Real-time Framework for On- and Off-line Multimodal Human-Human and Human-Robot Interaction

Frank Wallhoff, Jürgen Gast, Alexander Bannat, Stefan Schwärzler, Gerhard Rigoll
Human-Machine Communication, Department of Electrical Engineering and Information Technologies
Technische Universität München, Germany
{wallhoff,gast,bannat,schwaerzler,rigoll}@mmk.ei.tum.de
Cornelia Wendt, Sabrina Schmidt, Michael Popp, Berthold Färber
Institut für Arbeitswissenschaften, Fakultät für Luft- und Raumfahrttechnik
Universität der Bundeswehr München
{sa.schmidt,cornelia.wendt,michael.popp,berhold.faerber}@unibw.de
All authors contributed equally.

***Abstract*— In this paper we present a framework for real-time processing of multimodal data, which can be used for on- and off-line processing of perceived data in interactions. We propose the use of a framework based on the Real-time Database (RTDB). This framework allows easy integration of input and output modules and thereby concentrating on the core functionality of the module. Furthermore the asynchronous data from different sources is synchronized and can be recorded for off-line processing. This recorded data can be used to train recognition modules, which can then be used again on-line. Experiments and first results with off-line human-human and on-line human-robots are reported.**

I. Introduction

The focus of the excellence research cluster *Cognition for Technical Systems CoTeSys* is to make technical systems behave more intelligent and useful also in unseen situations, e.g. like a human. A cognitive system should be able to learn and perceive its environment. One important aspect is to make the interaction and communication between humans and technical systems more intuitive. Therefore it is necessary to gain knowledge and deeper understanding of multimodal human communication processes, especially in terms of automatic perception and recognition. In a fist step interaction processes between humans shall be studied in experiments. The obtained information can be used to build up a working model for the system interaction. Furthermore the model should be able to optimize its reactions with respect to the user and adapt his behaviour. This model can then be evaluated in experiments.

The general idea is to engineer a common framework for recording human-human experiments to collect data and use the measured data to train and construct a working model. The same framework will then also be used to observe human-robot experiments. The prior trained modules shall react to the actions of the human and control the robot accordingly. In order to perceive the interaction process as good as possible multiple modalities have to be used, i.e. several video sources, audio signals, physiological data, range sensors, haptic, eyetracking, etc. All these sensors have different sampling rates, which makes it necessary to synchronize the data for further processing. Such a framework is essential for example for human-like information processing of emotions, since it has been shown that the coincidence between muscle activity during facial expressions is a key to detect real or faked emotional reactions.

Although in the literature several middleware architectures have already been introduced to process 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 appropriate to act as a tool to record multimodal appropriately. Therefore we propose to use the Real-time database (RTDB) as sensory buffer with a high data bandwidth that can be used 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]. It furthermore has served as a convincing integration platform where several groups of researches 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 more detail. in Section III we present some conducted human-human experiments, followed by a setup for on-line human-robot experiments in Section IV. The paper closes with a summary and an outlook over the next envisaged steps.

II. System Architecture

In this section a short overview of the System Architecture and the modules used for the recording of the experiments will be given. The system architecture consists of the Real-time Database as sensory buffer and communication backbone between the input and output modules. The used Architecture is depicted in Figure 1.

The Real-time Database buffers and synchronizes the input modules that generate data for the output modules. In this setup the input was displayed, so that e.g. the physiological

From: 1st Intern. Cotesys Workshop 2008, Technische Universität München

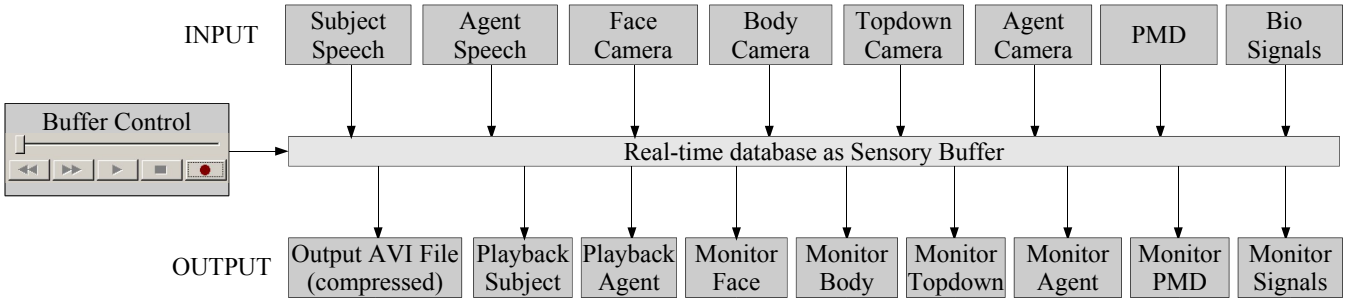

Fig. 1. Overview of the multi-modal recording scheme with the RTDB.

data could be monitored on-line. 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 real-time Database.

A. Real-time Database

The Real-time Database presented in [5] is able to deal with large amounts of sensor data (in our setup 47 Megabytes per second) 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 real-time-database manages objects that can be created and updated by input modules also called writers. These writers also have to submit a timestamp for the committed data. Thereby it is possible for the real-time database 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 analysed 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.

B. Video Recording

The video recording modules deliver raw RGB data from different sources like firewire cams or USB cams in a common representation based on OpenCV [6]. The OpenCV library has been chosen, because it is widespread and numerous output modules working exist, e.g. to localize a face or hands. In addition to these modules it is planned to train and implement gesture recognition like [7]. To obtain better results in training and recognition it has been decided to record the video data uncompressed. In our setup we use three webcams and one firewire cam and record 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 can be controlled on-line.

C. Audio Recording

Audio recording has been done using the Advanced Linux Sound Architecture (ALSA) library. The Pulse-Code-Modulated (PCM) audio data from the input device was buffered and stored in the audio object. Furthermore the sampling rate and other device specific parameters have also been saved. Thereby it is possible for an audio reader to play the recorded sound also on hardware that differs from the one used for recording. In our setup we used two high-quality head-worn microphones and recorded in stereo (using one channel per person) with 16 signed bits, little endian and a sampling rate of 44.1 kHz. This audio input can not only be used to play back the conversation, but also to train a speech recognition module. The same module can then be used for speech recognition and natural language understanding like in [8].

D. Physiological Data

Three physiological dimensions have been measured at a sampling rate of 256 Hz, i.e. the heart rate, skin conductance and pulse. The information was obtained from sensors on the ear, the knee and the upper part of the subject's body. These sensors were connected to a mobile bio signal 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. These signals have been recorded to have an objective measurement and deeper view of the subjects current status, e.g. if the person is stressed from the current task.

E. 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 [9]. However, because the camera is sensitive for infrared light it can also be used to provide intensity based gray scale image.

F. Transcription

For the transcription of the data ANVIL [10] is used. It is a free video annotation research tool and the annotation can be done frame-accurate and hierarchical in multiple layers. The annotation schemes can be freely defined by the user. It has been originally 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 all multimodal information with ANVIL we compile a combined video stream with all video and audio channels into one common AVI-file. Other data like the depth map of the range sensor and the biological signals are either coded in gray values or visualised as a function over time. The result of this process is depicted in Figure 2.

By having all relevant information in one window at a glance, the work with 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) [11], is considered in the future.

III. Human-Human Experiments

A. Experimental Setup

In order to improve human-robot interaction, a near to naturalistic interaction between two human actors was observed. One of them took the role of the robot and the other one the role of the human co-worker in a kind of factory setting. 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 (see Figure 3). Although the employment of LEGO for the experiment might give the impression of toys we chose it because there is always a definite plan. Furthermore the Mindstorms system is very high developed, so that building and programming a huge amount of different robots or machines is possible. That means working with this LEGO system is complex enough for a robot to be a useful support. On the other hand a human is needed, because some LEGO parts are too small or too tricky for a robot to be assembled.

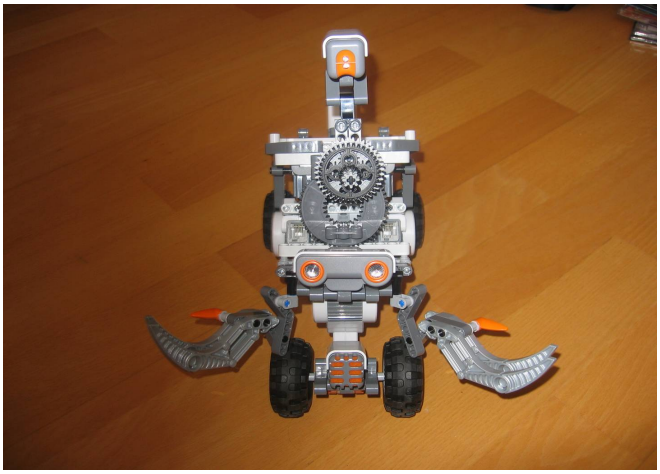

Fig. 3. Lego-Mindstorms robot.

The instructed companion, i.e. the ”robot”, had been instructed to act like an ideal kind of robot in terms of being supportive, not getting impatient, handing over the LEGO parts just in time, explaining difficult construction steps, or trying to cheer the co-worker up. ”It” was able to point, to speak, hear or react like a human, apart from two constraints: like a real robot, it could not put together the 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 for both of them. Thus, mimic information could still be used.

B. Procedure

To make sure that every participant had a comparable level of experience with the Mindstorm Lego parts, there was a practice phase before the experimental phase. In this practice task the participants had to construct the word ”TEST” from available Lego bricks under supervision of the ”robot”. Thus, they could make themselves acquainted with the instructions, the different LEGO parts and the overall setting. Before the test phase, participants were asked to complete a questionnaire surveying their former experience with LEGO and their affinity to mechanics in general. Technical affinity as personality variable might influence the completion of the rather technical LEGO-task, the type and quality of the interaction with the robot and the evaluation of the interaction (maybe not with a human, but especially with a real robot). Furthermore, the well-being of the human co-worker was measured by a list of adjectives the participant had to answer.

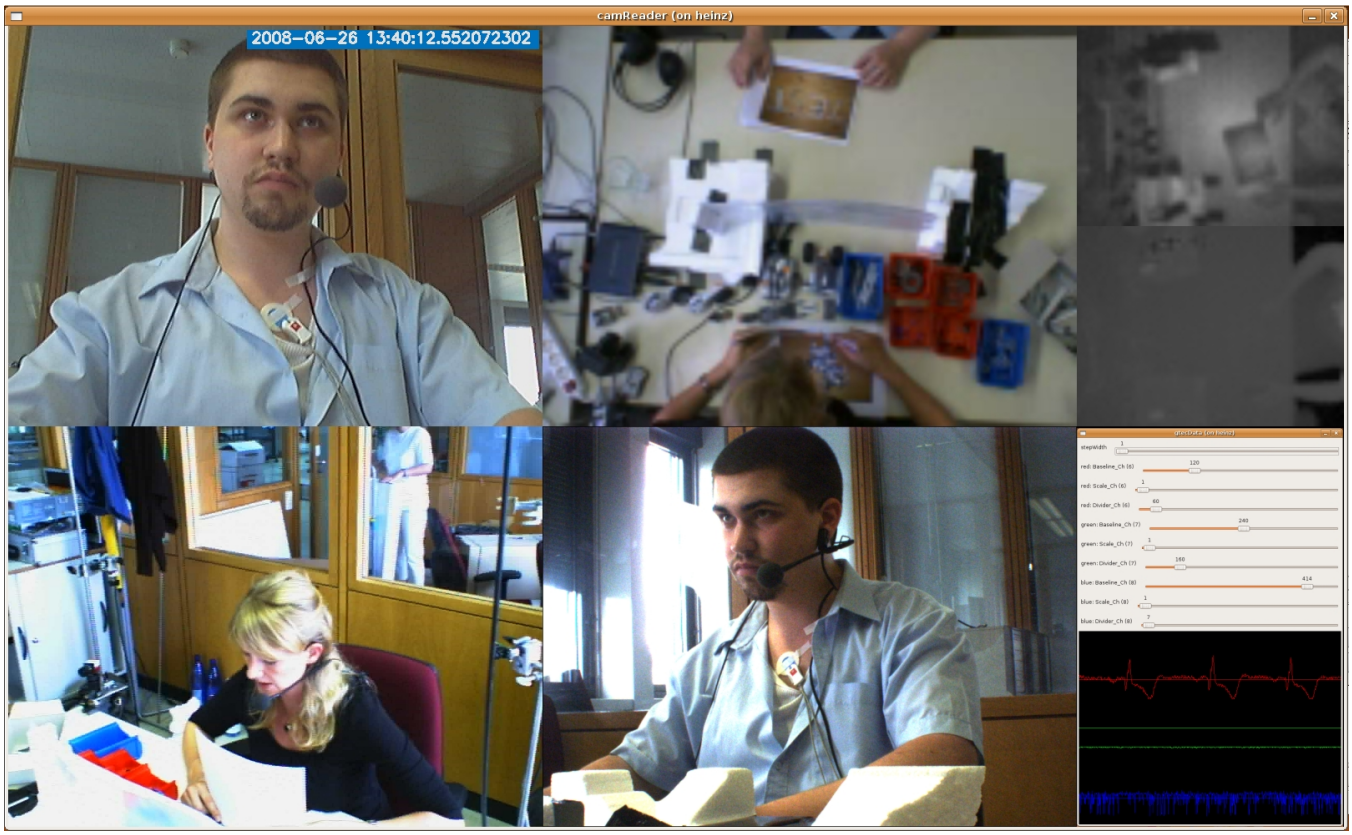

Fig. 2. Simultaneous visualisation of different video channels and other modalities.

After the experiment, a second questionnaire had to be completed concerning the affective state again, the experienced quality of the interaction, as well as ratings of the task and the support given by the robot.

In addition to the aforementioned subjective measures, objective criteria like the number of errors, time to complete the task etc. were also collected. These data will then be used as a benchmark that allows the comparison with ratings after interacting with a real robot, with or without certain abilities.

Comparing the answers concerning the affective state before and after the interaction allows for estimating the influence of the interaction process given the initial mood. Secondly, the initial mood effects the interaction, since just asking afterwards simply means only using half of the available information.

C. data collection

During the experiment (setup shown in Figure 4), recorded data included the speech of the participant and the robot, physiological data (heart rate, skin conductance, pulse) and camera views from 4 different angles (cf. Figure 2). For the analysis of facial expression there was one perspective 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 don't, or to judge which information source should be weighted higher in certain situations. This might also give hints regarding the compensation for the probable malfunction or loss of one of the measures.

D. data analysis

With regard to the video data, a qualitative data analysis is planned, comparable to those suggested by Kahn et al. (2003) [12], Zara et al. (2007) [13], or Dautenhahn & Werry (2002) [14]. As categories are highly context-dependent, we will develop a system of our own with diverse dimensions. Coding the different modalities independently will be the most basic step. This comprises not only facial expressions and speech, but also gestures, certain movements, as well as changes in physiological parameters. Beyond those micro-events, we are also interested in their psychologically relevant interplay or certain timely orders (e.g. ”event a” always shortly happens before ”event b”). With regard to speech, the complete dialogue structure will be analysed. On this meta-level, the

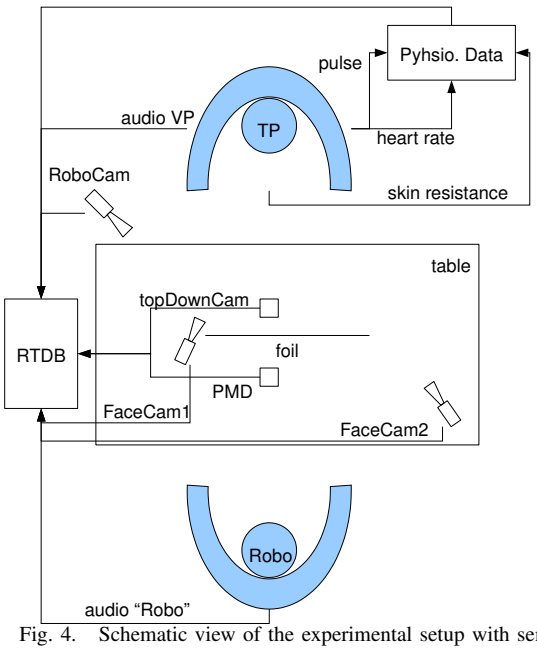

Fig. 4. Schematic view of the experimental setup with sensors.

dynamics of initiative are another interesting aspect. One point that will be realised in the robot is the possibility of mixed initiative. That means the robot can also be the initiator of an action. Usually robots are only reacting or following a given script. Of course the robot should be able to do all that, but here he might also take action, for example offering help to the co-worker when realising problems with the task, or suggesting another tempo when the co-worker gets annoyed. The categorisation process will help in two different ways: Firstly, to find 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 behaviours, actions and events that have to be recognised in the interaction process. Thus, it shall be possible to train the technical recognition algorithms with the important parameters.

This first experiment serves as benchmark for an optimal situation. More 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.

IV. Human-Robot Experiments

In this section we briefly describe how the above described software framework can be used in conjunction with on-line experiments. In contrast to human-human experiments the framework is now embedded in a human-robot interaction scenario similar to the setup of the CoTeSys Project *Joint Action of Humans and Industrial Robots* (JAHIR) presented in [15]. A typical joint action scene with a human worker and an industrial robot arm is depicted in Figure 6.


Fig. 6. Working cell with human worker and industrial robot arm.

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. This implies that the above mentioned modules are expanded in the way that for example several different computer vision algorithms can access the same video source without blocking.

In this scenario three processing units share the same top-down camera input, i.e. 1) a hand gesture recognizer, 2) a pointer tracking module and 3) a box localisation module. This setup has been presented during the AUTOMATICA exhibition 2008 with great success. Besides the video source also other modalities are processed in parallel. Without going into all details the actual configuration is depicted in Figure 5.

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 behaviour 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 behaviour 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

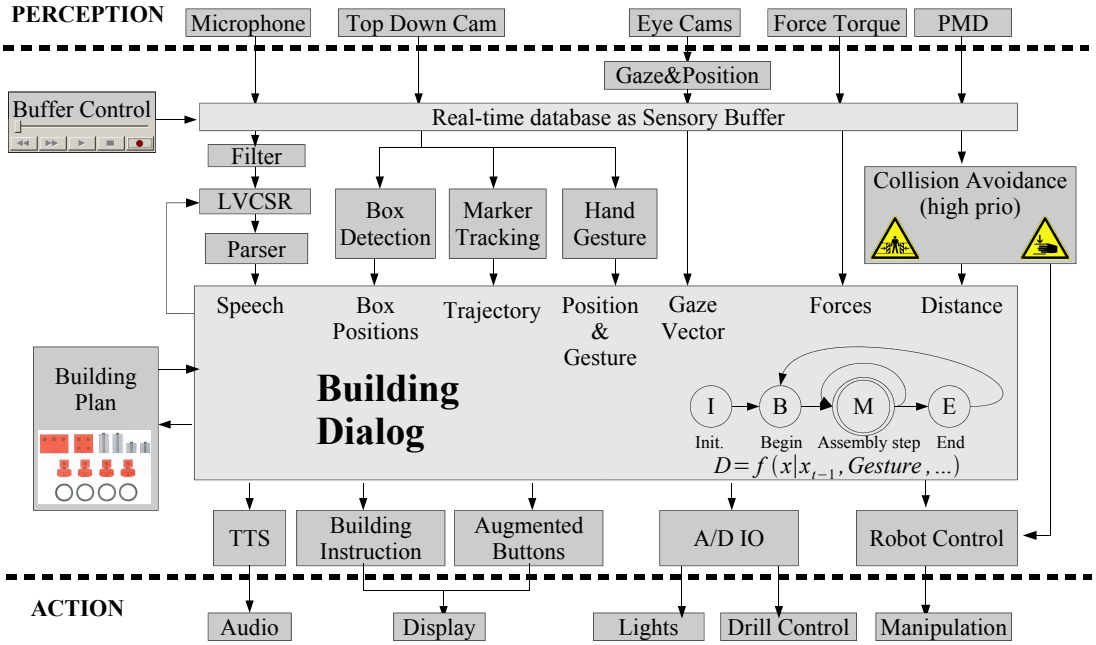

Fig. 5. Functional system overview with perception, cognition and output based on the RTDB.

the parameters of a sophisticated machine learning algorithm, which in turn can then be used in a cognitive system understanding its user's intentions. 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. Acknowledgement

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. The experiments were conducted in the interdisciplinary project MuDiS – A Multimodal Dialog Manager. Big thanks to all project partners for the many fruitful discussions.

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] OpenCV, "Open computer vision library," http://opencvlibrary.sourceforge.net.

[7] 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.

[8] S. Schwärzler, J. Schenk, F. Wallhoff, and G. Ruske, "Natural Language Understanding By Combining Statistical Methods And Extended Context-free Grammars," in *Proc. of 30th DAGM Symposium*, ser. LNCS 5096, G. Rigoll, Ed. Heidelberg, Germany: Springer, 2008, pp. 254 – 263.

[9] 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.

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

[11] 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.

[12] 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.

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

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

[15] 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.