Towards Semantic Grounding via Video Data: The Case of *Push* & *Pull*

Benjamin Burkhardt

CRC 991, University of Düsseldorf

Outline

- Development of grounded 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..."

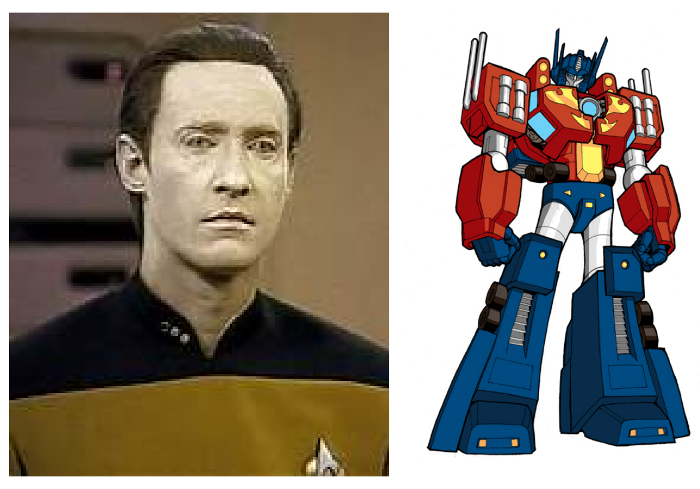

Figure 1: Imagination

Figure 2: Reality

- **The robot's owner:** Research Group at the Bernstein Center for Computational Neuroscience Göttingen (Project Leaders: Prof. Dr. Florentin Worgötter & 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

Video Capture and Analysis

- 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
- **Original goal:** Enable the robot to learn and recognize various manipulation types based on their prototypical visual properties.

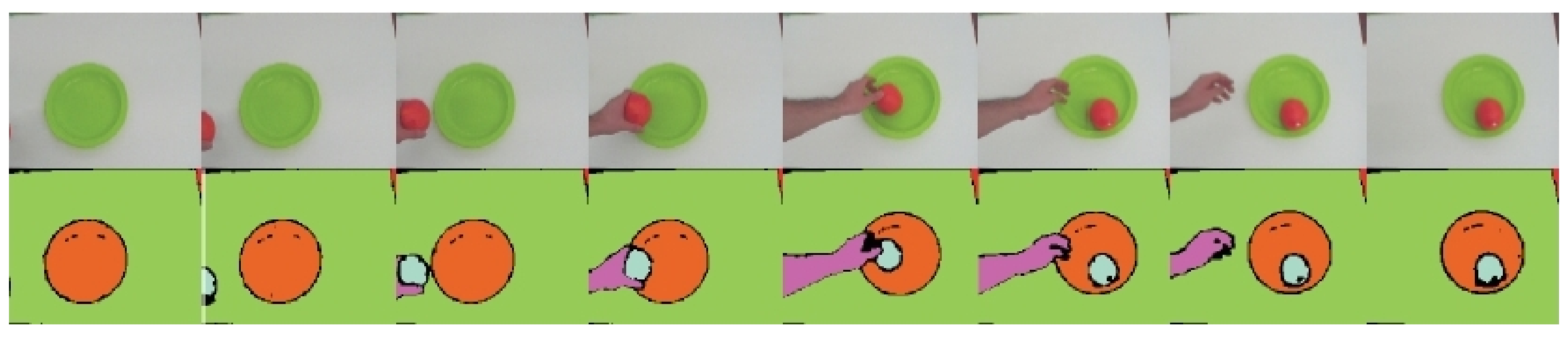

Figure 3: Object recognition & tracking

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
- **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.

$$\left(\begin{array}{cccc} \text{Object pairs} & \text{key frame 1} & \text{key frame 2} & \text{key frame 3} \\ bench, box & 1 & 1 & 1 \\ bench, hand & -1 & 0 & 0 \\ box, hand & -1 & 0 & 1 \end{array}\right)$$

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

$$M_{Push} = \begin{pmatrix} object\ 1, object\ 2 & -1 & 0 & 1 & 0 & -1 \end{pmatrix}$$

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.

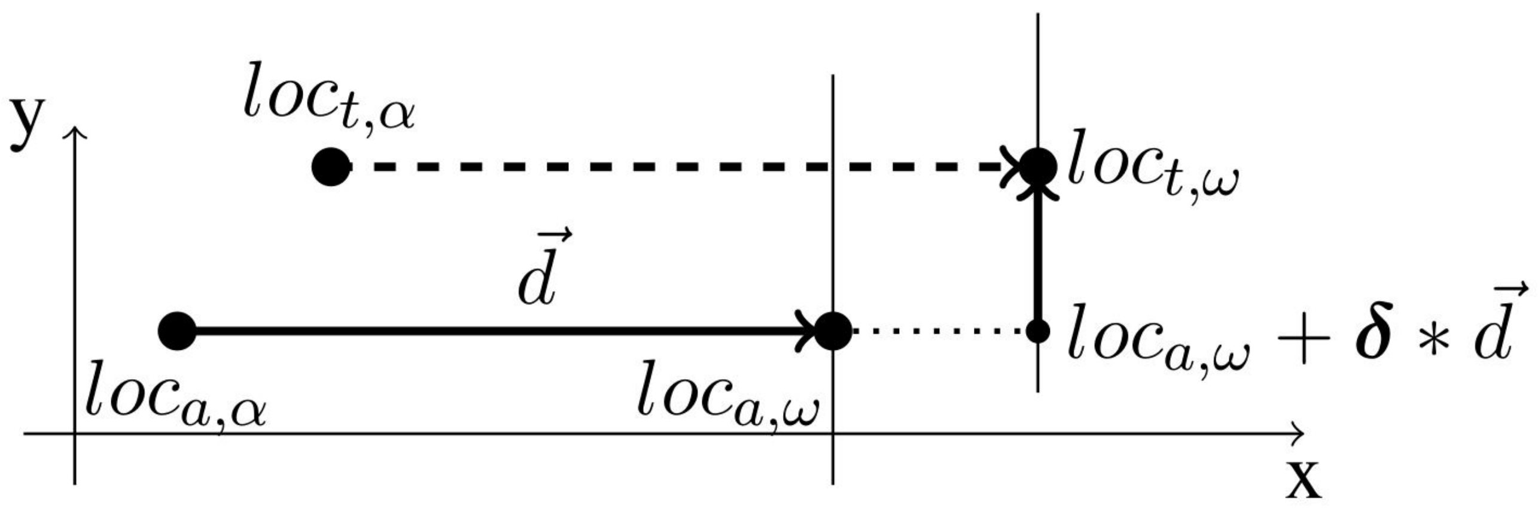

Figure 6: Push graph representation; agent and theme object travel along parallel paths.

- Starting from the final position of the agent, we either need to add or subtract some fraction of the length of the agent's movement path to get to theme's final position: $loc_{t,\omega} = loc_{a,\omega} + \boldsymbol{\delta} * \vec{d}$ with $\boldsymbol{\delta} \in \mathbf{R}$
- **Generalization:** If δ is positive, we can identify a manipulation as Push. If δ 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

push:

$$\begin{pmatrix} themeID, agentID \,\big|\, \langle -1, loc_t, []] \rangle \,\big|\, \ldots \,\big|\, \langle 1, loc_{t,\alpha}, loc_{a,\alpha} \rangle \,\big|\, \ldots \,\big|\, \langle 0, loc_{t,\omega}, loc_{a,\omega} \rangle \,\big|\, \ldots \\ \text{and } \exists \delta \in R \wedge \boldsymbol{\delta > 0} : \vec{d} \bullet [loc_{t,\omega} - [loc_{a,\omega} + \delta * \vec{d}]] = 0 \end{pmatrix}$$

pull:

$$\begin{pmatrix} themeID, agentID \,\big|\, \langle -1, loc_t, []] \rangle \,\big|\, \ldots \,\big|\, \langle 1, loc_{t,\alpha}, loc_{a,\alpha} \rangle \,\big|\, \ldots \,\big|\, \langle 0, loc_{t,\omega}, loc_{a,\omega} \rangle \,\big|\, \ldots \\ \text{and } \exists \delta \in R \wedge \boldsymbol{\delta < 0} : \vec{d} \bullet [loc_{t,\omega} - [loc_{a,\omega} + \delta * \vec{d}]] = 0 \end{pmatrix}$$

References: E. E. Aksoy, M. Tamosiunaite, and F. Wörgötter. 2015. Model-free incremental learning of the semantics of manipulation actions. Robotics and Autonomous Systems 71. 118–133 • E. E. Aksoy, A. Abramov, J. Dörr, N. Kejun, B. Dellen, and F. Wörgötter. 2011. Learning the semantics of object-action relations by observation. The International Journal of Robotics Re- search 30. 1229–1249 • K. Pauwels, N. Krüger, M. Lappe, F. Wörgötter, and M V. Hulle. 2010. A cortical architecture on parallel hardware for motion processing in real time. Journal of Vision 10(10). 1–21