Commanding a humanoid to move objects in a multimodal language

T. Oka*, K. Sugita and M. Yokota

**Nihon University, 1-2-1 Izumicho, Narashino, Chiba, 275-8575 JAPAN*
Fukuoka Institute of Technology, 3-30-1Wajiro-higashi, Higashi-ku, Fukuoka, 811-0295 JAPAN
*(*Tel : 81-47-474-9693; Fax : 81-47-474-2669)*
(oka.tetsushi@nihon-u.ac.jp, sugita@fit.ac.jp, yokota@fit.ac.jp)

Abstract: This paper describes a study on a humanoid robot that moves objects on the requests of its users. The robot understands commands in a multimodal language which combines spoken messages and two types of hand gestures. All of the ten novice users directed the robot using gestures when they were asked to spontaneously direct the robot to move objects after learning the language for a short period of time. The success rate of multimodal commands was over 90 % and the users completed their tasks without trouble. They thought that gestures were more preferable than and as easy as verbal phrases to inform the robot action parameters such as direction, angle, step, width, and height. The results of the study show that the language is fairly easy for non-experts to learn and can be more effective for directing humanoids to move objects by sophisticating the language and our gesture detector.

Keywords: humanoid, multimodal, command language speech, gesture

I. INTRODUCTION

In recent years, various humanoid robots have been developed for the purpose of realizing robots which work for humans in homes, offices, hospitals, etc. Humanoids have advantages for a multi-purpose robot which helps people. As they look like humans and their structures are similar to us, it is easier for us to communicate with them and for them to work in our environments. On the other hand, because they have at least 15 degrees of freedom, it is difficult to operate them with a conventional interface device. Thus, humanoids in the future need certain autonomy and a new kind of intuitive user interface.

The authors have been developing a multimodal command language for home robot users which combines speech, gestures, body touches, and key press actions [1] and conducting studies on robots including humanoids [2] that can be directed in the language. The results of these studies show that the language can be useful for realizing cost-effective home-use robots for various purposes. This study focuses on combining speech and hand gestures in order to direct humanoids to move objects such as boxes and chairs.

II. MULTIMODAL LANGUAGE

The multimodal command language is based on the Japanese language and two types of hand gestures. It is a set of multimodal commands which consist of a spoken message and a hand gesture. The language is defined by a grammar for spoken messages and a set of gesture events, which enables Japanese speakers to command a humanoid in a fairly natural way to pick up and place objects such as boxes, take steps forward and backward, turn left and right, step aside, and push and pull chairs. Table 1 shows actions that can be commanded in the language.

The grammar for spoken messages defines a set of spoken commands including words to specify an action. Thus, one can command a humanoid robot by giving a spoken command without a gesture in the language.

A single hand waving gesture generates a gesture event containing three parameter values: *direction*, *amplitude*, and *count* values (see Table 2). Single hand gesture events substitute spoken phrases to convey action parameter values such as *step*, *direction*, and *angle* values in Table 1. For instance, a single hand movement to the right means "to the right" for action types *sidestep*, *turn*, and *slide*. Table 3 shows the mapping of the *amplitude* and *count* values of gesture events to the *step* and *angle* values of actions.

A both hand gesture event occurs when a user moves the hands simultaneously up and down. It contains two parameter values, *distance* and *count*, and conveys action parameter values, *width* and *height* (Tables 1 and 2). The *distance* value of a both hand gesture specifies the size of an object, *small*, *medium* (the distance between the shoulders of the robot itself), or *large*, to be picked up. The count of a both hand gesture conveys

©ISAROB 2010

one of four *height* values: *the floor*, *the knees* (of the robot), *the table*, and *the hips*.

Table 4 shows how hand gestures substitute verbal phrases of spoken action commands and constitutes multimodal commands. As one may notice, multimodal commands include a word or phrase that specifies an action type and a gesture event for one or more action parameters. The commands can include verbal phrases for one or more action parameter values as well, which always override parameter values in gesture events.

Table 1. Actions to move objects

	Parameters	Examples
moveforward	*step*	*mf_3steps*
movebackward	*step*	*mb_2steps*
turn	*direction, angle*	*turn_l_30deg turn_r_much*
sidestep	*direction, step*	*sstep_r_2steps sstep_l_short*
pickup	*width, height*	*pu_30cm_20cm pu_small_table*
place	*height*	*place_table*
push/pull	*height step*	*push_h_2steps pull_l_3steps*
slide	*height step direction*	*slide_h_3steps_r*

Table 2. Gesture events

	Parameters	Examples
Single hand	*direction amplitude count*	*sh_l_long_3 sh_r_short_2*
Both hand	*distance count*	*bh_me_4 bh_short_2*

Table 3. Mapping of amplitude and count

amp. \ count	1	2	3
short	one step 15 deg.	4 steps 60 deg.	6 steps 90 deg.
long	two steps 30 deg.		

Table 4. Multimodal and spoken commands

Spoken command	Multimodal command
Take 3 steps!	sh_l_long_1 + Walk!
Pick the medium-size object from the table!	bh_medium_3 + Pick (that) up!
Turn right by 15 degrees!	sh_r_short_1 + Turn!
Place it on the floor!	bh_short_1 + Place it!
Slide that left at the height of your hips by 4 steps!	sh_l_short_2 + Slide it at the height of your hips!

III. USER EVALUATION METHOD

1. Simulated humanoid that moves objects

We developed a 23-DOF simulated humanoid robot (Fig. 1) using a humanoid model on Webots [3] robot simulator. The robot can execute action commands of our multimodal language (Table 1) to move boxes and chairs in simulated environments.

The humanoid can interpret multimodal commands using a multi-agent command understanding system [2] on top of OAA [4], which include an OpenCV based hand gesture detector using a web camera we developed by ourselves and a grammar based speech recognition system (Julius4.1.1) [5].

Fig.1. Simulated humanoid that moves objects

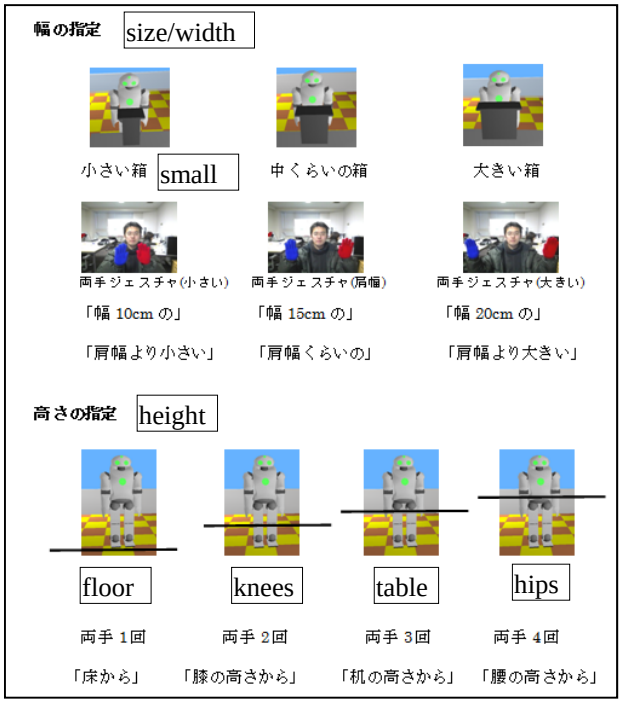

Fig.2. A page of the leaflet

2. User evaluation

Our humanoid robot was evaluated with ten students of Fukuoka Institute of Technology, who had never commanded our robot in the multimodal language.

At the beginning of each user evaluation session, we explained the user how to command the humanoid by speech alone using a leaflet illustrating spoken and multimodal commands (Fig. 2) within five minutes. Then, we demonstrated how to use gestures in the language and gave the user time to practice and learn to use them. Finally, we took ten minutes to teach the user multimodal commands combining gestures and speech.

After the demonstrations and practice, we gave the user the first task to estimate speech and gesture recognition rates and command success rates (Task1). The user read out 23 spoken commands printed in a sheet of paper, made 15 hand gestures as instructed, and gave 20 multimodal commands. The user had to use gestures given only spoken phrases specifying action parameter values in order to give the 20 multimodal commands. After this first task, the user was allowed to practice multimodal commands for ten minutes in order to learn to command the robot successfully without misrecognition and human errors. At this point, the user gave 13 spoken and 15 multimodal commands instructed in another sheet of paper in the same manner as Task1 (Task2).

The third task given to the user was to achieve three goals commanding the robot in the language:

1. Moving a medium-size box on a table down in a specified area on the floor
2. Move a box down to the floor (without information about its size and height)
3. Operate the robot following oral instructions

The user was allowed to consult the leaflet.

After Task3, the user was given the same task as Task1 to demonstrate how well commands and gestures by the user work. Finally, the user was asked how easy it was to specify action parameter values by speech and using hand gestures in 7 point scales from very difficult (1) to very easy (7).

IV. RESULTS

Tables 5-7 show speech recognition rates, gesture recognition rates, and command success rates, respectively, in each task. Gesture error rates for each action parameter in multimodal commands are shown in Table 8. The average ratings of the users about how easy to specify action parameters are shown in Table 9.

Eight of the ten users answered that they preferred multimodal commands to spoken commands. In fact, all of the commands given in Task3 were multimodal commands including both a hand gesture and a spoken message.

In Task3, all the users achieved each goal within three minutes. Six of them commanded the robot without consulting the leaflet. In total, 211 multimodal commands were given and there were two false alarms by the speech recognizer. Some single hand long gestures were misrecognized as the hand went out of the camera view.

Table 5. Speech recognition rates

	Task1	Task2	Task3	Task4
Spoken	92.2 %	88.5 %	-	96.9 %
Multimodal	97.5 %	98.7 %	96.7 %	99.0 %

Table 6. The Gesture recognition rates

	Task1	Task2	Task3	Task4
Gesture	95.2 %	-	-	98.0 %
Multimodal	93.0 %	98.0 %	93.4 %	99.0 %

Table 7. Command success rates

	Task1	Task2	Task3	Task4
Spoken	93.0 %	90.0 %	-	97.4 %
Multimodal	92.0 %	96.7 %	90.1 %	98.0 %

Table 8. Gesture error rates for action parameters

	Task1	Task2	Task3	Task4
direction	0.0 %	0.0 %	4.4 %	0.0 %
angle	2.5 %	3.3 %	5.5 %	0.0 %
step	10.0 %	0.0 %	10.7 %	1.4 %
width	8.0 %	2.5 %	3.1 %	0.0 %
height	0.0 %	1.7 %	1.5 %	1.1 %

Table 9. Average user ratings (7 point scale)

	Speech	Gesture
direction	6.3	6.5
angle	5.4	5.6
step	5.8	5.6
width	6.0	5.8
height	6.0	5.7

V. DISCUSSION

The multimodal language can be effective for the purpose of commanding humanoids to move objects by sophisticating the language and our gesture detector.

©ISAROB 2010

First, novice users were able to achieve given goals in Task3 without troubles. Secondly, over 90 % of the multimodal commands were successful (Table 7) in spite of the fact that some gestures were misrecognized when the hand went out of sight. Besides, the success rate of multimodal commands in Task4 was about 98%, which is higher than the success rate of spoken commands in the same task. The results of Task4 show that a novice user can successfully direct humanoids to move objects by speech and using gestures within a short period of time.

The speech recognition rates of multimodal commands are higher than the rates of spoken commands (Table 5) because most of spoken messages in multimodal commands include only a word or phrase to specify an action type. Novice users need a little practice commanding in the spoken language since some actions with two or three parameters are difficult even for Japanese speakers to command by speech alone. They also have to adapt to the speech recognizer, speak clearly, and use the microphone properly.

The results in Table 6 show that novice users can successfully use the two types of gestures with some practice. The recognition rates in Task4 imply that the users were better and better at using gestures. Some gestures were unsuccessful in Task3 probably because the users had to concentrate on looking at the humanoid on the computer screen and sometimes they failed to move their hands properly. The users had to move their hands a lot due to the limitations of our gesture detector. A better gesture detector using a stereo vision which can precisely detect subtle movements would make the language easier to learn.

Novice users need more experiences in order to successfully and quickly achieve goals in various situations as the success rate of multimodal commands in Task3 was lower than the rates of Task2 and Task4 (Table7) due to the lower recognition rates (Tables 5 and 6). In Task3, the users were not given printed words for action parameter values or action types; they had to find the right gestures for parameter values and the right words to inform the robot action types. In addition, some single hand gestures failed because the hand went out of the camera view.

Tables 8 and 9 indicate that it was slightly more difficult to specify *step* and *angle* values using single hand gestures than to specify values of the other parameters. There are some possible reasons for this. First, the mapping in Table 3 was not very natural and required the users some effort to learn it. Second, the long gestures were physically difficult and not very natural because they had to move their hand horizontally about 50cm; in Task3, long gestures failed for this reason. Designing a better mapping and allowing shorter gestures may help users. Another solution is allowing users to cue the robot using a simple gesture whenever they want to stop it.

VI. CONCLUSION

This paper described the results of a study on a humanoid robot that can be directed by its users in a multimodal language to moves objects. Ten novice users successfully directed the robot in multimodal commands to achieve given goals. The success rate of multimodal commands was over 90 % and the users thought that gestures were more preferable than and as easy as verbal phrases to inform the robot action parameters to move objects. The results of the study show that the language can be easier for non-experts to learn and effective for directing humanoids.

ACKNOWLEDGMENT

The authors would like to thank Toyokazu Abe and Kazuki Furihara for their kind help. This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) 19500171.

REFERENCES

[1] Oka T, Abe T, Sugita K, Yokota M (2009) RUNA: a multi-modal command language for home robot users. Journal on Artificial Life and Robotics 13-2:455-459

[2] Oka T, Abe T, Shimoji M, Nakamura T, Sugita K, Yokota M (2008) Directing humanoids in a multi-modal command language. The 17th International Symposium on Robot and Human Interactive Communication 580-585

[3] Michael O (2004) Cyberbotics Ltd – Webots™: Professional Mobile Robot Simulation. International Journal of Advanced Robotic Systems 1-1:39-42

[4] Cheyer A, Martin D (2001) The open agent architecture. Journal of Autonomous Agents and Multi-Agent Systems, 4-1/2:143-148

[5] Lee A, Kawahara A, Shikano K (2001) Julius --- an open source real-time large vocabulary recognition engine. The 7th European Conference on Speech Communication and Technology 1691-1694