2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly

Jinhyeok Jang, Dohyung Kim*, Cheonshu Park, Minsu Jang, Jaeyeon Lee, Jaehong Kim

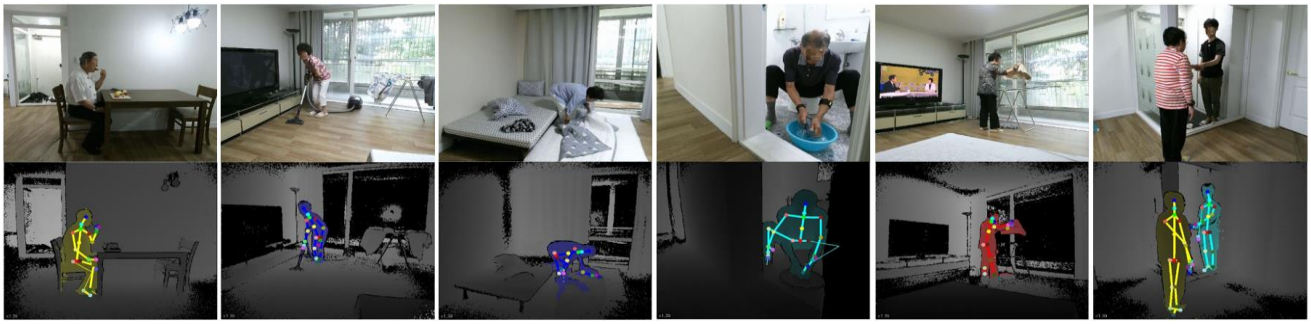

Figure 1. Examples of daily actions in the proposed dataset are displayed along with the corresponding depth map and skeleton information obtained from the Kinect v2 sensors. Actions (from left to right): *eating food with a fork, vacuuming the floor, spreading the bedding, washing a towel by hands, hanging out laundry, and hand shaking*. The entire dataset is available for download at the following link: https://ai4robot.github.io/etri-activity3d-en.

***Abstract*— Deep learning, based on which many modern algorithms operate, is well known to be data-hungry. In particular, the datasets appropriate for the intended application are difficult to obtain. To cope with this situation, we introduce a new dataset called ETRI-Activity3D, focusing on the daily activities of the elderly in robot-view. The major characteristics of the new dataset are as follows: 1) practical action categories that are selected from the close observation of the daily lives of the elderly; 2) realistic data collection, which reflects the robot's working environment and service situations; and 3) a large-scale dataset that overcomes the limitations of the current 3D activity analysis benchmark datasets. The proposed dataset contains 112,620 samples including RGB videos, depth maps, and skeleton sequences. During the data acquisition, 100 subjects were asked to perform 55 daily activities. Further, the domain difference between both groups of age was verified experimentally.**

I. Introduction

The tremendous success of deep learning approaches has brought a substantial improvement in several computer vision tasks. Because these approaches are heavily dependent on large and reliable datasets, efforts have been made to create such datasets, resulting in several publicly available datasets. This is also the case with human action understanding [1, 2, 9, 10, 11]. However, the characteristics of the reported datasets are biased toward their intended applications.

In this study, we employ the recognition of daily activities from a robot's point of view. We believe that the elderly in particular would be the first serious users of robot services. Therefore, they are primarily assumed to be the users. However, evidently, the reported datasets do not fit in this important application.

Therefore, ETRI-Activity3D, which precisely targets the application of recognition of daily activities of the elderly from the robot's point of view, is employed. The unique characteristics of this proposed dataset over the existing ones are provided in the following.

A new visual dataset based on observations of the daily activities of the elderly: A close understanding of what the elderly actually do in their daily lives is essential for constructing a useful dataset. Therefore, we visited the homes of 53 elderly people over the age of 70 years and carefully monitored and documented their daily behavior from morning to night. Then, we selected 55 most frequent actions. Further, while constructing the dataset, we recruited 50 elderly people aged between 64 and 88 years, which led to a realistic intra-class variation of the actions. Additionally, we recruited 50 young people in their 20s. This composition allowed us to perform various comparative experiments between the behavior of the elderly and that of the young. Thus, we were provided with a deeper understanding of the behavioral characteristics of the elderly.

J. Jang, D. Kim, C. Park, M. Jang, J. Lee and J. Kim are with the Human Robot Interaction Research Lab, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea ({jjh6297, dhkim008, bettle, minsu, leejy, jhkim504}@etri.re.kr).

* Corresponding author

A realistic dataset considering the service situation of human-care robots: The aim of the proposed dataset is to be utilized in practical research that can be applied to real-world environments. Therefore, while designing the dataset, probable situations that can occur when robots are in service are closely investigated. The data acquisition is performed in an apartment that has conditions quite similar to the living conditions of the elderly. Each subject is advised to perform the given action in his/her own way. Additionally, the subjects' actions are carried out in diverse environmental conditions such as the time of the day or places in the apartment and in various human postures. We also acquired data from the probable locations where the robot is supposed to be when it serves humans. For instance, in small spaces such as kitchens or bathrooms, the proposed dataset includes scenes where humans face their backs while performing the actions. Although recognizing the actions from their backs is difficult, we believe that these actions are realistic situations that may occur frequently. These considerations make the proposed dataset more challenging, but they make the dataset more practical.

A large-scale RGB-D action recognition dataset that overcomes the limitations of the reported datasets: A sufficient amount of data is crucial for developing deep learning algorithms. However, building a sufficiently large dataset is extremely expensive and difficult, especially for 3D actions, because there is no proper way to leverage video-sharing services or crowdsourcing as in their 2D counterparts. Because of these difficulties, only a limited number of 3D datasets have been reported in several aspects. Among them, the major drawbacks are a small number of subjects and action categories and monotonous environmental conditions. As presented in Table I, the proposed dataset consists of 112,620 video samples, which is comparable to the current largest 3D action dataset, NTU RGB+D 120 [10]. A dataset of this scale with realistic variations may aid in researching extensively from a variety of perspectives, which are expected to contribute to the advancement of robotics intelligence research.

In summary, the major contributions of this study are as follows: 1) a new large-scale dataset for the action recognition of the elderly is introduced, 2) the efficiency and usefulness of the dataset are validated using extensive experiments.

II. Related work

In this section, we briefly review the publicly available datasets for 3D daily activity recognition and the recent deep learning methods for human action recognition.

A. 3D daily activity recognition datasets

Table I shows the most popular public datasets captured indoors using the Kinect sensors. Although each one has its own unique characteristics, it has limitations.

RGBD-HuDaAct [3] is one of the largest datasets for the home-monitoring-oriented activity recognition. It contains 1,189 synchronized color-depth video streams that provide rich intra-class variations. However, because the background is limited to a single lab environment with a fixed camera, the dataset is not suitable for practical benchmarks.

TABLE I. Comparison between ETRI-Activity3D and other publicly available datasets for 3D daily activity recognition. Data modalities: (RGB)video, (D)epth, (S)keleton, (I)R.

Datasets	#Samples	#Sub	#Act	Modalities
RGBD-HuDaAct [3]	1,189	30	13	RGBD
MSRDailyActivity3D [4]	320	10	16	RGBDS
Act4^2 [5]	6,844	24	14	RGBD
CAD-120 [6]	120	4	10+10	RGBDS
Office Activity [7]	1,180	10	20	RGBD
UWA3D Multiview II [8]	1,075	10	30	RGBDS
NTU RGB+D [9]	56,880	40	60	RGBDSI
NTU RGB+D 120 [10]	114,480	106	120	RGBDSI
Toyota Smarthome [11]	16,129	18	31	RGBDS
ETRI-Activity3D	**112,620**	**100**	**55**	RGBDS

MSRDailyActivity3D [4] is designed to include the human daily activities in the living room. It contains 320 samples of 16 activities performed by 10 actors, either sitting on the sofa or standing close to it. However, the small number of samples and fixed camera viewpoints are the limitations of this dataset.

CAD-120 [6] focuses on high-level activities and object interactions. It comprises 120 long-term activities, such as *making cereal* and *microwaving food*. Each video is annotated with human skeleton tracks, object tracks, object affordance labels, sub-activity labels, and high-level activities.

Toyota Smarthome [11] is a real-world video dataset for activities of daily living. It consists of 16,129 RGB+D clips of 31 activity classes performed by the elderly in a smart home. Unlike the other datasets, the videos were completely unscripted in this dataset; this brought out several real-world challenges. However, the limited number of subjects and activity classes are the drawbacks of this dataset.

NTU-RGB+D 120 [10], an extension of NTU-RGB+D [9], is the current largest benchmark dataset for 3D action recognition. It includes 114,480 video samples collected from 120 action classes in multi-view settings. This large-scale dataset has contributed to the development and evaluation of data-driven learning methods. However, this dataset was acquired in a laboratory environment, and the activities were performed with a strict guidance; these do not reflect the realistic challenges that exist in daily activities.

III. ETRI-Activity3D

The ETRI-Activity3D dataset is collected using Kinect v2 sensors, and it consists of three synchronized data modalities: RGB video, depth map, and skeleton sequence. The resolution of RGB video is 1920 × 1080, and the depth map is stored frame by frame in a 512 × 424 resolution. The skeleton sequence contains the 3D locations of 25 body joints of tracked human bodies. There are 55 action categories, of which 52 are derived from the observation of daily activities (eating, cleaning, reading, etc.) of the elderly and the rest 3 are human-robot interaction specific actions (waving, beckoning, and pointing). Among them, there are five mutual actions such as handshaking and hugging.

The number of subjects was 100, of which 50 were senior citizens, and the rest were young adults. The age of the elderly subjects ranged from 64–88 years with an average of 77 years, whereas the young subjects were in their 20s with an average of 23 years. Among the elderly, 17 were men and 33 were women, whereas the numbers of men and women were 25 for young adults.

When collecting the data, the expected robot view was considered. That is, the capturing device was located at heights of 70 cm and 120 cm, which are based on the typical height of human-care robots, as shown in Fig. 2. Four capturing platforms (each one capturing from both the aforementioned heights) were arranged to capture various views of the action simultaneously. Additionally, the distance between the sensor and subject varied from 1.5–3.5 m. The actions that could be carried out independent of places (e.g., taking medicine or talking on the phone) were captured up to five times at all different places. In this way, we could provide additional intra-class variations in views and backgrounds.

The subjects were advised to ignore the cameras to capture actions as natural as possible. Different postures were recommended while carrying out the action (e.g., taking medicine while sitting or standing, etc.). Additionally, different shapes of relevant objects were provided to the subjects, and they were asked to hold the objects with their right or left hand.

Thus, we collected a large-scale RGB-D dataset with 112,620 samples. Detailed information on the dataset can be found at the following link: https://ai4robot.github.io/etri-activity3d-en.

Table II shows the differences between actions performed by the elderly and adults observed on the ETRI-Activity3D dataset. The first two rows clearly indicate that the elderly act slower than the adult. The frame length and motion differentials were calculated using normalized skeletons, and three action classes were excluded because of the strong noise. This statistical analysis indicates that the elderly act quite differently from the young, which suggests that the elderly subjects should be included in building realistic datasets.

IV. Experiments

We evaluated the many state-of-the-art methods using ETRI-Activity3D and NTU RGB+D [9] datasets. Additionally, we analyzed the ETRI-Activity3D to reveal the differences between the data of the elderly and adults.

Further, the data augmentation such as 3D rotation, body-shape variation, and noise was applied. Moreover, the random sampling based on the different lengths of data sequence provided the effect of data augmentation. For the NTU RGB+D dataset, the input length ranged from 32–128, and for the ETRI-Activity3D dataset, it ranged from 32–200.

A. Action recognition on NTU RGB+D and ETRI-Activity3D

For an objective comparison with the other reported networks, FSA-CNN was applied to the NTU RGB+D dataset and our new ETRI-Activity3D dataset. The accuracies

TABLE II. Performances of state-of-the-art methods on NTU RGB+D and ETRI-Activity3D.
CS: Cross-Subject, CV: Cross-View.

Method	NTU RGB+D		ETRI-Activity3D
	CS (%)	CV (%)	CS (%)
IndRNN [16]	81.8	88.0	73.9
Beyond Joint [15]	79.5	87.6	79.1
SK-CNN [12]	83.2	89.3	83.6
ST-GCN [18]	81.5	88.3	86.8
Motif ST-GCN [19]	84.2	90.2	89.9
Ensem-NN [14]	85.1	91.3	83.0
MANs [17]	83.0	90.7	82.4
HCN [13]	86.5	91.1	88.0
FSA-CNN[20]	88.1	92.2	90.6

of other methods on NTU RGB+D have already been reported in their studies. However, their accuracies on ETRI-Activity3D have been measured in our experiments using their publicized codes.

The NTU RGB+D dataset suggests two experimental protocols: cross-subject and cross-view. We followed the same protocols. The experimental results and comparisons are reported in Table III.

Additionally, we followed the "cross-subject" protocol for the ETRI-Activity3D dataset. To do so, we divided the entire subjects' data into a training and test sets. The training set

TABLE IV. Performances of FSA-CNN when trained on three domains: the Elderly, Adults, and Mixed.

Train \ Test	Elderly	Adults
Elderly	87.7	69.0
Adults	74.9	85.0
Mixed	84.8	82.1

consisted of 67 subjects', and the test set consisted of the remaining 33 subjects' data. The subject IDs of the test set are {3, 6, 9, 12 … 99}, and the remaining are of the training set. As shown in Table III, the FSA-CNN achieved the best accuracy, i.e., 90.6%, on the ETRI-Activity3D. Note that the size of ETRI-Activity3D is approximately twice that of NTU RGB+D.

All the methods presented in Table III use only the skeletal data for recognition. Beyond Joint [15], IndRNN [16], and MANs [17] attempted to use an RNN for action recognition. Because RNN-based methods use sequential data, they are not restricted by a fixed input length. However, they are highly sensitive to noise, and they should have the ability to forget the past data at an appropriate time. Owing to these reasons, RNN-based approaches often show lower accuracies, but they are highly robust to variations in input length.

The architectures of SK-CNN [12], Ensem-NN [14], and HCN [13] and FSA-CNN[20] typically extract spatial and temporal features and combine them to recognize human

actions. However, many of these methods suffer from overfitting and slow operating speeds.

ST-GCN [18] and Motif ST-GCN [19] adopted a GCN as a classifier. The GCN is a CNN having links with each node considering the natural connections of joints in the human body. Thus, the classifiers achieve good performances by considering domain knowledge. However, they have some disadvantages like needs of pre-defined adjacency matrix. Moreover, it is impossible to apply to skeleton data with different structures.

B. Analytical experiments

We applied FSA-CNN to verify the following issues. We divided the ETRI-Activity3D dataset into two: the elderly and adults, and trained the two separate networks using each subset to evaluate the cross-age performance.

1) Analysis of domain difference

The ETRI-Activity3D dataset consisted of 50 elderly and 50 young adults. For this cross-age experiment, the training and test sets of the cross-subject protocol were split again to generate 4 sets: training and test sets for the elderly and the corresponding two sets for young adults. The subject IDs of the elderly test set were {3, 6, 9, 12 … 48}, and those of the adult test set were {51, 54, 57 … 99}. Then, the two separate networks were trained using the training sets of the elderly and young adults, respectively, and they were then tested using both the test sets.

Additionally, we trained a network using mixed data. In this case, only half of the mixed training dataset was used so that the number of subjects remains the same as the other two. The subject IDs were {1, 4, 7, 10 … 97}. As shown in Table IV, the network trained using the data of the elderly exhibited significantly higher performance on the corresponding test dataset and vice versa; this indicates that there are noticeable differences between the actions of the two domains. The network trained on the combined dataset exhibited relatively good performance in both domains; this proves the need of a dataset containing the data of the elderly.

V. CONCLUSION

This study proposes a new dataset that includes the actions of the elderly and a novel network for recognizing the actions, and it provides analytical experimental results. ETRI-Activity3D, a new large-scale 3D action recognition dataset containing 112,620 video samples collected using 55 action classes and 100 subjects (50 elderly people and 50 adults), is used. This dataset is expected to be useful for research on action recognition, robotic intelligence, and elderly care. Additionally, the domain differences between actions of the elderly and adults were verified both statistically and experimentally.

ACKNOWLEDGMENT

This work was supported by the ICT R&D program of MSIP/IITP. [2017-0-00162, Development of Human-care Robot Technology for Aging Society]. The protocol and consent of data collection were approved by the Institutional Review Board(IRB) at Suwon Science College.

REFERENCES

[1] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "Rgb-d-based action recognition datasets: A survey," in *Pattern Recognition*, vol. 60, 2016.

[2] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "Rgb-dbased human motion recognition with deep learning: A survey," in *Computer Vision and Image Understanding*, vol. 171, 2018.

[3] B. Ni, G. Wang, and P. Moulin, "Rgbd-hudaact: A color-depth video database for human daily activity recognition," in *IEEE Int. Conf. Computer Vision Workshops*, 2011.

[4] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[5] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian, "Human daily action analysis with multi-view and color-depth data," in *European Conf. Computer Vision*, 2012.

[6] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," in *Int. J. Robotics Research*, vol. 32, no. 8, 2013.

[7] L. Wang, Y. Qiao, and X. Tang, "Action recognition and detection by combining motion and appearance features," THUMOS, 2014.

[8] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian, "Histogram of oriented principal components for cross-view action recognition," in *IEEE Tran. Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, 2016.

[9] A. Shahroudy, J. Liu, T. T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2016.

[10] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Y. Duan, and A. K. Chichung, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," in *IEEE Tran. Pattern Analysis and Machine Intelligence*, 2019.

[11] S. Das, R. Dai, M. Koperski, L. Minciullo, L. Garattoni, F. Bremond, and G. Francesca, "Toyota smarthome: Real-world activities of daily living," in *IEEE Int. Conf. Computer Vision*, 2019.

[12] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks." in *IEEE Int. Conf. Multimedia &Expo Workshops*. 2017.

[13] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation." in *Int. Joint Conf. Artificial Intelligence*. 2018.

[14] Y. Xu, J. Cheng, L. Wang, H. Xia, F. Liu, and D. Tao, "Ensemble one-dimensional convolution neural networks for skeleton-based action recognition." in *IEEE Signal Processing Letters*. vol. 25, no. 7, 2018, pp. 1044-1048.

[15] H. Wang, and L. Wang, "Beyond joints: Learning representations from primitive geometries for skeleton-based action recognition and detection." in *IEEE Trans. Image Processing*. vol. 27, no. 9, 2018, pp. 4382-4394.

[16] S.Li, W. Li, C. Cook, C. Zhu, and Y. Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn." in *IEEE Conf. Computer Vision and Pattern Recognition*. 2018.

[17] C. Xie, C. Li, B. Zhang, C. Chen, J. Han, C. Zou, and J. Liu "Memory attention networks for skeleton-based action recognition." in *Int. Joint Conf. Artificial Intelligence*. 2018.

[18] S. Yan, X. Yuanjun and L. Dahua, "Spatial temporal graph convolutional networks for skeleton-based action recognition." in *AAAI conf. Artificial Intelligence*. 2018.

[19] Y. H. Wen, L. Gao, H.Fu, F. L. Zhang, and S. Xia, "Graph CNNs with motif and variable temporal block for skeleton-based action recognition." in *AAAI Conf. Artificial Intelligence*. 2019.

[20] J. Jang, D. Kim, C. Park, M. Jang, J. Lee, and J. Kim, "ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly." arXiv preprint arXiv:2003.01920. 2020