance, DGU-HAU, is introduced in this paper. Our dataset provides 14,116 data samples of motion capture data, 1352 RGB videos, their textualized scenario script based on the audio data extracted from the video, and the 2408 annotation data samples in JSON format. The human actions are based on 14 scenarios occurring in daily life, and these scenarios include about ten action classes per each one. Therefore, there are 142 action classes in our dataset. Also, 166 subjects recorded action classes in various combinations according to age group, gender, and body shape. 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 generative model, Action2Motion [6], but it is possible to apply various models, such as human action recognition and human–object interaction. Action2Motion is a 3D human action generation model that leverages Lie Algebra for physically plausible human action. In the experimental results, the FID values for real motion showed good results in the following order: NTU-RGB+D, our dataset, CMU Mocap, and HumanAct12. Additionally, the FID values for generated motion showed good results in the following order: NTU-RGB+D, our dataset, HumanAct12, and CMU Mocap. Our dataset showed a difference of 0.01 and 0.66 in real motion and generated motion, respectively, from the NTU-RGB+D dataset that showed the best results, while differences of 0.05 in real motion and 1.89 in generated motion were found from the dataset that showed the worst results. While NTU-RGB+D consists of various types of actions that can occur in everyday life, our data consist of a series of actions that can occur in the flow of conversation. Therefore, because there is continuity of motion, even different motions within one scenario may have similar motions. For these reasons, the NTU-RGB+D has clear distinctions between each operation. Still, our data sometimes have actions that overlap or occur twice in the scenario, so the distinction between each operation may be less clear than in NTU-RGB+D. Additionally, in our data, we were able to confirm that the performance was slightly smaller than NTU-RGB+D because the scale of the action was not large. Therefore, we were able to confirm that our dataset was well-created and verified well.

In future work, we plan to apply various models that study 3D human actions in different ways, such as the human action recognition model, to our dataset.

Author Contributions: Funding acquisition, D.K.; Investigation, J.P. and K.P.; Project administration, D.K.; Software, K.P.; Supervision, D.K.; Writing—original draft, J.P.; Writing—review and editing, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. S-2021-A0496-00167). This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-2020-0-01789), and the Artificial Intelligence Convergence Innovation Human Resources Development (IITP-2023-RS-2023-00254592) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed Consent was obtained from all the human subjects who participated in the data collection.

Data Availability Statement: Dataset access: [26]. Dataset access for non-Koreans: https://github.com/CSID-DGU/NIA-MoCap-2 (accessed on 10 October 2023).

Acknowledgments: The dataset was built by DTAAS consortium.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.06950.
2. Zhang, Z. Microsoft kinect sensor and its effect. *IEEE Multimed.* **2012**, *19*, 4–10. [CrossRef]
3. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334.
4. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
5. Liu, J.; Shahroudy, A.; Perez, M.; Wang, G.; Duan, L.Y.; Kot, A.C. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2684–2701. [CrossRef] [PubMed]
6. Guo, C.; Zuo, X.; Wang, S.; Zou, S.; Sun, Q.; Deng, A.; Gong, M.; Cheng, L. Action2motion: Conditioned generation of 3d human motions. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2021–2029.
7. Kiasari, M.A.; Moirangthem, D.S.; Lee, M. Human action generation with generative adversarial networks. *arXiv* **2018**, arXiv:1805.10416.
8. Wang, Z.; Yu, P.; Zhao, Y.; Zhang, R.; Zhou, Y.; Yuan, J.; Chen, C. Learning diverse stochastic human-action generators by learning smooth latent transitions. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12281–12288.
9. Yu, P.; Zhao, Y.; Li, C.; Yuan, J.; Chen, C. Structure-aware human-action generation. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28. 2020, Proceedings, Part XXX*; Springer: Cham, Switzerland, 2020; pp. 18–34.
10. Petrovich, M.; Black, M.J.; Varol, G. Action-conditioned 3D human motion synthesis with transformer VAE. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10985–10995.
11. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 1325–1339. [CrossRef] [PubMed]
12. Ji, Y.; Xu, F.; Yang, Y.; Shen, F.; Shen, H.T.; Zheng, W.S. A large-scale varying-view rgb-d action dataset for arbitrary-view human action recognition. *arXiv* **2019**, arXiv:1904.10681.
13. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–60.
14. Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I. Improving Language Understanding by Generative Pre-Training. OpenAI. 2018. Available online: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf (accessed on 10 October 2023).
15. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
16. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
17. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
18. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.
19. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv* **2019**, arXiv:1910.13461.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–15.
21. Ginosar, S.; Bar, A.; Kohavi, G.; Chan, C.; Owens, A.; Malik, J. Learning individual styles of conversational gesture. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3497–3506.
22. Ahuja, C.; Lee, D.W.; Nakano, Y.I.; Morency, L.P. Style transfer for co-speech gesture animation: A multi-speaker conditional-mixture approach. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28. 2020, Proceedings, Part XVIII*; Springer: Cham, Switzerland, 2020; pp. 248–265.

23. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph. (TOG)* **2015**, *2*, 851–866. [CrossRef]
24. Zou, S.; Zuo, X.; Qian, Y.; Wang, S.; Guo, C.; Xu, C.; Gong, M.; Cheng, L. Polarization human shape and pose dataset. *arXiv* **2020**, arXiv:2004.14899.
25. Zou, S.; Zuo, X.; Qian, Y.; Wang, S.; Xu, C.; Gong, M.; Cheng, L. 3D human shape reconstruction from a polarization image. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28. 2020, Proceedings, Part XIV*; Springer: Cham, Switzerland, 2020; pp. 351–368.
26. AIHub. DGU-HAU: A Dataset for 3D Human Action Analysis on Utterances. Available online: https://www.aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=71419 (accessed on 10 October 2023).
27. Park, J. Full Action Classes Table for Presentation Scenario (Table 5). Available online: https://github.com/CSID-DGU/NIA-MoCap-2/blob/main/docs/action_classes_presentation_scenario(Table5).pdf (accessed on 10 October 2023).
28. Park, J. Full Action Classes Table for Conversational Scenario (Table 6). Available online: https://github.com/CSID-DGU/NIA-MoCap-2/blob/main/docs/action_classes_conversational_scenario(Table6).pdf (accessed on 10 October 2023).
29. CMU. *CMU Graphics Lab Motion Capture Database*; CMU: Pittsburgh, PA, USA, 2003.