GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

# Finding Regions of Interest from Multimodal Human-Robot Interactions

*Pablo Azagra, Javier Civera, Ana C. Murillo*

DIIS-i3A, University of Zaragoza, Spain
{pazagra,jcivera,acm}@unizar.es

## Abstract

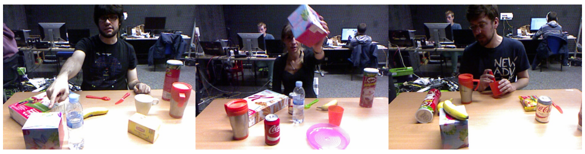
Learning new concepts, such as object models, from human-robot interactions entails different recognition capabilities on a robotic platform. This work proposes a hierarchical approach to address the extra challenges from natural interaction scenarios by exploiting multimodal data. First, a speech-guided recognition of the type of interaction happening is presented. This first step facilitates the following segmentation of relevant visual information to learn the target object model. Our approach includes three complementary strategies to find Regions of Interest (RoI) depending on the interaction type: Point, Show or Speak. We run an exhaustive validation of the proposed strategies using the recently published Multimodal Human-Robot Interaction dataset [1]. The currently presented pipeline is built on the pipeline proposed with the dataset and provides a more complete baseline for target object segmentation on all its recordings.

**Index Terms**: multi-modal data, Human-Robot interaction, object recognition

## 1. Introduction

Service and assistance robotic platforms need to be able to work in real user environments. One key step towards this goal is to enable a natural and seamless human-robot interaction. This topic has been notably approached by many researchers in the past [2, 3, 4], that identified the most relevant requirements for such interactions: learning world models, affordances and capabilities from the user's knowledge and behavior [5]. Natural human interaction with a robot presents challenging and complex situations. However, it also provides the advantages of rich multimodal data, because humans typically communicate combining speech and gestures. This multimodal nature can be exploited to improve the recognition performance in autonomous robotic platforms and is one of the focuses of this work.

The main goal of this work is to enable the robot to learn new visual models from human interaction. An essential task to solve this problem is to automatically identify which scene information is relevant for the model being learned. This information can be used to train offline models or to update incremental models of the target objects. This last step is out of the scope of this paper. We are currently focusing on how to take advantage of the multimodal nature of the human robot interaction to easily identify the relevant information in the scene. It is well known that multimodal data can enable tasks that are not feasible using a single sensor data. As a motivation example, Fig. 1 shows images from different user interactions to teach new object models to a robot. The general appearance of the scene is very similar in all of them, but the strategies to find the target or object of interest in the image should be adapted depending on the interaction. Therefore, we investigate how to split the problem of identifying relevant information into a hierarchy of smaller tasks: 1) recognize the action that the user is performing; 2) extract information from the acquired sensory data which is relevant to the object of interest.



(a) *Point* (b) *Show* (c) *Speak*

Figure 1: *Sample images from the three user interaction types considered. The user is saying: (a) "This is a box", while pointing to the box; (b) "This is a box", while holding the box; (c) "The banana is on top of the box."*

We propose a simple but effective pipeline that boosts the possibilities and effectiveness of a visual recognition pipeline thanks to basic speech processing integration. Figure 2 shows an overview of the proposed pipeline. The experiments presented in this work provide a more complete baseline of results for the recently presented MHRI dataset [1] and demonstrate how the proposed approach effectively segments Regions of Interest (RoI) on a multimodal dataset from natural human-robot interactions (HRI). As future work, we aim to investigate how to make incremental learning approaches more robust to the noisy training data that the robot can get from natural interactions with a human user.

## 2. Related Work

There are multiple relevant topics for human-robot interaction. We focus on the type of interaction occurring between the robot and the human and the multimodal data processing.

Multimodal data intuitively seems to provide important advantages to advance towards the goal of seamless communication between robots and humans. For example, Matuszek et al. [6] shows an approach similar to our goals of learning from human-robot interactions, but considers a simpler scenario. The user gives orders to pick pieces to the robot using both the hands and the speech. Their results are an example of how multimodal information improves the robot recognition performance. More recently Whitney et al. [7] presented an approach for fetching objects using as well multimodal data from interactions between the user and the robot. The user asks the robot for an item and the robot does an estimation based on the information it obtains from the user. The robot keeps asking until the user gives him an approval.

Vatakis et al. [8] present a dataset of multimodal recordings, but the purpose of their dataset was to capture the reactions of users to stimuli with objects or images on a screen. In our work, we use the more recently presented MHRI dataset [1], which is composed of multimodal recordings of several users teaching different object classes to the robot. It contains synchronized data from multiple sensors (two RGB-D cameras, one HD camera and a microphone) and captures different types of interactions, which reflect the most natural human interactions to show or

10.21437/GLU.2017-15

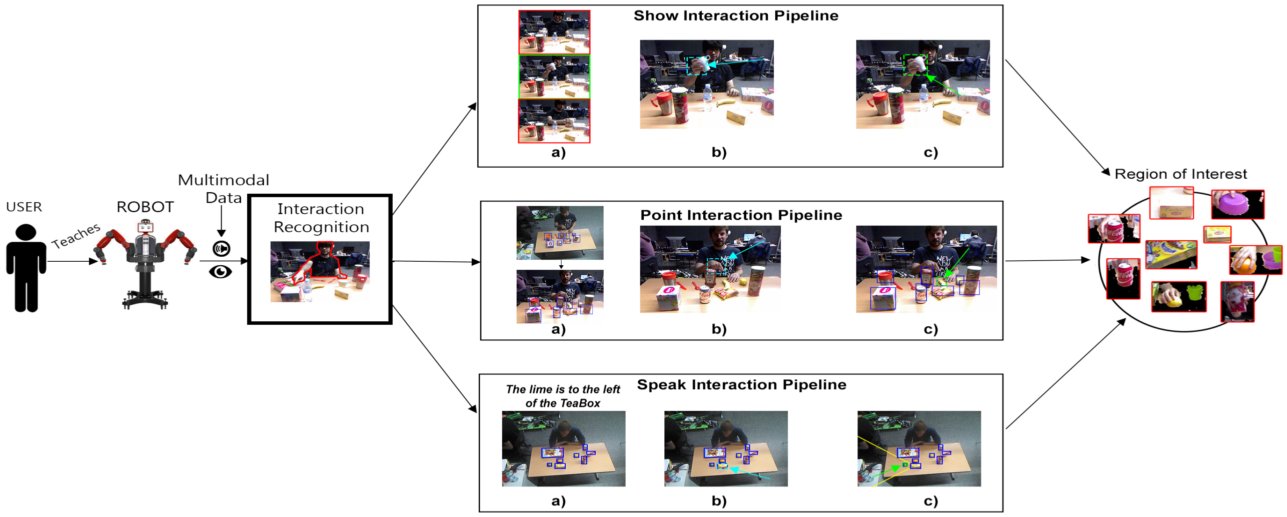

USER
Teaches
ROBOT
Multimodal Data
Interaction Recognition
Show Interaction Pipeline
a)
b)
c)
Point Interaction Pipeline
a)
b)
c)
The lime is to the left of the TeaBox
Speak Interaction Pipeline
a)
b)
c)
Region of Interest

Figure 2: *Region of Interest Extraction from Human Robot Interaction. We use the speech and the visual data to recognize the type of interaction. Each pipeline is divided in three steps: a) Initial information: Pre-process step to discard useless information or obtain possible candidates, b) Reference step: Obtaining the reference patch depending on the interaction and c) RoI extraction: Extraction of the Region of Interest based on the Reference step.*

teach objects: *Point*, *Show* and *Speak (Describing)*. Figure 1 shows an example of each of these three types of interactions.

As we can see in prior work, such as Abdullah et al. [9], combining language and image data can significantly boost the user action recognition. Our work shows how incremental object learning and recognition can benefit of this combination as well.

## 3. Strategies to segment Regions of Interest from Multimodal HRI

This section describes our approach to extract relevant information from multimodal HRI, in the context of learning new visual object models. The main motivation is to take advantage of utilizing more than one data modality acquired from natural user interactions while teaching the robot.

We consider three types of common interactions: *Point*, *Show*, and *Speak*. The motivation for our approach is that an initial classification of the type of interaction can simplify and optimize the object proposal task slightly different in each of these cases: 1) if the user is pointing to an object, we are likely to find that object by just following the finger directions; 2) if the user is grabbing an object and showing it to the robot, we know the object is held by the user's hand. In this case, the system should take into account that the hand is likely to occlude the object partially; 3) lastly, if the user is just talking, we can process the user speech to identify where the target object may be in the scene.

### 3.1. Recognizing types of human interactions

This section gives an overview of the proposed classifier for interaction type recognition, which is an improved version of the approach presented in [1]. It combines a sliding window approach with hierarchical classification for images and a winner-takes-all voting strategy to label the videos.

**Speech feature.** First, we identify the initial word that the user utters in any given interaction. This first word typically is:

- *"That"* when the user is pointing at a distant object
- *"This"* when the user is pointing at an object that is close, or showing something to the robot
- Any other word, usually *"the"*, if the user is describing something to the robot

This first word itself works as an effective feature to discriminate the *Speak* interaction from the Point and Show ones. To discriminate between *Point* and *Show*, as shown in [1], this simple step is not enough. Therefore, we make use of additional visual descriptors to discriminate between them.

**Visual features.** To distinguish between *Point* and *Show* interactions we need to analyze the user gesture information. First, before computing any visual feature, we remove the large irrelevant image regions, such as the background. Due to the nature of the dataset used [1], we can apply two filters to the image: one based on a standard background removal procedure and one using the depth map, in which we remove all the image regions that are further away than the user.

One we have applied such masks to the image, we run a simple hand gesture classifier. Specifically, ours is a sliding window-based approach (window size 100x100 pixels with a step of 10) to find the user hands on this filtered image. Finally, we describe the candidate hand-gesture patches using the following features:

- **Color histograms** $HC = [H_r \ H_g \ H_b]$, being $H_i = \sum_{x,y} p_i(x,y) \bmod B$ where $p_i$ is pixel $i$ component value and $B$ the number of bins.
- **HOG features** as described in [10].

With the above features we propose the following interaction recognition pipeline, based on two nested classifiers and voting across all frames:

1. SVM patch classification into *hand* vs *no-hand* classes. The patches are obtained with a sliding window approach, and the SVM is trained with random patches from the

dataset and manually selected patches of hands. This step only uses the $HC$ descriptor described above because it is easy to compute as a first pruning step.

2. SVM classification of resulting *hand* patches into *point* or *show* classes. Here we use both the $HC$ and the HOG descriptors.

3. Assign a label, *point* or *show*, to each video according to the label obtained by the majority of its frames. All windows from each video are labeled as that action for following steps.

### 3.2. Detecting regions of interest according to the type of interaction detected

In order to identify the relevant visual information for the object model to be learned, we take advantage of the previous interaction type classification. This classification allows us to build different strategies to find the regions of interest (RoI), as summarized in Fig. 2. All the strategies start from a reference known patch that serves as anchor or guide point to find the target RoI.

**Show interactions.** Following the outputs from previous step (Sec. 3.1), we have the patch with the user hand holding the shown object in multiple frames, which is processed further as follows.

First, we analyze the hand movement across frames, to select the hand patches from the frames where the user hand is on its highest position. This is the best moment to separate hand and object from the scene clutter. Then, we use SLIC [11] superpixels to segment image regions around the user hand and find a more accurate object area. We estimate the hand orientation as the direction of the first-order moment in the hand patch. Following this direction, we select which hand neighbouring superpixel is more likely to be our target object segment. Figure 2 shows an illustrative example of this *Show interaction pipeline* steps.

**Point interactions.** To process data from Point interactions, we use the advantages of the two RGB-D Cameras available in the dataset.

First, we use the top camera for searching the candidates in the table, since top view images are usually less cluttered. We use the frontal view camera to chose one candidate from the interaction recognized in the previous step. We extract the table plane from the image and take the blobs that are inside the table space as a candidate. Since we have calibrated the plane table homography, we can map the contour of the objects from the top camera to the frontal one. Then, using the orientation from the hand we pick the closest candidate that intersects this direction. Each considered frame votes to a candidate, and the most voted is chosen. A visual illustration of this *Point interaction pipeline* can be seen in Figure 2.

**Speak interactions.** This type of interaction presents different challenges than the other two. Since the visual part of the action is irrelevant, we analyze the user speech to get the information needed to obtain the candidate patch. We assume simple user descriptions, where using standard speech processing tools we can extract the name of the object the user is teaching, as well as the object used for reference and the relative direction or position. We use a very similar approach to find the RoI that the one for the *Point* interaction. The difference is that in this case the reference patch will be a known object, and the direction will be given by the user relative position rather than the user pointing direction. Also, we run the template search on the top camera, where it is easier to find non-occluded object candidates.

As before, we run an object detector on all candidate objects segmented on the top view. This has been trained on already known objects, and this prior information provides the patch that will serve as reference in this case. Finally, using the direction the user gives, we define a target area using the limits of the reference patch and the corner of the images. We select our target candidate as the closest one to the reference patch that is included inside that target area. Figure 2 shows a visual example of the proposed *Speak interaction pipeline*.

## 4. Experiments

This section presents experimental results that demonstrate the quality of the target patches obtained using the proposed strategies to analyze the multimodal interactions between the human and the robot. We have run the hierarchical pipeline proposed in this work, Fig. 2, on all the sequences of the dataset [1].

### 4.1. Interaction type classification.

We first evaluate the improved interaction type classification. As shown in Table 1, we obtain an improvement from the approach presented in [1]. While those results presented a **33%** at worst, we obtain almost **80%** of accuracy for the recognition of each interaction type.

| | **Point** | **Show** | **Speak** |
|---|---|---|---|
| **Point** | **85,71%** | 22,03% | 0,00 |
| **Show** | 14,29% | **77,97%** | 0,00 |
| **Speak** | 0,00% | 0,00% | **100,00%** |

Table 1: *Confusion matrix for interaction recognition approaches. Left: ground truth, top: prediction.*

### 4.2. Target objects RoI segmentation

As described, we apply different strategies to segment the RoI patches depending on the type of interaction:

- Point interactions: we consider all the frames and each one votes a candidate as explained before. The candidate with most votes is chosen. The ROI are obtained from the first five frames, since they have less hand occlusion to the object.
- Show interactions: we consider the frames that correspond to the highest hand location and obtain the ROI on them.
- Speak interactions: although the scene is more or less static, we consider the first five frames to obtain robust candidates. The output is the chosen ROI in the first frame.

We present results for the three types of interaction in the dataset. Fig. 3 shows the scene view and the output of the main steps for examples of the different cases. Figure 4 presents failure examples. In the example for *Show*, the hand is occluded by the object, so the hand recognized is the other one and the ROI is not useful. In the example from *Point*, we can see the finger is pointing in a direction that does not follow the dominant orientation of the hand. In the example from *Speak*, we can see that the reference object (bottle) is very close to the target object. In this case the segmentation considers both objects as one large

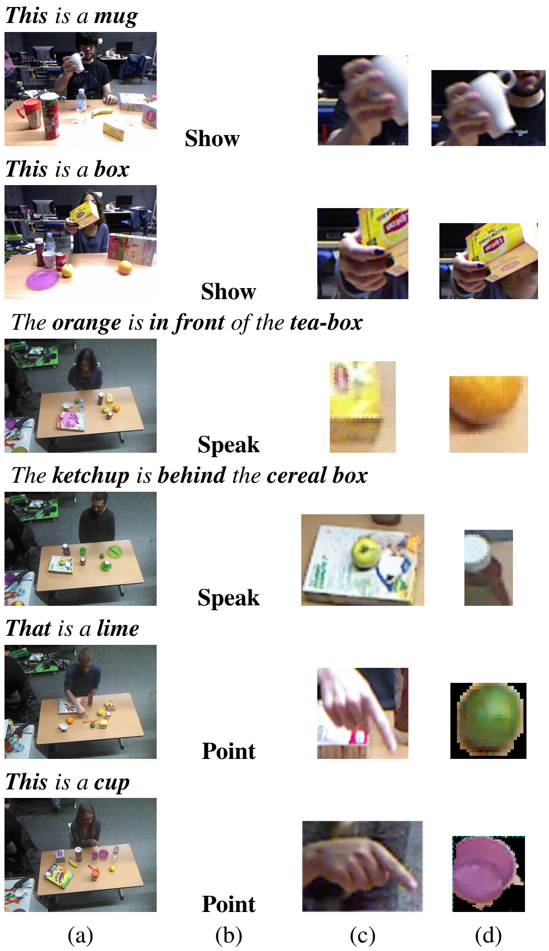


Figure 3: *Examples of target object segmentation from the different types of interactions: (a) Input data: scene image observed from the robot camera & speech; (b) Interaction type recognized; (c) Reference patch detected depending on the interaction type (hand or reference object); (d) target object patch segmented.*

reference segment, and leaves only a small part of the actual target object as such.

Fig. 5 shows preliminary results of merging not only information from speech and video but from the multiple cameras available in the dataset. For small occlusions, it is possible to match candidate object segments in more than one camera view, and therefore incorporate to the learning process multiple views of the target object at once.

Table 2 summarizes the percentage of valid object patches over the total found. We consider valid a patch that contains at least half of the object being referred to in the interaction. We obtain almost **70%** for the show approach and around **50%** for pointing and speaking. These results show a large number of correct object patches that can be incorporated into following object learning pipelines, but it also highlights the need of incremental learning methods which are robust to noisy data and able to detect outliers.

## 5. Conclusion

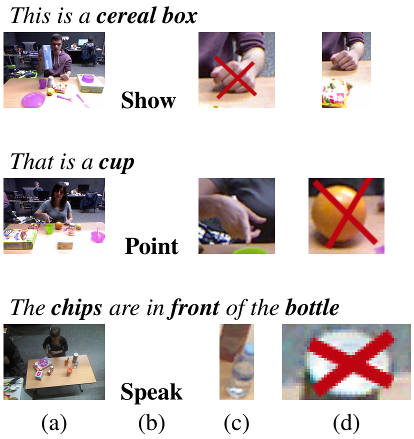


Figure 4: *Examples of incorrect target object segmentations: (a) Input image; (b) Interaction type recognized; (c) Reference patch detected; (d) RoI segmented from the target object. Usually these errors consist of an incorrect (first row) or noisy (second row) segmentation of the intermediate reference patch, or the target object is not fully included in the final RoI (third row).*

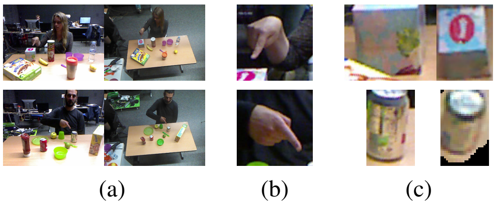


Figure 5: *Two examples from Point interactions using Two Cameras, where we can obtain two patches of each target object: (a) initial images; (b) reference patch (c) RoI segmented from the two cameras. This strategy can provide richer RoIs, but it is also more challenging due to miss-alignments and occlusions.*

| | % Valid Patches |
|---|---|
| **Show** | 68,15 |
| **Point** | 52,24 |
| **Speak** | 50 |

Table 2: *Amount of correct target object patches among all the patches found for each type of interaction.*

This work presents an approach to identify relevant information about new object models from multimodal images (visual and speech) data. We validate our proposed strategies with the realistic scenarios from the MHRI dataset. We propose to first recognize the type of interaction, and then propose a different pipeline for each type of interaction. We evaluate the quality of the segmented target object patches, which is a clear example of how simple multimodal integration can provide significant advantages. We also analyze the advantages and weaknesses of the current approach. Future lines of work will extend the presented approach by integrating the target object segmentation with incremental object model learning and improving the hand orientation taking account of the user's head position.

**Acknowledgments.** This research has been partially funded by the European Union (CHIST-ERA IGLU, PCIN-2015-122), the Spanish Government (projects DPI2015-67275, DPI2015-65962-R, DPI2015-69376-R), the Aragon regional government (Grupo DGA T04-FSE), the University of Zaragoza (JIUZ-2015-TEC-03).

# 6. References

[1] Azagra Pablo, Yoan Mollard, Florian Golemo, Ana Cristina Murillo, Manuel Lopes, and Javier Civera. A multimodal human-robot interaction dataset. In *FILM Workshop on NIPS 2016*, 2016.

[2] Ulrich Reiser, Christian Pascal Connette, Jan Fischer, Jens Kubacki, Alexander Bubeck, Florian Weisshardt, Theo Jacobs, Christopher Parlitz, Martin Hägele, and Alexander Verl. Care-o-bot® 3-creating a product vision for service robot applications by integrating design and technology. In *IROS*, volume 9, pages 1992–1998, 2009.

[3] J. Dumora, F. Geffard, C. Bidard, N. A. Aspragathos, and P. Fraisse. Robot assistance selection for large object manipulation with a human. In *2013 IEEE International Conference on Systems, Man, and Cybernetics*, pages 1828–1833, Oct 2013.

[4] J. Baraglia, M. Cakmak, Y. Nagai, R. Rao, and M. Asada. Initiative in robot assistance during collaborative task execution. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 67–74, March 2016.

[5] Jivko Sinapov, Connor Schenck, and Alexander Stoytchev. Learning relational object categories using behavioral exploration and multimodal perception. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5691–5698. IEEE, 2014.

[6] Cynthia Matuszek, Liefeng Bo, Luke Zettlemoyer, and Dieter Fox. Learning from unscripted deictic gesture and language for human-robot interactions. In *AAAI*, pages 2556–2563, 2014.

[7] David Whitney, Miles Eldon, John Oberlin, and Stefanie Tellex. Interpreting multimodal referring expressions in real time. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 3331–3338. IEEE, 2016.

[8] Argiro Vatakis and Katerina Pastra. A multimodal dataset of spontaneous speech and movement production on object affordances. *Scientific Data*, 3:150078 EP –, Jan 2016. Data Descriptor.

[9] Lili Nurliyana Abdullah and Shahrul Azman Mohd Noah. Integrating audio visual data for human action detection. In *Computer Graphics, Imaging and Visualisation, 2008. CGIV'08. Fifth International Conference on*, pages 242–246. IEEE, 2008.

[10] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 886–893. IEEE, 2005.

[11] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.