A Multimodal Human-Robot Interaction Dataset

Pablo Azagra, Yoan Mollard, Florian Golemo, Ana Cristina Murillo, Manuel Lopes, Javier Civera

A Multimodal Human-Robot Interaction Dataset

Pablo Azagra
Universidad de Zaragoza
pazagra@unizar.es

Yoan Mollard
Inria Bordeaux Sud-Ouest
yoan.mollard@inria.fr

Florian Golemo
Inria Bordeaux Sud-Ouest
fgolemo@gmail.com

Ana Cristina Murillo
Universidad de Zaragoza
acm@unizar.es

Manuel Lopes
INESC-ID, Univ. de Lisboa, Portugal
manuel.lopes@inria.fr

Javier Civera
Universidad de Zaragoza
jcivera@unizar.es

Abstract

This works presents a multimodal dataset for Human-Robot Interactive Learning. The dataset contains synchronized recordings of several human users, from a stereo microphone and three cameras mounted on the robot. The focus of the dataset is incremental object learning, oriented to human-robot assistance and interaction. To learn new object models from interactions with a human user, the robot needs to be able to perform multiple tasks: (a) recognize the type of interaction (pointing, showing or speaking), (b) segment regions of interest from acquired data (hands and objects), and (c) learn and recognize object models. We illustrate the advantages of multimodal data over camera-only datasets by presenting an approach that recognizes the user interaction by combining simple image and language features.

1 Introduction

Motivation. In the latest years we are witnessing a widespread adoption of robotics in multiple aspects of our daily life and activities, for example in household assistance devices and autonomous cars. One of the key aspects in service robotics is comfort / intuitive control in human-robot interaction, a topic notably approached by [10, 3, 5] among others. Some of the most relevant requirements for such interactions are learning world models, affordances and capabilities from the user's knowledge and behavior –as seen in [11] for example.

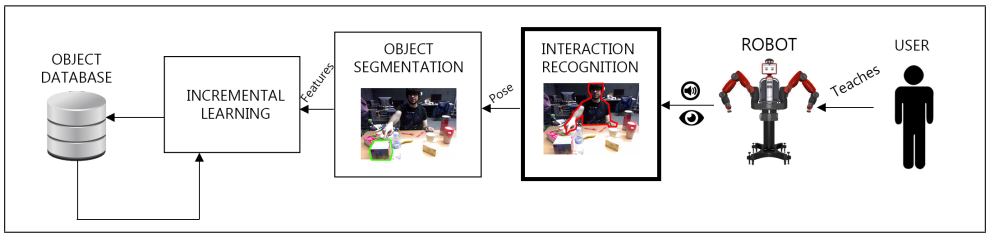


Figure 1: An interactive human-robot framework for incremental learning requires multiple modules: action recognition, object segmentation, incremental learning. This work includes an illustrative experiment of the action recognition step utilizing the presented MHRI dataset.

From the robot's perspective, a framework for incremental object learning from human interaction should contain multiple modules: 1) recognizing the action that the user is performing, 2) extracting information (from the acquired sensory data) that is relevant to the object of interest (as for example in [9]), 3) assigning a class to the object of interest (with works like [7]) and incrementally update the model for each object class (with an approach similar to [4]). Figure 1 shows a schematic overview

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

of this pipeline. Each of these modules is challenging on its own, specially if we address realistic environments and interactions. This work presents a public dataset for incremental object learning from human-robot interactions. We focus on multimodal data, since it intuitively seems to provide important advantages to advance towards the goal of seamless communication between robots and humans. However, multimodal data also presents additional challenges.

Contributions. The presented dataset contains synchronized acoustic (stereo microphone), visual (three cameras), and depth information (two of the cameras contain infrared sensors). To our knowledge, this is the first publicly available dataset that contains user interaction data from the robot's perspective with synchronized multi-camera and microphone data. We present an illustrative experiment showing the benefits of multimodal data.

Related Work. There are multiple aspects that are of relevance for interactive learning. In our particular case, we focus on the type of interaction and the multimodal data. In a realistic human-robot interactive learning scenario, the training data is expected to be presented in a very different way than standard datasets. The data should capture the human operator and his actions. That means the data should be recorded from the point of view (POV) of the robot. Recognizing a pedestrian from a close-up view from service robot is immensely different from performing the same task with the raw video data from a distant wide-angle surveillance camera.

Vatakis et al. [12] shows multimodal recording approach similar to ours, but the purpose of their dataset was to capture the reactions of users to stimuli with objects, or images on a screen. Datasets like [8] or [6] capture human-robot interaction from a third-person POV. This is useful in lab environments, but since we are working with service robots, information must be taken from the onboard sensors. Additionally, many datasets lack additional sensor data that is easy to find in human-robot interactive scenarios like speech from the user. As detailed in the next section, our dataset is focused not only on capturing the user's face, but also on capturing the scene information (hands, arms, desk, objects, etc.) related to the object classes being taught to the robot. This additional information is useful for future work, but for the purpose of this article we only labeled the interaction type in each clip.

2 The Multimodal Human-Robot Interaction (MHRI) dataset

The MHRI dataset[1] is composed of recordings of several users teaching different object classes to the robot. It contains synchronized data from the following sensors:

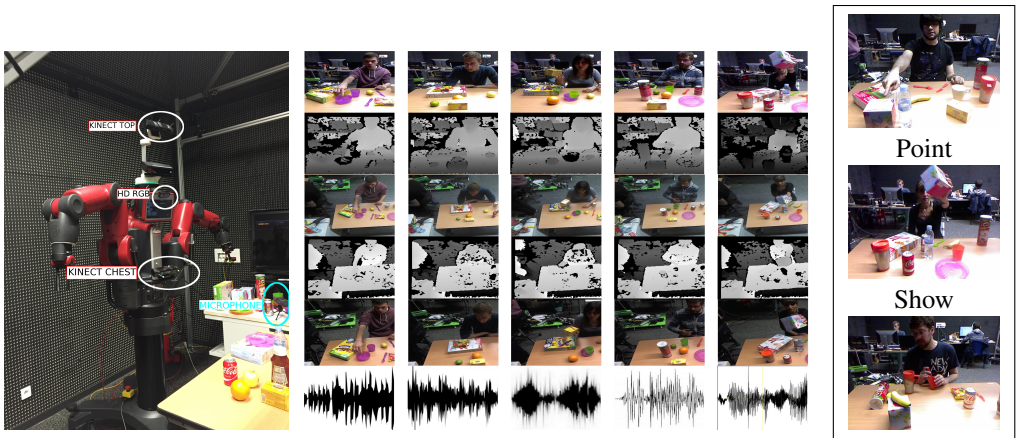
- *Chest-Kinect*. Microsoft Kinect v1.0 (640×480). This first camera is mounted on the chest of the robot and is focused on the frontal interaction with the user.
- *Top-Kinect*. Microsoft Kinect v1.0 (640×480). This camera is mounted on the head of the robot and gives a top global view including the user, the workspace and the objects.
- *Face-HDcam*. 1280×720 *RGB* camera mounted on the screen of the Baxter robot and focused on the user's face.
- *Audio*. Speech from the user, recorder with a USB microphone situated on the side of the table.

The cameras are mounted so that the user and the table on which the interaction occurs, are perfectly covered. Figure 2a shows the placement of the cameras in the Baxter robot used for the acquisition and Figure 2b shows some examples of the recordings.

The types of interactions captured in the dataset reflect the most common ways users show or teach objects to the robot: *Point*, *Show*, and *Speak*. The *Point* interaction captures a user pointing at an object and calling out its name to the robot. The *Show* interaction describes a user grabbing an object and bringing it approximately in front of the robot's torso camera while announcing the object's name. The *Speak* interaction captures a user describing to the robot where a certain object is placed. Figure 2c shows an example of each of these three types of interactions.

Table 1 contains a summary of the contents of the dataset. We recorded 10 users, each of them performing 10 object interactions for each of the 3 tasks (point, show, speak) for a total of 300

[1]The dataset can be downloaded at `http://robots.unizar.es/IGLUdataset/`



(a) (b) (c)

Figure 2: (a) The Baxter robot setup used to record the data. (b) Examples from the dataset. Each column shows the same event captured from the 4 sensors. The first row contains the rgb-depth pair of images from the *Chest-Kinect*; the second row shows the rgb-depth pair from the *Top-Kinect*; the third one is the *Face-HDcam* and the last one represents the audio recorded from the user.(c) Sample images from the three user interaction types considered. The user is saying (from top to bottom) : "This is a box", while pointing at the box; "This is a box", while holding the box; "The box is next to the chips and has a banana on top."

recordings. Each user got assigned a varied set of 10 objects out of a pool of 22 objects on the table, so that all objects have a roughly equal presence in the dataset. They were given unspecific instructions on the interactions, meaning no exact phrasing. As a result, there is a natural variation in the language usage between speakers.

Table 1: MHRI Dataset content summary.

Users	10	
Interaction Types (Actions)	3	*Point, Show, Speak*
Interactions per User	30	10 of each type. 1 random object per interaction.
Objects	22	*Apple, Banana, Bottle, Bowl, Cereal Box, Coke, Diet Coke, Ketchup, Kleenex, Knife, Lemon, Lime, Mug, Noodles, Orange, Plate, Spoon, Tall Mug, Tea Box, Vase*

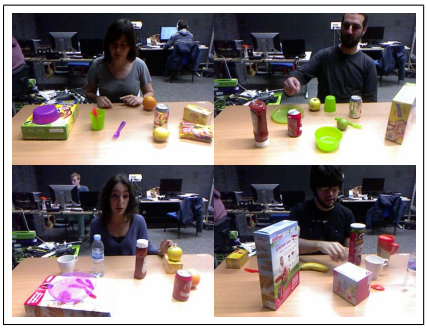
3 Multimodal Action Recognition

This section describes an illustrative experiment showing the advantages of utilizing more than one modality of data acquired from the user interactions while teaching the robot. As we can see in prior work, such as Abdullah et al. [1], combining language and image data can significantly boost the user action recognition.

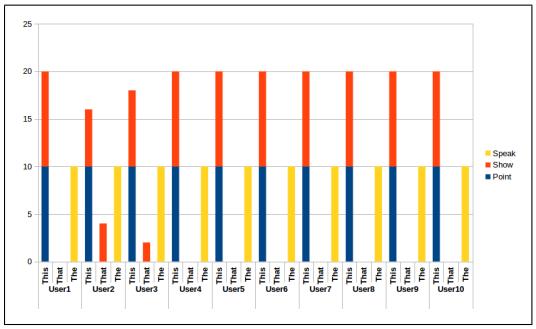
Our dataset includes three types of interaction: *Point*, *Show*, and *Speak*. A rough classification of the type of interaction can simplify and optimize follow-up tasks, for example segmenting the object. If the user is pointing at an object, our system could estimate an image region where the object is very likely to be. 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.

This experiment aims to recognize the three types of interactions described combining information from language and visual cues. Since we are seeking a simple pre-filter for an incremental object learning, and in order to avoid costly spatio-temporal video analysis, we propose a low-cost single-

frame classifier. Figure 3a highlights the difficulty of the interaction recognition problem from a single frame. Classifying a single frame into *Point*, *Show*, or *Speak* can be difficult even for a human.



(a)



(b)

Figure 3: Challenges of interaction recognition with a single data source. (a) With visual information, it is often hard to distinguish from a single frame if the user is pointing at an object, about to pick it up, or just narrating. (b) Stacked graphic showing the language feature distribution on all recorded user interactions. Clearly, *Point* and *Show* actions can not be separated using only this language feature.

Visual data features Our image descriptor is computed as follows. First, we segment the image into SLIC [2] superpixels. We classify each superpixel as skin/not skin using color and depth and fuse the adjacent skin superpixels into blobs. Afterwards we divide the image into a 5×5 grid, with the descriptor being a histogram counting the skin votes of the blobs per each image cell.

Language data features Our language feature is the first word of each user sentence, which is 1) "*that*" when the user is pointing to something far; 2) "*this*" when the user is pointing to something close or showing something to the robot; 3) any other word if the user is just describing something to the robot, usually "*the*". Figure 3b shows the language feature distribution for each user and interaction. Notice that the feature is not discriminative for the classes *Point* and *Show*. A classifier cannot be trained with only this feature because the performance on those classes will simply be random.

Interaction classifier We trained a RBF-SVM classifier first using only visual data and then using the combination vision-language. We report the confusion matrix (10-fold cross-validation) for both cases. Table 2a is vision-only and Table 2b contains the vision&language model. When comparing Table 2a and Table 2b, we can see the improvement in the division of *Speak* with the combination of visual-language data. In case of *Point* and *Show* there is also improvement, mostly focus on the *Show* interaction which is more discriminant with the language data.

Table 2: Confusion matrix for (a) interaction recognition using only visual data or (b) combining both visual and language features. Each row represents the ground truth and each columns the SVM-prediction.

	Point	**Show**	**Speak**
Point	**72,85%**	12,36%	14,78%
Show	76,01%	**12,88%**	11,11%
Speak	55,46%	20,00%	**24,54%**

(a)

	Point	**Show**	**Speak**
Point	**73,94%**	26,06%	0,00%
Show	66,45%	**33,55%**	0,00%
Speak	0,00%	0,00%	**100%**

(b)

4 Conclusion

This work presents an annotated multimodal dataset for Human-Robot interaction. A remarkable feature of the dataset is the synchronized recording of multimodal data, specifically two *RGB-D*

cameras, one high resolution *RGB* camera, and audio data. We have shown how even a simple task, like classifying the type of user interaction from single-frame descriptors, can significantly benefit from multimodal data. The interaction recognition is a crucial step to facilitate future processing steps, such as automatic object and region of interest segmentation in images and videos. In addition to the interaction, our dataset presents other challenges like occlusions and low object resolution. Future lines of work, taking advantage of the multimodal dataset presented, will extend the presented approach by working on the next steps of an incremental and interactive learning framework.

Acknowledgments

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

References

[1] L. N. Abdullah and S. A. M. 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.

[2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

[3] 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.

[4] R. Camoriano, G. Pasquale, C. Ciliberto, L. Natale, L. Rosasco, and G. Metta. Incremental object recognition in robotics with extension to new classes in constant time. *arXiv preprint arXiv:1605.05045*, 2016.

[5] 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.

[6] W. Gong, J. Gonzàlez, J. M. R. S. Tavares, and F. X. Roca. *A New Image Dataset on Human Interactions*, pages 204–209. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[7] P. Iravani, P. Hall, D. Beale, C. Charron, and Y. Hicks. Visual object classification by robots, using on-line, self-supervised learning. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1092–1099. IEEE, 2011.

[8] B. S. A. S. Jaeyong Sung, Colin Ponce. Human activity detection from rgbd images. *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*, 2011.

[9] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 1377–1382. IEEE, 2009.

[10] U. Reiser, C. P. Connette, J. Fischer, J. Kubacki, A. Bubeck, F. Weisshardt, T. Jacobs, C. Parlitz, M. Hägele, and A. 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.

[11] J. Sinapov, C. Schenck, and A. 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.

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