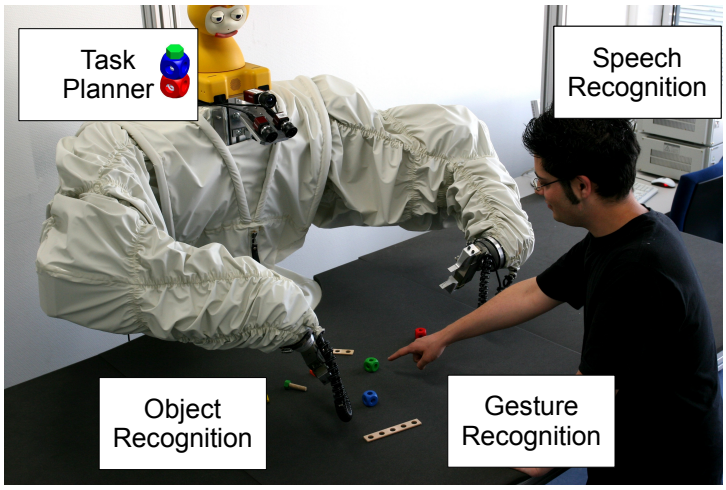
Combining Classical and Embodied Multimodal Fusion for Human-Robot Interaction

Manuel Giuliani
fortiss GmbH Munich
giuliani@fortiss.org

Jan de Ruiter
Universität Bielefeld
jan.deruiter@uni-bielefeld.de

Human-Robot Interaction System

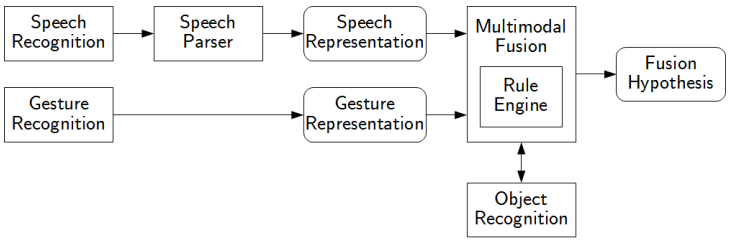


Classical artificial intelligence (CAI) and embodied cognition (EC) were successfully applied in different areas: CAI has its strength in fields such as planning or high-level cognition, which require a precise computation that involves logical inference; EC produced excellent approaches for sensori-motor coupling, which requires a robust and flexible computation. The question that we are following here is if the preciseness of CAI can be integrated with the robustness of EC?

We present two approaches for multimodal fusion on a human-robot interaction system: the classical multimodal fusion (CMF) uses methods from CAI to integrate information from a robot's input channels; the embodied multimodal fusion (EMF) is based on ideas from embodiment. Evaluations show that both approaches have their strengths and weaknesses, thus we propose a combination of both approaches to fuse their advantages.

For both implementations we used the human-robot interaction that we present in the picture on the left. Here, human and robot work together on a joint construction task, in which they have to assemble target objects from a wooden toy construction set. For that, the robot has modalities that provide information about human utterances, speech and gesture recognition, as well as modalities that provide knowledge about the robot's environment and given task, object recognition and task planner.

Classical Multimodal Fusion

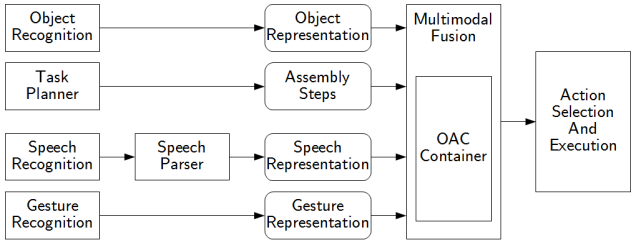


The classical multimodal fusion (CMF) is focused on the modalities that provide information about the utterances by the human, in our case speech and gesture recognition. CMF processes the information from the robot's modalities sequentially: the results from speech recognition are analysed by a speech parser, for example to extract deictic expressions. The information from gesture recognition is represented by a gesture type, a location at which the gesture took place, and the start and end time of the gesture. Both of these representations are fused with the help of a rule engine. During the fusion process, objects that the human talked about are grounded into information from object and gesture recognition. Finally, CMF generates a so-called fusion hypothesis that is then further processed by a frame-based dialogue manager. These hypotheses can be of one of four types: resolved, unresolved, ambiguous, and conflicting.

+ Long-term planning
+ Predictable
+ Modular

- Inflexible
- Error-prone
- Non-autonomous robot

Embodied Multimodal Fusion



The embodied multimodal fusion (EMF) is focused on modalities that provide information about objects in the robot's environment and objects that belong to assembly plans the robot need to follow, in our case object recognition and task planner. EMF processes the information of the robot's modalities in parallel: using the information from object recognition and task planner, EMF generates a set of so-called object-action complexes (OAC) [3]. OACs represent objects in combination with the actions that the robot can execute with these objects, a concept which is inspired by Gibson's Affordances [4]. EMF uses the information from all modalities to calculate a relevance score for each generated OAC. This score represents the relevance of an OAC in a given situation. Finally, EMF uses an action selection mechanism to decide which OAC should be executed by the robot in the next step.

+ Robust
+ Autonomous robot
+ Extensible

- Short-term interaction
- Inpredictable behaviour
- Individual implementation

Both presented approaches for multimodal fusion have their strengths and weaknesses. Thus, we argue that we have to combine aspects of both methods to yield a superior multimodal fusion approach. To reach this goal, we use the parallel processing and representation of information in the form of OACs from EMF, which guarantees a robust and error-tolerant input processing. From CMF, we adopt the planning of long-term goals and interaction patterns, to produce a controllable and predictable robot behaviour.

References

[1] Manuel Giuliani, Mary Ellen Foster, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. Situated reference in a hybrid human-robot interaction system. In Proceedings of the 6th International Natural Language Generation Conference (INLG 2010), Dublin, Ireland, July 2010.

[2] Manuel Giuliani and Alois Knoll. Evaluating supportive and instructive robot roles in human-robot interaction. In Proceedings of the International Conference on Social Robotics 2011 (ICSR 2011), Amsterdam, Netherlands, November 2011.

[3] N. Krüger, J. Piater, F. Wörgötter, C. Geib, R. Petrick, M. Steedman, A. Ude, T. Asfour, D. Kraft, D. Omrcen, et al. A Formal Definition of Object-Action Complexes and Examples at Different Levels of the Processing Hierarchy. PACO+ Technical Report, available from http://www.paco-plus.org, 2009.

[4] J. Gibson. The ecological approach to visual perception. Lawrence Erlbaum, 1986.

fortiss
Universität Bielefeld