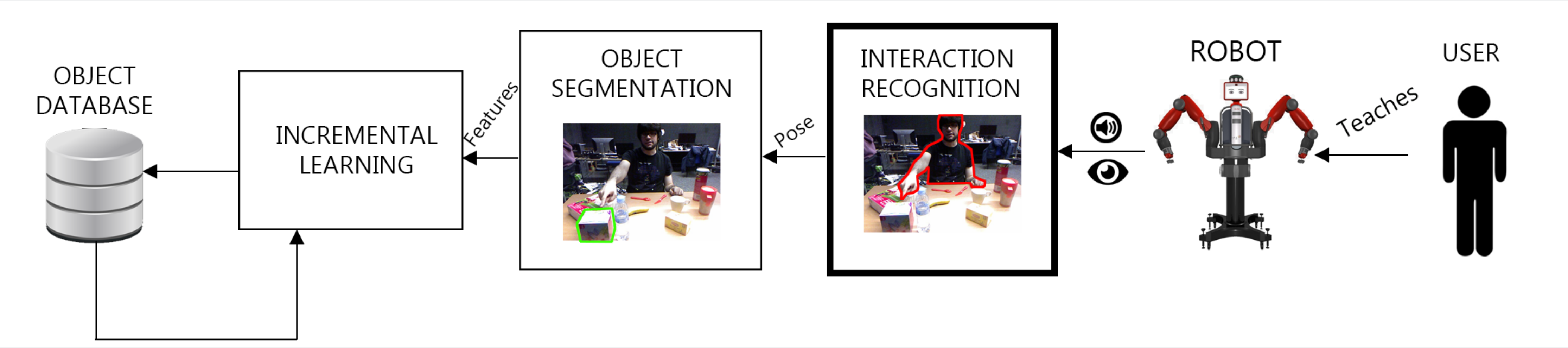
A Multimodal Human-Robot Interaction Dataset

Pablo Azagra[1*], Yoan Mollard[2], Florian Golemo[2],Ana C. Murillo[1], Manuel Lopes[2],Javier Civera[1]
[1] DIIS-i3A, Universidad de Zaragoza, Spain. ([*]pazagra@unizar.es).
[2]Centre de Recherche Inria Bordeaux - Sud-Ouest, France.

Introduction

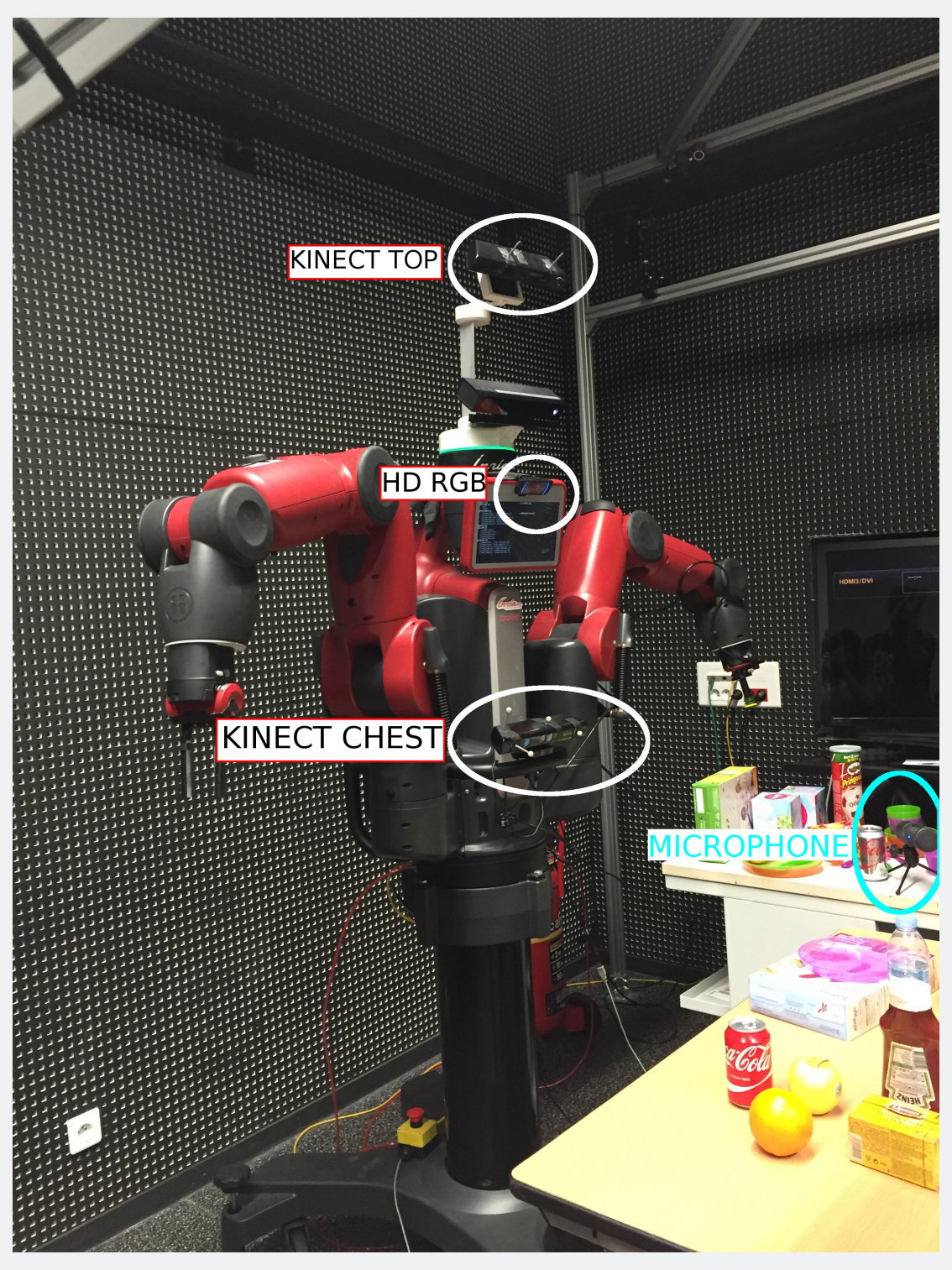
A framework for incremental object learning from human-robot interaction, from the robot's perspective, should contain the multiple modules shown in the figure. This work presents a public dataset for incremental object learning from human-robot interactions.



The Multimodal Human-Robot Interaction (MHRI) dataset

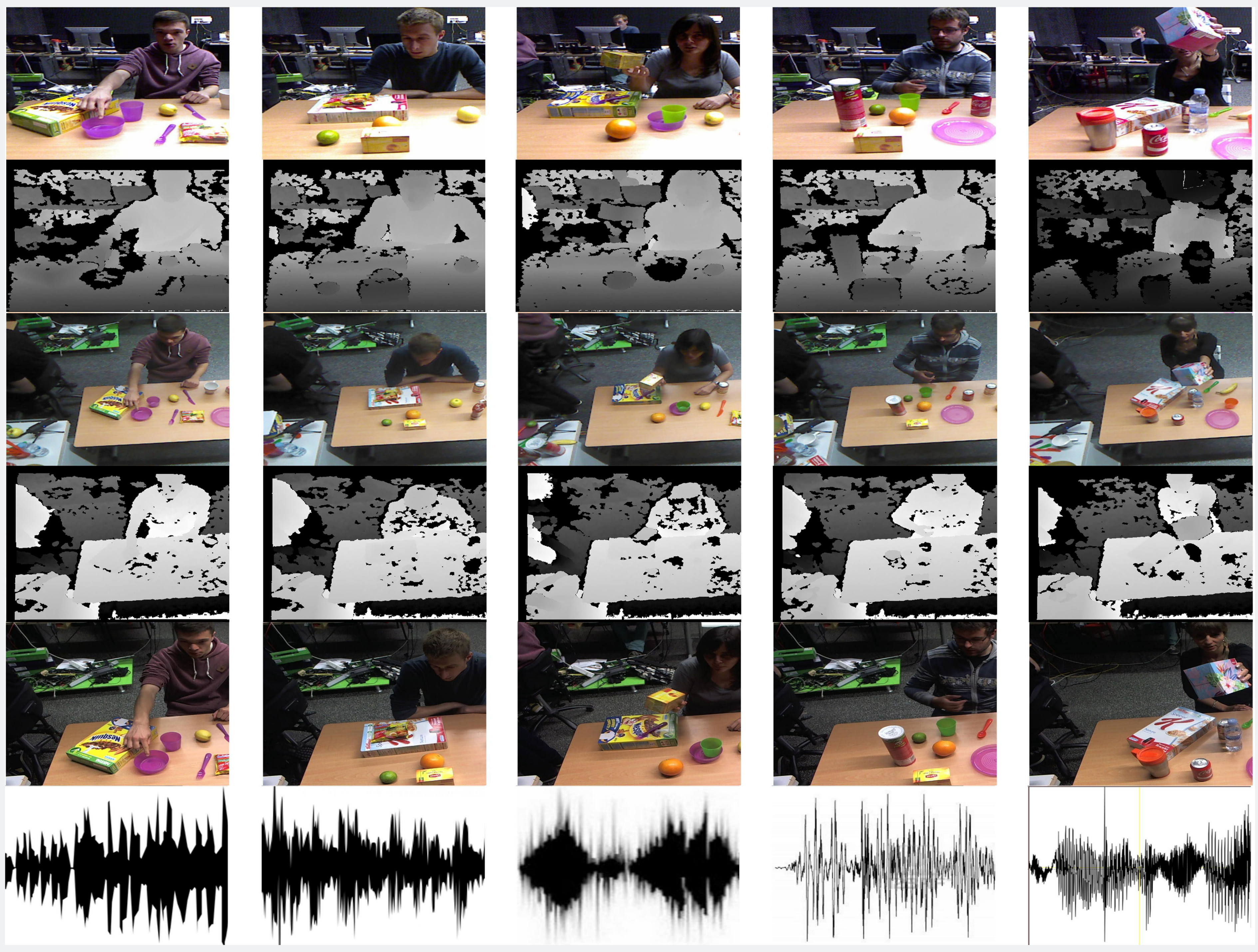
The MHRI dataset[a] contains recordings of users teaching objects to the robot Baxter using synchronized data from the following sensors:

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



Summary of the contents of the dataset:

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*



[a]can be downloaded at `http://robots.unizar.es/IGLUdataset/`

Acknowledgements

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

Related Work

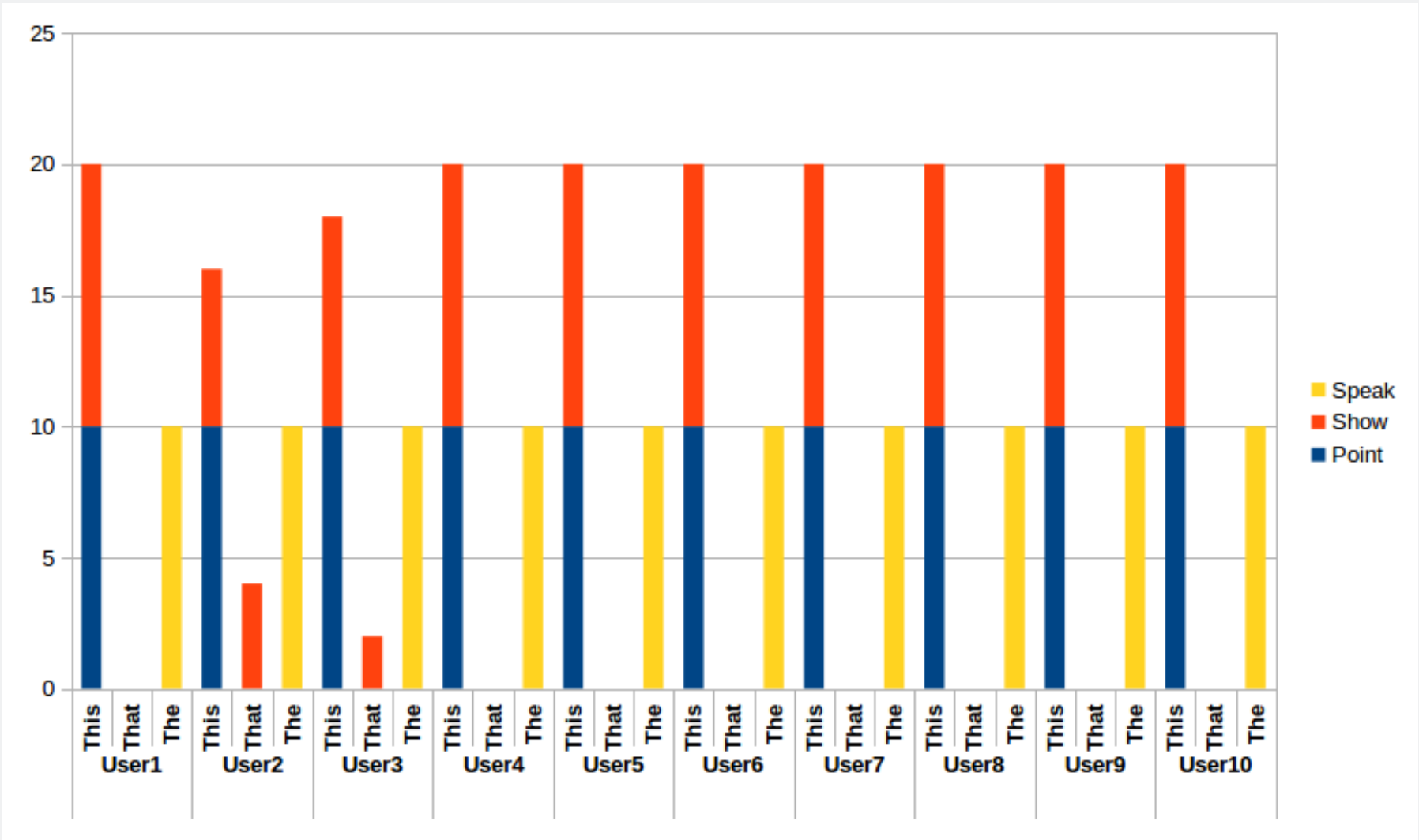
Other datasets:

- Vatakis et al. [4] shows multimodal recording approach similar to ours, but with other purpose.
- Gong et al. [2] and Sung et al. [3] capture human interaction from a third-person POV.

Multimodal Action Recognitio

SVM classifier trained with:

Language Feature: First word in the speech.



Visual Feature: SLIC [1] superpixels + skin/not skin with color and depth + 5 × 5 grid = histogram summing the skin votes.

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

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

Conclusions

- Annotated **multimodal dataset for Human-Robot interaction** (Two *RGB-D* cameras, one high resolution *RGB* camera, and audio data).
- Our dataset presents **challenges** like occlusions and low object resolution.
- Simple interaction classification improved by **Multimodal data**.
- Future lines of work will extend the presented approach by working on the next steps of an **incremental and interactive learning framework**.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *PAMI*, 2012.

[2] W. Gong, J. Gonzàlez, J. M. R. Tavares, and F. X. Roca. *A new image dataset on human interactions*, pages 204–209. 2012.

[3] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. *AAAI workshop PAIR*, 2011.

[4] A. Vatakis and K. Pastra. A multimodal dataset of spontaneous speech and movement production on object affordances. *Scientific Data*, 2016.

More Information

Sample data Video

Dataset download