A study of robot semantic localization based on multimodal HRI

Álvaro Villena[1], Ismael García-Varea[1], Jesus Martínez-Gómez[12], Luis Rodríguez-Ruiz[1], and Cristina Romero-González[1]

[1] Computer Systems Department. University of Castilla-La Mancha.
Campus Univ. s/n. 02071. Albacete, Spain
[2] Computer Science and Artificial Intelligence Dept., University of Alicante.
P.O. Box 99. 03080. Alicante, Spain
{Alvaro.Villena,Jesus.Martinez,Ismael.Garcia
Luis.RRuiz,Cristina.RGonzalez}@uclm.es

Abstract. Semantic localization describes the surrounding of a robot by using semantic labels. These labels are used to identify neighboring objects, but also the category of the place where the robot is located. Multimodal human-robot interaction refers to the communication between humans and robots by means of several information sources. This paper presents a bibliographical revision about these two important robotic-related topics: semantic localization and multimodal human-robot interaction. In addition, this paper proposes a method for performing semantic localization using information gathered from the interaction with humans. This method models a lifelong system where robot learns from scratch, using sensors and information obtained from humans.

Keywords: Semantic localization, multimodal HRI, human-robot interaction, sensors, lifelong learning

1 Introduction

Robot localization is a key problem on mobile and autonomous robotics. Traditionally, localization has been solved following a topological approach, that is, using range finder sensors to find geometrical features in a predefined map of the environment. When the map is also built at the same time the robot is trying to localize itself, the problem is known as SLAM. However, humans do not use exact map representations to figure out where they are located. This localization can be done by just observing some important elements of the environment, in such a way that if we return to a previously visited place, we can recognize it by just re-observing those elements. In addition to this, the information extracted from a fluent process of human-robot interaction (HRI) can also be helpful. That is, the knowledge about objects and their relation with the place where they are located can be used to improve current and future localizations. For example, the presence of cutlery commonly indicates that a robot is located in a kitchen, even if it has never been there before. Moreover, both localization and human-robot interaction rely on an internal representation that should be maintained by means of a lifelong adaptive cognitive architecture. This paper presents a brief current state of the art in multimodal HRI and semantic localization. As far as we know, there exists very few works dealing with these two problems in a common framework. Then, we propose some ideas about how to perform robot localization by fusing the semantic information to be gathered from the environment, but also from the interaction process with humans. What is more, our proposal is intended to start from total uncertainty about the environment, and without any previous kind of human-robot interaction knowledge.

2 Multimodal Human-Robot Interaction

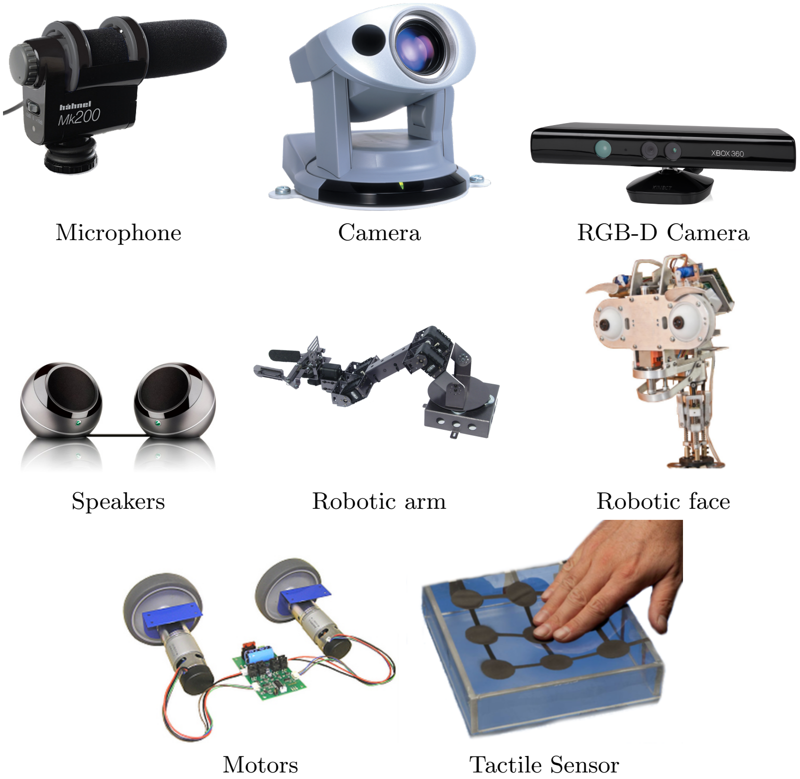
Human-robot interaction (HRI) is the research field that studies all the possible ways that can be used between robots and humans to interact. While first communication approaches were based on computer specific commands, recent social robots can interact with humans using natural language, facial expressions, body language or even expressing emotions. The main objective of HRI is to develop a natural communication system with robots, allowing them to accomplish interactive tasks in human environments in an easier way.

It may be considered that verbal communication is enough to solve interaction problems between humans and robots, but it has been proved that the use of more input data sources improves the performance, as well as it allows for a more natural interaction [4, 9]. In this way, robots can use visual information to detect the human body language, or audio sensors and actuators to perceive and simulate voice intonation [1].

2.1 Sensors and Actuators

Interaction between humans is composed of many elements, such as verbal language and intonation, facial gestures or body language. In order to achieve a similar multimodal interaction between robots and humans, it is indispensable to provide the robot with the appropriate set of sensors and actuators.

Sensors are used to receive input signals like sound and images. These signals should be processed to obtain valid information about human interlocutors and, in consequence, improve the interaction process. Due to the importance of verbal communication in human interaction, microphones are one of the key sensors used by robots to receive information. Other sensors with special relevance are cameras. They can be used to detect emotions, gestures or characteristics from humans. Some cameras can also capture depth data, and they complement or even substitute sonar and laser devices. Tactile sensors can also be used to obtain data, such as tactile areas or complete skins [11, 5]. On the other hand, actuators are the devices used to execute the desired output commands. For instance, actuators will be in charge of movement and sound output. The robot will use its actuators to provide feedback to human interlocutors. Speakers are commonly used to incorporate verbal communication output to the robot, and can also be used to provide simple sound feedback too. Other elements like robotic arms or faces can be used to express emotions and gestural information, but also with manipulation purposes. Some common sensors and actuators used in robotics are shown in Fig. 1.



Microphone

Camera

RGB-D Camera

Speakers

Robotic arm

Robotic face

Motors

Tactile Sensor

Fig. 1. Exemplar sensor and actuators.

2.2 Proposals

Recently, improved and novel interaction systems have been proposed to enhance human-robot communication. A very interesting interaction system is presented in [29], where a haptic creature is used to introduce touching as an interaction method. This system defines a set of touch gestures that humans can use to interact with the robotic creature, and each gesture causes a different reaction. Physical HRI is also discussed in [14], which presents a machine learning algorithm focused on improving physical interaction between robot and human. Both articles highlight the importance of physical contact to improve the interaction process. Interaction can be also enhanced by using complex robotic human-like heads to express emotions and gestural information, as presented in [7]. This robotic head includes eyes, eyebrows, mouth and neck, and is capable to detect, classify, imitate and generate facial expressions in real time.

Special attention should be paid to correctly incorporate different interaction types to achieve multimodal communication. The work presented in [25] mixes speech recognition and synthesis with head and gestures detection. More recent articles have improved this communication scheme by introducing a more complete set of body gestures detection and expression [28], or a mechanism to give positive and negative feedback using gestures and speech based on rewards [3]. Multimodal interaction systems have been also developed for commercial robots. For example, Nao robots (Fig. 2) can be programmed to improve their communication capabilities, as can be seen in [15, 8, 6]. This robot can be adapted to be used in therapy by mixing speech, gestural and tactile information as shown in [24].

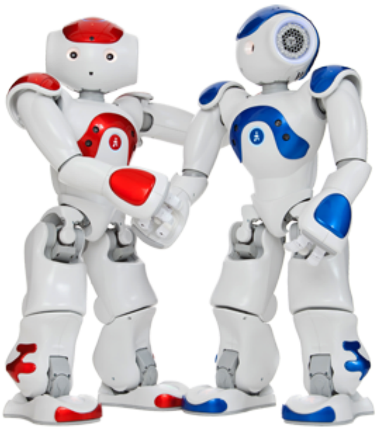


Fig. 2. Aldebaran Nao humanoid robot.

Another key aspect is how to model HRI, this is a research topic that has also received some attention lately. In [23], a Dynamic Neural Field is used to model cognitive processes and link them with robot sensors and actuators in order to enable interaction. In [17], a multimodal dialogue system is constructed using a POMDP-based system.

3 Robotic Semantic Localization

Semantic localization [27] consists of answering the question "where am I?" by means of tags or labels describing the place where the robot is located. This kind of localization is similar to human localization, and it differs from traditional navigation-oriented localization techniques like distance-based or topological-based ones. Semantic maps can complement geometrical ones by adding information about objects and areas on them. This semantic information consists of natural language tags, which facilitates the communication between robots and humans.

Semantic localization involves two different tasks, scene categorization and object identification, based on robots perceptions. On the contrary to topological localization, where range data is the main source of information (laser or sonar readings), images are the best alternative for semantic localization. Scene categorization and object identification can be performed independently. However, they are expected to be jointly managed by taking advantage of the inherent relationship between areas and objects. For instance, if a fridge has been identified, the fact that the robot is placed in a kitchen is very likely. Fig. 3 shows the semantic annotation for an image acquired in a warehouse.

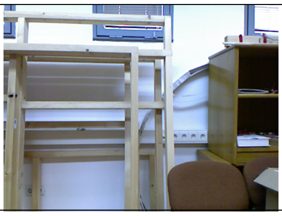
Input Image	Ground Truth
	**Scene**: Warehouse
	Objects
	Table
	Chair
	Book
	Socket
	. . .

Fig. 3. Semantic Localization Annotations for an exemplar image.

Semantic localization is commonly managed as a set of classification problems where input data correspond to robot perceptions and classes to scenes/objects [27]. Initial classification models can be generated from some of the existing datasets like KTH-IDOL [18], COLD [21] or ViDRILO [19]. Input data should be processed to extract an appropriate representation, which is usually done through computer vision techniques. GIST [20] and Histograms of Gradients (HoG [10]) are some of the well-known global image descriptors used in semantic localization. Regarding the classification model, Support Vector Machines (SVMs) is the most common approach due to its proper handling of numeric input data and fast classification time. However, other approaches as Bayesian classifiers can also be applied [22].

Semantic localization can also be performed along with spatial localization, as the proposal presented in [13]. The link between spatial and semantic representations is generated by means of anchoring [12, 16].

4 Relation between Multimodal HRI and Semantic Localization

Semantic information of the environment can be acquired by the robot in several ways. Usually, this task is performed using vision and range-finder sensors.

However, that detected semantic information can also be obtained and improved with HRI. This interaction complements the semantic localization, providing additional information about uncertain features and helping to reach a correct conclusion [26]. This kind of active learning model is specially useful in changing environments, where the robot would be able to adapt to new objects and places.

The general pipeline of a task based on robot vision will include the following steps: a) capture an image, b) segment the image and/or identify the keypoints with more relevant information, and c) generate either a local representation of the information encoded around the keypoints or a global representation. These image descriptors can be used for other tasks like matching or registration. For object detection purposes, the most common approach would segment the input image and generate a global descriptor of each segment to use with a classification method. This, however, requires the use of a initial training database for classifying. Here we propose to take advantage of the HRI to gather information about relevant objects and locations, given an initially unknown knowledge about the environment. Thus, the robot can ask the human about objects to obtain information and update its internal representation model. After obtaining enough information of an object/location, the robot will be able to detect them without further human intervention. This multimodal approach would enable the robot to learn about new objects and places, as well as to discover the inherent relationships between them.

Then, semantic localization is expected to be performed by labeling places based on object detection. Once the robot knows that a place called 'kitchen' contains objects tagged as 'cup' and 'spoon', it will know that it is placed in a kitchen when these objects are detected. Similarly, if the robot knows that it is placed in a bathroom, it should look for objects like toilet, but not fridge. Regarding this, it is important to highlight that semantic localization works in multiple ways. On the one hand, if a human ask the robot to perform an action involving an object in a known place, the robot internal representation model will be updated to link this object with this location. On the other hand, if a human talks about a known item and the robot is in an unknown place, the robot can infer its location given the relationship between this object and a place. In Fig. 4 we show an example of how the combination of different information sources in HRI can improve semantic localization.

In general, this approach is useful to improve both HRI and semantic localization. For example, the robot can ask a human about an unknown detected object in order to classify it with a semantic tag (the name of the object) and, in consequence, improve the object classification model and the semantic knowledge about the robot location. And this new information can then be used to enhance interaction with humans, as the robot will have a higher understanding of its surrounding. Furthermore, the robot can ask about a known detected object to improve communication and interaction capabilities based on the vocabulary used by the human to describe this known object. These capabilities can be used to construct a lifelong learning model. Fig. 5 describes some situations in which semantic localization benefits from HRI.

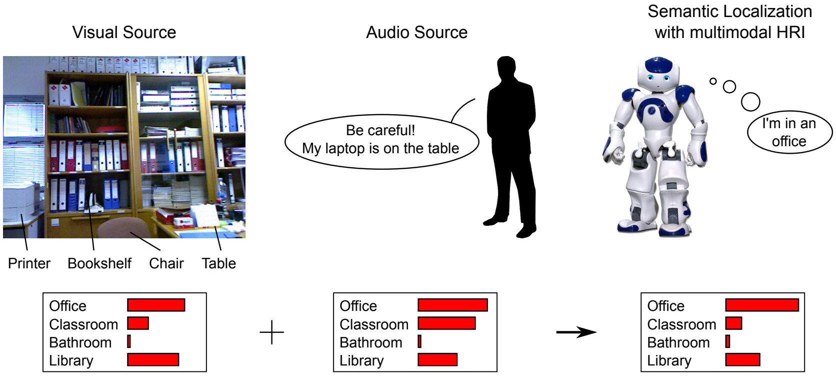


Fig. 4. Semantic Localization based on multimodal HRI.

Recent research on this topic include systems where semantic labels are obtained through conversation with humans [26]. This semantic labels are mixed with topological information obtained with range-finder sensors. However, the vast majority of the systems proposed assume an initial state with prior information, so results depend on the initial knowledge base. Another interesting idea can be found in [2], where a graph-based vision system is used to guide an spoken conversation with a human, mixing object detection and multimodal HRI.

5 Discussion

The purpose of this paper is to propose a lifelong semantic localization system for mobile robots integrating information obtained by interacting with humans. This system will be based in an initially empty knowledge base that has to be completed using the information obtained through the different information sources to create a representation of the environment mainly based on semantic tags.

On start up, the robot will have no knowledge about where it is placed, so it should start listening and observing its environment. This independent exploration of its surroundings, will help the robot to create an unsupervised first basic model. Later, to incorporate an actual multimodal HRI, the robot will perform a human detection process. Once a human has been detected, the robot must determine if the person is willing to answer its questions. Then, it can ask the appropriate questions to fill the gaps in its knowledge model. Additionally, the robot will also ask about known elements sometimes to improve its knowledge base. For example, a conversation with a human about a known item can be used to improve the speech abilities. This way, the robot can learn how the human talks about this specific item, and use it to define this or other objects in the future.

In conclusion, the main goal of this proposal is to build a system able to learn new objects, spaces and the relations between them. Then, in the appro-

HRI	Semantic Localization
Robot detects an unknown object and asks a human about it	– A semantic tag is added to the object. – If the object has the same name than a previously known object, the robot updates the information about it. – The object is included into the map representation. – If the robot is located in a known area, the object probability to appear in this kind of area increases.
Robot detects a known object and asks a human about it	– If the human confirms that the robot prediction is correct, robot's confidence to detect this type of object improves. – If the human informs that the robot prediction is not correct, robot's confidence to detect this type of object decreases, and the spotted object is tagged using the new information provided by the human.
Robot asks the name of the room/place.	– The robot tags the place and links the spotted known objects with this place.
A human talks about an object in the room/place.	– The robot links the object with the place. – If the robot hasn't seen the object, it will try to spot it for some time. If the object is not found, the robot asks the human where the object is.

Fig. 5. Situations in which semantic localization benefits from HRI.

priate situations, the robot can use the input provided by a human interlocutor to improve both, its semantic representation of the world and its interaction capabilities.

Acknowledgements This work has been partially funded by FEDER funds and the Spanish Government (MICINN) through projects TIN2013-46638-C3-3-P and DPI2013-40534-R and by Consejería de Educación, Cultura y Deportes of the JCCM regional government through project PPII-2014-015-P. Cristina Romero-González is funded by the MECD grant FPU12/04387, and Jesus Martínez-Gómez is also funded by the JCCM grant POST2014/8171.

References

1. Aarestrup, M., Jensen, L.C., Fischer, K.: The sound makes the greeting: Interpersonal functions of intonation in human-robot interaction. In: 2015 AAAI Spring Symposium Series (2015)

2. Aguilar, W., Pineda, L.A.: Integrating graph-based vision perception to spoken conversation in human-robot interaction. In: Bio-Inspired Systems: Computational and Ambient Intelligence, pp. 789–796. Springer (2009)
3. Austermann, A., Yamada, S.: good robot,bad robotanalyzing users feedback in a human-robot teaching task. In: Robot and Human Interactive Communication, 2008. RO-MAN 2008. The 17th IEEE International Symposium on. pp. 41–46. IEEE (2008)
4. Burger, B., Ferrané, I., Lerasle, F., Infantes, G.: Two-handed gesture recognition and fusion with speech to command a robot. Autonomous Robots 32(2), 129–147 (2012)
5. Cannata, G., Maggiali, M., Metta, G., Sandini, G.: An embedded artificial skin for humanoid robots. In: Multisensor Fusion and Integration for Intelligent Systems, 2008. MFI 2008. IEEE International Conference on. pp. 434–438. IEEE (2008)
6. Cheng, L., Sun, Q., Su, H., Cong, Y., Zhao, S.: Design and implementation of human-robot interactive demonstration system based on kinect. In: Control and Decision Conference (CCDC), 2012 24th Chinese. pp. 971–975. IEEE (2012)
7. Cid, F., Moreno, J., Bustos, P., Núñez, P.: Muecas: a multi-sensor robotic head for affective human robot interaction and imitation. Sensors 14(5), 7711–7737 (2014)
8. Csapo, A., Gilmartin, E., Grizou, J., Han, J., Meena, R., Anastasiou, D., Jokinen, K., Wilcock, G.: Multimodal conversational interaction with a humanoid robot. In: Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference on. pp. 667–672. IEEE (2012)
9. Cuayáhuitl, H., Kruijff-Korbayová, I.: Towards learning human-robot dialogue policies combining speech and visual beliefs. In: Proceedings of the Paralinguistic Information and its Integration in Spoken Dialogue Systems Workshop. pp. 133–140. Springer (2011)
10. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. vol. 1, pp. 886–893 vol. 1 (2005)
11. Duchaine, V., Lauzier, N., Baril, M., Lacasse, M.A., Gosselin, C.: A flexible robot skin for safe physical human robot interaction. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on. pp. 3676–3681. IEEE (2009)
12. Fritsch, J., Kleinehagenbrock, M., Lang, S., Plötz, T., Fink, G.A., Sagerer, G.: Multi-modal anchoring for human–robot interaction. Robotics and Autonomous Systems 43(2), 133–147 (2003)
13. Galindo, C., Saffiotti, A., Coradeschi, S., Buschka, P., Fernández-Madrigal, J., Gonzalez, J., et al.: Multi-hierarchical semantic maps for mobile robotics. In: Intelligent Robots and Systems, 2005.(IROS 2005). 2005 IEEE/RSJ International Conference on. pp. 2278–2283. IEEE (2005)
14. Ikemoto, S., Amor, H.B., Minato, T., Jung, B., Ishiguro, H.: Physical human-robot interaction: Mutual learning and adaptation. Robotics & Automation Magazine, IEEE 19(4), 24–35 (2012)
15. Jokinen, K., Wilcock, G.: Multimodal open-domain conversations with the nao robot. In: Natural Interaction with Robots, Knowbots and Smartphones, pp. 213–224. Springer (2014)
16. Lemaignan, S., Ros, R., Sisbot, E.A., Alami, R., Beetz, M.: Grounding the interaction: Anchoring situated discourse in everyday human-robot interaction. International Journal of Social Robotics 4(2), 181–199 (2012)
17. Lucignano, L., Cutugno, F., Rossi, S., Finzi, A.: A dialogue system for multimodal human-robot interaction. In: Proceedings of the 15th ACM on International conference on multimodal interaction. pp. 197–204. ACM (2013)

18. Luo, J., Pronobis, A., Caputo, B., Jensfelt, P.: The KTH-IDOL2 Database. Tech. Rep. CVAP304, KTH Royal Institute of Technology, CVAP/CAS, Stockholm, Sweden (Oct 2006), http://www.pronobis.pro/publications/luo2006idol2
19. Martinez-Gomez, J., Cazorla, M., Garcia-Varea, I., Morell, V.: ViDRILO: The Visual and Depth Robot Indoor Localization with Objects information dataset. International Journal of Robotics Research (2015)
20. Oliva, A., Torralba, A.: Building the gist of a scene: The role of global image features in recognition. Progress in brain research 155, 23–36 (2006)
21. Pronobis, A., Caputo, B.: Cold: The cosy localization database. The International Journal of Robotics Research 28(5), 588–594 (2009)
22. Rubio, F., Flores, M.J., Gómez, J.M., Nicholson, A.: Dynamic bayesian networks for semantic localization in robotics. In: XV Workshop of physical agents: book of proceedings, WAF 2014, June 12th and 13th, 2014 León, Spain. pp. 144–155 (2014)
23. Sandamirskaya, Y., Lipinski, J., Iossifidis, I., Schöner, G.: Natural human-robot interaction through spatial language: a dynamic neural field approach. In: RO-MAN, 2010 IEEE. pp. 600–607. IEEE (2010)
24. Shamsuddin, S., Yussof, H., Ismail, L., Hanapiah, F.A., Mohamed, S., Piah, H.A., Zahari, N.I.: Initial response of autistic children in human-robot interaction therapy with humanoid robot nao. In: Signal Processing and its Applications (CSPA), 2012 IEEE 8th International Colloquium on. pp. 188–193. IEEE (2012)
25. Stiefelhagen, R., Fügen, C., Gieselmann, P., Holzapfel, H., Nickel, K., Waibel, A.: Natural human-robot interaction using speech, head pose and gestures. In: Intelligent Robots and Systems, 2004.(IROS 2004). Proceedings. 2004 IEEE/RSJ International Conference on. vol. 3, pp. 2422–2427. IEEE (2004)
26. Woo, J., Kubota, N.: Recognition of indoor environment by robot partner using conversation. JACIII 17(5), 753–760 (2013)
27. Wu, J., Christensen, H.I., Rehg, J.M.: Visual place categorization: Problem, dataset, and algorithm. In: Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on. pp. 4763–4770. IEEE (2009)
28. Xiao, Y., Zhang, Z., Beck, A., Yuan, J., Thalmann, D.: Human-robot interaction by understanding upper body gestures. Presence 23(2), 133–154 (2014)
29. Yohanan, S., MacLean, K.E.: The role of affective touch in human-robot interaction: Human intent and expectations in touching the haptic creature. International Journal of Social Robotics 4(2), 163–180 (2012)