Real-time Framework for Multimodal Human-Robot Interaction

Jürgen Gast, Alexander Bannat, Tobias Rehrl, Frank Wallhoff, Gerhard Rigoll
Institute for Human-Machine Communication
Department of Electrical Engineering and Information Technologies
Technische Universität München
Munich, Germany

Cornelia Wendt, Sabrina Schmidt, Michael Popp, Berthold Färber
Institut für Arbeitswissenschaften
Fakultät für Luft- und Raumfahrttechnik
Universität der Bundeswehr München
Munich, Germany

***Abstract*— This paper presents a new framework for multimodal data processing in real-time. This framework comprises modules for different input and output signals and was designed for human-human or human-robot interaction scenarios. Single modules for the recording of selected channels like speech, gestures or mimics can be combined with different output options (i.e. robot reactions) in a highly flexible manner. Depending on the included modules, online as well as offline data processing is possible. This framework was used to analyze human-human interaction to gain insights on important factors and their dynamics. Recorded data comprises speech, facial expressions, gestures and physiological data. This naturally produced data was annotated and labeled in order to train recognition modules which will be integrated into the existing framework. The overall aim is to create a system that is able to recognize and react to those parameters that humans take into account during interaction. In this paper, the technical implementation and application in a human-human and a human-robot interaction scenario is presented.**[1]

I. Introduction

The general motivation of the presented framework is to collect data from human-human experiments to train a working model. The same framework will then be utilized for human-robot scenarios. The prior trained modules shall react to the actions of the human and control the robot accordingly.
In order for the interaction process to be perceived as naturally as possible, multiple modalities have to be exploited, i.e. several video sources, audio signals, physiological data, range sensors, haptics, eyetracking, etc. All these sensors have different sampling rates making it necessary to synchronize the data for further processing.
Although in the literature several middleware architectures have already been introduced to deal with multimodal data streams, most of them are suffering from transparency or from real-time capabilities, e.g. [1].

Especially distributed systems with a high overhead of network data exchange are not suited to act as a multimodal recording-tool appropriately. Therefore, we propose to use the Real-time Database (RTDB) as sensory buffer with a high data bandwidth applicable for on-line as well as off-line experiments. The underlying RTDB architecture has established itself as a reliable platform in conjunction with Cognitive Vehicles [2], [3], [4]. Furthermore, it has served as a convincing integration platform, where several groups of researchers simultaneously work on the same framework.

The rest of this paper is organized as follows: In Section II we introduce the integrated system framework basing on the RTDB with its interface modules in greater detail. In Section III-A we present some conducted human-human experiments, followed by a setup for on-line human-robot scenarios in Section III-B. In Section IV the transcription process is described in more detail. The paper closes with a summary and an outlook over the next planned steps.

II. System Architecture

In this section a short overview of the System Architecture and the modules used for both scenarios – human-human interaction, human-robot interaction – will be given. The system architecture consists of the RTDB as a sensory buffer and communication backbone between the input and output modules.

The RTDB buffers and synchronizes the input modules that generate data for the output modules. Further, an output module has been created that collects all input information in one compressed AVI-file for the transcription process, as explained in more detail later. The developed input modules will be explained after a short introduction to the RTDB.

The RTDB presented in [5] is able to deal with large amounts of sensor data (in our setup 47 Megabytes per second for the human-human experiments) and can provide data exchange and recording in real-time on a Linux PC equipped with an AMD Phenom 2.2GHz quad-core and four gigabyte RAM. In cognitive autonomous vehicles the database is used to manage all sensor inputs to keep the vehicle on track. The RTDB manages objects that can be created and updated by

[1] All authors contributed equally.

From: Proc. HSI 2009, Catania, Italy, IEEE

input modules also called writers. These writers also have to submit a timestamp for the committed data. Thereby it is possible for the RTDB to synchronize the data coming in asynchronously from multiple sources and at different sample rates. Output modules (called readers) wait for new objects to process them. For example, a module can write the image of a camera and multiple other modules can analyze this image in parallel to generate information on a higher level and write this output back for other modules without blocking effects. These data-objects can be recorded in real-time, bringing up two major advantages:

- The recorded sensor-input can be taken for replay or simulation of certain situations. In addition, the gathered material can be analyzed by humans off-line, e.g. to reveal important gestures in the co-operation process, or to see if the worker is stressed.
- The data can also be used for benchmarking purpose. Different reader implementations can be tested on the same data under the same conditions. After the new reader proves to work better than the old one, they can be used on the real set-up on-line without any modification to the code. The recorded database of sensor data can also be used by other projects to evaluate their system or algorithms in an unknown environment.

In the following sections, we will have a closer look at the software-modules that can be linked to the RTDB framework.

A. Image-based Processing

The video modules deliver raw RGB data from different sources like firewire cams or USB cams in a common representation based on OpenCV. The OpenCV library has been chosen, because it is widespread and numerous output modules working exist, e.g. to localize faces or hands. In addition to these modules it is planned to train and implement gesture recognition like [6]. To obtain better results in training and recognition it has been decided to record the video data uncompressed. The following image-processing tools are currently available in RTDB framework.

- Video-Recording: The RTDB framework is capable to store video data in an AVI-container having a resolution 640 x 480 pixels. For the video-recording operation the RTDB-recorder is used.
- Box-Localization: For the storage of construction components boxes are used in both scenarios (human-human interaction, human-robot interaction). The color-coded boxes can be detected by a thresholding operation performed in the HSV-color-space and a subsequent analysis of geometrical characteristics of the obtained color regions, cf. [7]. The obtained box positions can be easily transformed into robot coordinates enabling grasping and hand-over operations.
- Hand-Localization: The hand of a human worker can be detected by evaluating the result of skin-color filter operation, cf. [8]. These results are compared with the geometrical constraints – size, relation of width and height – stored in the system to validate only real human hands.
- Soft-Buttons: Soft-Buttons are selective fields projected via the table-projector unit onto the workbench to address certain system inputs. These inputs are triggered via the hand-localization method. A field will be activated, if the human hand is hovering over the sensitive field or the field is fixated with the eyes for a certain amount of time.
- Gaze-Control: The worker wearing eye-tracking glasses [9] can control the system via gaze. The system interprets the intersection of the gaze trajectory and the workbench surface as input command for the Soft-Button interface.
- Range Maps: For a ”deeper” view of the scene, input from a camera providing range maps has been recorded. Based on the novel Photonic Mixer Device (PMD) technology, the camera collects depth-information in real-time by emitting infrared light and measuring the time-of-flight. Thereby the distances from the camera can be calculated. It has a resolution of 64 x 48 Pixels at 25 frames per second. This additional depth information can be used to improve segmentation tasks for image processing or detection of human activities like handovers. More information regarding this sensor and calibration techniques can be found in [10]. However, because the camera is sensitive for infrared light it can also be used to provide intensity based gray scale images.

B. Audio-based Processing

With the RTDB is is possible to process audio data. This data is encoded in the Pulse-Code-Modulated (PCM) audio data format.

- Audio-Recording: The processed audio-data via the RTDB can be stored either in the RTDB-AVI-container (mentioned above), or be exported in the WAV-file format.
- Speech Recognition and Synthesis: A commercial speech recognition and speech synthesis software is linked to the RTDB to have both tools incorporated.

C. Physiological Data

Three physiological signals were recorded at a sampling rate of 256 Hz, i.e. the heart rate (ECG), skin conductance (SCR) and pulse (photoplethysmograph). From the difference between pulse and ECG, a blood pressure equivalent could be calculated. Sensors were applied to the ear (pulse), the knee (SCR) and the upper part of the subject’s body (ECG), respectively. These sensors were connected to a mobile biosignal acquisition device (g.MOBIlab) where the input was converted to digital signals. This mobile device was connected to the Real-time Database using a serial connection via bluetooth. Thus, objective and continuous information about the subject’s current state like frustration, stress, or fear is made available.

III. SCENARIOS

A. Human-Human Experiments

In the following paragraphs, we will present our human-human experimental setup considering the hardware setup as well as the psychological aspects for the experiment.

1) Hardware Setup: In our human-human experiment setup, three webcams and one firewire cam recorded images in real-time at a resolution of 640 x 480 at 15 frames per second. In order to compensate for different lighting conditions, the gain of the cameras were controlled online during the experiments. Besides, the audio of both experiment subjects were recorded with above mentioned RTDB audio module. In this setup the input was displayed, so that e.g. the physiological data could be monitored online, this data was only collected in this experimental setup. The used Architecture is depicted in Figure 1.

2) Experimental Setup: The ideal case of a naturalistic interaction is the one between humans, as they are able to respond to multimodal information intuitively. Hence, our starting point are human-human experiments which allow for information collection concerning important situation characteristics and their dynamics. The gained insights will then be used to improve human-robot interaction by implementing such knowledge into the robot. Transferability of the results is a major concern, and thus, a near to naturalistic setting was chosen resembling a working situation in a factory. One of the humans took the role of the human co-worker and the other one the role of the robot, with the latter handing over single components and giving supporting information about each construction step if needed. In our experiment, the "robot" was an instructed companion whereas the human co-worker was a test person. "Robot" and participant sat face to face at a table and were asked to solve a joint construction task with the LEGO-Mindstorms system (setup shown in Figure 2). We chose to use a LEGO construction task for several reasons: First, there is always a definite plan about what to do, but it's still complex enough for a robot to be a useful support. Second, On the other hand a human is needed, because some LEGO parts are too small for a robot to be assembled. And finally, the Mindstorms system is very flexible and thus enables many variations of the task.

As we were interested in an ideal interaction situation, the "robot" had been instructed to be supportive, not to become impatient, to hand over the LEGO parts just in time, to explain difficult construction steps with or without demand, as well as to cheer the co-worker up in case that he or she might look frustrated. "It" was able to react like a human would in such a situation, apart from two constraints: like an industrial robot, it could not put together single LEGO parts, and the sight was artificially impaired by placing a semi-transparent foil between the two actors. The foil was about 30 cm high so that LEGO parts could be exchanged, and the faces of the interaction partners were still visible. Thus, mimic information could still be used.

3) Procedure: Before the experiment, participants could make themselves acquainted with the instructions and the overall setting during a practice phase. Furthermore, the well-being of the human co-worker was measured by a list of adjectives the participant was asked to answer.

After the experiment, a second questionnaire had to be completed concerning the actual affective state, the experienced quality of the interaction, as well as ratings of the task and the support given by the robot. In addition to this subjective information, objective criteria like the number of errors, time to complete the task etc. were also collected. These data will then be used as evaluation criterion that allows the comparison with data from interacting with a real robot with varying abilities.

4) Data Collection: Recorded data included the speech of participant and "robot", physiological data (heart rate, skin conductance, pulse) of the participant, and camera views from 4 different angles (cf. Figure 6). For the analysis of facial expression, one camera focused explicitly on the face of the participant. This data will be used for the inference of emotional states or nonverbal cues in the interaction (e.g. frowning as a sign of irritation or confusion). Another perspective showed the table from the top for gesture recognition purposes and for coding the task progress. To be able to analyze the (appropriate) reactions of the "robot" there was also a camera aiming at him. The fourth camera recorded the whole interaction scene from a more distant point of view. This allows for an identification of important dynamic events occurring between the interaction partners. Physiological data were also used for the analysis of the participant's affective state, additionally to the facial expression and the self rating from the questionnaire. By using redundant information it is possible to reveal where the different measures match and where they do not, or to judge which information source should be weighted higher in certain situations. This might also give hints regarding the compensation in case of malfunction or data loss of one channel.

B. Human-Robot Scenario

In this section, we will briefly describe how the above described software framework can be used in a human-robot interaction scenario similar to the setup of the CoTeSys Project *Joint Action of Humans and Industrial Robots* (JAHIR) presented in [11]. Analogue to the above described setup for human-human experiments the framework has to be adapted in order to not only view the sensor's output but also to control the outputs and react to the worker's actions. Therefore, the above mentioned modules have to be extended in the following manner that several different computer vision algorithms can access the same video input signal in parallel (shared-memory).

A typical joint action scene with a human worker and an industrial robot arm is depicted in Figure 3. The area of application for the human-robot scenario is situated in a hybrid assembly task in the so called Cognitive Factory [12]. In this case, the robot is used as an assistant for the human worker similar to the above described human-human experiment. Besides, the robot helps fulfilling the worker's

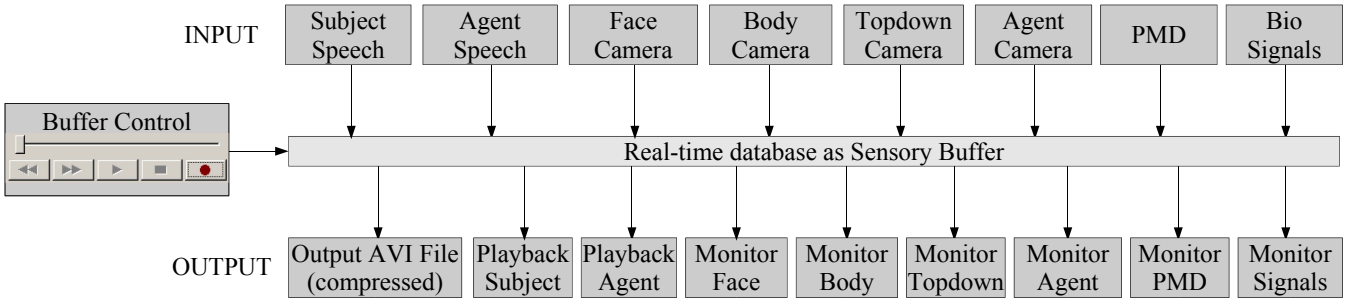

Fig. 1. Overview of the multi-modal recording scheme with the RTDB.

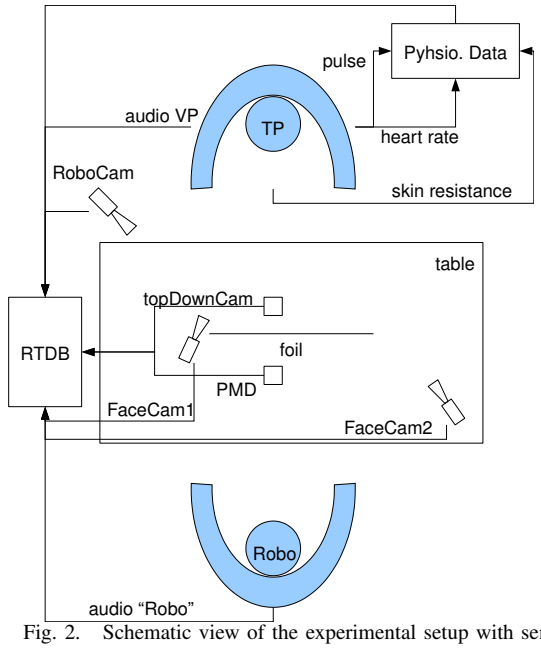

Fig. 2. Schematic view of the experimental setup with sensors.

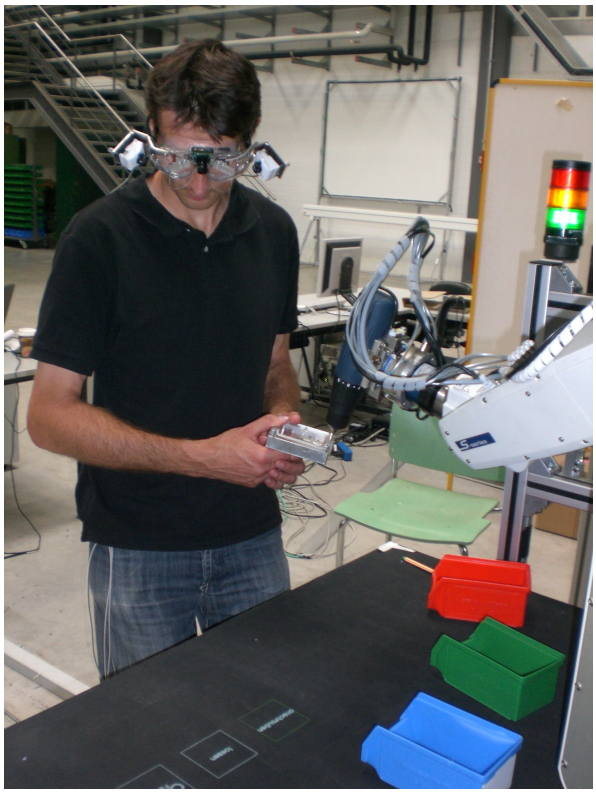

Fig. 3. Human worker and industrial robot in hybrid assembly.

task. The required hardware setup will be described in the following in more detail. Afterwards, the use-case will be considered.

1) Hardware Setup: In our setup we used one webcams and one firewire cam and recorded their images in real-time at a resolution of 640 x 480 at 15 frames per second. In order to compensate different lighting conditions at our setup during the experiments the gain of the cameras was controlled on-line.

In Figure 5 the hardware setup of the human-robot interaction demonstrator is shown. The Mitsubishi robot RV-6SL constitutes the foundation for the robot manipulator platform for the scenario situated in the industrial context. The Mitsubishi robot RV-6SL has six degrees of freedom and can heave objects with a maximum weight of six kilograms. A radius of 0.902 m around its body constitutes workspace of the Mitsubishi robot RV-6SL. The tool point is equipped with a force-torque-sensor and a tool-change unit. In addition, the currently installed gripper of the robot can be replaced by the robot itself at a station (see the left side of the table in Figure 5). Four different kinds of manipulators are capable of performing specific operations. The tools stored in the station are: two finger parallel gripper, electronic drill, camera unit for automatic observations, and a gluer. This gives the robot the capabilities of being able to solve entirely different tasks, like screwing and lifting.

Two kind of camera devices (webcam, PMD) as well as a table projector unit are directly mounted above the workbench, which has an overall range of nearly 0.70 square meters. The webcam is used as a top-down view camera delivering an overview of the entire workbench area, besides, this area is additionally monitored in a three-dimensional manner with the PMD.

For the information presentation the table projector unit is

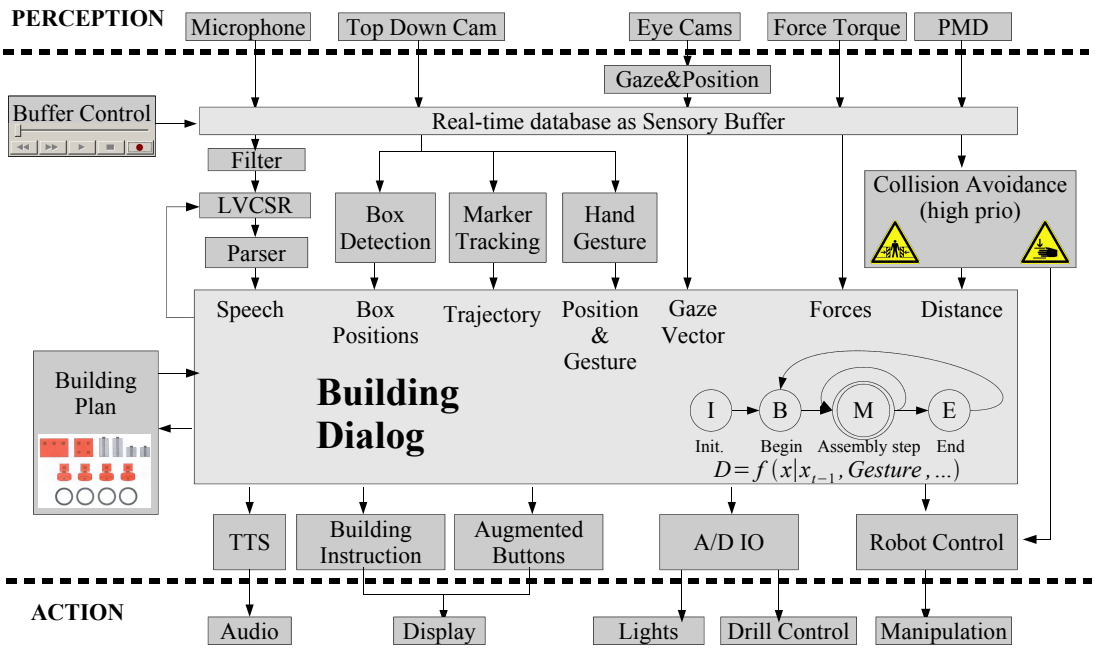

Fig. 4. Functional system overview with perception, cognition and output based on the RTDB.

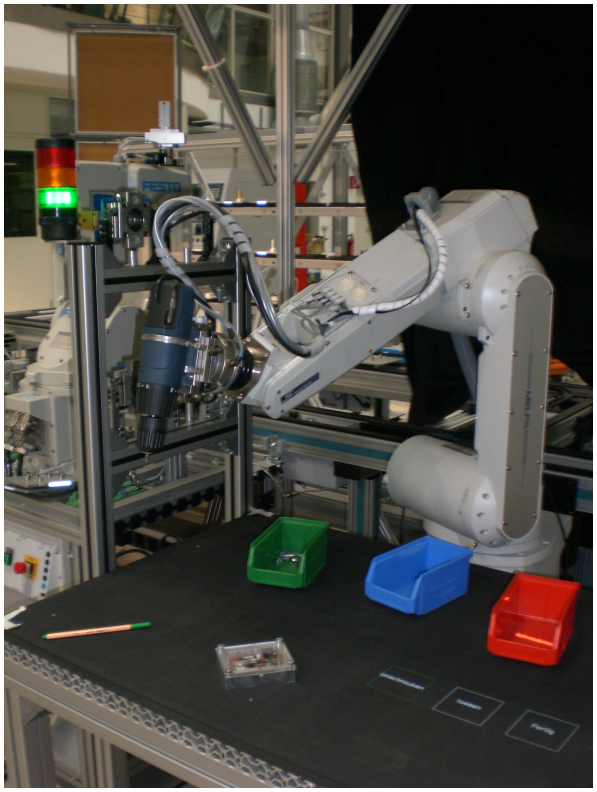

Fig. 5. Hybrid assembly station: tool station, robot arm (with electric drill) and assembly-table.

used to show different work instructions and system status feedbacks directly in the worker's field of view.

2) Demonstration Setup: In the human-robot demonstration scenario is located in a hybrid assembly task. In this task the human worker builds together with an industrial robot an exemplary non-working HF-transmitter. For this demonstration scenario further experimental evaluations are planned.

The first manufacturing step – triggered by user input – between the human and the robot is the base plate delivering. Having the required work pieces available at hand, the worker starts to teach in a glue-line. The track of the glue-line on the base plate is taught with *Programming by Demonstration* (PbD). This is done by tracking a colored pointer, similar to the box-detection. While the line is perceived, its trajectory is on-line projected back onto the work piece as a direct feedback for the worker. On completion of this step the robot changes its tool device from the gripper to the gluer according to the next step in the work plan. After the PbD, the robot protracts the glue on the work piece. As per assembly instructions the robot reaches out for the electronic parts.

Fine motor skills are required for assembling the electronic parts. Therefore, the next step is solely done by the human. In spite of the fact that the robot does not give any active assistance in this assembly step, the system supports the worker via presenting the manufacturing instructions for the insertion of the electronic parts into the base plate. After the worker has acknowledged the completion of the current step via Soft-Button or a speech based command, the robot fetches the four screws for the final assembly step. While the worker is pre-fitting the screws in the designated mount ports, the robot retrieves the automatic drill device from the tool changer station. The velocity of the drill is adjusted to the contact pressure of the work piece against the drill. The more pressure is applied, the faster the drill goes. As soon as the

human recognizes that the screw is fixed – the rattling noise of the slipping clutch – he will loosen the former conducted pressure. This modality allows an intuitive screwing behavior of the system.

IV. Transcription

For the transcription of the data, ANVIL [13] is used. This is a free video annotation tool that allows for a frame-accurate and hierarchical data annotation in multiple layers. The annotation schemes can be freely defined by the user. It has originally been developed for gesture research, but has also proven to be suitable for research in human-computer interaction, linguistics, psychotherapy and many other fields.

In order to process the complete multimodal information with ANVIL, we compiled a video stream combining all video and audio channels into one common AVI-file. Additional data like the depth map of the range sensor and the biological signals are either coded in gray values or visualized as a function over time. The result of this process is depicted in Figure 6.

By having all relevant information in one window at a glance, the use of the annotation board becomes very intuitive, because it also displays color-coded elements on multiple tracks together with its time-alignment. Further features are cross-level links, non-temporal objects and a project tool for managing multiple annotations.

However, it is also necessary to annotate the recorded audio data. Therefore, data from PRAAT and XWaves, which allow precise and comfortable speech transcription, will be imported into ANVIL. The created annotation data is based on XML and can thereby easily be exchanged with other tools. A comparison and compatibility with the Interaction Analysis Tool (IAT), formerly called Interaction Protocol (IAP) [14], is considered in the future.

A qualitative video data analysis was conducted, comparable to those suggested by Kahn et al. (2003) [15], Zara et al. (2007) [16], or Dautenhahn & Werry (2002) [17]. As critical annotation categories are highly context-dependent, it was necessary to develop a system that suited to our specific application context.

The first step was the basic coding of the different modalities. This comprises not only facial expressions and speech, but also gestures, gaze, and changes in physiological parameters. Beyond those micro-events, we were also interested in the psychologically relevant interplay or certain timely orders of the modalities (e.g. ”event a” always shortly happens before ”event b” or ”gesture c” always goes hand in hand with ”utterance d”). Regarding speech, the complete dialog structure was transcripted. For this purpose, we used a hierarchical category system comprising different main categories with several subdivisions.

a) Facial expressions: This category includes the movement of eyebrows and forehead muscles (e.g. ”frowning”), eyes and ears, nose and cheeks, lips and corners of the mouth.

b) Head: Head movements and head-related actions like ”resting the head on one’s hand” are coded here. This includes for example nodding or head shaking, holding it askew or laying the neck in the back.

c) Physiology: This category contains changes in blood volume pressure, heart rate or skin conductance as correlates of the human’s actual state (e.g. the stress level).

d) Gestures: In the gesture category, hand and arm movements are coded. It is differentiated whether the gesture was object-related (e.g. ”pointing”), self- or other-related or whether it served as a non-verbal utterance (e.g.”cover one’s face with the hands”, ”tap one’s forehead”).

e) Gaze: This category was included since gaze is a crucial factor of interaction: Does the human look at the instructions, or into the robot’s face, waiting for help? Gaze was annotated from video data and not recorded by an eye-tracking system since this would have disturbed the recording of the facial expressions.

f) Speech: The speech category comprises the literally transcription of the spoken material. It is divided into semantic phrases to get meaningful data that can be related to the other non-verbal categories.

g) Emotional state: This important higher order category was derived from a combination of the categories ”facial expressions”, ”physiology” and ”speech”. It comprises states relevant for this specific context like interest, impatience, or confusion.

This annotation process served two goals: Firstly, we want to find those principles and structures in the human-human-interaction that have to be implemented for an improved human-robot-interaction. And secondly, it helps to give tags to the different behaviors, actions and events that have to be recognized in the interaction process. Thus, it shall be possible to train the technical recognition algorithms with the important parameters. Since our category system is very complex, the annotation process is still in progress.

V. Conclusions and Future Work

We have presented the implementation of a unified software framework based on the RTDB that efficiently allows for on- and off-line processing of multimodal, asynchronous data streams. The data can origin from audio, video, haptic, range sensors and also physiological signals. The proposed framework is foreseen to gain detailed knowledge about verbal and especially non-verbal interaction behavior from human-human experiments. This knowledge will later be exploited and transferred in order to improve the human-robot interaction.

Until now, this framework has been applied to record human-human experiments observing human behavior in a cooperative assembly scenario. It has further served as an effective middleware in a human-robot joint action scenario with real-time constraints.

The pending next steps are to identify important indicators that can be derived from human behavior in order to train the parameters of a sophisticated machine learning algorithm, which in turn can then be used in a cognitive system

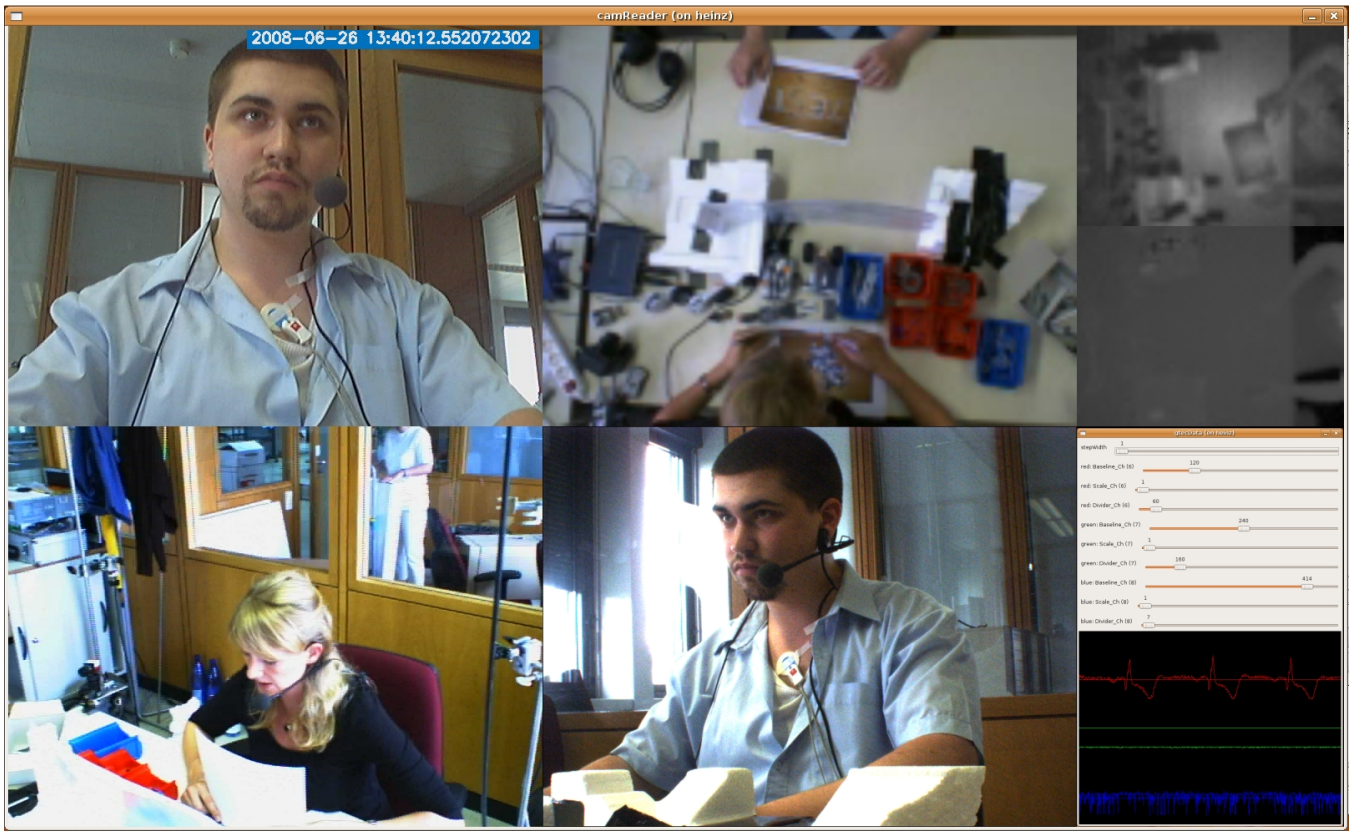

Fig. 6. Simultaneous visualization of different video channels modalities.

understanding its user's intentions. The first human-human experiment served as benchmark for an optimal situation. Further experiments are planned to include more situations and to further elaborate the algorithms. Hence, it is planned to conduct experiments with purposely induced errors on the robot side (handing the wrong parts, not being on time etc.). This will enable us to see how problems in dyads are dealt with. In such situations, much more interaction between the two partners is expected because they will have to jointly solve problems. Further, we want to simulate what happens if different modalities fail (no speech recognition or output, no facial recognition etc.) and which modalities could help to compensate for that.

Another demanding aspect is the definition of the perceived quality improvement over a system that only has reactive performance: "How can the performance of a cognitive system be measured?"

VI. Acknowledgment

This ongoing work is supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see www.cotesys.org for further details. The authors further acknowledge the great support of Matthias Göbl for his explanations and granting access to the RTDB repository. Furthermore, we would like to thank Joachim Schenk and Stefan Schwärzler for the RTDB-Audio-Reader and -Writer.

References

[1] D. S. Touretzky and E. J. Tira-Thompson, "Tekkotsu: A framework for aibo cognitive robotics," in *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*, Pittsburgh, PA., July 2005.

[2] M. Göbl and G. Färber, "Interfaces for integrating cognitive functions into intelligent vehicles." in *In Proc. IEEE Intelligent Vehicles Symposium*, June 2008, pp. 1093–1100.

[3] M. Thuy, M. Göbl, F. Rattei, M. Althoff, F. Obermeier, S. Hawe, R. Nagel, S. Kraus, C. Wang, F. Hecker, M. Russ, M. Schweitzer, F. P. León, K. Diepold, J. Eberspächer, B. Heißing, and H.-J. Wünsche, "Kognitive automobile - neue konzepte und ideen des sonderforschungsbereiches/tr-28," in *Aktive Sicherheit durch Fahrerassistenz*, Garching bei München, 7-8 April 2008.

[4] C. Stiller, G. Färber, and S. Kammel, "Cooperative cognitive automobiles," in *Intelligent Vehicles Symposium, 2007 IEEE*, June 2007, pp. 215–220.

[5] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," in *Intelligent Vehicles Symposium*. IEEE Press, June 2007, pp. 737–740.

[6] S. Reifinger, F. Wallhoff, M. Ablaßmeier, T. Poitschke, and G. Rigoll, "Static and dynamic hand-gesture recognition for augmented reality applications," in *Proceedings of the International Conference on Human-Computer Interaction*, C. Stephanidis, Ed. Beijing: Springer, July 2007.

[7] A. Bannat, J. Gast, G. Rigoll, and F. Wallhoff, "Event Analysis and Interpretation of Human Activity for Augmented Reality-based Assistant Systems," in *IEEE Proceeding ICCP 2008*, Cluj-Napoca, Romania, August 28-30 2008.

[8] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen, "Skin Detection in Video under Changing Illumination Conditions," in *Proc. 15th International Conference on Pattern Recognition*, Barcelona, Spain, 2000, pp. 839–842.

[9] S. Bardins, T. Poitschke, and S. Kohlbecher, "Gaze-based Interaction in various Environments." in *Proceedings of 1st ACM International*

Workshop on Vision Networks for Behaviour Analysis, VNBA 2008, Vancouver, Canada, October 31 2008.

[10] F. Wallhoff, M. Ruß, G. Rigoll, J. Göbel, and H. Diehl, "Surveillance and activity recognition with depth information," in *IEEE International Conference on Image Processing (ICIP)*, San Antonio, Texas, USA, September, 16-19 2007.

[11] C. Lenz, N. Suraj, M. Rickert, A. Knoll, W. Rösel, A. Bannat, J. Gast, and F. Wallhoff, "Joint actions for humans and industrial robots: A hybrid assembly concept." in *Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication*, August 2008.

[12] M. Zäh, C. Lau, M. Wiesbeck, M. Ostgathe, and W. Vogl, "Towards the Cognitive Factory," in *International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV)*, Toronto, Canada, July 2007.

[13] M. Kipp, "Anvil - a generic annotation tool for multimodal dialogue," in *7th European Conference on Speech Communication and Technology*, 2001, pp. 1367–1370.

[14] C. R. Burghart and A. Steinfeld, "Proceedings of metrics for human-robot interaction, a workshop at acm/ieee hri 2008," in *Technical Report 471, School of Computer Science, University of Hertfordshire*, Hatfield, UK, March 2008.

[15] P. H. Kahn, J. B. Friedman, N. Freier, and R. Severson, "Coding manual for children's interactions with aibo, the robotic dog – the preschool study (uw cse technical report 03-04-03)," Seattle: University of Washington, Department of Computer Science and Engineering, Tech. Rep., 2003.

[16] M. A. Zara, "Collection and annotation of a corpus of human-human multimodal interactions: emotion and others anthropomorphic characteristics." in *ACII*, 2007.

[17] K. Dautenhahn and I. Werry, "A quantitative technique for analysing robot-human interactions." in *Proc. Intl. Conf. on Intel. Rob. Sys.*, 2002.