2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
Macau, China, November 4-8, 2019

Multi-Layer Environmental Affordance Map for Robust Indoor Localization, Event Detection and Social Friendly Navigation

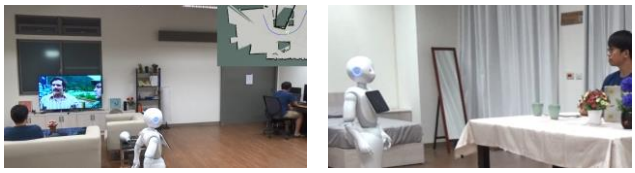
Ping-Tsang Wu, Chee-An Yu, Shao-Hung Chan, Ming-Li Chiang, and Li-Chen Fu, *Member, IEEE*

***Abstract*— In this paper, we propose a novel system architecture called multi-layer environmental affordance map for social and service companion robots. Based on this architecture, robots can organize the perception and inference information efficiently and generate social friendly navigation strategies. In other words, robots are able to strengthen their perception and inference abilities to interact with domestic environment and users under our efficient framework. The main feature of this architecture is that the relations between layers can be viewed as affordances to improve the accuracy and the robustness of the detection and inference. The results show that our architecture achieves robust indoor localization, scene localization, human event detection and socially friendly navigation in real time under limited computational resource.**

I. Introduction

In recent years, due to the growth of the elderly population, the use of social companion and service robots has received increasing attention [1][2]. For the application of these types of robot, the basic and important capabilities are the localization and navigation. On top of that, the ability of perception is also critical. As for normal mobile robots, it may be satisfactory for simply navigating safely and robustly from one place to another. Nevertheless, social robots, especially designing for household usage, should take human interactions into consideration while moving in the indoor environment [3]. For social robots that connect with people, it is practical for them to not only navigate safe and sound in the domestic environment, but also autonomously take human status into considerations while making decisions in real time [4]. Therefore, how a robot can store its inferences and perceive human-beings based on its past knowledge and current observation in real time become crucial and challenging. Furthermore, efficient high-level path planning for mobile robots that combines sensor perception and human-robot interaction (HRI) is required for robots in our daily lives [5].

Although a social robot may navigate smoothly using heuristic searching and avoid obstacles with local path planning methods given, it is still challenging to move smoothly in a human living environment. One of the main reasons is that the robot tends to ignore human status while moving, leading to unpleasant user experience. In other words, while navigating in the human social environment, a social robot should take object detection, scene recognition, and human activities into consideration instead of planning merely with heuristic algorithms as A* [6]. Other than "social navigation" that takes human-beings more than dynamic obstacles in [7] and [8], here we focus more on path planning and decision making for "social friendly navigation." As far as we concern, it is unsuitable for robot to over-interfere human. For example, the robot should avoid crossing the living room while there are people watching television inside, as shown in Fig. 1(a). To deal with such problem, we purpose a *multi-layer environmental affordance map* architecture that combines visual perceptions and human reactions while processing autonomous mobile robot navigation in the household surroundings to achieve social friendly navigation in complex real world environment. In addition, the perception results can be stored in our system during the Simultaneous Localization and Mapping (SLAM) procedure and recalled efficiently in the navigation procedure. That is to say, the robot can build the static map and memorize the inferences based on the detected objects in real time. On top of that, these inference results can be modified dynamically according to the new detection outcomes.



(a) (b)

Fig. 1 Two scenarios of robot navigation using our proposed multi-layer environmental affordance map architecture. (a) shows that the robot is moving to the office based on user command while avoiding interruption of person watching television. (b) is the case that robot recognizing human's engagement and moving toward him.

In this paper, we consider the affordance concept [9] in our system to illustrate the semantics and inference outcomes. For instance, a sofa can be taken as a sitting spot for a person to watch television. Since the affordances relate the agent's actions to their effects on the surrounding objects, it can be used as the prior knowledge so that the robots can have stronger inference ability. In [10], the authors proposed a hierarchical probabilistic representation of space by using a global topological representation of places with object graph. Nonetheless, the work preset the regions for scene recognition and may failed when the indoor environment is an open space without walls. On the contrary, the scene layer in the proposed *multi-layer environmental affordance map* provides regions boundaries such that the robot can recognize the size of the scene even in the open area. In [11], object affordance is used to predict human activities. Nevertheless, the work does not provide a mapping concept like our work to localize objects, scene, and human activities on a static map. In [12], the authors

Ping-Tsang Wu is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan (e-mail: r05921013@ntu.edu.tw).
Chee-An Yu is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan (e-mail: r07921010@ntu.edu.tw).
Shao-Hung Chan is with the Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan (e-mail: r06921017@ntu.edu.tw).
Ming-Li Chiang is with the Graduate Institute of Automation Technology, National Taipei University of Technology, Taipei, Taiwan(e-mail: minglichiang@ntu.edu.tw)
Li-Chen Fu, Director of NTU Center for Artificial Intelligence & Advanced Robotics, Taipei, Taiwan (e-mail: lichen@ntu.edu.tw)

978-1-7281-4004-9/19/$31.00 ©2019 IEEE

proposed a topological mapping using object detection as well as it affordances. However, topological mapping may lead to limited navigation feasibilities. In this paper, we implement our system on the basis of grid map, which is more comprehensive for robot to navigate in the domestic environment. To sum up, the proposed architecture strengthens the perception and inference abilities of the service and social companion robots such that they can interact with environment and users under an efficient framework. In comparison to robot equipped with merely low-level heuristic planning algorithms, our system is more capable of human-involved task handling. It is worth mentioning that the proposed architecture utilizes computational resource containing only one GPU in real time, which is suitable for household usage.

II. System Overview

The multi-layer environmental affordance map is shown in Fig. 2. The feature of this architecture is that the relations between layers can be used as affordance to improve the accuracy and the robustness of the detection. With the proposed architecture, the calculation can be reduced so that the robot can generate suitable navigation strategies in real time based on its previous observations efficiently.

We design four layers in the architecture which are *static map layer*, *object layer*, *scene layer*, and *event layer*. As the affordances relate the agent's actions to their effects on the surrounding objects, they can be provided to robots as the prior knowledge so that the robots can have stronger inference ability by relating the affordances to the environment. In our system, top layers can be constructed while obtaining affordance from low layers. With the help of these relations between layers, the decisions made by the robot can be more efficient and robust.

The flow of our system is described as follows. First, during the mapping process, the robot not only builds *static map layer* using Simultaneous Localization and Mapping techniques, but also detects and localizes objects in the *object layer*. Then, the *scene layer* is built through comparing the distances and affordances provided by the *object layer*, and forms regions of the indoor environment. In addition, when robot detects people, it can classify their actions in the *event layer* based on the skeleton detections, the surrounding scene and objects such that the accuracy is improved. As for the navigation process, given a destination from the user, the robot can make the decision of a suitable path by recalling the information from our system and current observations, which is more user-friendly comparing to the shortest path that might interrupt outliers.

III. Methodology

This section introduces the methodology of each layer, which includes static map layer, object layer, scene layer, and event layer. After that, social friendly navigation strategies will be generated based on the proposed architecture.

A. Static Map Layer

To increase the robustness of the indoor localization ability, a methodology for SLAM fusion was proposed in our previous work [13]. The architecture can utilize different and relative weak sensors to achieve robust indoor localization. Thus, with the help of the current SLAM and our localization methods in [13], the robot can perform localization and navigation safely.

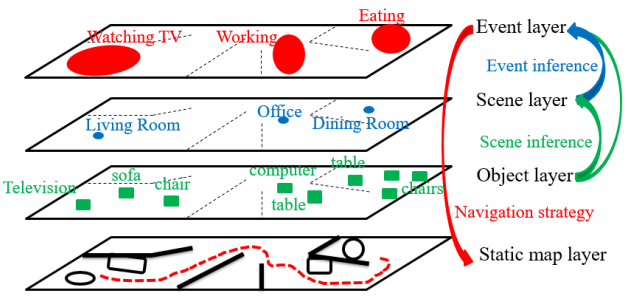


Fig. 2. Architecture of multi-layer environmental affordance map.

B. Object Layer

In order to detect and localize objects on the grid map, we apply the object layer to the system. We use a deep learning method to the object layer for the object detection called YOLOv3 [14]. Besides, the depth camera and the pinhole model are used to find the position in the 3-dimensional space. For the purpose of achieving high efficiency and low computational cost, object segmentation based on depth image is adopted such that the object in the bounding box can be segmented out from the background. Combining with the segmented objects and the 3 dimensional space point cloud, the object position in the space can be determined. While the localization of robot can be obtained from the static map layer, the object positions in the space can then be correlated through the use of coordinate transformation.

In this paper, the experiments are carried out with the humanoid robot *Pepper* developed by SoftBank Group, Corp. [15], which has 2 degree of freedom in the neck ($\theta_{\text{head\_pitch}}$ and $\theta_{\text{head\_yaw}}$) and 0.15 m camera offset. Therefore, a transformation relation corresponding to these two degrees of freedom needs to be applied:

$$\begin{bmatrix} z'_w \\ -x'_w \\ -y'_w \end{bmatrix} = R^T_{head\_pitch} R^T_{head\_yaw} \begin{bmatrix} z_w \\ -x_w \\ -y_w \end{bmatrix} - \begin{bmatrix} 0 \\ camera\_offset \\ 0 \end{bmatrix} \quad (1)$$

where the rotation matrices are

$$R_{head\_pitch} = \begin{bmatrix} \cos\theta_{head\_pitch} & 0 & \sin\theta_{head\_pitch} \\ 0 & 1 & 0 \\ -\sin\theta_{head\_pitch} & 0 & \cos\theta_{head\_pitch} \end{bmatrix}$$

$$R_{head\_yaw} = \begin{bmatrix} \cos\theta_{head\_yaw} & -\sin\theta_{head\_yaw} & 0 \\ \sin\theta_{head\_yaw} & \cos\theta_{head\_yaw} & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (2)$$

$$camera\_offset = 0.15 \text{ m}$$

Finally, with the localization provided in the static map layer, the object coordinates can be further transformed with the transformation matrix from the robot position, including displacement and yaw, under global coordination $(x_{disp}, y_{disp}, \theta_{yaw})$:

$$T_{robot\_pose} = \begin{bmatrix} R_{yaw} & t_{disp} \\ 0 & 1 \end{bmatrix}$$

$$R_{yaw} = \begin{bmatrix} \cos\theta_{yaw} & -\sin\theta_{yaw} & 0 \\ \sin\theta_{yaw} & \cos\theta_{yaw} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; t_{disp} = \begin{bmatrix} x_{disp} \\ y_{disp} \\ 0 \end{bmatrix} \quad (3)$$

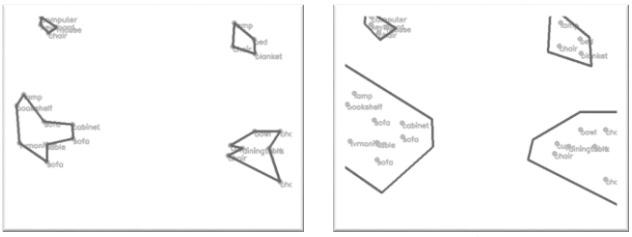
C. Scene Layer

The purpose of scene layer is to formulate the region and recognize the specific scene based on the information from the object layer. The proposed method is based on the fact that human knowledge (affordance) can be applied to help the robot infer the scene efficiently. For instance, a computer is likely to be put in the bedroom or shouldn't appear in the bedroom; a sofa can afford sleeping or only for sitting, which depend on the users' preferences. On the other hand, the prior knowledge (human preference/affordance) can lead the robot to think differently flexibly. In this paper, the scene inference is based on the object map so that not only the scene can be inferred but also its location can be realized. The connection among static map, object map and scene map can help the robot realize where the scenes and objects are in this environment, which improves the navigation strategy afterwards.

After knowing where the objects are spread in the environment from the previous object layer, one can partition the environment into regions according to the location of the objects. This system simply uses a threshold distance to separate the regions, so that the hallway (where is lack of objects) can be isolated. With the presented method, the relation between a scene and the static map can then be established. Algorithm 1 shows the pseudo code for the process of separating the objects into regions given the 2D object map. There are two steps in the program, namely, "group the objects" and "merge the nearby groups". The location of a scene depends on the inference from objects, and therefore, it is necessary to generate the biggest possible region when considering one of the separated groups of objects. We first sort a group of objects in a counter-clockwise order and tries to find the largest convex shape that can be produced by the connection of the objects. The reason we concern the shape as convex is to avoid sharp angles while encircling the scene region. Then, the region will further be expanded considering the volume of those objects. The separation example is shown in Fig. 3.

Algorithm 1: Separate Object with Distance Threshold

Input: 2D object map
Initialize: storages for grouped objects $\boldsymbol{G} = \boldsymbol{\phi}$
distance threshold D_{thres}
// group the objects
For ***object*** in 2D object map:
  If dis(***object***, any object in G_i) < D_{thres}: ($G_i \in \boldsymbol{G}$)
    Store ***object*** in to G_i
  Else:
    Generate a new group G_{n+1} (n is the number of existing group)
// merge the nearby groups
Calculate the mean position of each group ***M***.
For $\boldsymbol{M}_i$ and M_j ($i \neq j$, $i<j$) in ***M***:
  If *distance*(M_i, M_j) < D_{thres}:
    Give number i as label to both M_i and M_j
  Else:
    Give number i as label to M_i
    Give number j as label to M_j
  If multiple labels exist in one group:
    Choose the smaller label for the group.
Merge the same label group.
Return $\boldsymbol{G} = \{G_1, G_2, G_3, \ldots\}$



(a) (b)

Fig. 3 The region separation from not convex shapes to convex shapes. (a) Object connection with order sorting. (b) Object connections in convex shapes.

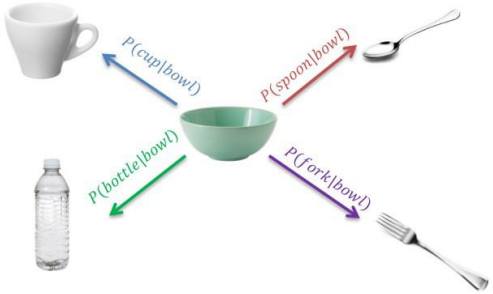


Fig. 4. The example for using affordance to speculate the objects may appear. The probability P implies the existence likelihood of certain objects given the observed item as well as its affordance.

While the object detection method detects only limited classes of object, the robot may have difficulty to recognize all objects in the environment. However, with the help of affordance, some objects are speculated to appear while some objects have already been detected. For instance in Fig. 4, when a bowl is recognized, a cup, a spoon, a fork, and a bottle may have some probabilities to appear at the same time since they have similar eating and drinking functionalities. These probabilities will then be calculated from the given affordance. In this paper, reviewing the target of the present thesis focuses, there are 18 affordances, which are "pourable", "be able to sit on", "can sustain drinks", "can be used to cut", "supportable", "storable", "dining usage", "for entertainment", "for decoration", "for bedding", "office using", "bath using", "for pet", "for cuisine", "edible", "reclinable", "for dressing", and "for toilet".

While the robot has to know the categories of objects under specific scene, the information should be pre-defined. The ways of constructing a knowledge to store such information in the present thesis are two steps: (1) reviewing the object categories that YOLOv3 (pre-train model) has since the property specifies the detection ability the robot possesses, and then categorizing them into possible scenes; (2) searching the suggested or possible furniture which is not able to be recognized by the robot under the specific scenes. With the presence of affordance provided by users, even though there are some objects which cannot be recognized by the robot, they are still be speculated to exist with some possibility, which

may allow the robot to possess better surmise on scene inference and event inference.

After constructing the object and affordance knowledge, we propose a probability model for inferring the scene while the object and affordance knowledge are considered simultaneously, written as Equation (4):

$$P_{S_i} = P(S_i \mid \mathbf{O}) \times \prod_{j=1}^{J} \frac{\sum_{k=1}^{K} P(O_{ja_k} \mid A, O_j) \times P(S_i \mid \mathbf{O}, O_{ja_k})}{Z}$$
$$Z = \sum_{i} \sum_{k=1}^{K} P(O_{ja_k} \mid A, O_j) \times P(S_i \mid \mathbf{O}, O_{ja_k}) \quad (4)$$

The denominator Z represents a normalized factor; J denotes the number of the observed objects; K denotes the number of possessed affordance of the object; A denotes the affordance knowledge set provided by human; S_i denotes the i-th scene; $\mathbf{O}$ denotes the set of observed objects; O_{ja^k} denotes the possible object that can be expected when object O_j is observed; and P_{Si} denotes the probability distribution for scene labeling.

The physical meaning of $P(O_{ja_k} \mid A, O_j)$ describes the probability at which an object may appear given that an object has already been recognized, named as affordance probability. Taking Fig. 4 as an example, when a bowl is recognized by the robot, there exists chances for a cup, a spoon, a bottle, and a fork to appear at the same time. This relation is generated by the affordance. That is to say, a bowl has affordances "pourable," "can sustain drinks," and "dining usage". On the other hand, a fork has the single affordance "dining usage." Consequently, the affordance probability $P(O_{fork} \mid A, O_{bowl})$ is 1/3.

Fig. 5 illustrates an example for descripting how the propose scene inference model and works. The process can be separated into two levels given observed objects O_1 and O_2. The first level inference tries to use the recognized objects to inference the scene, leading to P_{Si}^1. As for the second level, the system first infers the probability of existence of unseen objects through comparing their similarities of affordances with those observed objects. With the help of our second level inference based on the effect of affordance, the robot can speculate more even when some objects are not recognized.

D. Event Layer

The purpose of event layer is trying to combine all the information discussed previously as the affordance in order to improve the robustness of event detection ability. There are lots of works which focus on predicting the human activities or actions based on depth image or skeleton information. However, since a robot is a moving agent, it may lose observation often. Under the lack of observation, a robot cannot make a good judgment. For instance, for a skeleton detection based action recognition system, the system cannot perform well when the skeleton is unseen or noisy. The functionality of event layer then plays an important role as a gatherer by combining all the information discussed above to provide a robust resource for a robot to make decision.

The inputs to the event layer are *Human Pose State* coming from skeleton detection, combining information from

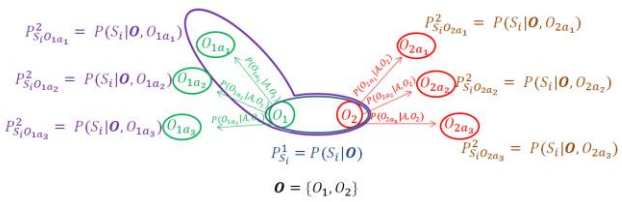


Fig. 5 An illustration of how scene inference works. The robot performs two level scene inference on the basis of $\boldsymbol{O}$

(a) (b) (c)

Fig. 6 The example of engage level. (a) Not engaged: The person is not neither facing to the robot nor turning to the robot. (b) Half engaged: The person is facing to the robot but not turning to the robot. (c) Engaged: The person is facing and turning to the robot at the moment.

previous *Object Layer* and *Scene Layer*. With the combination of different types of resource assigning to event layer, the detection is expected to be robust. Besides, the human's engagement level is also considered. Inspired by the human body languages when humans are interacting each other together [16], this work takes three kinds of human body languages into account to represent the engagement level of a human as shown in Fig. 6.

While inspired by the work [17], the system tries to calculate the angles between limbs. However, as mentioned previously, since the robot is a moving agent, it may not observe enough information all the time, the "Unknown" state is kept consequently. Therefore, the robot will just take the observed information to make the inference instead of guessing on the unseen/undetected joints. Although the "Unknown" state means missing information, with the information from *Object Layer* and *Scene Layer*, the robot can still make good inference.

Also, in order to obtain the engagement level, we derive the relation as Equation (5):

$$\theta_{engaged} = \arccos\left(\frac{shoulder\ length}{C \times arm\ length}\right) \quad (5)$$

if $\theta_{engaged} < \theta_{thres}$ and find human face: turning to the robot.

We define $\theta_{engaged}$ as the engaged angle, with human torso fully facing the robot being zero. The constant C is the ratio between the length of a human's shoulder and arm when human torso is fully facing the robot. By normalizing the shoulder length with arm length, we can calculate the tilt angle of human shoulder with regard to the fully facing one. θ_{thres} is the threshold angle for confirming the human is turning to the robot. Therefore, the robot can detect whether or not the human is likely to interact with it by calculating the engaged angle and check the faces from skeleton detection.

Finally, the system calculates the possible human event

probability based on the information from the *Object Layer*, scene layer, and human pose. The event probability $P(event)$ can separate into three independent probabilities which are $P(event,scene)$, $P(event,object)$, and $P(event,skeleton)$ respectively as shown in Equation (6).

$$P(event) = P(event,scene) \times P(event,object) \times \sigma(P(event,skeleton)) \quad (6)$$

σ: softmax function

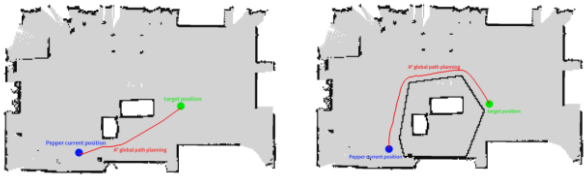
The probability $P(event,scene) = P(event \mid scene) \times P(scene)$ where $P(event \mid scene)$ comes from human prior knowledge that some events may only happen in certain scenes. There are six events: *standing*, *eating*, *talking*, *working*, *sleeping*, and *watching TV*. $P(scene)$ is the scene probability obtained from the scene layer. The probability $P(event,object)$ is calculated by making statistical analysis on the observed objects and the related affordances. The probability and $P(event,skeleton)$ is generated by making statistical analysis on the affordances from observed skeleton. Besides, we apply the softmax function to this term to gain larger weighting.

E. Social Friendly Navigation

The last piece of our system is to conduct social friendly navigation for service robot using the stored knowledge in the multi-layer environmental affordance map. The social friendly navigation focuses on whether to interrupt people or not. The robot will determine whether a location is able to be interrupted by checking human's engagement level, the scenes and the events occurring that stored in the multi-layer environmental affordance map. While the human's engagement level is "not engaged" or the scene is "office", or the events are "sleeping" and "working", the separated region will be formed as virtual obstacles so that A* algorithm will generate alternative path which will not intrude the regions as Fig. 7 shows.

IV. Experiments and Results

In this section, we implement the proposed multi-layer environmental affordance map on real robot in the testing scenario to validate the performance. The robot we use is the social interaction robot Pepper equipped with RGBD camera and laser range finder. The sensing data from Pepper are sent to laptop server with Intel® Core™ i7-8550U (1.80 GHz x 8) CPU and a single GPU (Nvidia GeForce GTX 980) through wireless connection.



(a) (b)

Fig. 7 Illustration of social friendly navigation. The mobile robot is trying to move from the current location (blue point) to the target position (green point). (a) is the shortest global path in the static map. (b) is the global path considering region intrusion from scene layer where the robot observes human inside is unwilling to be interrupted.

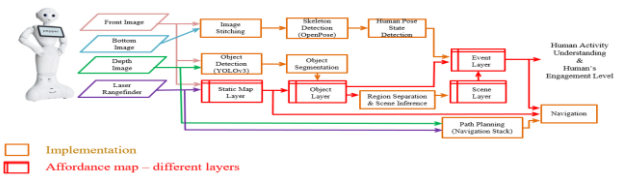


Fig. 8 The implementation details. The system takes robot sensor data as input, generates perceptions through functions in brown boxes, organizes information with layers and outputs social friendly navigation.

TABLE I. EVENT INFERENCE RESULTS

	Skeleton Detection				Object Layer		
							
	stand	*eat*	*talk*	*work*	*sleep*	*watch TV*	*Result*
$P(event,scene)$	**0.33**	**0.33**	**0.33**	<0.01	<0.01	<0.01	-
$P(event,object)$	< 0.01	**0.5**	<0.01	0.25	<0.01	0.25	eat
$P(event,skeleton)$	< 0.01	<0.01	<0.01	**0.5**	<0.01	**0.5**	work/ watch TV
$P(event)$	< 0.01	**0.94**	<0.01	<0.01	<0.01	0.06	**eat**
Ground Truth							**eat**

The whole system is built under Robot Operating System (ROS). The implementation detail is shown in Fig. 8. The inputs are the RGB images from the front and bottom camera, the depth image from the depth camera, and the laser range finder. The boxes in orange color are the functions for perceptions, while the red ones are the multi-layer environmental affordance map that organizes the information. The output of our system is the human activity understanding and human engagement level which can be applied for social friendly navigation. SLAM fusion and the robust localization in the static map layer have been addressed in our previous work [13].

A. Human Pose, Engagement, and Event in Event Layer

With the help in [18], we are able to robustly obtain human skeleton. Through our human pose detection algorithm, we can infer human event by information revealed from posture. TABLE I shows the event inference results in a dining room. The objects detected by Pepper robot are: chair, dining table, potted plant, fork, cup, and spoon. The scene inferred by Pepper is dining room with confidence of 0.9997. The final event detections, based on all the information from our system, are eating and standing with probabilities of 0.94 and 0.63 respectively.

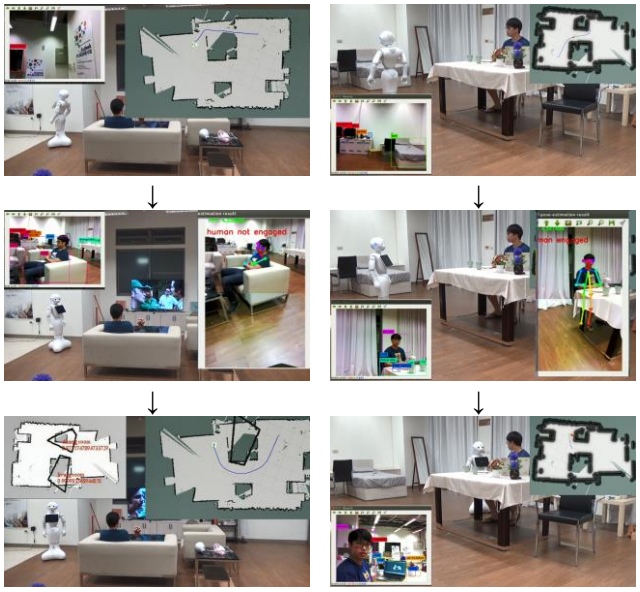
B. Social Friendly Navigation with Proposed Architecture

A practical application for the proposed multi-layer

environmental affordance map is to generate social friendly navigation that is suitable in the household environment. The follows are the demonstration scenarios for our system.

First is the case that Pepper tries to approach to the master in the office without interrupting the person watching television in the living room, as shown in Fig. 9(a). After Pepper finds out the shortest path, it soon recognizes that it is about to cross the living room based on the information provided by the scene layer. Furthermore, Pepper also detects that there is a human watching TV with low engagement level according to the event layer. As a result, it marks the living room region as blocked and regenerate a social friendly navigation strategy. The second demonstration showing in Fig. 9(b) is that the master calling Pepper in the dining room.

While Pepper stops and finds out the master is facing toward it, meaning high engagement level. Thus, it will mark the dining room as traversable and move toward human. We then provide a short video clip to demonstrate that Pepper can perform social friendly navigation robustly in real time based on the proposed system.



(a) (b)

Fig. 9 The demonstration for our social friendly navigation based on the multi-layer environmental affordance map architecture. Figures in the left column are the first case that Pepper navigates to the office without interrupting the person watching television. Figures in the right column are the second case that Pepper stops its original navigation and moves toward human for further interactions.

V. Conclusion

In this paper, a multi-layer environmental affordance map architecture which includes "static map layer", "object layer", "scene layer", and "event layer" is proposed. The architecture is designed to achieve robust indoor localization, scene localization, human event detection and socially friendly navigation in an efficient way under limited computational resource. The main characteristic of the proposed architecture is that the observation and inference results can be organized and stored efficiently such that the robot is capable of generating high-level navigation strategies. The experimental results show that the robot can not only move robustly but also reinforce its perception and inference abilities based on the multi-layer environmental affordance map. In the future, we expect to handle more practical HRI tasks robustly and accurately using the proposed architecture.

VI. Acknowledgment

This research was supported by the Joint Research Center for AI Technology and All Vista Healthcare under Ministry of Science and Technology of Taiwan, and Center for Artificial Intelligence & Advanced Robotics, National Taiwan University, under the grant numbers of 108-2634-F-002-016, 108-2634-F-002-017 and 108-2218-E-027-014.

References

[1] C. Y.Yang, M. J.Lu, S. H.Tseng, and L. C.Fu, "A companion robot for daily care of elders based on homeostasis," *2017 56th Annu. Conf. Soc. Instrum. Control Eng. Japan, SICE 2017*, vol. 2017–Novem, pp. 1401–1406, 2017.

[2] D.Hebesberger, T.Koertner, C.Gisinger, J.Pripfl, and C.Dondrup, "Lessons learned from the deployment of a long-term autonomous robot as companion in physical therapy for older adults with dementia: A mixed methods study," *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2016–April, pp. 27–34, 2016.

[3] X. T.Truong and T. D.Ngo, "Toward Socially Aware Robot Navigation in Dynamic and Crowded Environments: A Proactive Social Motion Model," *IEEE Trans. Autom. Sci. Eng.*, vol. 14, no. 4, pp. 1743–1760, 2017.

[4] T.Kruse, A. K.Pandey, R.Alami, and A.Kirsch, "Human-aware robot navigation: A survey," *Rob. Auton. Syst.*, vol. 61, no. 12, pp. 1726–1743, 2013.

[5] R.Alterovitz, S.Koenig, and M.Likhachev, "Robot Planning in the Real World: Research Challenges and Opportunities," *AI Mag.*, vol. 37, no. 2, p. 76, 2017.

[6] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael, "Formal Basis for the Heuristic Determination eijj ," *Syst. Sci. Cybern.*, no. 2, pp. 100–107, 1968.

[7] B.Okal and K. O.Arras, "Learning Socially Normative Robot Navigation Behaviors with Bayesian Inverse Reinforcement Learning."

[8] D. S.Wettergreen and T. D. B.Editors, *Field and Service Robotics*. 2011.

[9] E.Şahin, *State-of-the-Art and Formalization for Robotics*, no. December 2007. 2008.

[10] S.Vasudevan, S.Gächter, M.Berger, and R.Siegwart, "Cognitive Maps for Mobile Robots – An Object based Approach," in *IEEE International Conference on Intelligent Robots and Systems (IROS)*.

[11] H. S.Koppula and A.Saxena, "Anticipating Human Activities Using Object Affordances for Reactive Robotic Response," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 14–29, 2016.

[12] K. M.Varadarajan, "Topological mapping for robot navigation using affordance features," *ICARA 2015 - Proc. 2015 6th Int. Conf. Autom. Robot. Appl.*, pp. 42–49, 2015.

[13] S.H.Chan, P.T.Wu, and L.C.Fu, "Robust 2D Indoor Localization Through Laser SLAM and Visual SLAM Fusion," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018, pp. 1263–1268.

[14] J.Redmon and A.Farhadi, "YOLOv3: An Incremental Improvement," *CoRR*, vol. abs/1804.0, 2018. [15] V.Perera, T.Pereira, J.Connell, andM.Veloso, "Setting Up Pepper For Autonomous Navigation And Personalized Interaction With Users," 2017.

[15] Pepper Aldebaran Documentation of NAOqi 2.4.3 Pepper, 2017, [online] Available: doc.aldebaran.com/2-4/home_pepper.html.

[16] C.Rich, B.Ponsleur, A.Holroyd, and C. L.Sidner, "Recognizing engagement in human-robot interaction," *2010 5th ACM/IEEE Int. Conf. Human-Robot Interact.*, p. 375, 2010.

[17] Z.Uddin, S.Member, N. D.Thang, and T.Kim, "Human Activity Recognition via 3-D Joint Angle Features and Hidden Markov Models," *Training*, pp. 713–716, 2010.

[18] Z.Cao, T.Simon, S.-E.Wei, and Y.Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *CVPR*, 2017.