An efficient activity recognition for homecare robots from multi-modal communication dataset

Mohamad Yani [a,b,1,*], Nao Yamada [a,2], Chyan Zheng Siow [a,3], Naoyuki Kubota [a,4]

[a] Graduate School of Systems Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo, 191-0065, Japan
[b] Institut Teknologi Telkom Surabaya, Jl. Ketintang No.156, Ketintang, Kec. Gayungan, Kota SBY, Jawa Timur 60231, Indonesia
[1] mohamad-yani@ed.tmu.ac.jp, mai@ittelkom-sby.ac.id; [2] 0809n.ysoccer@gmail.com; [3] siow-chyanzheng@ed.tmu.ac.jp, [4] kubota@tmu.ac.jp
* corresponding author

ARTICLE INFO

Article history
Received August 26, 2022
Revised September 26, 2022
Accepted December 22, 2022
Available online March 31, 2023

Keywords
Homecare robots
Activity prediction
Graph neural network
RGB-D camera
ROS

ABSTRACT

Human environments are designed and managed by humans for humans. Thus, adding robots to interact with humans and perform specific tasks appropriately is an essential topic in robotics research. In recent decades, object recognition, human skeletal, and face recognition frameworks have been implemented to support the tasks of robots. However, recognition of activities and interactions between humans and surrounding objects is an ongoing and more challenging problem. Therefore, this study proposed a graph neural network (GNN) approach to directly recognize human activity at home using vision and speech teaching data. Focus was given to the problem of classifying three activities, namely, eating, working, and reading, where these activities were conducted in the same environment. From the experiments, observations, and analyses, this proved to be quite a challenging problem to solve using only traditional convolutional neural networks (CNN) and video datasets. In the proposed method, an activity classification was learned from a 3D detected object corresponding to the human position. Next, human utterances were used to label the activity from the collected human and object 3D positions. The experiment, involving data collection and learning, was demonstrated by using human-robot communication. It was shown that the proposed method had the shortest training time of 100.346 seconds with 6000 positions from the dataset and was able to recognize the three activities more accurately than the deep layer aggregation (DLA) and X3D networks with video datasets.

1. Introduction

The problem of a large elderly population and a low birth rate has become a serious social issue in several nations. Thus, developing a homecare robot to do work is being considered as a potential solution to the problem of the increase in aging societies across the globe. Many homecare robots are equipped with various sensors, such as RGB-D cameras, 2D LiDAR, stereo cameras, and arms, to understand a situation and carry out physical tasks in a dynamic environment. Furthermore, the Robot Operating System (ROS) is designed as a software system architecture to manage sensor inputs and outputs in many robotics applications. Many researchers are interested in developing a learning model for understanding the context of the environment from object detection and human behaviour estimations [1]–[4]. This capability affects the level of interaction with the environment and the action that is performed. For instance, a homecare robot at home is expected to take the initiative to help in the daily household chores, such as performing tasks in the kitchen to prepare food or wash the dishes, etc. Then, the question arises: What information will the robot need to help humans in their daily activities? Some

of that information is defined as 1) the robot must obtain information on the objects in the home environment, including the class, position, and orientation of the objects; and 2) the robot needs to know the position of the human and recognize the activity when the human interacts with several objects at home.

However, it is no easy task to obtain this high level of ability in homecare robots, such as understanding human activity from semantic information and human-object interaction (HOI). Many efforts have been made to recognize human activities using images and videos with 2D and 3D visual information [5]–[8]. Semantics and the context of a situation are usually used for classification, which involves typical HOI [9]–[11]. However, most of their frameworks require massive labelled datasets and much training time to achieve a high level of accuracy.

Recently, studies on the analysis of graphs using machine learning have been gaining more attention because of the outstanding expressive power of graphs. For instance, graphs can be used to denote a large number of applications across different scopes, including the extraction of topologies and geometries from 3D object detection or point cloud [12]–[14], natural science problems [15], [16], pharmaceutical research [17], [18] and other areas [19]. Graph computation, which is a unique form of data structure for supervised and unsupervised learning strategies, focuses on tasks, such as clustering, link prediction, and classification. Graph neural networks (GNNs) are deep learning-based approaches in the graph domain. Due to their effective performance, GNNs have lately become a widely used strategy for the analysis of graphs. Therefore, formulating these interactions into a graph representation based on the detected object and human features will be of significant help to the robot in learning human activities and deciding on the appropriate tasks.

In this paper, an efficient learning method was proposed to predict daily human activities in an indoor environment. The approach used a GNN, which was trained based on activities recorded directly from the Toyota HSR (Human Support Robot). The activities were labelled directly from human speech. The real-time YOLOv3 object detection [20] and MediaPipe [21] face detection frameworks were implemented using an X-tion RGB-D camera from the HSR head to obtain the object class, geometry, and human face position in a real indoor environment. Human activity is a huge topic in robotics. Therefore, the aim of this paper is to create a method to learn and predict three specific activities, namely, eating, reading, and working, from utterances. With this method, the robot can be trained to predict daily human activities in real-world applications using ROS environment. A human activity dataset recorded directly from a human-robot communication (HRC) system was also introduced. To summarize, the contributions of this work are as follows:

- A GNN model was proposed that utilized the prior information from HOI and HRC to classify human activities, especially eating, reading, and working, in a very similar environment.
- The proposed method can efficiently construct the labelled dataset using human-robot communication. When the robot detects a human doing an activity with some objects, the centre of the detected 2D bounding box object is transformed into a 3D position. Then, the human can ask the robot to collect the dataset of the current activity and train the teaching data using utterances.
- The method was applied using a Toyota HSR robot with an ROS environment system in a real-world indoor environment. The proposed method can tackle the problem of activity recognition in a very similar environment, which is difficult to classify using conventional 2D CNN.

The rest of the paper is arranged as follows. Section 2 discusses the method and explains the related works and the GNN strategy, along with explanations of the network configuration and various features. Following this, section 3 presents an evaluation of the approach in a real-world indoor environment, as well as a comparison with multiple dataset combinations. The conclusions are given in section 4.

2. Method

2.1. Human Activity Prediction

Human activity recognition (HAR) is one of the most difficult tasks in robotics and computer vision since it requires assigning a label to each activity. HAR can be divided into three types: (1) by vision, (2) by sensors, and (3) by waves/radio. (1) Vision typically refers to a camera that utilizes red, green, and blue (RGB) colours from a video to identify a human activity [4], [22]. Nevertheless, RGB is just a 2D vision, and does not indicate how far the object is. Therefore, a depth camera was added to the vision to make it RGB-D [23], [24]. By having depth, the 3D position of an object or human can be efficiently recognized, and the HAR can be realized using the 3D pose and visualization [10]. (2) Sensors refer to devices that do not use RGB-D data [25], such as mobile phone accelerometers and gyroscopes. The data is numeric and reliable, but the drawback of this method lies in the coverage issue, where the individual is required to always carry the device. (3) Finally, waves/radio is the latest technology that utilizes Wi-Fi signals to predict human activity [26]. The recognition coverage is more extensive and there are no problems with privacy. However, this method is still new and requires more extensive research to make it more reliable.

2.2. Human-Object Interactions to Recognize Human Activities

Recognizing human activities using graph convolutional neural networks has attracted the attention of many robotics and computer vision researchers in recent years. Activity recognition is required for homecare robots that are taking care of children, the elderly, or persons with disabilities. With this implementation, the inference of human activities using perceptual information plays an important role in human-robot interactions, smart surveillance, and content-based video analysis. A lot of work has been conducted toward predicting human activities in 2D and 3D images and videos where the overall technique observes and correlates with HOI. [27]–[30]. The principal method for predicting HOI is extracting visual characteristics from instance detectors and spatial knowledge to instantiate multi-streams of deep neural networks. Each stream includes detected human and contextual objects. The last step is designed for inference application. The work of [27] presented the forecasting of a human activity by predicting the possible trajectory movement towards a targeted object from RGB and depth sensors data. Using this strategy, the predicted trajectory for the ongoing action can be visualized. However, the camera should be set up at a certain distance and height to avoid the broader field of view that may lead to occlusions and poor activity prediction. Wang *et al.* [31] proposed a fully-convolutional approach that predicts the interactions between human-object pairs from RGB images. The network can predict the context of human activities by localizing the interaction points from the object, human and pairwise streams.

GNN has been utilized to predict human activities from HOI by extracting the image features into graph structures. Qi et al. [28] used a graph parsing neural network to detect HOI and predict human activity from various RGB datasets. Morais et al. [29] introduced asynchronous-sparse interaction graph networks, which are constructed from the temporal structure and content label of human-object interaction activities. This method uses a graph attention network model for HOI detection in the RGB dataset. Their approach involves the construction of nodes and edges from visual features. An adjacency matrix defines the structure and properties of the network, and is updated by a weighted sum of the messages from the other nodes. Finally, for interaction inference, a node readout function is employed. Simonovsky and Komodakis [32] proposed the edge-conditioned convolution (ECC) GNN, a spatial domain operation on graph signals in which filter weights are conditioned on edge labels and dynamically formed for each input sample. It was demonstrated that this strategy could generalize the traditional convolution on graphs if edge labels are suitably chosen, and this claim was empirically tested on MNIST. Moreover, this method was also evaluated for point cloud classification, achieving a new state-of-the-art performance on the Sydney dataset. The current work was inspired by [32], and their model was used for the proposed method. A 3D object and human pose of point clouds data were used to construct the edge features for the graph data by transforming the centre of the 2D bounding boxes of the YOLO and Mediapipe face detection frameworks into the 3D point cloud by utilizing the ROS library features. The data was labelled using human speech when the robot questioned the human to learn the HOI. In the

application for this study, a 3D object and human pose data were used to construct the graph data by transforming the centre of the 2D bounding boxes into a 3D point cloud by utilizing the ROS library features. The data was labelled using human utterances in a human-robot communication scenario for the data collection. The data was collected during the ongoing activity, and the robot asked the human, "What are you doing now?". Then, the uttered answer would be specified as an activity label. After collecting the data, a graphical representation was constructed for the training process.

2.3. Data Preparation

This section explains the approach for predicting specified human activities at home from a homecare robot system in real-world applications. The dataset was pre-processed and collected from the YOLO and MediaPipe face detection frameworks, and each detected object and activity were classified using direct communication with the Toyota HSR robot. The activity labels were specified into 3 categories; eating, reading, and working. Next, a graph convolutional neural network proposed by [32] was used. The data collection process for the approach is shown in Fig. 1, where the dataset from the YOLO object detection and MediaPipe face detection frameworks was collected using an X-tion RGB-D camera from the Toyota HSR head.

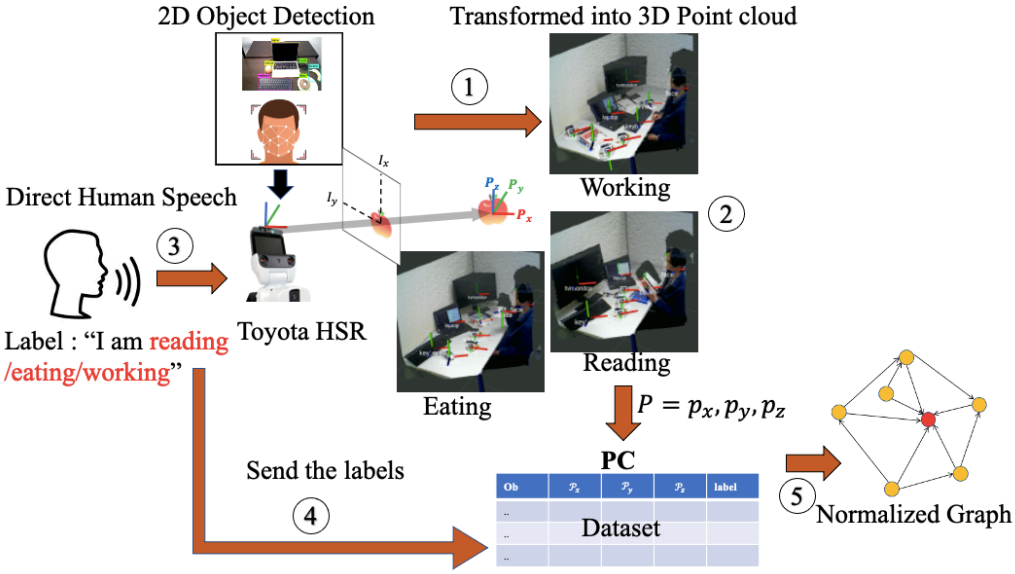

Fig. 1. Data collection system of proposed method and graph data representation

To obtain the 3D position of each detected object and a human face, the centre of the bounding boxes of each detected object and a human face was transformed, using the point cloud data type service in the ROS environment, into a 3D real-world position using tf2 of the ROS library [33]. Next, the object and human positions could be constructed in a graphical representation. During the labelling process, the robot only had a Japanese voice system. Therefore, the human-robot communication for label collection was conducted in Japanese, such as in this video [https://youtu.be/WI8vl_UJwCc]. The Google Cloud Speech-to-Text framework [URL: https://cloud.google.com/speech-to-text] was used to recognize the human voice and convert it to text. The collected dataset was stored on the server of the PC for graph construction and training.

2.4. Graph Construction and Features

In this paper, we represented the HOI containing the relationships between human positions and the surrounding objects as a graph form $G = (V, E)$, where V is the set of $|V| = N$, and N denoted as the centroid of 3D detected object $p \in P$. Meanwhile, E is a set of edges with $|E| = M$, where M is the number of edges. Then, we specify each detected object to graph signal by $X^0(i) = X_P(p)$. Then, we connect each node i to all nodes j by directed edge (j, i). In our works, we specify an edge from j to i by calculating the distance of each number of centroids from the detected object by

$d = P_j - P_i$ where d represented in cartesian coordinates as 3D edge label vector $L(j,i) = (d_x, d_y, d_z)$ and $L: E \to \mathbb{R}^{M \times C}$ is feature matrix with features number of each edges C. Afterwards, the node embedding is constructed from the name of the YOLOv3 object class by creating the vocabulary dataset, which encodes the object names with their IDs. The ID is an integer (index) that identifies a word's position in the vocabulary dataset such as shown in the subgraph in Fig. 2. The vocabulary dataset to construct the node embedding is arranged as follows:

{"face": 0, "tvmonitor": 1, "laptop": 2, "mouse": 3, "keyboard": 4, "book": 5, "soup": 6, "sandwich": 7, "salad": 8, "pizza": 9, "cup": 10}

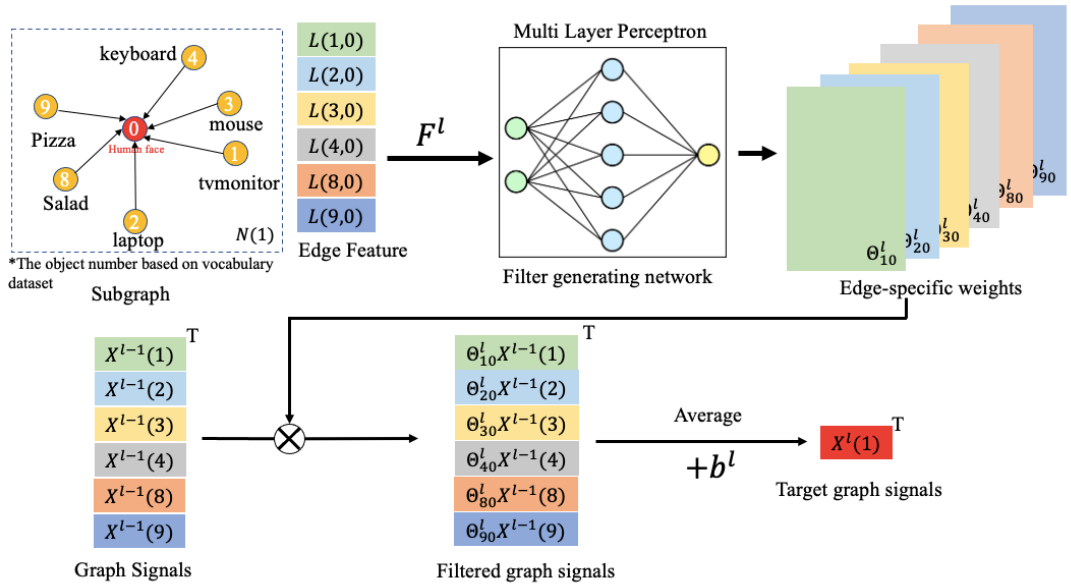

Fig. 2. Processing a subgraph in edge conditional convolution method

2.5. Graph Neural Network with Edge-Conditioned Convolution

A GNN can solve many complex problems, especially for activity classifications [30], social network recommendations [19], and predicting chemical and molecular properties [32]. In this paper, a GNN was used with the ECC by [32] to process the normalized graph structure data. This approach was used [32] to perform the graph classification to predict human activities, such as eating, working, and reading. The entire architecture was made up of 4 ECC layers and 2 fully-connected (FC) layers. The 4 ECC layers were used to aggregate the information from each of the human interactions and objects, while the 2 FC layers were used to perform the final recognition. There are three aggregations (or message-passing) approaches in the ECC, which are "add", "mean", and "max", where "add" is to calculate the total neighbours' features, "mean" is to calculate the average, and "max" is to use the maximum features during the message-passing. In this approach, the "mean" for all the ECC layers was used to obtain stable features. The overview of the GNN architecture is shown in Fig. 3. Moreover, the Cross-Entropy Loss function was utilized in the proposed GNN, as well as the Pytorch Geometric package [34], a python library dedicated to the GNN. The graph neural network can be described in the following form [32], [35] as:

$$X^l(i) = \frac{1}{|N(i)|} \sum_{j \in \mathbb{N}(i)} F^l\left(L_{ji}; w^l\right) X^{l-1}(j) + b^l$$
$$= \frac{1}{|N(i)|} \sum_{j \in \mathbb{N}(i)} \Theta^l_{ji}\, X^{l-1}(j) + b^l \quad (2)$$

where $X^l(i)$ denotes the node embedding corresponding to the i-th vertex in layer l, and $|N(i)|$ is the total number of neighborhood node of i-th node. Each layer l includes a multi-layer perceptron as a filter generating network F^l with learnable w^l and bias b^l that implements aggregation between nodes i and j with edge embeddings L_{ji}. The computed $X^l(i)$ are used to train two FC networks for final prediction. We illustrated the ECC and the overall architecture of the proposed method in Fig. 2 and Fig. 3 respectively.

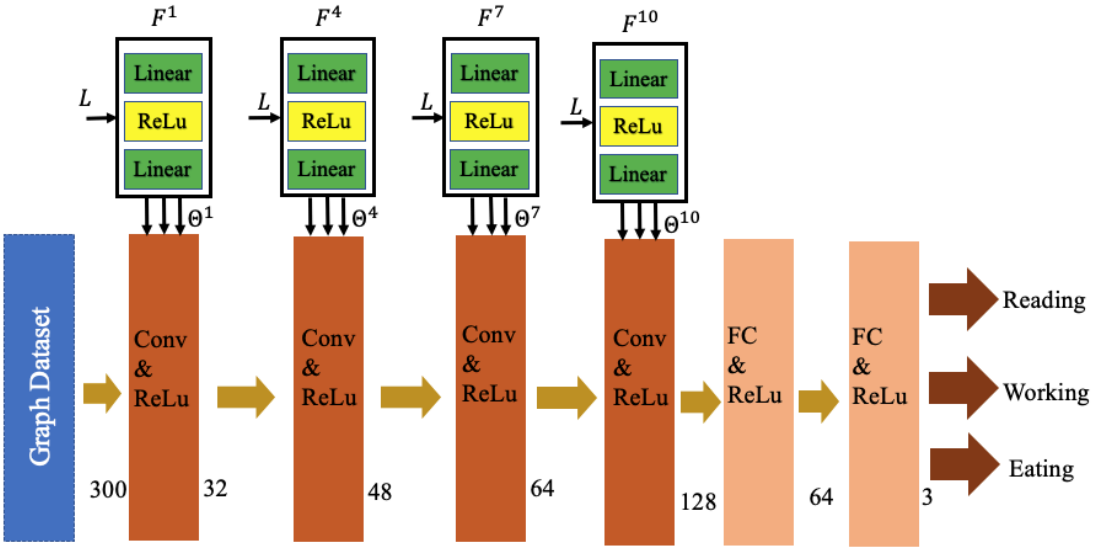

Fig. 3. The proposed GNN with ECC for the activity recognition

2.6. Positional Data HOI and Scenarios

To decrease overfitting of the small dataset, the dataset was collected by making the data slightly different from one group to another group of class of data. During the dataset collection, 2 scenarios (A and B) were specified for each activity, and the subject was asked to follow scenarios A and B in each specified task, such as eating, reading, and working. Table 1 describes in greater detail the scenarios of human-object interaction for the data collection. It was decided to use 8 objects for the 3 activities. The $P_0 = [p_x, p_y, p_z]$ was collected from the human face and object position during the human interaction with an object. When the data reached 1000 for each scenario of A and B, the utterance was used to label the collected dataset. The human utterance was converted into text using Google Speech-to-Text [URL: https://cloud.google.com/speech-to-text]. This automatic speech recognition can accurately convert more than 125 languages. In total, 6000 data were available to train the GNN. Next, the collected dataset was transformed into a graph data representation. 80% of the dataset was used for training the GNN, while 20% was used for testing the networks.

Table 1. Scenarios of the data collection

Activities		Laptop	Tvmonitor	Keyboard	Mouse	Book	Salad	Sandwich	soup
working	*A*	■	■	■	■	■			
	B	■	■	■	■	■			
Eating	*A*	■	■			■	■	■	■
	B	■	■				■	■	■
Reading	*A*					■			■
	B	■		■	■	■			

■ Object always used
■ Object sometimes used

2.7. Comparison Method

In this paper, the proposed method for the recognition of eating, reading, and working in a similar environment was compared with the DLA [36] and X3D networks [37] approach using different datasets. These two architectures commonly use an image or video dataset to make an image or video classification. Therefore, only 6 videos were collected for each specified activity, where 4 videos were for training and 2 for testing. The video dataset was conducted with a slightly different scenario of object and environment, such as described in Fig. 4. Moreover, the video was labelled manually in these methods without a speech-to-text recognition framework.

Fig. 4. Video dataset of eating, reading and working to train and test the DLA and X3D networks as comparison methods.

3. Results and Discussion

3.1. Proposed Method Performance

This section will discuss the experimental setup and several results. The object and human face detection was carried out using MediaPipe and YOLOv3. The centre of the object and face detection was converted into the 3D position using the ROS tf library, as shown in Fig. 5. The subjects were asked to use several objects to perform the specified tasks, such as eating, reading, and working. The Toyota HSR would ask what kind of activity was being done, and the subject would answer by speech.

Fig. 5. Experimental setup, collecting the dataset and visualizing the centre of the object into 3D Position

The robot would collect the detected object and face position with the utterance information. To visualize the brief experimental setup and the results in greater detail, the experimental example, data collection and inferencing were uploaded to this URL: [https://youtu.be/WI8vl_UJwCc]. In the proposed method, 100 epochs were used to train 4000 positional datasets, and the proposed architecture was tested with 2000 positional graph datasets. However, by using DLA and X3D, 50 epochs were used to train the collected videos. Generally, video classification frameworks are trained with huge datasets with a high number of epochs, such as by using UCF101[38] and kinetics dataset [39]. However, this study had its limitations in terms of the dataset collection and computational cost.

Table 2. Comparison of results between the proposed method, DLA and X3D

Network Architecture	Total Training Data	Total Testing Data	Epoch	Training Time(seconds)	Training Loss	Avg Training Accuracy (%)	Avg Testing Accuracy (%)
GNN with ECC(Ours)	4000 positional objects	2000	100	100.346	0.0145	99.789	91.26
DLA	4 videos	2 Videos	50	4030.784	0.306	99.96	22.67
X3D	4 videos	2 Videos	50	129.456	7.187	91.00	64

3.2. Training Results

This section presents the results of the GNN with ECC by using positional datasets. Table 2 gives a comparison of the efficiency of the proposed method with the DLA and X3D network architectures. The details of the DLA architecture are presented in [https://github.com/kuangliu/pytorch-cifar] and those of the X3D network are presented in [https: //github.com/facebookresearch/SlowFast]. It can be seen from Table 2 that the proposed method can efficiently recognize reading, eating, and working from the dataset collection in general. The positional dataset was collected and labelled in csv file format. Therefore, less time was spent on processing this collected dataset than on training the video dataset as only the 3D position of the object had to be processed with 3 discrete classifications. However, the object and face positions had to be extracted from the YOLOv3 and MediaPipe frameworks beforehand. However, it took the longest time to train the 4 videos using the DLA architecture, among other methods. 2D CNN, which commonly requires many datasets and takes a longer time to achieve high accuracy, was used in the DLA architecture. In the X3D, a similar dataset as in the DLA was used, but the X3D showed a better performance with a shorter training time and higher accuracy than the DLA. The outstanding performance of the 3D CNN in X3D architecture is discussed in [37]. Feature extraction of the object and face detection was not required to classify the activity in X3D. The video only had to be resized for the dataset pre-processing, such as in the DLA. It could be concluded from the results that different datasets and tools have different training times and accuracy scores. Nevertheless, the proposed method outperformed the other methods, where the GNN with ECC was able to detect eating, reading, and working based on 3D position data, and achieved a shorter training time. The confusion matrix for each tool is shown in Fig. 6.

Fig. 6. Confusion Matrix for proposed method, DLA and X3D

The X3D network may achieve a similar result as the proposed method if the parameters are tuned and the number of epochs is increased. However, the duration of the training time may have to be increased. This will be investigated in a future work. By using this method, the robot will always be able to detect activities accurately while the objects and humans are in a 3D global position. Therefore, this method can be used directly for navigation targets or manipulation targets based on the current situation. For the inferencing test using an HSR camera, the trained proposed network was tested for the detection of working, eating, and reading activities, as depicted in Fig. 7.

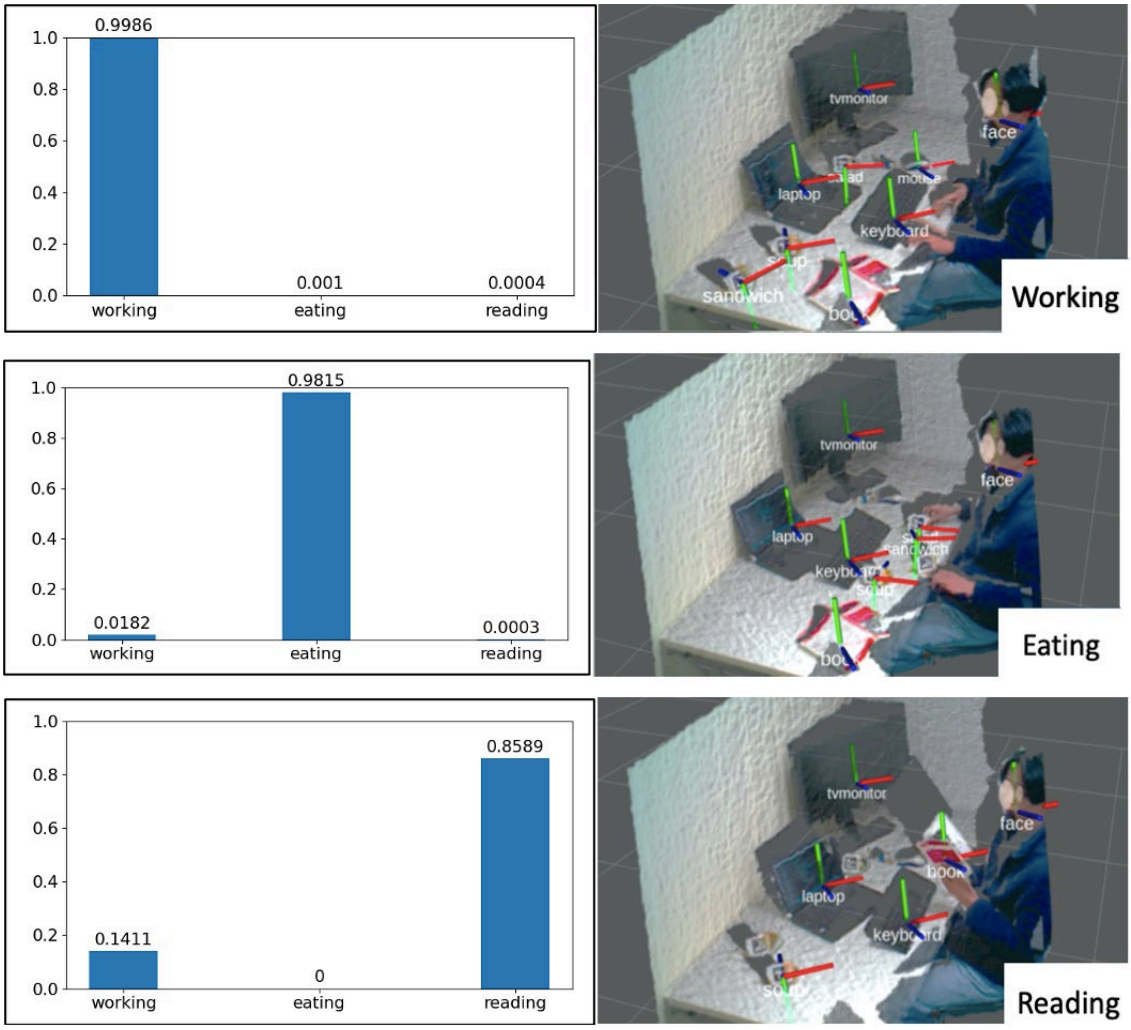

Fig. 7. Inferencing test by using HSR camera

4. Conclusion

In this paper, a human activity prediction using a GNN with ECC model was proposed based on the concept of human-object interactions from pragmatic research on human-robot communication. To obtain a homecare robot that can help with human activities, the proposed model can teach a robot to recognize human-specific activities using utterances and object information. It was also shown that the proposed system can recognize the tasks for each specified activity and detect the object related to the selected activity. From the viewpoint of the graph with ECC layers, it was assumed that the human activity consisted of the positional relationship with objects around the target human and the meaning of those objects. The object position was obtained by combining object recognition and face detection, and it was converted into node embedding using a distributed representation of words. ECC layers was used to construct the graph representation for the prediction model. Since human prediction requires instruction, the system was configured with human speech instructions to start the data collection and

training. Through experiments to determine the performance of the prediction model, it was confirmed that the model could achieve an accuracy of 91% in the testing stage. The proposed method was also compared with the DLA and X3D networks. The proposed method had a lower training time as only the positional dataset collected from human interactions with objects was used to extract the features of the 2D object and face detections. However, the limitation of the proposed method only used the human 3D position. To detect more complex activities and discriminate different activities with a similar object, we need to use the whole human body movement dataset, such as the skeleton data. For example, the robot detects a human holding a book and standing in front of a bookshelf. This activity can be classified as reading or arranging the books on a bookshelf in our proposed method. Therefore, the skeleton dataset representing a particular activity with a similar object should be considered to discriminate the different activities with a similar object in future work. Thus, the model can generate the robot task based on the current situation and environment.

Acknowledgment

The main author would like to thank the Japan Ministry of Education, Culture, Sports, Science and Technology (MEXT) for providing the financial support through its scholarship. This work was partially supported by the Japan Science and Technology Agency (JST), Moonshot R&D, with grant number JPMJMS2034.

Declarations

Author contribution. All authors contributed equally to this paper.
Funding statement. This work was partially supported by the Japan Science and Technology Agency (JST), Moonshot R&D, with grant number JPMJMS2034.
Conflict of interest. The authors declare that there is no conflict of interest.
Additional information. No additional information is available for this paper.

Data and Software Availability Statements

The applications were developed using the Python 3 and Ubuntu 20.04 operating system. Open-source frameworks, including Robot Operating System (ROS) Noetic Version, YOLOv3, Pytorch Geometric, and MediaPipe were used for tracking the human face. The Toyota HSR and their software package was used in this study. This software can only be used by the Toyota HSR developer. Our codes are also provided for the public at https://github.com/Nao-Y1996/gnn_state_recognition.

References

[1] M. Yani, A. R. A. Besari, N. Yamada, and N. Kubota, "Ecological-Inspired System Design for Safety Manipulation Strategy in Home-care Robot," 2020. doi: 10.1109/CcS49175.2020.9231354.

[2] H. Riaz, A. Terra, K. Raizer, R. Inam, and A. Hata, "Scene Understanding for Safety Analysis in Human-Robot Collaborative Operations," *2020 6th Int. Conf. Control. Autom. Robot. ICCAR 2020*, pp. 722–731, 2020, doi: 10.1109/ICCAR49639.2020.9108083.

[3] A. Carolina and H. Silva, "Scene Understanding for Autonomous Robots Operating in Indoor Environments by," 2021. Available at : E-archivo.

[4] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1036–1043, 2011, doi: 10.1109/ICCV.2011.6126349.

[5] S. Wan, L. Qi, X. Xu, C. Tong, and Z. Gu, "Deep Learning Models for Real-time Human Activity Recognition with Smartphones," *Mob. Networks Appl.*, vol. 25, no. 2, pp. 743–755, 2020, doi: 10.1007/s11036-019-01445-x.

[6] K. Li, J. Wu, X. Zhao, and M. Tan, "Real-Time Human-Robot Interaction for a Service Robot Based on 3D Human Activity Recognition and Human-Mimicking Decision Mechanism," *8th Annu. IEEE Int. Conf. Cyber Technol. Autom. Control Intell. Syst. CYBER 2018*, pp. 498–503, 2019, doi: 10.1109/CYBER.2018.8688272.

[7] M. Latah, "Human action recognition using support vector machines and 3D convolutional neural networks," *Int. J. Adv. Intell. Informatics*, vol. 3, no. 1, pp. 47–55, 2017, doi: 10.26555/ijain.v3i1.89.

[8] Y. A. Andrade-Ambriz, S. Ledesma, M. A. Ibarra-Manzano, M. I. Oros-Flores, and D. L. Almanza-Ojeda, "Human activity recognition using temporal convolutional neural network architecture," *Expert Syst. Appl.*, vol. 191, no. March 2021, p. 116287, 2022, doi: 10.1016/j.eswa.2021.116287.

[9] A. R. A. Besari, A. A. Saputra, W. H. Chin, Kurnianingsih, and N. Kubota, "Finger Joint Angle Estimation With Visual Attention for Rehabilitation Support: A Case Study of the Chopsticks Manipulation Test," *IEEE Access*, vol. 10, no. September, pp. 91316–91331, 2022, doi: 10.1109/ACCESS.2022.3201894.

[10] N. Khalid, Y. Y. Ghadi, M. Gochoo, A. Jalal, and K. Kim, "Semantic Recognition of Human-Object Interactions via Gaussian-Based Elliptical Modeling and Pixel-Level Labeling," *IEEE Access*, vol. 9, pp. 111249–111266, 2021, doi: 10.1109/ACCESS.2021.3101716.

[11] M. Hassan, P. Ghosh, J. Tesch, D. Tzionas, and M. J. Black, "Populating 3D scenes by learning human-scene interaction," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 14703–14713, 2021, doi: 10.1109/CVPR46437.2021.01447.

[12] Y. Tian, L. Chen, W. Song, Y. Sung, and S. Woo, "Dgcb-net: Dynamic graph convolutional broad network for 3d object recognition in point cloud," *Remote Sens.*, vol. 13, no. 1, pp. 1–20, 2021, doi: 10.3390/rs13010066.

[13] L. Shi, S. Li, Q. Zheng, L. Cao, L. Yang, and G. Pan, "Maximum Entropy Reinforcement Learning with Evolution Strategies," 2020, doi : 10.1109/IJCNN48605.2020.9207570.

[14] S. A. Tailor, R. De Jong, T. Azevedo, M. Mattina, and P. Maji, "Towards Efficient Point Cloud Graph Neural Networks Through Architectural Simplification," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2021-Octob, pp. 2095–2104, 2021, doi: 10.1109/ICCVW54120.2021.00237.

[15] A. Sanchez-Gonzalez, J. Godwin, T. Pfaff, R. Ying, J. Leskovec, and P. W. Battaglia, "Learning to Simulate Complex Physics with Graph Networks," , 2020, doi : 10.48550/arXiv.2002.09405.

[16] I. W. McBrearty and G. C. Beroza, "Earthquake Location and Magnitude Estimation with Graph Neural Networks," pp. 1–5, 2022, doi : 10.1109/ICIP46576.2022.9897468.

[17] P. Ruiz Puentes *et al.*, "Predicting target–ligand interactions with graph convolutional networks for interpretable pharmaceutical discovery," *Sci. Rep.*, vol. 12, no. 1, pp. 1–17, 2022, doi: 10.1038/s41598-022-12180-x.

[18] J. Xiong *et al.*, "Multi-instance learning of graph neural networks for aqueous pKa prediction," *Bioinformatics*, vol. 38, no. 3, pp. 792–798, 2022, doi: 10.1093/bioinformatics/btab714.

[19] W. Fan *et al.*, "Graph neural networks for social recommendation," *Web Conf. 2019 - Proc. World Wide Web Conf. WWW 2019*, pp. 417–426, May 2019, doi: 10.1145/3308558.3313488.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 2016. doi: 10.1109/CVPR.2016.91.

[21] C. Lugaresi *et al.*, "MediaPipe: A Framework for Building Perception Pipelines," 2019, doi : 10.48550/arXiv.1906.08172.

[22] W. Luo, W. Liu, and S. Gao, "Normal graph: Spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection," *Neurocomputing*, vol. 444, pp. 332–337, 2021, doi: 10.1016/j.neucom.2019.12.148.

[23] B. Parsa, A. Narayanan, and B. Dariush, "Spatio-Temporal Pyramid Graph Convolutions for Human Action Recognition and Postural Assessment," in *Proceedings - 2020 IEEE Winter Conference on Applications of Computer Vision, WACV 2020*, 2020, pp. 1069–1079. doi: 10.1109/WACV45572.2020.9093368.

[24] T. Ahmad, H. Mao, L. Lin, and G. Tang, "Action Recognition Using Attention-Joints Graph Convolutional Neural Networks," *IEEE Access*, vol. 8, pp. 305–313, 2020, doi: 10.1109/ACCESS.2019.2961770.

[25] S. Mekruksavanich and A. Jitpattanakul, "LSTM Networks Using Smartphone Data for Sensor-Based Human Activity Recognition in Smart Homes," *Sensors*, vol. 21, no. 5, pp. 1–25, 2021, doi: 10.3390/s21051636.

[26] M. Muaaz, A. Chelli, A. A. Abdelgawwad, A. C. Mallofré, and M. Pätzold, "WiWeHAR: Multimodal human activity recognition using Wi-Fi and wearable sensing modalities," *IEEE Access*, vol. 8, pp. 164453–164470, 2020, doi: 10.1109/ACCESS.2020.3022287.

[27] V. Dutta and T. Zielinska, "Prognosing Human Activity Using Actions Forecast and Structured Database," *IEEE Access*, vol. 8, pp. 6098–6116, 2020, doi: 10.1109/ACCESS.2020.2963933.

[28] S. Qi, W. Wang, B. Jia, J. Shen, and S. C. Zhu, "Learning human-object interactions by graph parsing neural networks," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 11213 LNCS, pp. 407–423, 2018, doi: 10.1007/978-3-030-01240-3_25.

[29] R. Morais, V. Le, S. Venkatesh, and T. Tran, "Learning asynchronous and sparse human-object interaction in videos," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 16036–16045, 2021, doi: 10.1109/CVPR46437.2021.01578.

[30] Z. Liang, J. Liu, Y. Guan, and J. Rojas, "Visual-Semantic Graph Attention Networks for Human-Object Interaction Detection," *2021 IEEE Int. Conf. Robot. Biomimetics, ROBIO 2021*, pp. 1441–1447, 2021, doi: 10.1109/ROBIO54168.2021.9739429.

[31] T. Wang, T. Yang, M. Danelljan, F. S. Khan, X. Zhang, and J. Sun, "Learning Human-Object Interaction Detection using Interaction Points," vol. 1, 2020, doi : 10.1109/CVPR42600.2020.00417.

[32] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017, vol. 2017-Janua, pp. 29–38. doi: 10.1109/CVPR.2017.11.

[33] T. Foote, "Tf: The transform library," 2013. doi: 10.1109/TePRA.2013.6556373.

[34] M. Fey and J. E. Lenssen, "FAST GRAPH REPRESENTATION LEARNING WITH PYTORCH GEOMETRIC," in *The International Conference on Learning Representations (ICLR)*, 2019, no. 1, pp. 1–9, doi : 10.48550/arXiv.1903.02428.

[35] L. Wu, P. Cui, J. Pei, and L. Zhao, *Graph Neural Networks: Foundations, Frontiers, and Applications*. 2022. doi: 10.1007/978-981-16-6054-2.

[36] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep Layer Aggregation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2403–2412. doi: 10.1109/CVPR.2018.00255.

[37] C. Feichtenhofer, "X3D: Expanding Architectures for Efficient Video Recognition," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 200–210, 2020, doi: 10.1109/CVPR42600.2020.00028.

[38] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," no. November, 2012, doi : 10.48550/arXiv.1212.0402.

[39] W. Kay *et al.*, "The Kinetics Human Action Video Dataset," 2017, doi : 10.48550/arXiv.1705.06950.