

AIR-Act2Act: Human-human interaction dataset for teaching non-verbal social behaviors to robots

Journal Title
XX(X):1-6
©The Author(s) 2019
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Woo-Ri Ko¹, Minsu Jang¹, Jaeyeon Lee¹ and Jaehong Kim¹

Abstract

To better interact with users, a social robot should understand the users' behavior, infer the intention, and respond appropriately. Machine learning is one way of implementing robot intelligence. It provides the ability to automatically learn and improve from experience instead of explicitly telling the robot what to do. Social skills can also be learned through watching human-human interaction videos. However, human-human interaction datasets are relatively scarce to learn interactions that occur in various situations. Moreover, we aim to use service robots in the elderly-care domain; however, there has been no interaction dataset collected for this domain. For this reason, we introduce a human-human interaction dataset for teaching non-verbal social behaviors to robots. It is the only interaction dataset that elderly people have participated in as performers. We recruited 100 elderly people and two college students to perform 10 interactions in an indoor environment. The entire dataset has 5,000 interaction samples, each of which contains depth maps, body indexes and 3D skeletal data that are captured with three Microsoft Kinect v2 cameras. In addition, we provide the joint angles of a humanoid NAO robot which are converted from the human behavior that robots need to learn. The dataset and useful python scripts are available for download at <https://github.com/ai4r/AIR-Act2Act>. It can be used to not only teach social skills to robots but also benchmark action recognition algorithms.

Keywords

Social robot, machine learning, human-human interaction, the elderly

Introduction

To better interact with users, a social robot should understand their behavior, infer the intention and formulate appropriate responses. For instance, if a user is crying in the bedroom, a robot should approach slowly and hug her shoulder gently. Toward this need, many researchers, e.g. [Huang and Mutlu \(2012\)](#), [Qureshi et al. \(2016\)](#) and [Hemminahaus and Kopp \(2017\)](#), have focused on implementing social intelligence for robots. However, these studies have a limitation that the robot repeats only predefined behaviors since they are more about behavior selection rather than behavior generation.

Machine learning is one way of overcoming the above limitation. It provides the robots with the ability to automatically learn and improve from experience instead of explicitly telling them what to do. In recent years, this methodology has shown good performances benefiting from the increased availability of demonstration data, as well as computational advancements brought on by deep learning. [Ko et al. \(2018\)](#) showed that it was feasible to a deep neural network for robots to learn social skills. Two social behaviors, i.e., *handshake* and *wait*, were learned from 583 human-human interaction videos of *NTU RGB+D* dataset introduced by [Shahroudy et al. \(2016\)](#). The dataset used was large enough to generate two behaviors, but a larger dataset is essential to generate more behaviors. However, the existing datasets are not large enough to learn the interactions that occur in various situations. Moreover, there has been no dataset collected for the elderly, so it is difficult to use in the elderly-care domain, which is what we are aiming to do.

For this reason, we introduce a human-human non-verbal interaction dataset, *AIR-Act2Act*. It was collected as a part of project *AIR (AI for Robots)* which aims to provide socially assistive services to the elderly. We recruited 100 elderly people and two college students to perform 10 interactions in an indoor environment. With three Microsoft Kinect v2 cameras, depth maps, body indexes, and 3D skeletal data were captured concurrently. In addition, the behaviors of the people to be learned were converted into the robot's joint angles. In summary, the dataset has the following strengths:

- (i) It is the only interaction dataset of the elderly;
- (ii) It provides robotic data to be learned;
- (iii) It is one of the largest interaction datasets that provides 3D skeletal data;
- (iv) It can be used to not only teach social skills to robots but also benchmark action recognition algorithms.

Related datasets

Several human-human interaction datasets are accessible for research purposes, and they contain RGB videos, depth maps, and 3D skeletal data. Example frames of the datasets appear in [Figure 1](#).

¹Electronics and Telecommunications Research Institute (ETRI), KR

Corresponding author:

Woo-Ri Ko, ETRI, 218 Gajeong-ro, Yuseong-gu, Daejeon, 34129, KR.
Email: wrko@etri.re.kr

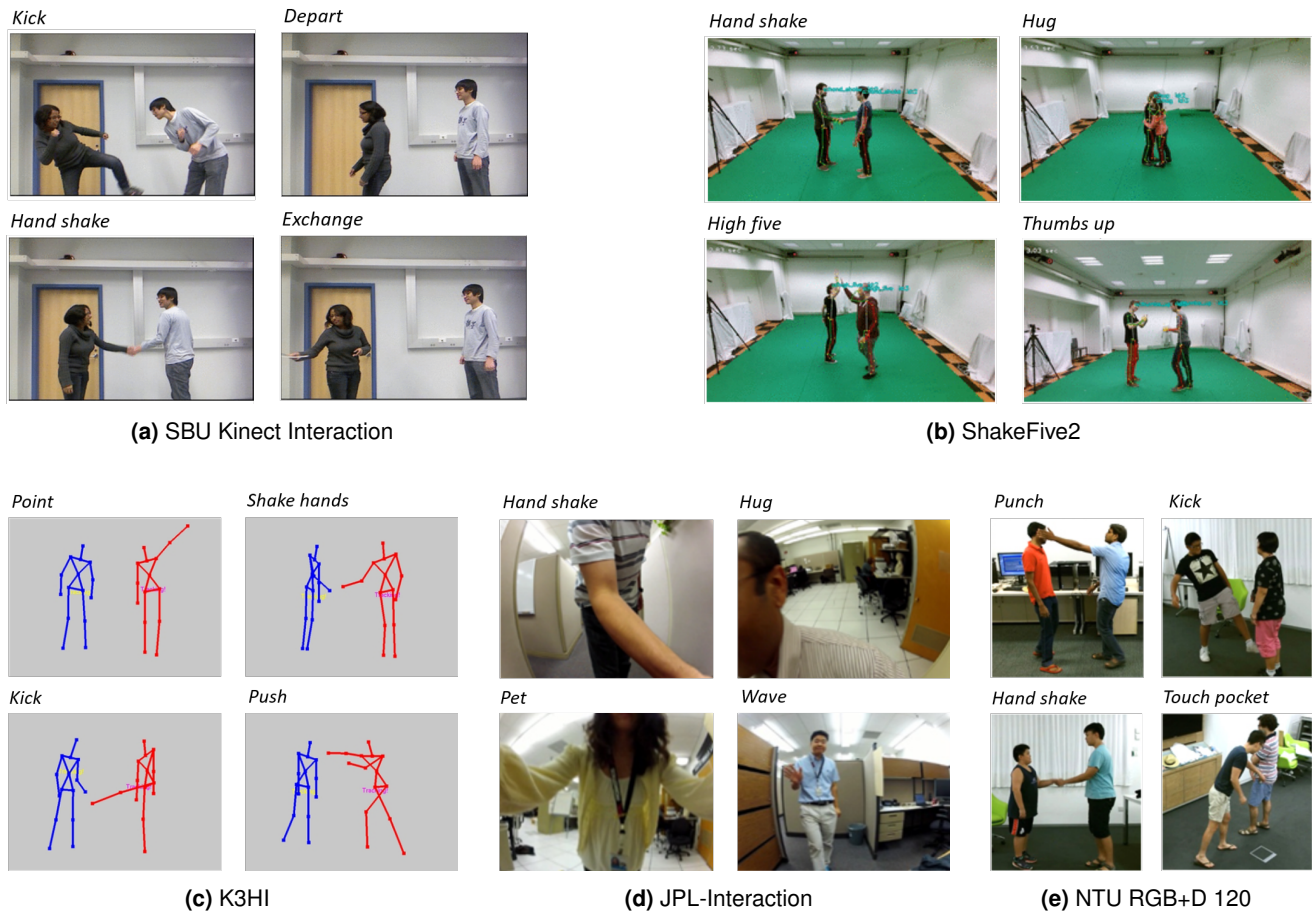


Figure 1. Related datasets provided for research purposes.

SBU Kinect Interaction

[Yun et al. \(2012\)](#) introduced the *SBU Kinect Interaction* dataset captured by Microsoft Kinect. It contains 300 videos of eight interactions: *push*, *kick*, *punch*, *pass object*, *hug*, *hand shake*, *approach* and *depart*. The RGB videos and depth maps were recorded in 640×480 resolution. The skeletal data contain 3D coordinates of 15 joints per person.

ShakeFive2

[Van Gemeren et al. \(2016\)](#) introduced a collection of human interaction clips captured by Microsoft Kinect v2 camera. There are eight interaction classes in the dataset: *fist bump*, *handshake*, *high-five*, *hug*, *pass object*, *thumbs up*, *rock-paper-scissors* and *explain*. Further, 153 videos were captured in lab conditions and encoded at a resolution of 1280×720 . The skeletal data contains 3D coordinates of 25 joints per person.

K3HI

[Hu et al. \(2013\)](#) introduced the *K3HI* dataset captured by Kinect. It provides 312 skeletal files of eight interactions: *approach*, *depart*, *exchange*, *kick*, *point*, *punch*, *push*, *shake*. The skeletal data contains 3D coordinates of 15 joints per person.

JPL-Interaction

The *JPL-Interaction* dataset is a first-person human-robot interaction dataset introduced by [Ryoo et al. \(2015\)](#). A GoPro2 camera is attached to the head of a humanoid model and human participants are asked to interact with the humanoid. The dataset has eight videos containing 180 executions of seven activities: *shake hand*, *hug*, *pet*, *wave hand*, *point*, *punch* and *throw object*. The RGB videos and depth maps were recorded in 640×480 and 320×240 resolutions, respectively. The skeletal data contains HOJ3D features obtained by [Xia et al. \(2012\)](#).

NTU RGB+D 120

The *NTU RGB+D 120* dataset is an action recognition dataset introduced by [Liu et al. \(2019\)](#). It contains 8,276 action samples of 26 two-person actions, e.g. *punch*, *kick*, *push*, *pat*, etc. The RGB videos, depth maps and 3D skeletal data are captured by three Microsoft Kinect v2 concurrently. The RGB videos and depth maps were recorded in 1920×1080 and 512×424 resolutions, respectively. The skeletal data contains 3D coordinates of 25 major body joints per person.

Other datasets without skeletal data

From skeletal data, temporal patterns of interactions can be learned without having to consider viewpoint and person appearance. However, some datasets contain RGB videos but no 3D skeletal data or depth maps depending on the cameras

used. If not available, the skeletal coordinates of the joints can be extracted from RGB frames by a 3D pose estimation algorithm such as that proposed by Tome et al. (2017). The datasets that can be used in this way are described as follows:

- *UT-Interaction*: Ryoo and Aggarwal (2010) introduced videos of continuous executions of six classes of human-human interactions: *shake-hands, point, hug, push, kick, and punch*.
- *TV Human Interaction*: Patron-Perez et al. (2010) provided short video segments of four classes taken from popular TV series: *handshake, hug, kiss, and highfive*.
- *Hollywood2*: Marszalek et al. (2009) introduced movie clips of four classes: *fight, handshake, hug, and kiss*.
- *DeepMind Kinetics*: Kay et al. (2017) provided 10-s clips from YouTube videos of 11 interaction classes including *handshake, hug, and massage feet*.

Our dataset: AIR-Act2Act

In this section, we introduce the details of the *AIR-Act2Act* dataset, which is the only human-human interaction dataset that elderly people have participated in as performers.

Subjects

We recruited 100 elderly people and two college students for our data collection. The elderly people were recruited at the senior welfare centers, based on the following criteria: (1) age over 60; (2) healthy enough to stand and walk for a few minutes. The mean age of the elderly people was 77 years (ranging from 64 to 88) and 39% were males. The college students were recruited online to perform as partners of the elderly. All subjects completed an institutional review board (IRB)-approved consent form prior to participating in a data acquisition session.

Scenarios

We asked participants to perform each scenario, described in Table 1, five times. Each interaction scenario is defined as a pair of coordinated behaviors: an *initiating* behavior performed by an elderly person, and a *responsive* behavior performed by a partner. The initiating behaviors consisted of eight greeting behaviors motivated by Heenan et al. (2014) and an additional two behaviors of *high-five* and *hit*. The responsive behaviors were designed so that, when performed by service robots, they would be acceptable to people as natural and humble reactions. Since we did not instruct the participants to act in an exact pattern, there were large variations in intra-class action trajectories.

Collection setups

Our interaction data were collected in an apartment and a senior welfare center where service robots are likely to be used. Figure 2 shows the locations of participants and cameras in the apartment environment. Scenarios 1 and 10 were performed at the front door to allow elderly people to enter and exit through the door, and the other scenarios were performed in the living room, which was equipped with a TV and a sofa. Figure 3 shows the senior welfare

Table 1. 10 interaction scenarios. (E: elderly person, R: partner performing as a robot)

Scenario	
1	E: enters into the service area through the door. R: bows to the elderly person.
2	E: stands still without a purpose. R: stares at the elderly person for a command.
3	E: calls the robot. R: approaches the elderly person.
4	E: stares at the robot. R: scratches its head from awkwardness.
5	E: lifts his arm to shake hands. R: shakes hands with the elderly person.
6	E: covers his face and cries. R: stretches his hands to hug the elderly person.
7	E: lifts his arm for a high-five. R: high-fives with the elderly person.
8	E: threatens to hit the robot. R: blocks the face with arms.
9	E: beckons to go away. R: turns back and leaves the service area.
10	E: turns back and walks to the door. R: bows to the elderly person.

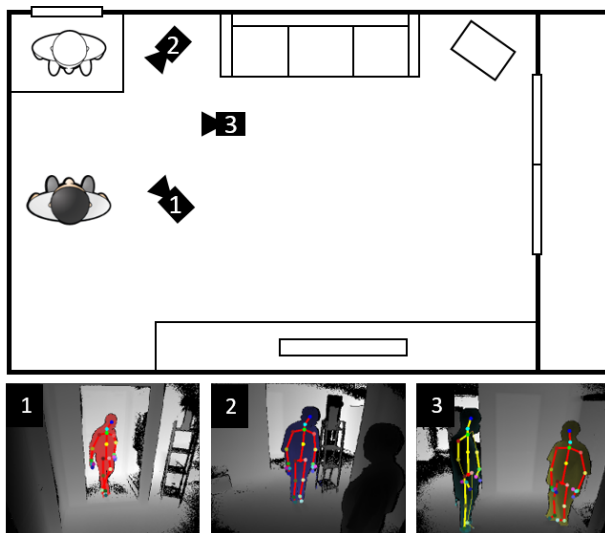
center environment, which is a meeting room with a door. For each scenario, three cameras were set up at the same height; however, were positioned to capture different views. Two cameras were placed next to each person to capture the behaviors from the other person’s point of view. The last camera was placed in a position where both participants were visible in order to gather information of the participants relative to each other. The position of each camera was adjusted each time to take into consideration the movement range of the participants. In total, the entire dataset has 5,000 interaction samples with three different views, where each view lasts for about 6 s.

Data modalities

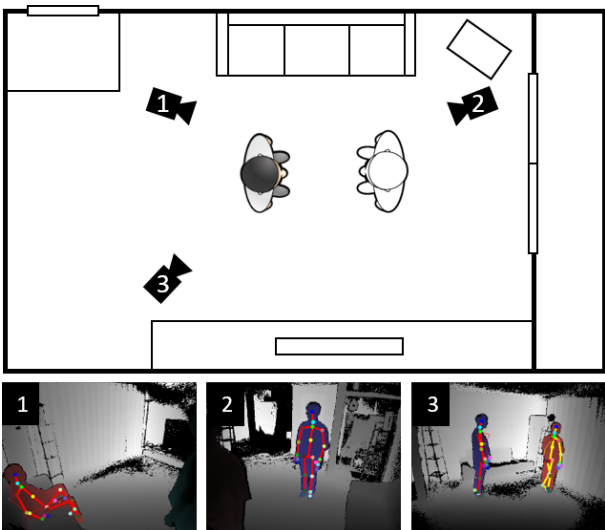
We used three Microsoft Kinect v2 sensors to collect our dataset. Each Kinect sensor provided depth maps, body indexes, and 3D skeletal data by using the APIs of Microsoft Corp. (2014). The depth maps and body indexes have a resolution of 512×424 pixels and a frame rate of 30 fps. The 3D coordinates of 25 joints were obtained via trained and randomized decision tree forests proposed in Shotton et al. (2011). We refined the skeletal data to check the existing tracking failures by manually checking the whole dataset. In addition, the human behaviors to be learned are transformed into the behaviors of a NAO robot. The detailed information of each data file is explained in the following.

Description of depth map

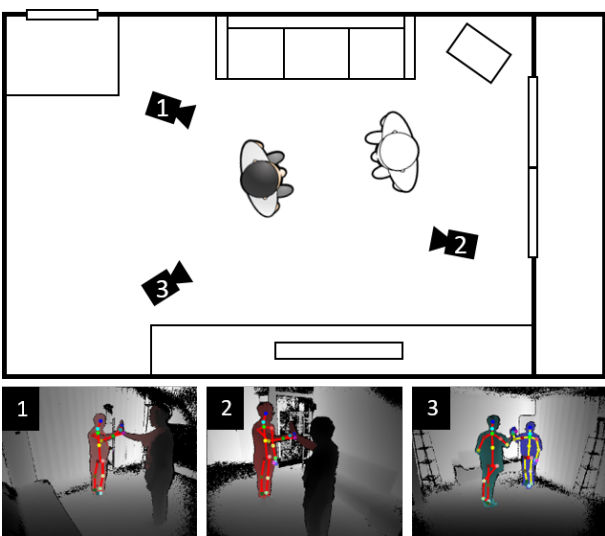
The depth map is a sequence of two-dimensional depth values in millimeters, where each of the depth value ranges from 0 to 8,000 mm. Each individual map is stored in a separate *.png* file, which is a 16-bit grayscale image with a resolution of 512×424 pixels. Figures 2 and 3 show an example of depth maps displayed in grayscale.



(a) Scenarios 1 and 10



(b) Scenarios 2, 3 and 4



(c) Scenarios 5, 6, 7, 8 and 9

Figure 2. The locations of participants and cameras in the apartment environment. The elderly person and his/her partner are denoted as white and colored people, respectively.

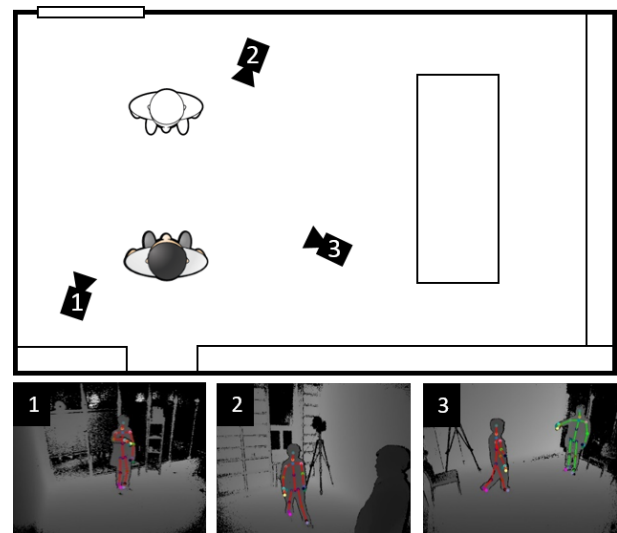


Figure 3. The senior welfare center environment.

Description of body indexes

The body index frame indicates which of the tracked subjects the depth pixels belong to. Since Kinect sensors can detect up to six people, the body index of each person is set to a value between 0 and 5; any other value indicates that it is the background. Each individual body index frame is stored in a separate *.png* file which is an 8-bit grayscale image. Figure 2 shows an example of body indexes denoted by different colors on the depth maps. Note that the body indexes were captured only in the apartment environment.

Description of skeletal data

The skeletal information consists of 3D coordinates of 25 major body joints for each detected human body in the scene. The original skeletal data were stored in a *~joint* file (JSON format) and contained the following information for each body:

- *bodyID*: tracking ID for the body;
- *trackingState*: tracking state of the body;
- *leanX*, *leanY*: lean vector of the body;
- x_j, y_j, z_j : 3D location of the j th joint in camera space;
- *trackingState_j*: tracking state of the j th joint;
- *orientationX_j*, *orientationY_j*, *orientationZ_j*, *orientationW_j*: orientation of the j th joint;
- *depthX_j*, *depthY_j*: 2D location of the j th joint on the depth map.
- *depthZ_j*: depth value of the j th joint.

The example skeletons are denoted by dots and lines on the depth maps in Figures 2 and 3.

Refinement of skeletal data

The skeletons extracted by the Kinect sensor are mostly accurate; however, there are some tracking failures in some frames. To account for this, we manually refined the skeletal data and stored it in a *joint* file (JSON format). Firstly, we restored the incorrectly inferred or missing skeleton by interpolating the front and back frames. Figure 4 shows the example frames before and after the interpolation.

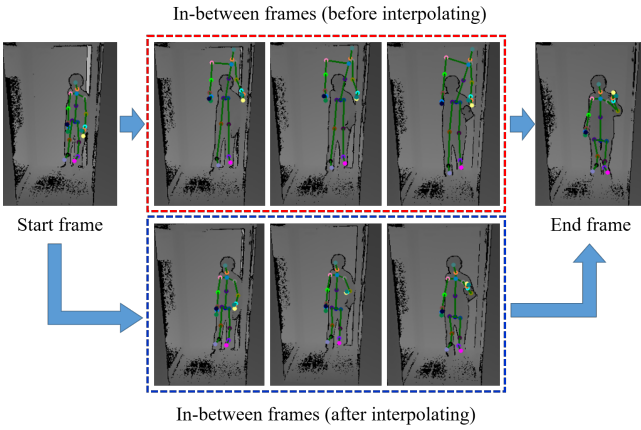


Figure 4. Example frames before and after the interpolation.

Secondly, additional skeletons other than those the camera should capture were removed, as shown in Figure 5. This was done to ensure that in the videos taken by camera 1, there was only the skeleton of the elderly participant remaining, and in the videos taken by camera 2, there was only the skeleton of the partner remaining. In the videos taken by camera 3, the additional skeleton was removed to retain the skeletons of the elderly participant and their partner only.

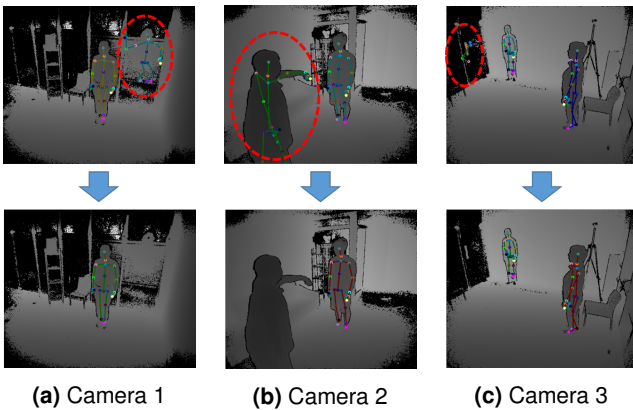


Figure 5. Example frames before and after removing skeletons.

Finally, if the 3D joint locations in the camera space, i.e., x_j , y_j and z_j (floats), were missing or incorrectly inferred, they were re-inferred from the 2D joint locations in the depth space and depth values. The 2D locations and depth values were manually refined in advance.

Generation of robotic data

Since the purpose of our dataset is to train robot intelligence, we transformed the behavior of the human that robots should learn, i.e., the partner of the elderly person, into the action of the robot. We choose the NAO robot as the target robot platform. This is a humanoid robot developed by Aldebaran Robotics (2012). It is designed to show human-like movements with two arms, two legs and a head. Because of the balancing problem, we did not consider the lower body movement; therefore, 10 joint angles, i.e., pitch of hip and head, pitch and roll of L/R shoulders, yaw and roll of L/R elbows, were analytically calculated. The joint angles were stored in a *.nao* file (JSON format). Figure 6 shows an example of the 3D skeletal data of a partner and the transformed pose of a NAO robot.

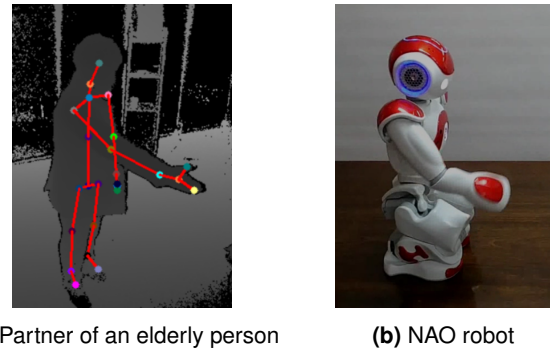


Figure 6. Example 3D skeletal data of a partner, and the transformed pose of a NAO robot.

Comparison with other datasets

The comparison of our dataset with the existing human-human interaction datasets is presented in Table 2. The numbers of subjects, actions, samples and views, and data modalities are summarized. Our dataset is the only interaction dataset of the elderly that provides robotic data to be learned. Moreover, it is one of the largest datasets, containing 5,000 interaction samples of 100 subjects. Each interaction sample provides depth maps, body indexes, and 3D skeletal data captured from three different points of view.

Data usage

Our dataset can be used to train various robot intelligence.

Table 2. Comparison of our dataset with existing human-human interaction datasets. (D: depth, S: skeleton, B: body index)

	#Subjects	#Actions	#Samples	#Views	Data Modalities	Year	Description
<i>Hollywood2</i>	-	12	2,517	1	RGB	2009	YouTube
<i>UT-Interaction</i>	6	6	120	1	RGB	2010	Outdoor
<i>TV Human Interaction</i>	-	4	300	1	RGB	2010	YouTube
<i>SBU Kinect Interaction</i>	7	8	300	1	RGB+D+S	2012	
<i>K3HI</i>	15	8	312	1	S	2013	
<i>JPL-Interaction</i>	8	9	180	1	RGB+S	2015	Human-robot
<i>ShakeFive2</i>	33	8	153	1	RGB+S	2016	
<i>DeepMind Kinetics</i>	-	11	6,378	1	RGB	2017	YouTube
<i>NTU RGB+D 120</i>	106	26	8,276	3	RGB+D+S	2019	
<i>AIR-Act2Act</i>	100	10	5,000	3	D+S+B+Robotic	2019	The elderly

Social behavior generation

The dataset provides the skeletal data of two human participants that interact with each other. Among them, the poses of an elderly person can be used as training input to determine which behavior to generate. At the same time, the poses of the other person can be used as the ground truth robot behaviors after retargeting to a humanoid robot.

Human action recognition

The dataset provides the videos captured from three different points of view. One is the third-person's point of view and the other two are the first-person point of view of each participant. The videos captured from the third person's point of view can be used as training input to recognize interactive behaviors, and the rest to recognize one-person actions. At the same time, the behaviors defined in Table 1 can be used as recognition results.

Data access method

Each subset of the *AIR-Act2Act* and useful python scripts are available for download at <https://github.com/ai4r/AIR-Act2Act>. The scripts include loading data into arrays and plotting skeletal data on depth maps, which will serve as good examples of how to utilize the dataset. The information of the license details can also be found on the webpage.

Acknowledgements

This work was supported by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2017-0-00162, Development of Human-care Robot Technology for Aging Society)

References

- Aldebaran Robotics (2012) Nao. http://doc.aldebaran.com/2-1/home_ nao.html.
- Heenan B, Greenberg S, Aghel-Manesh S and Sharlin E (2014) Designing social greetings in human robot interaction. In: *Proceedings of the 2014 conference on Designing interactive systems*. ACM, pp. 855–864.
- Hemminahaus J and Kopp S (2017) Towards adaptive social behavior generation for assistive robots using reinforcement learning. In: *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 332–340.
- Hu T, Zhu X, Guo W and Su K (2013) Efficient interaction recognition through positive action representation. *Mathematical Problems in Engineering* 2013.
- Huang CM and Mutlu B (2012) Robot behavior toolkit: generating effective social behaviors for robots. In: *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, pp. 25–32.
- Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P et al. (2017) The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.
- Ko WR, Yoon Y, Jang M, Lee J and Kim J (2018) End-to-end learning-based interaction behavior generation for social robots. In: *ICSR2018 Workshop on Social Human-Robot Interaction of Service Robots*.
- Liu J, Shahroudy A, Perez M, Wang G, Duan LY and Kot AC (2019) Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* DOI:10.1109/TPAMI.2019.2916873.
- Marszalek M, Laptev I and Schmid C (2009) Actions in context. In: *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, pp. 2929–2936.
- Microsoft Corp (2014) Kinect for Windows SDK 2.0 Documentation.
- Patron-Perez A, Marszalek M, Zisserman A and Reid ID (2010) High Five: Recognising human interactions in TV shows. In: *BMVC*, volume 1. Citeseer, p. 2.
- Qureshi AH, Nakamura Y, Yoshikawa Y and Ishiguro H (2016) Robot gains social intelligence through multimodal deep reinforcement learning. In: *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. IEEE, pp. 745–751.
- Ryoo MS and Aggarwal J (2010) UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). In: *IEEE International Conference on Pattern Recognition Workshops*, volume 2. p. 4.
- Ryoo MS, Fuchs TJ, Xia L, Aggarwal JK and Matthies L (2015) Robot-centric activity prediction from first-person videos: What will they do to me? In: *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Portland, OR, pp. 295–302.
- Shahroudy A, Liu J, Ng TT and Wang G (2016) NTU RGB+D: A large scale dataset for 3D human activity analysis. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1010–1019.
- Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A and Blake A (2011) Real-time human pose recognition in parts from single depth images. In: *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, pp. 1297–1304.
- Tome D, Russell C and Agapito L (2017) Lifting from the deep: Convolutional 3d pose estimation from a single image. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2500–2509.
- Van Gemeren C, Poppe R and Velthkamp RC (2016) Spatio-temporal detection of fine-grained dyadic human interactions. In: *International Workshop on Human Behavior Understanding*. Springer, pp. 116–133.
- Xia L, Chen CC and Aggarwal JK (2012) View invariant human action recognition using histograms of 3d joints. In: *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, pp. 20–27.
- Yun K, Honorio J, Chattopadhyay D, Berg TL and Samaras D (2012) Two-person interaction detection using body-pose features and multiple instance learning. In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, pp. 28–35.