2017 6th National Conference on Technology and Management (NCTM) January 27, 2017. Malabe, Sri Lanka.

A Multi-modal Approach for Enhancing Object Placement

P.H.D. Arjuna S. Srimal and A.G. Buddhika P. Jayasekara
Robotics and Control Laboratory
Department of Electrical Engineering
University of Moratuwa
Moratuwa 10400, Sri Lanka
arjuna@elect.mrt.ac.lk, buddhika@elect.mrt.ac.lk

***Abstract*—Voice commands have been used as the basic method of interaction between humans and robots over the years. Voice interaction is natural and require no additional technical knowledge. But while using voice commands humans frequently use uncertain information. In the case of object manipulation on a table, frequently used uncertain terms "Left, "Right", "Middle", "Front"...etc. These terms fail to depict an exact location on the table and the interpretation is governed by the robots point of view. Depending solely on vocal cues is not ideal as it requires the users to explain the exact location with more words and phrases making the interaction process cumbersome and less human like. However, using hand gestures to pinpoint the location is as natural as using the voice commands and frequently used when manipulating items on a surface. When compared to voice commands use of hand gestures is a more direct and less cumbersome approach. But when used alone hand gestures can result in errors while extracting the pointed location making the user dissatisfied. This paper proposes a multi-modal interaction method which uses hand gestures combined with voice commands to interpret uncertain information when placing an object on a table. Two fuzzy inference systems have been used to interpret the uncertain terms related to the two axes of the table.The proposed system has been implemented on an assistive robot platform. Experiments have been conducted to analyze the behaviour of the system.**

***Index Terms*—multi-modal human robot interaction, object manipulation, interpreting uncertain information**

.

I. Introduction

The elderly population has increased dramatically in the recent years. And it is growing even faster than ever in the history [1]. The main problem associated with this rapid growth lies with the augmented requirement of assistance needed to guide and support the elderly in their day to day activities. Owing to this a high concern has been exerted upon how to incorporate technology to assist the elderly in domestic environment. In this context, a growing body of literature highlights the importance of deploying domestic service robots designed specially for human assistance [2]–[7].

The main challenge is to enhance the quality of the interactions in between human and robot while assisting the human with higher integrity. If encountered properly, this concept would engender an era of more human-like robots which can serve the elderly in their day to day activities while providing both physical support and cognitive assistance. Furthermore the person requiring assistance should be able to control the robot with least amount of technical knowledge [3]. Voice based human interactions mostly include uncertain terms which habitually emphasize the qualitative information rather than the quantitative information. These qualitative instructions mostly include uncertain terms . Thus the competence of the robot to respond after properly evaluating these uncertain terms must be ensured when it is operated solely on voice commands. Even though voice commands are known to be the basic method of interaction between humans and robots, when describing spatial information humans tend to use both hand gestures and voice commands which significantly reduces the number of words required to explain a task. For most object placement cases direct pointing has been much more effective rather than voice commanding. For an example a scenario when a robot is being asked to fetch something from identical number of items can be taken. Here direct pointing would easily solve the complexity while instructing the robot to fetch the required item compared to vocal commands. Thus the importance of designing service robots which can be operated by combined voice and gesture commands must come under close scrutiny as it can effectively help the elderly to correlate among vocal instructions and spatial locations; an ability that the elderly get deprived of with age. Furthermore this approach will motivate the elderly to get accustomed to these human like systems and will positively influence them to seek help of service robots more [5].

Most of the research that have been done in order to build more human like robots in domestic environment, showcase attempts to understand and interpret uncertain information based on spatial and environmental factors. Deployment of fuzzy logic based approaches to aid in with the complex task of understanding uncertain information has been attempted successfully in [8]–[11]. However the drawbacks imposed by the unimodal interaction model have curtailed the effectiveness of these systems. The deployment of gestures based interactions can be found in [4], [5], [12] whereas more human like systems that can identify hand pointing gestures have been developed in [4], [12]. In these systems RGB - D camera is used to get the information of the hand pointing. However using 3D depth sensing and OPEN NI to perform skeletal tracking to

978-1-5090-4729-1/17/$31.00 ©2017 IEEE

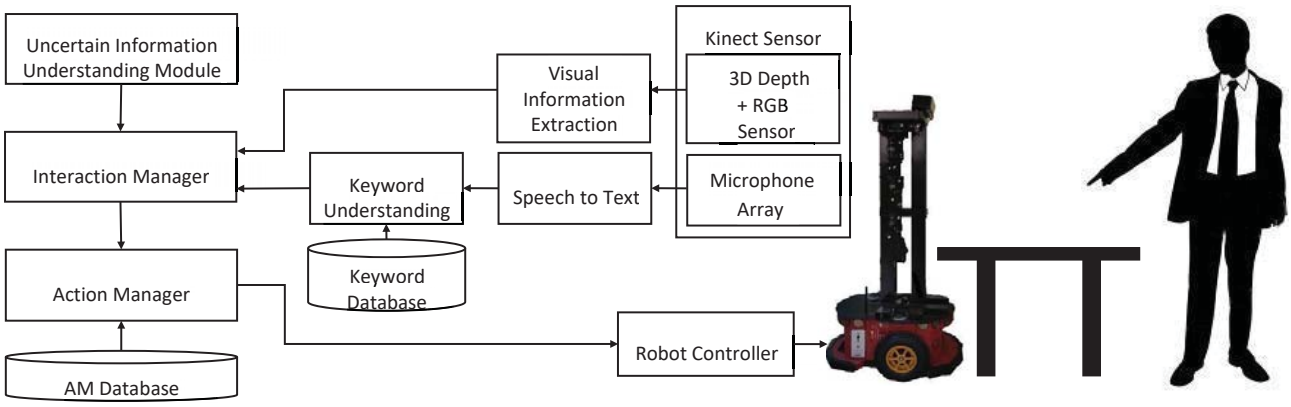

Fig. 1. System Overview

obtain the body joints is the most commonly used method. In household environment, objects manipulation on planar surfaces is a frequent task. Object manipulation on a table has been implemented on the scale of domestic service robots in [12]. Here the 3D depth data is used to identify the tabletop whereas segmentation is used to understand the objects placed on the table. Despite the effectiveness, this system lacks the ability to understand and respond to uncertain information which is a critical factor when interacting with humans. This paper suggests a methods to use multimodal interaction with the human user to understand uncertain information related to object placement on a table. Proposed system incorporates two fuzzy inference systems to interpret two axes of the table. Section II of this paper presents the details of the system. Uncertain information understanding is described in section III. Section IV provides details of the experiments and results while section V concludes the paper.

II. System Overview

Overview of the system is shown in Fig.1. The goal of the system is to provide an effective way to understand uncertain information related to object manipulation on a table surface. The main components of the system and their tasks are described in following subsections.

A. Visual Information Extraction

Visual information is extracted under two sub categories. They are extraction of the hand gesture pointing and extraction of table surface information.

1) Table Surface Information: The input is taken from the RGB color camera of the Kinect sensor. The information is processed to identify the width and the length of the table.

2) Hand Gesture Pointing: The Kinect sensor provides skeleton data for 25 joints on the body. When pointing their hand towards an object humans place their hand along the line of the object and their eyes. This is shown in the Fig.2(a). In the skeleton modal the corresponding vector goes through the head joint and the palm joint as shown in Fig.2(b). The Kinect gives the 3D coordinates of the body joints reference to the center of the Kinect as the origin. The position and the

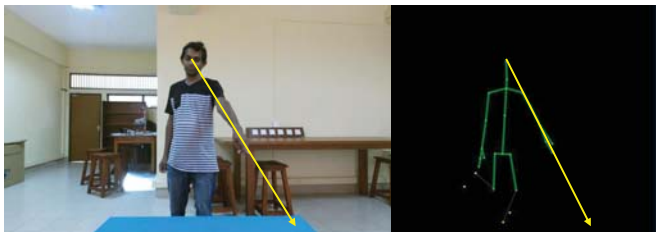

Fig. 2. Vector extended through the head joint and the palm joint is used to get the direction of pointing.

orientation of the table is fixed. The coordinates of the position where the hand is pointing is taken by the location where the extended vector is intersecting the table plane.

B. Keywords Understanding Module

The language used by humans contains various words and lexical symbols, which makes it difficult for the system to clearly understand the expressed idea. As a solution to this problem, the system extracts the keywords found in the vocal command. In an object placement task words expressing spatial information are identified as keywords. For example, “Left”, “Middle”, “Center”...etc. Implementing a system that can extract such keywords helps the users to give commands to the system without adhering to a strict grammar rule base.

Keywords understanding module uses a keyword database which contains the synonyms for the most frequently used keywords. For example, for the keyword “Middle” , “Center” can be used as a synonym. This helps the users to widen their vocabulary usage rather than adhering to a regulated set of commands. Namely there are two types of keywords; “Action keywords” and “Spatial Keywords”. The action keywords include words that are used to command a task, such as “Keep”, “Move”, “Place”...etc.

The spatial keywords include spatial information for the positioning of the object,such as “Left”,“Right”,“Middle”,“Front”,“Back”and“Corners”. The

main classified table areas according to the spatial keywords are shown in Fig.3.

C. Uncertain Information Understating Module

This module interprets the uncertain information in the commands.This contains three submodules as following

1) Voice Based Positioning Module: This module deals with the placement of objects when they are presented with voice based position information.

2) Hand Gesture Based Positioning Module: Hand gesture based position information is handled by this module.

3) Combined Positioning Module: When the position information is given by a combined voice and gesture based command, this module is used to interpret the location of the object.

The detailed explanation of this UIU module is given under section III.

D. Interaction Manager

This module handles the interactive tasks between the robot and human counterpart. There are three identified set of commands for the three submodules mentioned in UIUM, that the user can use in order to convey the positioning information of the object. They are

- Voice based position commands.
- Hand gesture based position commands.
- Combined positioning commands.

Here the combining positional commands give a combination of the voice based and hand gesture based positioning information. This module identifies theses three categories based on the keywords provided and the direction where the hand gesture is pointing. This modules searches for the action keywords in the received vocal commands. If there are action keywords with spatial keywords and hand gesture positioning then they command type will combined positing command. If there are only spatial keywords available with the action keyword then it will be a voice based positioning command. If there are only hand gesture position information available then it will be taken as hand gesture based command. If either of these does not occur they module will ignore the keyword. This way they system will not react to the unwanted dialogues and hand gestures it receives.

E. Action Manager

This module manages the actions that are to be performed when placing the object on the table. It includes two sub modules. One module for the controlling of the robot manipulator and the other for simple navigational tasks.

1) Navigational Manager: The placement of the table and the starting location of the robot is mapped inside the Action Manager Database using Mapper3 software. This module helps the robot to plan a collision free path around the table when it is required to reach the table in different directions. This is required since the robots reachability depends on how close it can get to the table in different directions.

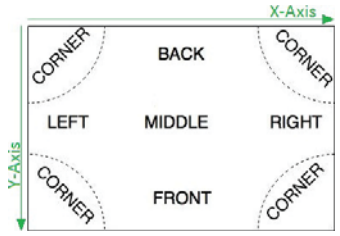

Fig. 3. Figure shows the commonly classified areas of a table.

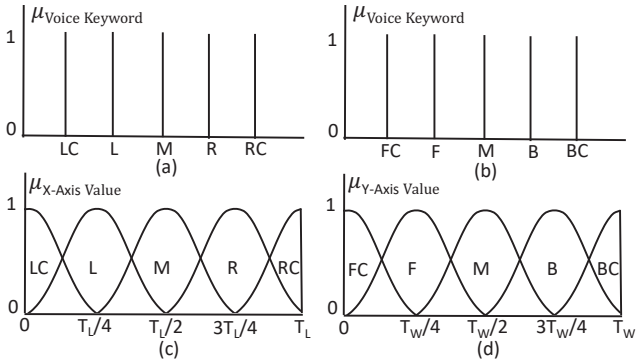

Fig. 4. (a) and (b) shows the input membership functions for the module 1 in UIUM. (c) and (d) shows the output membership functions for the module 1. Where T_L and T_W are table length and table width. Fuzzy labels are defined as LC:Left+Corner, L:Left, M:Middle, R:Right, RC:Right+Corner, FC:Front+Corner, F:Front, M:Middle, B:Back, BC:Back+Corner

2) Object Handling Manager: The UIUM returns a coordinate of the table where the item has to be placed. So in order to place the item without colliding with the table, the path of manipulator's end effector has to be planed. This submodule manages the end effector's path as well as the objects handling and placing.

F. Robot Controller

This module handles the interfacing between the software and hardware of the system. Both the navigation platform and the manipulator is controlled via microcomputer based controllers. This module interfaces them together.

TABLE I
FUZZY RULE BASE FOR VOICE BASED POSITION INFORMATION.

X-Axis					
Input Memberships	*Voice Command Keyword*				
	Left+ Corner	*Left*	*Middle*	*Right*	*Right+ Corner*
Output Memberships	VL	L	M	R	VR
Y-Axis					
Input memberships	*Voice Command Keyword*				
	Front+ Corner	*Front*	*Middle*	*Back*	*Back+ Corner*
Output Memberships	VF	F	M	B	VB

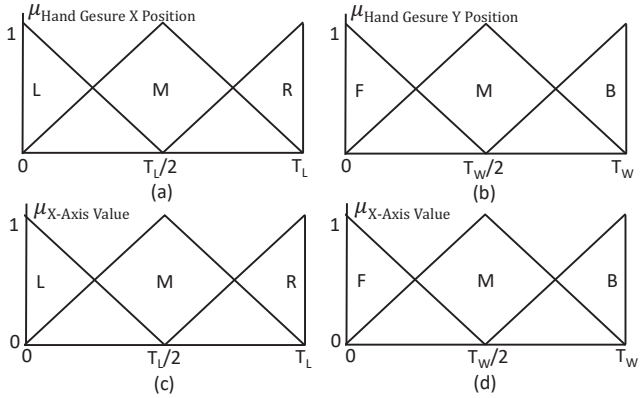

Fig. 5. (a) and (b) shows the input membership functions for the module 2 in UIUM. (c) and (d) shows the output membership functions for the module 2. Where T_L and T_W are table length and table width. Fuzzy labels are defined as L:Left, M:Middle, R:Right, , F:Front, M:Middle, B:Back, BC:Back+Corner

TABLE II
FUZZY RULE BASE FOR GESTURE BASED POSITION INFORMATION.

X-Axis			
Input Memberships	*Hand Gesture*		
	Left	*Middle*	*Right*
Output Memberships	L	M	R
Y-Axis			
Input memberships	*Hand Gesture*		
	Front	*Middle*	*Back*
Output Memberships	F	M	B

III. EVALUATION OF UNCERTAIN INFORMATION

Understating uncertain information is implemented in three sub modules. They are for voice based position information, gesture based position information and for combination of those two.

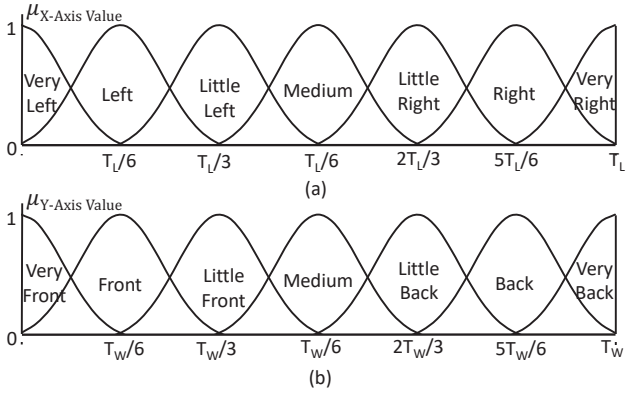

Fig. 6. (a) shows the output membership function for the X-Axis for the module 3 in UIUM and (b) shows the output membership function for the Y-Axis for the module 3.Where T_L and T_W are table length and table width.

TABLE III
FUZZY RULE BASE FOR COMBINED POSITION INFORMATION.

X-Axis						
Input Memberships		*Voice Command Keyword*				
		Left+ Corner	*Left*	*Middle*	*Right*	*Right+ Corner*
Hand Gesture	*Left*	VL	L	LL	-	-
	Middle	L	LL	M	LR	R
	Right	-	-	LR	R	VR
Y-Axis						
Input Memberships		*Voice Command Keyword*				
		Front+ Corner	*Front*	*Middle*	*Back*	*Back+ Corner*
Hand Gesture	*Front*	VF	F	LF	-	-
	Middle	F	LF	M	LB	B
	Back	-	-	LB	B	VB

A. Voice Based Position Information.

This module deals with commands that provide position information based on voice. For example, "place the item on the front left corner of the table" can be considered. For this type of commands, the fuzzy membership functions shown in Fig.4 is used. The input membership functions are singleton functions since they are classified based on the spatial keywords. They are shown in Fig.4 (a) and (b). Shown in (a) are the input membership functions for the X-Axis of the table as shown in Fig.3. Shown in (b) are the membership functions for Y-Axis of the table. Conceder an example command "Keep the item on the left front corner"; membership function "Left + Corner" will be selected as the X-Axis input and "Front + Corner" will be selected for the Y-Axis input. The output membership functions are shown in Fig.4 (c) and (d). Gaussian shaped output membership functions were used because the distribution of each area doesnt have strict boundaries and due to the fact that they are much suitable to represent natural human tendencies. The rule base for the two fuzzy inference systems are given in Table I.

B. Hand Gesture Based Position Information

Even though this type of position information provides a specific coordinate on the table surface, the determination of the exact location would be uncertain to some extent. In most cases owing to human errors and system errors, the exact location desired by the user may not be extracted. By using a fuzzy inference system to interpret the hand pointing location, these errors can be minimized. The input membership functions of the fuzzy inference system are given in Fig.5(a) and(b). The output membership functions are given by Fig.5(c) and (d). Here triangular shaped output membership functions are used ,as they represent distances. Rule base for these two fuzzy systems are given in TableII.

C. Combined Position Information

As mentioned before this module extracts position information from both hand gestures and voice based commands. The two fuzzy inference systems used for this module consists of same input membership functions used in module 1 and 2 shown in Fig.4(a),(b) and Fig.5(a),(b). The output membership

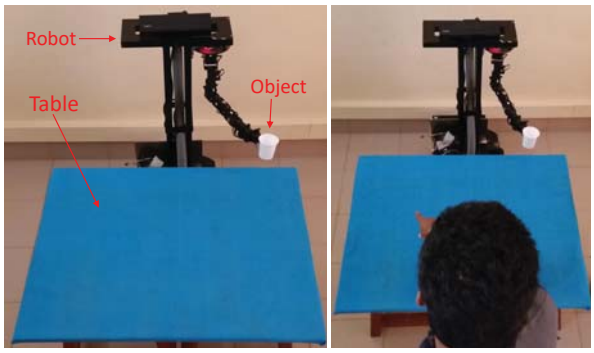

Fig. 7. Figure (a) shows the table that was used for the experiments where the robot platform is at the stating position with respect to the table and the the robot is holding object which was manipulated. (b) shows the user giving a hand gesture command.

functions are depicted in Fig.6 and the rule base for this module is given in TableIII.

IV. Results and Discussion

A. Research Platform and Experiments

The system has been implemented on Mirob platform described in [13].The experiments have been carried out on a domestic writing table with table length T_L of 900mm and table width T_W of 710mm. The performance of the system was measured by giving position information using all three methods described in section III. The selected set of results from the experiments are presented in table that highlights the key facts of the systems behaviour.

B. Behaviour of The System

The experiments were conducted by placing the robot in the starting position shown in Fig.7(a). and giving positioning commands to place the objects. The robot then performs the task and come back to its starting position. Table IV shows the results of selected set of commands. The spatial keywords for voice based or combined commands are given in the tab "Spatial Keywords". The users were asked to show a desired place for the objects to be placed. This location was recorded in the X and Y coordinates and are given in the tab "Desired Location" in mm. Then they were told to give the robot commands to place the object using three submodules described in section III. After the placement was done, the user was asked to give feedback on the performance of the robot, which is given under the column "User Rating"of the Table IV. If the user was satisfied about the system outcome the result is marked as "OK". For commands which contain voice based positioning information the keywords are given in the table and for commands involving hand gesture positioning information, the location of the detected hand gesture position on the table is given in mm. All the distance values are in reference to the X and Y axis shown in the Fig.8. The resulting position and the desired position of the object is shown in the Fig.8 as a object map. The blue "X" show the desired position and the red "X" shows the actual placement. Following key features can be identified by analyzing the results of the system.

When placing the object just depending on the voice based portion information, it can be clearly seen that some locations of the table is not reachable. As for example the commands 1 and 7 can be pointed out. There the desired position of the object is (320,243), but when the only voice commands were used the resulting position was (450,355) since there is no exact way to describe the location of the object. But when hand gestures were used together with the voice positioning information, much better outcome has resulted. Using combined positioning information provide a better outcome even in cases where the voice positioning information is able to provide a reasonably acceptable positioning. This can be seen when comparing commands 2 and 5 also in commands 3 and 6.

Another main concern is the safety of the object that is to be placed. Specially when using hand gesture based positioning, human errors or system errors can result in placing the object on a vulnerable location, for example placing the item on the edge of the table where the object can fall down after placement. The usage of the fuzzy inference systems can avoid this kind of misinterpretations. Command 8 shows this type of scenario. But when using both hand gestures and voice commands the item can be placed with a better confidence and better positioning. Command 9 can be pointed out as an example for this case. Even though the hand gesture is as the same as the one used in command 8, the desired location can be adjusted with the usage of the voice information.

Commands 10,11 and 12 shows a scenario where the hand gesture position is considerably off from the desired location, But when the system is used with both voice and gesture information, the posting of the object is comparatively better. This can conclude that depending purely on gesture based positioning can provide unsatisfying results. When the system is presented with a wrong voice keyword can result in a wrong placement of the object. but when hand gestures are combined, this erroneous placement can be minimized. In command 4 the correct keywords should be "Right" and "Back" but the keywords "Middle" and "Back" has been used. But since the hand gesture was used the item has position has moved toward the desired location. Further more some specific locations on the table like the center, can be easily reached by using voice commands rather than combining them with hand gestures or just using hand gestures. This scenario is shown by commands 13 and 14.

Commands that only used voice based positional information gave successful results for 60% while hand gesture based positional information commands were satisfactory for 66% and combined positional information gave satisfactory rate of 83.3%. Therefore, it can be concluded that the uncertain information understanding capability of the system has been improved when the voice position information is combined with the hand gesture position information. The user satisfaction has been improved in the cases where the combined commands have been used.

TABLE IV
MY CAPTION

Command No.	UIU Submodule	Desired Location (mm)	Spatial Keywords		Hand Gesture	System Output (mm)	User Rating
			X Axis	Y Axis			
1	1	(320,243)	Middle	Middle	-	(450,355)	NOT OK
2	1	(110,590)	Left+Corner	Front+Corner	-	(73,652)	NOT OK
3	1	(640,384)	Right	Middle	-	(673,355)	OK
4	3	(520,176)	Middle	Back	(511,184)	(625,180)	NOT OK
5	3	(110,590)	Left+Corner	Front+Corner	(119,442)	(107,594)	OK
6	3	(640,384)	Right	Middle	(561,422)	(641,383)	OK
7	2	(320,243)	-	-	(132,89)	(333,249)	OK
8	2	(810,640)	-	-	(920,780)	(753,594)	NOT OK
9	3	(810,640)	Right+Corner	Front+Corner	(920,780)	(852,672)	OK
10	3	(280,350)	Left	Middle	(361,310)	(266,335)	OK
11	2	(280,350)	-	-	(361,310)	(441,351)	OK
12	1	(280,350)	Left	Middle	-	(226,355)	OK
13	3	(450,355)	Middle	Middle	(356,303)	(412,333)	OK
14	1	(450,355)	Middle	Middle	-	(450,355)	OK

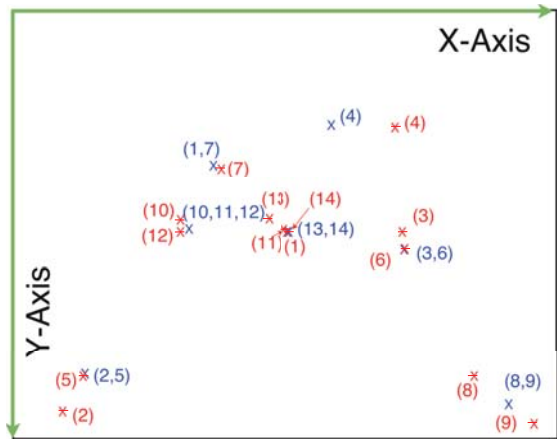

Fig. 8. Positions of the object placements and the desired positions of the user that were recorded during the experiments. Shown by blue 'X' are the desired position and the red '*' shows the position of placement by the robot.

V. CONCLUSION

A multi-modal system has been proposed that can place the objects on desired position of a table effectively using fuzzy based approach. The system can interpret uncertain user commands using hand gesture positioning information and voice information. A novel approach to identify hand pointing position has been introduced. The voice command interface is not relying on a strict grammar rule which in turn give the ability to the user to use language freely. This system improves the human-like object placement capability of the robot hence improve the human robot interaction.

ACKNOWLEDGMENT

This work was supported by University of Moratuwa Senate Research Grant Number SRC/CAP/14/16 and SRC/CAP/16/03.

REFERENCES

[1] "World population ageing 2015," Population Division, Department of Economic and Social Affairs, United Nations, ST/ESA/SER.A/390, 2015.

[2] G. Bugmann and S. N. Copleston, "What can a personal robot do for you?" in *Towards Autonomous Robotic Systems*, 2011, pp. 360–371.

[3] J. Forlizzi and C. DiSalvo, "Service robots in the domestic environment: a study of the roomba vacuum in the home," in *Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction*. ACM, 2006, pp. 258–265.

[4] D. Whitney, M. Eldon, J. Oberlin, and S. Tellex, "Interpreting multimodal referring expressions in real time," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 3331–3338.

[5] H. Osawa, J. Orszulak, K. M. Godfrey, M. Imai, and J. F. Coughlin, "Improving voice interaction for older people using an attachable gesture robot," in *19th International Symposium in Robot and Human Interactive Communication*. IEEE, 2010, pp. 179–184.

[6] T. Borangiu, "Advances in robot design and intelligent control," *Switzerland: Springer International Publishing*, 2014.

[7] P. Tsarouchi, S. Makris, and G. Chryssolouris, "Human–robot interaction review and challenges on task planning and programming," *International Journal of Computer Integrated Manufacturing*, vol. 29, no. 8, pp. 916–931, 2016.

[8] A. B. P. Jayasekara, Watanabe, K. Watanabe, K. Kiguchi, and K. Izumi, "Interpretation of fuzzy voice commands for robots based on vocal cues guided by user's willingness," in *2010 IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, 2010, pp. 778–783.

[9] M. A. V. J. Muthugala and A. G. B. P. Jayasekara, "Interpreting fuzzy linguistic information in user commands by analyzing movement restrictions in the surrounding environment," in *2015 Moratuwa Engineering Research Conference (MERCon)*, April 2015, pp. 124–129.

[10] A. B. P. Jayasekara, K. Watanabe, K. Kiguchi, and K. Izumi, "Interpreting fuzzy linguistic information by acquiring robot's experience based on internal rehearsal," *Journal of System Design and Dynamics*, vol. 4, no. 2, pp. 297–313, 2010.

[11] S. Schiffer, A. Ferrein, and G. Lakemeyer, "Fuzzy representations and control for domestic service robots in golog," in *Intelligent Robotics and Applications*. Springer, 2011, pp. 241–250.

[12] J. Stückler, D. Droeschel, K. Gräve, D. Holz, J. Kläß, M. Schreiber, R. Steffens, and S. Behnke, "Towards robust mobility, flexible object manipulation, and intuitive multimodal interaction for domestic service robots," in *Robot Soccer World Cup*. Springer, 2011, pp. 51–62.

[13] M. V. J. Muthugala and A. B. P. Jayasekara, "Mirob: An intelligent service robot that learns from interactive discussions while handling uncertain information in user instructions," in *2016 Moratuwa Engineering Research Conference (MERCon)*. IEEE, 2016, pp. 397–402.