國立臺灣大學電機資訊學院電機工程學系

碩士論文

Department of Electrical Engineering

College of Electrical Engineering and Computer Science

National Taiwan University

Master Thesis

居家型機器人運用位置訊息進行人類活動與習慣理解

Household Robot Utilizing Location Information for Human Activity and Habit Understanding

林子涵

Tzu-Han Lin

指導教授：傅立成 博士

Advisor: Li-Chen Fu, Ph.D.

中華民國 112 年 2 月

February 2023

doi:10.6342/NTU202300261

國立臺灣大學碩士學位論文
口試委員會審定書
MASTER'S THESIS ACCEPTANCE CERTIFICATE
NATIONAL TAIWAN UNIVERSITY

居家型機器人運用位置訊息進行人類活動與習慣理解

Household Robot Utilizing Location Information for Human Activity and Habit Understanding

本論文係___林子涵___(姓名)__R09921108__（學號）在國立臺灣大學電機工程學系完成之碩士學位論文，於民國_112_年_1_月_12_日承下列考試委員審查通過及口試及格，特此證明。

The undersigned, appointed by the Department of Electrical Engineering on __12th__ January, 2023 have examined a Master's thesis entitled above presented by___Lin, Tzu-Han_____ (name)_____R09921108_____ (student ID) candidate and hereby certify that it is worthy of acceptance.

口試委員 Oral examination committee:

傅立成　　陳祝嵐　　杨冬海

（指導教授 Advisor）

張文中　　曾士桓

系主任 Director: 李建模

致謝

光陰似箭，轉眼間在台大這兩年的碩士即將告一個段落，並邁向人生的下一段旅程，這篇研究得以順利完成，要感謝非常多在這兩年內給予我支持、指導與協助的人。首先，我想非常感謝我的指導教授傅立成教授，教授在這兩年來的指導不只幫助我奠定了研究所需必備的基礎與能力，更啟發我在研究中的熱忱，教授對研究上的嚴謹與堅持更深深的影響我在研究中的態度。除此之外，也很感謝教授提供非常多的資源，使我們在研究中不必因為缺乏設備而感到困擾，在這兩年內，教授也提供非常多機會，從參與開會和展示研究成果給業界專業人士到擔任機器人學助理，都是研究以外難得的寶貴經驗。在尖端控制實驗室這兩年中可說是受益良多。

除了傅老師之外，我也非常感謝實驗室的同學們，大家不僅會互相鼓勵打氣與提供協助外，也會互相給予意見與指導來共同向上提升自我的研究能力。首先我想感謝和我同時進入機器人組的黃珮涵同學在研究上給予我非常多的指教與協助，同時也提出許多建設性的問題和建議使我在研究中擁有更完整的探討。感謝博班的安陞學長、曉蓓學姊與甫浩學長在開會期間總是在簡報上點出我需改進的部分。也非常感謝顥云與俊傑學長在我們碩一的時候帶領我們進入實驗室並樹立良好的學習風氣與榜樣，在離開之後也仍然時常關心著我們，並給予我們適時的建議與幫助。感謝盛安、思銓、原凱、冠毓、承軒、志淵、嘉懿、體淮、建婷在每一次的報告中提出寶貴的建議並點出重要缺失以利於我在研究上有顯著的進步與提升，平時大家也會排出時間一起聚餐來加油打氣和分享生活。感謝學弟妹祖霖、志宏、慈安、浩天、品竹、昇傑在我們研究忙不過來的時候給予幫助，並擔起實驗室重要事務的責任與各項實驗的順利運作，祝福你們接下來的時間也能找到自己喜歡的研究方向並在順利畢業前擁有一番作為。

doi:10.6342/NTU202300261

最後，我想非常感謝我的家人在這段期間對我的信任與包容，在我最忙的時候你們體諒我，在我遇到挫折的時候你們給予我心靈上的慰藉，你們無私的奉獻與支持讓我在研究的路上更能堅持到最後，並順利完成此研究，我將帶著感恩的心與你們的祝福，在我未來的人生規劃裡，繼續努力向前行。

林子涵 February, 2023 謹致於

國立台灣大學電機所尖端控制實驗室

doi:10.6342/NTU202300261

摘要

近年來技術的快速發展與提升，為機器人領域開闢了多項新的領域，許多研究大量投入居家機器人的應用及解決其可能遇到的問題，多年來，研究人員與專業人士將注意力轉向提升機器人的智慧能力，此提升不僅能提高程序的效率，更能使機器人更人性化，而居家機器人要有智慧就必須擁有理解周邊環境的能力與辨識人類行為和背後意義的能力，此外給予必要與適當的回饋與反應。

雖然在這個領域已經擁有非常多的研究，但還是存在許多須解決的問題和對人類行為理解上的挑戰。首先，對於一個居家型機器人而言，要具備有理解人類活動上的能力是至關重要的，因此，機器人須具備一個高準確度的人類活動辨識系統，除此之外，由於人類的活動與執行的地點有高度相關性，位置訊息被認為是有助於提高人類活動辨識系統的表現，因此，機器人也必須具備識別地點的能力，然而，當機器人感知影像中存在多個地點時，要獲得一個穩定且準確的位置估計技術是不容易也具挑戰性的。其次，為了理解一系列所觀察到的活動，計畫辨識扮演著至關重要的腳色，此外人類的習慣是影響序列中活動順序的重要因素，並能提高與改變計畫辨識的結果。最後，機器人與人類的互動是應用中最重要的部分，基於對於人類活動與目標的理解，機器人有必要運用這些知識和人類進行互動，並潛在的改變人類的行為與決策而使系統更有意義在。

本篇研究中，我們考慮了一個整合兩種不同估計地點模型的方法來獲得用於人類活動辨識系統的位置訊息，其中一個模型名為 ResNet50-Place365 是透過處理單一影像來估計地點，另一個是我們創建的模型，此模型利用人與物體之間的距離來決定人所在的位置。處此之外，我們提出一個基於適應性圖卷積及活動地點的人類活動辨識系統名為 AL-GCN，該系統透過輸入三維人體骨架和估計的地點來預測居家環境中人類的行為，以實現更好的準確性與強健性。我

doi:10.6342/NTU202300261

們還設計了一個計劃辨識系統來預測下一個活動、目標與計畫，為了提高計畫辨識的性能，我們透過創建一個包含人類習慣的知識庫讓系統可以將人類習慣作為辨識的資訊，該知識庫包含存儲各種活動序列的 plan library、用於預測下一個活動的 Loc-NextAct Tensor 與用於預測目標的 Loc-Objective Tensor，此外該知識庫也能夠獲得新訊息並做適當的更新來改進預測的結果與適應不同的使用者。最後，偵測到的活動與預測將作為機器人響應模組的輸入來向用戶提供建議、警告與提醒。最終我們將系統部屬到我們自己的機器人上，以執行現實世界中會發生的場景。

在實驗中，我們進行了兩種不同類型的評估，一個是對數據集做的評估，另一個則是在真實世界中對我們的機器人所做的評估。在數據集的評估中，我們所提出將兩種不同位置估計融合的方法可獲得 92.83%的準確率，而這個準確率都比僅使用其中任何一個都更準確，此外我們的 AL-GCN 模型透過結合位置訊息與人體骨架在 Cross-subject 的評估上可以達到 94.33%的準確率，而我們的計畫辨識系統透過考慮位置訊息與更新知識庫來改進預測，並可得知系統在活動序列的前幾次觀測就可以獲得較高準確的預測。在真實世界中，我們的位置估計方法在估計客廳上可以達到 98%的準確率，至於提出的 AL-GCN 模型加上位置訊息後準確率提高了 10%到 20%，最後，在執行不同的計劃時，我們提出的計畫辨識系統所作的預測會隨著知識庫的更新而有顯著的提高。

在未來研究上，我們的系統可以與 Re-ID 做結合使機器人能分辨不同的人物，並進一步使用系統所儲存的個別個人習慣進行預測，除此之外，機器人響應可以進一步考慮習慣時間上的資訊，並提供一個在時間上更精確地回應。

關鍵字：人類活動辨識、計畫辨識、地點估計、居家機器人、人類習慣、機器人響應

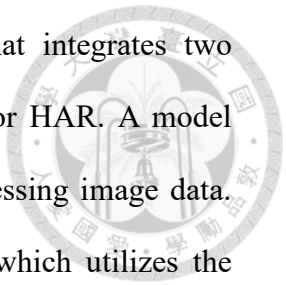
doi:10.6342/NTU202300261

ABSTRACT

The rapid development and growth of technology has widened up and opened new frontiers of the field of robotics. Many researches have devoted in investigating indoor robotic applications and the problems they may encounter. Throughout the years, researchers and professionals have turn their focus on producing robots that are becoming more and more intelligent. The increase of intelligence in robots not only allow the program to be more efficient, but also allow the robots to be more humanized. For household robots to be more intelligent, they require the ability to understand the surrounding environment and human behavior. Moreover, give necessary and suitable feedback and responses when interacting with a human.

In this field of research, there are still much to investigate and many problems to resolve. First, a household robot that possess the ability to understand human activity is crucial. Thus, a human activity recognition (HAR) system that has a high recognition accuracy is desired. Moreover, due to the high relationship between indoor activities and location, the location information is considered helpful for improving the HAR. Hence, the robot also needs to possess the ability to recognize location. However, obtaining a stable and accurate location estimating technique is challenging when the robot perceives multiple locations in an image. Second, to understand a sequence of perceived activities, plan recognition plays a vital role. Moreover, human habit is an essential element that affects the order of activities in the sequences which can enhance the plan recognition performance. Last but not least, the interaction between robot and user is a major part in robot application. Upon the knowledges from the understanding of human activity and plan recognition, the robot is necessary to utilize these knowledges to interact with humans and potentially impact human decision which makes the system more meaningful.

doi:10.6342/NTU202300261



In this study, we consider a location estimation method that integrates two different models to obtain the location information that is used for HAR. A model called ResNet50-Place365 is utilized to estimate location by processing image data. The other model is created by us called the location estimator, which utilizes the distance between human and objects to determine which location the human is in. Moreover, we propose a human activity recognition system called activity-location graph convolutional neural network (AL-GCN) based on adaptive graph convolutional network that incorporates location information to understand human behaviors. This model predicts the human activity in a household environment by taking the three-dimensional human skeleton and the estimated location as input to achieve better prediction accuracy and robustness. We also propose a plan recognition system that gives various predictions, such as next activity, objective, and plan. In order to enhance the performance of plan recognition, we take the human habit into account by creating a knowledge base that consists of a plan library, which stores various sequence of activities, a Loc-NextAct Tensor, which is used for predicting the next activity, and a Loc-Objective Tensor, which is used for predicting the objective. Moreover, the knowledge base can be updated as new information are obtained to adapt to different users and improve prediction results. Furthermore, a response module that takes the activity and predictions results from HAR and plan recognition as inputs is created to give advices, warnings, or reminders to the user. Finally, our system is physically deployed onto our own home-made robot for executing real-world scenarios.

In our experiment, two types of evaluations are conducted. One is the evaluation on the datasets, and the other is the evaluation on our physical robot in real-world scenarios. In the dataset evaluation, our proposed method that fuses ResNet50-Place365 and location estimator receives an accuracy of 92.83%, which is more

doi:10.6342/NTU202300261

accurate than only using either one of them. By incorporating the location information, our proposed AL-GCN HAR model achieves a 94.33% accuracy on the cross-subject evaluation. Moreover, the predictions of our plan recognition system are improved by updating the knowledge base and considering the location information. As a result, the predictions can achieve high accuracy in the first few observations. In the real-world experiment, our location estimation model achieves a 98% accuracy in estimating the living room. As for our AL-GCN Model, with the location information added, the accuracy of different activities gain improvement ranging from 10% to 20%. Finally, our plan recognition results in real-world show that with the updated knowledge base, the accuracy prediction increases significantly.

In Future work, our system can integrate with Re-ID to allow the robot to recognize multiple people and further apply their personal stored habit for predictions. Moreover, robot responses can take the habit time into consideration and provide a more precise response.

Keywords: Human activity recognition, plan recognition, location estimation, household robot, human habit, robot response

doi:10.6342/NTU202300261

TABLE OF CONTENTS

doi:10.6342/NTU202300261

doi:10.6342/NTU202300261

doi:10.6342/NTU202300261

doi:10.6342/NTU202300261

LIST OF FIGURES

doi:10.6342/NTU202300261

doi:10.6342/NTU202300261

doi:10.6342/NTU202300261

LIST OF TABLES

doi:10.6342/NTU202300261

Chapter 1 Introduction

In this chapter, the overall proposed concept and the system will be introduced. The contents of this chapter consist of the motivation of this work, the objective of this research, works that are related to this research, our contribution, and finally, the overview of this thesis.

1.1 Motivation

In the past several years, the rapid development of technology has driven the field of robotics and their applications to grow exponentially along the way [1]. Researches and studies regarding this field have been conducted widely and have paved the way to the technological world that we are now living in. Recalling to the early development and usage of robots that first appeared in industrial factories, the purpose of these robots is to alleviate pressure and workload as well as increase productivity since they can work in high-accuracy performance continuously for a long time without exhaustion. Nowadays, applications on robotics have expanded to a variety of fields including, food industry [2], medical industry [3], tourism industry [4], etc.

In recent years, the rise of deep/machine learning and fast improvement of hardware have made robots become more intelligent and made more sophisticated computations applicable. Eventually, smart robots made its way to home environments which became a major part of the development of home service robots. Home robots, therefore, expanded rapidly starting with the well-known application, Robot Vacuum Cleaner [5]. However, having robots in a household environment is rather challenging as they need to understand the surrounding environment and being able to navigate through it. The previous work on simultaneous localization and

doi:10.6342/NTU202300261

mapping (SLAM) such as, Gmapping [6], and navigation algorithm such as, Time Elastic Band (TEB) [7] has made it possible for robots to be successfully deployed in household environments. These kinds of works have improved the efficiency and performance of household robots, for example, Gmapping allows the Robot Vacuum Cleaner to know the full sweeping area and allow the system to apply better and more efficient sweeping algorithms.

Besides localization and navigation applications, household robots are now used for more sophisticated tasks such as, caretaking, human robot collaborating, danger detecting, or medical diagnosing. These applications require more than just understanding the information of its surroundings, but also requires to understand the human behavior in order to provide necessary services.

On that note, fields including, environmental understanding, human activity recognition, and plan recognition became major studies in household robot applications. However, many problems regarding these areas are yet to be investigated and solved. Firstly, household robots neglect the meaning of the places where it is situated in. For example, robots would know its location on the map and whether there are obstacles around it, but would not know if it is situated in a bedroom when a bed is detected, or in a kitchen if a refrigerator is observed. This information is rather important due to the fact that human activities are highly related to where it is performed in a household environment. Secondly, many human activity recognitions use activity datasets that are not relevant to the location where the activity is supposed to happen due to that the purpose of gathering data is purely interested only in the activity itself. However, in order to produce a household robot in the real-world, one needs to consider the relationship between activities and locations. Thirdly, the understanding of the human intent and plans is essential for the next step of human habit understanding and human robot collaboration. Most studies analyze HAR and

doi:10.6342/NTU202300261

plan recognition separately which is a pity since plan recognition is dependent on what activities are observed. Plan recognition investigations tend to focus on a sequence of actions of one activity in a single location. For example, [8] investigates the actions and goal for cooking in the kitchen. Seldom has plan recognition investigated a longer term of activity sequence that involves various locations which reflects more strongly in human's daily behavior. In addition to considering location information, human habit is also an important factor that is rarely considered, but highly essential and influential towards plan recognition. Every individual has his/her own preferred sequence of activities he/she would execute when trying to reach specific objectives. These habits may alter the outcome of the plan recognition which can target at different individuals. Lastly, the observed and predicted outcome from HAR and plan recognition are essential knowledge that can be utilized for robotic application purposes which is rarely being performed.

This work aims to focus on resolving these difficulties that allow the robot to better perform and understand human activities and their objectives. Finally, a system is developed and employed onto a robot with real-world scenarios that has the ability to give a better solution and robot response.

1.2 Research objectives

The present thesis attempts to build a human activity recognition and plan recognition system on a home-made robot, called OREO utilizing graph convolutional network and plan recognition algorithm that incorporates the knowledge of environmental information and human habit in order to achieve the following objectives:

(1) Location estimating system for environmental understanding: Prior to understanding the activity a human is executing, the robot needs to first

doi:10.6342/NTU202300261

understand where the activity is being performed. Therefore, this thesis tries to implement a location estimating system that utilizes an RGB-D camera for human and object detection to estimate where the activity is held.

(2) Human activity recognition model considering location information: As mentioned, indoor activities are highly related to the location where they are executed in. For example, activities such as cooking and washing dishes are more relevant to kitchen rather than bedroom. The human activity recognition system receives skeletal data as inputs as well as the location information to better learn the relationship between them.

(3) Plan recognition system considering human habit and activity location: After activities are detected, they can be utilized as important information for plan recognition to predict the following activity, the final objective, and the plan, which is the full sequence of activities leading to the objective. This thesis additionally takes the location information as well as the human habit into account for better recognizing and adapting to users.

(4) Robot responding to user module: With the given output from HAR and plan recognition, the robot should give different types of responses based on what kind of outputs they get, including giving advice, warning, and reminding.

1.3 Related Works

In this section, we will go through the backgrounds and researches related to our work. Our thesis can be categorized into two major parts: Human Activity Recognition and Plan Recognition. Each will be discussed in details in the following sections.

doi:10.6342/NTU202300261

1.3.1 Human Activity Recognition (HAR)

In this Sub-section, different types of data, including acceleration data, RGB-D, and skeleton data will be introduced as well as different types of models, such as LSTM-based, CNN-based, and GCN-based models that are chosen for these data. At the end, we will discuss the method this thesis took regarding the kind of data and model adopted.

In the field of HAR, various kinds of sensors such as, accelerometer, RGB camera, and RGB-D camera are used to receive different types of data, including acceleration, image, or video data. The work in [9] uses Accelerometer, which is a wearable sensor, for detecting the movements of the user. Accelerometer is in use for receiving acceleration data of moving actions through reference angles. It is suitable for detecting behaviors such as, moving, sitting down, rising, and ascending. Figure 1-1 shows the raw acceleration data of various activities and Table 1-1 shows the activities and locations that are used. However, the input information with specifically only the x-axis, y-axis, or z-axis accelerations of the wrist seems lacking. Even though the location in which the action is taken is considered, activities with similar wrist motions may cause confusion, for example, “putting on glasses” vs. “putting on a hat”. Moreover, in real-life, we are refrained from adding additional accessories to users. Not all users can accept wearing of an additional watch or a wristwatch type of accelerometer if they have already worn one.

Table 1-1 DHA Dataset of Human Activities [9]

Activity	Abbr.	Location	Number of Raw Accelerometer Data
Office working	Ow	Office	62,711
Reading	Re	Office	36,976
Writing	Wr	Office	27,677
Taking a rest	Tr	Office	31,265
Playing a game	Pg	Office	51,906
Eating	Ea	Kitchen	46,155
Cooking	Co	Kitchen	10,563
Washing dishes	Wd	Kitchen	10,712
Walking	Wa	Outdoors	25,768
Running	Ru	Outdoors	6,452
Taking a transport	Tt	Outdoors	28,483

doi:10.6342/NTU202300261

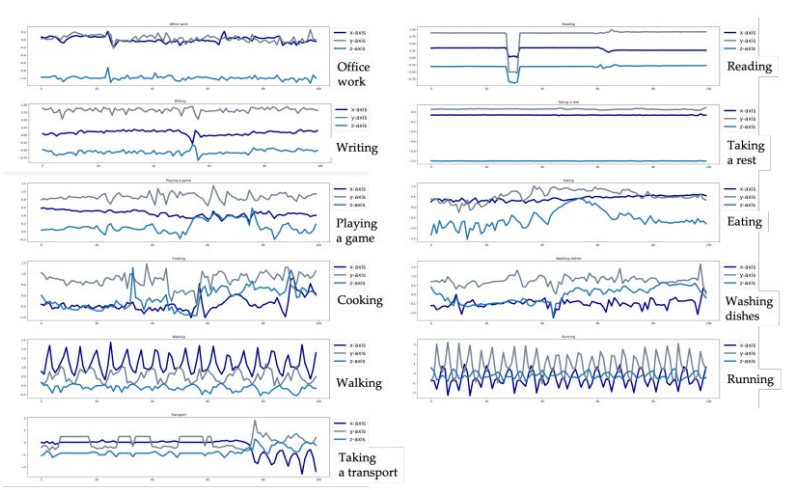


Figure 1-1 Line plot of samples from DHA dataset [9]

Image and video input data are utilized more nowadays for solving HAR problems. The major advantages are the rich information that can be extracted and learned with deep learning method, as well as preventing the users from wearing additional devices. The major difference between image and video data is the temporal information that is available. The work in [10] uses still images as input and classic CNN model for feature extraction. However, by using only still images for training can be challenging due to insufficient learning features. Researches in [11], [12], on the other hand, utilize video data as inputs which consist of temporal knowledge. Thus, [12] utilizes 3D CNN architecture that can extract and learn temporal information, see Figure 1-2.

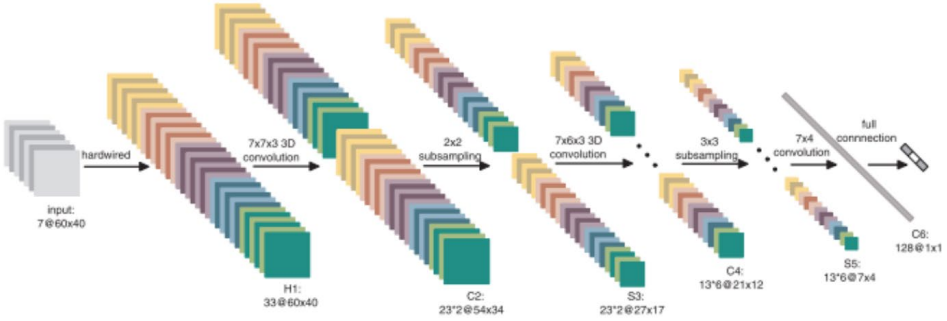


Figure 1-2 3D CNN architecture for activity recognition [12]

doi:10.6342/NTU202300261

HAR based on Skeleton Data

Aside from using RGB-D videos as input data, in recent years, skeletal data has become more and more popular in HAR. The main reason for this is due to the benefit of requiring decent or better accuracy with small data size that only contains information regarding skeleton joints.

Many models have been developed in recent years, including LSTM-based, CNN-based, and GCN-based models. Each of them is modified for tackling different characteristics of datasets to better improve the recognition accuracy of HAR. LSTM-based models are suitable for modeling temporal dynamics. Researches in [13], [14], [15] use skeleton as vectors that is formed by the concatenation of 3D human joint coordinates. In the work of [14], the author separates their system into two small subnetworks as shown in Figure 1-3. The first subnetwork is called "View Adaptation Subnetwork" that automatically determines the observation viewpoint. The following subnetwork is the "Main LSTM Network" which learns the temporal dynamics and performs the feature extractions from the skeleton data for activity recognition. The work in [15] combines GCN with Bi-LSTM to model human motion dynamics and learns past and future contexts. The author believes that future information is as important as past knowledge that can be useful in predicting future motion. Figure 1-4 shows the architecture of the system used in [15].

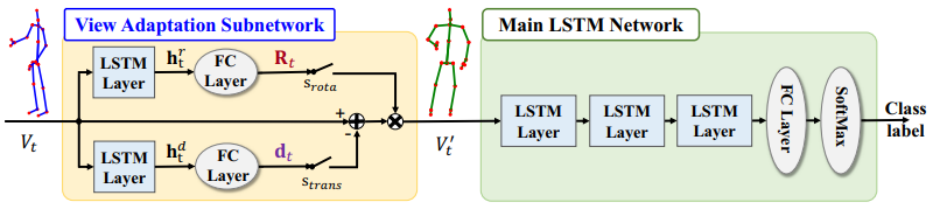


Figure 1-3 Architecture of end-to-end view adaptive RNN [14]

doi:10.6342/NTU202300261

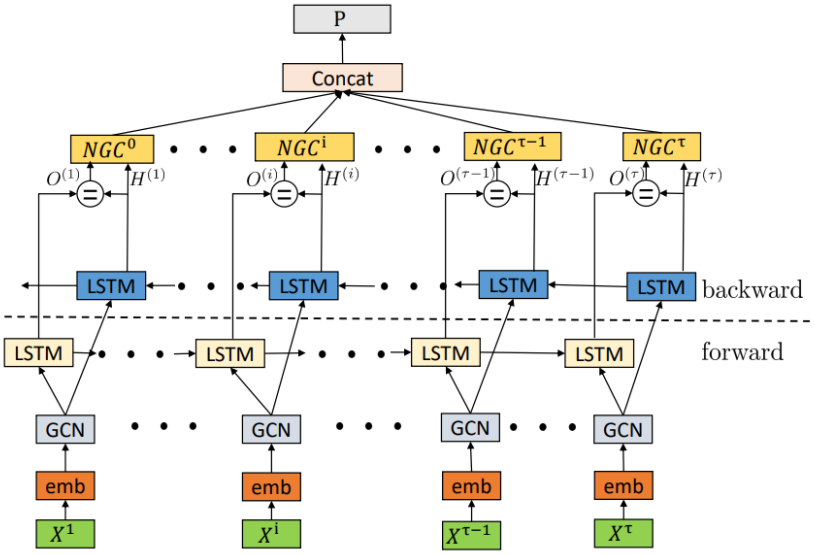


Figure 1-4 Architecture of GCN-LSTM encoder module [15]

CNN-based models are more commonly used for HAR since they receive both images from videos or skeleton data as inputs [17] [18]. However, the feature of grid structure in CNN models makes it suitable for image data rather than skeleton data. To process skeleton data through CNN models, they are transformed into pseudo-images by re-organizing joint coordinates into a 3D map with joint information along the horizontal axis and temporal information along the vertical axis, see Figure 1-5. The work in [16], [17] utilize CNN model to learn spatial and temporal information. Figure 1-6 shows the architecture of the proposed method in [17]. The author transforms skeleton sequence into a collection of 3 clips each consists of several images for corresponding 3 channels. Then CNN model and temporal mean pooling are used to extract a compact representation of each frame of the clip resulting in feature vectors. These feature vectors are fed into their proposed multi-task learning network (MTLN) for activity recognition.

doi:10.6342/NTU202300261

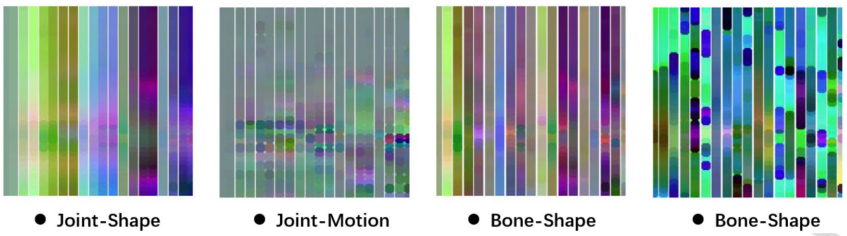


Figure 1-5 Pseudo image of skeleton data [16]

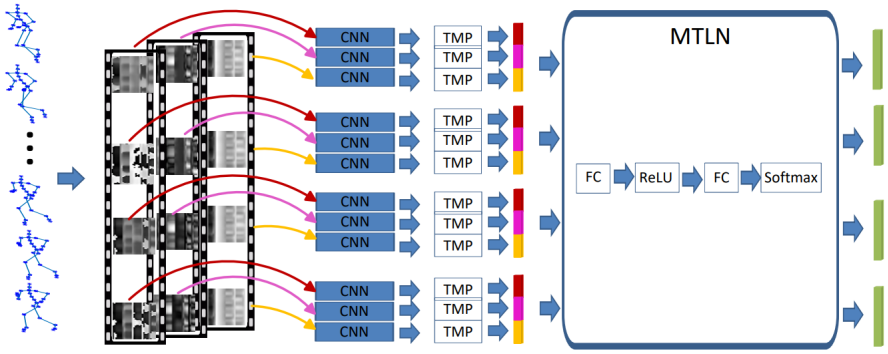


Figure 1-6 Architecture of proposed system [17]

In recent years, GCN-based models are more preferred when it comes to using skeleton data as input [19], [20], [21], [22]. The main reason is that GCN models does not suffer from grid structure restriction, but can be in arbitrary structures, as shown in Figure 1-7, which is similar to skeleton structure. The research in [19] focuses on enhancing the robustness of activity recognition models that encounter incomplete skeletons. Thus, the author proposes a multi-stream GCN for exploring sufficient discriminative feature distributed over all skeleton joints. Each stream is responsible for learning features from inactivated joints. However, the increase of streams in the processing procedure will increase the computational time which steers away from the objective to use it in real-world scenarios. The work in [20], on the other hand applies temporal attention to learn informative frames. The work shows that even without the full frames in a video being processed, the HAR accuracy maintains high. The decrease in temporal frames allow the computational time to be more efficient which

doi:10.6342/NTU202300261

benefits in real-world applications. Researches in [22] propose a multi-stream interaction network (MSIN) that consist of 3 streams that process skeleton joints, surrounding objects, and both together, respectively, see Figure 1-8. The first stream is a classical HAR process with skeleton data as input and GCN as the processing model. This work also considers the objects that surrounds the observed human along with the interaction between them. Therefore, the dynamic appearance of the object through time is important information that this work also considers and is processed through the second stream. The last stream is similar to the first with a minor difference of adding object, which is seen as a joint or node, to the skeleton joints.

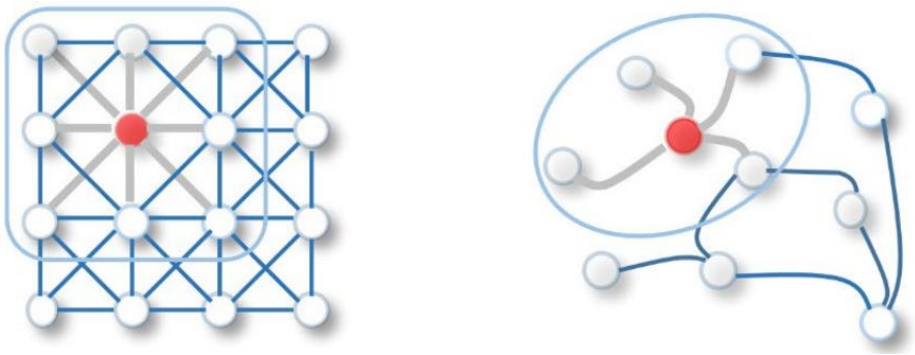


Figure 1-7 Illustration of 2D convolution (left) and Graph convolution (right) [23]

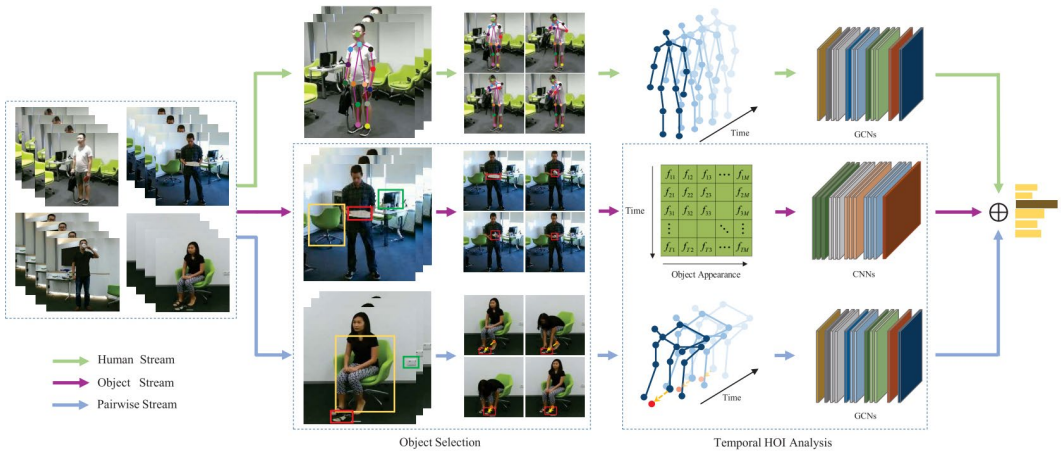


Figure 1-8 Overall structure of MSIN [22]

doi:10.6342/NTU202300261

● HAR using Location Information

Location information is usually omitted when it comes to human activity recognition. The main reason is that some dataset regarding HAR does not necessarily take location into account when they are being recorded. Therefore, activities do not have any correlation with where they were executed. However, some research, such as [9], utilizes wrist watch worn by users to gather acceleration data in human motion of different activities from various locations. DHA dataset in Table 1-1 mentioned in the previous Sub-section shows the activities that are executed in different locations, for example, cooking is usually done in the kitchen rather than in the office. Figure 1-9 also depicts their architecture that the location can be viewed as their system input.

In this thesis, we utilize ETRI 3D Activity Dataset [24] which contains skeleton data that is recorded in a household environment. The location recorded includes, bedroom, living room, kitchen, bathroom, *etc.*, see Figure 1-10. The location information will be used in both HAR as well as Plan Recognition.

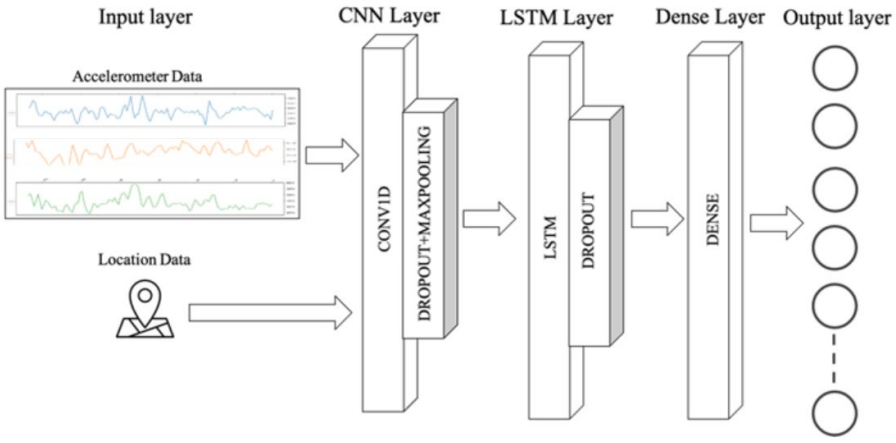


Figure 1-9 Architecture of location-based CNN-LSTM model [9]

doi:10.6342/NTU202300261

Figure 1-10 Locations of activities recorded in ETRI [24]

As mentioned in this section, different types of data and models have been investigated for improving the accuracy of HAR. In this thesis, skeleton data are used along with GCN-based model which is suitable for arbitrary structures. Moreover, due to the relationship between activities and location in a household environment, we also take into consideration of them in our HAR model which will be introduced in details in Chapter 3.

1.3.2 Plan Recognition

Plan recognition is the understanding of the objective and the execution process of an actor that is being observed of his/her actions. The applications of plan recognition span from Artificial Intelligence (AI) to cyber-security. In Recent years, plan recognition has made its way to HAR problems. It is related to both activity recognition and goal recognition [25]. First, we will explain the definition of activity and objective recognition as well as their role in plan recognition.

doi:10.6342/NTU202300261

- Activity recognition: It is the analysis of a sequence of human movement. In plan recognition, the observation of what a human is doing consist of sequences of such activities. In this thesis, such activities come from the HAR as illustrated in Sub-section 1.3.2.
- Objective recognition: It is the analysis of the human intention given a sequence of activities performed by them. The difference between objective recognition and activity recognition is in the predictive component. The former infers figuring of the final objective of the observant, whereas the latter focuses on the activity executing at that instant.
- Plan recognition: It provides a more sophisticated solution. Objective recognition predicts the final objective of an observed activity with detailed plan ignored. On the other hand, plan recognition not only includes objective recognition, but also evaluates the sequence of observed or predicted activities leading to such objective.

● Plan Library based Plan Recognition Algorithms

Most plan recognition task includes plan library. A plan library is a compact representation of all the possible plans and objectives that an observed individual can execute and reach. These plans can be expressed as graphs or AND/OR trees, see Figure 1-11 for example, and conventionally there are, two commonly used algorithms, called Symbolic Behavior Recognition (SBR) [26] and Probabilistic Hostile Agent Task Tracker (PHATT) [27]. SBR is superior than PHATT in terms of computation time, and it is designed for fast performance in order to be integrated into a robot. Although SBR is faster and could give partial answers about the current state, these answers might not be the same in future observations. PHATT, on the other hand, is a more sophisticated algorithm and could process complex inputs and outputs.

doi:10.6342/NTU202300261

However, its computation time is somewhat exponentially slower as compared to that of SBR. Note that SBR and PHATT are considered as the baseline for many future algorithms to compare against.

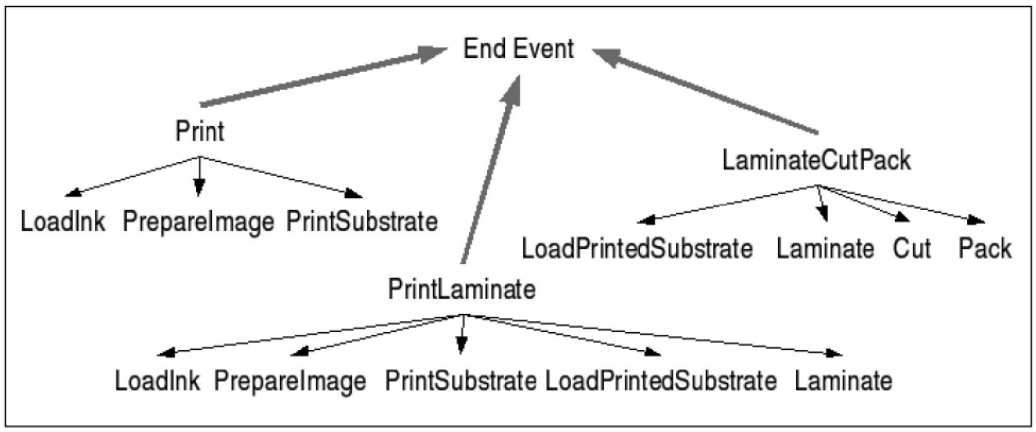


Figure 1-11 Simple plan library for packing industry [28]

The work in [29] formalizes and discusses a more efficient probabilistic plan recognition algorithm called Yet Another Probabilistic Plan Recognizer (YAPPR), which is based on weighted model counting. Model construction is rather computationally expensive. YAPPR uses string rewriting instead of other tree building methods, as compared to PHATT, has a significant reduction in algorithm's runtime, see Figure 1-12.

Order	Algorithm	Mean	Std. dev.
Unord	Yappr	15728	78.4
First	Yappr	6.6	2.31
	PHATT	19.3	2.92
Last	Yappr	20.3	7.54
	PHATT	192.7	49.1
Total	Yappr	7.40	3.77
	PHATT	40.7	60.5

Figure 1-12 Summary of run times (in milliseconds) [29]

doi:10.6342/NTU202300261

Most researches on analyzing plan recognition and activity recognition remains separate. [30]; however, these two are integrated by first analyzing the raw video data, see Figure 1-13, through a CNN architecture to locate the activities, and then they are processed by SBR to find the final goal of the video. Their plan library consists of possible objectives and plans, see Figure 1-14, based on the KSCGR dataset [31]. This thesis is inspired by this work to combine HAR with plan recognition to allow the robot system to be more intelligent in terms of understanding of the meaning behind a sequence of observed activities.

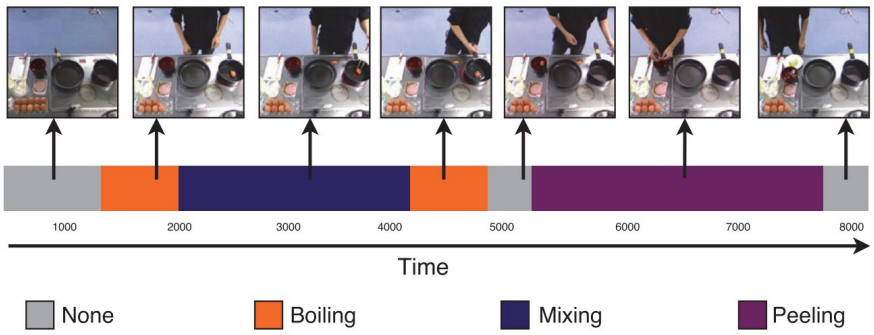


Figure 1-13 Timeline containing the frame/activity sequence for the Boiled-Egg menu [30]

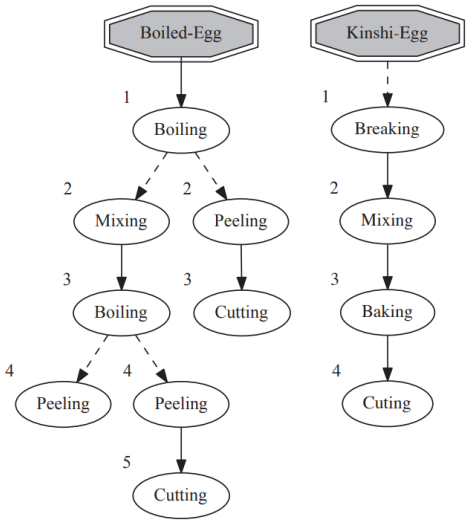


Figure 1-14 Set of plans to achieve the menus *Boiled Egg* and *Kinshi-Egg* [30]

doi:10.6342/NTU202300261

- **Plan Recognition Considering Human Habit**

Traditional plan recognition regarding activity recognition only considers a sequence of activities and goals that a human can execute and achieve. However, human factor can influence the outcome of the objective and the probabilities of plans achieving such objective. The works in [32], [33] consider human habits that can be further analyzed and explored.

The author in [32] creates a Dynamic Daily Habit Modeling (DDHM) model to learn the daily habit of elders from their daily activities. DDHM generates two-layer tree structure with the first layer specifying the predefined activities and their different time periods in the second layer, see Figure 1-5. To build this structure, start time, end time, and the duration need to be precisely identified and distinguished. Hence, by evaluating the similarity between the detected activities and modeled daily habit, a new node can be created if a similarity is lower than a certain threshold. In addition, the similarity evaluates how close the start times and end times are and how much the overlap is between the detected activity and modeled daily habit.

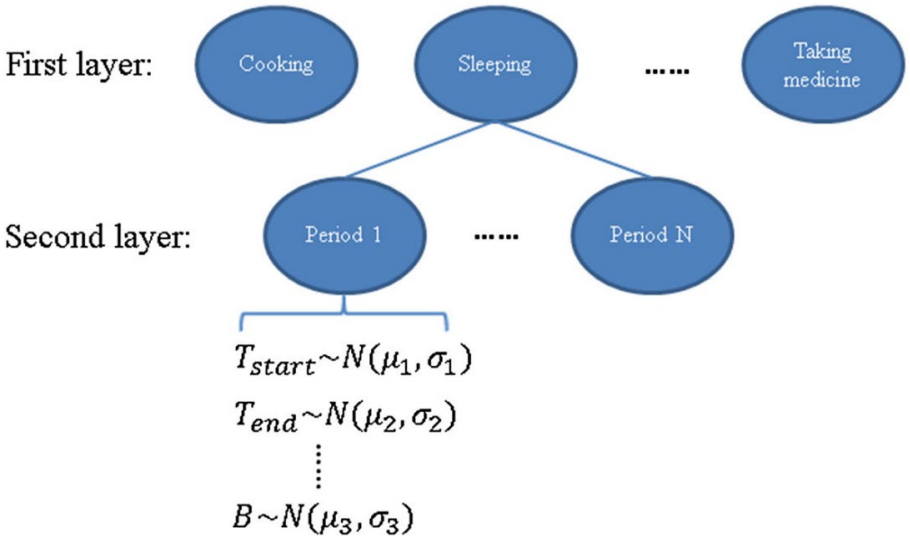


Figure 1-15 Structure of the dynamic daily habit modeling (DDHM) model [32]

doi:10.6342/NTU202300261

In the work of [33], the author takes into account the human preferences on plans. The author creates historical rating scores on the subset of plans given by humans and proposes a plan recognition approach called Human-Aware plan Recognition (HARE). See Figure 1-6 for example. This thesis is inspired by this notion of recording humans' past behavior for future prediction.

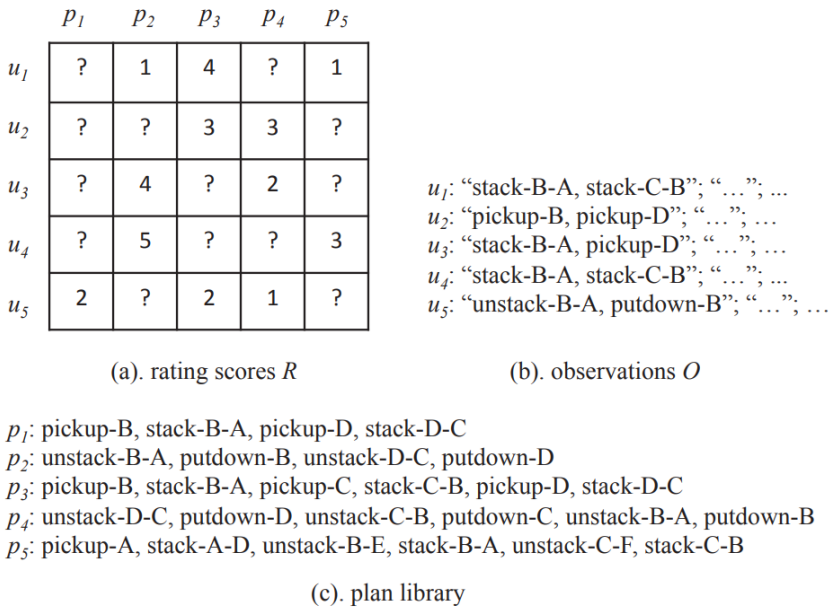


Figure 1-16 An example of home-aware plan recognition problem [33]

1.4 Contributions

In this thesis, we propose a novel activity recognition system that receives different inputs from skeleton data to estimate location from a household environment. The association between different perceptions and inputs can leverage each other to achieve a robust human activity recognition system. We also introduce a plan recognition system that updates its knowledge base to adapt to different individuals and acknowledge different habits that the user develops. Moreover, our models and systems are deployed onto a physical robot to give various responses to users. The main contributions are listed below:

doi:10.6342/NTU202300261

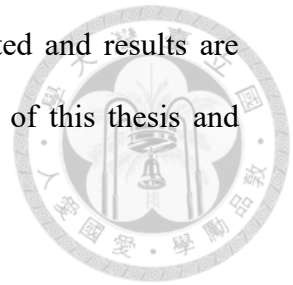
- We propose a novel graph convolutional network-based human activity recognition model called Activity-Location Graph Convolutional Network (AL-GCN) that integrates both human skeleton data in 3D space and the location where it is performed to improve and accurately predict human activities.
- We propose a plan recognition system that is able to update its knowledge base according to the observation perceived to adjust to future predictions. The system receives the activity and location detected from our proposed HAR system and location estimation module and outputs the prediction of the next activity, objective, and plan. Moreover, the system can identify habits of different users and stores them into a habit knowledge base.
- The human activity recognition and plan recognition are integrated with a robot response module that is able to provide responses to users. The response module consists of various responses that is dependent on what kind of outputs it receives from our HAR system and plan recognition system.

1.5 Thesis Overview

This thesis is organized as follows: Chapter 1 provides the introduction which consists of the motivation, research objectives, related works, and contributions. Details on the research background and related works on two different recognitions are elaborated and discussed. In Chapter 2, the preliminary of our system are illustrated. In Chapter 3, the system overview, as well as the different functionalities, including location estimation, human activity recognition, plan recognition, and robot response module are introduced and explained in details. In this chapter, we discuss how the location information is retrieved and implemented with human skeleton data as inputs for our proposed AL-GCN model. Moreover, the impact of human habits towards plan recognition and how the robot response to various observations and

doi:10.6342/NTU202300261

predictions are elaborated. In Chapter 4, experiments are conducted and results are presented to verify our proposed systems. Finally, the conclusion of this thesis and future works are presented in Chapter 5.



doi:10.6342/NTU202300261

Chapter 2 Preliminary

In this chapter, various deep learning-related methods are introduced, including different graph convolutional network, human skeleton detection, and object detection. Moreover, knowledge, such as plan library, which is commonly used, as well as different algorithms regarding plan recognition are also illustrated. Some of these materials are utilized and highly related to the developed architecture in this thesis.

2.1 Graph Convolutional Network (GCN)

Unlike CNN mentioned in previous chapter, GCN model is not bounded by inputting grid structured data. This is useful for skeleton data which can be viewed as graphs. Hence, the basic definition and construction of a graph will be illustrated in detail in this section.

A graph is composed of sets of nodes and sets of links between the nodes. Each node is called a vertex and each link is called an edge. Each edge connects two vertices. Graph can be categorized as two types: directed and undirected graph, see Figure 2-1. A graph can also be written as:

$$G = (V, E) \tag{2-1}$$

where G, V, and E represents a graph, a set of vertices, and a set of edges respectively. To formulate a graph, first we need to identify three 2 by 2 matrices.

The first matrix is called the adjacency matrix A. This matrix records the neighboring nodes that are connected to each node and can be shown as:

$$\begin{cases} A_{ij} = 1 & if\left(v_i, v_j\right) \in E(G) \\ A_{ij} = 0 & otherwise \end{cases} \text{ [34]} \tag{2-2}$$

doi:10.6342/NTU202300261

where i is the row and j is the column of the adjacency matrix. Matrix A is always a skew-symmetric matrix. With the example of the undirected graph in Figure 2-1, the adjacency matrix can be written as:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{bmatrix} \quad (2\text{-}3)$$

The second matrix is called the degree matrix D. This matrix records the number of nodes that connects to each node and is defined as:

$$\begin{cases} D_{ij} = \deg(v_i) & if\ i = j \\ D_{ij} = 0 & otherwise \end{cases} \text{ [35]} \quad (2\text{-}4)$$

where $\deg(v_i)$ is the number of vertices that is connected to vertex i. Matrix D is always a diagonal matrix. By applying the same example above, the degree matrix D can be expressed as:

$$D = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix} \quad (2\text{-}5)$$

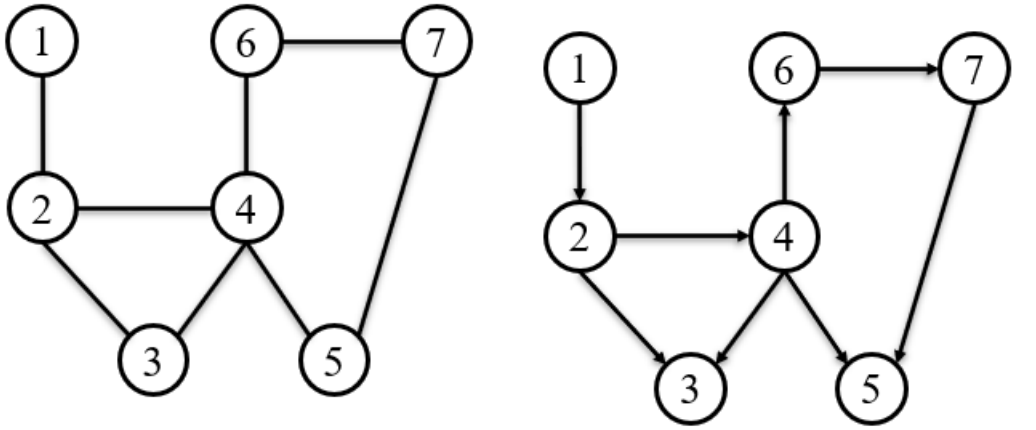
The final matrix is called the Laplacian matrix L. Unlike previous matrices, Laplacian matrix records the information of both adjacency and degree matrix and is defined as follow:

$$L = D - A \text{ [36]} \quad (2\text{-}5)$$

With the same example, the Laplacian matrix L can be written as:

$$L = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 3 & -1 & -1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 4 & -1 & -1 & 0 \\ 0 & 0 & 0 & -1 & 2 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 & 2 & -1 \\ 0 & 0 & 0 & 0 & -1 & -1 & 2 \end{bmatrix} \quad (2\text{-}6)$$

doi:10.6342/NTU202300261



(a) (b)

Figure 2-1 Undirected graph (a) and directed graph (b)

To convolute over a grid structured data is easy to comprehend. However, to convolute over an arbitrary structure, such as a graph is somewhat different. Given the Laplacian matrix, we need to find the eigenvectors of L to transform features to spectral domain. This transformation turns convolution into a multiplication computation which is easier to handle. The following equation (2-8) and (2-9) are the transformation of features:

$$\begin{bmatrix} F_1 \\ \vdots \\ F_N \end{bmatrix} = \begin{bmatrix} v_{11} & \cdots & v_{1N} \\ \vdots & \ddots & \vdots \\ v_{N1} & \cdots & v_{NN} \end{bmatrix} \begin{bmatrix} f_1 \\ \vdots \\ f_N \end{bmatrix} \tag{2-7}$$

$$F = v^T f \quad \forall v = [v_1, \cdots, v_N] \in R^{N \times N} \tag{2-8}$$

where F is the feature after transformation, $v_i = [v_{i1}, \ldots, v_{iN}]$ is the eigenvector of the Laplacian matrix L, and $f_i = [f_1, \ldots, f_N]$ is the N-dimensional feature on the vertices of the graph. Next, we can perform multiplication by multiplying equation (2-9) with h, which is the convolution function, resulting in $hv^T f$. Finally, by multiplying v, we perform the inverse transformation from spectral domain back to the original domain. The final equation can be expressed as:

$$f * h = vhv^T f \tag{2-9}$$

doi:10.6342/NTU202300261

In the early days, graph convolutional network is proposed by Bruna et al. [37] where it can be written as:

$$f_{out} = \sigma(v\ diag(\theta)\ v^T) f_{in} \tag{2-10}$$

where f_{in} represents the input feature and f_{out} represents the output feature of the graph. θ represents the kernel. Later, equation (2-11) has developed into a more stabilized expression, see equation (2-12), with the help of "renormalization trick" [38] that prevents gradient explosion and gradient vanishing problem by selecting $\tilde{A} = A + I$ and $\widetilde{D}_{ii} = \sum_j \tilde{A}_{ij}$.

$$f_{out} = \theta \widetilde{D}^{-\frac{1}{2}} \tilde{A} \widetilde{D}^{-\frac{1}{2}} f_{in} \tag{2-11}$$

Finally, equation (2-13) is derived to be the current form of graph convolution process, where W is the kernel of the convolution.

$$f_{out} = \widetilde{D}^{-\frac{1}{2}} \tilde{A} \widetilde{D}^{-\frac{1}{2}} W f_{in} \tag{2-12}$$

2.2 Spatial-Temporal Graph Convolutional Network (ST-GCN)

Traditional research on action recognition utilizes CNN models to learn from not only image data, but also skeleton data. As graph convolutional network becomes more and more powerful and efficient for processing skeleton data, Yan et, al. [39] saw the opportunity to use GCN to learn spatial information in each time stamp, along with CNN to learn temporal information. As shown in Figure 2-2, the orange box indicates a partial graph where the features of a skeleton data can be learnt by GCN. On the other hand, the green edges indicate the temporal dynamics of the same vertex where the features can be learnt by CNN. The combination of GCN and CNN for learning information from both dimensions has become a baseline model called "ST-GCN" for future GCN model comparisons.

doi:10.6342/NTU202300261

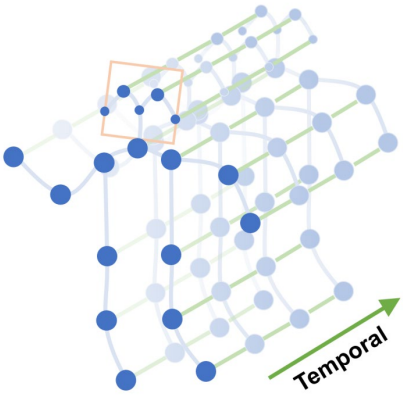


Figure 2-2 Spatial temporal graph of skeleton sequence used for ST-GCN [39]

2.3 Adaptive GCN

Even though ST-GCN mentioned in the previous section is a classic. It is not without flaws and needs improvement. GCN models are usually constructed with fixed graph structure. This confines the learning to predefined edges between nodes. However, in real world situations, this is unrealistic due the fact that graph structure may not be consistent. Shi et al. [40] proposes an adaptive structure called adaptive convolutional network. This structure is a modification of the fixed structure in order to be more flexible in processing inconsistent graphs.

Different from previous works, Shi et al. [40] rearranged the matrix operation from equation (2-13) into

$$f_{out} = WA_k f_{in} \tag{2-13}$$

by setting $A_k = \tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$. In order to allow the model to learn adjacency characteristic, two learnable matrices, B_k and C_k, are added to the formula, resulting in:

$$f_{out} = W(A_k + B_k + C_k)f_{in} \tag{2-14}$$

Matrix B_k has the same dimension as the adjacency matrix A_k. However, the difference is the elements of B_k are trainable and the graph is completely learned

doi:10.6342/NTU202300261

according to the training data. The element in the matrix can be an arbitrary value. This means that it not only indicates the existence of the connections, but also the strength of the connection.

Matrix C_k, on the other hand, is used to determine the connection between two vertexes and their strength by calculating their similarity using the following equation:

$$C_k = softmax(f_{in}^T W_\theta^T W_\phi f_{in}) \quad (2\text{-}15)$$

where W_θ and W_ϕ are the convolutions kernel that extracts the feature of each vertex. The similarity is performed by applying inner product on two output features and passes though a SoftMax function.

Instead of replacing A_k with the additional matrices, adding matrices B_k and C_k can strengthen the flexibility of the model without degrading the original one.

2.4 YOLO v4 Object Detection

Object detection has been widely known and utilized in various of applications, including human tracking, person re-identification, and human object interaction (HOI). In early days, object detections, such as [41][42][43], utilizes two-stage method where images process through region proposal, and classification and localization. Although this method outputs high accuracy, it results in slow inference rate that may not be desirable in real world scenarios. In recent years, one-stage method, for example, [44][45][46] has been created for a faster performance. In this thesis, we will be using YOLO v4 for object detect and the following are the details.

As mentioned in the previous paragraph, one-stage method is designed to have a fast performance. However, without the region proposal network, the accuracy is not as good as a two-stage method. Therefore, throughout the years, YOLO object detection has been developed and improved into multiple versions [47][48][49].

doi:10.6342/NTU202300261

Lately, version 4 has been released as an open-source package for real-time object detection application usage. YOLO v4 is an end-to-end training and inference model that inputs images and outputs the bounding box and its score. It possesses robust and fast performance that is now considered state-of-the-art model for object detection.

Figure 2-3 below shows the architecture of YOLO v4. It has four components namely, Input, Backbone, Neck, and Dense Prediction. The first component is the input which can be images, patches, or image pyramid. The second component is the Backbone, which CSPDarknet-53 [49] is utilized to improve the accuracy. The third component is the Neck, which PA-FPN [50] is chosen for performing attention mechanism. The final component is the Dense Prediction, which predicts the bounding box with the given feature maps from previous layer.

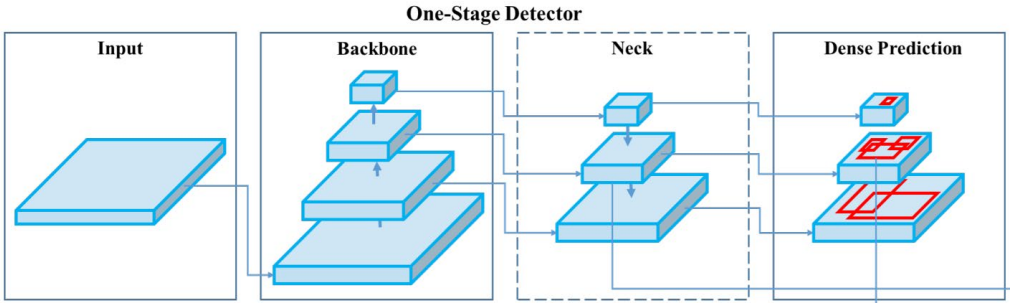


Figure 2-3 Architecture of YOLO v4 object detection [49]

Figure 2-4 shows the comparison of YOLO v4 with other object detection models in terms of mean average precision (mAP) and frames per second (fps). According to the figure, YOLO v4 runs two times faster than EfficientDet with comparable performance. It also improves significantly in terms of AP by 10% compared with YOLO v3 with decent frame rates.

doi:10.6342/NTU202300261

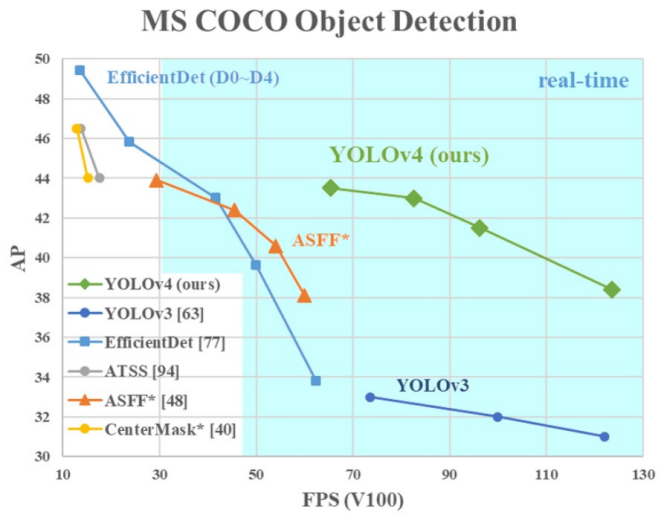


Figure 2-4 Comparison of YOLO v4 and other object detections [49]

2.5 Robot Operating System (ROS)

Robotics Operating System, so called ROS [51], is an open-source, meta-operating system for any robotic development. The difference between traditional operating systems, such as Windows, Linux, and Mac OS, and ROS is that it provides hardware abstraction, message transfer, software libraries, robotics functions, package management, to assist developers to create robot application that is simplified and easier to communicate through various platforms.

The major advantage of applying ROS to any robotic developments are the reusable and maintainable properties of ROS framework. ROS system works like a graph architecture which the processing takes place in nodes, and on the other hand, links between nodes represents the messages that are being transferred. With ROS applied, it is simple to modify or redirect a package to fit our input and output needs. In addition, it is possible to combine multiple processes with the same purposes in order to manage, redistribute, and modify more conveniently. In the ROS community, it has made possible for developers to effortlessly utilize many low-level resources and spent more precious times on their application developments.

doi:10.6342/NTU202300261

The main composition of ROS includes master, nodes, messages, topics, services, and actions. Figure 2-5 below shows the basic structure and flow of ROS. The master acts like a server that provides a list of registrations, such as IP address, port number, and topics. It also provides lookups for other computational graphs and manages the message flow between nodes. Without the master, nodes won't be able to communicate with each other. Nodes are series of data processing. Each program can be seen as a node and the data can be transferred through topics. However, in order to transfer messages in between nodes, publishers and subscribers are declared and utilized for sending and receiving data. For example, in this thesis, we created a location estimation node that is able to estimate the location a human is in by subscribing bounding boxes data from ZED2 camera. For further use of estimated location, the node will publish the location for HAR node to subscribe and use in it process. Another component in ROS is topics [52]. Topics are like buses that transfer people from one place to another, it transfers data between nodes. For a node to send messages, it needs to publish to a certain topic. On the other side, a node needs to subscribe that particular topic in order to receive such messages. The types of messages passed on a topic can vary and is user-defined.

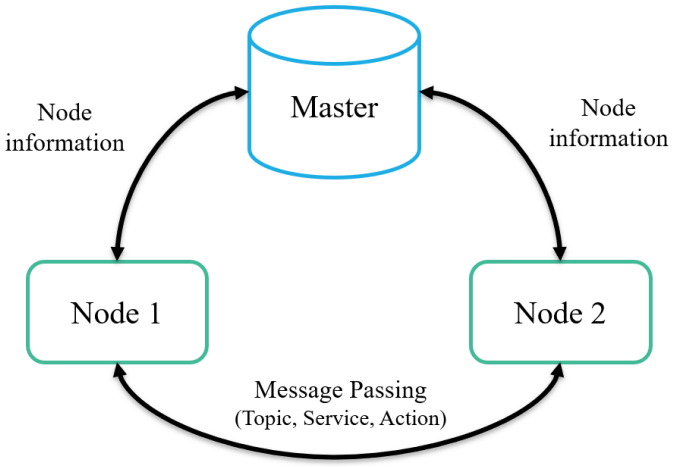


Figure 2-5 Basic communication flow of ROS between Master and Nodes

doi:10.6342/NTU202300261

2.6 3D Human Pose Estimation

3D human pose estimation has become a popular field of research that identifies the key joints in a human body. Essentially, it is a way to capture a set of coordinates for each joint, such as arm, head, torso, etc. Various kinds of sensors are used for human pose measurement, such as inertia measurement unit (IMU), RGB camera, and Depth camera. Among these sensors, RGB and Depth camera can be mounted on a robot. However, IMU needs to be worn by human, which in this case, is not desirable.

RGB and Depth sensing can be divided into two types. One is imaged-based and the other is video-based. The difference between image-based and video-based human pose detection is that the former outputs the human pose by single frame data while the latter outputs by multiple frames of data. Both methods have their strength. Video-based human pose estimation can achieve better result and lower temporal noise. Image-based human pose estimation, on the other hand, has a lower latency.

In the work of Lifting from the Deep [53], the authors proposed a new joint approach to 2D landmark detection and full 3D pose estimation from a single RGB image that takes advantage of reasoning jointly about the estimation of 2D and 3D landmark locations to improve both tasks. Figure 2-6 below shows the overall system of this work. In each stage, the model will produce a set of belief maps for location of the 2D landmarks. The belief map from each stage and the input image are used as inputs for the next stage.

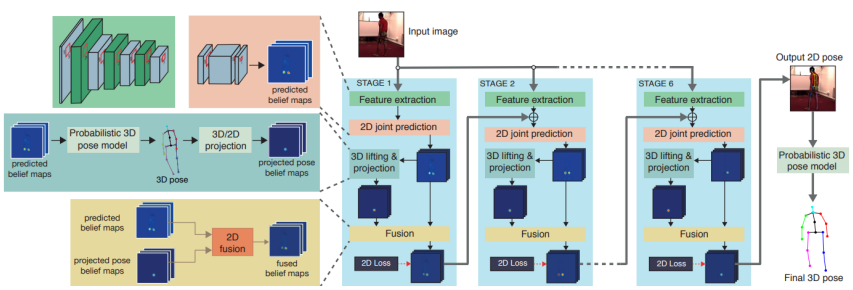


Figure 2-6 Overview of Lifting from the deep [53]

doi:10.6342/NTU202300261

Despite the mentioned work, this thesis will utilize human pose estimation model that is provided by StereoLab in the zed_ros_wrapper [54]. The main reason of this choice is due to the great compatibility between ZED2 camera and ROS. This model was trained with COCO Dataset 2017 and the skeleton joint was tracked by Extended Kalman filter. Figure 2-7 below shows the joint number of the skeleton.

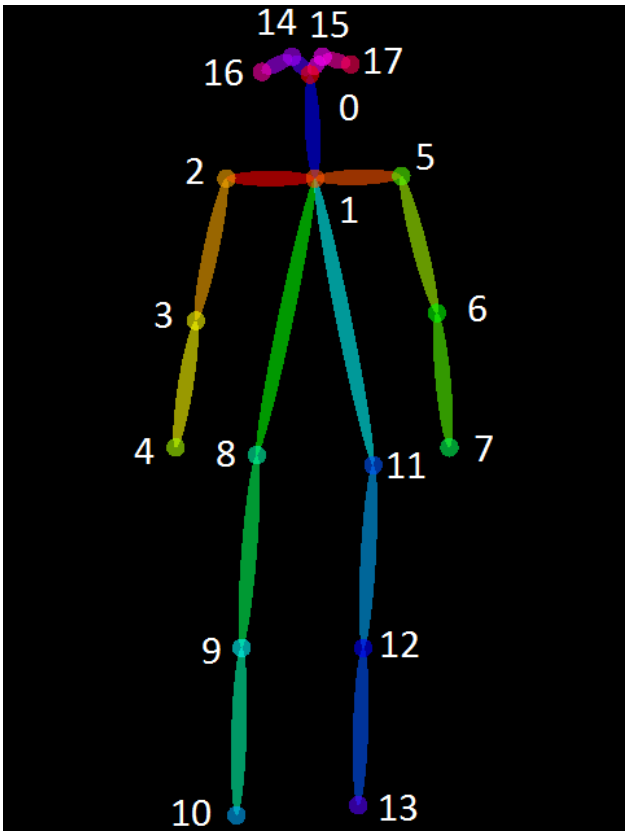


Figure 2-7 Human skeleton joint order [55]

2.7 Plan Library

In a plan recognition problem, a plan library is mostly defined and employed in plan recognition algorithms. Plan library can be designed based on different application if needed. However, the basic components in a plan library are indispensable. From [27], the plan library consists of major components including A, NT, G, and P, which will be described in the following with the assist of Figure 2-8:

doi:10.6342/NTU202300261

Definition of plan library: A plan library (PL) is a tuple PL = (A, NT, G, P) where

- A: is a finite set of terminal symbols;
- NT: is a finite set of non-terminal symbols;
- $G \subset NT$: is a set of goals;
- P: is a set of production rules in the form $\alpha \rightarrow S, \sigma$, with $\alpha \in NT$, S is a set of symbols from $A \cup NT$ and σ is a partial order of S.

A partial order σ is in the form of (i, j). This means that the i^{th} action S_i must be executed before the j^{th} action S_j in order to complete task α.

From the example in [56], we will elaborate on the content of plan library with Figure 2-8. Terminal symbols A may appear in the outputs of the production rules of a formal grammar and cannot be broken down into smaller units [57]. In a sequence of actions, terminal symbols correspond to the low-level actions. According to the figure, $A = \{$*Get teakettle, Fill with water, Get tea, Get mug, Fill mug, Get milk, Get Choco*$\}$, which cannot be broken down into smaller actions. Non-terminal symbols, on the other hand, are symbols that can be replaced and break down into smaller components. The non-terminal symbols are $NT = \{$*Boil water, tea making, Choco making*$\}$. The goal G belongs to the set of non-terminal symbols. Goal is the final intent or action of a sequence of actions. In the example, $G = \{$*Tea making, Choco making*$\}$ where there is no parent node that links to these actions. Finally, a production rule, P, is a set of rules that needs to be followed. For example, in the "Tea making" goal, boiling water, get tea, and get mug need to be performed before filling the mug. However, the three mentioned actions do not need to be executed in order. In this example, $P = \{$*Tea making* $\rightarrow$ *Boil water, Get tea, Get mug, Fill mug,* $\sigma = \{(1,4), (2,4), (3,4)\}$, $\{$*Choco making* $\rightarrow$ *Get milk, Get Choco, Get mug, Fill mug,* $\sigma = \{(1,4), (2,4), (3,4)\}$, $\{$*Boil water* $\rightarrow$ *Get teakettle, Fill with water,* $\sigma = \{(1,2)\}\}$

doi:10.6342/NTU202300261

In this thesis, our plan recognition system will be using plan library similar to this concept with additional information added. Moreover, our plan library will update itself through time to adapt to real world scenarios.

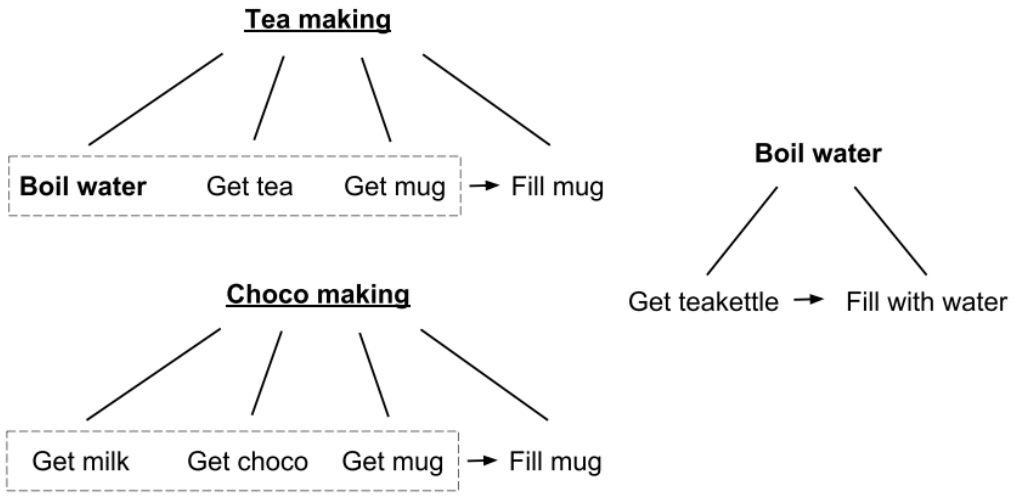


Figure 2-8 Example of a simple tree structure in plan library [56]

2.8 Particle Filter

A particle filter, also known as sequential Monte Carlo [58] method, is a set of sampling algorithms used to compute probabilities. It uses a set of particles, known as samples, to represent the posterior distribution of a stochastic process given the observations. In each iteration, particles are weighted and filtered with the given new observation and resampled to maintain the number of particles.

Particle filter is used in various field, including navigation, localization, or tracking. In this thesis, adaptive Monte Carlo Localization (AMCL) [59] is used to track the robot's pose in a map. AMCL is a probabilistic localization system that uses particle filter for robot moving in a 2D environment. Particles are sampled throughout the 2D map. With every iteration, the sensors on the robot picks up the neighboring environment features and compare them to the pre-swept 2D map. Weights are given

doi:10.6342/NTU202300261

high for those particles that have high similarities. On the other hand, weights are given low for those which are not. By filtering out the low particles, the exact pose of the robot will slowly emerge on the 2D map.

Despite the applications mentioned that utilizes particle filter, [60] managed to use particle filter on plan recognition. Figure 2-9 below shows the procedure for predicting the final goal in this work. First, the particles are initialized according to the probability of a goal that might be executed. This probability is predefined as 1/3 for tea making and 2/3 for Choco making. Next is the transition which distributes the particles evenly to the actions that need executing in each goal. Third, with the observed action, particles related to the actions are remained and the others are filtered out. Finally, the remaining particles are resampled to maintain the particles population. After that, the process iterates until a certain threshold of the probability is reached or a single goal is remained.

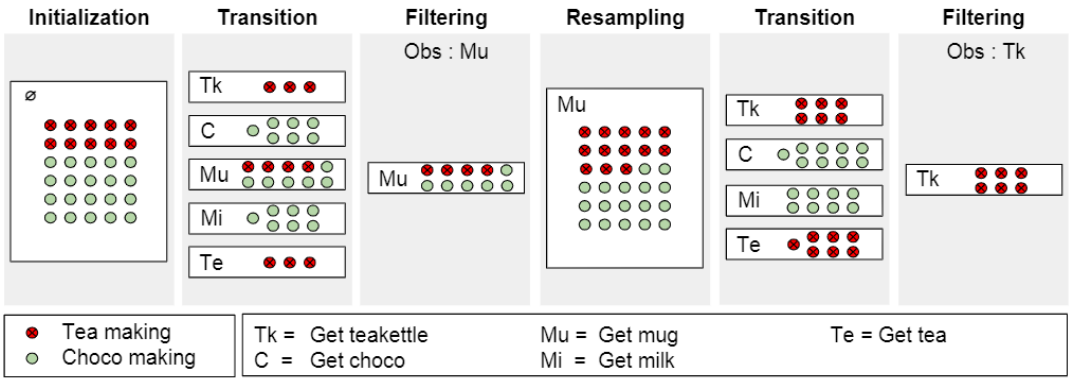


Figure 2-9 Procedure of plan recognition using particle filter [60]

doi:10.6342/NTU202300261

Chapter 3 Methodology

In this chapter, the overall method will be illustrated in detail. First, the system overview will be shown and explained, then followed by the investigation of its functionalities, including location estimation, recognition reposition, human activity recognition, plan recognition, and robot response.

3.1 System Overview

As shown in Figure 3-1, our system is composed of two major parts: “Server” and “Robot”, altogether consisting of totally six different functionalities. The “Robot” section deals with “Response Module” and “Navigation and Localization” which will be introduced in Section 3.7. On the other hand, the “Server” section processes “Location Estimation”, “Recognition Reposition”, “Human Activity Recognition”, and “Plan Recognition” which will be presented from Section 3.3 to Section 3.6 in details.

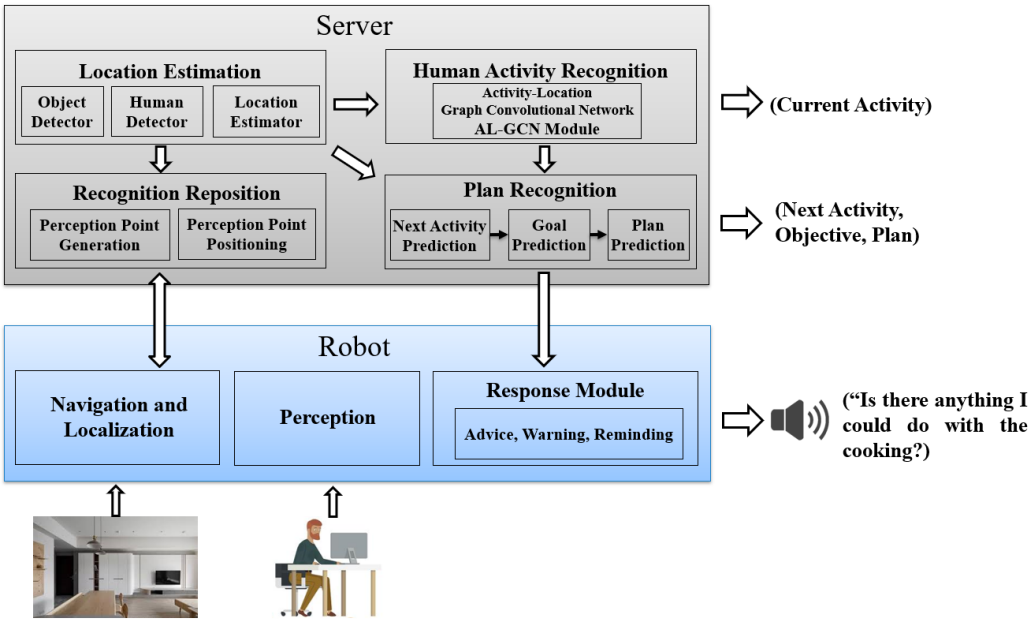


Figure 3-1 The overview of our system

doi:10.6342/NTU202300261

In order to operate this system fully, the robot needs to possess mobility and observability. “Location Estimation” and “Recognition Reposition” allow the robot to move around the environment. Meanwhile, “Human Activity Recognition” and “Plan Recognition” allow the robot to observe and understand the human activity and his/her intent.

3.2 Overview of Functionalities

As mentioned above in Section 3.1, the overall system contains several different functionalities: Location Estimation, Recognition Reposition, Human Activity Recognition, Plan Recognition, Response Module, and Navigation & Localization. Figure 3-2 shows the different functionalities and how they interact with each other. In the following Sub-sections, we will dive into the details of the first five functionalities according to this order.

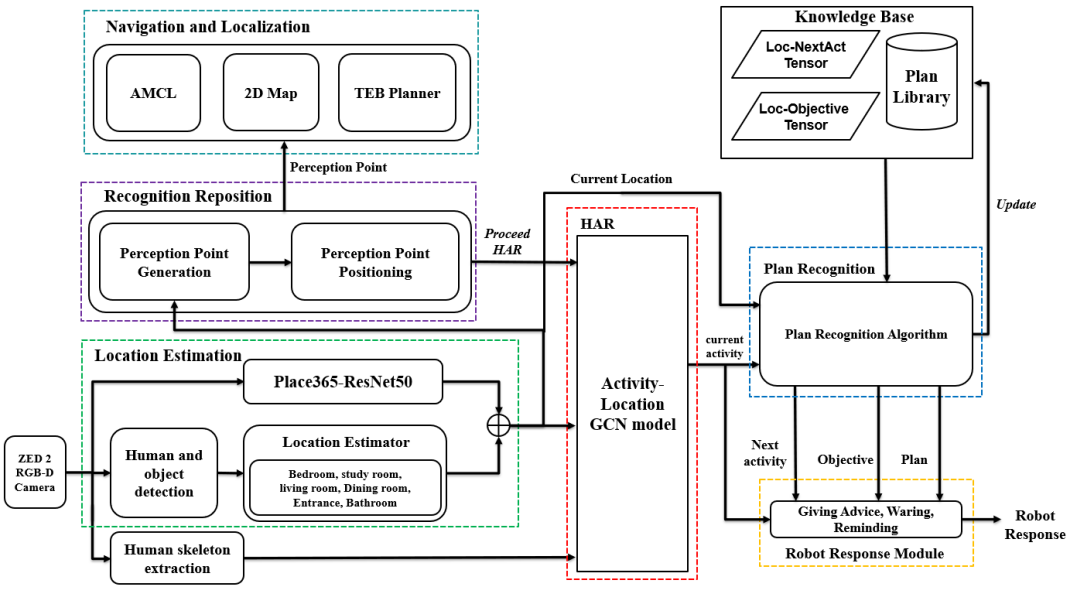


Figure 3-2 Overall flow of our system

doi:10.6342/NTU202300261

3.3 Location Estimation

In this section, the method of estimating what location a human is in will be illustrated. The definition of location and introduction of different types of objects will be discussed first, followed by the estimation method which will be elaborated in Sub-section 3.3.3.

3.3.1 Location Definition

Before activating the location estimation functionality, a few basic knowledge and definitions need to be addressed priorly. For a robot to understand a location in a household environment, it needs to understand what makes locations different. For example, what characterizes a space as a bedroom. Intuitively, one would define a location as a bedroom if one sees a bed inside within a space, or a kitchen when an oven and a refrigerator are observed. The work [61] claims that recognizing a place as a specific location is mainly based on the objects that are used as a key intermediate representation of places. Even nowadays, deep learning has developed models such as SceneNet [62] that is able to segment the objects inside an image and train models to recognize different rooms based on them.

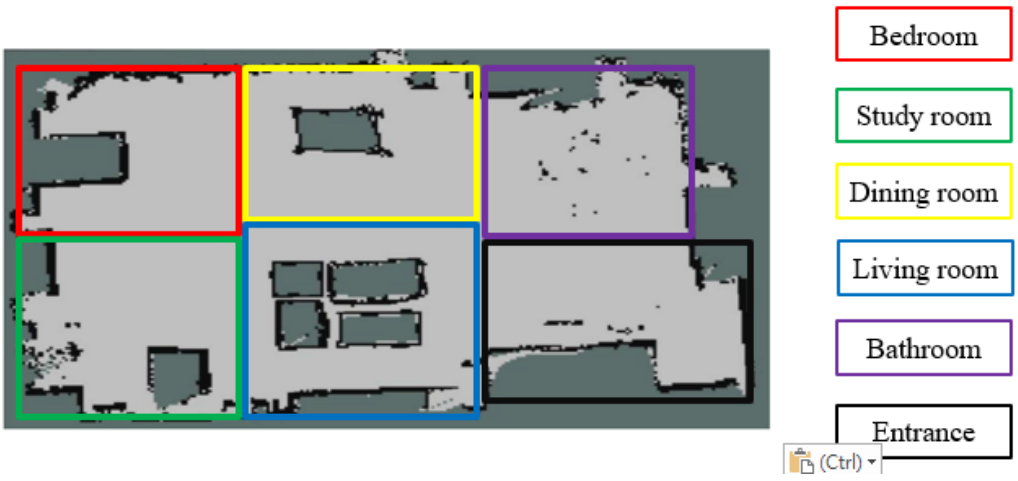


Figure 3-3 The 6 different locations in Yonglin 412 laboratory

doi:10.6342/NTU202300261

Therefore, the first step is to declare the locations where we draw perimeters on the grid map of the environment (an example in our lab is shown in Figure 3-3 for easy view). There are 6 different locations declared, namely, Bedroom, Study room, Dining room (Note that we combine dining room and kitchen), Living room, Bathroom, and Entrance. Next, we decided to use YOLO v4 object detection for detecting the objects in such locations. In the next section, we will declare which objects are special to which locations.

3.3.2 Representative Objects and Affordance Objects

As mentioned previously, the locations are defined according to the objects that are present in such locations. The researches in [63] evaluates the probability of various objects that appear in difference locations, see Figure 3-4. The author evaluates on SUN397 scene recognition benchmark [64] with 397 scene categories that consists of at least 100 images per class and 108,754 images in total. In the figure below, an example is that toilet and sink have a higher probability to appear in bathrooms. According to these probability of different objects in different locations, we can identify those objects which are more significant in representing a location.

%	Closet	Chair	Screen	Toilet	Sink	Lamp	Bed	Night Table	Desk	Drawers	Fridge	Microwave	Table	Oven	Wall Clock	Bin	Robot
Bathroom	0,075	0,025	0,006	0,365	0,453	0,006	0,006	0,006	0,006	0,006	0,006	0,006	0,006	0,006	0,006	0,006	0,006
Bedroom	0,055	0,018	0,055	0,005	0,005	0,258	0,442	0,083	0,018	0,028	0,005	0,005	0,005	0,005	0,005	0,005	0,005
Kitchen	0,099	0,043	0,012	0,003	0,154	0,003	0,003	0,003	0,003	0,056	0,142	0,154	0,056	0,259	0,003	0,003	0,003
Offices	0,039	0,284	0,265	0,003	0,003	0,045	0,003	0,003	0,239	0,058	0,003	0,003	0,013	0,003	0,006	0,026	0,003
Classroom	0,016	0,398	0,104	0,004	0,004	0,004	0,004	0,004	0,032	0,004	0,004	0,004	0,398	0,004	0,008	0,004	0,004
Laboratory	0,044	0,192	0,306	0,004	0,004	0,004	0,004	0,004	0,061	0,004	0,004	0,004	0,122	0,004	0,004	0,004	0,227

Figure 3-4 The probability of objects residing in various places [63]

In our work, we assume that each room is required with at least one major object that is not the same as any object in any other room so that the distinction can be explicit. These objects are defined as “Representative Objects”, and a few major

doi:10.6342/NTU202300261

aspects are considered when choosing them. The 1st aspect is that such objects should appear in a certain location very frequently, *e.g.*, bed, lamp, plants, teddy bear, *etc.* appear in a bedroom very commonly. Second, from the work [63], the objects' probabilities, along with their sizes and distinctiveness are taken into account as well, and a reason for that is those with larger sizes are easier to be recognized using, say, YOLO v4 detector, which explains why small objects tends to be ignored. By using the same example, the representative objects inside the living room would be "Sofa". For distinctiveness, some objects are considered definitely appearing in a specific location but not in others, *e.g.*, "Toilet" can only be present in the bathroom but no other locations.

Due to stability, the remaining small objects which are also present are considered for location estimation as well. These small objects are called "Affordance Objects". However, they remain selective because some of the small objects are too common and may appear in most locations which are pointless if considered, such as water bottles. The major difference between "Representative Objects" and "Affordance Objects" are the weights of their consideration when estimating the most possible location. Table 3-1 shows the partial list of the "Representative Objects" and "Affordance Objects" as well as their weights in 6 locations. The full list of objects in each location is shown in Appendix II.

Table 3-1 Different types of objects vs. different locations

Locations	Representative object (W_{Robj} = 10)	Affordance object (W_{Aobj}=1)
Bedroom	Bed	Backpack, Handbag, Suitcase, Teddy bear, *etc.*
Study room	Monitor (number of monitors detected)	Backpack, Handbag, Mouse, Laptop, *etc.*
Living room	Sofa, TV monitor	Remote, Potted plants, *etc.*
Dining room/kitchen	Dining table, Oven, Refrigerator	Sink, broccoli, Cake, *etc.*
Entrance	Umbrella	-
Bathroom	Toilet, Toothbrush	Sink, Cup, *etc.*

doi:10.6342/NTU202300261

3.3.3 Estimating Location

In traditional location estimation, such as the work in [72], a model-based method is used to recognize locations by inputting RGB images. In this work, we first adopt the model called ResNet50-Place365 from [72] which is trained with 10 million images of different locations from Place365 dataset. In this thesis, we only consider the 6 locations of its outcome. 10 images of each video clip from ETRI 3D Activity dataset are selected first and fed into the ResNet50-Place365 Network to get the likelihood of 6 different locations of each image. Then, the so-obtained values from the 10 images are averaged and normalized to get the final likelihood, Res, of the 6 locations about this video. An example of using ResNet50-Place365 for detecting dining room is shown in Figure 3-5.

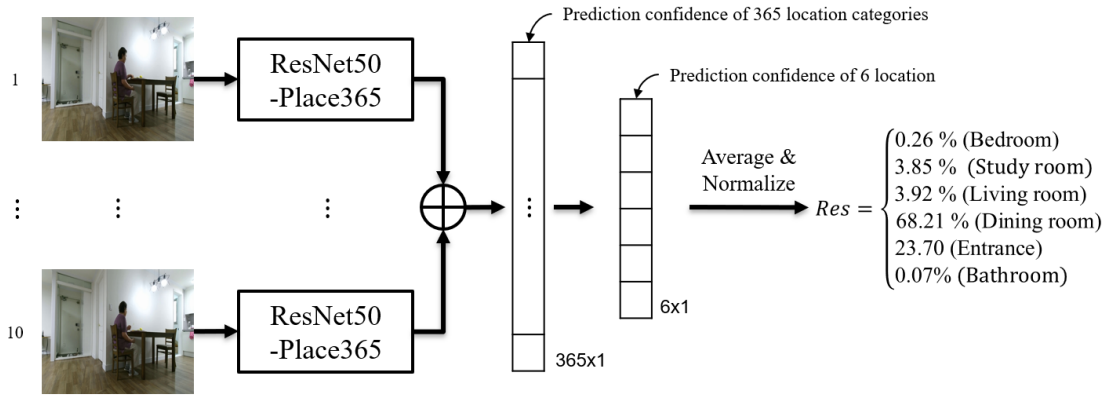


Figure 3-5 Network of using ResNet50-Place365 [72] on RGB videos

However, most models are trained with images containing only the pure scene view. It is seldom that they are trained with humans involved in the images since the original purpose is to solely recognize the location. Even so, problems, such as images containing two locations as shown in Figure 3-6, still exist. By understanding which location that the human is in in this problem, we resolve it by figuring out how close the human is to its surrounding objects. Fortunately, unlike other datasets, ETRI

doi:10.6342/NTU202300261

dataset provides the depth of the images, which we can take advantage of. We take the distance between the human and objects in 3D space into account to estimate which location the human is more likely to be in.

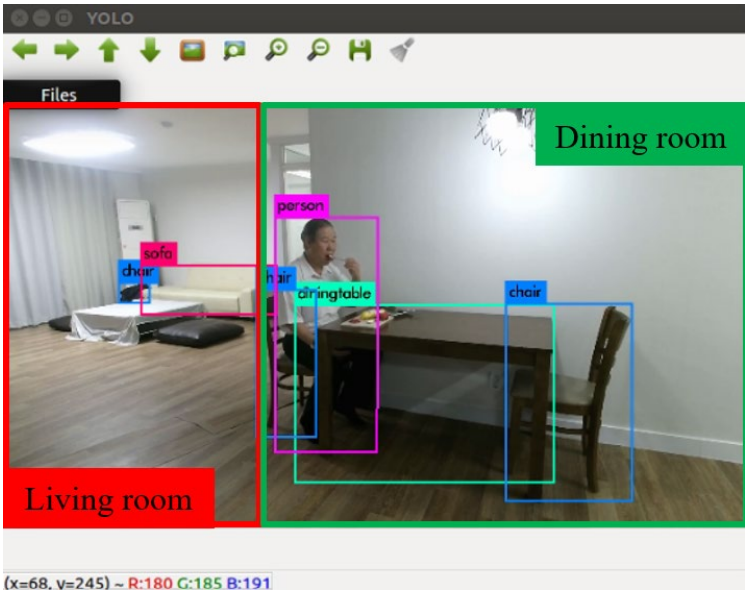


Figure 3-6 Example of two locations appearing in an image in ETRI dataset

The camera implemented is ZED2 RGB-D camera that contains depth information. The transformation from pixel coordinate to real world coordinate can be expressed as follows:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \quad (3\text{-}1)$$

where K is the given intrinsic parameter matrix, f_x and f_y are the focal lengths of the camera expressed in pixel units, and (c_x, c_y) is the principal point of the image frame, and $[u, v]^T$ indicates the pixel coordinate, and $\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}$ indicates the coordinate with respect to the camera.

Next is to calculate the distance between the human and the objects. As mentioned in the previous Sub-section, the objects are classified into “Representative

doi:10.6342/NTU202300261

Objects" and "Affordance Objects". Hence, the following equations show the distance calculation for both types of objects:

$$D_{Robj}^{I} = \sqrt{\left(x_c - x_{R\_oc}^{I}\right)^2 + \left(y_c - y_{R\_oc}^{I}\right)^2 + \left(z_c - z_{R\_oc}^{I}\right)^2} \tag{3-2}$$

$$D_{Aobj}^{J} = \sqrt{\left(x_c - x_{A\_oc}^{J}\right)^2 + \left(y_c - y_{A\_oc}^{J}\right)^2 + \left(z_c - z_{A\_oc}^{J}\right)^2} \tag{3-3}$$

where D_{Robj}^{I} and D_{Aobj}^{J} are the distances between the human and the representative object as well as that for the affordance objects, respectively, (x_c, y_c, z_c) is the center coordinate of the human, and $(x_{R\_oc}^{I}, y_{R\_oc}^{I}, z_{R\_oc}^{I})$ and $(x_{A\_oc}^{J}, y_{A\_oc}^{J}, z_{A\_oc}^{J})$ are the center coordinates of the representative object, I, and the affordance object, J, respectively.

Then, the likelihood of a person in a specific room is calculated through the following equations:

$$\begin{cases} V_k = W_{Robj} \sum_{i=1}^{N_k} \frac{1}{D_{Robj\_location}^{i} + 1} + W_{Aobj} \sum_{j=1}^{M_k} \frac{1}{D_{Aobj\_location}^{j} + 1} & if\ (N_k, M_k) \neq 0 \\ V_k = W_{Robj} \sum_{i=1}^{N_k} \frac{1}{D_{Robj\_location}^{i} + 1} & if\ N_k \neq 0,\ M_k = 0 \\ V_k = W_{Aobj} \sum_{j=1}^{M_k} \frac{1}{D_{Aobj\_location}^{j} + 1} & if\ N_k = 0,\ M_k \neq 0 \\ V_k = \sigma & otherwise \end{cases} \tag{3-4}$$

$$L_k = \frac{V_k}{\sum_{k \in location} V_k} \tag{3-5}$$

Eq. (3-4) calculates the value, V_k, for each room with the representative and

doi:10.6342/NTU202300261

affordance weights, W_{Robj} and W_{Aobj}, and distances, $D^{i}_{Robj\_location}$ and $D^{j}_{Aobj\_location}$, for different objects in the image of the observed location, N_k and M_k are the number of representative objects and affordance objects respectively of the k^{th} location appearing in the image, such as, bedroom, study room, living room, dining room, entrance, and bathroom, σ is a small value that prevents the equation from dividing zero. Eq. (3-5) calculates the likelihood for a person staying in one of the 6 rooms. Finally, the two likelihoods from the ResNet50-Place365 model and our location estimator are integrated to output the final likelihood result of the locations, see Figure 3-7. Note that in Figure 3-7, the output from ResNet50-Place365 is incorrect due to two locations detected. However, with the help of distance consideration in the location estimator, the correct location can be estimated.

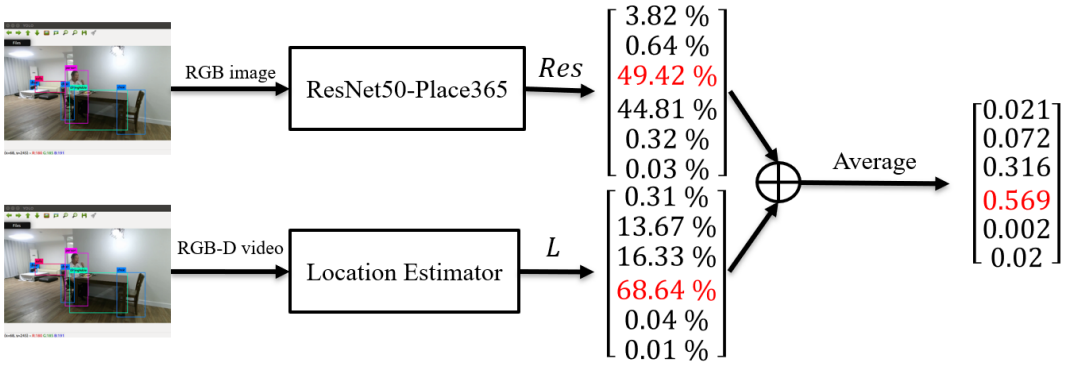


Figure 3-7 Integration of both likelihoods from two different location estimating methods

3.4 Recognition Repositioning

Continuing the last section with the estimated location, the robot would need to navigate to the location near the human and reposition itself to a good perception view angle and distance before it can proceed to human activity recognition. In this section, we will illustrate in details on how this module is processed and the method that is being utilized.

doi:10.6342/NTU202300261

3.4.1 Human Heading Detection

Before navigating to the perception point to proceed human activity recognition, we first need to figure out which position the robot must be towards the human in order to gain a good perception. Therefore, the heading of the perceived human is crucial for further robot navigation. In order to obtain the heading of a human, we utilize the human skeleton as a way to solve this problem.

From ZED 2 camera, the skeleton of a human, as shown if Figure 3-8, can be extracted and used to derive the heading after identifying crucial feature points. Out of the many joints in the skeleton, we figure out that the projection of the vector connecting the “neck” joint and the “nose” joint onto the ground plane is constantly aligned with the human’s heading. Thus, by calculating the projection of the vector in 3D space, we can infer the heading easily and effortlessly.

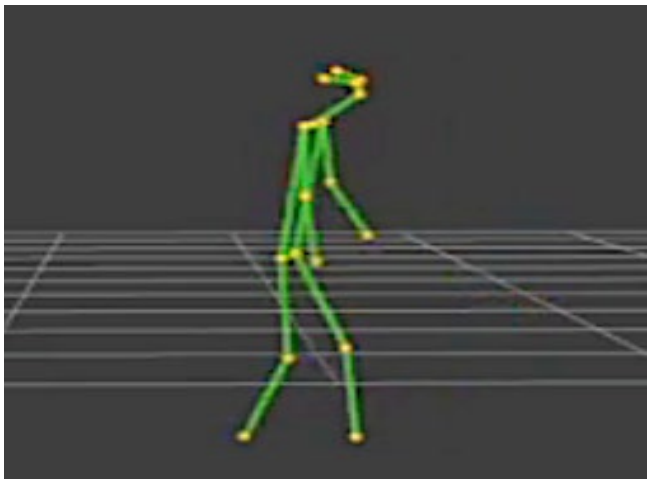


Figure 3-8 3D human skeleton data from ZED 2 camera

3.4.2 Perception Point: Distance and Section Selection

In addition to finding the heading of the human, the best angle and distance of the robot relative to the human for activity recognition is also important. Therefore, by

doi:10.6342/NTU202300261

conducting trials of recognizing different activities in different sections and distances, we are able to identify which zone surrounding the human is the best for recognition perception. Figure 3-9 below shows the activities tested vs. different distances and sections they are executed in. We conducted 10 trials for each of the 6 activities in each section at different distances using our home-made robot, OREO, in the real-world. Table 3-2 shows the overall result of the trials. It is shown that the HAR is likely to be more successful in recognizing activities in section 1 and 2 with a distance from 1.0 to 1.25 meters.

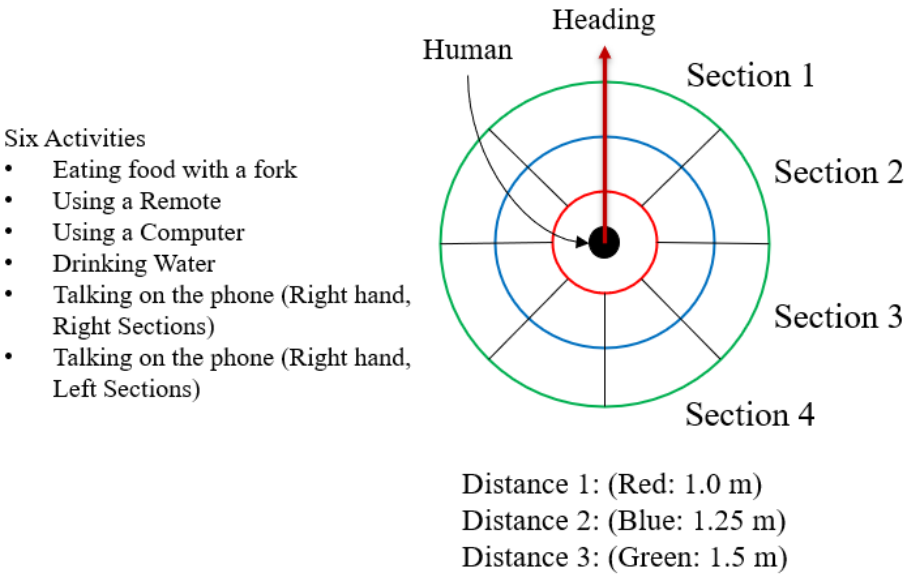


Figure 3-9 Different sections and distances tested with six activities

Table 3-2 Success rate of activity recognition in different sections and distances

	Section 1	Section 2	Section 3	Section 4
Distance 1	92%	92%	62%	14%
Distance 2	92%	90%	60%	12%
Distance 3	79%	86%	44%	10%

doi:10.6342/NTU202300261

3.4.3 Recognition Reposition Flow

After identifying the heading and the perception section and distance, the robot still needs to navigate to a certain point. Figure 3-11 below shows the steps of finding the perception spot that is best for perceiving the human's activity. First, the robot rotates to find the human. Next, the system generates 16 spots equally spaced around the human with a distance of 1.0 meters, which is shown in Figure 3-10 as purple spots. Moreover, the system will generate an additional spot that is closest to the robot on the 1.0-meter radius circle, which is the yellow spot in Figure 3-10. The purpose of the additional spot is to take into consideration of the shortest possible and available recognition spot for robot to navigate. Then, the system filters out the spots that are occupied and prioritizes the remaining spots according to Table 3-2 and the distances between the spot and robot location in a mixed way, where the latter is the shorter the better. Lastly, the robot will choose and navigate to the highest priority spot and adjust its view so that the human is in the middle of the image.

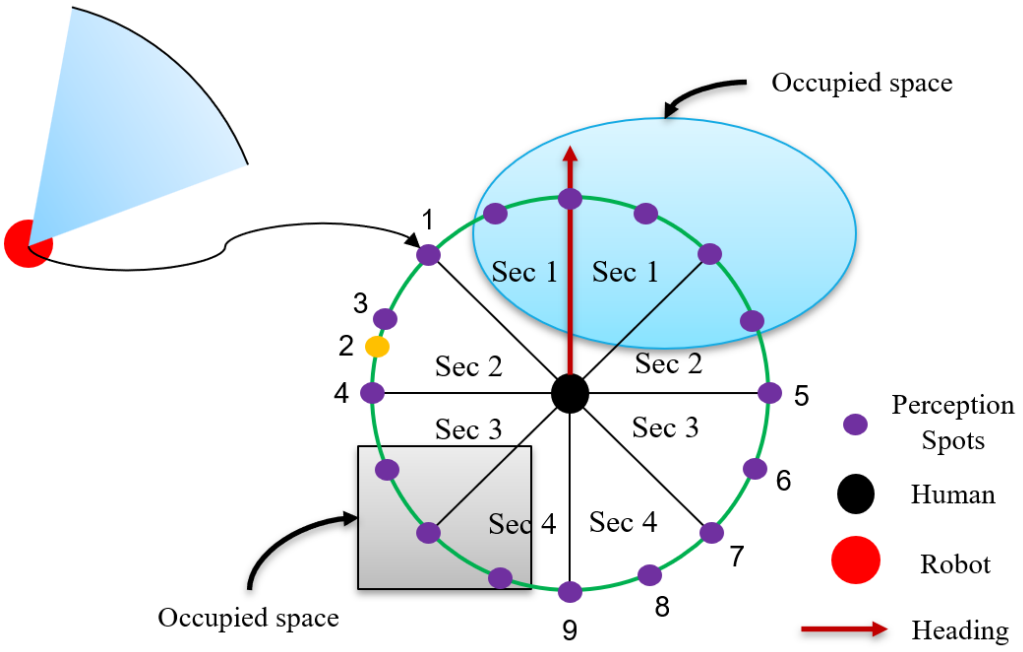


Figure 3-10 Generating and Navigating to perception spot

doi:10.6342/NTU202300261

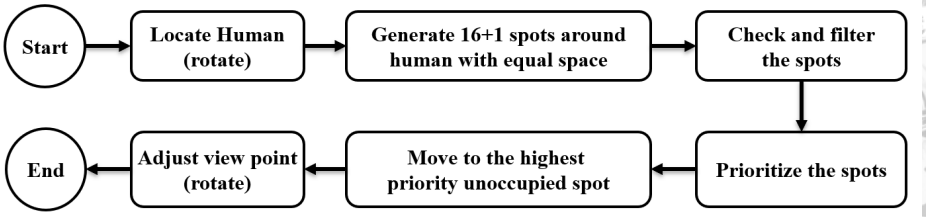


Figure 3-11 System flow of recognition reposition

3.5 Human Activity Recognition (HAR)

As soon as the robot comes to a halt after finishing adjusting its perception, it begins its process of recognizing human activity. In this section, we will proceed to the details of human activity recognition with adaptive GCN as well as our HAR model called Activity-Location GCN (AL-GCN).

3.5.1 Human Activity Recognition with Adaptive GCN

As mentioned in the preliminary, the major difference between adaptive GCN [40] and traditional GCN [39] is the trainable matrix that is added to the formula which not only is able to process arbitrary number of nodes, but also is able learn the correlation between nodes that are not predefined. From Eq. (2-15) the two matrices A_k and B_k can be written as:

$$A_k = \begin{bmatrix} 1 & \cdots & s_{A_{1,n}} \\ \vdots & 1 & \vdots \\ s_{A_{n,1}} & \cdots & 1 \end{bmatrix}_{n \times n} \tag{3-6}$$

$$B_k = \begin{bmatrix} s_{B_{1,1}} & \cdots & s_{B_{1,n}} \\ \vdots & \ddots & \vdots \\ s_{B_{n,1}} & \cdots & s_{B_{n,n}} \end{bmatrix}_{n \times n} \tag{3-7}$$

where $s_{A_{ij}}$ is the skeleton's joint-to-joint relationship defined in matrix A_k, $s_{B_{ij}}$ is the skeleton's joint-to-joint relationship that can be learned, and n is the number of skeleton joints.

doi:10.6342/NTU202300261

3.5.2 Human Activity Recognition Model (AL-GCN)

In this section, our model AL-GCN is proposed. Considering the activities executed in a household environment is highly related to different locations, for example, brushing teeth in the bathroom or watching TV in the living room, our model aims to take the location information in Section 3.3 into account to improve the recognition. In our proposed method, the model learns not only human joint-to-joint features, but also the relationship between the joints and its executed location. In this section, the architecture of the model will be shown.

First, to incorporate location information, the skeleton data and location information are inputted into a Data Pre-process block. This block inputs the skeleton data coordinate as well as the location likelihood and outputs the joint, difference, and bone that is fed into the ResGCN model [65] as f_{in} in (2-15). In this Data Pre-process block, we perform data augmentation on the skeleton data. We figure out that some activities are executed right-handed and others are left-handed which would limit the performance to one side when trained. Therefore, the data is applied with mirroring, which view the one side of the skeleton data as the other. In addition, the data is also applied with horizontal flip, which flips the image and data. Next, we multiply the likelihood that is estimated in Section 3.3 to the skeleton data to incorporate the likelihood of the joints in each location.

Originally, from [65], the dimension of the gathered input data being processed through ResGCN model [65] is 5 which consists of [I, C, V, T, M] matrices. I is the number of input branches which is the 3 branches shown in Figure 3-12 (a). C is the number of channels which is 3 for (x, y, z). V is the number of joints being processed which is 14 joints. T is the number of frames which is 400. M is the number of people which is set to 1. Through the Data Pre-process block, an additional dimension, L, which is the number of locations, is added resulting the input dimension of the

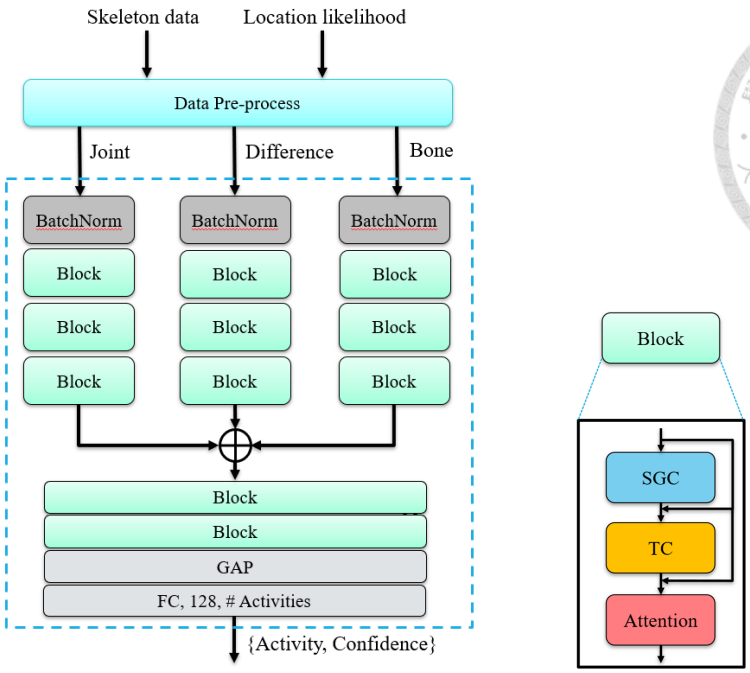
doi:10.6342/NTU202300261

ResGCN model [65] to change to [I, C, V, T, M, L].

Next, as mentioned above, we adopt ResGCN from [65] as the backbone network of our model. Figure 3-12 illustrates the architecture of the ResGCN [65]. The input of this model includes the joint, difference, and bone, where joint is the joint coordinate, difference is the joint coordinate differences compared to the previous frame, bone is the length between each joint in a single frame. This backbone incorporates the concept of residual network [66] which can prevent the network from encountering gradient explosion problem. Moreover, it utilizes spatial-temporal GCN architecture [39] which can learn spatial as well as temporal dynamic features. For more details, (b) in Figure 3-12 shows the structure of this model implementing SGC layer, which is spatial graph convolution layer, and TC layer, which is temporal convolution layer. Followed by these two layers is an attention layer. This attention layer is called Spatial Temporal Joint Attention (ST-JointAtt) which is able to distinguish the most informative joints in certain frames from the whole skeleton sequence.

However, this structure acts like the traditional GCN model that only processes static graph that only convolutes joints that are already linked by the pre-defined edges. Thus, inspired by [40], we incorporate adaptive mechanism allowing the model to be more flexible in learning features from different joints. To do so, the traditional method as shown in Eq. (2-14) is modified with the addition of B_k and C_k matrices resulting in Eq. (2-15). Finally, the model outputs the activity and its confidence which will be used in the next Section: Plan recognition.

doi:10.6342/NTU202300261



(a) (b)

Figure 3-12 The Overall structure of our AL-GCN model: (a) is the full model including the data pre-process receiving skeleton data and location likelihood, and the ResGCN [65] in the blue dotted line, (b) shows the structure inside the layers

3.6 Plan Recognition

In this section, we will investigate the plan of a sequence of observed activities and their locations. It is a major subject for a robot to understand what the objectives are behind these activities in order to give suitable and necessary responses. Moreover, human habits are a major factor to the decision of a plan to be made by that human and therefore will be considered in our plan recognition system. This section is divided into several sub-sections starting with the illustration of the system flow, followed by the definition and construction of the plan library. Then, the incorporation of human habit into our system is discussed. Followed by that is the illustration of the method and predicted outcome of our system. Finally, the update of the knowledge base is elaborated in the last section of this chapter.

doi:10.6342/NTU202300261

3.6.1 Plan Recognition System Flow

Figure 3-13 below shows the overall flow of our plan recognition system. The input of the system is the detected activity from our HAR model and the estimated location from our location estimation module. These two sources of information are being fed into the system to produce three major predictions as the outputs. The outputs include the prediction of the next activity that might be executed, the objective of the observed activities, and the final plan, which is the full sequence of activities that leads from the first observed activity to the final objective. After the whole observation is finished, the system will update the contents in the knowledge base for future use. The contents in the knowledge base will also be described and elaborated in the following sections.

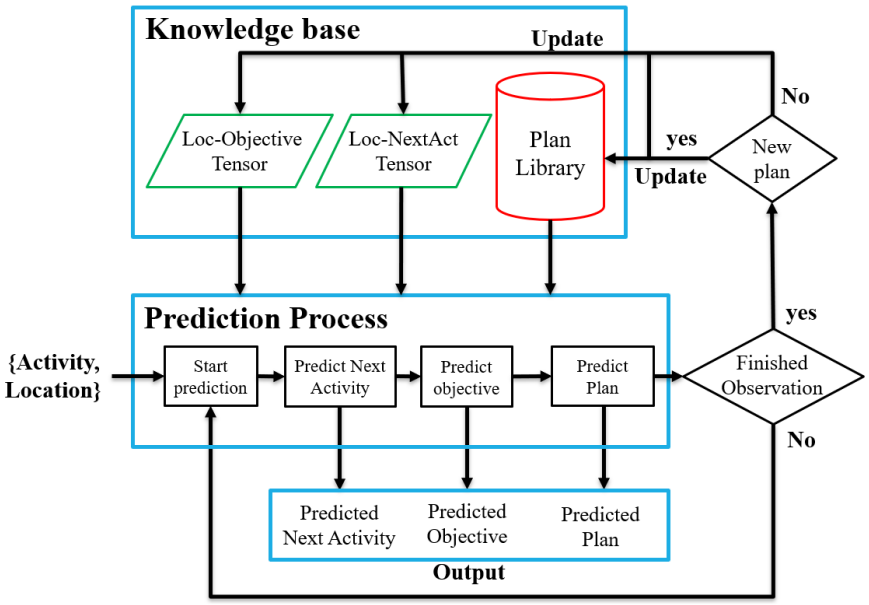


Figure 3-13 Overall flow of our plan recognition system

3.6.2 Plan Library Definition and Construction

From Section 2.7 the conventional plan library [27] is defined as a 4-tuple that consist of $\langle A, NT, O, P \rangle$, where A is a set of basic activities, NT is a set of complex

doi:10.6342/NTU202300261

activities, O is a set of objectives, P is a set of plans. To incorporate location information, in this thesis we add an additional index called L, which is a set of locations. Therefore, our plan library is defined as a 5-tuple $PL = \langle A, NT, O, L, P \rangle$.

Before constructing the plan library, we first need to define what kind of objectives that a human can achieve based on the activities in the ETRI dataset that we use. Table 3-3 below shows the pre-defined possible objectives that a human can achieve from ETRI dataset [24], which are further divided into 6 categories.

Table 3-3 Objectives that the ETRI activities can achieve

Categories	Possible Objectives
1	1) Relieve stress
2	2) Take medicine 3) Freshen up 4) Sleep 5) Rest
3	6) Going out doors 7) Meet with ppl 8) Mail letter 9) Doing makeup
4	10) Cook 11) Have meal 12) Eat fruit 13) Afterward cleaning (meal) 14) Do chores
5	15) Watch TV 16) Using computer 17) Turn on AC 18) Relax 19) Harm 20) Call for help 21) Call someone
6	22) Greet people 23) Say goodbye

The construction of the plan library is based on the common knowledge of the people. A survey was conducted asking people from different ages and sex which activities and locations they would prefer to achieve the objectives presented in Table 3-3. The detail of the survey will be illustrated in, Sub-section 4.1.2 of Chapter 4. Figure 3-14 below is an example of the partial construction of the plan library. In the example, the objectives are, “Cooking”, “Having Meal”, and “Eating Fruit”, while the rest are ETRI activities. Moreover, a sequence of activities achieving an objective is called a plan. With the same example, a sequence including “Washing hand” and “Peeling fruit” is a plan, whereas a sequence consisting “Washing hand” and “Trimming vegetable” is not a plan.

doi:10.6342/NTU202300261

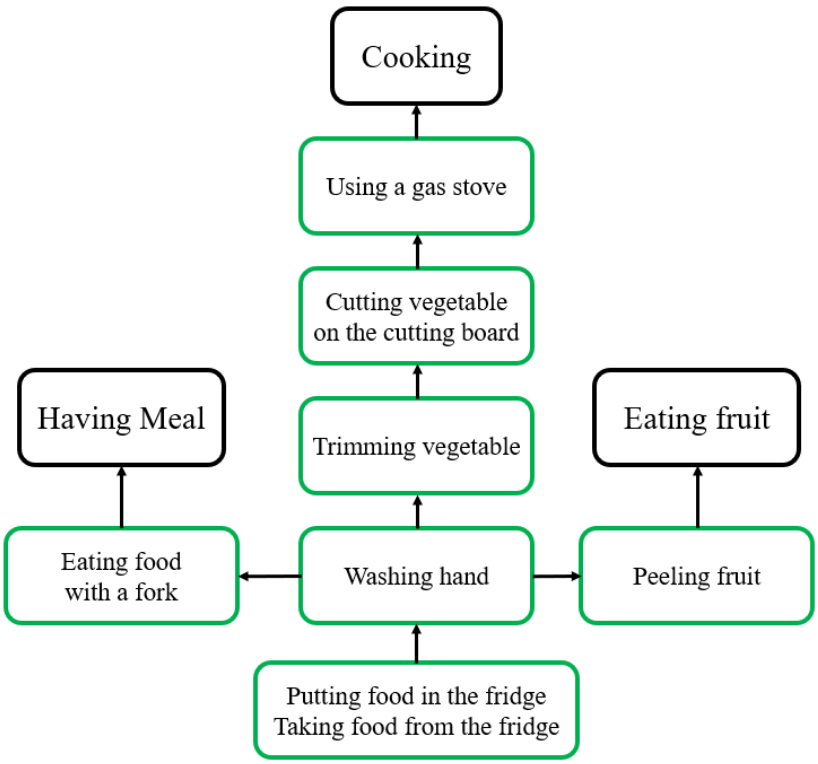


Figure 3-14 Partial tree structure of category 4 in plan library

3.6.3 Human Habit Incorporation

Human factor is considered important when it comes to analyzing plan recognition. Different individual has different habits of executing activities. In [67], the author suggested the definition of a habit as the order of a sequence of activities. [67], on the other hand, referred habit to the routines of the daily lives of humans. However, [32] used a more comprehensive way to define habit as a sequence of activity along with the duration of each activity. In this thesis, we view habit same as [73] with the addition of repetition. In other words, the same sequence of activity under an objective need to be executed repeatedly in order to be called a habit.

Individual habits can alter the order of the sequence of activities achieving an objective, which may influence the outcome of the prediction. Figure 3-15 below is an example that exhibits Jane's habit when doing chores. Although Jane may execute

doi:10.6342/NTU202300261

both “Hanging out the laundry” and “Vacuuming the floor”, recently, she has developed a habit of hanging out the laundry after she vacuums the floor. The sequence of activities will be stored in the knowledge base, called “Habit Knowledge Base” (HKB), as the habit of “Doing Chores” along with the time and duration. In this thesis, due to the experiment of our plan library method, the number of consecutive executions of the same sequence of activity under an objective to become a habit is 7. It is the number of updates in our update algorithm to finally achieve maximum prediction accuracy. Moreover, the habit can be a full plan or a sub-plan as long as they are executed consecutively for 7 times.

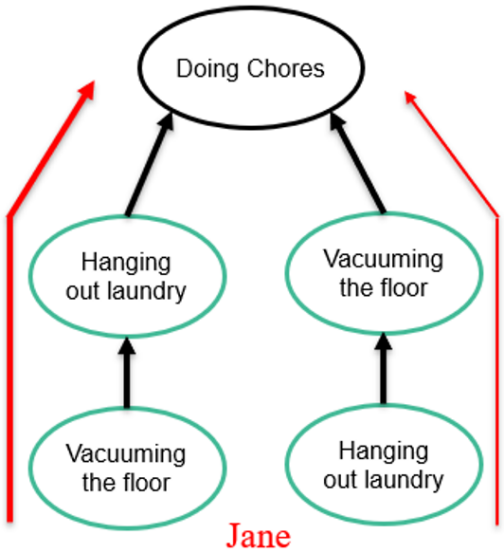


Figure 3-15 Example of an individual conducting two different orders of activities achieving an objective

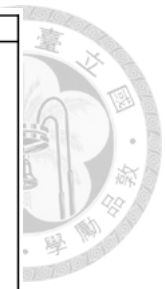
Inspired by [33] which keeps track of historical rating scores on plans given by humans, those scores reflect the preference of individuals executing certain plans. In this thesis, we utilize this method and define two tensors called, Loc-NextAct Tensor

doi:10.6342/NTU202300261

and Loc-Objective Tensor. These two tensors store utility values between 0 and 10 that can be used for the prediction of the next activity and the final objective. Moreover, these tensors are updated at the end of every sequence along with newly detected plan in the knowledge base. How the knowledge base is updated will be illustrated in Sub-section 3.6.5 in detail. Loc-Objective Tensor has the dimension of $55 \times 6 \times 23$, where 23 is number of objectives, 55 is the number of activities in the ETRI dataset, and 6 is the total number of locations. Thus, this tensor stores the relationship between each activity, location, and the final objectives. Loc-NextAct Tensor has dimension of $55 \times 6 \times 55 \times 6$, where the first 55×6 dimensions represent the indices of the current activities and their likely associated locations, whereas the next 55×6 dimensions represent the indices of the next possible activities and their likely associated locations. Hence, this tensor prescribes the implicit relationship among the current activity, current location, the next possible executed activity, and the next possible location.

Before putting the tensors into good use, they are first initialized by using the samples gathered via survey which the details will be elaborated in Chapter 4. In addition, shown below in Figure 3-16 and Figure 3-17 are the pseudo-code of both tensor initializations. For the Loc-NextAct initialization, the algorithm goes through every activity in each objective of every survey sample and locates the current observation and next observation of every plan. Then, it finds the indices of the current observation and the next observation and increment the entry with the indices of the tensor by one. On the other hand, the algorithm for Loc-Objective initialization finds the indices of the current observation and its objective and increment the entry with the indices of the tensor by one. Lastly, both tensors are normalized between 0 and 10 which results in the final initialized tensors for further use.

doi:10.6342/NTU202300261



ALGORITHM I: LOC-NEXTACT TENSOR INITIALIZATION

Input:	Survey: **Survey** – Survey of activity sequence leading to objectives. ACT: **ETRI_Activities** – List of ETRI activities. (55) LOC: **Locations** – List of Locations. (6) OBJ: **Objectives** – List of Objectives. (23) LNAT: **Loc-NextAct** Tensor.
Parameters:	S: Samples in the Survey. O: 23 Objectives in each sample. obs: Observations in each objective. obs_{curr}: Current observation in the sequence. obs_{next}: Next observation in the sequence. i: Index number of **ETRI_Activities** for current activity. j: Index number of **Locations** for current activity location. w: Index number of **ETRI_Activities** for next activity. r: Index number of **Locations** for next activity location. $a_{i,j,w,r}$: Value of Loc-NextAct Tensor.
Output:	LNAT_norm: Normalized Loc-NextAct Tensor.

```
for all S ∈ Survey do
    for all O ∈ Objectives do
        for all obs ∈ Observation do
            if obs_next == None then
                break
            else
                i ← Index(ACT(obs_curr[activity])).
                j ← Index(LOC(obs_curr[location])).
                w ← Index(ACT(obs_next[activity])).
                r ← Index(LOC(obs_next[location])).
                LNAT_{i,j,w,r} + 1 = a_{i,j,w,r}.
        end
    end
end
LNAT_norm ← normalize(LNAT). // Normalize between 0 and 10
return LNAT_norm.
```

Figure 3-16 Pseudo code of Loc-NextAct Tensor Initialization

ALGORITHM II: LOC-OBJECTIVE TENSOR INITIALIZATION

Input:	Survey: **Survey** – Survey of activity sequence leading to objectives ACT: **ETRI_Activities** – List of ETRI activities (55) LOC: **Locations** – List of Locations (6) OBJ: **Objectives** – List of Objectives (23) LOT: **Loc-Objective** Tensor
Parameters:	S: Samples in the Survey O: 23 Objectives in each sample obs: Observations in each objective i: Index number of **ETRI_Activities** j: Index number of **Locations** k: Index number of **Objectives** $o_{i,j,k}$: Value of Loc-Objective Tensor
Output:	LOT_norm: Normalized Loc-Objective Tensor

```
for all S ∈ Survey do
    for all O ∈ Objectives do
        for all obs ∈ Observation do
            i ← Index(ACT(obs[activity])).
            j ← Index(LOC(obs[location])).
            k ← Index(OBJ(O)).
            LOT_{i,j,k} + 1 = o_{i,j,k}.
        end
    end
end
LOT_norm ← normalize(LOT). // Normalize between 0 and 10
return LOT_norm.
```

Figure 3-17 Pseudo code of Loc-Objective Tensor Initialization

doi:10.6342/NTU202300261

3.6.4 Plan Predictions

After we have all the necessary components, including final constructed plan library and the two tensors in the previous section, we can begin our prediction. This system makes three predictions including next activity, objective, and plan prediction. These predictions are executed in this order. In addition, in the Plan Recognition Section of this thesis, "observation" is referred as the predicted activity from the HAR and its location from the location estimation module.

Therefore, given the current observation, which is the current activity and its location, the first step is to predict the next executing activity and its location. In the given example as shown in Figure 3-18, if the current observation is obs_8, the next activity and its location is predicted as obs_9. This comes from the Eq. (3-8), (3-9), and (3-10) as shown below. Let $Next\_Obs(w,r)$ be the possible next activity with index w and possible next location with index r, and A_{pred} is the prediction of the next observation, then

$$P\big(Next\_Obs(w,r)\big|\theta_{i,j}\big) = \left(\frac{a_{i,j,w,r}}{\sum_{n=1}^{55}\sum_{m=1}^{6} a_{i,j,n,m}}\right) \tag{3-8}$$

$$(w^*, r^*) = \underset{w,r}{argmax}\, P\big(Next\_Obs(w,r)\big|\theta_{i,j}\big) \tag{3-9}$$

$$A_{pred} = \{Next_{act},\ Next_{loc}\} = \{\boldsymbol{ETRI\_Act}[w^*], \boldsymbol{Location}[r^*]\} \tag{3-10}$$

where, $i, j, w,$ and r are the indices of the current activity, the current location, the next activity, and the next location respectively, $\theta_{i,j}$ is the current observation, $\boldsymbol{ETRI\_Act}[]$ is the list of ETRI activities, $\boldsymbol{Location}[]$ is the list of possible locations, $Next_{act}$ and $Next_{loc}$ are the predicted next activity and its location, $a_{i,j,w,r} \in [0,10]$

doi:10.6342/NTU202300261

is the utility value corresponding to the index of the next observation in the Loc-NextAct Tensor given the index (i, j) of the current observation. The higher the $a_{i,j,w,r}$ is, the more probable the next observation is with index (w, r). Eq. (3-8) is the calculation of the probability of the next activity with index w and its location with index r, while Eq. (3-10) outputs the final predicted observation.

Next step is the prediction of the objective given the initial observation, θ_{init}, and the predicted next observation, $Next\_Obs$. Eqs. (3-11), (3-12), (3-13), and (3-14) below illustrate the steps to objective prediction. Let O_{ach} be the achievable objectives given the predicted next observation, $O(k)$ is the possible objective with index k, and O_{pred} is the prediction of the objective, then

$$O_{ach} = PLsearch(Next\_Obs) \tag{3-11}$$

$$P(O(k)|\theta_{init}) = \left(\frac{o_{init,k}}{\sum_{l=1}^{23} o_{i,j,l}}\right) \tag{3-12}$$

$$k^* = \underset{k}{argmax}\, P(O(k)|(\theta_{init}, O_{ach})) \tag{3-13}$$

$$O_{pred} = \boldsymbol{Objective}[k^*] \tag{3-14}$$

where $\boldsymbol{Objective}[]$ is the list of totally 23 objectives , $PLsearch(\cdot)$ is to search the plan library given the input, $Next\_Obs$, O_{pred} is the probability of the final predicted objective, θ_{init} is the initial observation, $o_{init,k} \in [0,10]$ is the utility value corresponding to the index of the objective in the Loc-Objective Tensor. The higher the $o_{init,k}$ is, the more probable the final predicted objective is with index k. The predicted next activity is fed into our plan library to search for all achievable objectives, Eq. (3-11). Then, the probability of all possible objectives is calculated

doi:10.6342/NTU202300261

with Eq. (3-12). Finally, the predicted objective with the highest probability is the found with Eq. (3-13) and (3-14). However, there is a scenario which is shown in Figure 3-19 where the current activity is not within the next predicted observation in the previous observation, $\boldsymbol{obs_{prev}}[obs_{next}]$. In other words, obs_4 is not within the next predicted activity of obs_9, which are obs_{12} and obs_{13}. In this case, Eq. (3-15) is performed by setting the initial observation, obs_8, to the current observation, obs_4, which is then used in (3-12) for calculating the probability of the possible objectives.

$$\theta_{init} = \theta_{i,j} \quad \text{if } \theta_{i,j} \notin \boldsymbol{obs_{prev}}[obs_{next}] \tag{3-15}$$

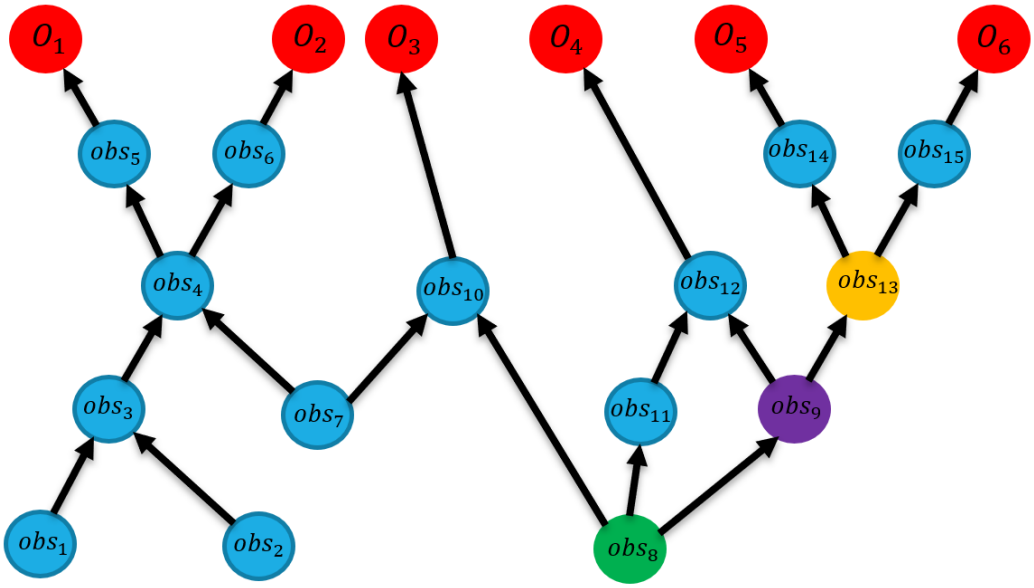


Figure 3-18 Example of observation flow: Initial observation (Green), second observation (Purple), current observation (Yellow), Objectives (Red), Unobserved node (Blue)

The final prediction is the plan, which is the full sequence of activities from the initial activity to the final objective. The predicted plan can be obtained by Eq. (3-16), which, after receiving the predicted objective, predicted next activity, and the current observation as input, searches the most likely plan through the plan library by matching the given inputs, and the outcome of the final predicted plan is

doi:10.6342/NTU202300261

then shown as below:

$$Plan = PLsearch(O_{pred}, A_{pred}, \theta_{i,j}) \tag{3-16}$$

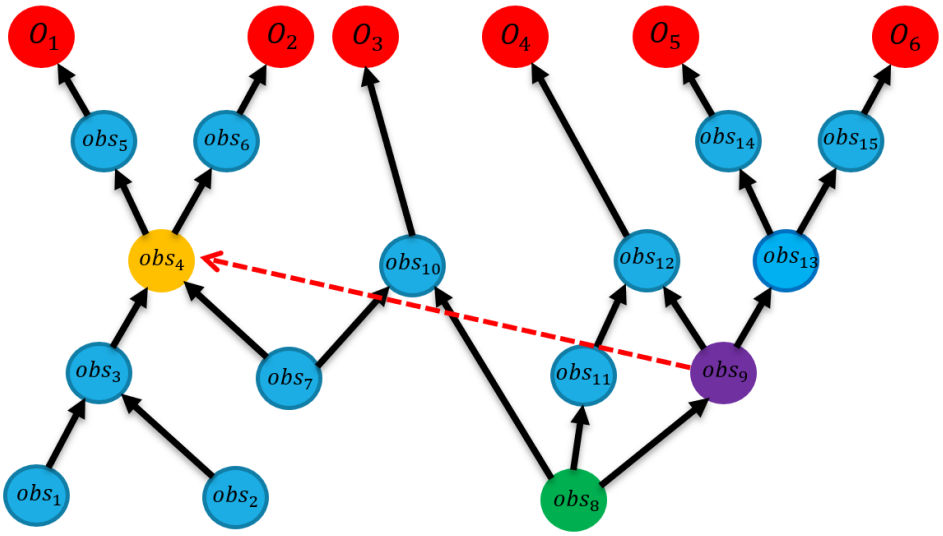


Figure 3-19 Scenario of new plan occurring: Initial observation (Green), second observation (Purple), current observation (Yellow), Objectives (Red), Unobserved node (Blue)

3.6.5 Knowledge Update

At the end of the full observation, the system will need to update the knowledge base to best adapt to the detected human. To do so, the system will need to determine whether the final objective is valid or not. Eq. (3-17) illustrates how the final objective is defined. The system will ask if the predicted objective is correct. If it is true, the system will automatically assign the predicted objective as the final objective. On the other hand, if the predicted objective is false, the system asks user to enter the correct objective as the final objective.

$$ask\ if\ O_{pred}\ is\ correct \begin{cases} objective = O_{pred} & if\ True \\ objective = user\ input & if\ False \end{cases} \tag{3-17}$$

Next, the system updates the Loc-Objective and Loc-NextAct Tensors, with Eq. (3-18) and (3-19):

doi:10.6342/NTU202300261

$$\text{Loc-Objective}_{new} = \textit{LO-Update}(\text{Loc-Objective}_{old}) \quad (3\text{-}18)$$

$$\text{Loc-NextAct}_{new} = \textit{LNA-Update}(\text{Loc-NextAct}_{old}) \quad (3\text{-}19)$$

where $\text{Loc-Objective}_{old}$ and Loc-NextAct_{old} are old tensors that have not yet been updated, $\text{Loc-Objective}_{new}$ and Loc-NextAct_{new} are new tensors that are updated after an iteration, and *LO-Update*$(\cdot)$ and *LNA-Update*$(\cdot)$ are operators that updates the tensors.

The update mechanism of both tensors follows the rule as illustrated. After obtaining the full sequence of activities and their locations. *LO-Update*$(\cdot)$ updates the Loc-Objective tensor by first, incrementing the utility value of related indices of each activity, its location, and the final objective. Second, the utility values in the tensor for the activity and its location corresponding to all objectives are multiplied by 0.9. Given an example of observation = (drinking water, bedroom) and the objective is Taking Medicine. The index of the tensor that needs to be incremented is (3, 0, 1). Then, all values of indices from (3, 0, 0) to (3, 0, 22) are multiplied by 0.9. This indicates that when drinking water in the bedroom is observed, the objective is more likely to be taking medicine. *LNA-Update*$(\cdot)$, on the other hand, updates the Loc-NextAct tensor by first incrementing the utility value of the indices of each observation corresponding to its next observation. If an observation is (pouring water, dining room) and its next observation is (drinking water, dining room), then the utility value of the corresponding index (1, 3, 3, 3) of tensor Loc-NextAct is incremented. Second, all utility values of all possible next observation of each observation from first to second last in the sequence is multiplied by 0.9. The multiplication indicates the downgrade of the relative importance between all the past observations through time. Given an example from Figure 3-18, the value of the index corresponding to

doi:10.6342/NTU202300261

(obs_8, obs_9) is incremented. Then, they are multiplied by 0.9 along with value of the index corresponding to (obs_8, obs_{10}) and (obs_8, obs_{11}). The pseudo code of the update is shown in Figure 3-20 below.

ALGORITHM I: LOC-NEXTACT & LOC-OBJECTIVE TENSOR UPDATE

Input: ACT: **ETRI_Activities** – List of ETRI activities (55).
LOC: **Locations** – List of Locations (6).
OBJ: **Objectives** – List of Objectives (23).
Observation: Full observation
objective: Final objective of the sequence of observations

Parameters: obs_{curr}: Current observation in the sequence.
obs_{next}: Next observation in the sequence.
i: Index number of **ETRI_Activities** for current activity.
j: Index number of **Locations** for current activity location.
w: Index number of **ETRI_Activities** for next activity.
r: Index number of **Locations** for next activity location.

Output: LNAT: **Loc-NextAct** Tensor.
LOT: **Loc-Objective** Tensor
PL: **Plan Library**.

```
for all obs ∈ Observation do
  if obs_next == None then
    break
  else
    i ← Index(ACT(obs_curr[activity])).
    j ← Index(LOC(obs_curr[location])).
    w ← Index(ACT(obs_next[activity])).
    r ← Index(LOC(obs_next[location])).
    k ← Index(OBJ(objective)).
    LOT_{i,j,k} + 1.                          LOT-Update
    for O ∈ OBJ do
      k ← Index(OBJ(O)).
      LOT_{i,j,k} * 0.9
    end
    LNAT_{i,j,w,r} + 1.
    for A ∈ ACT and L ∈ LOC do
      w ← Index(ACT(A)).
      r ← Index(LOC(L)).
      LNAT_{i,j,w,r} * 0.9                    LNAT-Update
    end
end
if Plan_new == True then                      PL-Update
  PL ← (obs, objective)
return LNAT, LOT, PL
```

Figure 3-20 Pseudo code of updating Loc-NextAct and Loc-Objective Tensors

doi:10.6342/NTU202300261

Plan library, on the other hand, is updated only if a new plan is detected. The new plan will be added to the plan library for future predictions using Eq. (3-20):

$$\text{PlanLibrary}_{new} = PL\text{-}Update(\text{PlanLibrary}_{old}) \quad (3\text{-}20)$$

where PlanLibrary_{old} is the old plan library, PlanLibrary_{new} is the newly updated plan library, and $PL\text{-}Update(\cdot)$ adds the newly detected plan into the plan library according to the final objective. The pseudo code of Plan Library update is shown in Figure 3-20 above.

3.7 Robot Response Module

In the field of robotics, human robot interaction (HRI) and human robot collaboration (HRC) are main purposes of investigating robot applications. In this thesis, with the given knowledge from the outputs of HAR and plan recognition, the final step is to utilize them to interact with humans. We create a robot response module that gives various kinds of responses in terms of the current activity, predicted next activity, and the final objective.

3.7.1 Different Reponses for Different Activities

The responses in the response module are classified into 3 categories, namely, advice, warning, and reminder. Different activities react to different responses. Table 3-4 below shows the responses of three different ETRI activities. On the other hand, Table 3-5 shows the responses of three different objectives. As an activity is being observed, the response module outputs a response corresponding to the activity. The priority of outputting responses when responses is detected is current activity, next activity, then objective. If activity contains advice, warning, and reminder responses,

doi:10.6342/NTU202300261

the sequence of choosing which kind of responses is warning, reminder, and advice. If multiple responses are detected in the chosen category, the module will randomly choose one response to output. Finally, the response is marked as chosen to avoid responding the same sentence repeatedly.

Table 3-4 Responses reacting to different activities

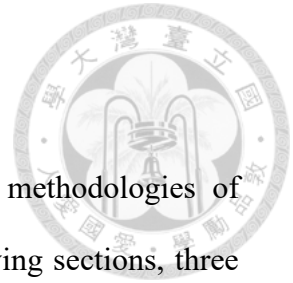
Activities	Putting/taking food in/from the fridge	Trimming vegetables	Eating food with a fork
Advice	"Check if the food is expired or gone bad"	-	-
Warning	-	"Be careful of sharp knife"	-
Reminder	"make sure to close the fridge"	"remember to wash vegetable afterwards"	"make sure to clean up after the meal"

Table 3-5 Responses reacting to different objectives

Objectives	Watch TV	Going out doors	Cooking
Advice	-	"Bring an umbrella in case of a rainy day"	-
Warning	-	-	"Be careful of fire when using the stove"
Reminder	"Don't watch too long, remember to rest your eyes"	"Remember to bring your phone"	-
Speak	-	-	"is there any thing I can do to help with the cooking?"

doi:10.6342/NTU202300261

Chapter 4 Experiment



In this chapter, the experiments are set up to validate the methodologies of several major parts proposed in the previous chapter. In the following sections, three experimental parts, namely, location estimator, AL-GCN, and plan recognition, will be shown in details. The location estimator will first be tested, followed by our activity recognition model. Finally, plan recognition is evaluated to test the predicted outcomes. However, before diving into these experiments and validations, the experimental setup will first be described.

4.1 Experimental Setup

In this section, the experimental setup for our human activity recognition and plan recognition are discussed, including the utilized datasets, the implemented robot, the mounted sensors, and the experimental environment.

4.1.1 Human Activity Recognition

In our experiment for human activity recognition, the ETRI Activity 3D dataset [24] is used to train and validate our proposed AL-GCN model. This dataset targets on collecting daily activities that elders perform in a household environment. The wide range of variety in terms of subject, age, sex, environment, and camera view makes it more suitable for real-world applications.

The dataset consists of different types of data, including RGB videos with 1920 x 1080 resolution, depth video with 512 x 424 resolution, and 3D skeleton information with 25 joint features of a human body. Moreover, 55 activity classes are being executed and recorded (see Appendix I). There are a total number of 104,200 out of

doi:10.6342/NTU202300261

112,238 samples collected, and 100 subjects consisting of 50 elders and 50 young people are being recorded with a setup of 8 different cameras, see Figure 4-1. The activities are conducted in 7 different locations in a household environment, and the total size of the final dataset is 4.44 TB.

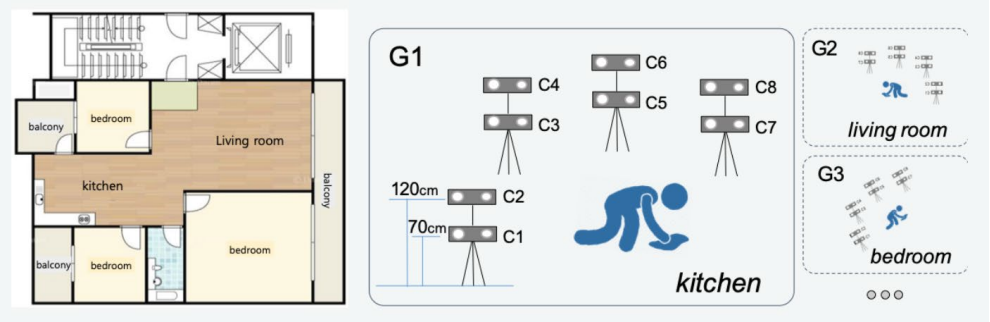


Figure 4-1 Layout of the rooms and configuration of the data acquisition system [24]

This dataset can be evaluated through cross-subject and cross-view as has been proposed by [68]. The cross-subject is trained and tested on videos of different subjects. The dataset is separated by different subject ID numbers. Those whose ID number is divisible by 3 is used for testing. The remaining subjects are used for training. See top bar of Figure 4-2 for data arrangement for cross-subject test. On the other hand, the cross-view is trained and tested on different camera setups. The total number of cameras in their setup is 8 with camera indices 1~6 having used for training and camera indices 7~8 having used for testing. See bottom bar of Figure 4-2 for data partitioning for cross-view test.

doi:10.6342/NTU202300261

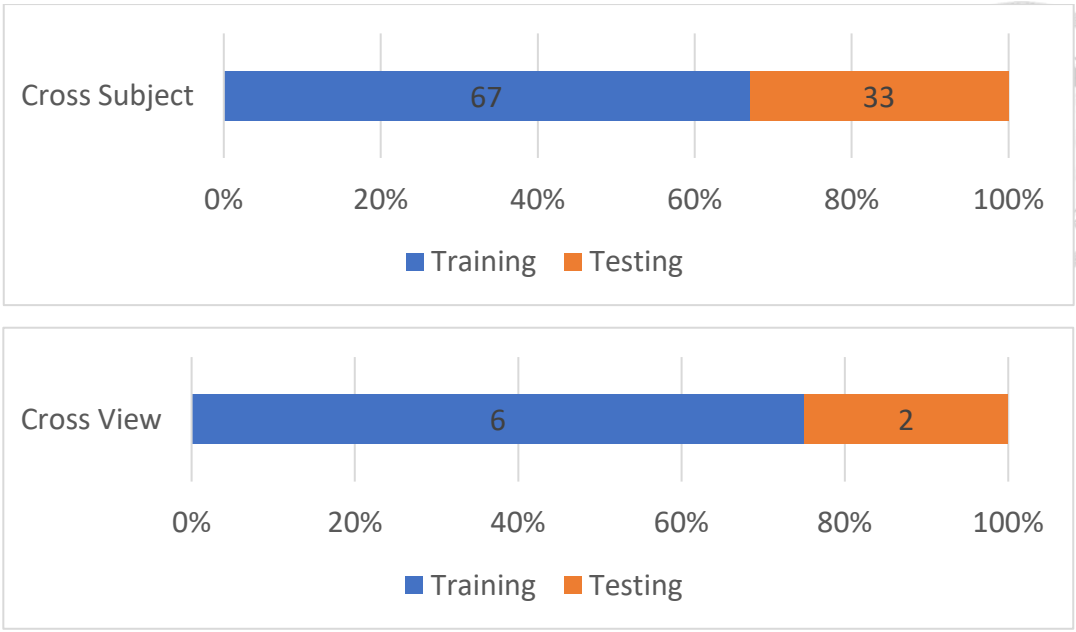


Figure 4-2 Division between training and testing set for Cross-subject and Cross-view validation

4.1.2 Plan Recognition

In our plan recognition experiment, the dataset is gathered through conducting of a survey based on the ETRI 3D activities. The survey as shown in Appendix III, consisting of multiple questions aiming at collecting a wide variety of plans that are executed towards different objectives in a household situation.

The total number of individuals of the dataset is 30, consisting 16 females and 14 males. The age range is between 18 and 64. The top figure of Figure 4-3 indicates the percentage of both sex in the survey, while the bottom indicates the percentage of each age range. The survey, as shown in Appendix III, consists of questions for 23 objectives, each containing relevant activities and locations that can be chosen to achieve them. For each objective, the participants choose and list the activities in order to achieve the objective in their daily life. The survey samples are used to construct the plan library and to initialize our Loc-NextAct and Loc-Objective Tensors, as mentioned in the previous Chapter. Figure 4-4 below shows a partial construction of the plan library with the objective: Relieve stress.

doi:10.6342/NTU202300261

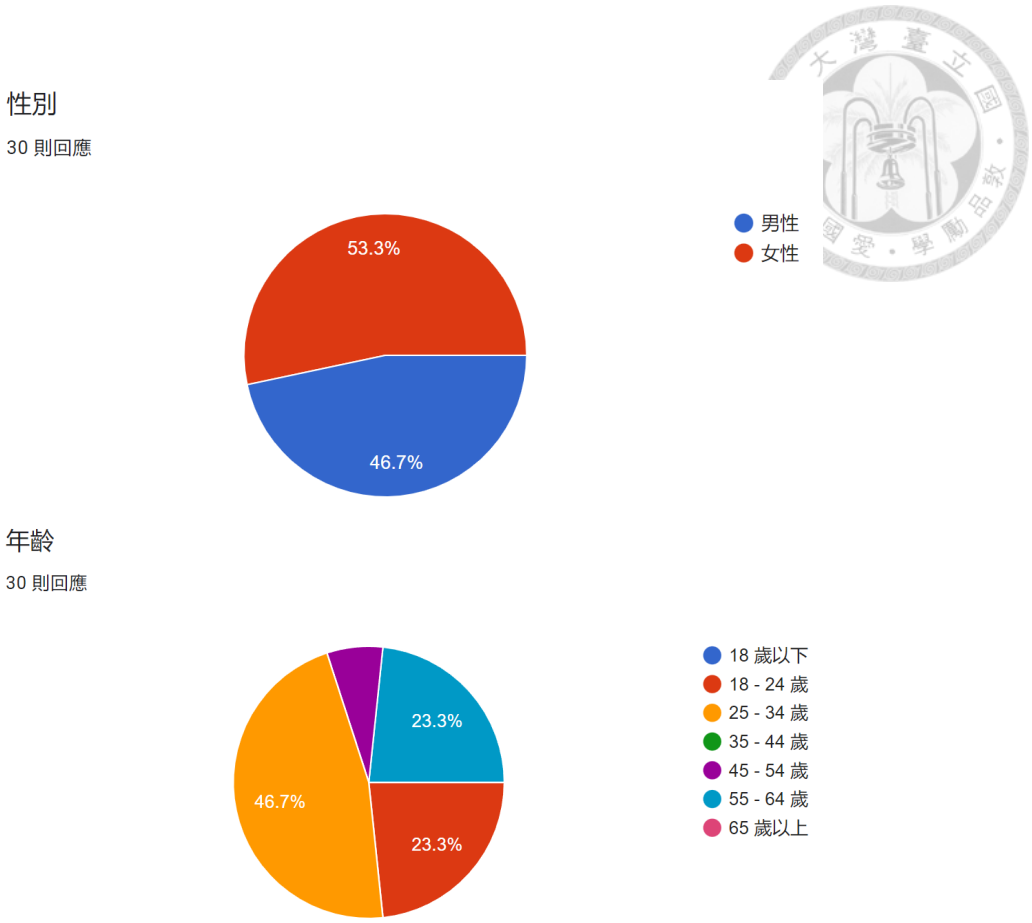


Figure 4-3 Samples collected regarding sex and ages

```
<Recipes>
    <Recipe Objective="Relieve Stress" id="RS">
        <Params id="0">
            <Param activity="doing neck roll exercise" location="Living room" />
            <Param activity="massaging a shoulder oneself" location="Living room" />
            <Param activity="clapping" location="Study room" />
            <Param activity="doing free hand exercise" location="Living room" />
            <Param activity="massaging a shoulder oneself" location="bedroom" />
            <Param activity="smoking" location="Entrance" />
            <Param activity="doing neck roll exercise" location="bedroom" />
            <Param activity="doing free hand exercise" location="bedroom" />
            <Param activity="clapping" location="Living room" />
            <Param activity="massaging a shoulder oneself" location="Study room" />
            <Param activity="doing free hand exercise" location="Entrance" />
            <Param activity="clapping" location="bedroom" />
            <Param activity="doing neck roll exercise" location="Study room" />
            <Param activity="doing free hand exercise" location="Study room" />
            <Param activity="clapping" location="Dining room/kitchen" />
            <Param activity="massaging a shoulder oneself" location="Entrance" />
            <Param activity="doing neck roll exercise" location="Dining room/kitchen" />
            <Param activity="doing neck roll exercise" location="Bathroom" />
            <Param activity="doing free hand exercise" location="Dining room/kitchen" />
            <Param activity="smoking" location="Living room" />
        </Params>
        <Params id="1">
            <Param activity="doing neck roll exercise" location="Living room" level="1" />
            <Param activity="massaging a shoulder oneself" location="Living room" level="2" />
        </Params>
        <Params id="2">
            <Param activity="clapping" location="Study room" level="1" />
            <Param activity="doing free hand exercise" location="Living room" level="2" />
            <Param activity="doing neck roll exercise" location="Living room" level="3" />
            <Param activity="massaging a shoulder oneself" location="bedroom" level="4" />
            <Param activity="smoking" location="Entrance" level="5" />
        </Params>
        <Params id="3">
            <Param activity="massaging a shoulder oneself" location="Living room" level="1" />
            <Param activity="doing neck roll exercise" location="bedroom" level="2" />
            <Param activity="doing free hand exercise" location="bedroom" level="3" />
        </Params>
```

Figure 4-4: Partial construction of the plan library

doi:10.6342/NTU202300261

4.1.3 Robot setup: OREO

The robot that used in our conducted real-world experiment is an omnidirectional robot called OREO. As shown in Figure 4-5, three wheels are mounted on the robot to allow it to move in any direction without turning. Moreover, the robot is equipped with various sensors, including Sick LMS 100 2D Laser on the robot bottom to carry out tasks, such as localization and mapping, and a StereoLab ZED 2 Camera on the robot top for receiving RGB-D videos. Behind the Sick Laser is an industrial personal computer (IPC) running Ubuntu 16.04 along with robot operating system (ROS) Kinetic version.

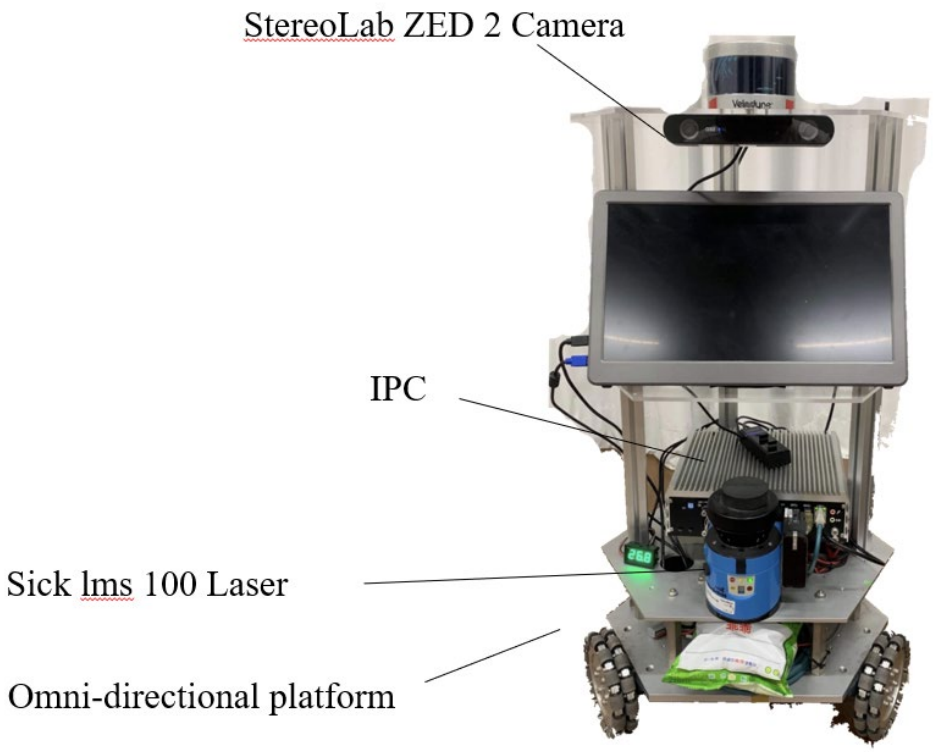


Figure 4-5 Our mobile robot: OREO

4.1.4 Sensor setup: ZED 2 Camera

The sensor we utilized for capturing images and videos is ZED 2 camera [69], which is a camera provided by a company called “StereoLab”. The two camera lenses

doi:10.6342/NTU202300261

apart reconstruct the depth information of the image by performing triangulation algorithm. Table 4-1 below is the specifications of ZED 2 camera.

As mentioned earlier, ZED 2 camera is mounted on top of OREO, which is utilized to capture the RGB image of the environment and the depth of objects and human whereby the distances among them can be calculated in order to infer which location the human is currently in. Moreover, it provides skeleton data that consists of joint features of the observed human that is being processed in our human activity model. ZED 2 is also compatible with ROS so that its data can be more easily obtained, processed, and utilized.

Table 4-1 Specification of ZED 2 Camera

	
Output Resolution	• (2208x1242)@15fps • (1920x1080)@30fps • (1280x720)@fps • (672x376)@fps
Field of View	Max. 110°(H) x 70°(V) x 120°(D)
Depth range and accuracy	Range: 0.3 m to 20 m (1 to 65.6 ft) Accuracy: <1% up to 3m <5% up to 15 m
Dimension and weight	Dimension: 175 x 30 x 33 mm Weight: 166g
Sensors	Motion: Gyroscope, Accelerometer, Magnetometer Environmental: Barometer, Temperature

doi:10.6342/NTU202300261

4.1.5 Environmental setup: Yonglin Lab

In the real-world experiment, it is conducted in our Yonglin lab which is renovated as a household environment, see Figure 4-6. The whole area size is around 6.8 m x 11.8 m. The environment is divided into 6 areas, including, bedroom, study room, living room, dining room, entrance, and bathroom. In the environment, Gmapping [6] is utilized to scan the surrounding and provides a 2D grid map, as shown in Figure 4-7.

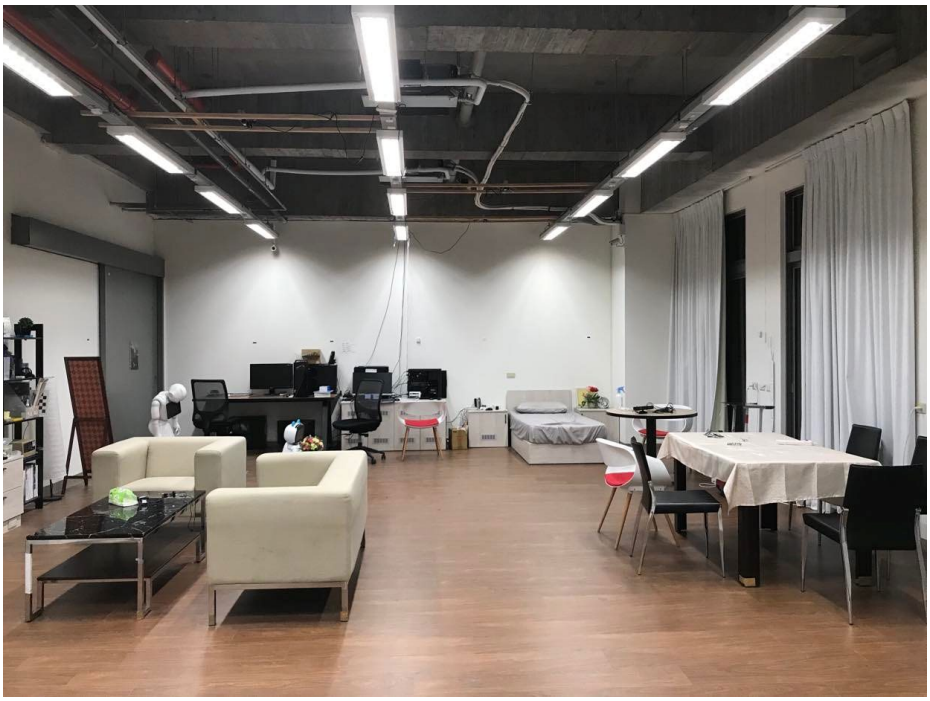


Figure 4-6 Real world experiment environment: Yonglin Lab

By using Adaptive Monte Carlo Localization (AMCL) [59], the robot is able to localize its position on the 2D grid map. AMCL produces a population of particles that represents the possible position of the robot on the map. As the robot receives new measurements, the particles converge with the highest confidence particle acts as the estimated robot pose. By using transformation matrix, the human and object poses can be found in terms of the map's coordinate.

doi:10.6342/NTU202300261

Figure 4-7 2D grid map of Yonglin Lab

4.2 Location Estimation

4.2.1 Experiment and Evaluation on ETRI Dataset

In this section, we tested our location estimator on the ETRI Activity 3D dataset. The test is performed under 100 individuals with a total of 112,238 videos processed that are spread into 6 locations. Table 4-2 below shows the accuracy of utilizing different kinds of models. Moreover, each location is tested with different number of samples: 18,253 for Bedroom, 5,588 for Study room, 47,930 for Living room, 31,688 for Dining room, 4,744 for Entrance, and 4,035 for Bathroom. It can be seen that with the consideration of distances between people and the surrounding objects, the model is able to more accurately estimate samples with multiple locations involved. On the other hand, ResNet50-place365 can enhance the location estimation model when samples lack objects. This is apparent for Entrance samples which lacks objects that can be detected.

doi:10.6342/NTU202300261

Table 4-2 Accuracy of different models for location estimation

Locations (# of videos)	ResNet50-place365	Location Estimator	**ResNet50-place365 + Location Estimator**
Bedroom (18,253)	81.31%	84.28%	**93.72%**
Study room (5,588)	84.17%	85.71%	**89.01%**
Living room (47,930)	82.51%	91.35%	**93.81%**
Dining room (31,688)	85.84%	88.27%	**91.36%**
Entrance (4,744)	85.12%	79.93%	**88.57%**
Bathroom (4,035)	87.46%	84.85%	**94.10%**
Total Accuracy	84.41%	85.73%	**92.83%**

In addition, the average likelihood values of the correct videos for each room using our proposed model are calculated and shown in Table 4-3 below. The total number of correct videos is 104,200 out of 112,238. Each room has a different number of correct videos calculated: 17,107 for bedroom, 5,178 for study room, 44,963 for living room, 28,953 for dining room, 4,202 for entrance, and 3,797 for bathroom videos. Table 4-3 below shows the average likelihood values of each location. As shown in the table below, living room has the highest likelihood followed by bathroom. This can be expected due to the fact that, when testing the videos, sofa, TV monitor, and toilet are more stably detected by YOLO v4 object detection as compared with detection of bed or dining table. Hence, the estimated performance on bedroom and dining room is worse. Entrance, on the other hand, lacks big representative objects, that affects the stability of the estimation. Lastly, although Study room does not have a representative object, it is augmented by the number of Affordance objects.

doi:10.6342/NTU202300261

Table 4-3 Average likelihood of different locations tested on ETRI dataset

Locations (# of videos)	Bedroom	Study room	Living room	Dining room	Entrance	Bathroom
Bedroom (17,107)	**84.48%**	9.59%	2.96%	0.89%	1.22%	0.86%
Study room (5,178)	3.46%	**85.82%**	3.83%	2.58%	2.32%	1.98%
Living room (44,963)	1.04%	4.80%	**88.40%**	2.04%	2.06%	1.66%
Dining room (28,953)	5.38%	4.18%	1.72%	**84.42%**	2.32%	1.98%
Entrance (4,202)	3.82%	3.55%	4.36%	5.01%	**76.04%**	7.22%
Bathroom (3,797)	0.19%	1.20%	0.12%	6.44%	5.26%	**86.80%**

4.3 Human Activity Recognition (AL-GCN Model)

4.3.1 Experiment and Evaluation on ETRI Dataset

In this section, our proposed model for HAR is validated on the ETRI Activity 3D dataset and it is mounted on our robot, OREO. The validation is in terms of the metric, accuracy, [70] that is defined below

$$Accuracy = \frac{TP}{TP + FP} \tag{4-1}$$

where:

TP: Number of True Positive

FP: Number of False Positive

In this experiment, our model is evaluated and compared with different models that process the same dataset, see Table 4-4. The compared models vary from different architectures, including CNN and GCN based models, to different types of inputs, such as videos and skeletons. These models leverage the strength of model architectures and input information to increase the activity recognition accuracy. RGB

doi:10.6342/NTU202300261

3D-CNN [71] utilizes CNN architecture that is suitable in processing grid structure input and learns temporal dynamic features. Thus, their input is RGB Videos. ST-GCN [39] is considered as the base-line for GCN-based model that processes skeleton data and learns spatial-temporal features. The graph-like structure of skeleton joints is adequate for GCN-based model. FSA-CNN [24] transforms the skeleton data into a pseudo image with vertical axis as the joint coordinates and horizontal axis as the temporal information. Then, the image is processed through a CNN-based architecture. Ensemble DNN [71] utilized 3 streams each of which has a different back-bone network of processing different inputs. The first stream is a 2D-CNN structure processing skeleton data. The second stream is a 2D-CNN + LSTM structure processing RGB video data. The last stream is a 3D-CNN structure processing RGB video data. Each stream produces a probability of the activities and the sum of these probabilities is the final ensemble probability. As we can see in Table 4-4, our model utilizes GCN-based structure that process skeleton data incorporating location information. Our proposed model outperforms the others.

Table 4-4 Comparison between our HAR model with other existing models

Related Works	**Input**	**Cross View Accuracy**	**Cross Subject Accuracy**
RGB 3D-CNN [71]	RGB Video	-	79.20%
ST-GCN [39]	Skeleton	77.88% (Result from [71])	83.36% (Result from [71])
FSA-CNN [24]	Skeleton	86.44 %	90.6 %
Ensemble DNN [71]	RGB Video Skeleton	-	93.20%
AL-GCN w/o Location [65]	Skeleton	94.02%	94.15%

doi:10.6342/NTU202300261

AL-GCN w/ Location (Ours)	Skeleton Location	**94.27%**	**94.33%**
Note: - Result is not available			

Next, we evaluate our method against the previous work that utilizes object information to improve the activity accuracy. The model is called Human Object Interaction Graph Convolutional Network (HOI-GCN). This model utilizes YOLO v4 object detection to extract the objects' bounding boxes in the dataset videos and define the center of the bounding boxes as the reference points of the objects. By using the pixel coordinate of YOLO v4 and depth of ZED 2 camera, the objects' (x, y, z) coordinate can be obtained and viewed as a node that is being processed into the model along with the skeleton data. However, only the object that is closest to the hand is being chosen for processing since activity is usually related to the object that is interacted by hand.

HOI-GCN processes 200 frames, fewer than the standard 400 frames adopted commonly in the literature for comparison. Generally speaking, with fewer frames processed, the information should be less sufficient and may decreases the accuracy. However, on the other hand, with fewer frames, the inference time can be reduced which is beneficial for real-world applications. Moreover, even with the drop of accuracy, the accuracy is still considerably high, not to mention with the consideration of objects. Thus, we compare our model with HOI-GCN also in terms of frame numbers. As shown in Table 4-5, we compared our model using only 200 frames with HOI-GCN model. Without any additional information, the model we utilized from [65] has a higher accuracy as compared with the model, HOI-GCN without incorporating objects, under the condition with 200 frames. With additional location information, our model accuracy is higher than HOI-GCN model with object information included.

doi:10.6342/NTU202300261

Finally, the best model is our model with location information included under the condition with 400 frames.

Table 4-5 Comparison in terms of input frame number

Model type	**Input**	**Frames**	**Cross View Accuracy**	**Cross Subject Accuracy**
HOI-GCN w/o Object	Skeleton	200	93.28%	92.99%
AL-GCN w/o Location [65]	Skeleton	200	93.39%	93.16%
HOI-GCN w/ Object	Skeleton Object	200	93.44%	93.46%
AL-GCN (ours)	Skeleton Location	200	93.57%	93.61%
AL-GCN w/o Location [65]	Skeleton	400	94.02%	94.15%
AL-GCN (ours)	Skeleton Location	400	**94.27%**	**94.33%**

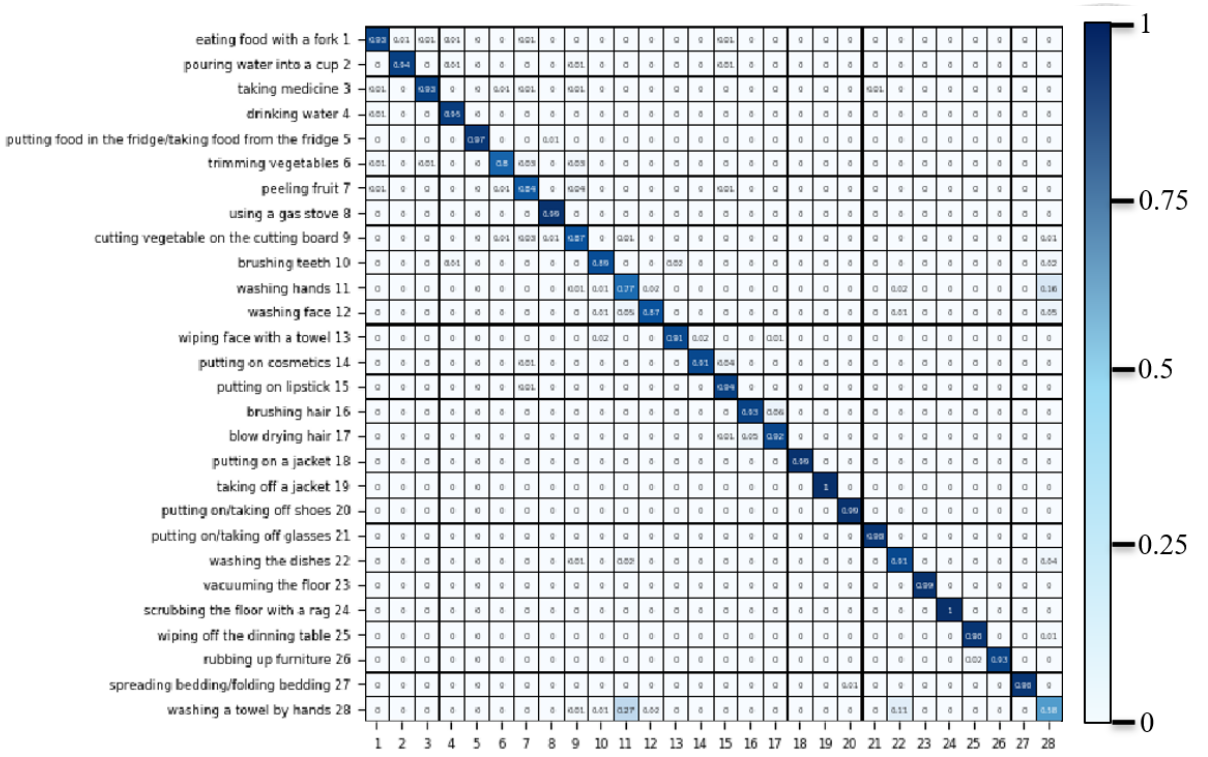
As mentioned previously, the number of frames processed can influence the inference rate which would be a major concern in real-world applications. The inference rate of HOI-GCN is tested to be around 10 seconds in terms of 200 frames. However, in our case, the inference rate of our AL-GCN model is tested to be around 13 to 15 seconds in terms of 400 frames, which still are acceptable in real-world applications, such as monitoring elder's safety.

Finally, we visualize the evaluation of our model with and without the incorporation of location. Figure 4-8 below is the confusion matrix of our model without including the location information. Figure 4-8 (a) is the diagonal section of the first 28 activities, whereas Figure 4-8 (b) is the diagonal section of the remaining

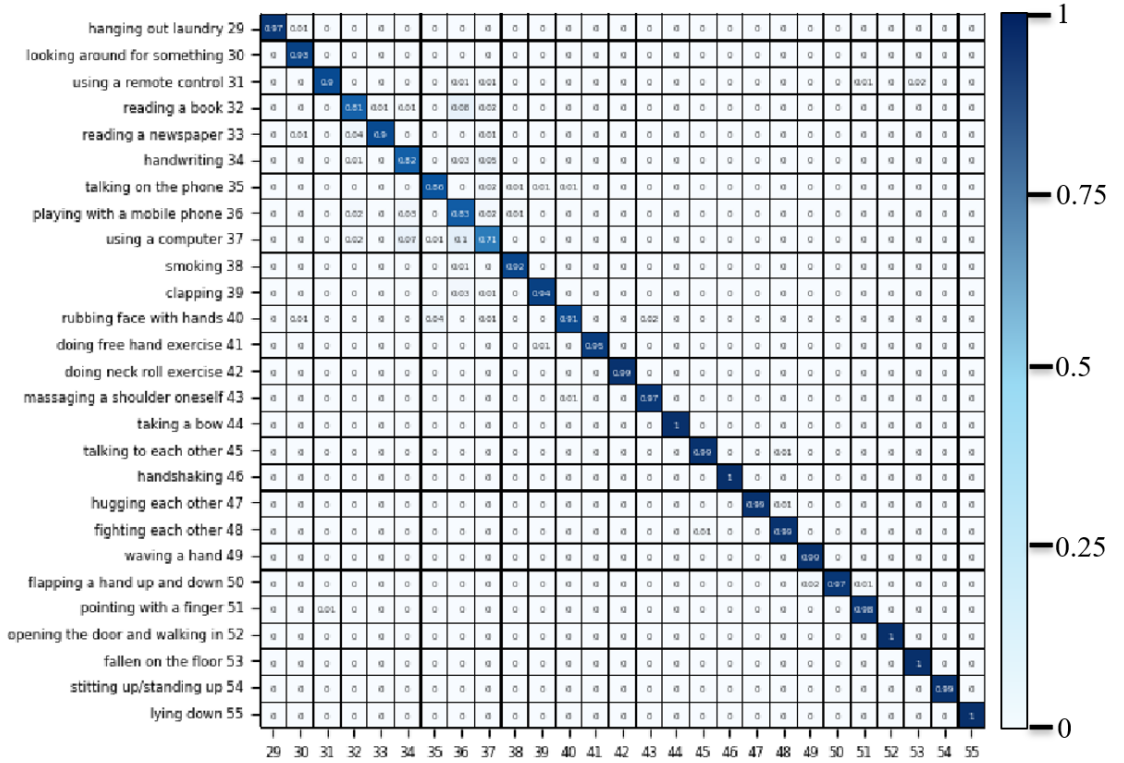
doi:10.6342/NTU202300261

27 activities. Figure 4-9, on the other hand, is the confusion matrix with including the location information. Same as Figure 4-8, Figure 4-9 (a) is the diagonal section of the first 28 activities and Figure 4-9 (b) is the diagonal section of the remaining 27 activities. These two matrices are the results trained under cross-subject circumstances. By comparing these two matrices, the false detected activities are reduced when location information is included. In other words, some of the indices that are off the diagonal are decreased. However, to identify the differences and improvement more easily, the accuracy of recognizing each activity along the diagonal of the confusion matrix of Figure 4-8 is shown in Figure 4-10 and that of Figure 4-9 is shown in Figure 4-11. It can be visualized that on average the recognition accuracy has all been increased once the location information has been incorporated. Table 4-6 below shows the exact percentage and differences of the evaluation of the two cases, one w/ and the other w/o location information. Through this evaluation, the most improved activities are those executed in the kitchen. We believe that it is mainly due to the distinctiveness of these activities, that is, since they are executed only in the kitchen but not anywhere else, the model will not confuse them with other activities. Therefore, the performance increase from w/o case to w/ case is the most significant. On the contrary, there are also activities in Table 4-6 that show a drop in recognition accuracy, which are those that either lie in the videos where two people are involved that may generate noises and uncertainties when training and testing the AL-GCN model, or their involved motions are similar to those that can be detected in multiple locations.

doi:10.6342/NTU202300261



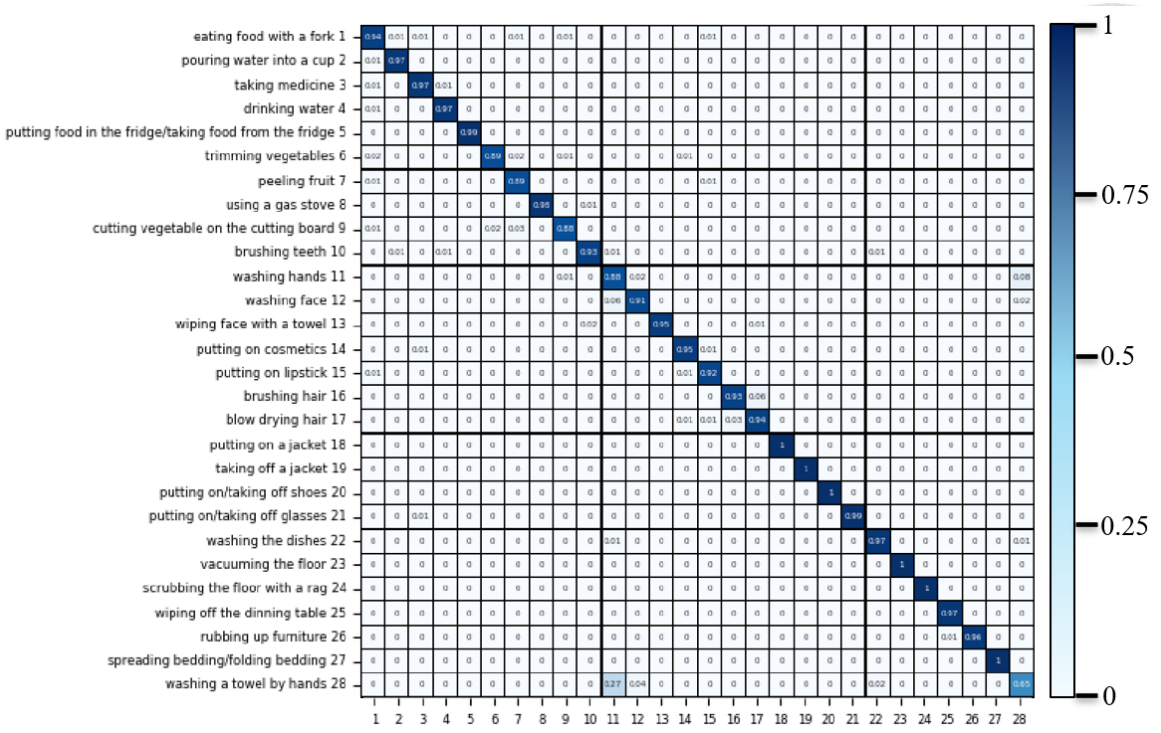
(a)



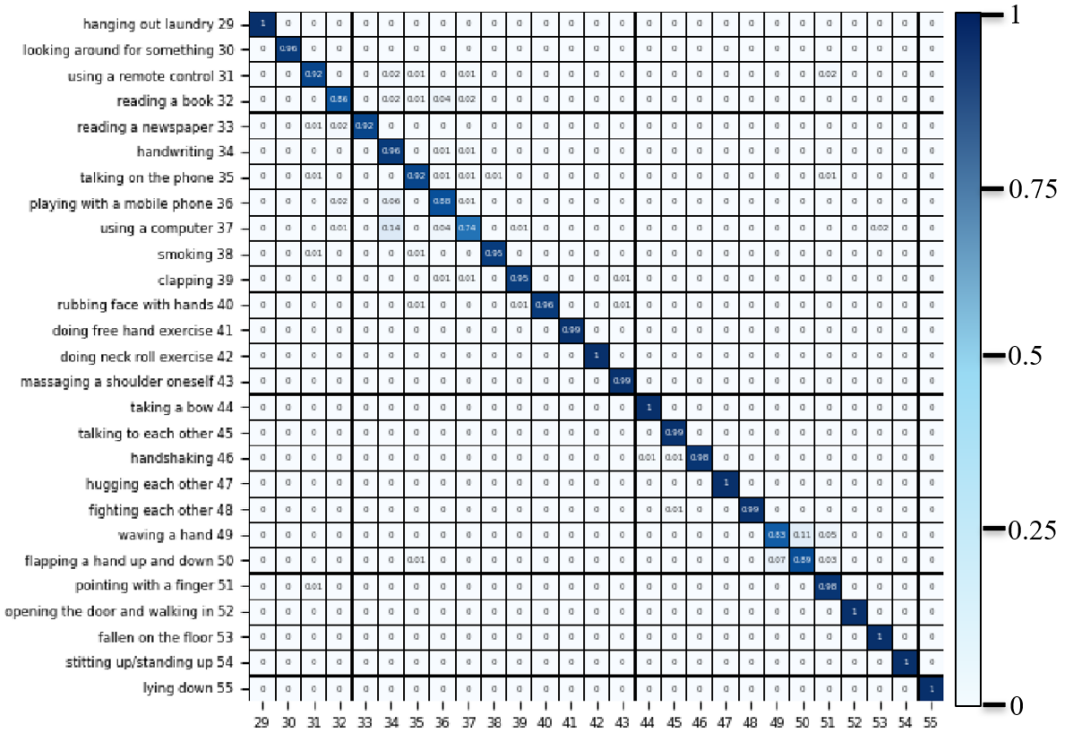
(b)

Figure 4-8 Confusion matrix without location information

doi:10.6342/NTU202300261



(a)



(b)

Figure 4-9 Confusion matrix with location information

doi:10.6342/NTU202300261

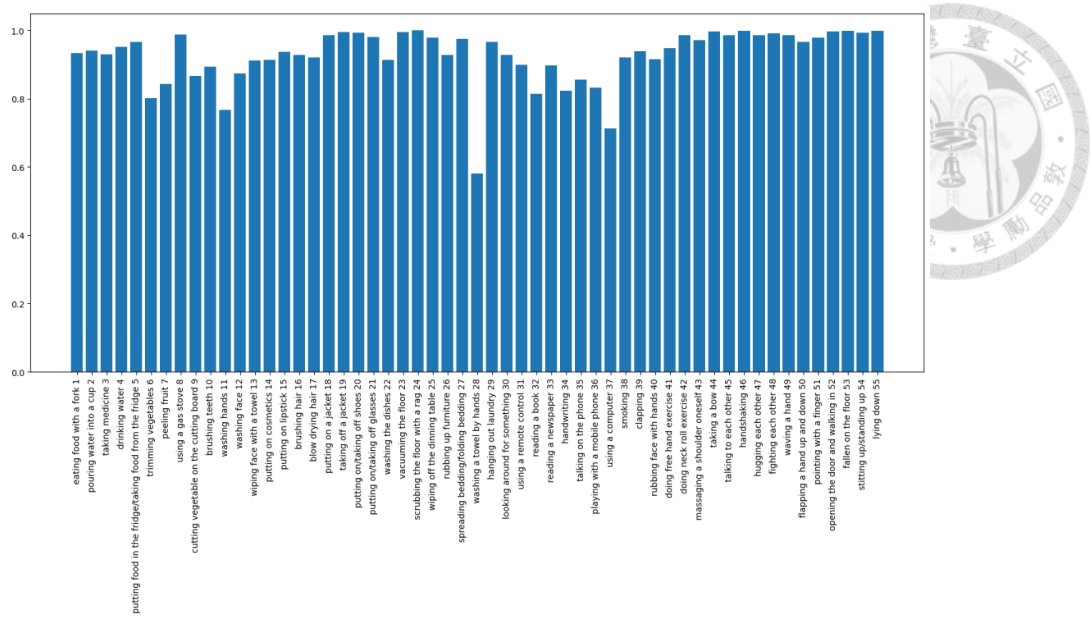


Figure 4-10 Bar chart of activity accuracies without location information

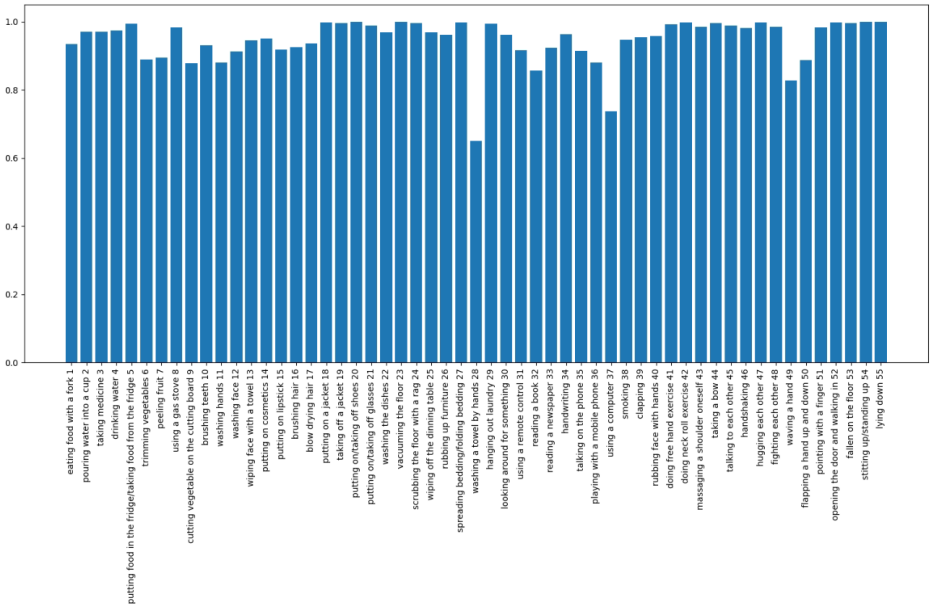


Figure 4-11 Bar chart of activity accuracies with location information

Table 4-6 Accuracy of activity improvements and degradation

Activities	With location	Without location	Sum
1. Eating food with a fork	96.85%	93.37%	+2.48%
2. Pouring water into a cup	98.23%	94.15%	+4.08%
3. Taking medicine	98.20%	93.02%	+5.18%
4. Drinking water	97.07%	95.10%	+1.97%
5. Putting food in the fridge/ taking food from the fridge	99.38%	96.69%	+2.69%
6. Trimming vegetables	91.12%	80.11%	+11.01%

doi:10.6342/NTU202300261

7. Peeling fruit	93.72%	84.33%	+9.39%
8. Using a gas stove	99.50%	98.71%	+0.79%
9. Cutting vegetable on the cutting board	92.38%	86.58%	+5.80%
10. Brushing teeth	94.01%	89.41%	+4.60%
11. Washing hands	86%	76.75%	+9.25%
12. Washing face	91%	87.41%	+3.59%
13. Wiping face with a towel	94.10%	91.14%	+2.96%
14. Putting on cosmetics	96.74%	91.30%	+5.44%
15. Putting on lipstick	93.31%	93.63%	-0.32%
16. Brushing hair	92.89%	92.76%	+0.13%
17. Blow drying hair	93.37%	92.02%	+1.35%
18. Putting on jacket	99.92%	98.53%	+1.39%
19. Taking off jacket	99.83%	99.51%	+0.32%
20. Putting on/taking off shoes	99.88%	99.29%	+0.59%
21. Putting on/taking off glasses	99.28%	98.03%	+1.25%
22. Washing the dishes	98.24%	91.33%	+6.91%
23. Vacuuming the floor	100%	99.42%	+0.58%
24. Scrubbing the floor with a rag	99.87%	100%	-0.13%
25. Wiping off the dining table	96.88%	97.79%	-0.91%
26. Rubbing up furniture	96.50%	92.81%	+3.69%
27. Spreading bedding/ folding bedding	100%	97.52%	+2.17%
28. Washing a towel by hands	72.50%	58.01%	+14.49%
29. Hanging out laundry	100%	96.69%	+3.31%
30. Looking around for something	97.53%	92.82%	+4.71%
31. Using a remote control	93.23%	89.84%	+3.39%
32. Reading a book	88.55%	81.47%	+7.08%
33. Reading a newspaper	95.00%	89.69%	+5.31%
34. Handwriting	97.25%	82.35%	+14.90%
35. Talking on the phone	91.77%	85.60%	+6.17%
36. Playing with a mobile phone	89.73%	83.18%	+6.55%
37. Using a computer	79.38%	71.21%	+8.17%
38. Smoking	94.82%	92.11%	+2.71%
39. Clapping	96.71%	93.95%	+2.76%
40. Rubbing face with hands	96.67%	91.49%	+5.18%
41. Doing free hand exercise	99.69%	94.73%	+4.96%
42. Doing neck roll exercise	99.73%	98.54%	+1.19%
43. Massaging a shoulder oneself	98.04%	97.22%	+0.82%
44. Taking a bow	99.88%	99.63%	+0.25%
45. Talking to each other	94.17%	98.63%	-4.46%
46. Handshaking	99.12%	99.82%	-0.70%
47. Hugging each other	99.89%	98.68%	+1.21%

doi:10.6342/NTU202300261

48. Fighting each other	94.11%	99.22%	-5.11%
49. Waving a hand	83.50%	98.53%	-15.03%
50. Flapping a hand up and down	90.95%	96.69%	-5.74%
51. Pointing with a finger	98.56%	97.79%	+0.77%
52. Opening the door and walking in	100%	99.63%	+0.37%
53. Fallen on the floor	99.62%	99.91%	-0.29%
54. Sitting up/standing up	99.92%	99.39%	+0.53%
55. Lying down	100%	99.88%	+0.12%

4.4 Plan Recognition

In this section, our plan recognition is being experimented and evaluated on our survey samples from Appendix IV. There are a total of 282 samples that are being tested, each consists with various length between 2 and 8 of activity sequences. Moreover, this section is partitioned into several parts of prediction evaluations, namely Next Activity Prediction, Objective Prediction, and Plan Prediction, and finally the evaluation on the influence when knowledge update is in play.

4.4.1 Next Activity Prediction

The prediction of next activity is the first major part that plays in our plan recognition algorithm. The prediction of next activity of initial observation may affect the plan prediction since it utilizes this knowledge. The experiment is performed under 8 observations and 8 iterations of every sequence of activities from the initial setup. That is, the 282 samples from the survey shown in Appendix IV are used for evaluation. Each sample is executed through the system 8 consecutive times with an update after each iteration. The system outputs a predicted next activity when an activity in each sequence is being read. The correct prediction of each sequence in each iteration is recorded and the accuracy of each iteration are calculated to view the performance. The average percentage of correct prediction in each iteration is shown

doi:10.6342/NTU202300261

in Figure 4-12 to view the influence as updates are being made. Since the first iteration predicts from the initial setup without any knowledge acquired yet, the prediction is somewhat low. However, as the same sequence of activities being executed, the accuracy of predicting the next activity starts to increase. As it reaches the 8th iteration, the knowledge base would have updated 7 times and the prediction of next activity would be accurate for every observation in such sequence.

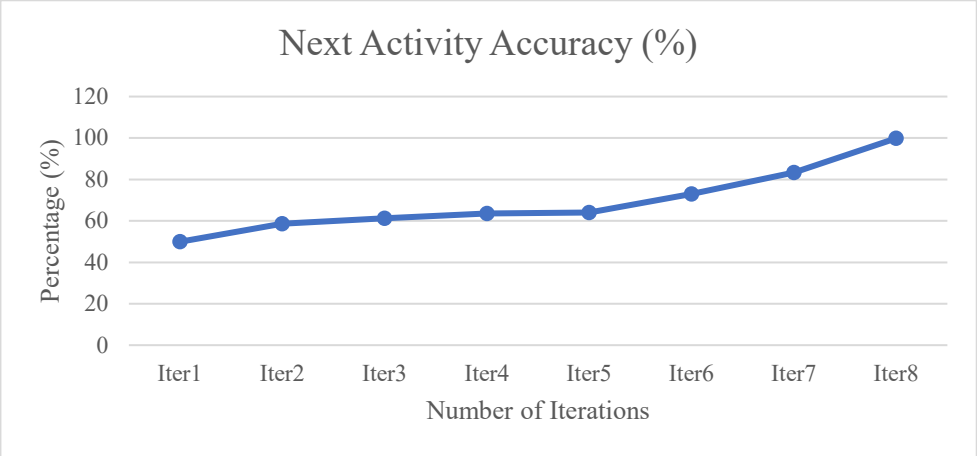


Figure 4-12 Average accuracy of next activity prediction

4.4.2 Objective Prediction

The objective prediction depends on the predicted next activity. The evaluation is somewhat different from the evaluation on the next activity and plan prediction. Objective prediction is only evaluated on the activity sequences that contains activities that lead to more than one objective. Therefore, the total number of samples evaluated from Appendix IV is 220 out of 282. The same performance is executed as previous, instead the predicted objective is being recorded for each observation in the sequences. The correct objective predictions of each observation of the 220 sequences are averaged for each iteration. That is, we can view the average accuracy of each observation from iteration 1 to 8.

doi:10.6342/NTU202300261

To validate our objective prediction module, we see it is appropriate to compare our work against method [56]. We apply both methods onto our plan library and see how the performance is. However, method [56] only considers the current observation when predicting the objective and does not update the knowledge base for the next iteration. Therefore, in Figure 4-13, only the first iteration of our method is shown to compare with the particle filter method. The maximum population of the particles used is 460 with 20 particles representing each objective when initialized. The particles are then updated as new observations are made. As a result, our method has a higher accuracy due to the consideration of next activity that rules out unnecessary objectives when predicting. In addition, Table 4-7 below shows the improved accuracy of the two methods. On the other hand, Figure 4-14 below shows the comparison between the first iteration, which is the system that has not been updated, and the average prediction of each observation across 8 iterations. As shown in the Figure 4-14, the first observation already has a near 60% accuracy in predicting the objective. This means that when the update is processed, next activity would become more accurate, which leads to increased accuracy when filtering out the unwanted objectives. As soon as the observation reaches 4, it can be expected that the predicted objective is considered correct.

Table 4-7 Objective accuracy comparison with Particle Filter

Observations	**Ours (%)**	**Particle Filter (%)**
Obs. 1	41.8531	33.751
Obs. 2	62.5671	51.738
Obs. 3	81.7329	75.1392
Obs. 4	92.1052	83.4419
Obs. 5	100	97.3157
Obs. 6	100	99.1475
Obs. 7	100	100
Obs. 8	100	100

doi:10.6342/NTU202300261

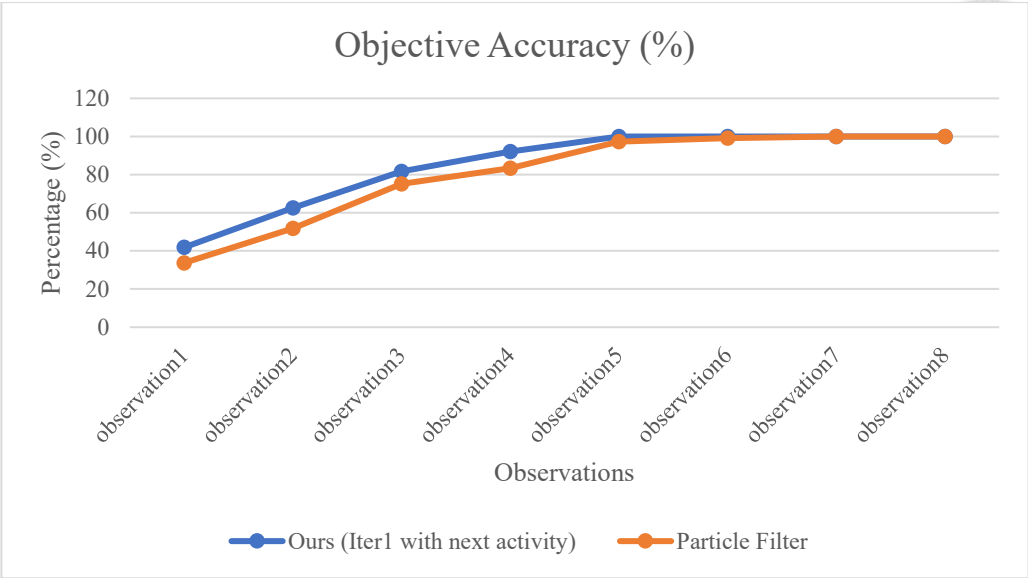


Figure 4-13 Comparison between two objective prediction method

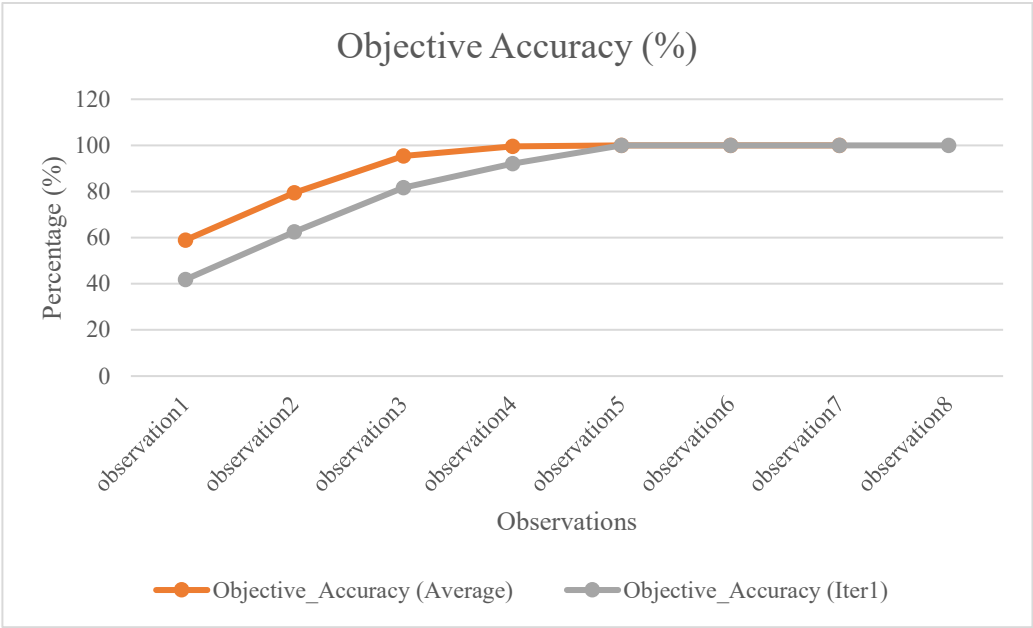


Figure 4-14 Comparison between average accuracy and first iteration accuracy

4.4.3 Plan Prediction

In this section, the accuracy of plan prediction is evaluated. The inputs to plan prediction include the current activity, next predicted activity, and the predicted objective, and the most likely plan is searched through the plan library.

The evaluation setup is the same as next activity prediction with a total number of

doi:10.6342/NTU202300261

282 samples from Appendix IV tested. Each sequence of activities is processed through 8 iterations to view the influence of knowledge update. Same as the objective prediction evaluation, in plan prediction evaluation, the correct predicted plan of each observation in the sample sequences are recorded and the accuracy calculated for each iteration. As viewed in Figure 4-15, it can be observed that the average plan prediction accuracy is lower than the average objective prediction accuracy. This is because the predicted observation may have more than one plans leading to the predicted objective; thus, resulting in predicting the incorrect plan. However, as more observations are made, the accuracy increases. In Figure 4-16, the accuracy of plan prediction across 8 iteration is illustrated to show the effect of updating the knowledge base. It can be seen that as each update is performed the accuracy in each observation prediction increases. As exhibited in Table 4-8, plan prediction accuracy can be low in the first observation. However, a significant increase is apparent when the second observation is made. Moreover, after 4 iteration updates, the plan accuracy is boosted up to 75% when only two observation is made. Even when there are no updates (Iter1), the accuracy is above 72% after the second observation is made.

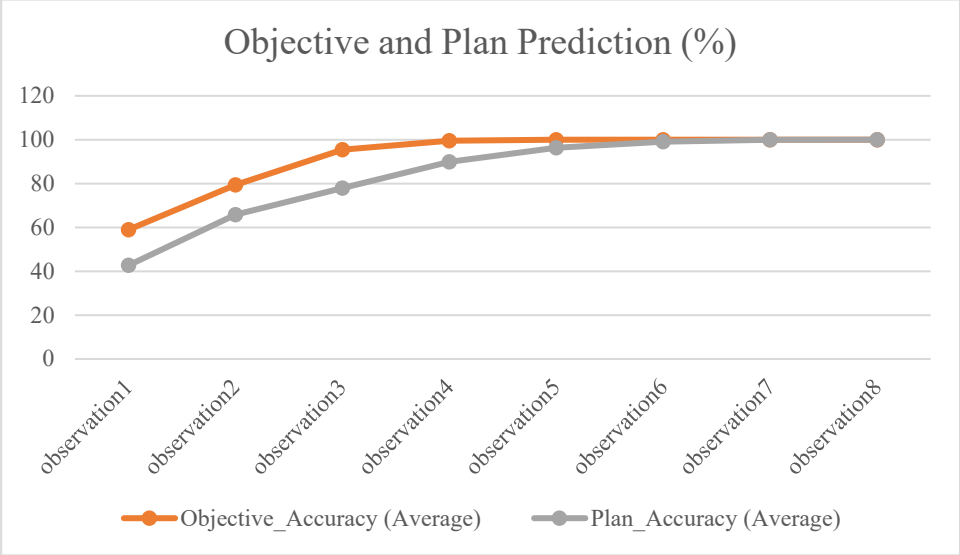


Figure 4-15 Average accuracy between objective and plan prediction

doi:10.6342/NTU202300261

Table 4-8 Exact accuracy of plan prediction across 8 iterations

	Iter1	Iter2	Iter3	Iter4	Iter5	Iter6	Iter7	Iter8
observation1	23.1756%	31.9584%	33.5704%	36.1480%	38.0034%	46.0643%	55.4431%	77.0934%
observation2	72.3201%	73.1534%	74.3553%	74.8183%	75.8342%	76.2786%	76.5009%	77.4268%
observation3	86.9881%	87.6547%	87.6547%	88.2103%	89.1436%	89.6770%	89.9436%	90.2214%
observation4	93.2607%	93.2607%	93.2607%	93.2607%	93.2607%	93.2607%	93.2607%	93.2607%
observation5	98.2456%	98.2456%	98.2456%	98.2456%	98.2456%	98.2456%	98.2456%	98.2456%
observation6	100%	100%	100%	100%	100%	100%	100%	100%
observation7	100%	100%	100%	100%	100%	100%	100%	100%
observation8	100%	100%	100%	100%	100%	100%	100%	100%

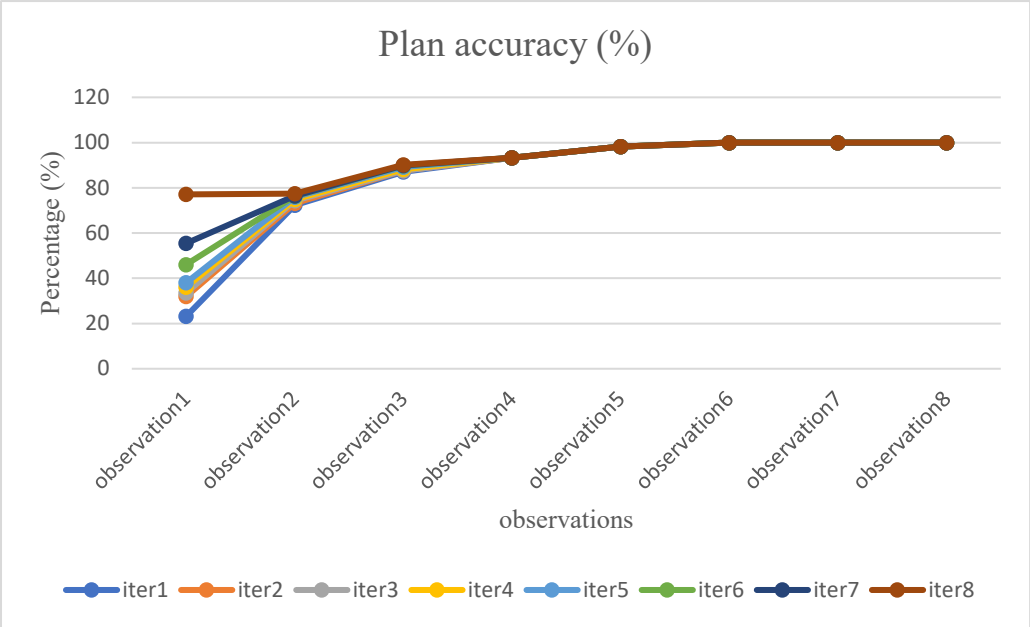


Figure 4-16 Plan prediction accuracy across 8 iterations

4.5 Overall System Result in Real World Robot

In this section, the overall system results, which are tested on our OREO robot, are illustrated for both HAR and Plan recognition. But first, the process of our system performed on our robot will be illustrated to give as sense of how the experiments are conducted. Then, various experiments are performed in each recognition along with different setups. The results and comparisons are shown in the following tables.

doi:10.6342/NTU202300261

4.5.1 System Flow Performed by OREO Robot in Experiment

Before discussing the results of our experiment on OREO robot, we first introduce the flow of how the robot proceeds in the real-world to execute the experiments. Figure 4-17 below shows the flow that will be illustrated in the following.

Locating a human is the first important step before proceeding any analysis. Thus, the robot needs to first find the human in the environment. Next, through "Recognition Reposition", the robot will navigate to the human's vicinity. When the robot arrives near the human and the individual is in the center of the camera image, human activity recognition activates and start processing. After the HAR process is finished, the system checks whether the HAR result is a valid one or not. If not, the system proceeds to check if the perception of the robot needs adjustment. In other words, the system checks whether the human is still near the robot and in the center of the camera image. If the human has moved to another location or is not in the center of the camera image, the system activates "Recognition Reposition" to navigate to another perception point. On the other hand, if no adjustment is needed, the system activates HAR to continue detecting the human activity. If the HAR result is a valid one, the system proceeds to plan recognition for prediction. Next, when the HAR and plan recognition results are available, the system continues to "Robot Response" to allow the robot to make different responses. After the response is been made, the system checks if the observation is finished or not. If not, the system proceeds to check if perception adjustment is necessary and repeats the cycle until the observation is finished. Finally, with the sequence of activities and objective obtained, the system proceeds to the final updating step and updates the knowledge base.

doi:10.6342/NTU202300261

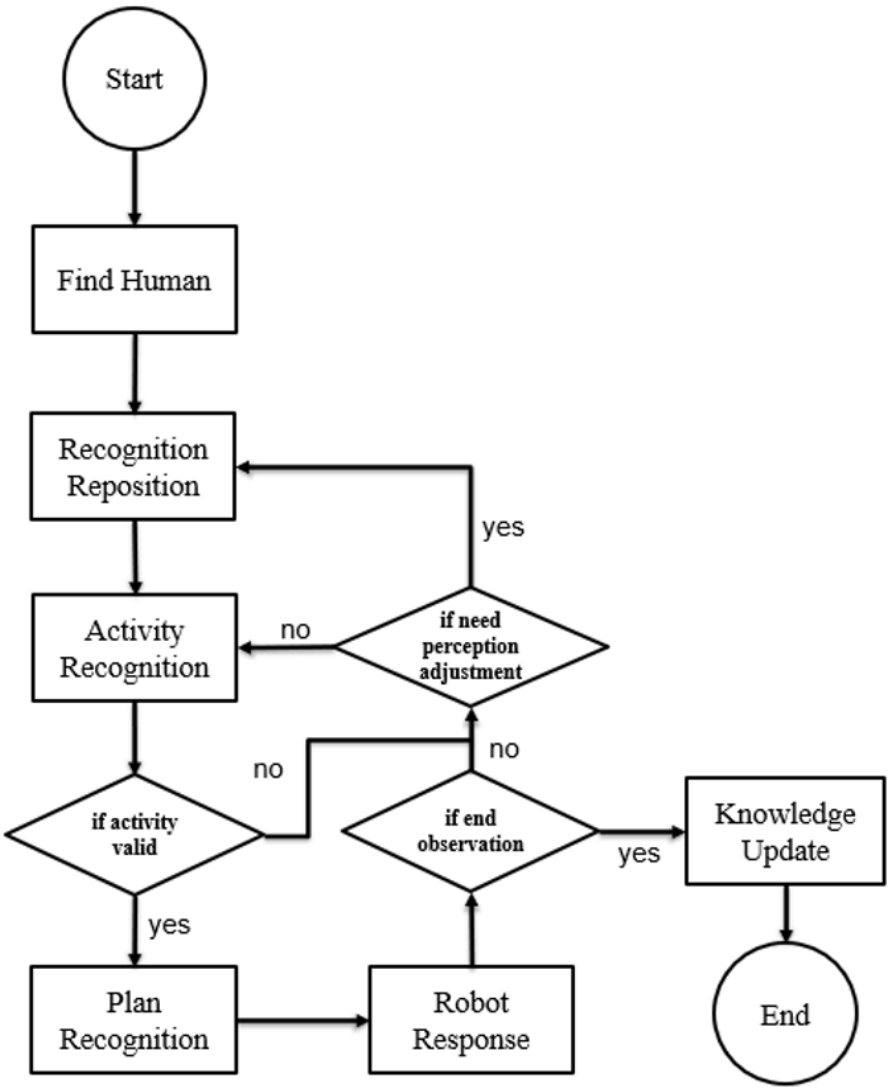


Figure 4-17 System flow of robot in real-world

4.5.2 Location Estimator on OREO Robot

To further test our location estimator in the real world. We ran our estimator on our household robot, OREO, and our Yonglin lab. The experiment is conducted on 4 locations, each with 10 different activities with a total of 200 trials executed (20 trial for each activity). The robot will find, locate, and navigate to the location where the experimental subject is in to execute location estimation while the subject performs an activity. The experiment is to validate the accuracy of the integrated use of the estimator with ResNet50-place365 and our location estimation.

doi:10.6342/NTU202300261

Table 4-9 below shows the accuracy result for the 4 locations. The accuracy of the living room is expected to be high because the detection of the sofas detected in such environment tends to be stable. In our living room, there are 2 sofas; even if the detection of one of the sofas is unstable, the other one can aid the ill detection and stabilize the estimation. This can also explain that the estimated location accuracy is different for the same activity but in different locations. Moreover, the results exhibited in the table which is higher compared to the testing result in Sub-section 4-2 is due to the stability and accuracy of object detection. In the real-world experiment, the detection of dining table, bedroom, or other representative objects tend to be more stable and correct. Whereas, some of the representative objects in the ETRI datasets are detected incorrectly due to their special shapes or similarity to each other.

Table 4-9 Accuracy of location estimation in Yonglin lab

Activities (# from ETRI dataset)	**Locations**	**Activities (# from ETRI dataset)**	**Locations**
	Bedroom		*Study room*
1. Pouring Water into a cup (2)	19/20 (95%)	1. Brushing hair (16)	19/20 (95%)
2. Putting on jacket (18)	20/20 (100%)	2. Putting on/Taking off glasses (21)	20/20 (100%)
3. Vacuuming the floor (23)	19/20 (95%)	3. Reading a book (32)	19/20 (95%)
4. Scrubbing floor with a rag (24)	19/20 (95%)	4. Handwriting (34)	18/20 (90%)
5. Reading a book (32)	18/20 (90%)	5. Talking on the phone (35)	17/20(85%)
6. Talking on the phone (35)	20/20 (100%)	6. Playing mobile phone (36)	18/20 (90%)
7. Doing free hand exercise (41)	20/20 (100%)	7. Using a computer (37)	20/20 (100%)
8. Doing neck rolls (42)	20/20 (100%)	8. Rubbing face with hands (40)	18/20 (90%)
9. Sitting up /Standing up (54)	20/20 (100%)	9. Doing free hand exercise (41)	17/20 (85%)

doi:10.6342/NTU202300261

10. Lying down (55)	17/20 (85%)	10. Sitting up/Standing up (54)	17/20 (85%)
Average	192/200 (96%)	Average	183/200 (91.5%)
	Living room		*Dining room*
1. Eating food with a fork (1)	20/20 (100%)	1. Eating food with a fork (1)	20/20 (100%)
2. Pouring water into a cup (2)	20/20 (100%)	2. Pouring water into a cup (2)	20/20 (100%)
3. Take medicine (3)	20/20 (100%)	3. Take medicine (3)	19/20 (95%)
4. Drinking water (4)	20/20 (100%)	4. Drinking water (4)	20/20 (100%)
5. Putting on jacket (18)	20/20 (100%)	5. Washing hands (11)	20/20 (100%)
6. Vacuuming the floor (23)	18/20 (90%)	6. Putting on jacket (18)	17/20 (85%)
7. Using a remote (31)	18/20 (90%)	7. Wiping off the dining table (25)	18/20 (90%)
8. Talking on the phone (35)	15/20 (75%)	8. Talking on the phone (35)	19/20 (95%)
9. Using a computer (37)	19/20 (95%)	9. Doing free hand exercise (41)	19/20 (95%)
10. Sitting up /Standing up (54)	19/20 (95%)	10. Sitting up /Standing up (54)	19/20 (95%)
Average	196/200 (98%)	Average	191/200 (95.5%)

4.5.3 HAR System on OREO Robot

Our proposed HAR system is implemented on our household robot and tested through various activities, locations, and trials. The experiment is to analyze the integrated accuracy, including whether the robot can successfully locate the human, navigate to the location where the human is in, and output the correct perceived activity. The experiment is conducted over the same 10 activities of each location in Table 4-9 and the results for HAR are shown in Table 4-10. Each location is tested with different activities according to their classification in ETRI dataset.

From experiments and the results in Table 4-10 below, some activities' results have exceptionally high accuracies, such as "Doing free hand exercise", "Putting on jacket", and "Sitting up/Standing up". Turning back to the datasets, we found out that

doi:10.6342/NTU202300261

the skeleton movement of the two activities are massive and distinct. For "Doing free hand exercise", the human arm movement are big as they swing back and forth. Similar for "Putting on jacket", the upper limbs movement is huge. On the other hand, "Sitting up/Standing up", has big movement on the lower limbs of the human body. However, it can be viewed that same activity can have different accuracies across different rooms. This is due to the data differences of locations when they are treated as inputs to train the AL-GCN model. Hence, different locations output different accuracy for the same activity.

Table 4-10 HAR accuracy of 10 various activities in 4 different locations

Activities (# from ETRI dataset)	**Locations**	**Activities (# from ETRI dataset**	**Locations**
	Bedroom		*Study room*
1. Pouring Water into a cup (2)	18/20 (90%)	1. Brushing hair (16)	17/20 (85%)
2. Putting on jacket (18)	17/20 (85%)	2. Putting on/Taking off glasses (21)	17/20 (85%)
3. Vacuuming the floor (23)	17/20 (85%)	3. Reading a book (32)	16/20 (80%)
4. Scrubbing floor with a rag (24)	19/20 (95%)	4. Handwriting (34)	17/20 (85%)
5. Reading a book (32)	16/20 (80%)	5. Talking on the phone (35)	17/20 (85%)
6. Talking on the phone (35)	16/20 (80%)	6. Playing mobile phone (36)	17/20 (85%)
7. Doing free hand exercise (41)	19/20 (95%)	7. Using a computer (37)	18/20 (90%)
8. Doing neck rolls (42)	18/20 (90%)	8. Rubbing face with hands (40)	17/20 (85%)
9. Sitting up /Standing up (54)	18/20 (90%)	9. Doing free hand exercise (41)	20/20 (100%)
10. Lying down (55)	17/20 (85%)	10. Sitting up /Standing up (54)	19/20 (95%)
	Living room		*Dining room*
1. Eating food with a fork (1)	19/20 (95%)	1. Eating food with a fork (1)	18/20 (90%)

doi:10.6342/NTU202300261

2. Pouring water into a cup (2)	17/20 (85%)	2. Pouring water into a cup (2)	16/20 (80%)
3. Take medicine (3)	16/20 (80%)	3. Take medicine (3)	17/20 (85%)
4. Drinking water (4)	18/20 (90%)	4. Drinking water (4)	18/20 (90%)
5. Putting on jacket (18)	19/20 (95%)	5. Washing hands (11)	19/20 (95%)
6. Vacuuming the floor (23)	15/20 (75%)	6. Putting on jacket (18)	20/20 (100%)
7. Using a remote (31)	18/20 (90%)	7. Wiping off the dining table (25)	19/20 (95%)
8. Talking on the phone (35)	15/20 (75%)	8. Talking on the phone (35)	16/20 (80%)
9. Using a computer (37)	19/20 (95%)	9. Doing free hand exercise (41)	19/20 (95%)
10. Sitting up /Standing up (54)	19/20 (95%)	10. Sitting up /Standing up (54)	20/20 (100%)

In addition to the accuracy evaluation of utilizing our proposed HAR model, another evaluation is performed to compare the accuracy between using location information or not. As shown in Table 4-11 below, 10 activities are tested each on 2 of their most performed locations from ETRI dataset. As a result, the accuracy of the 10 activities using location information is higher than those that does not. The 10 activities are later utilized in the evaluation of plan recognition in the next section.

However, to show the improvement more clearly, we compare activities which took places in specific and different locations with similar skeleton movement. Three comparisons are made and tested in three different locations, namely, Entrance, Bedroom, and Living room to show the influence of considering the location information. The experiment is performed with 20 trials, each on the activities with the * symbol at the location mentioned in Table 4-12 below. Note that only the 2 activities that are compared are extracted for testing.

An example in the table is “putting on/taking off shoes” and “Standing up/sitting up” where they have similar skeleton movements. However, their performed location

doi:10.6342/NTU202300261

is somewhat different, where the former is executed at the Entrance whereas the latter at other locations. Thus, by executing the former activity at the entrance when accounting the location gives a higher accuracy, while without the location considered, the skeleton movement can be more likely to be confused with the latter.

Table 4-11 HAR model accuracy w and w/o location information in real-world

Activities	**Location**	**Accuracy w/o location information**	**Accuracy with location information**	**Difference**
Eating food with a fork	(D, L)	32/40 (80%)	37/40 (92.5%)	12.5%
Pouring water	(B, L)	29/40 (72.5%)	34/40 (85%)	12.5%
Take Medicine	(D, L)	25/40 (62.5%)	33/40 (82.5%)	20%
Drinking water	(D, L)	31/40 (77.5%)	36/40 (90%)	12.5%
Putting on jacket	(D, L)	34/40 (85%)	38/40 (95%)	10%
Using remote control	(B, L)	30/40 (75%)	36/40 (90%)	15%
Handwriting	(S, L)	28/40 (70%)	33/40 (82.5%)	12.5%
Using computer	(S, L)	31/40 (77.5%)	37/40 (92.5%)	15%
Stand up/sitting up	(B, L)	34/40 (85%)	38/40 (95%)	10%
Putting on/taking off shoes	(E, L)	26/40 (65%)	34/40 (85%)	20%
Tested over 10 activities, B: Bedroom, L: Living room, D: Dining room, S: Study room, E: Entrance				

Table 4-12 Similar skeleton movement comparison

Activities	**Accuracy w/o location information**	**Accuracy with location information**
Stand up/sitting up vs. Putting on/taking off shoes* (Entrance)	15/20 (75%)	18/20 (90%)
Wiping off the dining table vs. Spreading bedding/folding bedding* (Bedroom)	12/20 (60%)	16/20 (80%)
Brushing teeth vs. talking on the phone* (Living room)	16/20 (80%)	18/20 (90%)
* Tested activity, () Tested location		

doi:10.6342/NTU202300261

4.5.4 Plan Recognition System on OREO Robot

The system is analyzed under the influence of the HAR correctness. The correctness of the HAR of perceiving the human activity is vital to the later performed plan recognition. Thus, we test the plan recognition accuracy with the HAR output as its input. The experiment is to evaluate the accuracy of the system that can successfully output the correct full plan. Here, 4 plans in 4 objectives are experimented and 15 trials are conducted for each plan, as shown in Table 4-13. The plans are tested on the original non-updated version and updated version. The non-updated version uses the two tensors after their initialization with the surveys; whereas, the updated version utilizes tensors after they are updated 7 times with the correct plans.

Table 4-13 Plan accuracy in real-world

Objective	Plan	Accuracy w/o Update	Accuracy full Update
Have Meal	1. Washing hand (Dining room) 2. Eating food with a fork (Dining room)	9/15 (60%)	14/15 (93.3%)
Taking Medicine	1. Pouring water into a cup (Bedroom) 2. Drinking water (Dining room) 3. Taking Medicine (Dining room)	5/15 (33.3%)	11/15 (73.3%)
Going Out Door	1. Sitting up/standing up (Study room) 2. Putting on jacket (Living room) 3. Putting on /taking off shoes (Entrance)	7/15 (46.6%)	12/15 (80%)
Mailing Letter	1. Handwriting (Study Room) 2. Sitting up/standing up (Study room) 3. Putting on jacket (Dining room) 4. Putting on /taking off shoes (Entrance)	5/15 (33.3%)	10/15 (66.6%)

According to the results of Table 4-13 above, the performance of the plan recognition system deployed on robot and real-world scenarios is influenced by the number of activities in a sequence, the accuracy of the activities itself, and whether the activity is under multiple objectives. By comparing objectives “Taking Medicine” and “Going Out Doors”, with the same number of activities, the higher the accuracy

doi:10.6342/NTU202300261

of the activities lead to higher accuracy of the plan recognition whether using location information or not. On the other hand, objectives “Going Out Doors” and “Mailing Letter” have similar but different number of activities, which affects the overall performance.

doi:10.6342/NTU202300261

Chapter 5 Conclusion and Future Works

In this thesis, a system containing location estimation, human activity recognition, plan recognition, and robot response is developed. A human activity recognition system that is based on the location where the detected human is in called, AL-GCN is proposed. With the help of location estimation, the proposed AL-GCN model receives the location information along with the skeleton data to perform a better accuracy recognition. Moreover, the location information and activity output from AL-GCN are used as the input data in the next plan recognition system. The proposed plan recognition system contains knowledge base that records the preference of activities, objectives, and plans for individuals. This is inspired by [33] which keeps rating scores on the plan. However, in this thesis, the scores are used on next activity prediction and objective prediction with two tensors: Loc-NextAct Tensor and Loc-Objective Tensor. With the ability to update knowledge base, the system can adapt to different individuals and predict each human more accurately. In addition, the adaptation also allows the human to change habits in a certain objective and the robot can change its prediction. A habit knowledge base is also created to store the habit information, including the full plan or sub-sequences of activities along with their execution time and duration. Finally, a response module is created to receive predictions from the proposed two systems and gives necessary responses. Furthermore, the whole system is implemented onto our mobile robot called Oreo. Oreo is a homemade robot that runs on omnidirectional wheels and allows us to implement our system on it for real-world applications.

In the experiment result, the performance of AL-GCN model is evaluated on the ETRI 3D Activity dataset. The result shows that the accuracy has improved considering the location information. Especially for the activities that depends highly

doi:10.6342/NTU202300261

on the locations. The more unique the activity is, the more improvement it has. On the other hand, the performance of our plan recognition system is evaluated on the samples that is gathered through survey. It can be shown that as the update performs in each iteration, the average accuracy improves in all predictions. In the real-world experiment, it can also be viewed that the location estimation has a high accuracy due to the detection stability of the representative objects. Moreover, by experimenting and comparing similar skeleton movement, the accuracy of utilizing the location information is higher than those that does not for HAR. Lastly, with the updated version of our matrices, it is more likely to predicted the correct path compared to the version without update.

For future works, we see the advantages that semantic map and scene recognition can bring us. With the aid of scene recognition, the variety of scenes can be considered with higher accuracy. Moreover, with the help of semantic map, the objects can be allocated beforehand and could solve occlusion problem in the observed scene. In regard of the plan recognition, more samples are expected to be gathered to produce a more general plan library and initialized tensors. Moreover, the time factor can be considered in these activities that is being executed to allow the robot to better know when is more suitable to make responses. The information of habit knowledge can be further utilized in various applications and analysis. Lastly, our plan recognition can integrate with Re-ID to allow the system know which human habit is suited for making predictions. To conclude, we provide a robot system to perceive human and their activities to make certain predictions with high success rate in a household environment.

doi:10.6342/NTU202300261

REFERENCES

[1] Ifr., "Service Robots Record: Sales Worldwide up 32%." *IFR International Federation of Robotics*, https://ifr.org/ifr-press-releases/news/service-robots-record-sales-worldwide-up-32.

[2] Mueller, M., Kuhlenkoetter, B., & Nassmacher, R., "Robots in Food Industry Challenges and Chances: Semantic Scholar." *Undefined*, 1 Jan. 1970, https://www.semanticscholar.org/paper/Robots-in-Food-Industry-Challenges-and-Chances-Mueller-Kuhlenkoetter/2197ce8501badc66f645267bc84a9d9fb4f34af0.

[3] Kar, Suman, "Robotics in Healthcare." *2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC)*, 2019, https://doi.org/10.1109/peeic47157.2019.8976668.

[4] Samala, N., Katkam, B. S., Bellamkonda, R. S., & Rodriguez, R. V., "Impact of AI and Robotics in the Tourism Sector: A Critical Insight." *Journal of Tourism Futures*, vol. 8, no. 1, 2020, pp. 73–87., https://doi.org/10.1108/jtf-07-2019-0065.

[5] "Robotic Vacuum Cleaner." *Wikipedia*, Wikimedia Foundation, 15 July 2022, https://en.wikipedia.org/wiki/Robotic_vacuum_cleaner.

[6] G. Grisetti Dept. of Comput. Sci., Grisetti, G., Sci., D. of C., Stachniss, C., Profile, V., Burgard, W., & Metrics, O. M. V. A., "Improved Techniques for Grid Mapping with Rao-Blackwellized Particle Filters." *IEEE Transactions on Robotics*, vol. 23, no. 1, 2007, pp. 34–46.

[7] Rösmann, C., Hoffmann, F., & Bertram, T., "Integrated Online Trajectory Planning and Optimization in Distinctive Topologies." *Robotics and Autonomous Systems*, vol. 88, 2017, pp. 142–153.

[8] Granada, R., Monteiro, J., Gavenski, N., & Meneguzzi, F., "Object-Based Goal Recognition Using Real-World Data." *Advances in Soft Computing*, 2020, pp. 325–337.

[9] Mekruksavanich, S., Promsakon, C., & Jitpattanakul, A., "Location-Based Daily Human Activity Recognition Using Hybrid Deep Learning Network." *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2021.

[10] Snehitha, B., Sreeya, R. S., & Manikandan, V. M., "Human Activity Detection from Still Images Using Deep Learning Techniques." *2021 International Conference on Control, Automation, Power and Signal Processing (CAPS)*, 2021.

doi:10.6342/NTU202300261

[11] Babiker, M., Khalifa, O. O., Htike, K. K., Hassan, A., & Zaharadeen, M., "Automated Daily Human Activity Recognition for Video Surveillance Using Neural Network." *2017 IEEE 4th International Conference on Smart Instrumentation, Measurement and Application (ICSIMA)*, 2017.

[12] Archana, N., and K. Hareesh., "Real-Time Human Activity Recognition Using ResNet and 3d Convolutional Neural Networks." *2021 2nd International Conference on Advances in Computing, Communication, Embedded and Secure Systems (ACCESS)*, 2021.

[13] Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J., "An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[14] Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., & Zheng, N., "View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data." *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[15] Shi, L., Zhang, Y., Cheng, J., & Lu, H., "Non-Local Graph Convolutional Networks for Skeleton-Based Action Recognition." *ArXiv.org*, 4 July 2018, https://arxiv.org/abs/1805.07694v2.

[16] Ren, B., Liu, M., Ding, R., & Liu, H., "A Survey on 3D Skeleton-Based Action Recognition Using Learning Method." *ArXiv.org*, 14 Feb. 2020, https://arxiv.org/abs/2002.05907.

[17] Ke, Q., Bennamoun, M., An, S., Sohel, F., & Boussaid, F., "A New Representation of Skeleton Sequences for 3D Action Recognition." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] Bagate, Asmita, & Medha Shah., "Human Activity Recognition Using RGB-D Sensors." *2019 International Conference on Intelligent Computing and Control Systems (ICCS)*, 2019.

[19] Song, Y.-F., Zhang, Z., Shan, C., & Wang, L., "Richly Activated Graph Convolutional Network for Robust Skeleton-Based Action Recognition." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 5, 2021, pp. 1915–1925.

[20] Heidari, Negar, & Alexandros Iosifidis., "Temporal Attention-Augmented Graph Convolutional Network for Efficient Skeleton-Based Human Action Recognition." *2020 25th International Conference on Pattern Recognition (ICPR)*, 2021.

doi:10.6342/NTU202300261

[21] Li, S., Yi, J., Farha, Y. A., & Gall, J., "Pose Refinement Graph Convolutional Network for Skeleton-Based Action Recognition." *IEEE Robotics and Automation Letters*, vol. 6, no. 2, 2021, pp. 1028–1035.

[22] Wang, H., Yu, B., Li, J., Zhang, L., & Chen, D., "Multi-Stream Interaction Networks for Human Action Recognition." *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, 2022, pp. 3050–3060.

[23] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S., "A Comprehensive Survey on Graph Neural Networks." *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, 2021, pp. 4–24.

[24] Jang, J., Kim, D., Park, C., Jang, M., Lee, J., & Kim, J., "ETRI-Activity3D: A Large-Scale RGB-D Dataset for Robots to Recognize Daily Activities of the Elderly." *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[25] Van-Horenbeke, Franz A., & Angelika Peer., "Activity, Plan, and Goal Recognition: A Review." *Frontiers in Robotics and AI*, vol. 8, 2021.

[26] Avrahami-Zilberbrand, D., Kaminka, G.A., & Zarosim, H., "Fast and Complete Symbolic Plan Recognition: Allowing for Duration, Interleaved Execution, and Lossy Observations." *International Joint Conference on Artificial intelligence (IJCAI)*, 2005.

[27] Geib, Christopher W., & Robert P. Goldman., "A Probabilistic Plan Recognition Algorithm Based on Plan Tree Grammars." *Artificial Intelligence*, vol. 173, no. 11, 2009, pp. 1101–1132.

[28] Marchetta, M. G., & R. Q. Forradellas., "A New Model for Automatic Generation of Plan Libraries for Plan Recognition". *Brazilian Journal of Operations & Production Management*, vol. 3, no. 2, Feb. 2010, pp. 5-19, https://bjopm.org.br/bjopm/article/view/BJV3N2_2006_P1.

[29] Geib, C.W., Maraist, J., & Goldman, R.P., "A New Probabilistic Plan Recognition Algorithm Based on String Rewriting." *ICAPS*, 2008.

[30] Granada, R.L., Pereira, R., Monteiro, J., Barros, R.C., Ruiz, D.D., & Meneguzzi, F., "Hybrid Activity and Plan Recognition for Video Streams." *AAAI Workshops*, 2017.

[31] Shimada, A., Kondo, K., Deguchi, D., Morin, G., & Stern, H., "Kitchen Scene Context Based Gesture Recognition: A Contest in ICPR2012." *Advances in Depth Image Analysis and Applications*, 2013, pp. 168–185.

[32] Meng, L., Miao, C., & Leung, C., "Towards Online and Personalized Daily Activity Recognition, Habit Modeling, and Anomaly Detection for the Solitary Elderly through Unobtrusive Sensing." *Multimedia Tools and Applications*, vol. 76, no. 8, 2016, pp. 10779–10799.

doi:10.6342/NTU202300261

[33] Zhuo, Hankz., "Human-Aware Plan Recognition." *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[34] "Adjacency Matrix." *Wikipedia*, Wikimedia Foundation, 7 June 2022, https://en.wikipedia.org/wiki/Adjacency_matrix.

[35] "Degree Matrix." *Wikipedia*, Wikimedia Foundation, 26 Oct. 2021, https://en.wikipedia.org/wiki/Degree_matrix.

[36] "Laplacian Matrix." *Wikipedia*, Wikimedia Foundation, 11 June 2022, https://en.wikipedia.org/wiki/Laplacian_matrix.

[37] J. Bruna, W. Zaremba, A. Szlam, Y.L. Cun, "Spectral Networks and Deep Locally Connected Networks on Graphs," *International Conference on Learning Representations (ICLR)*, 2014.

[38] Kipf, T. N., & Welling, M., "Semi-Supervised Classification With Graph Convolutional Networks," *International Conference on Learning Representations (ICLR)*, 2017.

[39] Yan, S., Xiong, Y., & Lin, D., "Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition," *Association for the Advancement of AI (AAAI)*, 2018.

[40] Shi, L., Zhang, Y., Cheng, J., & Lu, H., "Two-Stream Adaptive Graph Convolutional Networks for Skeleton-Based Action Recognition," *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Girshick, R., Donahue, J., Darrell, T., & Malik, J., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[42] Girshick, Ross., "Fast R-CNN." *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015.

[43] Ren, S., He, K., Girshick, R., & Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, 2017, pp. 1137–1149.

[44] C. Liu, Y. Tao, J. Liang, K. Li & Y. Chen., "Object Detection Based on Yolo Network." *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 2018.

[45] Shuai, Qianjun, & Xingwen Wu., "Object Detection System Based on SSD Algorithm." *2020 International Conference on Culture-Oriented Science & Technology (ICCST)*, 2020.

[46] Del Prete, R., Graziano, M. D., & Renga, A., "RetinaNet: A Deep Learning Architecture to Achieve a Robust Wake Detector in SAR Images." *2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI)*, 2021.

doi:10.6342/NTU202300261

[47] Redmon, Joseph, & Ali Farhadi., "Yolo9000: Better, Faster, Stronger." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[48] Li, S., Tao, F., Shi, T., & Kuang, J., "Improvement of yolov3 Network Based on Roi." *2019 IEEE 4th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, 2019.

[49] J. Redmon, S. Divvala, R. Girshick, & A. Farhadi, "YOLOv4: Optimal Speed and Accuracy of Object Detection," *arXiv:1506.02640 [cs.CV]*, 2015.

[50] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J., "Path Aggregation Network for Instance Segmentation." *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[51] "ROS/Introduction – ROS Wiki," Accessed on: April. 20th, 2021 [Online]. Available: http://wiki.ros.org/ROS/Introduction.

[52] "Robot Operating System." *Wikipedia*, Wikimedia Foundation, 5 Apr. 2022, https://en.wikipedia.org/wiki/Robot_Operating_System.

[53] Tome, D., Russell, C., & Agapito, L., "Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[54] Stereo Labs, "zed-ros-wrapper" [Source code] Available: https://github.com/stereolabs/zed-ros-wrapper, 2021.

[55] G. Hidalgo, Z. Cao, T. Simon, S.E. Wei, Y. Raaj, H. Joo, Y. Sheikh., "Openpose," Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose.

[56] Massardi, J., M. Gravel, & Éric Beaudry., "Error-Tolerant Anytime Approach to Plan Recognition Using a Particle Filter". *Proceedings of the International Conference on Automated Planning and Scheduling*, vol. 29, no. 1, May 2021, pp. 284-91, https://ojs.aaai.org/index.php/ICAPS/article/view/3490.

[57] "Terminal and Nonterminal Symbols." *Wikipedia*, Wikimedia Foundation, 20 Apr. 2022, https://en.wikipedia.org/wiki/Terminal_and_nonterminal_symbols.

[58] Cappe, O., Godsill, S. J., & Moulines, E., "An Overview of Existing Methods and Recent Advances in Sequential Monte Carlo." *Proceedings of the IEEE*, vol. 95, no. 5, 2007, pp. 899–924.

[59] Zhang, B., Liu, J., & Chen, H., "AMCL Based Map Fusion for Multi-Robot Slam with Heterogenous Sensors." *2013 IEEE International Conference on Information and Automation (ICIA)*, 2013.

[60] Massardi, J., Gravel, M., & Beaudry, E., "Parc: A Plan and Activity Recognition Component for Assistive Robots." *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

doi:10.6342/NTU202300261

[61] Oh, Jung H., & Heung-Jae Lee., “Object-Based Place Recognition for Mobile Robots Using Laplace's Rule of Succession.” *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, 2020.

[62] Handa, A., Patraucean, V., Badrinarayanan, V., Stent, S., & Cipolla, R., “Understanding Realworld Indoor Scenes with Synthetic Data.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[63] Hernandez, A. C., Gomez, C., Barber, R., & Martinez Mozos, O., “Object-Based Probabilistic Place Recognition for Indoor Human Environments.” *2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO)*, 2018.

[64] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A., “Sun Database: Large-Scale Scene Recognition from Abbey to Zoo.” *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010.

[65] Song, Y.-F., Zhang, Z., Shan, C., & Wang, L., “Constructing Stronger and Faster Baselines for Skeleton-Based Action Recognition.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, pp. 1–1.

[66] He, K., Zhang, X., Ren, S., & Sun, J., “Deep Residual Learning for Image Recognition.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[67] Leotta, F., Mecella, M., Sora, D., & Catarci, T., “Surveying Human Habit Modeling and Mining Techniques in Smart Spaces.” *Future Internet*, vol. 11, no. 1, 2019, p. 23.

[68] Shahroudy, A., Liu, J., Ng, T.-T., & Wang, G., “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[69] StereoLabs, “ZED 2,” Available: https://www.stereolabs.com/zed-2/.

[70] “Triangulation.” *Wikipedia*, Wikimedia Foundation, 18 July 2022, https://en.wikipedia.org/wiki/Triangulation.

[71] Byeon, Y.-H., Kim, D., Lee, J., & Kwak, K.-C., “Ensemble Three-Stream RGB-s Deep Neural Network for Human Behavior Recognition Under Intelligent Home Service Robot Environments.” *IEEE Access*, vol. 9, 2021, pp. 73240–73250.

[72] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, & A. Torralba, “Places: A 10 Million Image Database for Scene Recognition,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 6, pp. 1452-1464, 1 June 2018.

[73] Lee, Jaeryoung, & Nicholas Melo., “Habit Representation Based on Activity Recognition.” *Sensors (Basel, Switzerland)* vol. 20,7 1928. 30 Mar. 2020.

doi:10.6342/NTU202300261

Appendix I

Activities in ETRI Activity 3D dataset [24]

ID	Activity	ID	Activity
1	Eating food with a fork	29	Hanging out laundry
2	Pouring water into a cup	30	Looking around for something
3	Taking medicine	31	Using a remote control
4	Drinking water	32	Reading a book
5	Putting food in the fridge/taking food from the fridge	33	Reading a newspaper
6	Trimming vegetables	34	Handwriting
7	Peeling fruit	35	Talking on the phone
8	Using a gas stove	36	Playing with a mobile phone
9	Cutting vegetable on the cutting board	37	Using a computer
10	Brushing teeth	38	Smoking
11	Washing hands	39	Clapping
12	Washing face	40	Rubbing face with hands
13	Wiping face with a towel	41	Doing freehand exercise
14	Putting on cosmetics	42	Doing neck roll exercise
15	Putting on lipstick	43	Massaging a shoulder oneself
16	Brushing hair	44	Taking a bow
17	Blow drying hair	45	Talking to each other
18	Putting on a jacket	46	Handshaking
19	Taking off a jacket	47	Hugging each other
20	Putting on/taking off shoes	48	Fighting each other
21	Putting on/taking off glasses	49	Waving a hand
22	Washing the dishes	50	Flapping a hand up and down (beckoning)
23	Vacuuming the floor	51	Pointing with a finger
24	Scrubbing the floor with a rag	52	Opening the door and walking in
25	Wiping off the dining table	53	Fallen on the floor
26	Rubbing up furniture	54	Sitting up/standing up
27	Spreading bedding/folding bedding	55	Lying down
28	Washing a towel by hands		

doi:10.6342/NTU202300261

Appendix II

Representative Objects and Affordance Objects from YOLO v4 object detection in 6 locations

Locations	Representative Objects	Affordance Objects
Bedroom	1. Bed	1. Backpack 2. Handbag 3. Suitcase 4. Potted plant 5. Teddy bear 6. Laptop
Study Room	1. Monitor (Number of monitors detected)	1. Backpack 2. Handbag 3. Suitcase 4. Bench 5. Chair 6. Laptop 7. Mouse 8. Remote 9. Keyboard
Living Room	1. Sofa, TV monitor	1. Backpack 2. Handbag 3. Suitcase 4. Cup 5. Chair 6. Potted plant 7. Laptop 8. Remote
Dining Room/kitchen	1. Dining table 2. Oven 3. Refrigerator	1. Wine glass 2. Cup 3. Fork 4. Knife 5. Spoon 6. Bowl 7. Chair 8. Broccoli 9. Cake 10. Laptop 11. Microwave 12. Toaster 13. Sink
Entrance	1. Umbrella	-
Bathroom	1. Toilet 2. Toothbrush	1. Cup 2. Sink 3. Hair drier

doi:10.6342/NTU202300261

Appendix III

Following is the survey that has been conducted for plan library construction and tensor initialization.

Activities for different objectives

首先，非常感謝您抽空填這個表單！

這個表單總共有23個區段，各區端中包含：

1. 所需達到的目標
2. 達到目標所需的各種行為
3. 各種行為所執行的地點

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

vincent0131lin@gmail.com (未分享) 切換帳戶

性別

- 男性
- 女性

年齡

- 18 歲以下
- 18 - 24 歲
- 25 - 34 歲
- 35 - 44 歲
- 45 - 54 歲
- 55 - 64 歲
- 65 歲以上

doi:10.6342/NTU202300261

目標1

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Relieve Stress_緩解壓力

	clapping_拍手	doing free hand exercise_做手部運動	doing neck roll exercise_做頸部運動	massaging a shoulder oneself_按摩肩膀	smoking_抽菸
step1	○	○	○	○	○
step2	○	○	○	○	○
step3	○	○	○	○	○
step4	○	○	○	○	○
step5	○	○	○	○	○

請填入Relieve Stress_緩解壓力中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○

目標2

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Taking medicine_服藥

	taking medicine_吃藥	drinking water_喝水	pouring water into a cup_將水倒入杯中
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Taking medicine_服藥中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標3

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Freshening up_梳洗

	pouring water into a cup_將水倒入杯中	drinking water_喝水	brushing teeth_刷牙	brushing hair_梳頭髮	blow drying hair_吹頭髮	wiping face with a towel_用毛巾擦拭臉頰	washing face_洗臉	putting on/taking off glasses_戴上或拿下眼鏡
step1	○	○	○	○	○	○	○	○
step2	○	○	○	○	○	○	○	○
step3	○	○	○	○	○	○	○	○
step4	○	○	○	○	○	○	○	○
step5	○	○	○	○	○	○	○	○
step6	○	○	○	○	○	○	○	○
step7	○	○	○	○	○	○	○	○
step8	○	○	○	○	○	○	○	○

請填入Freshening up_梳洗中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○
step6	○	○	○	○	○	○
step7	○	○	○	○	○	○
step8	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標4

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Taking a rest_休息

	putting on/taking off glasses_戴上或拿下眼鏡	rubbing face with hands_用手揉臉部	taking off jacket_脫下外套	taking off shoes_脫下鞋子	opening the door and walk in_打開門後進入	lying down_躺下
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○
step6	○	○	○	○	○	○

請填入Taking a rest_休息中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○
step6	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標6

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

having a meal_用餐

	washing hand_洗手	eating food with a fork_用叉子吃食物
step1	○	○
step2	○	○

請填入having a meal_用餐中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○

目標5

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Cooking_煮飯

	putting food in the fridge/taking food from the fridge_從冰箱放入或拿出食物	washing hand_洗手	trimming vegetable_處理蔬菜	cutting vegetable on the cutting board_在砧板上切蔬菜	using a gas stove_用瓦斯爐
step1	○	○	○	○	○
step2	○	○	○	○	○
step3	○	○	○	○	○
step4	○	○	○	○	○
step5	○	○	○	○	○

請填入Cooking_煮飯中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標7

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Eating fruit_吃水果

	putting food in the fridge/taking food_從冰箱放入或拿出食物	washing hand_洗手	peeling fruit_削水果
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Eating fruit_吃水果中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

目標8

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Afterward cleaning (meal)_餐後清潔

	washing dishes_洗碗	wiping off the dining table_擦拭桌子	washing a towel by hands_用手洗毛巾
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Afterward cleaning (meal)_餐後清潔中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標10

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Mailing letter_寄信

	putting on/taking off glasses_戴上或拿下眼鏡	handwriting_手寫信	sitting up/standing up_做起來或站起來	putting on jacket_穿上外套	putting on/taking off shoes_穿上或脫下鞋子
step1	○	○	○	○	○
step2	○	○	○	○	○
step3	○	○	○	○	○
step4	○	○	○	○	○
step5	○	○	○	○	○

請填入Mailing letter_寄信中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○

目標9

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Doing chores_做家事

	hanging out laundry_曬衣服	vacuuming the floor_吸地板	scrubbing the floor with a rag_用抹布擦拭地板	rubbing up furniture_擦拭家具	spreading/folding bedding_鋪床單或收床單	washing dishes_洗碗
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○
step6	○	○	○	○	○	○

請填入Doing chores_做家事中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○
step6	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標11

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Going outdoor_出門

	sitting up/standing up_做起來或站起來	putting on jacket_穿上外套	putting on/taking off shoes_穿上或脫下鞋子
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Going outdoor_出門中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

目標12

請在該區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Meeting with people_與人碰面

	putting on cosmetics_化妝	putting on lipstick_擦口紅	sitting up/standing up_做起來或站起來	putting on jacket_穿上外套	putting on/taking off shoes_穿上或脫下鞋子
step1	○	○	○	○	○
step2	○	○	○	○	○
step3	○	○	○	○	○
step4	○	○	○	○	○
step5	○	○	○	○	○

請填入Meeting with people_與人碰面中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○
step5	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標14

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Greeting_打招呼

	Handshake_握手	hugging each other_擁抱	waving a hand_揮手	taking a bow_鞠躬
step1	○	○	○	○
step2	○	○	○	○
step3	○	○	○	○
step4	○	○	○	○

請填入Greeting_打招呼中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○

目標13

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Doing make up_化妝

	putting on cosmetics_化妝	putting on lipstick_擦口紅
step1	○	○
step2	○	○

請填入Doing make up_化妝中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標16

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Using computer _用電腦

	putting on/taking off glasses_戴上或拿下眼鏡	using a computer_用電腦
step1	○	○
step2	○	○

請填入Using computer _用電腦中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○

目標15

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Saying goodbye_說再見

	Handshake_握手	hugging each other_擁抱	waving a hand_揮手	taking a bow_鞠躬
step1	○	○	○	○
step2	○	○	○	○
step3	○	○	○	○
step4	○	○	○	○

請填入Saying goodbye_說再見中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標17

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Watch TV_看電視

	putting on/taking off glasses_戴上或拿下眼鏡	using remote control_使用遙控器
step1	○	○
step2	○	○

請填入Watch TV_看電視中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○

目標18

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Relaxing_放鬆

	reading a book_讀書	reading a newspaper_讀報紙	play with mobile phone_使用手機	putting on/taking off glasses_戴上或拿下眼鏡
step1	○	○	○	○
step2	○	○	○	○
step3	○	○	○	○
step4	○	○	○	○

請填入Relaxing_放鬆中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標20

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Sleep_睡覺

	putting on/taking off glasses_戴上或拿下眼鏡	lying down_躺下	spreading bedding/folding bedding_鋪床單/收床單
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Sleep_睡覺中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

目標19

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Calling someone_打給某人

	flapping a hand up and down_招手	talk on the phone_講電話	play with mobile phone_使用手機
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Calling someone_打給某人中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標22

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Harm_傷害

	pointing with a finger_用手指點	fighting each other_互相打架
step1	○	○
step2	○	○

請填入Harm_傷害中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○

目標21

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Turn on AC_開啟空調

	looking around for something_尋找某樣東西	using remote control_使用遙控器	taking off a jacket_脫外套
step1	○	○	○
step2	○	○	○
step3	○	○	○

請填入Turn on AC_開啟空調中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○

doi:10.6342/NTU202300261

目標23

請在每個區段以順序性(Step1~n)選擇各個行為和執行的地點，(注意：並非一定要選擇所有行為，但最少請選擇一個)

Call for help_呼叫救援

	fallen on the floor_跌倒在地上	playing with a mobile phone_使用手機	talk on the phone_講電話	flapping a hand up and down_招手
step1	○	○	○	○
step2	○	○	○	○
step3	○	○	○	○
step4	○	○	○	○

請填入Call for help_呼叫救援中每個活動發生的地點

	bedroom_臥室	Study room_書房	Living room_客廳	Dining room/kitchen_餐廳/廚房	Entrance_玄關	Bathroom_浴室
step1	○	○	○	○	○	○
step2	○	○	○	○	○	○
step3	○	○	○	○	○	○
step4	○	○	○	○	○	○

doi:10.6342/NTU202300261

Appendix IV

Below are the samples used from survey for plan recognition evaluation. There are a total of 282 samples. Note that samples from Objective 1, 8, 14, 15, and 22, which are colored in brown, are not utilized in objective prediction. The remaining samples, which are colored in green, on the other hand, are utilized in all predictions.

Objective1: Relieve stress (20 samples)

Sample1:
activity="clapping" location="Study room"
activity="doing free hand exercise" location="Living room"
activity="doing neck roll exercise" location="Living room"
activity="massaging a shoulder oneself" location="bedroom"
activity="smoking" location="Entrance" level="5"

Sample2:
activity="massaging a shoulder oneself" location="Living room"
activity="doing neck roll exercise" location="bedroom"
activity="doing free hand exercise" location="bedroom"

Sample3:
activity="doing neck roll exercise" location="Living room"
activity="clapping" location="Living room"
activity="massaging a shoulder oneself" location="Study room"
activity="doing free hand exercise" location="Living room"

Sample4:
activity="doing free hand exercise" location="Entrance"
activity="doing neck roll exercise" location="Living room"
activity="massaging a shoulder oneself" location="bedroom"
activity="clapping" location="bedroom"
activity="massaging a shoulder oneself" location="Living room"

Sample5:
activity="doing free hand exercise" location="bedroom"
activity="doing neck roll exercise" location="Study room"
activity="massaging a shoulder oneself" location="Living room"
activity="doing free hand exercise" location="Living room"
activity="doing neck roll exercise" location="Living room"

Sample6:
activity="doing neck roll exercise" location="Living room"
activity="doing free hand exercise" location="Study room"
activity="massaging a shoulder oneself" location="bedroom"
activity="clapping" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample7:

activity=”doing neck roll exercise” location=”Living room”
activity=”doing free hand exercise” location=”Study room”
activity=”massaging a shoulder oneself” location=”Entrance”
activity=”clapping” location=”bedroom”

Sample8:

activity=”doing free hand exercise” location=”bedroom”
activity=”doing neck roll exercise” location=”Study room”
activity=”massaging a shoulder oneself” location=”Living room”
activity=”doing neck roll exercise” location=”Dining room/kitchen”
activity=”clapping” location=”Living room”

Sample9:

activity=”doing free hand exercise” location=”bedroom”
activity=”doing neck roll exercise” location=”Bathroom”
activity=”massaging a shoulder oneself” location=”Living room”
activity=”clapping” location=”Living room”

Sample10:

activity=”doing free hand exercise” location=”Living room”
activity=”doing neck roll exercise” location=”Living room”
activity=”doing free hand exercise” location=”Dining room/kitchen”
activity=”clapping” location=”Study room”
activity=”doing neck roll exercise” location=”Living room”

Sample11:

activity=”doing neck roll exercise” location=”bedroom”
activity=”massaging a shoulder oneself” location=”bedroom”
activity=”doing free hand exercise” location=”bedroom”

Sample12:

activity=”doing neck roll exercise” location=”Study room”
activity=”massaging a shoulder oneself” location=”Study room”>
activity=”doing free hand exercise” location=”bedroom”

Sample13:

activity=”smoking” location=”Living room”
activity=”doing neck roll exercise” location=”bedroom”
activity=”massaging a shoulder oneself” location=”bedroom”
activity=”doing free hand exercise” location=”bedroom”
activity=”clapping” location=”bedroom”

Sample14:

activity=”clapping” location=”Living room”
activity=”doing free hand exercise” location=”bedroom”
activity=”doing neck roll exercise” location=”Living room”
activity=”doing free hand exercise” location=”bedroom”

doi:10.6342/NTU202300261

activity=”clapping” location=”Living room”

Sample15:
activity=”massaging a shoulder oneself” location=”bedroom”
activity=”doing neck roll exercise” location=”Living room”
activity=”doing free hand exercise” location=”Study room”
activity=”clapping” location=”Dining room/kitchen”
activity=”smoking” location=”Entrance”

Sample16:
activity=”doing free hand exercise” location=”Study room”
activity=”doing neck roll exercise” location=”Study room”
activity=”massaging a shoulder oneself” location=”Study room”

Sample17:
activity=”doing free hand exercise” location=”bedroom”
activity=”doing neck roll exercise” location=”Living room”
activity=”clapping” location=”Study room”
activity=”massaging a shoulder oneself” location=”Study room”

Sample18:
activity=”clapping” location=”Study room”
activity=”doing free hand exercise” location=”Study room”
activity=”doing neck roll exercise” location=”Living room”
activity=”massaging a shoulder oneself” location=”Living room”

Sample19:
activity=”clapping” location=”Living room”
activity=”doing free hand exercise” location=”Living room”
activity=”doing neck roll exercise” location=”bedroom”
activity=”massaging a shoulder oneself” location=”Study room”
activity=”smoking” location=”Entrance”

Sample20:
activity=”clapping” location=”bedroom”
activity=”doing free hand exercise” location=”bedroom”
activity=”doing neck roll exercise” location=”bedroom”
activity=”massaging a shoulder oneself” location=”bedroom”

Objective2: Take Medicine (13 samples)

Sample1:
activity="pouring water into a cup" location="Dining room/kitchen"
activity="taking medicine" location="Living room"
activity="drinking water" location="Living room"

Sample2:
activity="pouring water into a cup" location="Dining room/kitchen"
activity="drinking water" location="Living room"

doi:10.6342/NTU202300261

activity="taking medicine" location="Living room"

Sample3:
activity="pouring water into a cup" location="Living room"
activity="taking medicine" location="bedroom"
activity="drinking water" location="Dining room/kitchen"

Sample4:
activity="pouring water into a cup" location="Dining room/kitchen"
activity="drinking water" location="Dining room/kitchen"
activity="taking medicine" location="Dining room/kitchen"

Sample5:
activity="pouring water into a cup" location="Living room"
activity="drinking water" location="bedroom"
activity="taking medicine" location="Dining room/kitchen"

Sample6:
activity="pouring water into a cup" location="Living room"
activity="taking medicine" location="Living room"
activity="drinking water" location="Living room"

Sample7:
activity="pouring water into a cup" location="bedroom"
activity="drinking water" location="Dining room/kitchen"
activity="taking medicine" location="Dining room/kitchen"

Sample8:
activity="taking medicine" location="Dining room/kitchen"
activity="pouring water into a cup" location="Study room"
activity="drinking water" location="bedroom"

Sample9:
activity="pouring water into a cup" location="bedroom"
activity="drinking water" location="bedroom"
activity="taking medicine" location="bedroom"

Sample10:
activity="taking medicine" location="Living room"
activity="drinking water" location="bedroom"
activity="pouring water into a cup" location="Living room"

Sample11:
activity="drinking water" location="Dining room/kitchen"
activity="taking medicine" location="bedroom"
activity="pouring water into a cup" location="Living room"

doi:10.6342/NTU202300261

Sample12:

activity="pouring water into a cup" location="bedroom"
activity="taking medicine" location="bedroom"
activity="drinking water" location="bedroom"

Sample13:

activity="pouring water into a cup" location="Living room"
activity="taking medicine" location="bedroom"
activity="drinking water" location="bedroom"

Objective3: Freshen Up (24 samples)

Sample1:

activity="putting on/taking off glasses" location="Living room"
activity="brushing hair" location="Living room"
activity="washing face" location="Bathroom"
activity="wiping face with a towel" location="Bathroom"
activity="brushing teeth" location="Bathroom"
activity="pouring water into a cup" location="Dining room/kitchen"
activity="drinking water" location="Dining room/kitchen"

Sample2:

activity="pouring water into a cup" location="Bathroom"
activity="brushing teeth" location="Bathroom"
activity="drinking water" location="Bathroom"
activity="washing face" location="Bathroom"
activity="wiping face with a towel" location="Bathroom"
activity="putting on/taking off glasses" location="Bathroom"
activity="brushing hair" location="Bathroom"
activity="blow drying hair" location="Bathroom"

Sample3:

activity="putting on/taking off glasses" location="Living room"
activity="washing face" location="Bathroom"
activity="wiping face with a towel" location="Bathroom"
activity="brushing teeth" location="Bathroom"
activity="brushing hair" location="Bathroom"
activity="blow drying hair" location="Bathroom"
activity="pouring water into a cup" location="bedroom"
activity="drinking water" location="bedroom"

Sample4:

activity="pouring water into a cup" location="Bathroom"
activity="brushing teeth" location="Bathroom"
activity="washing face" location="Bathroom"
activity="wiping face with a towel" location="Bathroom"
activity="putting on/taking off glasses" location="bedroom"

doi:10.6342/NTU202300261

Sample5:

activity=”putting on/taking off glasses” location=”bedroom”
activity=”pouring water into a cup” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”brushing hair” location=”bedroom”
activity=”drinking water” location=”bedroom”

Sample6:

activity=”pouring water into a cup” location=”bedroom”
activity=”drinking water” location=”bedroom”
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”brushing hair” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”

Sample7:

activity=”wiping face with a towel” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”blow drying hair” location=”bedroom”

Sample8:

activity=”pouring water into a cup” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”brushing hair” location=”Bathroom”
activity=”drinking water” location=”Dining room/kitchen”

Sample9:

activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”pouring water into a cup” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”

Sample10:

activity=”pouring water into a cup” location=”Bathroom”
activity=”drinking water” location=”Entrance”
activity=”brushing teeth” location=”Dining room/kitchen”
activity=”brushing hair” location=”Living room”
activity=”blow drying hair” location=”Study room”
activity=”wiping face with a towel” location=”bedroom”
activity=”washing face” location=”Study room”
activity=”putting on/taking off glasses” location=”Living room”

doi:10.6342/NTU202300261

Sample11:

activity=”putting on/taking off glasses” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”blow drying hair” location=”Living room”
activity=”brushing hair” location=”Living room”
activity=”pouring water into a cup” location=”Living room”
activity=”drinking water” location=”Living room”

Sample12:

activity=”drinking water” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”

Sample13:

activity=”putting on/taking off glasses” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”drinking water” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”brushing hair” location=”Bathroom”

Sample14:

activity=”putting on/taking off glasses” location=”bedroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”

Sample15:

activity=”pouring water into a cup” location=”bedroom”
activity=”brushing teeth” location=”Living room”
activity=”pouring water into a cup” location=”bedroom”
activity=”brushing teeth” location=”Living room”
activity=”pouring water into a cup” location=”bedroom”
activity=”brushing teeth” location=”Living room”
activity=”pouring water into a cup” location=”bedroom”
activity=”brushing teeth” location=”Living room”

Sample16:

activity=”putting on/taking off glasses” location=”Bathroom”
activity=”pouring water into a cup” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”

doi:10.6342/NTU202300261

activity=”brushing hair” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”drinking water” location=”Bathroom”

Sample17:
activity=”putting on/taking off glasses” location=”bedroom”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”bedroom”
activity=”blow drying hair” location=”bedroom”
activity=”brushing hair” location=”bedroom”
activity=”pouring water into a cup” location=”Living room”
activity=”drinking water” location=”Living room”

Sample18:
activity=”brushing teeth” location=”Bathroom”
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”pouring water into a cup” location=”bedroom”
activity=”drinking water” location=”bedroom”

Sample19:
activity=”brushing hair” location=”bedroom”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”bedroom”
activity=”blow drying hair” location=”bedroom”>
activity=”pouring water into a cup” location=”Dining room/kitchen”
activity=”drinking water” location=”Dining room/kitchen”
activity=”putting on/taking off glasses” location=”bedroom”

Sample20:
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”pouring water into a cup” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”drinking water” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom
activity=”blow drying hair” location=”Bathroom”
activity=”brushing hair” location=”Bathroom”

Sample21:
activity=”drinking water” location=”Living room”
activity=”brushing teeth” location=”Bathroom”
activity=”washing face” location=”Bathroom”

doi:10.6342/NTU202300261

activity=”wiping face with a towel” location=”Bathroom”

Sample22:
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”blow drying hair” location=”Bathroom”
activity=”brushing hair” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”

Sample23:
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”washing face” location=”Bathroom”
activity=”wiping face with a towel” location=”Bathroom”
activity=”blow drying hair” location=”bedroom”
activity=”brushing hair” location=”bedroom”
activity=”brushing teeth” location=”Bathroom”
activity=”drinking water” location=”Living room”
activity=”pouring water into a cup” location=”Living room”

Sample24:
activity=”putting on/taking off glasses” location=”Bathroom”
activity=”brushing teeth” location=”Bathroom”
activity=”pouring water into a cup” location=”Bathroom”
activity=”drinking water” location=”Bathroom”
activity=”washing face” location=”Bathroom”

Objective4: Rest (24 samples)

Sample1:
activity="opening the door and walk in" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
activity="taking off jacket" location="Entrance"
activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="bedroom"
activity="lying down" location="bedroom"

Sample2:
activity="opening the door and walk in" location="bedroom"
activity="taking off jacket" location="bedroom"
activity="putting on/taking off shoes" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="bedroom"
activity="lying down" location="bedroom"

Sample3:
activity="putting on/taking off shoes" location="Entrance"
activity="opening the door and walk in" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"

doi:10.6342/NTU202300261

activity="taking off jacket" location="bedroom"
activity="rubbing face with hands" location="bedroom"
activity="lying down" location="bedroom"

Sample4:
activity="opening the door and walk in" location="Entrance"
activity="putting on/taking off glasses" location="Living room"
activity="putting on/taking off shoes" location="Living room"
activity="taking off jacket" location="Living room"
activity="lying down" location="Living room"

Sample5:
activity="opening the door and walk in" location="Entrance"
activity="taking off jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Dining room/kitchen"
activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="bedroom"
activity="lying down" location="bedroom"

Sample6:
activity="opening the door and walk in" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
activity="taking off jacket" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="bedroom"
activity="lying down" location="bedroom"

Sample7:
activity="putting on/taking off shoes" location="Living room"
activity="opening the door and walk in" location="bedroom"
activity="taking off jacket" location="Study room"
activity="putting on/taking off glasses" location="Dining room/kitchen"
activity="rubbing face with hands" location="Entrance"
activity="lying down" location="Living room"

Sample8:
activity="putting on/taking off shoes" location="bedroom"
activity="opening the door and walk in" location="bedroom"
activity="taking off jacket" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="Study room"
activity="lying down" location="Study room"

Sample9:
activity="opening the door and walk in" location="bedroom"
activity="putting on/taking off shoes" location="Living room"
activity="taking off jacket" location="Study room"
activity="lying down" location="bedroom"

doi:10.6342/NTU202300261

Sample10:

activity="putting on/taking off glasses" location="Bathroom"
activity="rubbing face with hands" location="Entrance"
activity="taking off jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Living room"
activity="opening the door and walk in" location="Study room"
activity="lying down" location="bedroom"

Sample11:

activity="putting on/taking off shoes" location="Entrance"
activity="opening the door and walk in" location="Living room"
activity="taking off jacket" location="bedroom"
activity="lying down" location="bedroom"

Sample12:

activity="putting on/taking off glasses" location="bedroom"
activity="rubbing face with hands" location="Living room"
activity="lying down" location="Living room"

Sample13:

activity="opening the door and walk in" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
activity="taking off jacket" location="Entrance"
activity="putting on/taking off glasses" location="bedroom"
activity="lying down" location="bedroom"

Sample14:

activity="opening the door and walk in" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
activity="taking off jacket" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="lying down" location="bedroom"

Sample15:

activity="opening the door and walk in" location="bedroom"
activity="putting on/taking off shoes" location="Study room"
activity="taking off jacket" location="Study room"

Sample16:

activity="putting on/taking off glasses" location="bedroom"
activity="taking off jacket" location="Living room"
activity="putting on/taking off glasses" location="bedroom"
activity="taking off jacket" location="Living room"
activity="putting on/taking off glasses" location="bedroom"
activity="taking off jacket" location="Living room"

doi:10.6342/NTU202300261

Sample17:

activity=”putting on/taking off glasses” location=”Study room”
activity=”rubbing face with hands” location=”Study room”
activity=”taking off jacket” location=”Living room”
activity=”lying down” location=”bedroom”

Sample18:

activity=”putting on/taking off shoes” location=”bedroom”
activity=”opening the door and walk in” location=”bedroom”
activity=”taking off jacket” location=”bedroom”
activity=”putting on/taking off glasses” location=”bedroom”
activity=”lying down” location=”bedroom”

Sample19:

activity=”opening the door and walk in” location=”Entrance”
activity=”putting on/taking off shoes” location=”Entrance”
activity=”taking off jacket” location=”Living room”
activity=”putting on/taking off glasses” location=”Living room”
activity=”lying down” location=”Living room”
activity=”rubbing face with hands” location=”Living room”

Sample20:

activity=”opening the door and walk in” location=”Entrance”
activity=”putting on/taking off shoes” location=”Entrance”
activity=”taking off jacket” location=”Living room”
activity=”putting on/taking off glasses” location=”Living room”
activity=”rubbing face with hands” location=”Living room”
activity=”lying down” location=”Living room”

Sample21:

activity=”opening the door and walk in” location=”Entrance”
activity=”putting on/taking off shoes” location=”Entrance”
activity=”taking off jacket” location=”Living room”
activity=”putting on/taking off glasses” location=”bedroom”
activity=”rubbing face with hands” location=”bedroom”
activity=”lying down” location=”bedroom”

Sample22:

activity=”opening the door and walk in” location=”Living room”
activity=”putting on/taking off shoes” location=”Living room”
activity=”taking off jacket” location=”Living room”
activity=”putting on/taking off glasses” location=”Living room”
activity=”rubbing face with hands” location=”Living room”
activity=”lying down” location=”Living room”

Sample23:

activity=”opening the door and walk in” location=”Entrance”
activity=”putting on/taking off shoes” location=”Entrance”

doi:10.6342/NTU202300261

activity=”taking off jacket” location=”Entrance”
activity=”opening the door and walk in” location=”bedroom”
activity=”lying down” location=”bedroom”

Sample24:
activity=”opening the door and walk in” location=”Entrance”
activity=”putting on/taking off shoes” location=”Entrance”
activity=”taking off jacket” location=”bedroom”
activity=”lying down” location=”bedroom”

Objective5: Cook (5 samples)

Sample1:
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="washing hand" location="Dining room/kitchen"
activity="trimming vegetable" location="Dining room/kitchen"
activity="cutting vegetable on the cutting board" location="Dining room/kitchen"
activity="using a gas stove" location="Dining room/kitchen"

Sample2:
activity="washing hand" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="trimming vegetable" location="Dining room/kitchen"
activity="cutting vegetable on the cutting board" location="Dining room/kitchen"
activity="using a gas stove" location="Dining room/kitchen"

Sample3:
activity="washing hand" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Living room"
activity="trimming vegetable" location="Study room"
activity="cutting vegetable on the cutting board" location="Dining room/kitchen"
activity="using a gas stove" location="Dining room/kitchen"

Sample4:
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="washing hand" location="Bathroom"
activity="trimming vegetable" location="Dining room/kitchen"
activity="cutting vegetable on the cutting board" location="Living room"
activity="using a gas stove" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample5:
activity=”washing hand” location=”Bathroom”
activity=”putting food in the fridge/taking food from the fridge” location=”Dining room/kitchen”
activity=”trimming vegetable” location=”Dining room/kitchen”
activity=”cutting vegetable on the cutting board” location=”Dining room/kitchen”
activity=”using a gas stove” location=”Dining room/kitchen”

Objective6: Have Meal (6 samples)

Sample1:
activity="washing hand" location="Dining room/kitchen"
activity="eating food with a fork" location="Living room"

Sample2:
activity="washing hand" location="Dining room/kitchen"
activity="eating food with a fork" location="Dining room/kitchen"

Sample3:
activity="washing hand" location="Living room"
activity="eating food with a fork" location="Dining room/kitchen"

Sample4:
activity="washing hand" location="Dining room/kitchen"
activity="eating food with a fork" location="bedroom"

Sample5:
activity="washing hand" location="Bathroom"
activity="eating food with a fork" location="Living room"

Sample6:
activity="washing hand" location="Bathroom"
activity="eating food with a fork" location="Dining room/kitchen"

Objective7: Eat Fruit (9 samples)

Sample1:
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="washing hand" location="Dining room/kitchen"
activity="peeling fruit" location="Dining room/kitchen"

Sample2:
activity="washing hand" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="peeling fruit" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample3:
activity="washing hand" location="Living room"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="peeling fruit" location="Study room"

Sample4:
activity="washing hand" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Living room"
activity="peeling fruit" location="Living room"

Sample5:
activity="washing hand" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="peeling fruit" location="Study room"

Sample6:
activity="putting food in the fridge/taking food from the fridge" location="Living room"
activity="washing hand" location="Living room"
activity="peeling fruit" location="Living room"

Sample7:
activity="washing hand" location="Dining room/kitchen"
activity="peeling fruit" location="Dining room/kitchen"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"

Sample8:
activity="putting food in the fridge/taking food from the fridge" location="bedroom"
activity="peeling fruit" location="Living room"
activity="putting food in the fridge/taking food from the fridge" location="bedroom"

Sample9:
activity="washing hand" location="Bathroom"
activity="putting food in the fridge/taking food from the fridge" location="Dining room/kitchen"
activity="peeling fruit" location="Dining room/kitchen"

Objective8: Afterward Cleaning (11 samples)

Sample1:
activity="washing dishes" location="Dining room/kitchen"
activity="wiping off the dining table" location="Living room"
activity="washing a towel by hands" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample2:
activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Dining room/kitchen"
activity="wiping off the dining table" location="Dining room/kitchen"

Sample3:
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Dining room/kitchen"

Sample4:
activity="washing dishes" location="Dining room/kitchen"
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing a towel by hands" location="Dining room/kitchen"

Sample5:
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Study room"

Sample6:
activity="washing a towel by hands" location="Living room"
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing a towel by hands" location="Study room"

Sample7:
activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Living room"
activity="washing dishes" location="Dining room/kitchen"

Sample8:
activity="washing a towel by hands" location="Living room"
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing dishes" location="Dining room/kitchen"

Sample9:
activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Living room"
activity="wiping off the dining table" location="Dining room/kitchen"

Sample10:
activity="wiping off the dining table" location="Dining room/kitchen"
activity="washing a towel by hands" location="Dining room/kitchen"
activity="washing dishes" location="Dining room/kitchen"

Sample11:
activity="wiping off the dining table" location="Living room"

doi:10.6342/NTU202300261

activity="washing dishes" location="Dining room/kitchen"
activity="washing a towel by hands" location="Dining room/kitchen"

Objective9: Do Chores (21 samples)

Sample1:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="bedroom"
activity="scrubbing the floor with a rag" location="bedroom"
activity="rubbing up furniture" location="bedroom"
activity="hanging out laundry" location="Living room"
activity="spreading/folding bedding" location="bedroom"

Sample2:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="hanging out laundry" location="Bathroom"
activity="spreading/folding bedding" location="bedroom"

Sample3:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="hanging out laundry" location="Bathroom"
activity="rubbing up furniture" location="Living room"

Sample4:
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"
activity="hanging out laundry" location="Entrance"

Sample5:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="hanging out laundry" location="Entrance"

Sample6:
activity="washing dishes" location="Dining room/kitchen"
activity="spreading/folding bedding" location="bedroom"
activity="hanging out laundry" location="Entrance"

doi:10.6342/NTU202300261

activity="rubbing up furniture" location="Living room"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"

Sample7:
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Dining room/kitchen"
activity="rubbing up furniture" location="bedroom"
activity="hanging out laundry" location="Bathroom"
activity="spreading/folding bedding" location="Study room"
activity="washing dishes" location="Dining room/kitchen"

Sample8:
activity="rubbing up furniture" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="hanging out laundry" location="Entrance"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Sample9:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="Living room"
activity="hanging out laundry" location="Bathroom"
activity="scrubbing the floor with a rag" location="bedroom"
activity="spreading/folding bedding" location="Study room"
activity="rubbing up furniture" location="Living room"

Sample10:
activity="hanging out laundry" location="bedroom"
activity="vacuuming the floor" location="Study room"
activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Dining room/kitchen"
activity="spreading/folding bedding" location="Entrance"
activity="washing dishes" location="Bathroom"

Sample11:
activity="washing dishes" location="Dining room/kitchen"
activity="vacuuming the floor" location="Living room"
activity="hanging out laundry" location="Living room"
activity="rubbing up furniture" location="Living room"
Sample12:
activity="spreading/folding bedding" location="bedroom"
activity="rubbing up furniture" location="Study room"
activity="hanging out laundry" location="Living room"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="washing dishes" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample13:
activity="hanging out laundry" location="bedroom"
activity="scrubbing the floor with a rag" location="bedroom"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Sample14:
activity="hanging out laundry" location="Living room"
activity="vacuuming the floor" location="Study room"
activity="rubbing up furniture" location="bedroom"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Sample15:
activity="scrubbing the floor with a rag" location="bedroom"
activity="hanging out laundry" location="Living room"
activity="vacuuming the floor" location="Study room"
activity="scrubbing the floor with a rag" location="bedroom"
activity="vacuuming the floor" location="Living room"
activity="hanging out laundry" location="Study room"

Sample16:
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="hanging out laundry" location="Living room"
activity="washing dishes" location="Dining room/kitchen"

Sample17:
activity="washing dishes" location="Dining room/kitchen"
activity="hanging out laundry" location="Living room"
activity="vacuuming the floor" location="Living room"
activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Living room"
activity="spreading/folding bedding" location="bedroom"

Sample18:
activity="washing dishes" location="Dining room/kitchen"
activity="rubbing up furniture" location="Living room"
activity="vacuuming the floor" location="Study room"
activity="scrubbing the floor with a rag" location="Study room"
activity="hanging out laundry" location="bedroom"
activity="spreading/folding bedding" location="bedroom"

Sample19:
activity="hanging out laundry" location="Living room"
activity="vacuuming the floor" location="Living room"

doi:10.6342/NTU202300261

activity="scrubbing the floor with a rag" location="Living room"
activity="rubbing up furniture" location="Living room"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Sample20:
activity="hanging out laundry" location="Entrance"
activity="vacuuming the floor" location="bedroom"
activity="scrubbing the floor with a rag" location="bedroom"
activity="rubbing up furniture" location="bedroom"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Sample21:
activity="rubbing up furniture" location="Living room"
activity="vacuuming the floor" location="bedroom"
activity="scrubbing the floor with a rag" location="bedroom"
activity="hanging out laundry" location="bedroom"
activity="spreading/folding bedding" location="bedroom"
activity="washing dishes" location="Dining room/kitchen"

Objective10: Mail Letter (16 samples)

Sample1:
activity="sitting up/standing up" location="Living room"
activity="putting on/taking off glasses" location="Living room"
activity="handwriting" location="Living room"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"

Sample2:
activity="putting on/taking off glasses" location="bedroom"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Living room"

Sample3:
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample4:
activity="handwriting" location="Study room"
activity="putting on/taking off glasses" location="Study room"
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Living room"

doi:10.6342/NTU202300261

activity="putting on/taking off shoes" location="Entrance"

Sample5:
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on/taking off shoes" location="Entrance"
activity="putting on jacket" location="Entrance"

Sample6:
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

Sample7:
activity="putting on/taking off glasses" location="Living room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Entrance"

Sample8:
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Entrance"

Sample9:
activity="putting on/taking off glasses" location="bedroom"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Entrance"

Sample10:
activity="handwriting" location="bedroom"
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample11:
activity="sitting up/standing up" location="bedroom"
activity="putting on/taking off glasses" location="Living room"
activity="handwriting" location="bedroom"
activity="sitting up/standing up" location="Living room"
activity="putting on/taking off glasses" location="bedroom"

doi:10.6342/NTU202300261

Sample12:
activity="sitting up/standing up" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="handwriting" location="Study room"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample13:
activity="sitting up/standing up" location="Study room"
activity="handwriting" location="Study room"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

Sample14:
activity="sitting up/standing up" location="Study room"
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"

Sample15:
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Study room"
activity="putting on/taking off shoes" location="Study room"

Sample16:
activity="putting on/taking off glasses" location="Study room"
activity="handwriting" location="Study room"
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Study room"
activity="putting on/taking off shoes" location="Entrance"

Objective11: Going out Doors (14 samples)

Sample1:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
Sample2:
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Living room"

Sample3:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

doi:10.6342/NTU202300261

Sample4:
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample5:
activity="sitting up/standing up" location="Living room"
activity="putting on/taking off shoes" location="Dining room/kitchen"
activity="putting on jacket" location="Entrance"

Sample6:
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

Sample7:
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Dining room/kitchen"

Sample8:
activity="sitting up/standing up" location="Dining room/kitchen"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample9:
activity="sitting up/standing up" location="Entrance"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"

Sample10:
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Entrance
activity="putting on/taking off shoes" location="Entrance"

Sample11:
activity="sitting up/standing up" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"
activity="putting on jacket" location="Entrance"

Sample12:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Study room"
activity="putting on/taking off shoes" location="bedroom"

Sample13:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Living room"

doi:10.6342/NTU202300261

activity="putting on/taking off shoes" location="Entrance"

Sample14:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Living room"

Objective12: Meeting with people (16 samples)

Sample1:
activity="sitting up/standing up" location="Living room"
activity="putting on lipstick" location="Living room"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"

Sample2:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Living room"

Sample3:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

Sample4:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample5:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="sitting up/standing up" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Entrance"
activity="putting on jacket" location="Entrance"

Sample6:
activity="sitting up/standing up" location="Living room"
activity="putting on jacket" location="Dining room/kitchen"
activity="putting on/taking off shoes" location="Entrance"

Sample7:
activity="sitting up/standing up" location="Study room"
activity="putting on jacket" location="Living room"

doi:10.6342/NTU202300261

activity="putting on/taking off shoes" location="Entrance"

Sample8:
activity="putting on cosmetics" location="Dining room/kitchen"
activity="putting on lipstick" location="Living room"
activity="sitting up/standing up" location="Dining room/kitchen"
activity="putting on lipstick" location="Living room"
activity="sitting up/standing up" location="Dining room/kitchen"

Sample9:
activity="sitting up/standing up" location="bedroom"
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample10:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="sitting up/standing up" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

Sample11:
activity="sitting up/standing up" location="bedroom"
activity="putting on jacket" location="Entrance"
activity="putting on/taking off shoes" location="Entrance"

Sample12:
activity="putting on cosmetics" location="Living room"
activity="putting on lipstick" location="bedroom"
activity="sitting up/standing up" location="Study room"
activity="putting on lipstick" location="Living room"
activity="sitting up/standing up" location="bedroom"
Sample13:
activity="sitting up/standing up" location="bedroom"
activity="putting on cosmetics" location="Bathroom"
activity="putting on lipstick" location="Bathroom"
activity="putting on jacket" location="Living room"
activity="putting on/taking off shoes" location="Entrance"

Sample14:
activity="sitting up/standing up" location="bedroom"
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"
activity="putting on jacket" location="bedroom"
activity="putting on/taking off shoes" location="Entrance"

doi:10.6342/NTU202300261

Sample15:
activity=”putting on cosmetics” location=”bedroom”
activity=”putting on lipstick” location=”bedroom”
activity=”sitting up/standing up” location=”bedroom”
activity=”putting on jacket” location=”Living room”
activity=”putting on/taking off shoes” location=”Entrance”

Sample16:
activity=”putting on cosmetics” location=”bedroom”
activity=”putting on lipstick” location=”bedroom”
activity=”sitting up/standing up” location=”bedroom”
activity=”putting on jacket” location=”bedroom”
activity=”putting on/taking off shoes” location=”bedroom”

Objective13: Doing Makeup (7 samples)

Sample1:
activity="putting on cosmetics" location="Living room"
activity="putting on lipstick" location="Living room"

Sample2:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="bedroom"

Sample3:
activity="putting on cosmetics" location="Entrance"
activity="putting on lipstick" location="Dining room/kitchen"

Sample4:
activity="putting on cosmetics" location="bedroom"
activity="putting on lipstick" location="Living room"

Sample5:
activity="putting on cosmetics" location="Study room"
activity="putting on lipstick" location="Study room"

Sample6:
activity=”putting on cosmetics” location=”Study room”
activity=”putting on cosmetics” location=”Study room”

Sample7:
activity=”putting on cosmetics” location=”Bathroom”
activity=”putting on lipstick” location=”Bathroom”

Objective14: Greet People (12 samples)

Sample1:
activity="waving a hand" location="Entrance"
activity="hugging each other" location="Entrance"

doi:10.6342/NTU202300261

Sample2:

activity="waving a hand" location="Entrance"
activity="taking a bow" location="Entrance"
activity="Handshake" location="Living room"
activity="hugging each other" location="Living room"

Sample3:

activity="waving a hand" location="Entrance"
activity="Handshake" location="Entrance"
activity="hugging each other" location="Entrance"
activity="taking a bow" location="Entrance"

Sample4:

activity="waving a hand" location="Living room
activity="Handshake" location="Dining room/kitchen"
activity="hugging each other" location="Dining room/kitchen"
activity="taking a bow" location="Dining room/kitchen"

Sample5:

activity="waving a hand" location="Living room"
activity="Handshake" location="Living room"
activity="hugging each other" location="Living room"

Sample6:

activity="waving a hand" location="Living room"
activity="taking a bow" location="Entrance"
activity="Handshake" location="Entrance"

Sample7:

activity="Handshake" location="Entrance"
activity="hugging each other" location="Entrance"
activity="waving a hand" location="Entrance"

Sample8:

activity="waving a hand" location="Living room"
activity="hugging each other" location="bedroom"
activity="Handshake" location="bedroom"
activity="waving a hand" location="Living room"

Sample9:

activity="hugging each other" location="Entrance"
activity="Handshake" location="Living room"
activity="taking a bow" location="Living room"
activity="waving a hand" location="Entrance"

Sample10:

activity="waving a hand" location="Entrance"
activity="Handshake" location="Entrance"

doi:10.6342/NTU202300261

activity="taking a bow" location="Entrance"

Sample11:
activity="waving a hand" location="Entrance"
activity="Handshake" location="Living room"
activity="hugging each other" location="Living room"

Sample12:
activity="Handshake" location="bedroom"
activity="hugging each other" location="bedroom"
activity="waving a hand" location="bedroom"
activity="taking a bow" location="bedroom"

Objective15: Say Goodbye (13 samples)

Sample1:
activity="hugging each other" location="Entrance"
activity="waving a hand" location="Entrance"

Sample2:
activity="Handshake" location="Living room"
activity="hugging each other" location="Living room"
activity="taking a bow" location="Entrance"
activity="waving a hand" location="Entrance"

Sample3:
activity="Handshake" location="Living room"
activity="hugging each other" location="Living room"
activity="waving a hand" location="Entrance"
activity="taking a bow" location="Entrance"

Sample4:
activity="waving a hand" location="Living room"
activity="hugging each other" location="Dining room/kitchen"
activity="Handshake" location="Study room"
activity="taking a bow" location="Dining room/kitchen"

Sample5:
activity="Handshake" location="Living room"
activity="waving a hand" location="Living room"
activity="taking a bow" location="Living room"

Sample6:
activity="Handshake" location="bedroom"
activity="hugging each other" location="Study room"
activity="hugging each other" location="Dining room/kitchen"
activity="Handshake" location="Living room"

doi:10.6342/NTU202300261

Sample7:
activity="Handshake" location="Entrance"
activity="waving a hand" location="Living room"

Sample8:
activity="Handshake" location="Entrance"
activity="hugging each other" location="Entrance"
activity="waving a hand" location="Entrance"

Sample9:
activity="waving a hand" location="bedroom"
activity="hugging each other" location="Study room"
activity="waving a hand" location="Living room"
activity="hugging each other" location="bedroom"

Sample10:
activity="hugging each other" location="Living room"
activity="taking a bow" location="Living room"
activity="waving a hand" location="Entrance"

Sample11:
activity="Handshake" location="Entrance"
activity="waving a hand" location="Entrance"
activity="hugging each other" location="Entrance"

Sample12:
activity="waving a hand" location="Living room"
activity="hugging each other" location="Entrance"

Sample13:
activity="Handshake" location="bedroom"
activity="hugging each other" location="bedroom"
activity="waving a hand" location="bedroom"
activity="taking a bow" location="bedroom"

Objective16: Using Computer (6 samples)

Sample1:
activity="putting on/taking off glasses" location="Living room"
activity="using a computer" location="Living room"

Sample2:
activity="putting on/taking off glasses" location="Study room"
activity="using a computer" location="Study room"

Sample3:
activity="putting on/taking off glasses" location="Living room"
activity="using a computer" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample4:
activity="using a computer" location="Living room"
activity="putting on/taking off glasses" location="Study room"

Sample5:
activity="putting on/taking off glasses" location="Dining room/kitchen"
activity="using a computer" location="Living room"

Sample6:
activity="using a computer" location="Living room"
activity="putting on/taking off glasses" location="Living room"

Objective17: Watch TV (6 samples)

Sample1:
activity="putting on/taking off glasses" location="Living room"
activity="using remote control" location="Living room"

Sample2:
activity="using remote control" location="Living room"
activity="putting on/taking off glasses" location="Living room"

Sample3:
activity="putting on/taking off glasses" location="Dining room/kitchen"
activity="using remote control" location="Living room"

Sample4:
activity="putting on/taking off glasses" location="bedroom"
activity="using remote control" location="bedroom"

Sample5:
activity="putting on/taking off glasses" location="Study room"
activity="putting on/taking off glasses" location="Study room"

Sample6:
activity="putting on/taking off glasses" location="Dining room/kitchen
activity="using remote control" location="Dining room/kitchen"

Objective18: Relax (10 samples)

Sample1:
activity="putting on/taking off glasses" location="Living room
activity="reading a book" location="Living room"
activity="play with mobile phone" location="Living room"

Sample2:
activity="putting on/taking off glasses" location="bedroom"
activity="play with mobile phone" location="bedroom"
activity="reading a book" location="Study room"
activity="reading a newspaper" location="Study room"

doi:10.6342/NTU202300261

Sample3:

activity="putting on/taking off glasses" location="Living room"
activity="reading a newspaper" location="Living room"
activity="play with mobile phone" location="Living room"
activity="reading a book" location="Study room"

Sample4:

activity="putting on/taking off glasses" location="Study room"
activity="reading a newspaper" location="Study room"
activity="reading a book" location="Study room"
activity="play with mobile phone" location="bedroom"

Sample5:

activity="putting on/taking off glasses" location="Living room"
activity="play with mobile phone" location="Dining room/kitchen"
activity="reading a book" location="bedroom"
activity="play with mobile phone" location="Study room"

Sample6:

activity="reading a newspaper" location="Study room"
activity="reading a newspaper" location="Study room"
activity="reading a newspaper" location="Study room"
activity="reading a newspaper" location="Study room"

Sample7:

activity="play with mobile phone" location="bedroom"
activity="reading a book" location="Study room"
activity="play with mobile phone" location="Living room"
activity="reading a book" location="bedroom"

Sample8:

activity="putting on/taking off glasses" location="bedroom"
activity="reading a book" location="bedroom"
activity="reading a newspaper" location="bedroom"
activity="play with mobile phone" location="bedroom"

Sample9:

activity="putting on/taking off glasses" location="Living room"
activity="play with mobile phone" location="Living room"
activity="reading a newspaper" location="Living room"
activity="reading a book" location="Study room"

Sample10:

activity="reading a book" location="Living room"
activity="reading a newspaper" location="Living room"
activity="play with mobile phone" location="Living room"
activity="putting on/taking off glasses" location="Living room"

doi:10.6342/NTU202300261

Objective19: Call Someone (9 samples)

Sample1:
activity="play with mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample2:
activity="play with mobile phone" location="Living room"
activity="talk on the phone" location="Living room"
activity="flapping a hand up and down" location="Living room"

Sample3:
activity="flapping a hand up and down" location="Living room"
activity="play with mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample4:
activity="play with mobile phone" location="Study room"
activity="talk on the phone" location="Study room"
activity="flapping a hand up and down" location="Study room"

Sample5:
activity="play with mobile phone" location="Living room"
activity="talk on the phone" location="Dining room/kitchen"
activity="flapping a hand up and down" location="Study room"

Sample6:
activity="flapping a hand up and down" location="Living room"
activity="flapping a hand up and down" location="Living room"
activity="flapping a hand up and down" location="Living room"

Sample7:
activity="play with mobile phone" location="bedroom"
activity="talk on the phone" location="bedroom"

Sample8:
activity="flapping a hand up and down" location="bedroom"
activity="talk on the phone" location="Study room"
activity="play with mobile phone" location="Living room"

Sample9:
activity="flapping a hand up and down" location="bedroom"
activity="talk on the phone" location="bedroom"
activity="play with mobile phone" location="bedroom"

Objective20: Sleep (5 samples)

Sample1:
activity="spreading bedding/folding bedding" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"

doi:10.6342/NTU202300261

activity="lying down" location="bedroom"

Sample2:
activity="putting on/taking off glasses" location="Living room"
activity="lying down" location="Living room"

Sample3:
activity="putting on/taking off glasses" location="Study room"
activity="spreading bedding/folding bedding" location="bedroom"
activity="lying down" location="bedroom"

Sample4:
activity="spreading bedding/folding bedding" location="bedroom"
activity="lying down" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"

Sample5:
activity="spreading bedding/folding bedding" location="bedroom"
activity="putting on/taking off glasses" location="bedroom"
activity="lying down" location="Living room"

Objective21: Turn on AC (11 samples)

Sample1:
activity="looking around for something" location="Living room"
activity="using remote control" location="Living room"

Sample2:
activity="taking off a jacket" location="Living room"
activity="looking around for something" location="bedroom"
activity="using remote control" location="bedroom"

Sample3:
activity=”taking off a jacket” location=”Entrance”
activity=”looking around for something” location=”Living room”
activity=”using remote control” location=”Living room”

Sample4:
activity=”taking off a jacket” location=”Entrance”
activity=”looking around for something” location=”bedroom”
activity=”using remote control” location=”bedroom”

Sample5:
activity=”taking off a jacket” location=”Study room”
activity=”using remote control” location=”Study room”

Sample6:
activity=”taking off a jacket” location=”bedroom”
activity=”looking around for something” location=”bedroom”

doi:10.6342/NTU202300261

activity=”using remote control” location=”bedroom”

Sample7:
activity=”taking off a jacket” location=”Living room”
activity=”using remote control” location=”Dining room/kitchen”

Sample8:
activity=”taking off a jacket” location=”Living room”
activity=”looking around for something” location=”Living room”
activity=”using remote control” location=”Living room”

Sample9:
activity=”looking around for something” location=”Living room”
activity=”using remote control” location=”Living room”
activity=”taking off a jacket” location=”Living room”

Sample10:
activity=”taking off a jacket” location=”Entrance”
activity=”looking around for something” location=”Study room”
activity=”using remote control” location=”Study room”

Sample11:
activity=”taking off a jacket” location=”Study room”
activity=”looking around for something” location=”Study room”
activity=”using remote control” location=”Study room”

Objective22: Harm (6 samples)

Sample1:
activity="pointing with a finger" location="Living room"
activity="fighting each other" location="Living room"

Sample2:
activity="pointing with a finger" location="bedroom"
activity="fighting each other" location="bedroom"

Sample3:
activity="fighting each other" location="Study room"
activity="pointing with a finger" location="Study room"

Sample4:
activity="fighting each other" location="Living room"
activity="pointing with a finger" location="Living room"

Sample5:
activity="fighting each other" location="Dining room/kitchen"
activity="pointing with a finger" location="Dining room/kitchen"

doi:10.6342/NTU202300261

Sample6:
activity=”pointing with a finger” location=”Study room”
activity=”fighting each other” location=”Study room”

Objective23: Call for Help (18 samples)

Sample1:
activity="fallen on the floor" location="Living room"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample2:
activity="fallen on the floor" location="Bathroom"
activity="playing with a mobile phone" location="Living room"

Sample3:
activity="fallen on the floor" location="Dining room/kitchen"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample4:
activity="fallen on the floor" location="Dining room/kitchen"
activity="flapping a hand up and down" location="Living room"
activity="talk on the phone" location="Living room"

Sample5:
activity="fallen on the floor" location="bedroom"
activity="flapping a hand up and down" location="bedroom"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample6:
activity="fallen on the floor" location="Bathroom"
activity="playing with a mobile phone" location="Bathroom"
activity="talk on the phone" location="Bathroom"

Sample7:
activity="fallen on the floor" location="Dining room/kitchen"
activity="flapping a hand up and down" location="Dining room/kitchen"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample8:
activity="fallen on the floor" location="Bathroom"
activity="flapping a hand up and down" location="Bathroom"
activity="playing with a mobile phone" location="Bathroom"

doi:10.6342/NTU202300261

Sample9:

activity="fallen on the floor" location="Living room"
activity="playing with a mobile phone" location="Study room"
activity="talk on the phone" location="Study room"

Sample10:

activity="fallen on the floor" location="bedroom"
activity="playing with a mobile phone" location="bedroom"
activity="talk on the phone" location="bedroom"

Sample11:

activity="fallen on the floor" location="bedroom"
activity="flapping a hand up and down" location="Living room"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample12:

activity="fallen on the floor" location="Dining room/kitchen"
activity="flapping a hand up and down" location="Dining room/kitchen"
activity="talk on the phone" location="Dining room/kitchen"

Sample13:

activity="fallen on the floor" location="Living room"
activity="flapping a hand up and down" location="Living room"
activity="talk on the phone" location="Living room"

Sample14:

activity="fallen on the floor" location="Bathroom"
activity="flapping a hand up and down" location="Bathroom"
activity="playing with a mobile phone" location="Living room"
activity="talk on the phone" location="Living room"

Sample15:

activity="fallen on the floor" location="Study room"
activity="flapping a hand up and down" location="Study room"
activity="playing with a mobile phone" location="Study room"
activity="talk on the phone" location="Study room"

Sample16:

activity="fallen on the floor" location="Bathroom"
activity="flapping a hand up and down" location="Bathroom"
activity="talk on the phone" location="Bathroom"

Sample17:

activity="fallen on the floor" location="bedroom"
activity="flapping a hand up and down" location="bedroom"
activity="talk on the phone" location="bedroom"

doi:10.6342/NTU202300261

Sample18:

activity="fallen on the floor" location="bedroom"

activity="playing with a mobile phone" location="Living room"

activity="talk on the phone" location="Living room"

doi:10.6342/NTU202300261