FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D1.2.1

Formalism of the modality integration scheme

Due date of deliverable: 31/12/2004
Actual submission date: 31/01/2005

Start date of project: January 1st, 2004 **Duration:** 48 months

Organisation name of lead contractor for this deliverable:
Bielefeld University

Revision: Final
Dissemination Level: PU

Executive Summary

The objective of WP 1.2 in the first phase is to develop a data association model that establishes correspondences between interpretation results of spoken language utterances, deictic gestures, and basic visual features such as e.g. shape or colour in order to resolve multi-modal object references in the real environment.
In the first three months of this work package (which has started at T0+9) we have specified a data association model in close cooperation with RA 2. The main focus of the model is to resolve verbal object references that are accompanied by pointing gestures. According to this model, the dialogue module will provide verbal cues for pointing gestures which will then trigger the gesture recognition module. The results from the gesture recognition in terms of hand coordinates will significantly restrict the search area for the subsequent object detection or object recognition module. In certain cases additionally given verbal information about the objects, such as colour or shape, will be used to further restrict the search area and to find the intended object. All the information that have been collected by the different modalities will be stored in a single multi-modal representation of the object. The definition of the multi-modal representation scheme is currently work in progress and a first version will be implemented till the end of phase 1 (M 18).

Role of modality integration in Cogniron

In order to enable embodied communication on a robot system it is necessary to merge information from different modalities. This is because in situated communication humans tend to make heavy use of non-verbal communicative cues such as gestures or mimic without which a verbal utterance can not be understood. In the first phase of the project we have focused on developing a model for the multi-modal integration to associate information about objects from different modalities. This requires a close collaboration with RA 2 where gesture recognition is performed. In the second phase of the project we will start to also integrate spatial information such as topological maps which are provided by RA 5. Spatial information is not only important for navigation of the robot but also for communication with the user. It will therefore be necessary to define a multi-modal representation scheme that can be used by different modules such as, e.g., the navigation or the dialogue module. Therefore, the modality integration is an important part of the COGNIRON project because it will bring together results from different components.

Relation to the Key Experiments

Similar as WP 1.1 (Declarative dialogue model), WP 1.2 will mainly be concerned with KE 1. We will focus on those aspects of the Robot Home-Tour where the user interacts with the robot about places and objects. This includes situations where the robot is asked to build a map of the layout of the home or where it is taught the names of objects. These tasks do heavily rely on the usage of multi-modal information and are therefore the main focus of WP 1.2.

1 Integration of speech, deictic gestures, and basic visual features

During the development of the data association model a focus has been put on the specification of the interface between the dialogue component and the vision-based modules. Therfore, a close cooperation with RA 2 has been initiated where the CF-ROR (Resolving Object References) will be developed in phase 2. This cooperation is necessary since the data association model has to establish correspondences between speech, deictic gestures and salient visual features. It, therefore, needs to provide a mechanism that is able to synchronise input from different modalities and to explicitly search for input in one modality when a trigger is given in the other one. Hence, the processing in the dialogue module and in the other modalities has to be closely coordinated.

As shown in Fig. 1 the core of the modality integration is a finite state machine (FSM) that enables a smooth coordination of the different hardware components during the integration process. It provides a mechanism that can handle both known and unknown objects.
In order to determine whether an observed object in the scene belongs to an already known object type, an XML-based database [3] is queried. This database is called “scene model” and is the long-term memory of the system. It stores multi-modal information of objects such as their visual appearance and the features provided by verbal input, e.g., their names. The exact definition of this object representation as a basis for verbal interaction will be a major research issue in WP 1.2 in the next phase of the project. In the following we will discuss in more detail the mechanism of the model.

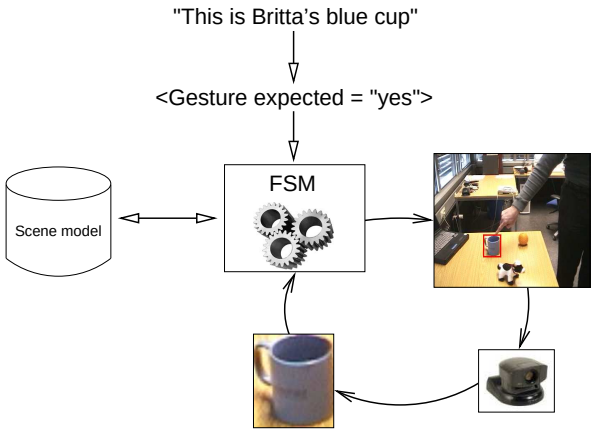

Figure 1: Processing chain of our FSM-based data association model for a given verbal input information

1.1 Preprocessing

The modality integration module is needed for utterances that verbally describe objects and that contain hints that other modalities might require for the analysis of the propositional content. For example, in the statement “This is Jane’s blue cup” the demonstrative pronoun “this” is a hint for a potential gesture. The perception-based dialog as described in the deliverable WP 1.1 will process this input and forward it to the modality integration module to resolve the meaning of the word “this”. Upon this request from the dialogue module the FSM (see Fig. 2) will then switch from its idle state *Object Alertness* (ObjAlert) to the *Input Analysis* (IA) state. In this state the XML-encoded message from

the dialogue module containing a frame-based semantic representation of the verbally specified object will be parsed. Fig. 3 shows an example of such a message for the utterance “This is Jane’s blue cup”. Because the word “this” indicates a possible involvement of a deictic gesture the modality integration module activates the gesture recognition module to search for it. It also checks the scene model to find out if the object type “cup” is already known to the system. Depending on the result of this search two different processes can follow.

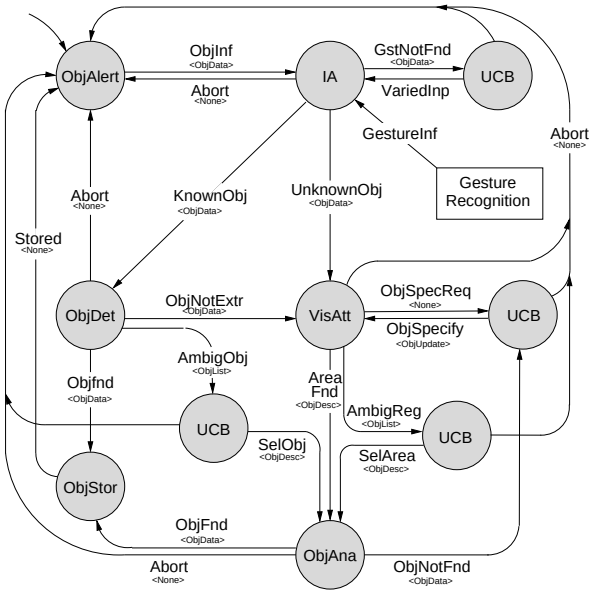

Figure 2: Overview of the finite state machine

1.2 Handling of known object types

In case that the object is an instance of a known object type the FSM switches to the *Object Detection* (ObjDet) state in order to search for a known object. For this searching procedure the hand coordinates from the gesture recognition module provide helpful information in order to confine the search area for the object detection.

For the object detection itself a template-based object recognition concept is adopted that makes use of the perception-based semantic approach which will be the basis of our object representation scheme. It uses a fast normalised cross-correlation algorithm as described in [2] and tries to match image patterns of the known object type retrieved from the scene model with the current camera image or the relevant area of it. Note that this component is used only during the development phase of the modality integration module and will be replaced in the course of the project by more sophisticated object recognition algorithms being developed in RA 5.

If the object is found a message will be sent to the dialogue system which will then decide on further interaction steps according to the dialogue context and history. If ambiguities occur or no object is found in spite of additional searching (as described in the following section) the dialogue system will receive a corresponding message and can initiate clarifying questions or make suggestions to the user to improve the interaction success rate.

```
< OBJECT type = "MOVABLE"
          name = "cup" >
  < GESTURE_EXPECTED value = "yes" />
  < ATTRIBUTE name = "color"
              value = "blue" />
  < OWNER value = "Jane" />
</ OBJECT >
```

Figure 3: Example of XML-encoded frame-based semantic representation

1.3 Handling of unknown object instances

Learning new objects is especially interesting for the KE 1 Robot Home-Tour. If the user references an unknown object, the FSM of the modality integration module will receive a message from the scene model that the object was not found. It will then switch from the IA state to the *Visual Attention* state (VisAtt). In this state the system will make use of the additionally given verbal information such as the colour. This is done by using different filters which are similar to the "attention maps" proposed in [1] to extract salient image features from the camera image as, e.g., distinctive colour information provided verbally by the user.
If the user points to the object and the gesture recognition module can help to confine the search area, the probability of a successful search will increase considerably. In this case the modality integration module can set a bounding box close to the hand position and search efficiently for the specified colour. If a new object is found its view is stored in the scene model along with other object attributes provided by the user's verbal input, e.g., the owner of the object. The search result will then be sent to the dialogue system which can generate appropriate feedback to the user. In case of a search failure the dialogue can ask the user to point to the object.
As already mentioned in the previous section, this mechanism of handling unknown object instances is also activated after a search failure of a known object type.

2 Future Work

In the next phase this model will be implemented in close cooperation with RA 2. A pre-requisite for this is the specification of the multi-modal representation scheme of the objects. Therefore, most of the work in this work package in the second phase will be dedicated to specifying and implementing an object representation scheme that allows the representation of multi-modal information that is suitable for such different functionalities as object referencing, navigation, or dialogue management. This will require close collaboration with the research areas RA 2 and RA 5.
Also, the functionality of the modality integration framework will have to be extended to other modalities. It is planned to incorporate topological maps into the representation scheme in order to allow for communication of locations but also to use them for navigation of the robot. The adequate activation of these functionalities will also have to be supported by the dialogue module and is being implemented in close collaboration with WP 7.2 (KE 1 Robot Home-Tour).

3 References

3.1 Applicable documents

3.2 Reference documents

[1] Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 20(11):1254–1259, 1998.

[2] J.P. Lewis. Fast template matching. In *Proc. Conf. on Vision Interface*, pages 120–123, Quebec, Canada, 1995.

[3] S. Wrede, M. Hanheide, C. Bauckhage, and G. Sagerer. An Active Memory as a Model for Information Fusion. In *Proc. 7th Int. Conf. on Information Fusion*, number 1, pages 198–205, 2004.

Annexes