Gesture detection based on 3D tracking for multimodal communication with a life-supporting robot

Tetsushi Oka[1], Ryuichi Kibayashi[2], and Hirosato Matsumoto[2]

[1]Nihon University, Chiba 275-8575, Japan
(Tel/Fax: 81-47-474-9693)
[2]Graduate School of Industrial Technology, Nihon University, Chiba 275-8575, Japan
(Tel/Fax: 81-47-474-9693)

[1]oka.tetsushi@nihon-u.ac.jp

Abstract: This paper reports some recent results from a study on multimodal communication between life-supporting robots and their users. In the study, novice users understood how to use four types of hand gestures, raising, lowering, pushing, and pulling in a short period of time. They successfully conveyed their intensions to a robot using gestures, after watching a video for four minutes that explained how to use gestures and practicing for less than four minutes. The robot guided the users by displaying messages on its front screen and detected gestures by tracking the user's head and right hand based on depth and color information. The results show that novice users can learn quickly how to convey intentions to our robot and imply that it is easy for untrained users to combine hand gestures and spoken messages, in order to make a robot to turn to them, move forward to them, back away, approach them, and so on.

Keywords: communication, gesture detection, life-supporting, multimodal, robot

1 INTRODUCTION

These days, there are robots that work in various kinds of places. We predict that in near future, there will be more and more life-supporting robots, which are supposed to help non-experts, especially in aging societies, including Japan. Millions of robot cleaners are already in homes and offices all over the world and help people who do not have special knowledge or training experiences. However, many robots are still operated by experts and it is difficult for untrained people to command them.

This study focuses on communication between a life supporting robot and its user, in particular, commanding the robots to look at the user, come close to the user, and move forward and backward to keep appropriate distance from each other, which would be necessary before the user gives a task to the robot, receive help from it, or work on a cooperative task.

We believe that multimodal communication using spoken messages and three dimensional hand gestures is effective when we want a life-supporting robot to turn to us or come close to us. The robot can also send verbal and nonverbal messages on its screen and through its speakers.

We realized a 3D hand gesture detection system for life-supporting robots that understand multimodal commands and conducted a user study in order to verify hypotheses that novice users can convey their intensions through 3D hand gestures in a short period of time and that our system can recognize their gestures without errors. There have been studies in multimodal languages, combining speech and key input, speech and body touch, or speech and 2D gestures, for untrained users of home-use and life-supporting robots [1-4], which have shown that the languages are beginner friendly for some groups of tasks, such as moving the robot and moving an object. However, it is difficult to specify 3D positions and distances in the real world in those multimodal languages.

We have proposed a multimodal language in which one can command robots using a touch screen and a speech interface [5]. However, the language requires a hand-held device and it is difficult to let robots know one's own location or distance information. In this paper, we propose a multimodal language which combines speech and 3D gesture, which allows users to intuitively convey their own location, goals, and short distances. For example, one can raise the hand and say "turn to me" in order to turn a robot to the person. Without gestures, they would give a spoken command, say, "make a 73-degree right turn," which is not very natural and difficult for untrained users: they have to determine angles, distances, and other spatial quantities by eye. Without speech, it is also difficult to give robots a variety of commands: one needs to learn many types of gesture [6].

These days, it is possible to obtain depth images [7], find and track an object, and recognize spoken commands in real time using inexpensive computers and input devices. Therefore, one can develop and test robots that understand multimodal commands combining speech and 3D gesture.

© ISAROB 2012

Our gesture recognition system detects four types of 3D single-hand gestures: raising, lowering, pushing, and pulling, which we find important for communication between a life-supporting robot and its user. These types of gesture are useful when one wants to start communicating and collaborating with a robot and to maintain a comfortable distance during communication. Pushing and pulling gestures convey a 3D direction, forward or backward, and length information. A raising gesture can inform a robot of a 3D position. Finally, raising and lowering gestures can be used to start and stop robot actions.

In the rest of this paper, we describe our gesture detection system and a user study of it. Eleven novice users learned to use the four types of 3D gesture within ten minutes and all of their properly used gestures were classified correctly by our system. These facts and question sheets filled out by the users imply that our gesture detection system is of a great value for multimodal communication between life-supporting robots and their users, none the less because the system works well on a single laptop computer equipped with a microphone and an inexpensive motion input device to capture color and depth images.

2 GESTURE DETECTION SYSTEM

2.1 User detection

Our gesture detection system can find the nearest user sitting on a chair in the view in the following steps:

1. find every person in the view
2. mark each person's top of head
3. find the closest sitting person using 3D information

Our current system is built using OpenNI [8], a free SDK for Kinect motion sensor [9], and OpenCV [10] for image processing and GUI. OpenNI includes a function to obtain a depth map from a Kinect and a function to label image pixels by person. Both are incorporated into our system.

2.2 Hand detection and tracking

The system can find the right hand of the nearest sitting person, when the person moves the hand up close to the face. It traces the boundary of the sitting person's label looking for the right elbow, i.e. the lowest point, and then the right hand (see **Fig. 1**). If this does not work, the system assumes that the hand is near the label closest to the view

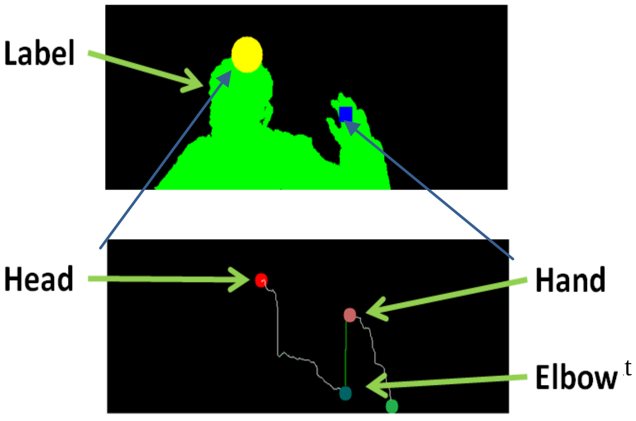

Fig. 1. Finding the right hand

2.3 3D gesture detection

Our system can detect raising, lowering, pushing, and pulling gestures based on the results of real-time 3D hand tracking described above. It looks for a 3D motion segment that fulfills several conditions of duration, speed, the location of the start/end point relative to the head position, distance between the start and end point, and direction. **Table 1** shows some of the criteria for each type of 3D gesture. For instance, to give a raising gesture, one must stop the right hand below the top of head ($y < 0$), move it fast and straight upward, and stop it above the top of head ($y > 0$).

Table 1. Gesture detection criteria

Gesture Type	raise	lower	push	pull
average speed [m/s]	0.3	0.3	0.15	0.15
top speed [m/s]	0.84	0.9	NA	NA
distance [m]	0.13	0.09	0.19	0.19
start point(x,y,z)[m]	y < 0	y > 0	z < 0.4	z < 0.4
end point (x,y,z)[m]	y > 0	y < 0	z > 0.4	z > 0.4
direction	+y	-y	-z	+z
	up	down	forth	back

2.4 Messages and images on the screen

Our system displays verbal messages, a face, and a hand on a PC screen (see **Fig. 2**). When the system finds a sitting person, a frontal face appears (bottom-left). When the right hand is successfully tracked, a hand and a "ready" sign are displayed near the face and up-left corner of the screen, respectively (top). The screen shows a verbal sign such as "up" and "back" and an image that depicts a gesture for two seconds after a gesture is detected (middle).

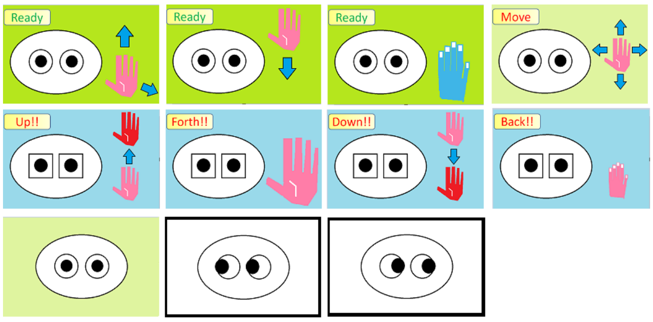

Fig. 2. Screen messages

3 USER STUDY

Fig. 3 shows Rocky, our life-supporting robot platform, which we used in the user study of our gesture detection system. It has a Kinect motion sensor on top, a 10 inch touch screen at front, and a laptop PC with an Intel Core i7 processor inside the body. The gesture detection system ran on the PC and the touch screen, connected with the PC via a USB cable, displayed messages and images to the users.

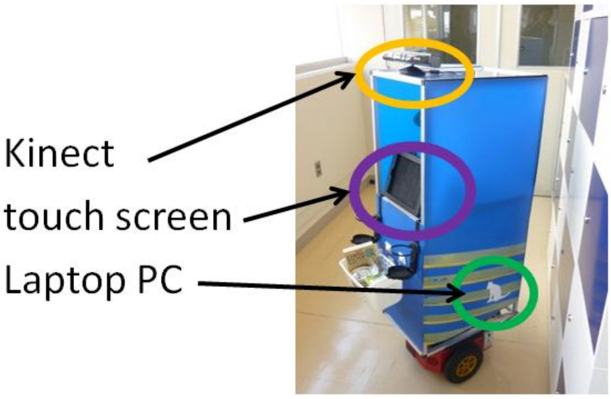

Fig. 3. Rocky, our life-supporting robot

We showed each of the eleven novice users a four-minute demonstration video, who then sat on a chair 2[m] away from the robot and moved the right hand as instructed.

The users were advised to move the right hand fast and straight when the screen displays a "ready" sign and let it down immediately after each gesture. We also suggested them to move the hand up or to the right when the system could not find it.

The first six users (Group A) practiced for two minutes after watching the demonstration video. Then, they were given pieces of advice described above and instructed to show the robot three gestures in a row for each type to test the system.

The other five users (Group B) practiced until we judged they learned the four gestures, since some of the former six users failed to practice all the types. We advised them during the practice and emphasized that they needed to wait for a "ready" sign. In the test phase, they were instructed to show gestures one by one rather than three in a row, in order to avoid unnecessary confusions.

Each of the eleven users filled out a question sheet which included eight questions (see **Table 2**).

4 RESULTS

Our system detected and correctly classified all the 3D gestures given properly by the eleven novice users. All of them moved their hand quickly and straight. However, some moved their hand when there was no "ready" sign on the screen, and also, there were three users, two in Group A and one in B, who did not put their hand down immediately after some of their pushing and pulling gestures. They told us that they had moved their hand "without thinking." In addition, some users moved their hand forward when our system happened to give a false alarm on the screen.

Table 3 shows the results of the first six questions in **Table 2**. A user (Group A) answered that the system failed to detect pulling and pushing gestures. Four in Group A thought that our system classified pulling as pushing and vice versa.

Six people found the screen messages comprehensible, and three thought they were slightly comprehensible. The other two's answers were neutral.

Table 2. Questions to the users

Id	Question
Q1	Do you understand how the four types of 3D gesture work?
Q2	Do you know exactly when to gesture?
Q3	Did our robot fail to detect your gesture during the final test?
Q4	Did our robot fail to classify your gesture?
Q5	During the final test, did the robot react to your unintended hand motion?
Q6	Have you mastered the four types of 3D gesture?
Q7	Were the pictures and signs on the screen comprehensible?
Q8	Any problems or comments?

Table 3. Answers to Q1-Q6

Id	Yes	No
Q1	11	0
Q2	10	1 (Group A)
Q3	1(Group A)	10
Q4	4 (Group A)	7
Q5	0	11
Q6	11	0

5 DISCUSSION

The results imply that our gesture detection system is of great value for multimodal communication between a life-supporting robot and its novice users. Even beginners can learn to use the four types of 3D gestures given some simple advice within less than ten minutes. Our system can detect 3D gestures properly given by beginners at a very high probability. It can be applied to understanding multimodal commands combining speech and 3D gesture. Our robot would be able to detect and classify multimodal commands at a high rate and reduce false alarms since it is unlikely that speech and gesture false alarms are given simultaneously.

Although our system ignored hand motions for two seconds after a gesture was detected, in multimodal communication it would not be necessary because only gestures that match spoken messages are in question. The users, mostly in Group A, who tried to give gestures when the screen was not displaying a ready message, were probably confused by images and messages that showed detected gestures (**Fig. 2**). Presumably, they tried to give a second gesture and our system detected a hand motion in the opposite direction as a gesture. We need to avoid such confusions when designing multimodal human-robot communication.

Users in Group B, who practiced given our advice, had better impression about our system, so it is important that novice users learn to wait for a ready message and give a gesture properly.

Although our current system is already highly reliable with respect to detection classification of four types of 3D gesture, we can improve it by reducing false positives and negatives. We presume that we can add more types without performance degradation. Finally, our 3D gestures are more intuitive than static hand gestures [11] for multimodal commands.

6 CONCLUSION AND FUTURE WORK

We built a 3D gesture detection system for multimodal communication with a life-supporting robot and conducted a user study. In the study, eleven novice users learned using 3D gestures properly within a short period of time, and our system correctly recognized the users' intentions when they moved their hand as instructed. The results show that our 3D gesture detection system can be readily applied to multimodal communication with life-supporting robots. We are currently developing a life-supporting robot one can command by combining spoken messages and 3D hand gestures.

ACKNOWLEDGMENT

This work was supported by KAKENHI Grant-in-Aid for Scientific Research (C) (22500179).

REFERENCES

[1] 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

[2] 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

[3] Oka T, Sugita K, Yokota M (2010) Commanding a humanoid to move objects in a multimodal language. Journal on Artificial Life and Robotics 15-1:17-20

[4] Oka T, Abe T, Sugita K, Yokota M (2011) User study of a life supporting humanoid directed in a multimodal language. Journal on Artificial Life and Robotics 16-2:224-228

[5] Oka T, Matsumoto H, Kibayashi R (2011) A multimodal language to communicate with life-supporting robots through a touch screen and a speech interface. Journal on Artificial Life and Robotics 16-3:292-296

[6] Wachs JB, et.al. (2011) Vision-based hand-gesture applications. Communications of the ACM 54-2:60-71

[7] Goth G (2011) I, domestic robot. Communications of the ACM 54-5:16-17

[8] OpenNI: http://www.openni.org/

[9] Kinect: http://www.xbox.com/en-US/Kinect

[10]OpenCV:
http://sourceforge.net/projects/opencvlibrary/files/

[11] Wang K, Wang L, Li R, Zhao L (2010) Real-time hand gesture recognition for service robot. Proc. of 2010 International Conf. on Intelligent Computation Technology and Automation

© ISAROB 2012