User Interaction for Guided Learning Supporting Object Recognition in Service Robots

Jan Dornig, Yunjing Zhao, and Xiaohua Sun(✉)

College of Design and Innovation, Tongji University, Shanghai, China
jandornig@gmail.com, ginzhao@foxmail.com,
xsun@tongji.edu.cn

Abstract. Under current technical conditions, robots often have difficulties or make errors in object recognition due to the complexity of the scene or the particularity of the object to be recognized. It is necessary for users to provide input or guided training for the robot during the recognition process. However, most of the input methods commonly used by researchers, is limited to using mouse and keyboard to mark outer and inner edges of the object on the screen. We introduce in this paper a survey of possible actions and input methods for designing human robot interactions supporting guided learning in object recognition. Specifically, we analyzed key factors and procedures in analyzing typical object recognition and proposed human robot interaction methods appropriate for each type of obstacle.

Keywords: Human robot interaction · Interaction design · Guided learning
Object recognition

1 Introduction

Object recognition plays a key role in the functioning of service robots, enabling the robots to perceive and interact with its surroundings. Research in object recognition of robots has seen rapid progress in recent years. Methods such as RGB/RGBD semantic segmentation [1] and SLAM [2] have been developed and widely tested in laboratory and real-life conditions. However, these methods often require manual labeling of large pre-defined datasets, which differs from the case of service robot's actual environment of use. The layout and content of such environments are mostly unknown, with many objects not covered by the dataset. Moreover, the complexity of the environments also poses a problem for effective object recognition in real-life scenarios.

One method currently used to overcome this problem is to adopt guided learning techniques- preparing a robot to receive additional, ongoing input from humans at a later stage during use. This method has been proven to be effective by extended research under laboratory conditions, but there are still problems hindering its real-world application and we have to differentiate between how end-users will interact with a robot and a situation that is fully controlled by researchers in a lab. The actual users of the robot lack the expertise and experience in utilizing complex interfaces to

A. Marcus and W. Wang (Eds.): DUXU 2018, LNCS 10919, pp. 427–436, 2018.
https://doi.org/10.1007/978-3-319-91803-7_32

accurately provide the information needed. This calls for designing of more natural, user-friendly interfaces and means of interaction. However, the way that users interact with the robot can be unpredictable and often multimodal, complicating the development of robots that are able to adaptively learn object recognition from user interaction.

In this paper, we aim to look at key parts of the interaction between human users and service robots and derive a framework by analyzing the situation and matching the appropriate reactions and interaction possibilities. The outcome will serve as a tool for designers and engineers to help develop service robots that are able to take advantage of user input to perform guided learning in its environment of use, while doing so in a natural, user-friendly way.

2 Recognition Obstacles

In order to deal with the recognition problems and determine corresponding interactions to support guided learning, we first start to analyze the kinds of recognition obstacles that the robot may encounter. A thorough analysis and prediction of most possible situations from which difficulties may arise will contribute to the effectiveness of the counter-actions that may follow.

The possible recognition obstacles can mainly be divided into five categories as below [3–7]:

O1. Object Similarity. It refers to the situation when there is an object which is similar to the target object and the robot is confused by these two objects. This confusion can stem from objects that are very similar, like an apple smartphone and an apple iPod, to point out an extreme case [3]. To situations where there are multiples of the same object but only one of them is referred to [4].

O2. Difficult to Distinguish from Surroundings. It includes different cases, the first one is that the color of the target object is similar to its surroundings, which makes the target object blend into the surroundings. The robot cannot find or cannot identify the particular shape. The second one is that the edges of the target object are complex and or blurry, such as hair, and the robot is unable to mark the edges successfully. Similarly, transparent objects like glass can appear in a way that is extremely difficult to process. Reflections and shapes appear in the object and distortions happen through the shape.

O3. Partially Occluded. If the target object is occluded partially by surroundings, especially when key elements are occluded, it can be difficult or impossible for the robot to identify the object [5].

O4. Change in Objects. When the feature of the target object has changed significantly over time, the robot is not likely to be able to identify it. For example, a flower, once bright and blooming, will become withered in several days. Then the color and shape of the flower will be totally different. It is hard for the robot to recognize it as the same flower or object as recorded before. Additionally, current image processing algorithms often show shortcomings with objects that are rotated [6], so if the algorithm

and training data doesn´t account for that, an object that is found later on laying on the floor instead of standing up, might be also difficult to recognize.

O5. Missing Knowledge/Understanding. The user might refer to objects in a particular way that the robot has not yet learned. Relationships between the user and the object and general context information of an object like "*my* cup" or "the *3 o'clock* pills" have to be taught to the robot. This personalization can also include regional colloquial names for objects [7].

Further obstacles exist in relation to the spoken word – misunderstanding of pronunciations can lead to many mistakes and is still an issue in natural language processing. Ambiguity in language, words that are the same but refer to different objects and can only be differentiated through context, are an issue. This paper considers this to be out of scope for the issue discussed here since it´s a stage prior - understanding commands, no matter what it refers to.

3 Instructional Actions

After summarizing the possible recognition obstacles that the robot will encounter, we are attempting to list the instructional actions for guiding the robot to learn. These actions are selected considering both physical and digital means, which can deal with the recognition problems effectively. The specific instructional actions are as follows [8–11]:

A1. Point at the Target Area. When the robot cannot distinguish the target object and another similar but different object, the user can point at the target object to let the robot know which the correct one is. This can be done over visual information that the robot provides, a sort of first person view, or for example use existing data like a map of the room, to point to the location of the object [8].

A2. Mark the Edges of the Target Area. A common procedure to help with image and object recognition is to help the robot understand the outline of an object – differentiating object and background. The user can help the robot to mark the edges of the target for the robot to collect the data of the target object and identify it more precisely [9]. The action can include different operations for editing the selected area, such as moving points and zooming.

A3. Move the Target Object. The user can move the target object by hand and show it to the robot clearly from all angles, enabling the robot to collect the data of the object more effectively. Besides, if the target object is hidden by surroundings, the user can also pick it up and show it to the robot, and then the robot can be aware of the existence of the object.

A4. Define the Target Object. When the user aims to let the robot know what the target object is, can help to define the object for easier identification [10, 11]. The user can input information that further define the object and its current state like the color, position, size, material, shape, etc. Additionally, the user might have to tell the robot about certain relationships and context attributes, "*my* cup", that cannot be known to

the robot before or are too difficult to reliably perform. For example, if the robot cannot recognize that the present withered flower and the past blooming flower are the same object, the user is supposed to define the withered flower as “the same”, letting the robot know they are the same object and remember it.

4 Operation Modality and Medium

Every user’s instructional action requires an operation medium in order to be conveyed to the robot. Even when performing the same instructional actions, a change in medium can influence the whole interaction process. Here we list possible operation mediums and modalities to use in constructing corresponding interactions between robot and human. They can be divided into several categories:

M1. Robot’s own Screen. It refers to the screen equipped on the robot. The robot can display the photo of target object and surroundings on this screen, and then the user can perform the actions like pointing at the target object or mark the edges of it.

M2. Robot’s own Projector. The robot is able to project its recognition area on the physical target object. The user can edit this area through other input methods and make like gesture or screen input.

M3. Screen-Based Device. It refers to devices such as smart phone, iPad and laptops, which are independent from the robot and use a screen. They can be very helpful for more detailed input, loosing ambiguity of interpretation of the medium that might occur in voice and gesture-based input. Screens are well suited for displaying visual information like the first-person view of the robot or maps and let the user act on that.

M4. Augmented Reality Device. It refers to the AR device like HoloLens or AR applications in Smartphones. Using AR with the robot would require a shared physical and digital environment. In a sense, it is comparable to two people using a HoloLens which are able to see the same holograms. Large amounts of information could be communicated in a user-friendly way like that. The human can also use the various input methods of the devices to input further information in form of commands or even 3d data like defining volumes.

M5. Voice. The user can directly instruct the robot via voice commands. It is convenient for the user to define the target object via voice but might be difficult for the robot to interpret.

M6. In-air Gesture. The user instructs the robot only by gesture without any other assistant equipment. For example, the user can just point at the target object by finger and let the robot remember it. This method can only be used if it´s ensured that the robot is reliably able to recognize such gestures as pointing to an area a couple meters away can be quite ambiguous.

M7. Physical Action and Assistant Tools. The user can use some assistant tools such as labels and laser pointers. He or she can attach a label on the target object, making it easy for the robot to recognize it. As for the laser pointer, the user can use it to point at

the target object, providing a more direct way than gesturing with the convenience of not having to physically walk around the room.

5 Obstacle, Action, Interface Analysis

After considering instructional actions and operation mediums for dealing with recognition obstacles and guiding the robot to learn, in order to design the interactions for different recognition obstacles, we are analyzing which actions and modalities are suitable for each recognition obstacle. First, it is necessary to list the possible combinations of actions and modalities as below (Table 1).

Table 1. Possible combinations of actions and media.

	M1	M2	M3	M4	M5	M6	M7
A1. Point at the target area	✓	✓	✓	✓	✓	✓	✓
A2. Mark the edges of the target area	✓	✓	✓	✓		✓	✓
A3. Move the target object							✓
A4. Define the target object	✓		✓	✓	✓		✓

For *A1. Point at the target area*, it is possible for the user to perform this action based on any media. Because the user can click at the target area on any digital device, and includes *M1. Robot's own screen*, *M3. Screen-based device* and *M4. Augmented Reality device*. The user can do it by gesturing, *M6. In-air gesture.* For *M2. Robot's own projector*, the user can also point at the physical target object first, which can be detected by robot's camera. Then the robot will project the recognition area on this object, so this action is feasible via the media of the projector. Although for *M5, Voice* the interpretation has to be stretched a bit and overlaps with *A4, Defining the object*, as the user is able to select the target object via descriptions like "the left one", verbally pointing. Finally, the user can point at the object with the aid of some physical assistant tools like the laser pointer, which demonstrates *M7. Physical action and assistant tools* is also possible.

Similarly, for *A2. Mark the edges of the target area*, the gesture and every digital device seem to be possible. The user can also use the laser pointer or other tools to mark edges. However, *M5. Voice* is hardly possible, although you could direct the robot to follow an object outline by directing it, we don´t see this implemented are user-friendly in any robot at the current time.

For *A3. Move the target object*, it can only be realized by physical action. *M7.*

As for *A4. Define the target object*, most means can provide additional information apart from *M6 In-air gesture*, gesturing and *M2 Robot's own projector*, since it's mainly an output medium.

Then we can analyze the suitable combinations of actions and modalities for each obstacle as below (Table 2).

In order to cope with *O1. Confused by another similar object*, first we should select the correct target object, which requires *A1. Point at the target area*. Secondly, we need

Table 2. Suitable actions and media for recognition obstacle O1.

	M1	M2	M3	M4	M5	M6	M7
A1. Point at the target area	✓	✓	✓	✓	✓	✓	✓
A2. Mark the edges of the target area	✓	✓	✓	✓		✓	✓
A3. Move the target object							✓
A4. Define the target object	✓		✓	✓	✓		

to show the difference between two objects to the robot and mark it, so this includes *A2. Mark the edges of the target area* (mark the key feature of the target object) and *A3. Move the target object* (show from all angles). Finally, we are supposed to let the robot know what the marked feature or object is and how to recognize it, which requires *A4. Define the target object*. So all the combinations are suitable for it.

As for *O2. With difficulty in marking edges*, what we need to do first is to mark the edges of the target object, and then let the robot know what the target object is. If the target object is hidden, we should pick it up and show it to the robot. Therefore, it is not necessary to point at anything, and *A1. Point at the target area* is not suitable for it (Table 3).

Table 3. Suitable actions and media for recognition obstacle O2.

	M1	M2	M3	M4	M5	M6	M7
A2. Mark the edges of the target area	✓	✓	✓	✓		✓	✓
A3. Move the target object							✓
A4. Define the target object	✓		✓	✓	✓		

To address the problems caused by *O3. Partially occluded*, we should instruct the robot to recognize it through the exposed part of the target object. We can mark it, a feature of it, or move the object (Table 4).

Table 4. Suitable actions and media for recognition obstacle O3.

	M1	M2	M3	M4	M5	M6	M7
A1. Point at the target area	✓	✓	✓	✓	✓	✓	✓
A2. Mark the edges of the target area	✓	✓	✓	✓		✓	✓
A3. Move the target object							✓
A4. Define the target object	✓		✓	✓	✓		

Similarly, with *O4. Object has changed*, we need to find the unchanged feature that can demonstrate that they are the same object, and then mark it for the robot. Finally, we should define the object. It requires *A2. Mark the edges of the target area* and *A4. Define the target object* (Table 5).

Table 5. Suitable actions and media for recognition obstacle O4.

	M1	M2	M3	M4	M5	M6	M7
A2. Mark the edges of the target area	✓	✓	✓	✓		✓	✓
A4. Define the target object	✓		✓	✓	✓		

For *O5. Missing Knowledge/Understanding*, when the robot misses some knowledge about the target object, the user is supposed to show the object to the robot and provide more information about it, which includes *A3. Move the target object* and *A4. Define the target object* (Table 6).

Table 6. Suitable actions and media for recognition obstacle O5.

	M1	M2	M3	M4	M5	M6	M7
A3. Move the target object							✓
A4. Define the target object	✓		✓	✓	✓		

It is not necessary that the user has to use all suitable actions to deal with a single recognition obstacle. On the contrary, during guided learning, we aim to instruct the robot with the least interaction effort, in order to optimize the user experience and improve the robot's learning efficiency.

6 Interaction Process

With the overview of possible obstacles, counter-actions and ways to interact, we can attempt outlining a process for the human-robot interaction with the goal to engage the user productively. In further research, this process must be tested, refined and overall validated.

There are in general two situation where a guided learning project will be triggered. In the case of a robot entering initially a new environment, it might be a standard procedure to take part in an extensive setup process, where the human figuratively introduces the new space to the robot. An example would be a newly bought or rented service robot being welcomed in the residence of a customer. This particular situation might be accompanied by a professional "robot-handler" but in any case, we assume that either the professional, the customer or both together would engage in the introductory setup process. In this initial stage, the robot would start scanning and identifying objects in the household while the present people would input the necessary data for the robot to start handling its assigned task. This can be specific information about the rooms that the robot will occupy as well as important objects and chores that need to be handled and taken care of. It is likely that during this process the first problems of object recognition occur and in the same vein since guided learning already happens because object relationships that could have not been known before are now introduced to the robot.

The second situation occurs during the regular use of the robot. Similarly to the setup process, the ongoing human-robot interaction might carry some elements of guided learning. Each time new tasks are introduced, and objects referred to, the robot will attempt to understand the instructions to its best ability. If the ambiguity in the occurring situation is very low, the robot might be able to infer already from part of the interaction about what it is needed to do. In the case the robot's prediction was correct, it can regard the rest of the interaction as additional input and extract for example object specific names or relationships from this. A person could call the robot to "bring me my bag" and indicate an area with only one bag. In this moment the robot would add the information that this is the person's personal bag and recall this information the next time the object is referred to and can act on it even without the same clear indication of location.

While this naturally occurring learning moments will be crucial to enable an intelligent and fluid use and interaction of the robot, it will be necessary to create a defined process to handle the situation for when errors occur. In the setup process as well as during the latter use, the robot will undeniable run into problems with translating instructions into action. In these moments, we propose the following steps could be considered to arrive to an adequate solution in terms of situation handling as well as solving the problem.

1. **Error occurrence.** It is possible that the robot itself realizes a lack of information hindering the fulfillment of a request or the failure to recognize an expected object. The information of error occurrence can also be sent to the robot by the human when they notice that the robot is making a mistake. In general, this means the robot has encountered one or more of the object recognition obstacles.
2. **Error identification.** After noticing that an error is occurring, the robot should attempt to analyze the cause of the error. It might be able to point to low confidence in the understanding of an instruction or the inability to match the instruction to any object in his surroundings.

   In the case that the human triggered the error notice, the robot might be able to take a second chance on accomplishing the task. The error notice itself might add enough information to enable a better understanding of what was the initial request like when two similar objects were present. Overall, in the scenario that the human informs the robot of the error, the robot would ideally restart the last task with the information that this object was the wrong one and see if it can solve it now, or be programmed to await further instruction from the user in that moment.
3. **Obstacle Communication.** The robot needs to convey the right information to enable a productive user interaction. The design of the robot and it´s system should provide it with the ability to communicate which part of the current task leads to interruption. For example, the robot might be able to identify a range of objects that fit the instruction and ask the user for a relatively simple clarification or share its own current view to the human, transmitted through the discussed media use.
4. **User Instruction.** In the fourth phase, the main part of guided learning takes place. The human acts on the given information. The user will take steps as discussed in the *Instructional Actions* section to deliver additional input to the robot or manipulate the surroundings to enable recognition on the part of the robot. The

choice of action will be foremost based on the encountered obstacle and secondly on the available input methods. Ideally a simple interaction can solve the problem but if the problem persists, the interaction might cycle through a set of possible solutions asking consequently more information and help from the user.

5. **Learning Feedback.** Lastly, the robot should confirm the newly learned information with the user. It should clarify to the human what happened and in case, what new information has been added to the database. This serves the purpose of avoiding learning something wrong, which could lead to greater complications later on. This feedback could be supported by knowledge graph representation.

7 Conclusion

Human-Robot interactions are composed of several key factors which provide a significant influence on the exploration of suitable interaction methods. A design solution identified on one aspect is not guaranteed to be a good solution for all aspects of the use scenario. Analyzing a scenario from multiple key factors can associate them into an integrated view which is a way to discover hidden problems and opportunities [12]. In this paper we analyzed the key factors (1) Obstacles (2) Actions (3) Media and came to propose a generalized process supporting guided learning for solving object recognition errors. With this information, we aim to provide a basis for designing successful guided learning interactions. Further research on detailed scenarios will be necessary and should involve additional key factors to complete the design. We suggest including the following factors: (4) User- mobility and capabilities (5) Environment- public, private (6) Human-Robot distance derived from the Design Information Framework [13].

Acknowledgments. This paper was supported by the Funds Project of Shanghai High Peak IV Program (Grant DA17003).

References

1. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 345–360. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_23
2. Ekvall, S., Jensfelt, P., Kragic, D.: Integrating active mobile robot object recognition and SLAM in natural environments. In: Proceedings of IEEE/RSJ International Conference Intelligent Robots and Systems, pp. 5792–5797 (2006)
3. Shectman, E., Irani, M.: Matching local self-similarities across images and videos. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2007)
4. Shi, G., Xu, T., Luo, J., Guo, J., Zhao, Z.: Alleviate Similar Object in Visual Tracking via Online Learning Interference-Target Spatial Structure, 19 October 2017. Published online
5. Gao, T., Packer, B., Koller, D.: A segmentation-aware object detection model with occlusion handling. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2011

6. Khurana, K., Awasthi, R.: Techniques for object recognition in images and multi-object detection. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **2**(4), 1383 (2013)
7. Darrell, T.: Overcoming Ambiguity in Visual Object Recognition. UC Berkeley Berkeley EECS Department & International Computer Science Institute (ICSI) (2010)
8. Chao, F., Wang, Z., Shang, C., Meng, Q., Jiang, M., Zhou, C., Shen, Q.: A developmental approach to robotic pointing via human–robot interaction, **283** (2014)
9. Anderson, H.: Edge Detection for Object Recognition in Aerial Photographs, February 1987
10. Cantrell, R., Schermerhorn, P., Scheutz, M.: Learning actions from human-robot dialogues. In: RO-MAN 2011, pp. 125–130. IEEE (2011)
11. Lauria, S., Bugmann, G., Kyriacou, T., Bos, J., Klein, E.: Training personal robots using natural language instruction **15**, 38–45 (2001)
12. Lim, Y., Sato, K.: Development of design information framework for interactive systems design. In: Proceedings of the 5th Asian International Symposium on Design Research, Seoul, Korea (2001)
13. Lim, Y.-K., Sato, K.: Describing multiple aspects of use situation: applications of Design Information Framework (DIF) to scenario development. Elsevier Ltd. (2005)