FP6-IST-002020

# COGNIRON

*The Cognitive Robot Companion*

Integrated Project

Information Society Technologies Priority

# D3.4.1
# Report on methods for recognising user intent

**Due date of deliverable**: December 31st
**Actual submission date**: January 15th

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

**Organisation name of lead contractor for this deliverable**:
Fraunhofer IPA

**Revision**: final
**Dissemination level**: PU

# Executive Summary

The overall objective of this work package is to recognise user intent. During the first period a conceptual framework was developed and a first disambiguation method implemented as prototypical software. Human data was extracted from medical publications. High-end sensors (depth and colour imaging sensors) were combined to achieve a maximally reliable and robust detection of human movements which in turn were used for the disambiguation.

In its first section this report describes the theoretical work carried out during the first project phase. Communication through gestures will be described as a type of transmission of intent. Then classification of gestures as part of human-robot interaction is discussed followed by a description of the interface and the human data used. The part closes with a note on the accuracy.

The second section describes the software developed so far. A visualisation tool for debugging, development and evaluation was written, able to use arbitrary data delivered from gesture recognition software. Secondly, a first attempt for disambiguating gestures using sensors and internal information was made. We will use this as a basis for future integration of additional data, delivered by external components.

Originally it was planned to give an exhaustive comparison of disambiguation methods with this report. However, IPA started a new research here and first basic concepts needed to be considered. As shown in the report there are many hints to the user's intent not all of which are detectable by current state-of-the-art sensors. Finding a good mapping between sensors and cues is therefore a crucial requirement for the successful completion of this work package. This turned out to be more time consuming than expected and the evaluation of different methods needed to be shifted to the next months. Hence, the report's title was rephrased leaving out the term experimental evaluation.

# Role of recognising user intent in Cogniron

In order for the robot to engage in social interaction, it needs to understand whether gestures are meant for communication or other purposes. A reliable scheme for disambiguation is identified as a basic competence of a cognitive robot companion.

# Relation to the Key Experiments

The main Cogniron Function this work contributes to is CF-ACT (Detection and interpretation of human activities and postures), in close relation to CF-PTA (Tracking of entire human body and detection of person's attention), CF-TBP (Tracking of human body parts for observation) and CF-GR (3D gesture recognition). This work relies on those skills. We also try to develop functionality based on achievements in RA2 by IPA. We plan to use our work in KE3, due to the sensor information and human intention detection needed in this experiment.

# 1 Recognition of User Intent

## 1.1 Theoretical Background

### 1.1.1 Transmitting intent – communication

To detect a user's intent first of all some kind of communication between the robot and human has to take place. Generally communication can be divided into three categories [1] : verbal communication, non-verbal communication and non-verbal communication in a broader sense (e.g. physical characteristics). While verbal communication is not a part of this work package, information delivered by speech recognition software will certainly be important data used for disambiguation of gestures in a later stage of development. Similar, paralinguistic (vocal non-verbal communication) like pitch of speak, laughter and pausing may give useful hints for decoding the user's intention.Currently our focus is on the non-vocal and non-verbal communication, i.e. body language, which makes up two thirds of communication [1]. Other studies even speak of 92 %.

Again body language is divided into several sub categories, e.g.:

- tactile communication,
- gaze,
- body posture (standing, sitting, crouching, kneeling, lying),
- translation (20 basic movements, e.g. crawl, slurp, walk, run, sprint [2]),
- proxemic space,
- and gestures.

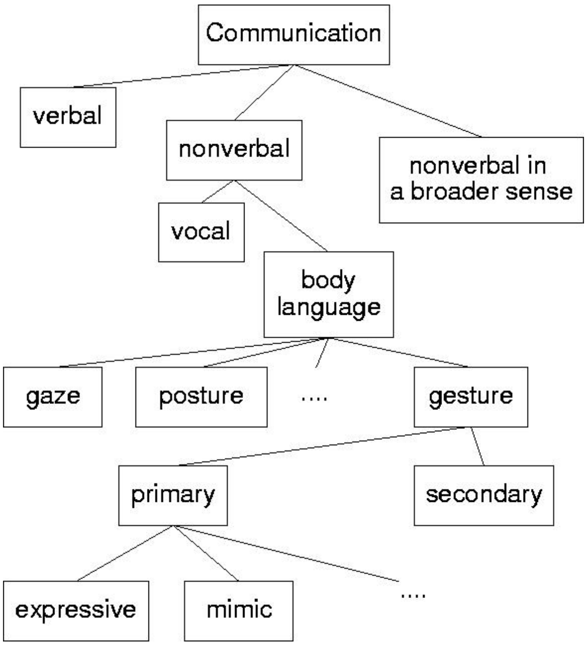

*Fig 1 Communication Tree*

Fig. 1 shows roughly how the terms stated above connect. Since we want to detect human's intent from gestures performed, we have analysed and categorized gestures, taking into consideration the robot's tasks, data currently delivered by a gesture recognition software and reliability in every day human-robot interaction.

### 1.1.2 Classifying gestures for use in human-robot-interaction

Incidental or secondary gestures like scratching, coughing or eating give the observer a lot of information on current mood and condition. These on the other hand are very difficult to detect by a robot and to interpret with today's gesture recognition software. We have therefore not considered these gestures currently.

Primary gestures are again dividable into several sub categories, from which only expressive gestures are not solely human.
Expressive gestures are common for all humans and animals, especially primates. A major part is facial expressions. Gestures imitating something or someone are considered to by mimic gestures. Of all mimic gestures we consider the “no-load” mimicry an important part of human-robot-interaction. These gestures perform actions without the associated object, e.g. to pretend to hold a glass and drink. Considering key experiment 3 “Learning Skill and Tasks”, where learning goals from observation is required, these kind of gestures need to be detected and interpreted.
Schematic gestures are standardized abbreviated mimic gestures and represent objects as opposed to symbolic gestures, indicating abstract expressions, e.g. moods and thoughts. These kind of gestures can vary greatly between cultures and are often misunderstood inter cultural. So called technical gestures are probably the most easily detectable gestures. They are defined and used by specialists in special situations. These could be a set of gestures for controlling a robot's movement. Closely related are coded gestures like sign languages.

While the above is a more general classification of gestures, we have defined types of gestures from the robot's point of view, similar as proposed by Nehaniv [3].

**Irrelevant gestures**
Since “one cannot not communicate” as stated by Watzlawick [6] the robot needs to distinguish between gestures that are relevant and irrelevant for it. Gestures, which are currently considered irrelevant are e.g. people communicating without the robot, moving of arms while walking, people being outside a certain distance and gestures which are not recognized by the gesture recognition software. The amount of irrelevant gestures in an ideal system should be close to zero, but will be very high in this early stage of development.

**Transitional gestures**
These are gestures showing the willingness of an user to interact, e.g. waving. These gestures will cause the robot to focus on a person. These gestures will also contain secondary gestures in a later stage of the development. A person playing with an empty coffee cup might mean the user wants more coffee and should therefore trigger a robot action.

**Symbolic gestures**
Symbolic gestures are gestures an underlying gesture recognition software system recognizes and are be contained in a dictionary within the software. Note that any necessary disambiguation will not be done by the recognition software, but by methods developed in this work package using additional information. The information we receive from the gesture recognition has a certain degree of arbitrariness depending on the gesture detected. Fig.2 e.g. could mean “pass me the object”, an offer to shake hands or a deictic gesture.

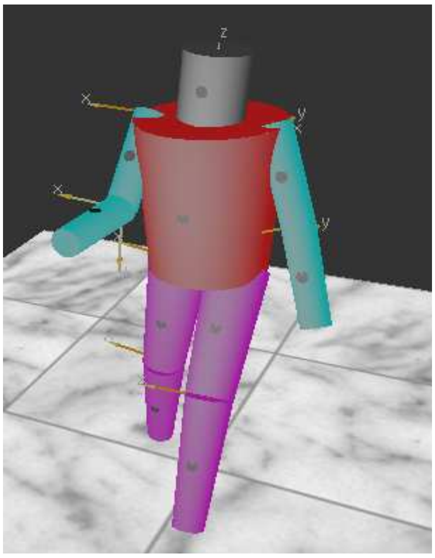

*Fig 2 Visualisation Tool*

### 1.1.3 Interface to gesture recognition software

Since WP 3.4 relies on the results of RA2 we have defined terms for the use in Cogniron. The terms defined are depicted in fig 3.
Since the visual capture of a human will result in discrete snapshots this basic block was defined as a "motion primitive". A "posture" or static gesture can then be identified, if no or only little change between consecutive motion primitives is detected (e.g. deictic gesture). (Dynamic) "gestures" are made up of a sequence of motion primitives (e.g. waving). Postures and gestures need to be interpreted as communication. In a more abstract level we then will speak of "activities" made up of a sequence of "postures" and "gestures". For data communication and compatibility a data structure and exchange format in XML were defined in cooperation with Cogniron partners from RA2.

Furthermore, we have implemented a visualisation tool for data delivered by a gesture recognition software for debugging and development purposes. It is based on OpenGL and binaries are available for Windows and Linux systems. Poses as well as dynamic gestures can be displayed.
The model is expandable depending on the amount of information available by the gesture detection software. The current minimal information agreed upon is as shown in fig. 2. The body is made up of ten limbs with one to three degree of freedom joints combining them. If e.g. a gesture recognition software gives information on hand position and orientation, this can also be simply displayed by defining the hand as a further limb attached to the lower arm with a certain joint.

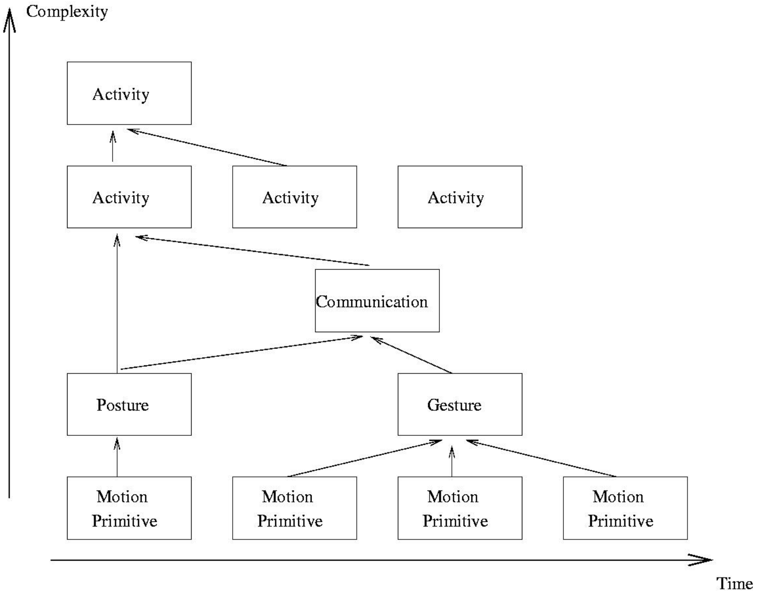

*Fig 3 Visualisation Tool*

What still needs to be defined in cooperation with RA2 is a common language for the positions of the limbs a gesture recognition software detects. The gesture recognition software needs to filter the detected positions to the relevant information and pass this to the disambiguation routines. Above position should be described in a technical way e.g. “right hand in front of body with bent arm”.

### 1.1.4 Data for disambiguation

For disambiguating gestures we need additional information that is not delivered by the gesture recognition software. Depending on the type of gesture recognition we might be able to extract information at a lower lever and use it for uncovering human's intentions:

- Proxemic space: Depending on how far a human is from the robot we can draw conclusions on the user's intent and attitude. Here we will rely on input delivered by other Cogniron partners working on personal distances. A basic rule is given by the following table: Intimate distance: approx. 1 arm length, social distance: 1- 3 mnormal communication distance, public distance: > 3 m.
- Orientation: How a person or persons are orientated in reference to the robot can gives us valuable information whether the robot is included in the communication.
- Foot orientation: Depending on the angle of a person's feet the robot is part of communication or not.
- Head pose: The user's intention might be measured by the tilt of his head, showing a degree of attention.
- Gaze detection and direction: By knowing what a user is looking at, the robot might be able to differentiate between gestures.
- Sound and speech detection and localisation: As mentioned above this is only a small part of communication but essential for gesture disambiguation.
- Tracking and recognition of objects: If objects addressed by deictic gestures are recognized, the recognition will be more robust.
- Information on the user: Information like face identification, colour of clothing, size, etc. will make gesture recognition more reliably and robust.

- Phases of gestures: All gestures have a preparatory phase, an actual gesture phase and a retraction phase. By having knowledge of these as well as of the actual gesture misunderstandings will be reduced. For example, a waving gesture requires first of all raising the hand above the head, the actual waving gesture and then the return of the hand, usually to its original position. The preparatory and retraction phases will be different from those of a person dusting some shelves using a similar motion in the actual gesture phase.
- Speed: The speed of a gesture performed gives information for disambiguation (attitude, urgency, etc.)
- History: By storing information on the interaction taken place between a certain user and the robot, some gestures detected will be more likely than others. This might be due to a user's individual preferences or even certain rituals interactions follow (e.g. shaking hands, when meeting the first time). Also information on objects and their changes over time will deliver valuable information.

The more of the above information we have available in our algorithms the more reliable our predictions on user intent will be.

### 1.1.5 Accuracy

Considering the sometimes severe consequences [4] occurring in inter-cultural human communication by misinterpreted gestures, we can be sure that we will not be able to achieve a 100 percent accuracy in detecting human intention, even concentrating solely on one culture group. Humans interpreting gestures from pictures alone usually have difficulties, if they have no further background information and knowledge of the person. Body language is never non-ambiguous.

## 1.2 Software Development

### 1.2.1 Visualisation Tool

As already mentioned in 1.1.1.3 we have developed a visualisation tool for debugging purposes and experimental set ups. By showing poses to people with this tool we can collect possible meanings of gestures with no other information that might be communicated subconsciously. Once we have a dictionary of gesture recognized by a gesture recognition software we can show these gestures to test candidates and collect possible interpretations. By adding additional information to each picture shown, we can then extract the necessary data for disambiguation.

### 1.2.2 Disambiguation experiments

For first experiments and proof of concept, we have set up a 3D sensor [5] to try disambiguating gestures. Since we don't want to recognize the actual gesture but its meaning, we have chosen two gestures that are simple to detect but allow for several different meanings. The gestures are shown in fig. 4 (“waving” and “bent arm”) as depth picture and 2D image.
We try to derive their meaning solely on additional information extracted from the 3D-Sensor and internal state of our imaginary robot.

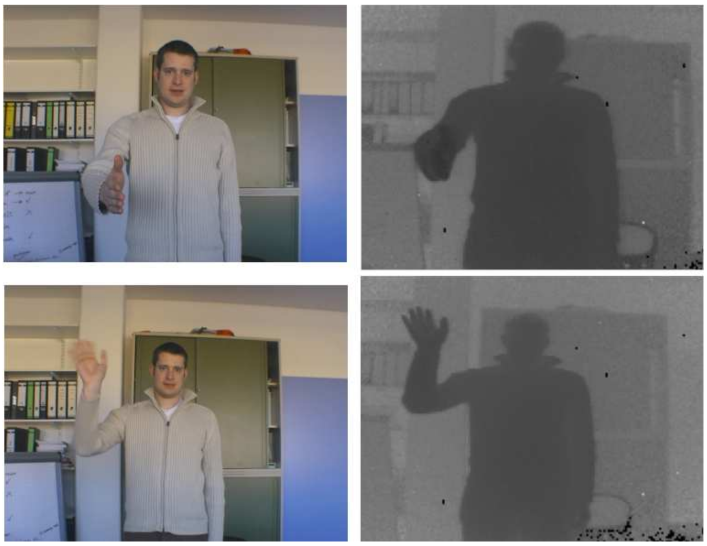

*Fig 4 Gestures in depth picture and 2D image*

The following information is extracted from the 3D-Sensor:

- Gesture is static or dynamic.
- Orientation of the human towards the robot and the distance.
- Hand is above or below the shoulder.
- Imaginary line defined by the centre of the body and hand.
- Factor for the size of the hand.

Also information on internal state is calculated:

- Robot has object.
- Person has not been greeted yet.

Fig. 5 shows the decision tree the robot takes to determine a person's intent. The functionality is limited but we will extend the functionality gradually by including more meanings and finer gesture recognition. Depending on our robot's abilities, we will create a dictionary on what a user can communicate and what the gestures will look like. The intentions detected above are based on the idea of a robot constructed for fetch-and-carry tasks, on which we will conduct our future experiments.

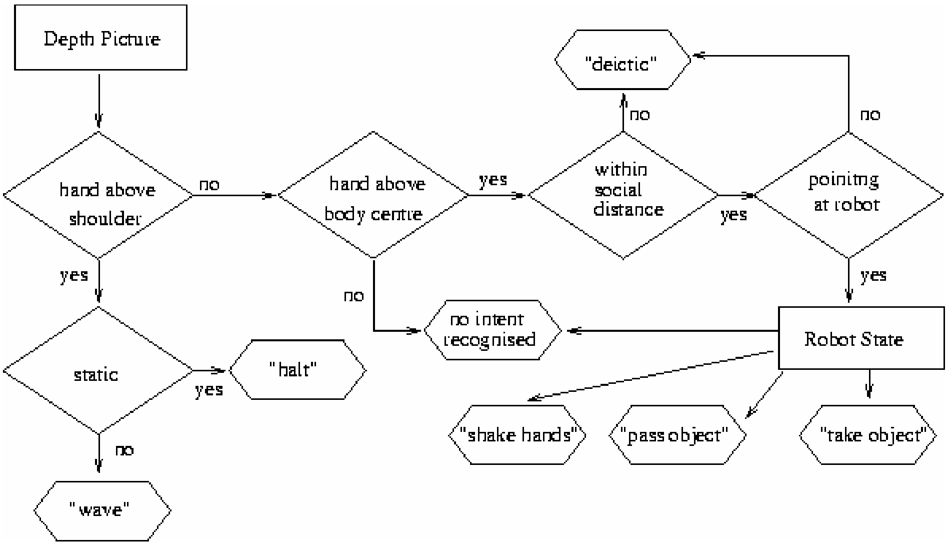

*Fig 5 Decision Tree*

Roughly, the 3D Sensor information is extracted with the help of a reduced depth histogram and blob detection. Starting from the centre of the largest blob after the person has been extracted from the depth picture; the head is determined by searching upwards. Searching below the head and slightly to the left and right, the approximate position of the shoulders can be found. Staying within the corridor defined by the arms, finally the hands can be found and their pixel size. This returns a size factor, if divided by the distance. Since the 3D sensor gives us information on the actual position of every pixel in space, the pointing direction of the human is easily determined. Another advantage of the sensor used is, that since it is an active device sending out light pulses, the detection is independent from environmental lighting. Fig. 6 shows the result of our algorithm.

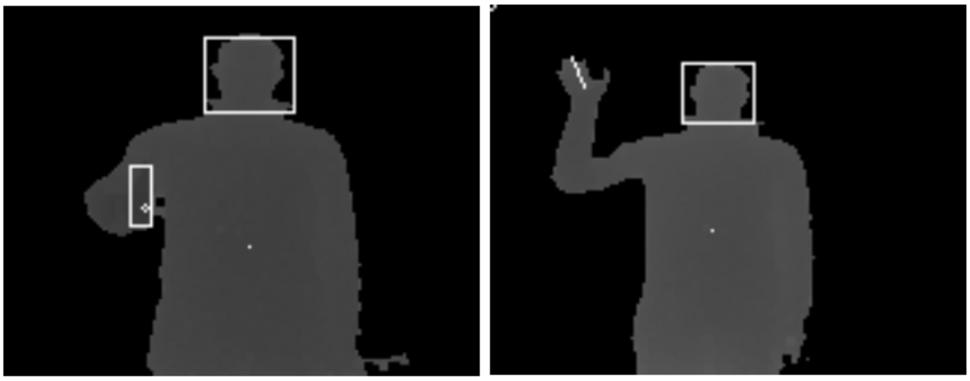

*Fig 6: Gesture detection*

Head, centre of body and right hand are marked. This data is used for identifying the gesture. Currently we are evaluating the robustness and reliability of our algorithm.

# 2 Future work

In the next period, we will set up a robotic platform with these sensors and software. A first scenario will be implemented. A person will attract the robot's attention by waving. The robot will turn towards the person and communicate its willingness to interact. Another gesture i.e. command will make the robot follow the person. The person will then be able to stop the robot and name objects and locations by pointing at them. The robot will store images of these for later processing.

Based on this set up, the interpretation of the recorded movements will become more and more sophisticated to achieve a more context based interpretation of commands. Possible key points could be the integration of results from other WPs (e.g. gaze detection and personal spaces), extraction of human features for classification and identification (e.g. child – adult and guest - teacher), analysis of body pose (attitude and personal spaces), recognition of similar situations and interpretation of user's intent by observation (e.g. playing with an empty coffee cup means somebody wants more coffee.)

We believe that, by combining multi modal input with observation over time, we can extract reliable information on human intention. The input will not only consist of gesture recognition, but also of speech recognition, current state and tasks of the robot, state of the environment, person identification, history on past interaction and so forth. What separates the work in this package from other approaches proposed is the fact that not only cues from different modalities are taken into account. Rather, sequential patterns of these cues and contextual knowledge will be used to disambiguate the robot's detection of human intention. If for instance the intention is to get into a dialogue with the robot, there might be typical pattern of activity before. The robot can learn from this and next time it will turn to the human in advance.

# 3 References

[1] Delhees, K. (1994) *Soziale Kommunikation. Psychologische Grundlage für das Miteinander in der modernen Gesellschaft:* Westdeutscher Verlag
[2] Morris, Desmond (1977): *Manwatching:* Jonathan Cape, London
[3] Nehaniv, Chrystopher (2004): *Classifying Types of Gestures and Inferring Intent:* Hertfordshire CS Tech Report 419
[4] Axtel, Roger (1998): *Gestures: The do's and taboo of body language around the world – revised and expanded:* Wiley, New York
[5] Swiss Ranger: www.csem.ch
[6] J.H. Beavin, D. D. Jackson, P. Watzlawick (1967): *Menschliche Kommunikation*