

# TOWARDS SEMANTIC GROUNDING VIA VIDEO DATA: THE CASE OF *PUSH* & *PULL*

Benjamin Burkhardt

CRC 991, University of Düsseldorf

## My Project

- Development of a semantic representations of the verbs *push* and *pull* based on video data
- Requirement:** manipulation descriptions and manipulation videos must be represented in such a way that the two can be compared.

## "These Neuroscientists Have a Robot..."

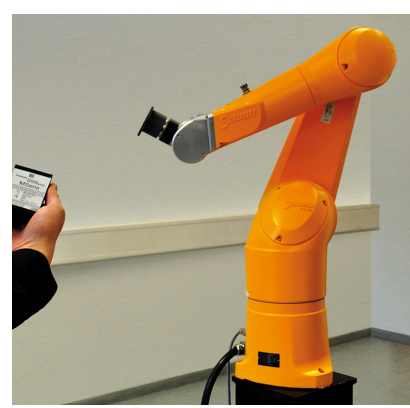
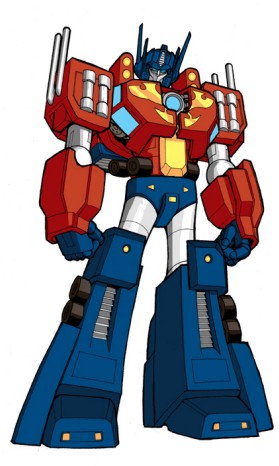


Figure 1: Imagination

Figure 2: Reality

- The robot's owner: Research Group at the Bernstein Center for Computational Neuroscience Göttingen (Project Leaders: Prof. Dr. Florentin Wörgötter & Dr. Eren Erdal Aksoy)
- Stereoscopic camera system for 3D vision
- Workbench to which the camera is mounted
- Computer to analyze the camera footage and control the robot arm

## What is it good for?

- Recording of 3D videos of simple manipulations: *PUSH*, *PUT*, *HIDE*, *STIR*, *CUT*, *CHOP*, *TAKE*, *UNCOVER*; manipulations were performed by 5 informants; each informant performed 3 versions of each manipulations
- Video Analysis:** object recognition in all frames, object tracking across video frames, object-relation-tracking
- Goal:** Enable the robot to learn and recognize various manipulation types based on their prototypical visual properties.

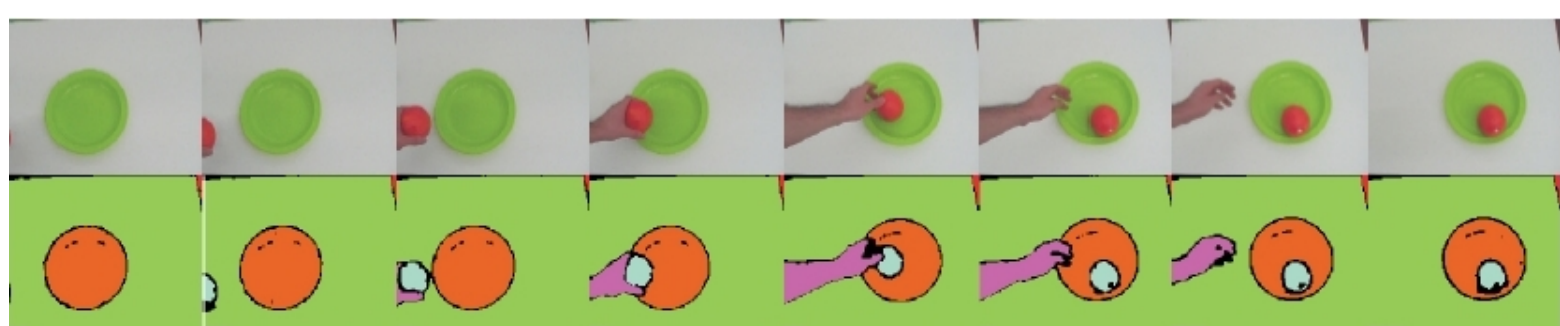


Figure 3: Object recognition & tracking

## Video Representation

- Videos are split into frames, every object receives a static object ID number, for every distinct pair of objects an algorithm determined the spatial relation between those objects
- Possible spatial relations:** Absent (-1), Non-Touching (0), Touching (1)
- Key frame:** a frame in the video in which at least one spatial relation changes compared to the previous frame
- Example scenario:** imagine a scene that shows a workbench table top with a box sitting on it. In the course of the video, a hand enters the scene, touches the box, and the video ends while hand and box still touch each other.

$$M'_{\text{scenario}} = \begin{matrix} & \begin{matrix} \text{keyframe 1} & \text{keyframe 2} & \text{keyframe 3} \end{matrix} \\ \begin{matrix} R_{\text{objectID}_{\text{bench}}, \text{objectID}_{\text{box}}} \\ R_{\text{objectID}_{\text{bench}}, \text{objectID}_{\text{hand}}} \\ R_{\text{objectID}_{\text{box}}, \text{objectID}_{\text{hand}}} \end{matrix} & \begin{pmatrix} 1 & 1 & 1 \\ -1 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix} \end{matrix}$$

Figure 4: A video-representation-matrix; first column: object tuples, key frame columns: object relations captured in the individual video frames

## PUSH vs PULL

- Learning Algorithm:** compares all videos that show the same manipulation type, represents their common properties as a manipulation-representation-matrix

$$M_{\text{push}} = \begin{pmatrix} R_{\text{objectID}_1, \text{objectID}_2} & -1 & 0 & 1 & 0 & -1 \end{pmatrix}$$

Figure 5: The learned representation for PUSH manipulations

- Problem I:** the dataset does not include videos of *PULL* manipulations
- Problem II:** judging from the spatial relations alone, *PUSH* and *PULL* cannot be differentiated (Intuition)
- Assumption:** *PUSH* and *PULL* are minimal pairs with respect to movement, the spatial relation changes are identical for the two manipulation types
- Problem III:** the learned representations do not include any explicit information about movement

## The Differentiating Factor

- Given the assumption that *PUSH* and *PULL* are so similar, the videos and video-representation-matrices for *PULL* manipulations were derived by reversing the videos and matrices of the *PUSH* manipulations.
- Video analysis raw data include every object's current position in each frame.
- Observation:** during *PUSH* manipulations the agent object always stays behind the theme object relative to the movement direction; after agent and theme have stopped moving, the theme object can be found on the agent's extended movement path. Vice versa for *PULL* manipulations.

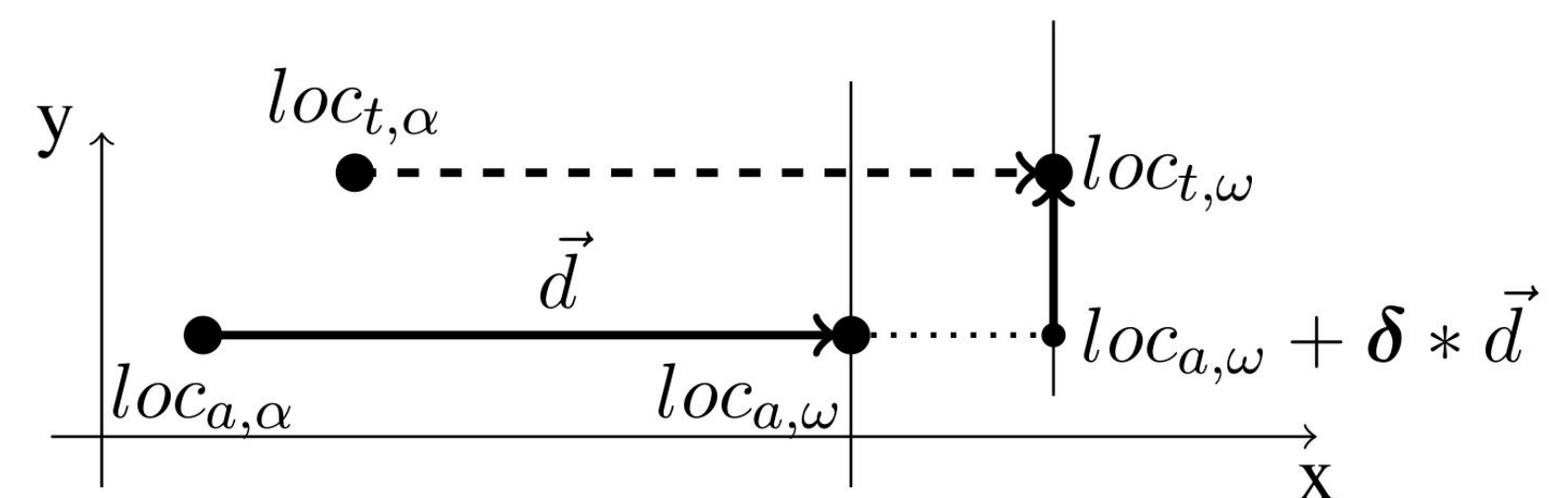


Figure 6: PUSH graph representation; agent and theme object travel along parallel paths.

- If  $\delta$  is positive, we can identify a manipulation as *PUSH*. If  $\delta$  is negative, we have a *PULL* manipulation.

## A Grounded Representation for PUSH & PULL

- The Representations for *PUSH* and *PULL* manipulations combine object relation and location information requirements, to distinguish between the two manipulations.
- In the set of *PUSH* and *PULL* manipulations, location information are essential to the distinction

$$\begin{matrix} \left( \text{themeID}, \text{agentID} \mid \langle -1, \text{loc}_t, [] \rangle \mid \dots \mid \langle 1, \text{loc}_{t,\alpha}, \text{loc}_{a,\alpha} \rangle \mid \dots \mid \langle 0, \text{loc}_{t,\omega}, \text{loc}_{a,\omega} \rangle \mid \dots \right) \\ \text{and } \exists \delta \in \mathbb{R} \wedge \delta > 0 : \vec{d} \bullet [\text{loc}_{t,\omega} - [\text{loc}_{a,\omega} + \delta * \vec{d}]] = 0 \\ \left( \text{themeID}, \text{agentID} \mid \langle -1, \text{loc}_t, [] \rangle \mid \dots \mid \langle 1, \text{loc}_{t,\alpha}, \text{loc}_{a,\alpha} \rangle \mid \dots \mid \langle 0, \text{loc}_{t,\omega}, \text{loc}_{a,\omega} \rangle \mid \dots \right) \\ \text{and } \exists \delta \in \mathbb{R} \wedge \delta < 0 : \vec{d} \bullet [\text{loc}_{t,\omega} - [\text{loc}_{a,\omega} + \delta * \vec{d}]] = 0 \end{matrix}$$

Figure 7: Semantic representations of *push* and *pull*. Differentiated only by the requirement on the value of  $\delta$ .