Combining classical and embodied multimodal fusion for human-robot interaction

Submission Type

Either oral or poster presentation

Abstract

Research in classical artificial intelligence (CAI) and embodied cognition (EC) has yielded many approaches to solve tasks of different kinds. Methods from CAI were for example successfully applied for playing chess, planning, and linguistic processing. However, CAI cannot solve tasks that require tight action-perception coupling, such as handing over an object or stable locomotion of a robot. EC is very good at solving some of these challenges, which led researchers like Rodney Brooks to the conclusion that EC might eventually replace CAI [1]. But is that really the case?

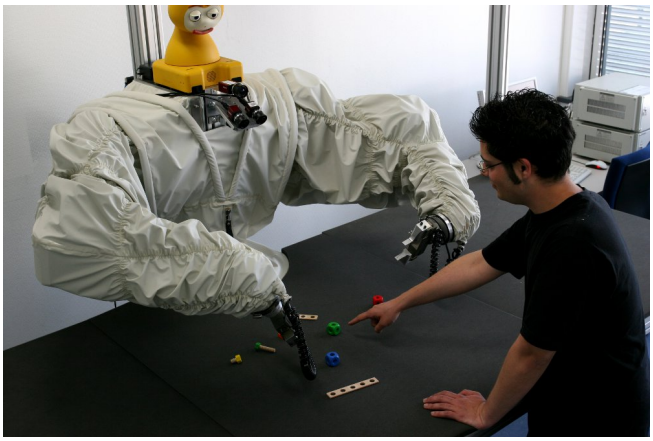

Fig. 1. The multimodal human-robot interaction system.

To investigate this question, we compared two approaches for multimodal fusion for a human-robot interaction system. Multimodal fusion is the process in which information from multiple modalities—for example speech, gesture, and object recognition—is combined to form a unified representation of the single channels. We implemented two multimodal fusion approaches for the robot that

we show in Figure 1: the *classical multimodal fusion* (CMF) processes input information sequentially. To fuse the information from several channels, it represents speech, gestures, and objects separately, and uses a rule engine to form hypotheses about the robot's environment. Here, the information from object recognition is used to ground symbols from the parsed spoken utterances by a human. In contrast to that, the *embodied multimodal fusion* (EMF) uses information from object recognition to form representations of the robot's actions that are coupled with the objects in its environment, which is inspired by Gibson's Affordances [2]. Based on these representations, EMF uses the information from speech and gesture recognition to calculate the relevance of the robot's actions in a given situation and selects the most relevant action for execution.

We used CMF and EMF in several experiments and found that both methods have their advantages and disadvantages: CMF behaves in a predictable way, which is the first precondition to implement a planning horizon in the robot's behaviour. However, CMF is not robust against input data that does not match the specifications in the rule engine. EMF processes input data more flexibly and continues to produce reasonable behaviour, even with faulty input. However, the robot's behaviour is not predictable, which makes this approach useless for contexts in which the robot needs to pursue a specified goal.

We conclude that the input processing of human-robot interaction systems needs methods from CAI as well as from EC. Especially in contexts in which the robot should interact with humans in a socially appropriate way, it could use EMF to robustly handle simple tasks like greeting humans or engaging in small-talk. When the robot needs to plan longer interactions, for example when it follows a joint plan with a human collaborator, it needs to be able to switch to a multimodal fusion component that is based on CMF.

References

1. R. Brooks. Elephants don't play chess. *Robotics and autonomous systems*, 6(1-2):3–15, 1990.
2. J. Gibson. *The ecological approach to visual perception.* Lawrence Erlbaum, 1986.