Action and Language: Grounding Visual Cues with Textual Information

Humans are able to understand a very large variety of complex actions and can easily relate them with language. In our cognitive development, language tightly interacts with perception and action. Language can affect perceptual categorization since it provides a generic structure for bootstrapping, which is fundamental for imitation learning. In this project, we aim at bridging the gap between language and visual perception of observed human demonstrated actions. We can represent observed actions with our recently introduced "Semantic Event Chain" (SEC) [Aksoy et al., 2011] concept, which captures the underlying spatiotemporal structure of an action invariant to motion, velocity, and scene context. SECs are directly using the visual cues in the scene and can categorize objects according to their roles in the action. In this regard, our main intent here is to set up a direct link between structures provided by SECs and language. Hence, we can achieve, for the first time, a basic grounding of visual cues with the given textual information.

In this project, the candidate has to conduct various experiments on already existing human action datasets. First, a large corpus of textual description has to be created for these existing datasets. A parser has to be implemented to extract the subject, object, and action information embedded in the provided textual description. The last step is to propose a matching method that binds the extracted textual information with SECs (Fig. 1). Thus, without employing any recognition method, the cognitive agent can learn not only the actual meaning of the perceived objects in the scene, but also all possible actions that those objects can afford.

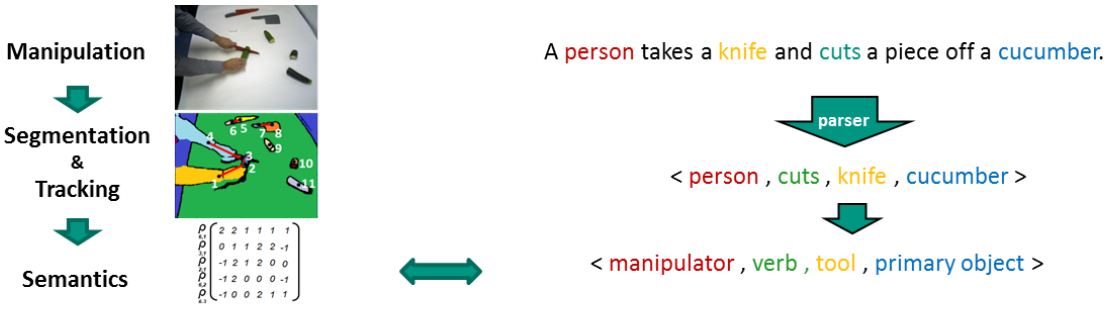
Manipulation
Segmentation
&
Tracking
Semantics
A person takes a knife and cuts a piece off a cucumber.
parser
< person , cuts , knife , cucumber >
< manipulator , verb , tool , primary object >

Figure 1: Action and Language: Grounding Visual Cues with Textual Information. On the left a snapshot of a human demonstrated cutting action is shown. Each object in the action is segmented and tracked to extract the respective semantic representation in a matrix format. The encoded semantics can be directly compared with the parsed textual data (shown on the right) from a given description regarding the same action. Thus, without employing any recognition method, the system can learn what each observed segment actually means and what actions each object can afford.

Aksoy et al., 2011: Aksoy E. E., Abramov A., Dörr J., Kejun N., Dellen B., and Wörgötter F. “Learning the semantics of object-action relations by observation”. The International Journal of Robotics Research, 2011.

Contact: Eren Erdal Aksoy (eren.aksoy@kit.edu) and Tamim Asfour (tamim.asfour@kit.edu)