Using Gestures to Interact with a Service Robot using Kinect 2

Harold Andres Vasquez[1], Hector Simon Vargas[1], and L. Enrique Sucar[2]

[1] Popular Autonomous University of Puebla, Puebla, Pue., Mexico
{haroldandres.vasquez,hectorsimon.vargas}@upaep.edu.mx
[2] National Institute of Astrophysics, Optics and Electronics, Puebla, Pue., Mexico
esucar@inaoep.mx

Abstract. In this paper is presented a proposal for multimodal interaction between the robot and the person, using voice and gestures, in order to make friendlier human-robot relationship. This functionality will be integrated into a service robot called Donaxi, from Robotics Lab of UPAEP. Due to this research is recent, only is showed some preliminary results what has been achieved so far. Some data from a dataset of gestures being built for service robots is shown.

Keywords: HRI, gestures,service robot, Kinect 2.

1 Introduction

Many countries, in special the European countries, are concerned about the care of the old people. This affect the social and economic aspects of any country. Given the aging of the population, in the future there will be a lack of caregivers for old people; service robots provide an alternative to assist senior persons. Due this, there are many support programs to help to develop this kind of robots, called services robots [1].

Because these robots are going to interact with a person, it is necessary a comfortable communication way so that person feels that the robot is able to understand very well the given instructions. The most natural way to achieve this is with voice, due it is the most common communication way between us. However, use only the voice to communicate is not enough, due some problems and also because not all people can use the voice.

Therefore, is necessary use other communication way. It is clear that gestures is one very useful alternative, still is not very common between persons. Gestures have been used in many projects with service robots and they have proven to be very efficient in noisy ambient, where the voice is not enough.

According to the above, we can exploit the benefits given by both of them: voice and gestures. Each one can solve the problem of another. When it's used two or more communication way with a machine, this is called multimodal interaction. Therefore, we propose a multimodal interaction with a service robot using voice and gestures.

On the first part of this document is presented the problem and the methodology proposed to solved it. Then the main expected contributions are described, followed by some results obtained at this moment and closed with the conclusions.

2 Problem

Gestures have proved be very useful to interact with a robot [2],[3],[4]. Many kind of devices have been used to achieved this, but the most commons are the video cameras and the Kinect. Among these, the Kinect is the most used in services robots, due the programmers don't have to deal with both segmentation and following persons. These functionalities are provided by the Kinect SDK, using a combination of hardware and software. With this, the problem reduced to identify when the gesture is being performing and what gesture performed the user. The "when" problem is called segmentation and the "what" problem is called recognition. Each problem can be solved by separated, but the aim is that both operate simultaneously, thus a more natural interaction is achieved [5].

Actually, this problem seems to have been solved with the Kinect version 2 (Figure 1), a device created by Microsoft Company, initially to be used with the game machine called XBox. This because the Microsoft team (the owners of Kinect) developed a IDE where is possible makes new systems with own gestures. This IDE is called Visual Gesture Builder (VGB). On this, after dispose of a good number of videos examples, we can obtain a database with the new gestures trained. With this Database of Gestures (DBG), it's possible develop a system where it's used these gestures.

Fig. 1. Kinect version 2 used for multimodal interaction with the robot using voice and gestures

In some tests conducted (showed in Section 5), this technique proved be very good solving the two problems described before. Nevertheless, all the previous work was very tedious and long. We have to capture many videos from different persons to achieve a good training of the DBG. After, we have to tag each video with the gestures where are performed.

The other hand, voice is the most natural way to interact with machines. The most common problem with this is the machine can't understand to the person. This because each person have different voice and pronunciation. Another issue that affect this is noisy ambients. This can be solved with: software only, hardware only or both. In those cases, is necessary complement the voice with another communication ways. Obviously, gestures is the best option for doing

this. Therefore, we can obtain a multimodal interaction with the service robot using voice and gestures.

The different situations in which the voice and gestures together can be used are:

1. Gesture is voice reinforcement, i.e. when the robot is not capable to understand the voice command, the person performs a gesture with the same significance of the voice command. For example, if the user commands with voice to robot to pay attention to him, and because they are in a Shopping Center, the robot can't hear its owner; then the user can wave his hand to call the attention of the robot.
2. Gesture is voice complement, i.e. at home environment, the user wants the robot gives him a new medicine unknowing by it, and therefore, the robot don't know where is. Then, with voice, the user asks for the medicine to robot and simultaneous, with hand, he shows the place where is the medicine.

In summary, this work objective is to contribute to solve these challenges of the multimodal interaction with the service robot.

3 Main contribution

There are many works about Multimodal Human Robot Interaction (MHRI) [2],[6]. Most of them use different devices by each input: voice and gesture, or only use the first combination of them described in Section 2.

We propose use only the Kinect 2 to treat both signals. This because, while fewer devices have connected the robot, it consumes less power and less weight will be loaded. Also, on software side, less communication with different devices is required and this reduces processing costs. This last is very difficult to achieve, due the problems with voice described before, so only can be solved on software side.

Another relevant aspect of our proposal is use the multimodal interaction in different ways, like was described in Section 2.

As result of this work, be going to create data sets for both signals, so these can be used by any research in robotics or associated fields.

4 Methodology

It is important mention that this work is a continuation of a previous work regarding simultaneous segmentation and recognition of gestures, which apparently is already solved with the VGB. Nonetheless, the experience obtained with that research has allowed a more rapid development of this proposal.

At this moment, this research is in the beginning phase. A general idea of a work plan to be followed, based on the model shown in Figure 2, is describe below.

As result from the model showed, the work plan to follow is:

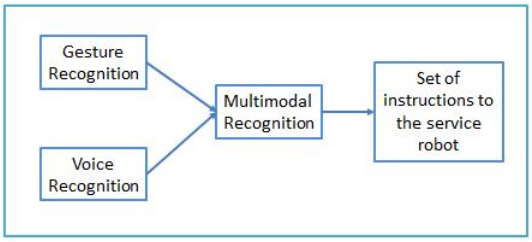


Fig. 2. General model to work

1. Know how the Kinect version 2 works with gestures and voice. Although we worked with the Kinect version 1 before, this new version has many different technologies and software that make it more capable to recognize gestures and voice separately. Therefore, it's necessary understand in depth how this works to exploit its benefits and uses them to achieve this research. One idea for this, is using the examples coming with the Microsoft Kinect SDK, designed for gestures and voice recognition.
2. Test the recognition quality from the Kinect 2 for both gestures and voice. It has to design formal experiments to measure the quality of both recognitions way and probe if these are enough in order to the service robot can interact with any person in not controlled ambients.
3. Identify the causes of any shortcomings found in existing recognition techniques from kinect 2. This will allow us determine our research hypothesis and then build proposal to solve them.
4. Once each recognition works well separately, be must to design a model to combine both results, as it showed in Figure 2. This is the multimodal recognition model, that is expected to be able to perform the service robot.
5. Finally, we must apply the experiments for evaluations, designed to prove that our robot can perform a natural interaction with any person in any ambient.

As will be shown in the next section, there are already some progress in the plan designed, allowing to demonstrate the feasibility of this work.

5 Results

This research is conducted in the UPAEP robotics laboratory , where there is a service robot called Donaxi. This robot is completely built from scratch by students from different college careers like electronics, mechatronics, bionics and computing. Donaxi will have:

1. Omnidirectional navigation system, with four wheels, each one with a DC motor with encoder. This allow to Donaxi moves in almost any direction.
2. Laser system to build navigation map. It consists of two laser, front and rear.

3. Vertical movement System, based on a rail and a motor, which moves a platform up or down.
4. One arm with five freedom degrees and a parallel gripper on the end.
5. One Kinect version 1 that moves with the vertical movement system, and it used for object recognition.
6. One Kinect version 2 for people recognition, whether the whole body, face or voice. This is fixed in the top of the robot. This it will use to multimodal recognition with gestures and voice.
7. Two laptops. One with Ubuntu and ROS, used for some of the functionality of the robot, and the second with Windows, used to recognize people, faces, gestures and voice. The laptops communicating with each other through TCP/IP messages.

Some of these devices are already available on the robot, but others are in the adaptation process. In Figure 3 is showed the preliminary version of Donaxi.

Fig. 3. The Service Robot Donaxi from the UPAEP Robotics Laboratory

Because in april this year was the Mexican Robotics Tournament (TMR2015)[3] and Donaxi team participated on it, it became necessary to have some work of both gestures and voice recognition separately. For this reason we have some progress on these two features, that was proven in a almost realistic house ambient in this tournament. In Figure 4 are showed the services robots competed at TMR2015.

For voice recognition, a Creative 3D Sense camera was used, which also was used to face recognition. The software used for this was the Intel RealSense SDK, created for these devices. Although not conducted rigorous testing of this feature, a very good preliminary results was obtained, even allowed Donaxi to win the competition in this category. In Figure 5 it showed this device.

[3] http://www.tmr2015.mx/

Fig. 4. The three services robots present in the Mexican Robotics Tournament 2015

Fig. 5. Creative Sense 3D used for voice and face recognition

For gesture recognition, the Microsoft Kinect version 2 was used. To achieve this, it was necessary complete the steps showed in Figure 6.

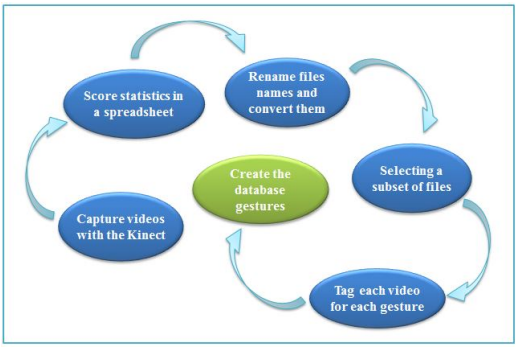


Fig. 6. Diagram where show the steps followed to obtain the DataBase of Gestures

First, we have to capture many videos from many different people in different places. This because for more different examples provided to the training algorithm, more overfitting is avoided and will provide the capacity to detect any person in any environment. Second, to keep track of the statistics of the video taken by the Kinect, it has been written relevant data into a spreadsheet. Third, to know the content of each video file without opening it, it should be rename them using a specific nomenclature, to then make a conversion using the tool KSConvert from the Kinect SDK. Fourth, because several kinds of catches with various kind of people was used, it was decided to use only part of the full set, to verify if this subset was enough for both training and testing of the recognizer. Fifth, to train each gesture, it should shown in each video where are the positive

examples and negative examples, which is called tagging and it was used the Microsoft Visual Gesture Builder. Finally, using the VGB also, can be created the DataBase of gestures, for build the application for the service robot.

In table 1 it shows the total numbers of gestures captured until now for this research.

Table 1. Total number of gestures in dataset

Gesture	Number
Stop	194
Come	177
Left	395
Right	407
Attention	417
Indication	363
Turn	207

Further details on this work will be shown in a publication about this dataset and thus put this material available to the scientific community.

Once this database of gestures is available, it can build an application for Donaxi, that can detect gestures being made by the user, though the Kinect 2. For this purpose, a sample in C# available in the Microsoft Kinect SDK was used, called "DiscreteGestureBasics". For the TMR competition, only were used three gestures from the seven contemplated. This because has not been tested the accuracy of the recognizer with the seven gestures together in the database. In Figure 7 is showed the three gestures used: attention, to make Donaxi know where is your master; right, to make Donaxi revolves on his right and stop, so Donaxi stop when approaching. These three gestures were used in TMR2015.

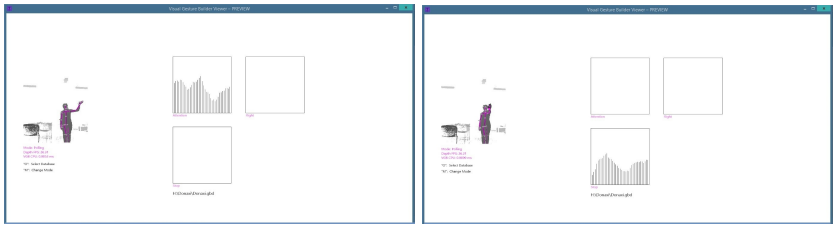
(a) Attention (b) Stop (c) Right

Fig. 7. The three gestures used at TMR2015 for Donaxi

In attention gesture, the user moves either hand from side to side, near to the head. In stop gesture, the user put the open hand, with arm extended, in front to him. In right gesture, the user extend his arm to right side. It can be appreciated in Figure 7 the arrow showing the movement at each case.

Nevertheless, the application developed with three gestures showed good results, making Donaxi speak only when the user completed any of the gestures. Only the gesture "stop" had some not detection problems (false negatives). For this reason, it is necessary to prepare a more rigorous test plan while is being adding more gestures and to evaluate the results.

In figure 8 can be appreciated the gesture recognition running in Microsoft VGB LivePreview tool, where the user perform each gesture at once.



(a) Attention (b) Stop

(c) Right

Fig. 8. ScreenShots from the Microsoft VGB LivePreview where the user perform each gesture at once

At this moment, is preparing Donaxi for her next challenge, the Rockin 2015 (http://rockinrobotchallenge.eu). This is an international competition of services robots, which this year will be held in Lisbon, Portugal in November. For this competition it will expected to have ready the speech and gesture recognizers fully functioning separately.

6 Conclusions

It is clear the need for service robots to interact with people in the most natural way possible. However, achieving this type of interaction is not so simple. Many challenges are looming to achieve this kind of robot-human relationship.

For now, most research on multimodal human robot interaction are using only voice and gestures, but obviously this set can grow to reach the expected goal.

This research aims to achieve this multimodal HRI using only the Kinect version 2 for detecting gestures and voice and allow to Donaxi can understand better its owner.

References

1. Aracil, R., Balaguer, C., and Armada, M.: Robots de servicio. Revista Iberoamericana de Automatica e Informatica Industrial 5, 6–13 (2008)
2. Goodrich, M. A. and Schultz, A. C.: Human-Robot Interaction: A Survey. Foundations and trends in human-computer interaction 1(3), 203–275 (2007)
3. Droeschel, D., Stckler, J., Holz, D and Behnke, S.: Towards joint attention for a domestic service robot person awareness and gesture recognition using time-of-flight cameras, In: Proc. of the IEEE Int. Conf. on Robotics and Automation (ICRA), pp. 1205–1210 (2011)
4. Yan, R., Tee, K. P., Chua, Y., Li, H., Tang, H.: Gesture Recognition Based on Localist Attractor Networks with Application to Robot Control. Computational Intelligence Magazine 7(1), 64–74 (2012)
5. Vasquez Chavarria, H., Escalante, H. J., and Sucar Succar, L. E.: Simultaneous segmentation and recognition of hand gestures for human-robot interaction. In: 16th International Conference on Advanced Robotics, pp. 1–6 (2013)
6. Pavlakos, G., Theodorakis, S., Pitsikalis, V., Katsamanis, A., Maragos, P.: Kinect-based Multimodal Gesture Recognition using a two-pass fusion scheme. In: Proc. International Conference on Image Processing, pp. 1495–1499 (2014)