Plan Recognition Using Multimodal Integration

Olga Vybornova (vybornova@gmail.com)
Kosta Gaitanis (gaitanis@tele.ucl.ac.be)
Benoit Macq (Benoit.Macq@UCLouvain.be)
UCL-TELE, Universite Catholique de Louvain,
Batiment Stevin, Place du Levant, 2, B-1348, Louvain-la-Neuve, Belgium

Keywords: multimodal interfaces; semantics; ontology; online probabilistic plan recognition; Dynamic Bayesian Networks.

General Overview and Method

We describe an ongoing research aimed at development of an anthropomorphic multimodal interface (meaning an interface for which the goal is not to precisely achieve a measurable task, like in instrumental interfaces, but more to provide a natural interaction) for recognition of a general intention, as well as of particular goals and plan fragments.

To recognize the communicative and overall practical intentions and goals of a person we use information arriving from natural language and gestures (actions). The result of the multimodal integration is a joint semantic meaning representation of the multimodal behavior and communicative intentions of the user. We use semantic frames (with slots to be filled, situational modeling) defining for every type of intention (goal) what can be said, with which gestures, actions emotions (using a multimodal vocabulary, adapted from Johnston & Bangalore, 2001)) for prediction and for recognition of this goal. The frame system provides a powerful mechanism for encoding information to support reasoning, action prediction, plan recognition and decision-making. The slots in frames are properties of the objects, actions of the objects and relationships of the objects with other frames. These relationships with other frames enable designing ontologies for particular contexts and domains. In our case we are developing an ontology of subjects, objects, activities and relations for a particular person while designing a multimodal interface helping the elderly and cognitive impaired people within their household environment.

Knowledge of past actions and future plans provide for successful communication. Thus plan recognition = knowledge of history + knowledge of goals (see Gorniak & Roy, 2005). In this respect we track the behavior of presuppositions in discourse, since presupposed information builds the semantic basis of discourse, its "history", enables to establish common ground (common knowledge) between the discourse participants. This presupposed information is resolved and accumulated within the framework of the Discourse Representation Theory (DRT). (Blackburn & Bos).

Movements of the crucial points of the human body (hands, feet, head and center of gravity) are analyzed in order to detect actions such as walking, bending down or displacing an object. These points are considered as cooperative agents forming a cooperative team (the whole body). A novel robust framework for online probabilistic plan recognition in cooperative multiagent systems (the MultiAgent Abstract Hidden Markov mEmory Model (M-AHMEM) is used to model the human body and detect the actions performed. (Gaitanis, Correa and Macq, 2006).

The action detection problem is phrased as an inference on the underlying Dynamic Bayesian Networks representing the process of executing the actor's plan. DBNs are employed to track the paths to goals (using hierarchical plan fragments) and modifications (strategy corrections, adaptations) of the initial goals.

Taking into account both the physical situation and the communicative intentions of a person in a general framework will provide ultimate understanding and goal recognition.

Acknowledgments

This work is supported by the European FP6 Network of Excellence SIMILAR (http://www.similar.cc).

References

Gorniak P. and Roy D. (2005) Probabilistic grounding of situated speech using plan recognition and reference resolution, *Proceedings of the International Conference for Multimodal Interfaces.*

Blackburn P. and Bos J. Working with Discourse Representation Theory. *An Advanced Course in Computational Semantics.* // Current draft at http://www.cogsci.ed.ac.uk/~jbos/comsem/book2.html.

Johnston M. and Bangalore S. Finite-state methods for multimodal parsing and integration, *in ESSLLI Workshop on Finite-state Methods*, Helsinki, Finland, 2001.

Gaitanis K., Correa P. and Macq B. (2006) Modelization of limb coordination for human action detection, submitted for publication in *Proceedings of the IEEE International Conference on Image Processing*.

Vybornova O. and Macq B. (2005) Towards multimodal treatment of presuppositions in natural dialog discourse, *Proceedings of the 10th International Conference on Speech and Computer (SPECOM'05),* Patras, Greece.