2019 SinfonIA Pepper Team Description Paper

Carlos A. Quintero Eyberth Rojas Fabián Pérez
Saith Rodríguez Santo Tomás Members Uniandes Members
Unimagdalena Members CdC-IA Members

February 17, 2019

Abstract. In this paper we introduce the SinfonIA Pepper Team for RoboCup Social Standard Platform League (SSPL). Our team is a joint effort between several Colombian institutions whose main objective is to lead the development of robotics and artificial intelligence in our country, so that new applications and technologies are introduced in people's daily life. Our current research interest is to improve the interaction between robots and humans by a) allowing the robots to learn new instructions in a natural way using multimodal inputs and b) developing an emotion recognition system that the robots can use to drive its interaction with the humans. In this paper, we also show our efforts implementing a shortened version of the Cocktail Party challenge from previous years, using a Pepper robot that demonstrates speech interaction, person recognition, object detection and task planning capabilities.

1 Introduction

The SinfonIA Pepper Team aims to participate in RoboCup 2019's Social Standard Platform League (SSPL). The team is a joint effort between 4 different institutions in Colombia, i.e., Universidad de los Andes, Universidad Santo Tomás, Universidad del Magdalena and the Bancolombia's Center of Competences on Artificial Intelligence, to promote research in robotics and artificial intelligence in our country. Due to the nature of the collaboration between, industry and academia, we are equally interested in both the contribution of scientific and academic value solutions as well as the implementation of highly relevant solutions for commercial and business applications.

Our main interests lie in the development of robots and systems capable of providing services to people in a variety of contexts and situations, domestically, at work, in hospitals, hotels, universities and schools among many others, and for a variety of services as assistance, entertainment, information provision and guidance. Our current research focus is to provide better decision-making and self-learning abilities to these systems by building data-driven models.

Finally, an important objective of our collaboration is to sensitize the colombian society about the way social robots and AI systems are used in people's daily lives. We anticipate that our developments and their socialization will have a profound impact on the perception the colombian society has about intelligent robots.

Our participation in RoboCup aims to join social robots academic international community to share and learn the latest innovations in the field. The following section briefly describes our background and experience in previous RoboCup competitions. After that, we show our main research interests by briefly mentioning recent works and publications in the field of social robotics and human-robot interaction. Then, we describe the study case presented for our participation in RoboCup 2019 and finally we talk about our future work plans.

2 Background and previous experiences in RoboCup

While the SinfonIA Pepper team and the joint work between the four institutions, started recently, the team consists of people and universities that have an academic trajectory in research groups and that have also been part of the RoboCup community for several years now. Our most important involvement is our participation in the Small Size League (SSL) from 2011 to 2017 with the STOx's team, where we reached the top 8 three times $(2013, 2014, 2017)$ and the top 4 once (2015), becoming the first and only latin-american team to enter the league's hall of fame up to date. Also, SinfonIA Pepper team leader, had previously served as TC member of the SSL twice, 2016 and 2018.

Our greatest achievements and contributions to the SSL include the design and implementation of our 3^{rd} generation of robots, a fleet of 12 omnidirectional robots that attained a high degree of robustness and reliability in terms of their electronic and mechanical design. An overall description of these robots can be found in our 2014 ETDP [1]. We have also devoted large efforts in the development of a data-driven based model to automatically identify and reconstruct chip kicks during games [2]. Strategies for creating dynamic attacking plays instead of fixed state-machine-like plays and a defensive strategy based on an optimal marker assignment algorithm were proposed and implemented during the tournament games with good results [2,3]. Finally, we presented a general multi-agent robot coordination methodology based on an optimal agent/task assignment [4] that evolved our game from static pre-factored plays to a completely dynamical game where all decisions and multi-agent behaviors were based on the solution of an optimization problem.

One contribution that's worth mentioning is our proposal of the Fast Path Planning algorithm for the problem of robotic soccer in the domain of the Small Size League [5]. The algorithm is a simple procedure that generates paths and is very fast compared to more traditional approaches such as the Rapidly-exploring Random Tree (RRT) algorithm. The Fast Path Planning algorithm showed improved behavior in terms of route's smoothness and the generation of shorter paths. Many teams in the league have implemented it as their solution to perform the path planning task [6,7,8,9].

3 Research interests and achievements

Our team's most important research interest is developing data-driven models for social robots that allow them to improve their behavior as they interact with humans. We are interested in providing social robots with the ability of automatically learn new ways to interact with humans without being explicitly programmed to do so. We also aim to come up with new methods to improve the robot's understanding of humans. We have tackled some of these problems in the last few years as described below.

3.1 Automatic emotion recognition for human-robot interaction

In the context of human-robot interaction and specifically for social robots, it has been proved that providing robots with the capability of recognizing human emotions during their interaction can be helpful. With this in mind, we have proposed an automatic classification model that recognizes four different human emotions using multimodal inputs, i.e., audio and video [10]. The video classifier uses a convolutional neural network (CNN) trained to recognize human emotions in images that are later merged to automatically classify the detected emotion in a video stream. The audio classifier uses data from human voice during the interaction to achieve a similar result, based on the extraction of features from the raw audio data using a set of coefficients with frequency relations that provide important information of the human voice. This combined classifier was tested as a case study of a robot acting as a sales agent. The recognized emotions were used by the robot to drive its speech during its interaction with the user and persuade him to purchase a specific product. In our work, we have achieved a high classification accuracy for the emotion recognition problem. However, there are still plenty of opportunities to improve on the way the robot uses such information to interact with users.

3.2 Learning multimodal instructions in a natural way

The problem of identifying an instruction given by humans in the context of HRI is a key element for further improvement of social robots' behavior for different applications. Although great advances have been made in this area, humans usually communicate in a complex manner, in order to achieve a more natural communication with humans and improve the robot's understanding, further research is needed. In this context, we have proposed a learning methodology that allows robots to use multimodal inputs (speech and gesture) to recognize a specific human instruction [11]. Our methodology uses a scalable learning architecture that is based on one-class classifiers (one for each instruction). Each instruction classifier is implemented using Support Vector Machines with special customized kernels that use mat exploiting for the specific input properties, i.e., a specific kernel for audio data and another for gesture. In [12] we proposed a methodology that uses such multimodal learning architecture and allow the user to teach new instructions to the robot in a natural way, i.e., without explicitly reprogramming the learning model, but by using natural interaction procedures with the robot.

4 Case study: Pepper's Cocktail Party

In order to illustrate the abilities developed by our team, we decided to make a simplified version of the Cocktail's Party challenge. Fortunately, we acquired our Pepper robot from SoftBank Robotics, therefore we present our work and progress with this robot. The test is designed to show Pepper's autonomous skills, including object and person recognition, audio recognition and navigation. Moreover, there were several tasks that we needed to accomplish for this challenge, including gathering orders from clients, identifying missing drinks at the bar, recognizing the clients that were being attended, maintaining a conversation with them and moving safely and accurately through the party's scenario.

To accomplish this challenge we designed a general architecture, which includes different abilities that were developed and the adopted tools in our work, for the RoboCup participation. In figure 1 we present the approach we made in order to solve all the previously mentioned tasks. The core of the solution lays in the Scheduler implemented in ROS (Robotic Operating System). In order to accomplish our goals we integrated ROS with Pepper, so we could be able to use Slam gmapping, and navigation techniques. This integration lays in the NAOqi framework that provides lookup services alllowing us to find methods and network access to be called an executed remotely from the robot. As shown in figure 1, the package responsable for the connection between ROS and the NAOQi Driver is the Robot Toolkit. This packages provide methods and functions that allows the Navigation, Interaction and Making Decisions tools access to all the sensors, cameras and actuators of the Pepper robot.

On the other hand, all the tool packages, found in the application layer, allow us to develop specific functions to build our case study, such as reaching a reference location, identifying clients, remember orders, etc. Finally, the Scheduler gathers all the applications in a state machine and decides when and which function to execute, reaching the construction of a successful Cocktail Party. It is accurate to mention that all the tools and applications were developed with the objective to use them for future applications in the area of social robotics.

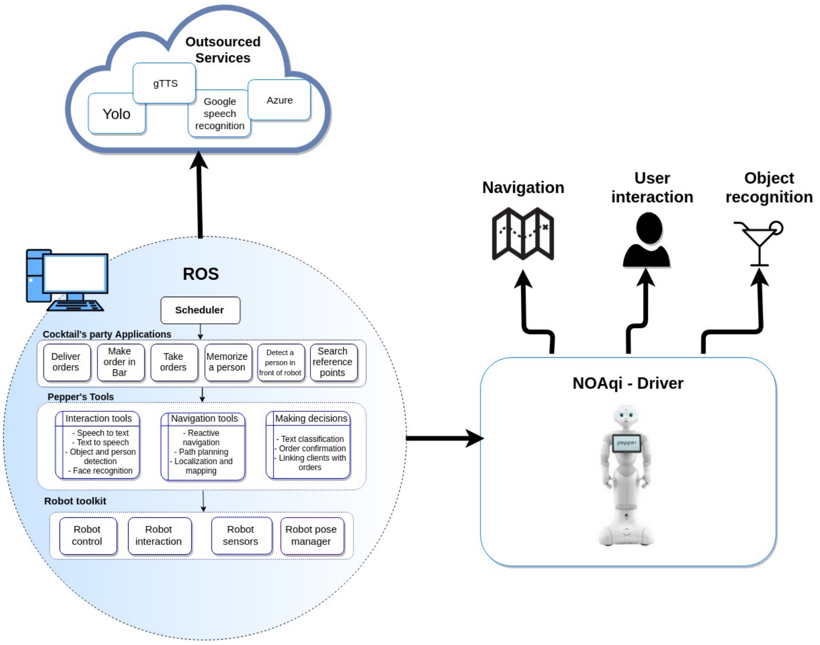

Fig. 1: General solution architecture for Pepper's Cocktail Party

4.1 Speech interaction

One of the main skills of Pepper is its ability to interact with people by means of speech. This ability allows our robot to reach out in a kind way. In the following sections, we explain with more detail the audio recording and speech recognition we established.

Speaking skill In order to make our robot speak with the user in a natural way, we used the Google API gTTS (Google Text-to-Speech) which creates a mp3 file from a predetermined text introduced as a parameter of the function. Despite the fact that we implemented Google Services, we designed an algorithm capable of analyzing all the possible answers and obtain the desired words. To integrate all these possible options and to diversify the answers for each client we implemented a method able to choose a different response appropriate for the interaction stage.

Speech recognition The other part of the human-robot communication is the human dialog speech recognition. In order to get a transcription of human speech we used the Google Cloud API Speech-To-Text, which allows the robot to transcript words of the person who is talking into a text file. Once the text file was created, we developed an algorithm to split and analyze all the words contained in the original speech. Therefore, in the algorithm, according to the meaning and position of the word, the robot is able to vary it's answer. The speech recognition works together with the audio recording skill, in order to communicate in a successful way with a client. The algorithm is based on our customized Cocktail Party. For example, Pepper gives an answer to the client when he or she makes an order.

4.2 Visual skills

Two important visual skills are required for social robots, specially for the Cocktail's Party application, namely person recognition and object detection. For this purpose we used the two robot's on-board 2D cameras to capture the images that are later processed to perform these tasks. However, since the robot and the people are in constant motion during the interaction, some of the captured images may not have the proper quality that is needed to successfully learn and detect people and/or objects. To solve this problem, we have implemented a methodology that allows the robot to select the less blurry images from a larger set. To this end, we have implemented the methodology in [14,15], where the focus level can be found by computing the standard deviation of the images Laplacian, which consists of its 2^{nd} spatial derivative. This focus measure allows us to select images with largest focus level from a set of pictures taken by the robot and use them to learn new faces or objects.

Person recognition and characterization: For the face recognition task we used the Microsoft Azure Face API, which allows Pepper to identify some characteristics of a person from a single photo. Despite the fact that we are using Microsoft services, we integrated this skill in our Pepper, with the intention of having a better performance in our personalized Cocktail Party. This function is implemented as a ROS service we called *FaceDetector*.

Object detection and recognition Other applications developed for RoboCup 2019 are object detection and recognition. To accomplish this, we used an interface called LightNet, which is based on the object detection models of YOLO. More specifically, LightNet uses the weights and the neural network configuration of YOLO. This library was chosen due to the features that it provides during the process of object detection, which include numpy arrays with bounding boxes and the probabilities of prediction done by the net. The developed system is based on a picture of the objects taken by the robot that is used to create labelled bounding boxes on the image. To obtain an accurate result, the image is used as the training data of a pre-trained network, previously loaded from YOLO. Through the training process the weights of the network are modified in order to get a higher detection score, provided in the form of probabilities for each of the objects found on the image. Prior information and domain knowledge are used for training the network, and to obtain better accuracy in the robot's predictions. For this function it exists one Service called *detect_objects*.

4.3 Navigation

Hoping to accomplish an effective navigation, we developed an application that could fulfill the requirements of path planning, obstacle avoidance, mapping and localization of the robot. In contrast with the other applications used for this challenge, we chose to design this application in a way in which we didn't have the need to depend on any source other than the robot's sensors. Due to logistic problems we received our Pepper two weeks before the qualifying deadline, with the downside that it was not enough time to implement a robust enough path planning algorithm. We tried implementing *A**, *RRT* and *Dijkstra* algorithms, but unfortunately none of these worked out as expected.

On the other hand, with the aid of the laser scan and odometry sensors, Pepper is able to locate itself in the map frame and determine it's position. We used *slam-gmapping* and *ira-laser-tools* in order to use all the laser sensors.

5 Conclusions and future work

Thanks to the work developed for the 2019 RoboCup qualification process, we accomplished the goals and objectives proposed among the group. The research and dedication within this work has allowed us to learn about different applications for different projects that we have been working on lately, such as applications like object detection, speech recognition, maintaining a conversation between humans and robots and navigation. The main difference between our packages is the fact that our navigation package was a technological implementation and integration without using an online service or resources.

Keeping in mind that the main objective of all the development and research presented in the current document is to improve the technological presence in people's daily lives, speccially in Latin American countries like Colombia, we decided to film the video in Spanish. In this way Pepper will be able to interact in a better way and much larger scale with the Spanish speaking community. Something we see as a challenge as the main language's for Pepper are English and Japanese.

Finally, we were able to successfully use ROS in this project. However, the main focus of our future work is to continue to develop the advantages that ROS provides for a better work with Pepper. For example, to solve the issues found when implementing the Navigation algorithm. We are working on making a cross compilation in our Pepper to install ROS together with the Naoqi Driver. Once this is done, it would be easier for us to work with tools like the Navigation Stack, which provides several solutions for Pepper's navigation.

References

1. S. Rodríguez, E. Rojas, K. Pérez, J. López, C. Quintero, and J.M. Calderón. STOx's 2014 Extended Team Description Paper. Joao Pessoa, Brazil 2014. Available for download in http://robocupssl.cpe.ku.ac.th/_media/robocup2014:etdp:stoxs_2014_etdp.pdf

2. S. Rodríguez, E. Rojas, K. Pérez, C. Quintero, H. Báez, O. Peña, J.M. Calderón. STOx's 2016 Extended Team Description Paper. Leipzig, Germany 2016. Available for download in http://wiki.robocup.org/images/0/0d/Small_Size_League_-_RoboCup_2016_-_TDP_STOx%27s.pdf
3. S. Rodríguez, E. Rojas, K. Pérez, C. Quintero, O. Peña, A. Reyes and J.M. Calderón. STOx's 2017 Team Description Paper. Nagoya, Japan 2017. Available for download in http://wiki.robocup.org/images/d/d7/Robocupssl2017-final16.pdf
4. S. Rodríguez, E. Rojas, K. Pérez, C. Quintero, O. Peña and A. Reyes. 2018 STOx's Extended Team Description Paper. Montreal, Canada 2018. Available for download in http://wiki.robocup.org/images/6/67/Robocupssl2018-STOxs.pdf
5. Rodríguez S., Rojas E., Pérez K., López J., Quintero C., Calderón J. (2015) Fast Path Planning Algorithm for the RoboCup Small Size League. In: Bianchi R., Akin H., Ramamoorthy S., Sugiura K. (eds) RoboCup 2014: Robot World Cup XVIII. RoboCup 2014. Lecture Notes in Computer Science, vol 8992. Springer, Cham
6. Bouchard S., Lachapelle F., Lebel P. and Verret B. (2018) ULtron 2018 Team Description Paper, RoboCup 2018.
7. Vedder K., Schneeweiss E., Rabiee S., Nashed S., Lane S., Holtz J., Biswas J. and Balaban D. (2017). UMass MinuteBots 2017 Team Description Paper, RoboCup 2017.
8. Shamsi M., Waugh J., Williams F., Ross A., Llofriu M. and Weitzenfeld A. (2015) RoboBulls 2015: RoboCup Small Size League, RoboCup 2015
9. Gao T., Wu Y., Yang T., Huang Z. and Xiong R. (2017) ZJUNlict Extended Team Description Paper for RoboCup 2017, RoboCup 2017.
10. Pérez A.K., Quintero C.A., Rodríguez S., Rojas E., Peña O., De La Rosa F. (2018) Identification of Multimodal Signals for Emotion Recognition in the Context of Human-Robot Interaction. In: Brito-Loeza C., Espinosa-Romero A. (eds) Intelligent Computing Systems. ISICS 2018. Communications in Computer and Information Science, vol 820. Springer, Cham
11. Rodriguez S., Pérez K., Quintero C., López J., Rojas E., Calderón J. (2016) Identification of Multimodal Human-Robot Interaction Using Combined Kernels. In: Snášel V., Abraham A., Krömer P., Pant M., Muda A. (eds) Innovations in Bio-Inspired Computing and Applications. Advances in Intelligent Systems and Computing, vol 424. Springer, Cham
12. Rodríguez S., Quintero C.A., Pérez A.K., Rojas E., Peña O., De La Rosa F. (2018) Methodology for Learning Multimodal Instructions in the Context of Human-Robot Interaction Using Machine Learning. In: Brito-Loeza C., Espinosa-Romero A. (eds) Intelligent Computing Systems. ISICS 2018. Communications in Computer and Information Science, vol 820. Springer, Cham
13. M. Ríos, C. Quintero, C. Gamarra and S. Rodríguez. Design and Implementation of an Automatic Object Recognition System using Deep Learning and an Array of One-Class SVMs. Accepted to appear in 17th International Conference on Machine Learning and Applications (ICMLA 2018), Orlando Florida, 2018.
14. Pertuz, Said and Puig, Domenec and Garcia, Miguel Angel, 2013 Analysis of focus measure operators for shape-from-focus. In: Pattern Recognition, Elsevier, 2013
15. Pech-Pacheco, José Luis and Cristóbal, Gabriel and Chamorro-Martinez, Jesús and Fernández-Valdivia, Joaquín, Diatom autofocusing in brightfield microscopy: a comparative study, in: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000
16. V. Perera, T. Pereira, J. Connell, M. Veloso. Setting Up Pepper For Autonomous Navigation And Personalized Interaction With Users. 2017. Taken from: https://arxiv.org/abs/1704.04797

Pepper Software and External Devices

We use a standard *SoftBank Robotics* Pepper robot unit.

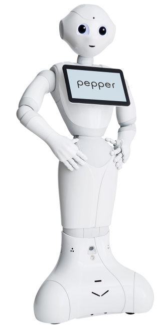

Fig. 2: Pepper Robot

Robot's Software Description

For Pepper robot we are using the following software:

- Platform: Ubuntu 9.1
- Face recognition: Based on Microsoft Azure Face API (See previous sections).
- Object recognition: Based on Lightnet and Yolo (See previous sections).
- Speech Interaction: Based on gTTS and Speech Recognition. (See previous sections).

External Devices

Pepper relies on the following external hardware:

- Atom E3845 Quad core 1.91 GHz
- Intel HD graphics up to 792 MHz
- 4 microphones, 2 RGB HD cameras, 5 tactile sensors, touch screen on the breast

Cloud Services

Pepper connects the following cloud services:

- Object detection and recognition: Yolo and Lightnet.
- Face Recognition: Microsoft Azure Face API.
- Speech recognition: Google Text-to-Speech (gTTs).

Team Members

Uniandes
Adelaida Zuluaga
César Daniel Garrido
Juan David García
Juan José García
Nicolás Rocha

Santo Tomás
Armando Mateus
Bryan Betancur
Camilo Camacho
Carolina Higuera
Diego Ibáñez
Lina Plazas
Nicolás Garzón
Sindy Amaya

Unimagdalena
Andrés Fornaris
Christian Carpio
Daniela Castillo
Dilan Solar
Ervinson Plata
Fabián Beltrán
Jim Cotes
Juan Camilo Salgado
Mario Acuña

CdC-IA
Diana Arismendy
Eduardo González
Estefany Montoya
John Alexander
Manuel Ríos
Óscar Salinas
Oswaldo Peña