Fatemeh Ziaeetabar

Postdoc Researcher at Georg-August-Universität Göttingen
Bernstein Center for Computational Neuroscience
Department for Computational Neuroscience
III Physikalisches Institut-Biophysik
Office E01.104
Friedrich-Hund Platz 1 37077, Göttingen Germany
Tel: +49 (0) 15253766570
E-mail: fziaeetabar@gwdg.de

A brief project description:

Automatic description of complex human actions from observation and/or from videos in the form of natural language remains a challenging problem especially when several actors are present. Methods for this may suffer from the high degree of diversity how any action can be performed and they are often subjective, depending on parameters adjusted by the observer. The goal of this study is to provide a generative framework for creating semantically different annotation-variants describing multi-agent actions in video test sequences. In our prior works we have introduced a framework – called “enriched semantic event chain (eSEC)” – that represents manipulation actions in an objective (observer independent) and semantically rich way. For this, we rely only on the changing static as well as dynamic spatial relations between the objects involved in a manipulation, including the actor’s hand. This framework was compatible with a context free grammar and, thus, allowed for generating action-describing sentences. Here, we will first extend our eSEC framework to full body actions. Next, we will create an ontology of fundamental utterances describing the different temporal phases of an action (e.g., a three-phase action: “The actor (1) picks up an object, (2) moves it, (3) puts it down”). Through a process based on statistical information from large text corpora, these utterances will be appropriately concatenated and supplied by alternative and more descriptive action verbs (e.g., same example: “…places an object”). eSECs provide us with static and dynamic spatial relational information, which can then be added (“…quickly places an object on another object”) and object information will then be provided by deep-learning-based object recognition (“…quickly places a cup on a

saucer”). Inter-agent interactions can be described using Region Connection Calculus (RCC-8) and partial sentences for each interacting agent will be linked by the temporal relations found with RCC-8. This framework, thus, can describe every action with different annotation variants (e.g., containing versus leaving out the static and/or dynamic aspects). Using a large set of video action chunks, we will create a set of such annotations for each chunk. We will validate our approach by quantifying the appropriateness of the automatically generated annotations using human rater-based assessments. This project will result in a novel algorithmic framework for multiple, semantically different possibilities for action annotation in video.