Multi Joint Action in CoTeSys - Setup and Challenges -

Technical report CoTeSys-TR-10-01

D. Brščić, F. Rohrmüller, O. Kourakos, S. Sosnowski, D. Althoff, M. Lawitzky, A. Mörtl, M. Rambow, V. Koropouli, J.R. Medina Hernández, X. Zang, W. Wang, D. Wollherr, K. Kühnlenz, S. Hirche and M. Buss [1]
{drazen, rohrm, omirosk, sosnowski, dalthoff, lawitzky, moertl, rambow, vicky, medina, xueliang_zang, wangwei, dirk, kuehnlen, hirche, buss}@lsr.ei.tum.de

M. Eggers, C. Mayer, T. Kruse, A. Kirsch, M. Beetz and B. Radig [2]
{eggers, mayerc, kruset, kirsch, beetz, radig}@in.tum.de

J. Blume, A. Bannat, T. Rehrl and F. Wallhoff [3]
{blume, bannat, rehrl, wallhoff}@tum.de

T. Lorenz and A. Schubö [4]
{lorenz, schuboe}@psy.lmu.de

P. Basili and S. Glasauer [5]
{p.basili,s.glasauer}@lrz.uni-muenchen.de

C. Lenz, T. Röder, G. Panin and A. Knoll [6]
{lenz,roeder,panin,knoll}@in.tum.de

W. Maier and E. Steinbach [7]
{werner.maier, eckehard.steinbach}@tum.de

[1] Institute of Automatic Control Engineering
Department of Electrical Engineering and Information Technology
Technische Universität München
Arcisstraße 21, 80333 München

[2] Intelligent Autonomous Systems
Department of Informatics
Technische Universität München
Boltzmannstraße 3, 85748 Garching bei München

[3] Institute for Human-Machine Communication
Department of Electrical Engineering and Information Technology
Technische Universität München
Arcisstraße 21, 80333 München

[4] Experimental Psychology Unit
Department of Psychology
Ludwig-Maximilians-Universität München
Leopoldstraße 13, 80802 München

[5] Center for Sensorimotor Research
Clinical Neurosciences and Department of Neurology
Ludwig-Maximilians-Universität München
Marchionistraße 23, 81377 München

[6] Robotics and Embedded Systems
Department of Informatics
Technische Universität München
Boltzmannstraße 3, 85748 Garching bei München

[7] Institute for Media Technology
Department of Electrical Engineering and Information Technology
Technische Universität München
Arcisstraße 21, 80333 München

Contents

1 Introduction

The Multi Joint Action demonstration scenario is one of the demonstrators in CoTeSys – Cognition for Technical Systems cluster of excellence. It serves as a showcase and focusing point for the research work inside the cluster that deals with or is connected to the area of multi joint action.

This report present a summary of the current work done inside the demonstrator, focusing on the description of the experimental setup and an overview of challenges in the specific topics that are being studied.

The rest of this section gives a brief introduction to multi joint action, as well as a description of two scenarios that serve as motivating examples for the research. Section 2 describes the details of the hardware setup, whereas section 3 presents the research challenges and preliminary results.

1.1 Multi joint action

The research area of multi joint action is concerned with actions in which multiple cognitive systems take part. In our case we are interested in actions that change the physical state of the world, where by cognitive systems we mean primarily humans and robots, with the final goal being the study of mixed human-robot teams.

Due to the presence of multiple actors, and interdependence and mutual influence of their actions, it is necessary to take a different perspective on multi joint action compared to actions taken by a single cognitive system. The distinctive cognitive aspects of multi joint action can be defined as follows:

- Shared knowledge
  - Perception
  - Communication
- Coordination
  - Self-awareness
  - Team-awareness
- Joint execution
  - Predictability
  - Adaptation

The first aspect stands for the part of the knowledge about the joint action that is shared, i.e. existing on all involved cognitive systems. The channels for acquisition and sharing of this knowledge are perception, which is similar to the single system case, and direct sharing through communication, which is typical for joint action. In the planning and coordination of actions inside a team important roles are played by the awareness of own capabilities, as well as awareness of the team partners' capabilities, actions, intentions, habits, etc. The last aspect defines the direct influences during joint execution: predictability is concerned both with the prediction of actions of the partner, as well as with the performing in a way that is predictable to the partner; adaptation includes the changes in own actions as immediate response to detected characteristics of the team or changes thereof.

The presented view on multi joint actions matches well the CoTeSys definition of cognitive system architecture, which is illustrated with the perception-cognition-action (PCA) loop, see Figure 1. The figure displays multi joint action as a connection of two PCA loops, one belonging to the considered system and the other representing the other system(s) in the team. As shown, in multi joint actions there exist direct connections between the knowledge, control, and action parts of the team members, which can be associated to the aspects of shared knowledge, coordination and joint execution, in the way they were described above. Note that for perception and learning / reasoning there is no direct connection between different cognitive systems: here the mutual influence is mostly indirect.

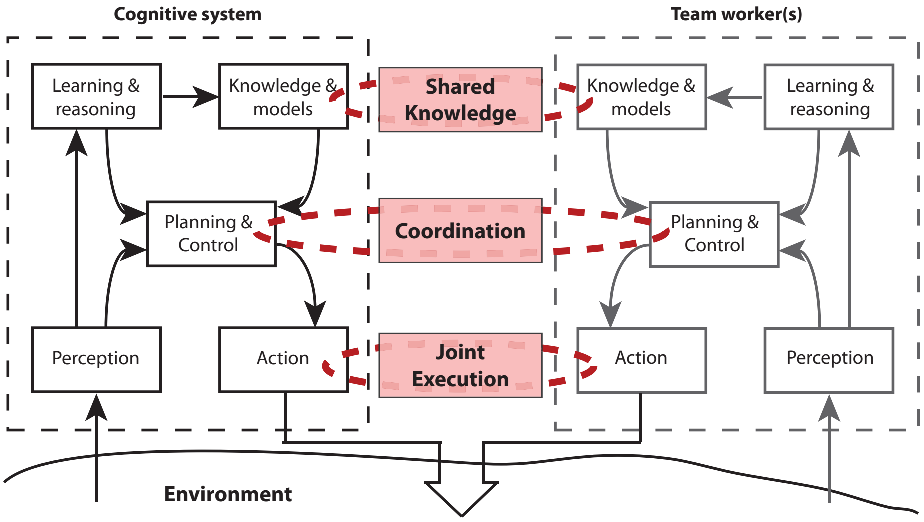


Figure 1: Multi joint action represented through the connection of perception-cognition-action loops of two (or more) separate cognitive systems.

Parallels to the presented framework for multi joint action can be found in definitions presented in the literature on teamwork, human-agent actions, and joint actions. For example, Sycara and Sukthankar in [68] give as important facets of human-agent interaction: team knowledge, mutual predictability, and directability (role assignment) and adaptation. In addition they also mention communication as another key aspect of teamwork. Sebanz et al. [60] give a summary of the psychological research of human-human joint actions and state that "successful joint action depends on the abilities: (i) to share representations, (ii) to predict actions, and (iii) to integrate predicted effects of own and others actions". In their description of the ATOM model of teamwork, Smith-Jentsch et al. [65] argue that besides the individual competence in domain-specic tasks, team members must have domain independent team expertise comprised of four different categories: information exchange, communication, supporting behavior, and team initiative/leadership. For all these definitions it is easy to observe the similarity with the above framework.

All described multi joint action aspects are present, albeit to varying degree, in the current research done in the Multi Joint Action demonstrator. The details are described in Section 3.

1.2 Motivating scenarios

Coffee Break This is a service scenario, where a team of robots providing services to human users is considered. The basic setting of the scenario is a typical coffee break, in which multiple robots are working as waiters, whereas humans are guests that can make orders for various beverages.

In a typical scene the robots wait until they detect a human who wants to be served. Upon detection one of the robots approaches the human and initiates a dialog. After finding out the human's wish, the robot team proceeds with the serving, coordinating their actions and sharing the work when necessary.

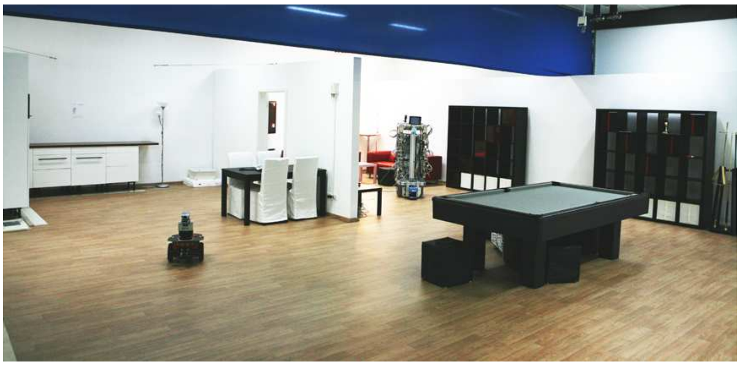


Figure 2: Experimental setup in the CoTeSys Central Robotics Laboratory .

Several capabilities play an important role for successful execution of this scenario. As discussed also in Section 1.1, knowledge needs to be acquired both through perception, as well as from communication with humans, such as in the case of an ordering dialog. Different research works in the demonstration scenario dealing with these topics are presented in Section 3.1, including multi-camera perception, multi-modal human-robot interaction, and expression recognition and generation. Next, inter-robot cooperation (Section 3.2.1) is a prerequisite for efficient operation in a team. Finally, appropriate robot action execution in a dynamic and human-populated environment is necessary, especially during navigation and handover, see Sections 3.3.1, 3.3.2, and 3.3.3.

Moving-In The second scenario considers the coordinated work in the occasion of a move into a new office or apartment. As opposed to the Coffee Break scenario where robots needed to do joint work only between themselves, here they act as members of a team consisting of both humans and robots. Therefore, apart from the topics mentioned for the first scenario, here there is also a strong focus on joint actions in such mixed teams.

Different aspects of human-robot joint action are analyzed in this setting. One the one hand it is important that the robots coordinate their actions with humans, which is why humans need to be included in the planning. Our approach is described in Section 3.2.2. On the other hand during action execution one can examine the joint performance in repetitive tasks with shared workspace, like in the case of piling up of objects or taking them from a heap, see Section 3.3.4. Another important aspect is physical cooperation between humans and robots, like for example when jointly carrying large furniture, where a haptic coupling between them exists – Section 3.3.5.

2 Experimental Setup

2.1 Living area mockup and equipment

Figure 2 shows the demonstration area which has been set up in the CoTeSys Central Robotics Laboratory – CCRL. It includes two mockup spaces: a kitchen and a living room.

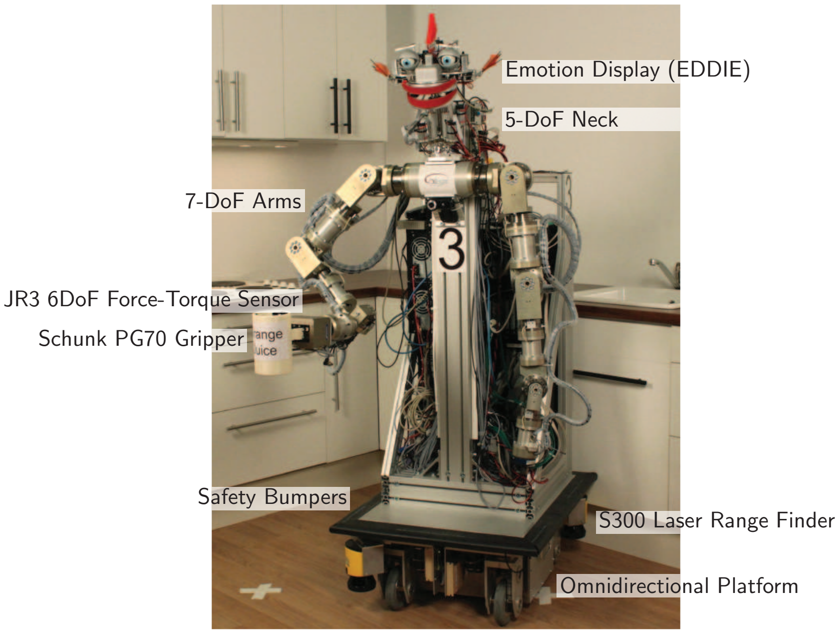


Figure 3: Robot hardware overview.

The experimental setup consists of several robots and a cluster of overhead cameras covering the complete operating area. These will be described in the following sections.

Additional equipment used in the experiments includes Visualeyez$^{\text{TM}}$ VZ4000 motion measurement and tracking system from PhoeniX Technologies Inc.[1], which covers part of the area. It enables precise tracking of active infrared markers, so it is utilized for determining the position of the robots, humans and other objects and for calibration of other equipment.

Furthermore an array of several laser range finders (Sick LMS200) can be setup to cover the whole operation area and used in addition to the ceiling cameras for tracking of humans and other objects in the environment.

For precise tracking of hand/arm trajectories in part of the experiments a Polhemus Liberty 240/8 system is used[2]. This is a magnetic motion tracker with a six degree-of-freedom range (X, Y and Z coordinates and the three rotation angles) that tracks with a constant sampling rate of 240 Hz.

2.2 CCRL robots

Our robots are intended to closely interact with humans and with each other in an everyday human household environment. They all have the same base configuration, but in order to cope with varying demands parts of their configuration as end-effectors or specialized sensors differ or can be easily exchanged. Figure 3 shows one of the robots used in the demonstration scenario.

Each robots' main chassis provides space for up to six 19-inch rack-mount

[1] http://www.ptiphoenix.com/
[2] http://www.polhemus.com/

cases. Here the battery trays, the manipulator amplifiers and the computational hardware for program execution are mounted.

Sufficient battery capacity for long autonomous runs under full load is provided by lithium-ion polymer batteries, which have been chosen for their high energy density and provide a total of 252 Ah at 52 V. A battery surveillance keeps track of the charge condition of the individual battery packs in order to provide status information – to humans or the robot – or give an acoustic signal in case a battery pack is damaged or the batteries have to be charged.

Each robot is equipped with three PCs,whereby one computer is responsible for the real-time actuator control, one for visual sensor data processing, and the third one for processing the data of the remaining sensors, communication with other robots and the environment, and planning. The PCs on a robot are connected via LAN and a WLAN switch provides communication to other robots and global sensors.

The main chassis is placed on a four-wheeled omnidirectional mobile platform [26], which offers human-like maneuverability and smooth motions. The platform can carry a payload of up to 200 kg and drives at a maximum velocity of 1.5 m/s.

Two identical anthropomorphic 7 degrees-of-freedom (DoF) arms in a mirrored configuration are front-mounted on the top of the main chassis to provide a human-like working space [67]. The arms are able to carry up to 7 kg of static load each and can be equipped with different end effectors that can vary in design and functionality. In Figure 3 a Schunk PG70 two-finger gripper is attached to the right arm. Additionally we use Schunk PVG two-finger grippers and a three-fingered BarrettHand for dexterous manipulation. Both arms are equipped with JR3 6 DoF force-torque sensors positioned before the end-effectors.

For environment sensing the robot uses two Sick S300 laser range finders. They have a 270° field of view and are placed in opposing corners of the chassis what allows circumferential planar obstacle detection during navigation. Furthermore it is possible to mount a small laser range finder (e.g. Hokuyo URG-04LX) or a web camera before the end effector for use in visual servoing and area sensing.

On top of one robot's body there is the emotion display head EDDIE [66]. The head is a 23 DoF device mounted on a 5 DoF actuated neck which allows dynamic and intuitive expression of the robot's emotional state and social gaze (see Section 3.1.2). EDDIE's facial features are derived from a mixture of anthropomorphic and zoomorphic features. EDDIE is also equipped with speech processing and synthesis what results in a very natural human-robot interaction. Furthermore, it contains a pair of Firewire cameras mounted in the eyes, which are used as additional sensors for face and gesture recognition (see Section 3.1.1) or visual servoing. The other robots are equipped with pan-tilt units with mounted stereo camera heads and additional sensors according to needs.

For safety of hardware and humans each robot is equipped with an emergency system switching off the power supply to all actuators in case of emergency. The system can be triggered by a human via a wireless emergency button. Furthermore bumpers cover the mobile platform and activate the emergency shut-down when coming into touch with environmental obstacles. For additional security during the interaction with humans we periodically charge the complete

Camera Type	Qty.	FPS	Resolution
Baumer TXG08c	30	28	1024 × 768 px
Basler Scout scA1000-30gm/gc	10	30	1024 × 768 px

Table 1: Specifications of the cameras used in the setup.

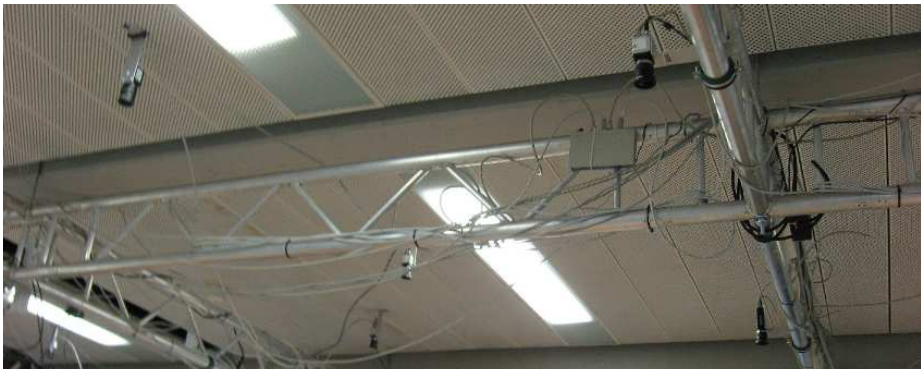


Figure 4: An impression of the camera setup covering the experimental area.

robot and measure changes in the capacitance, which allows for detection of unintentional contact of the robot with humans. The latter can be switched on and off via software.

2.3 Ceiling camera setup

As mentioned before, our experimental setup includes a networked cluster of optical sensors to address global perception tasks. To achieve a redundant survey of the scene, 40 ethernet-connected cameras were installed on a metal scaffolding at ceiling height, approximately 3.5 m above the floor level. Figure 4 provides an impression of the camera setup. The cameras' fields of view (FOV) cover the whole area from a top-down view. They were set up to achieve a coverage redundancy of approximately 75%, which is measured at a height of 1.7 m. This height was chosen since it is the approximate average height of an adult person, compare [52].

The cameras used in the setup provide images of 1024×768 pixels at a rate of 28 to 30 frames per second, for details see Table 1[3]. Image acquisition occurs asynchronously, using the GigE-Vision standard described in [12]. The cameras are grouped in threes and pairs respectively to form 14 camera groups, each of which is in turn linked via a Gigabit Ethernet (GigE) switch to one of 14 diskless client nodes [35], where image acquisition and processing take place. Cameras with adjacent FOVs are assigned to different camera groups. This helps to compensate for the observed fact that human beings in social scenarios such as the coffee-break demonstration scenario tend to flock together, rather than distribute evenly over the surveyed area. Furthermore, this procedure reduces

[3] For further information on the cameras refer to `http://www.baumer.com/` and `http://www.baslerweb.com/`

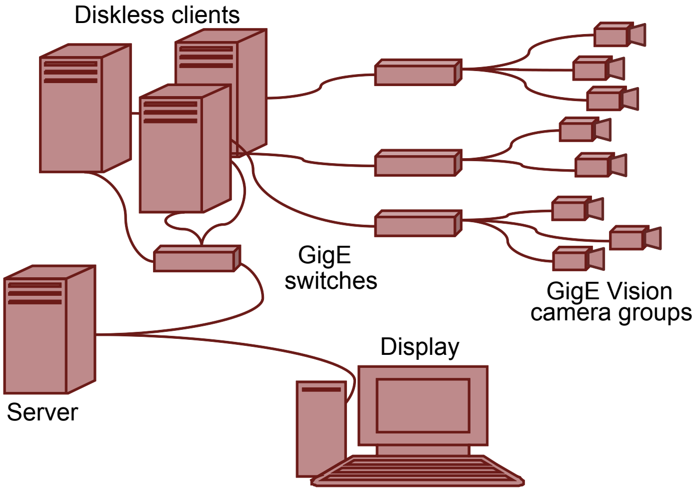


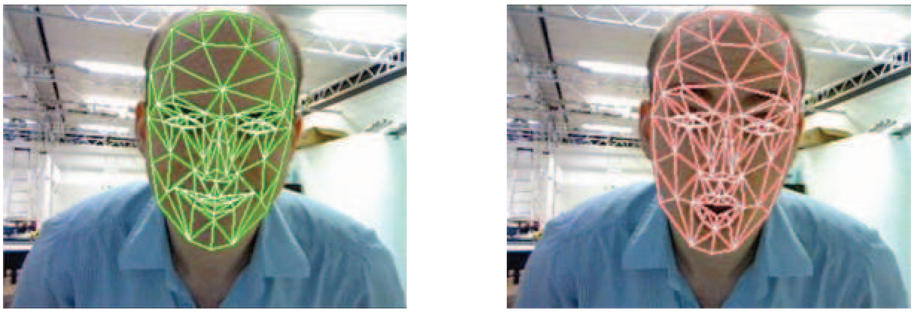
Figure 5: Schematic of the hardware and network setup for the experimental area survey.

the likelihood of adjacent cameras becoming unavailable simultaneously in case of problems caused by single processing nodes, thus improving system robustness and load-balancing between the image processing nodes.

The diskless client nodes consist of general purpose hardware and are powered by AMD Phenom^{TM} Quad-Core processors, operated by Debian Linux 5.0. The computers are equipped with two GigE network adapters each. One of these adapters connects the node to the local camera group network, while the other one connects to the client network via an HP ProCurve 2848[4] 48-port GigE switch. Figure 5 provides a comprehensive overview of the network component setup used in the scenario. The RTDB, as described below, is used for interprocess communication on the diskless nodes, which makes the images available to potential subscribers and facilitates the process of buffering and synchronizing the image data. Image processing libraries used are HALCON 9.0 [13] and OpenCV 2.0 [10].

In addition to the hardware described above, for we make use of a software framework for interprocess communication based on the Real-Time DataBase (RTDB) [24]. The RTDB offers functionality similar to a shared memory, while allowing maintenance of hierarchical data objects and providing real-time guarantees for management and exchange of data structures inside one computer. For information exchange between computers on the robots, as well as other hardware in the system, the ZeroC/ICE[5] middleware is utilized. More details on the communication framework and its use in the scenario can be found in [3] and [22].

[4] http://www.procurve.com/
[5] http://www.zeroc.com/



(a) Happiness

(b) Surprise

Figure 6: The Candide-3 face model fit to images of a human face displaying different basic emotions, where the model coloring intensity is used to visualize the intensity of the facial expression.

3 Challenges in Multi Joint Action

3.1 Shared knowledge

3.1.1 Communication: facial expression recognition

Everyday human communication relies on a large number of different communication mechanisms, like spoken language, facial expressions, body pose and gestures. This allows humans to pass large amounts of information in short time. In contrast, traditional human-machine communication is often unintuitive, and requires specifically trained personal. To complement human-robot communication, we establish a real-time capable framework that recognizes traditional visual human communication signals: head gestures (nodding and shaking one's head) are a convenient way to signal agreement or disagreement, facial expressions give evidence about the interaction partners' emotional states and hand gestures are a fast way of passing simple commands.

Head gestures are commonly used to signal agreement or disagreement [49]. Nonetheless, it is also imaginable to use head gestures for controlling operations, like document browsing [48]. Ekman and Friesen find six universal facial expressions that are expressed and interpreted independently of the cultural background, age or country of origin [14]. The Facial Action Coding System (FACS) precisely describes the muscle activity within a human face as it occurs during the display of facial expressions [15].

We follow a three-step approach: First, preprocessing of the data supports subsequent steps. Second, features are extracted from the raw image data. Third, high level information is obtained from these features. The raw image data, the extracted image features and the determined high level information are stored and buffered in the RTDB. Since there are no dependencies within a single work step, there is possibility for parallelization. For instance, the feature extraction for the recognition of head gestures is independent of the feature extraction for hand gestures and therefore both feature sets are extracted in parallel.

Preprocessing: We apply an implementation of the object detection approach published by Viola et al. [74] to detect human faces in the video images obtained from the RTDB memory. Since this information is required by multiple modules, it is also made available to them in the RTDB. We adapt a skin color model with regard to the obtained face image; please refer to [44] for further details on the approach.

Feature Extraction: To constrain the hand gesture recognition, we define a region of gesture action (ROGA), which is located on the right side of the interaction partner's head, and extract Hu moments [27] from the ROGA in the skin color image. To obtain information about face pose and shape, we integrate the Candide-III face model and rely on the work of Wimmer et al. for model fitting [1, 75]. To gain descriptive information about head gestures, the temporal changes of the in-plane transition of the face and the temporal change of the three rotation angles (pitch, yaw and roll) are extracted. Face shape is considered by static features, which are assembled by calculating the model parameters for a single image, and person-adapted features, which are calculated as the model parameter change between a neutral reference image of the person visible in the image sequence and the current image.

Classification: Classifiers are obtained using machine-learning techniques. We apply continuous Hidden Markov Models with a left-right structure for the classification of head and hand gestures. The HMM is presented with sequences of varying length by applying different sliding window sizes on the feature data buffered by the RTDB. Training classifiers for facial expression recognition usually relies on standard databases. For this purpouse, we integrate the Cohn-Kanade Facial Expression Database [34] and the MMI Face Database [53]. We apply decision trees and SVMs to map the feature vector obtained in the previous step to the corresponding facial expression.

Due to the utilization of the RTDB for communication, the data produced in each step is easily accessible. The recognition of facial expressions is combined with the emotion display EDDIE (see Section 3.1.2) to establish a closed-loop human-machine interaction scenario for demonstration. In the preliminary version, EDDIE is mirroring the facial expression recognized from the human interactant. However, it is intended to serve as a basis for more complex interaction scenarios.

3.1.2 Communication: robot emotions

Non-verbal communication and emotions play a key role in social scenarios, providing an intuitive human-machine interface that increases efficiency in working with the robot and requires no user training. The user perceives internal states of the robot in a simple and robust way, and the robot can recognize mental and/or emotional states of a user. For example, recognition of the intention of a user by the robot benefits from this, because our decision process is heavily influenced by emotions [8, 9, 42]. On the other hand, including emotions in the robot architecture can improve the robot's own decision process and create the possibility for adaptive human-robot interaction that accounts for the human emotional state, enabling the robot to react appropriately in a socially accepted context.

An effort to model and integrate emotions into robots has been made for many years. But despite the fact that many models have been proposed and several agents have been developed, there is still no out-of-the-box solution for implementing emotions on artificial intelligent systems. This is definitely due to the immense complexity that emotions and cognitive systems induce. However, even simpler parts that can be isolated, like expression of emergent emotions, are still not available as ready to use solutions. Referring to Mehrabian's 7%-38%-55% rule [45], non-verbal communication dominates the eliciting of emotional context in human-human interaction, so using non-verbal channels is of great importance. For implementations in agents, mainly two emotion representations are used.

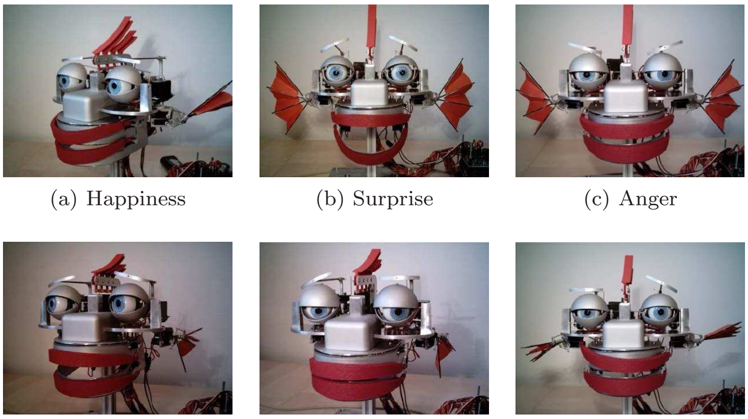
One is a discrete set of emotional states, quite often based on the basic emotions found by Ekman and Friesen [16] or Izard [31]. They are convenient, because they reduce the complexity of emergent emotions and expressions to a limited number of states, which are recognized independent of the cultural background of an observer.

The other form is a variety of mental/emotional state spaces. They allow a continuous representation either in a two-dimensional approach, for example Russell's circumplex model of affect with valence as axis of abscissae and arousal as axis of ordinates [58], or a three-dimensional version. Although there exist several three-dimensional models, the most noticeable are based on Russell's 2D model and extended by another orthogonal axis. The PAD model introduces a dominance axis (pleasure-arousal-dominance). The dynamic mental model [46, 47] has a certainty axis, with the certainty being either +1 or -1. In this case, we have two pleasure-activation planes. The mental vector, beginning in the origin and ending on one of the planes, describes the actual mental state. Breazeal extended the 2D model by adding a stance axis, with the 6 basic emotions on the corresponding axes, forming the state space [11]. Some of these representations are often closely related to the basic emotions, because they are divided into regions which are labeled according to the basic emotions. For instance, the dynamic mental model is divided into 7 regions, the pleasure-arousal-stance space is defined by the basic emotions according to Ekman.

Facial expression is the best researched non-verbal communication channel. The human face as primary indicator for emotional states is the most complex and versatile face in a species [11] and capable of generating up to 7000 different facial expressions [7]. The various expressions are a result of the interaction between different layers of skin and several muscles and muscle groups. To be able to make an objective distinction of the observable changes in the facial appearance, the facial action coding system (FACS) by Ekman and Friesen [16] is used. This theory defines so called action units, which are responsible for certain changes in the face. Actuating a robot face can be based on these action units. This is much simpler and more effective than rebuilding the whole facial muscle system.

For the facial expression synthesis, the robot head EDDIE, as described in Section 2.2, is used. With this head, 13 out of the 21 emFACS action units can be displayed (emFACS is a subset of the facial action coding system, including only action units which are involved in emotional facial expressions). The emotional state displayed can be controlled in two ways, referring to either of the two models described above. Using a state-space to joint-space mapping, EDDIE allows for arbitrary facial expressions with smooth transitions. For details we refer to [66]. The expression of the six emotional states can be seen in Figure 7.

The head is linked via the RTDB to the other systems, providing in-eye camera images for the visual processing and visualization of emotional and internal states (see Section 3.1.1) as well as visemes for the speech output.



(a) Happiness (b) Surprise (c) Anger

(d) Anxiety (e) Sadness (f) Disgust

Figure 7: EDDIE displaying the basic facial expressions proposed by Ekman et al. [14].

3.1.3 Communication: dialog engine

The main goal of the dialog engine is to make the interaction between the human and the robot more natural and human like. Therefore, the dialog engine is orchestrating the involved modules using a multi-modal communication backbone [22] to obtain a desired information from the human. In our motivating example scenario (c.f. Section 1.2), this could be to find out what drink a human would like to have. The dialog relevant modules for our Coffee Break scenario are depicted in Figure 8. The dialog engine is triggered by the controller of the robotic platform, when the robot is close enough to a new customer (e.g. detected via the ceiling cameras see Section 2.3) to start the order dialog. In the first step, the head and eyes of the robotic head EDDIE (see Section 3.1.2) are directed towards the customer and asks what drink the customer would like to have. Therefore, this question phrase is sent to the text-to-speech (TTS) module and the head controller, to make synchronous lip movements to the generated speech output.

As next step, the robot has to listen to the customer's order. Therefore, a speech recognition grammar is reduced to the current context of the dialog (here: ordering a drink). Included in this grammar are all available drinks, which are queried from a knowledge database containing all available drinks. With this reduced grammar the automatic speech recognition (ASR) module is configured to allow a more robust and remote speech recognition. At the same time the ears of the robotic head are widened to inform the user, that the system is now ready to receive speech input. After a speech command was perceived, the robot repeats the perceived phrase and asks the customer to confirm the utterance. In addition to speech, head gestures [21] (for more details, see Section 3.1.1) can be applied to confirm or cancel the received command. Basing on the positive or negative response of the user, the robotic head reflects the answer by showing happiness or sadness.

After the successful order, the desired drink is reported to the controller of the robotic platform to initiate the fetch-sequence. Meanwhile, the knowledge database is updated by retracting one instance of the ordered drink. These steps

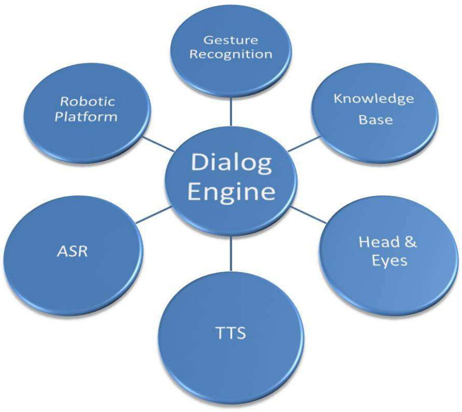


Figure 8: Schematic view of dialog engine and involved modules

complete the dialog order process and now the robotic platform controller is in charge of locating, fetching and handing over the drink. During the fetch process the dialog engine remains in a slave-mode to enable error or status reports of the robotic platform controller.

The dialog engine is represented using first-order logic. Tasks to solve are represented by a predicate of first-order logic with variables representing information needed to be determined during the dialog. Equivalence rules on these predicates are specified to navigate through the dialog by splitting a task into several subtasks. For the Coffee Break scenario explained above, the dialog engine tries to evaluate the predicate $orderDialog(P, D)$ with P being the person to be served and D being the drink. Therefore, the dialog engine inspects the equivalences of the system and determines a rule replacing the $orderDialog(P, D)$ predicate with $isIdentified(P)$ & $isOrdered(D)$ & $isConfirmed(D)$. In the next step, the dialog engine will determine the truth value $isIdentified(P)$ by inspecting the binding of P. At this point, interaction with the environment is required to determine what person is interacting with the robot and the predicate $isIdentified(P)$ evaluates to `true` at the same time determining the interaction partner by the binding of P. In this case, the predicate $isIdentified(P)$ is a C++ extension of the first-order logic interpreter, which communicates with image processing modules and returns the identified person as binding of P. In a similar manner, $isOrdered(D)$ is evaluated by binding the variable D to some name of a drink from the speech recognition system. Finally, the value of $isConfirmed(D)$ is determined from either speech recognition or head gesture recognition to confirm the acquired drink.

Note, that this first order logic representation together with the C++ extension allows to integrate further sensory modules that perceive information about the environment. Thereby, all modules report their information in a unified way as predicate, which enables higher-level processing of the multi-modal data.

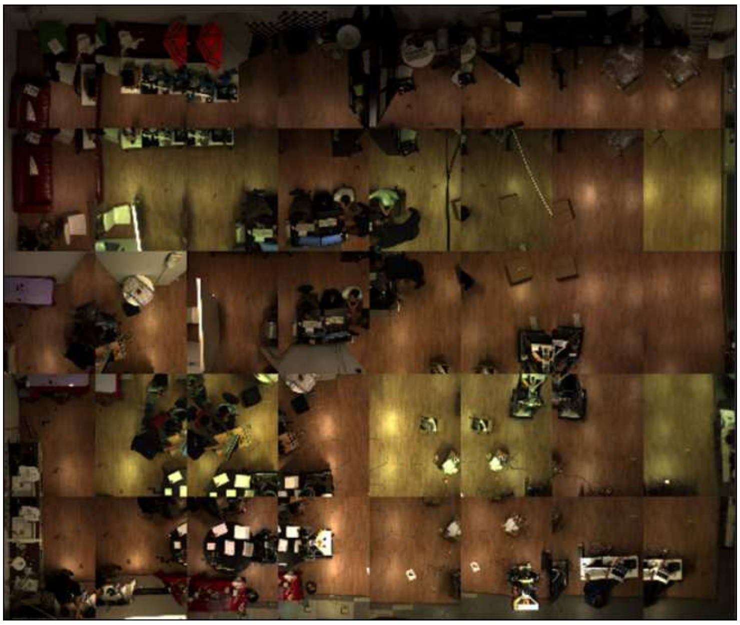


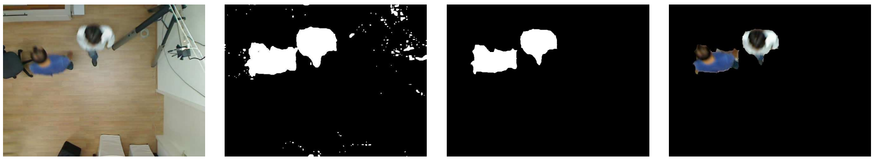
Figure 9: Tiled joint view of the 40 cameras observing the apartment mockup. Different illumination conditions are caused by object shadows, while variant chromatic appearances are caused by using different camera models. Specular reflections result from light sources above the ceiling-mounted cameras.

3.1.4 Perception: ceiling camera surveillance and visual tracking

A variety of perception tasks has to be addressed by the camera cluster described in Section 2.3. A common quality in all these tasks is that they benefit from a total survey of the scene to be executed effectively. To allow the robots to approach specific persons for interaction, humans in the scenario have to be detected and tracked across the whole apartment in real time, without confusing their identities in the process. To allow robots to plan and execute the manipulation of objects, viable candidates, such as tools or containers, have to be detected and identified. For robot movement planning, the experimental space has to be segmented into traversable and obstructed areas by obstacle detection and floor segmentation.

Intelligent camera surveillance is employed commonly both for security purposes as well as for smart rooms, which can autonomously act on perceived situations. Surveillance systems can operate both in real-time or focus on the post-processing of previously acquired video data. The state of the art for that kind of visual surveillance systems is is described in several surveys, such as Valera et al. [71] (with an emphasis on distributed systems) or Šegvić et al. [61]. However, for a large-scale many camera system using high-resolution images while operating in real-time, precedents prove to be hard to find.

Since the camera system is designed to cover the whole area, challenges start with the scope of the system which has to be designed and integrated. One can easily compute from the figures in Table 1 that at any single moment the full-size combined image for all cameras would measure 5120×6144 pixels, while the



(a) Original camera frame (b) Foreground and background segmentation (c) After applying morphological operations (d) Static pixel-wise fused result

Figure 10: Segmentation of foreground pixel regions that are likely to correspond to people. For this basic detector, a background subtraction procedure in combination with a morphological closing operation was used.

total data rate generated by the cameras amounts to approximately 7.6 Gbps. Since this exceeds the capacity of a GigE adapter by far, it is not an option to transfer the image data to a single computer. Consequently, image processing has to be distributed. This incurs challenges regarding the integration of data over all the processing nodes maintaining the cameras, such as the real-time exchange of extracted features to track persons and objects. Figure 9 provides an impression of exemplary image data on all 40 cameras at a single point in time.

Apart from the circumstances of the distributed setup, individual computer vision tasks retain their inherent challenges. Applicable approaches are limited for the scenario by real-time and stereo coverage constraints. Some popular approaches, like 3D reconstruction [55], cannot be realized under these conditions. Instead, we opt for algorithms operating on monocular images only, and integrate our data on the positional level rather than the image level. While the largest single challenge is posed by the real-time constraints for the algorithms, dynamic lighting conditions remain a challenge for all vision tasks, ranging from background subtraction and floor segmentation to person tracking and body-part identification for gesture recognition. Precise intrinsic and extrinsic calibration of all involved sensors is required to allow a consistent world model incorporating all cameras, and has proven to be a challenging task on its own. To facilitate this process, the Visualeyez infrared tracker (see Section 2.1) are used to accurately determine the pose of a calibration table with fiducials attached anywhere in the experimental area. Subsequently, the determined pose is used to perform the extrinsic calibration of the cameras.

As mentioned before, visual person tracking is a main focus in the scenario, and a prerequisite for further analysis of participants' actions. The visual tracking process itself typically consists of two distinct phases. In the detection phase, the object to track is detected and located within a single image, using techniques such as foreground segmentation or feature extraction, e.g. the Viola-Jones [74] algorithm. A summary of the person detection technique used is depicted in Figure 10. In the subsequent tracking phase, the trajectory of this object is established in successive images using a predictive algorithm, a process considerably less expensive than the previous detection. For a comprehensive survey on algorithms to be used during the tracking phase we refer to Elfring et al. [17], and for a comprehensive survey of visual tracking methods see Yilmaz et al. [76].

For such a tracking approach to be implemented on a distributed multi-camera system efficiently, exchange of world position and tracked features between the involved processing clients has to be dealt with to avoid repeated detection phases, and thus improve the performance of the system beyond the one of the sum of its parts. Unless a tracking-by-detection approach [4] is employed, a robust visual tracking algorithm relies on spatio-temporal consistency of the images depicting the object to track. Consequently, the challenge of tracking an object increases in a real-time environment, where spatio-temporal consistency of the image material is limited by the processing speed of the tracking system in addition to the camera frame rate. Ergo, the most challenging parts within the tracking process are those which require the most processing, i.e. the detection phase. With a distributed camera setup such as the one described in Section 2.3, spatio-temporal consistency is violated for targets switching camera FOVs. Currently, we approach this by initiating another detection phase for the object to track within the corresponding image. From an image processing point-of-view, this is the greatest challenge for person tracking in the distributed setup. For a more detailed description of our approach, along with quantitative results on tracking accuracy, we refer to Lenz et al. [41].

3.1.5 Perception: probabilistic appearance representations and their application to surprise detection

An environment where multiple robots interact with multiple humans is highly dynamic and changes over time. These changes may include the sudden and unexpected appearance of objects which are unknown to the robots. New objects which are relevant for the robots' tasks have to be represented in their internal environment model. In general, it is challenging to acquire accurate geometric models of transparent and glossy objects. Image-based modeling approaches, in contrast, provide realistic appearance representations of real-world environments [62]. It was shown that surprise attracts the attention of humans and is already generated at very early stages of visual information processing [30] if the perceived stimuli do not match the expectation inferred from past stimuli data. Hence, robots need similar mechanisms for attentional control in order to trigger learning processes. To this end, we develop algorithms which provide a quantitative measure for the level of surprise for each pixel of an image captured by the robot's camera. These surprise maps can be used in order to segment new objects from the familiar background which is represented in the internal model. In this context it is challenging to achieve a robust segmentation if the estimated pose of the robot's camera is inaccurate and thus the robot's camera image cannot be registered well to the internal representation.

In order to overcome this issue, we base the computation of surprise on a probabilistic appearance representation for the robot's environment. The appearance of the environment is represented as a dense series of viewpoints. For each viewpoint, the luminance and chrominance at each pixel are modeled by a Gaussian distribution, as illustrated in Figure 11. The mean of the Gaussian distribution is the expected luminance or chrominance at a given pixel and the variance expresses the uncertainty of the robot about the environment's appearance. A joint probability distribution over the mean and the precision (reciprocal variance) is obtained from the robot's past observations by Bayesian inference. In addition, a per-pixel depth map and the pose of the robot's camera head is stored for each viewpoint.

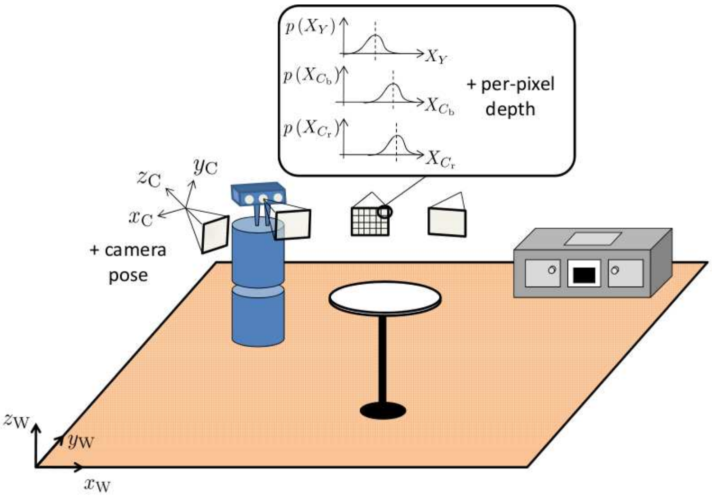


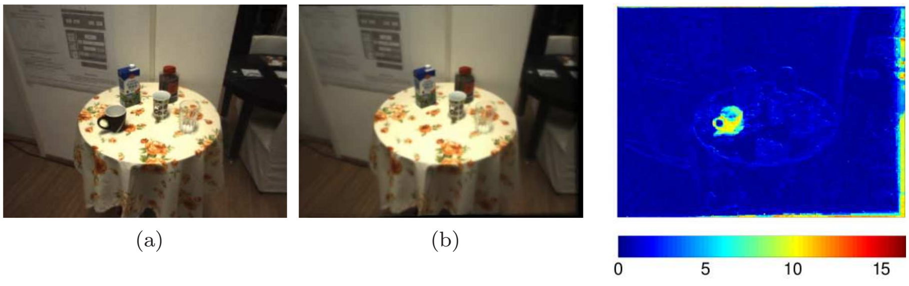
Figure 11: Our proposed appearance representation uses Gaussian models for the luminance and the chrominance of the environment for each pixel at a viewpoint. The Gaussian distributions are inferred from observations along the robot's trajectory. The representation also includes a depth map and the pose of the robot's camera head for each viewpoint.

When a robot captures a new camera image at a given viewpoint, the internal environment model provides prior knowledge about the appearance of the scene. Since the robot never exactly returns to a previous viewpoint, the probability distributions stored in the internal representation have to be interpolated in order to be used as priors for the pixels of the image acquired at the robot's current viewpoint.

The luminance or chrominance value at a pixel in the new image updates the interpolated prior distribution and provides a posterior distribution over the parameters of the Gaussian model. The posterior distribution is different from the interpolated prior distribution in image regions where the environment has changed. Hence, a quantitative measure for the surprise level at a pixel in the new observation is provided by the Kullback-Leibler divergence between the posterior and the interpolated prior distribution. The interpolation of the prior distribution and the computation of the Kullback-Leibler divergence are done in real-time on the robot's graphics hardware [43].

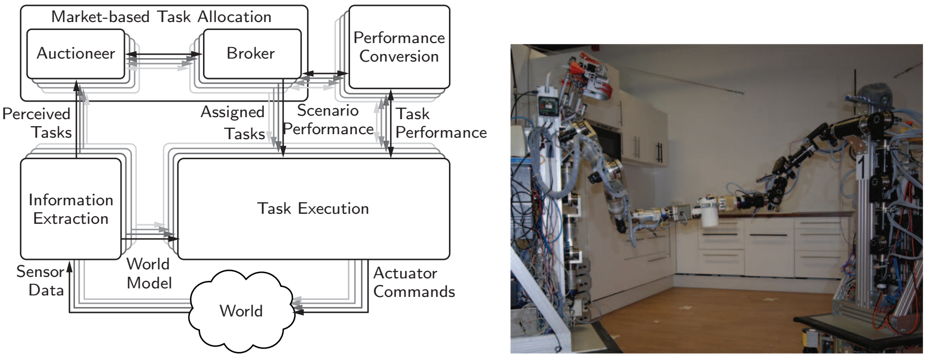
Figure 12 shows a camera image captured by the robot (Figure 12a), together with an image which shows the robot's expected appearance at this viewpoint (Figure 12b). The expected appearance is computed from the robot's internal environment representation. The surprise map in Figure 12c shows high surprise values in the region of the new cup whereas the surprise values are relatively low in the rest of the map.

In future work, we will use our method for the computation of surprise in order to selectively extract visual features from new objects and to build object databases which can be queried later for recognition. Whereas common approaches require that new objects be presented in front of a uniform background, we suppose that our method allows for the segmentation of new objects



(a) (b) (c)

Figure 12: (a) Image captured by the robot's camera. (b) The robot's expected appearance. (c) The surprise map indicates high Kullback-Leibler distance values in the region of the new cup.



(a) The multi-robot framework. Layers with different shades of gray denote instances on different robotic agents in the system.

(b) Cooperative robot-robot handover.

Figure 13: The multi-robot framework and an application example.

in front of arbitrary backgrounds.

3.2 Coordination

3.2.1 Multi-robot task allocation and execution

Cooperative human-robot and robot-robot interaction in complex robotic systems requires distributed decision making and cooperative execution. Furthermore, these two parts have to be efficiently connected to yield in a consistent performance optimization. We address these issues with the cooperation framework illustrated in Figure 13a. It shows the main modules and the interaction between them. The Task Allocation is responsible for negotiating and distributing the tasks among the robots, whereas the Task Execution takes care of both sequential and parallel execution and the supervision of tasks on one robot. For optimal operation a global performance metric has to be maximized. As the local task performance metrics defined by the Task Execution might be different from the global one, a performance metrics conversion procedure is included.

For the task allocation we use a market-based approach that follows the idea of economical markets [33], and provides a good trade-off between a fully decentralized system design, which allows only suboptimal solutions, and a fully centralized design, which is unfavorable in terms of complexity and due to a single point of failure. Following the taxonomy of [23], our approach deals with the instantaneous assignment of tasks (i.e. with no planning for future allocations) – whose completion may require either a single robot or multiple robots – to a set of robots which are capable of executing multiple tasks simultaneously. The assignment of a set of single-robot tasks as well as the sequential assignment of multi-robot tasks is solved optimally. The sequencing of the set of multi-robot tasks is done in a greedy manner.

While the task allocation might only have a symbolic representation of the tasks, the task execution must provide the actual realization. This is achieved by decomposing the symbolic task into a set of sub-tasks and forming a hierarchical combination of simpler tasks. The robots are assumed to be highly capable, i.e. each robot is a kinematic tree with multiple degrees of freedom and several branches that comprise the hardware resources, such as mobile platform, arm, end-effector, head etc. To take full advantage of the capabilities of the robot, the task execution is able to accommodate multiple concurrently running tasks which possibly interfere with each other. This is achieved by prioritizing tasks and combining them using null space projection.

Ongoing work focuses on the enhancement from an ability-aware system, i.e. resource management, failure recovery etc., to a capacity-aware system. With capacity-awareness we envision a self-aware system that is able to quantitively estimate and adapt its own performance online. In this respect, we utilize the probabilistic system interdependence analysis described in [57], which enables the learning of the metric interdependencies. This provides the functional relation to solve the consistent optimization problem and is encompassed in the performance conversion module in Figure 13a.

3.2.2 Planning of joint actions

At a high-level of control, a robot needs to decide what to do when. In a joint-action scenario, these decisions depend on both a global plan of what needs to be achieved as well as the current actions of humans in the vicinity.

Using the reactive planning language RPL [50], and a framework for integrated Planning and Learning [36] we describe plans in a way that allows each individual goal to be either achieved by a human or the robot. The robot will then in the partially ordered set of plan steps always strive to decide on a next task which best suits what the humans are doing.

Most of the preliminary work is done using a simulator[6], in order to be able to run a maximum of experiments without impacts of failures from perception and technical infrastructure. The results will then be validated in fewer experiments on a real robot.

In simulation, we introspect high-level interaction by having 2 agents represented in the simulator, one controlled by a human over the keyboard, and the other by a robot controller. The simulation appears as shown in Figure 14 The sequences of actions that emerge when both act to achieve common goals

[6] Gazebo: `http://playerstage.sourceforge.net/gazebo/gazebo.html` - part of the Player/Stage project

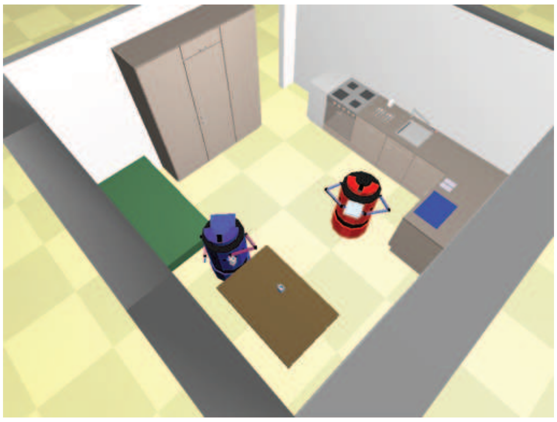


Figure 14: Example of two agents in the Gazebo simulator jointly setting table

can then be evaluated with respect to performance and the satisfaction of the human with his robot companion.

The main challenge for planning in multi joint action is different from classical planning as we are mainly focussed on deciding between several current alternatives in a given general plan rather than creating plans.

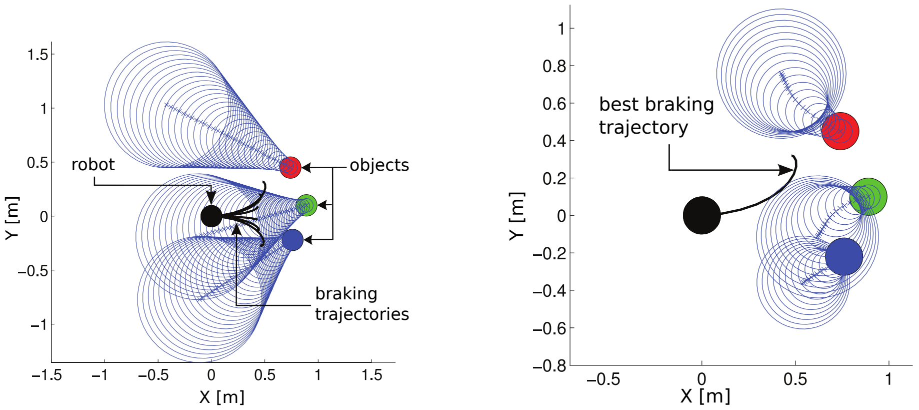
We assume a general plan exists that states the goal to be achieved, and the robots require a strategy to follow that plan while allowing for the human to take part in that plan.

We hope to find which algorithmic changes to a given robot controller yield the highest benefit for human robot interaction. As prerequisites, we depend on lower level routines of the robot such as perception of the human and human aware motion planning, without those working reliably, the experiments in multi joint action will not yield meaningful results.

3.3 Joint execution

3.3.1 Navigation in crowded environments

The major challenges for a robot navigating in human-populated areas, i.e. dynamic and often very crowded environments, are finding a valid path to its goal position and guaranteeing safety. In a static environment it is in general easy to plan a path, whose safety can then be verified by checking if the planned state trajectory does not enter the set of inevitable collision states (ICS) [20], which depends on the geometry of the obstacles and the dynamics of the robot. The concept of inevitable collision states can be extended to scenarios in which the future behavior of dynamic obstacles is exactly known, see e.g. [54]. However, in a scenario where the future behavior of other obstacles is unknown, one cannot decide if the current robot state is an ICS since it depends on the future actions of the dynamic obstacles. For this reason, the concept of ICS was extended to probabilistic scenarios [2].
In crowded environments, motion planning with ICS is not reasonable. Consider the scenario where a robot finds its way through many people in crowded environment. Since in this scenario the workspace objects are humans, their future occupancy is unknown and can only be predicted. Their motion and



(a) Blue ellipsoids depict object motion prediction ignoring robot's state and black lines representing possible braking trajectories of the robot

(b) Blue ellipsoids depict object motion prediction considering robot's state and black line represents braking trajectory with the smalles crash probability

Figure 15: Difference between active and passive workspace objects.

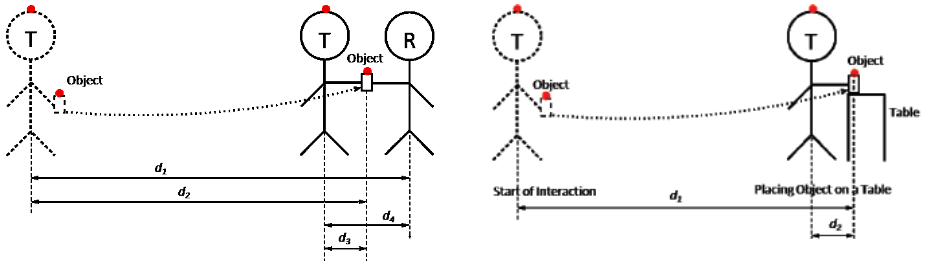
their future occupancy is specified by probability density functions.
Two different kinds of objects are considered:

- Passive objects: they *ignore* the robot's trajectory.
- Active objects: they *react* to the robot's trajectory in order to reduce the collision risk.

Modeling humans as active objects has two advantages. On the one hand it is easier to find a path to the goal position and on the other hand the calculation of the crash probability is less conservative. Figure 15 illustrates the usefulness of modeling humans as active objects. The crash probability of the robot's state could be reduced by 45% compared to passive objects. In [2] was shown that it is essential to take into account the avoidance possibilities of all workspace objects to solve the problem of navigation in crowded environments, which can be interpreted as an interactive process.

3.3.2 Intention recognition of approaching agent and dynamic robot-human handover

The human capability to interact smoothly and efficiently constitutes the reference standard for robotic agents in human-robot interaction. A promising way of achieving human-like performance is to mimic relevant features of human behavior to enable humans to intuitively react to and interact with the robot. Until now, several studies were done on human-robot interaction [5, 29, 37, 51, 59, 63] but only few of them [5, 29, 51] were based on the analysis of human-human experiments. We therefore analyzed one of the basic interaction scenarios in human daily life: one person approaches another one and hands over an item [6]. 26 subjects participated in the experiment. The delivering person approached the receiving subject from a distance of 4.2 meters. The object to be handed over and the head of the approaching subject were tracked using a motion track-



(a) Handover scenario.

(b) Table scenario.

Figure 16: Human-human handover and table scenarios. The red dots denote the tracking sensors, while T and R denote the transporting and receiving subjects, respectively.

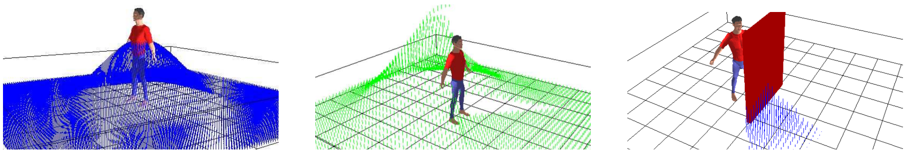
ing system (IS-600 Mark 2, InterSense Inc.[7], USA). The results showed that the approaching person started to raise her hand during the approach phase and well in advance of the handover. The approaching person was still in motion at the moment of the handover with an average head and hand speed of 30 cm/s and 12.5 cm/s, respectively. This blending of approach and handover actions has two advantages: the approaching subject can achieve a smooth and fast interaction, and the receiving subject is able to recognize the intention of the approaching subject early enough to swiftly react in order to take over the object. Thus, both the intuitive indication of intention by the transporting subject and the respective understanding by the receiving subject determine the seamless interaction.

When considering the handover position in an interaction related coordinate system with the origin being the midpoint between the interacting subjects, the handover position was close to this midpoint and slightly but significantly shifted to the right side of the receiving person (all subjects were right-handed). The same handover pattern was also observed in our previous studies, where two subjects sitting at a table were handing over a cube [28].

In order to analyze whether the behavior of the transporting subject depended on the receiving subject being present, the receiving person was replaced by a table (24 transporting subjects). A similar behavioral pattern as in the first scenario was observed suggesting that the way humans approach another person to handover an object depends mainly on their own sensorimotor contingencies rather than social factors. The only significant (t-test, $p < 0.05$) difference was found for peak hand velocity with a higher peak hand velocity in the table scenario. This is in accordance with Fitts' law [18], which states that the speed increases if the size of the final target becomes larger: handover position has to be more accurate when delivering an object to a person as when placing the same item on a table.

In summary, our results suggest that the dynamic blending of motor programs determine the efficiency of the handover interaction between two persons. However, the action of the approaching person it is not particularly affected by a receiving subject being present. The various parameters measured in our study are directly transferable to robotic systems. However, a human-like dynamical

[7] `http://www.intersense.com/`



(a) Safety costs (b) Visibility costs (c) Hidden zones costs

Figure 17: Cost functions for HANP A* search

behavior can only be achieved in robots by executing motor programs in *parallel* rather than successively.

3.3.3 Human-aware navigation for joint task achievement

Human-aware navigation planning has global and local impacts. While there is no formal separation between what is global and what is local planning, usually the boundaries depend on the available information. Global planning is based on a map of the whole environment including the current position of the robot and a target location.

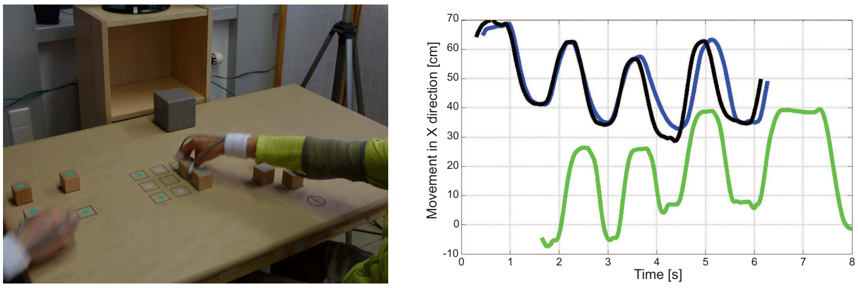
Using search algorithms on the map, motion plans such as sequences of waypoints can be found. Local planning on the other hand usually just takes a local motion target from a global plan (such as a waypoint), and also the sensor readings of the robot, to find a sequence of motoric commands that will reduce the distance to the next target.

Human presence affects both global and local planning. In global planning, human may act as obstacles in a map like furniture, however humans may be moving and they may be considered as movable obstacles. Taking into account the comfort of concerned humans using models of human discomfort, improved global motion plans can be achieved.

We introduce the human aware navigation planner HANP [64] to create global motion plans that take into account social costs of motion around humans as visualized in Figure 17. Original HANP has been developed and evaluated mostly for navigating around static humans in scenarios where the robot acts as a servant. We extended HANP to take into account moving humans and humans as cooperating agents [39].

In local planning, human presence naturally is a safety concern, a local planner must make sure no harm can be caused by a robot. Aside from that, local planning attempt to create robot motions that are legible and comfortable, in particular when approaching humans such as to hand over items or communicate. The velocity and acceleration of the robot determine the human understanding and satisfaction of the robot behavior. A robot approaching too fast can be perceived as threatening, a robot approaching too slow will will be annoying. A robot changing directions often will be hard to predict and thus not have very legible behavior in that respect, and a robot changing acceleration while approaching often similarly.

To find models of motion that appear natural, many researchers focus on mimicking aspects of human motion [37]. In the demonstration scenario we contribute by tracking human motion in controlled experiments. From that we can infer models about how humans plan and interpret motion by other agents,



(a) Example of experimental setup in HH-Interaction experiment.

(b) Mean trajectory in x-direction over time, comparison of left (green) and right sitting interaction partner (blue) vs. single condition (black).

Figure 18: Experimental setup for a human-human interaction task.

and get estimate of parameters for natural motion.

Aside from the technical challenges of creating safe experiments with moving robots and humans, and scientific challenges in proving ergonomic improvements of robot motion, we are also interested in combining local motion planning and high-level behavior. Traditionally, robot motion planning follows robot task planning, so the robot first decides what to do, and then where to go and how to get there.

However, with the dynamics introduced by humans, this approach causes annoying robot behavior. Instead we would like the robot to consider costs of global and local motion when deciding what to do, and to be able to change its mind about what to do next at given times. This human-aware navigation is not just used to reach a certain goal, but also to continuously estimate costs about how much discomfort moving to a certain target location will cause to humans at this given time.

3.3.4 Movement adaptation in joint action

Motivation: Our studies aim at understanding how humans coordinate their movements with a direct interaction partner. For a successful interaction between humans, it is necessary that people not only share the representation of the task - they must also be able to predict the actions of the interaction partner and integrate the predicted effects into the individual movement plan [60]. For a successful human-robot interaction, these mechanisms are of importance for several reasons. First, the human should be able to intuitively interact with the robot. Therefore, it is necessary that the robot acts and reacts in a way in which the human would expect him to move. Otherwise, it would be very irritating to a human interactant and an intuitive interaction might not be possible. Additionally, knowing what the counterpart will do is a safety issue: if the human expects the robot to act differently, the risk of collision and injury rises. To provide solutions to these problems, it is important to know how humans react in different situations [19, 32]. Our approach is to explore human-human interaction and find patterns and mechanisms of human interpersonal coordination. These behavioral rules will later be used as input to robotic development to improve direct human-robot interaction.

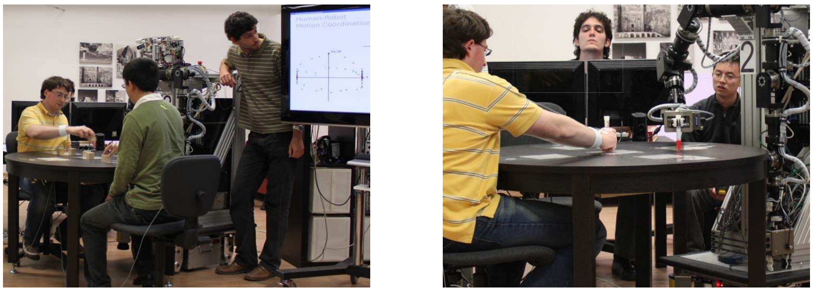


Figure 19: Demonstration of human-human and human-robot interaction in a sample motor task on CoTeSys October 2009 Workshop.

Human-Human Interaction: Exploring human-human interaction is pretty new and untackled in experimental psychology. Therefore, we start with simple experiments. For example, participants have to move bricks from a starting position to a target area [72, 73]. To obtain movement data, the trajectories of the hand and arm are recorded using the Polhemus Liberty motion tracker described in Section 2.1. The movement data is analyzed offline, calculating different spatial and temporal parameters. Figure 18 depicts an experimental setup and corresponding movement data.

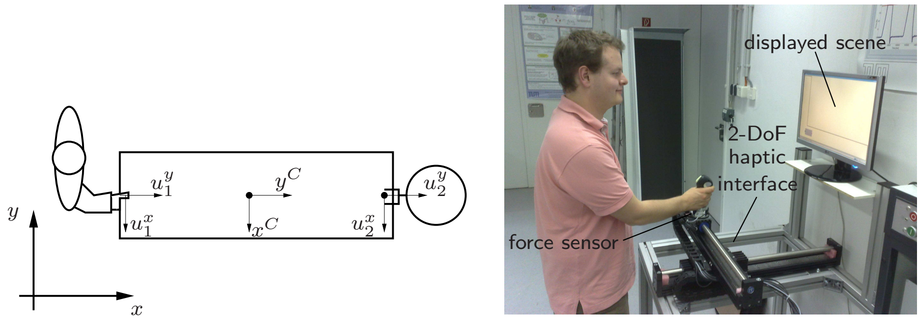
Human-Robot Interaction: From former experiments it is known that people synchronize their movements to those of an interaction partner when conjointly working on a repetitive motor task [56]. It is therefore assumed that in the present task, people establish an individual phase difference to their interaction partner over time that is kept constant throughout the experiment [25]. Here, robotic motion behavior may benefit from the found mechanisms. From online-tracking of human motion data this phase difference can be calculated and used to generate a robotic motion plan, as depicted in Figure 19.

In the next step, we want to evaluate this robotic control scheme in a direct human-robot interaction experiment in terms of coordinated task performance. As magnetic tracking with our robots is not possible due to interferences with metal parts, the infrared tracking system installed in the experimental area will be used (see Section 2).

3.3.5 Physical interaction

Physical human-robot interaction (pHRI) is an important aspect in multi joint action scenarios [38]. While speech and gestures can be used to negotiate tasks and goals explicitly, haptic interaction is mostly directed to simultaneously negotiate and solve tasks that require physical coupling [70]. Tasks demanding for sophisticated pHRI capabilities of robots can be generally referred to as joint manipulation of objects under environmental constraints where cooperation towards a common goal within a mixed human-robot team is needed. For realizing such robotic systems, several challenges must be addressed:

- Cooperation in overlapping subspaces of physical tasks requires efficient task sharing mechanisms.



(a) Planar example scenario of cooperative pHRI system.

(b) Experimental setup used for the evaluation of pHRI in [40].

Figure 20: Analysis and synthesis of task sharing strategies.

- Heterogenous, time-varying capabilities of participating agents demands for a sliding autonomy of the robotic partner.
- Tight coupling of dynamically changing, interacting systems give rise to adaptive control methods incorporating human models.
- Active contribution advancing the task state towards the common goal can be achieved with human-aware feedback planning only.

Task Sharing Strategies Our work in this field addresses the cooperative transport of a rigid bulky object with multiple humans and robots. We presented a system theoretic analysis of the human-robot cooperative manipulation problem and developed an experimentally confirmed strategy for static task sharing in terms of *consistent effort sharing policies* [40], see also Figure 20.

Next steps will involve an enhancement of our task sharing approach towards human-adapted sharing policies. User-adapted control of effort sharing will be performed in terms of an online-identification of the human's current input capabilities to the task.

Feedback Interaction Control: Human motor control is suggested to be executed on at least two levels to reduce task complexity on a higher level [69]. Additionally, humans have a very distinct sense of touch -called haptics- which is also used to communicate implicitly during interaction.

Schemes for fast *feedback interaction control* will be deployed on a low level of task execution, forming a finite set of haptic motor primitives: feedforward generation of motion following optimality principles derived from human modeling and adaptive feedback shaping of reactive behavior based on human dynamical models. As a common control objective, we seek to deploy metrics for cooperative task performance.

Planning for Cooperative Physical Tasks: Referring to our system-theoretic analysis of pHRI [40] we identified the need for cooperative planning for joint manipulation under environmental constraints. Challenging aspects of pHRI include:

- A human behavior model for joint physical tasks.
- Concurrent task execution and plan negotiation.
- The appropriate objective function for joint task performance, satisfying human ergonomic requirements as well as optimality criteria.

Human Behavior Modeling: To enhance the dyad's cooperative performance, an estimation of the human intention is necessary, as it minimizes the risks of conflict between partners and improves the performance of the robot's planner. Using a probabilistic approach for the human modeling we are able to learn not only at a continuous level but also at a symbolic level, segmenting the time series data into task primitives, which provides a higher level of abstraction prediction.

4 Conclusion

We presented an overview of the ongoing research in the cluster of excellence CoTeSys in the area of multi joint action and in connection with the MuJoA demonstration scenario. Multi joint action was defined as action that is done by two or more cognitive systems, whose distinctive features result from the interconnection between the systems. The different aspects of this interconnection were specified here as: shared knowledge, coordination and joint execution.

The various research efforts in the demonstration scenario that fall under these three topics have been described in more detail, focusing on the motivations and challenges of the individual research areas. Additionally, the already achieved results of the work, as well as work that is under way, have been introduced.

The report also presented the distinctive and quite unique experimental setup of the demonstration scenario, consisting of a human living area mockup, with multiple human-like robotic platforms and a large ceiling camera array.

A Collaborating CoTeSys Projects

A.1 List of active projects

- #136 System-Theoretic Modeling Approach to Perception-Cognition-Action Closed Control Loops
- #328 JAMIE – Joint Action in Multimodal Interaction Environments
- #410 MuJoA – Multi-Joint Action of Cognitive Systems
- #421 PARA – Planning for Adaptive Robot Assistance
- #425 Supporting Cognitive Processes on Mobile Platforms Using Joint Geometry-Based and Image-Based Environment Modeling

A.2 List of finished projects

- #327 ITrackU – Image-based Tracking and Understanding
- #344 MuDiS – A Multimodal Dialogue System for Intuitive Human-Robot Interaction
- #418 ACE – Action Planning and Decision Making for Cognitive Technical Systems: Mobile Vehicles and Humanoid Robots

References

[1] J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden, 2001.

[2] D. Althoff, M. Althoff, D. Wollherr, and M. Buss. Probabilistic collision state checker for crowded environments. In *IEEE International Conference on Robotics and Automation*, 2010.

[3] D. Althoff, O. Kourakos, M. Lawitzky, A. Mörtl, M. Rambow, F. Rohrmüller, D. Brščić, S. Hirche, and M. Buss. An architecture for real-time control in multi-robot systems. In *3rd International Workshop on Human-Centered Robotic Systems*, 2009.

[4] M. Andriluka, S. Roth, and B. Schiele. People-tracking-by-detection and people-detection-by-tracking. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, 2008.

[5] J. C. Balasuriya, K. Watanabe, and A. Pallegedara. Anfis based active personal space for autonomous robots in ubiquitous environments. In *International Conference on Industrial and Information Systems*, pages 523–528, 2007.

[6] P. Basili, M. Huber, T. Brandt, S. Hirche, and S. Glasauer. Investigating human-human approach and hand-over. In *Proceedings of the Third International Workshop on Human Centered Robotic Systems '09 , Cognitive Systems Monographs*, pages 151–160. Springer-Verlag, 2009.

[7] B. Bates and J. Cleese. *Gesichter: Das Geheimnis unserer Identität*. VGS Verlagsgesellschaft, 2001.

[8] H. Bless, N. Schwarz, and M. Kemmelmeier. Mood and stereotyping: The impact of moods on the use of general knowledge structures. *European Review of Social Psychology*, 7:63–93, 1996.

[9] G. Bower. Emotional mood and memory. *American Psychologist*, 31:129–149, 1991.

[10] G. Bradski and A. Kaehler. *Learning OpenCV, 1st edition*. O'Reilly Media, Inc., 2008.

[11] C. L. Breazeal. *Designing Sociable Robots*. MIT Press, 2001.

[12] G. Chamberlain. GigE Vision: Standard route to video over IP. *Industrial Ethernet Book*, 33(35), July 2006.

[13] W. Eckstein and C. Steger. The Halcon vision system: An example for flexible software architecture. In *In: 3. Meet. of Pract. Appl. on Real-Time Image Processing (org. by JSPE)*, 1999.

[14] P. Ekman. Universals and cultural differences in facial expressions of emotion. In *Nebraska Symposium on Motivation*. University of Nebraska Press, 1971.

[15] P. Ekman. Facial expressions. In T. Dalgleish and M. Power, editors, *Handbook of Cognition and Emotion*, New York, 1999. John Wiley & Sons Ltd.

[16] P. Ekman and V. W. Friesen. *Facial Action Coding Consulting*. Psychologist Press, 1977.

[17] J. Elfring, R. Janssen, and R. van de Molengraft. Data association and tracking: A literature survey. In *ICT Call 4 RoboEarth Project*. April 2010.

[18] P. M. Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology*, 47:381–391, 1954.

[19] T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and autonomous systems*, 42(3-4):143–166, 2003.

[20] T. Fraichard and H. Asama. Inevitable collision states. A step towards safer robots? *Advanced Robotics*, 18:1001–1024, 2004.

[21] J. Gast, A. Bannat, T. Rehrl, G. Rigoll, F. Wallhoff, C. Mayer, and B. Radig. Did I get it right: Head gestures analysis for human-machine interactions. In J. Jacko, editor, *Proc. Int. Conf. on Human-Computer Interaction HCI 2009, San Diego, CA, USA*, volume LNCS 5611, pages 170–177. Springer, 2009. 19.-24.07.2009, Human-Computer Interaction. Novel Interaction Methods and Techniques , ISBN 978-3-642-02576-1.

[22] J. Gast, A. Bannat, T. Rehrl, F. Wallhoff, G. Rigoll, C. Wendt, S. Schmidt, M. Popp, and B. Frber. Real-time framework for multimodal human-robot interaction. In *Proc. 2nd Int. Conf. on Human System Interaction HSI 2009, Catania, Italy*, pages 276 – 283. IEEE, 2009. 21.-23.05.2009.

[23] B. P. Gerkey and M. J. Matarić. A formal analysis and taxonomy of task allocation in multi-robot systems. *International Journal of Robotics Research*, 23:939–954, September 2004.

[24] M. Goebl and G. Farber. A real-time-capable hard-and software architecture for joint image and knowledge processing in cognitive automobiles. In *Proceedings of the Intelligent Vehicles Symposium (IVS)*, pages 734–740, Istanbul, Turkey, 2007. IEEE Press.

[25] H. Haken, J. Kelso, and H. Bunz. A theoretical model of phase transitions in human hand movements. *Biological cybernetics*, 51(5):347–356, 1985.

[26] U. D. Hanebeck, N. Saldic, and G. Schmidt. A modular wheel system for mobile robot applications. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 17–23, 1999.

[27] M. Hu. Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187, 1962.

[28] M. Huber, A. Knoll, T. Brandt, and S. Glasauer. Handing-over a cube: spatial features of physical joint action. *Annals of the New York Academy of Sciences*, 1164:380–382, 2009.

[29] M. Huber, M. Rickert, A. Knoll, T. Brandt, and S. Glasauer. Human-robot interaction in handing-over tasks. In *7th IEEE Ubterbatuibak Symposium on Robot and Human Interactive Communication*, pages 107–112, 2008.

[30] L. Itti and P. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009.

[31] C. E. Izard, J. Kagan, and R. B. Zsjonc. *Emotion-Cognition Relationships and Human Development, Emotion, Cognition and Behavior.* Cambridge University Press, 1984.

[32] M. Johnson, J. Bradshaw, P. Feltovich, C. Jonker, M. Sierhuis, and B. van Riemsdijk. Toward coactivity. In *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*, pages 101–102. ACM, 2010.

[33] N. Kalra, M. B. Dias, R. M. Zlot, and A. T. Stentz. Market-based multirobot coordination: A comprehensive survey and analysis. Technical Report CMU-RI-TR-05-16, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, December 2005.

[34] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *International Conference on Automatic Face and Gesture Recognition*, pages 46–53, France, March 2000.

[35] H. H. Kari. *Diskless Workstations in a Local Area Network.* License of science in technology, Helsinki University of Technology, Department of Electrical Engineering, 1989.

[36] A. Kirsch, T. Kruse, and L. Mösenlechner. An integrated planning and learning framework for human-robot interaction. In *4th Workshop on Planning and Plan Execution for Real-World Systems (held in conjuction with ICAPS 09)*, 2009.

[37] K. Koay, E. Sisbot, D. Syrdal, M. Walters, K. Dautenhahn, and R. Alami. Exploratory studies of a robot approaching a person in the context of handing over an object. In *Proc. AAAI - Spring Symposium 2007: Multidisciplinary Collaboration for Socially Assistive Robotics*, 2007.

[38] K. Kosuge and Y. Hirata. Human-robot interaction. In *Proceedings of the 2004 IEEE International Conference on Robotics and Biomimetics*, pages 8–11, 2004.

[39] T. Kruse, A. Kirsch, E. A. Sisbot, and R. Alami. Dynamic generation and execution of human aware navigation plans. In *Proceedings of the Ninth International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2010. accepted for publication.

[40] M. Lawitzky, A. Mörtl, and S. Hirche. Human-robot cooperative task sharing. In *19th IEEE International Symposium on Robot and Human Interactive Communication*, (accepted) 2010.

[41] C. Lenz, T. Röder, G. Panin, A. Knoll, M. Eggers, S. Amin, T. Kisler, and B. Radig. Distributed many-camera system for multi-person tracking in a human-robot-interaction scenario. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, October 2010. (submitted).

[42] J. S. Lerner and D. Keltner. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition and Emotion*, 14:473–493, 2000.

[43] W. Maier, E. Mair, D. Burschka, and E. Steinbach. Visual homing and surprise detection for cognitive mobile robots using image-based environment representations. In *Proc. IEEE Int. Conf. on Robotics and Automation (ICRA)*, pages 807–812, 2009.

[44] C. Mayer, M. Wimmer, and B. Radig. Adjusted pixel features for facial component classification. *Image and Vision Computing Journal*, 2009.

[45] A. Mehrabian. *Silent messages: Implicit communication of emotions and attitudes*. Wadsworth, Belmont, California, 1981.

[46] H. Miwa, K. Itoh, D. Ito, H. Takanobu, and A. Takanishi. Introduction of the need model for humanoid robots to generate active behavior. In *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)*, volume 2, pages 1400–1406, 27–31 Oct. 2003.

[47] H. Miwa, T. Okuchi, K. Itoh, H. Takanobu, and A. Takanishi. A new mental model for humanoid robots for human friendly communication introduction of learning system, mood vector and second order equations of emotion. In *Proc. IEEE International Conference on Robotics and Automation ICRA '03*, volume 3, pages 3588–3593, 14–19 Sept. 2003.

[48] L.-P. Morency and T. Darrell. Head gesture recognition in intelligent interfaces: the role of context in improving recognition. In *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006.

[49] C. Morimoto, Y. Yacoob, and L. Davis. Recognition of head gestures using hidden markov models. In *Proceeding of ICPR*, 1996.

[50] A. Müller, A. Kirsch, and M. Beetz. Transformational planning for everyday activity. In *Proceedings of the 17th International Conference on Automated Planning and Scheduling (ICAPS'07)*, pages 248–255, Providence, USA, September 2007.

[51] Y. Nakauchi and R. Simmons. A social robot that stands in line. *Autonomous Robots*, 12(3):313–324, 2002.

[52] C. Ogden, C. Fryar, M. Carroll, and K. Flegal. Mean body weight, height, and body mass index, United States 1960–2002. *Advance data from vital and health statistics*, 347, 2004.

[53] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int'l Conf. Multmedia and Expo (ICME'05)*, 2005.

[54] R. Parthasarathi and T. Fraichard. An inevitable collision state-checker for a car-like vehicle. In *Proc. of the IEEE International Conference on Robotics and Automation*, pages 3068–3073, 2007.

[55] M. Pollefeys, D. Nistér, J. M. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. J. Kim, P. Merrell, C. Salmi, S. Sinha, B. Talton, L. Wang, Q. Yang, H. Stewénius, R. Yang, G. Welch, and H. Towles. Detailed real-time urban 3D reconstruction from video. *Int. J. Comput. Vision*, 78(2-3):143–167, 2008.

[56] M. Richardson, K. Marsh, R. Isenhower, J. Goodman, and R. Schmidt. Rocking together: Dynamics of intentional and unintentional interpersonal coordination. *Human movement science*, 26(6):867–891, 2007.

[57] F. Rohrmüller, G. Lidoris, D. Wollherr, and M. Buss. System interdependence analysis for autonomous mobile robots. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009.

[58] J. A. Russell. A circumplex model of affect. *Personality and Social Psychology Review*, 1980.

[59] S. Satake, T. Kanda, D. Glas, M. Imai, H. Ishiguro, and N. Hagita. How to approach humans? strategies for social robots to initiate interaction. In *Proceedings of the 2009 ACM/IEEE International Conference on Human-Robot Interaction*, pages 109–116, 2009.

[60] N. Sebanz, H. Bekkering, and G. Knoblich. Joint action: bodies and minds moving together. *Trends in Cognitive Sciences*, 10(2):70–76, 2006.

[61] S. Šegvić and S. Ribarić. A software architecture for distributed visual tracking in a global vision localization system. In *Proceedings of the 3rd International Conference on Computer Vision Systems*, pages 365–375. Springer-Verlag, 2003.

[62] H.-Y. Shum, S.-C. Chan, and S. B. Kang. *Image-Based Rendering.* Springer, Spring Street, NY, 2007.

[63] E. A. Sisbot, A. Clodic, R. Alami, and M. Ransan. Supervision and motion planning for a mobile manipulator interacting with humans. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 327–334, 2008.

[64] E. A. Sisbot, L. F. Marin-Urias, R. Alami, and T. Simeon. A human aware mobile robot motion planner. *IEEE Transactions on Robotics*, 23:874–883, 2007.

[65] K. Smith-Jentsch, J. Johnson, and S. Payne. Measuring team-related expertise in complex environments. In J. Cannon-Bowers and E. Salas, editors, *Decision making under stress: Implications for individual and team training*. American Psychological Association, 1998.

[66] S. Sosnowski, A. Bittermann, K. Kühnlenz, and M. Buss. Design and evaluation of emotion-display EDDIE. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS '06)*, pages 3113–3118, Beijing, China, Oct. 2006.

[67] B. Stanczyk. *Developement and Control of an Anthropomorphic Telerobotic System*. PhD thesis, Technische Universität München, Institute for Automatic Control Engineering, 2006.

[68] K. Sycara and G. Sukthankar. Literature review of teamwork models. Technical Report CMU-RI-TR-06-50, Robotics Institute, Carnegie Mellon University, Pittsburgh, USA, 2006.

[69] E. Todorov. Optimality principles in sensorimotor control. *Nature Neuroscience*, 7(9):907–915, 2004.

[70] T. Tsumugiwa, T. Yokogawa, and K. Hara. Variable impedance control based on estimation of human arm stiffness for human-robot cooperative calligraphic task. In *Proceedings of the 2002 IEEE International Conference on Robotics and Automation*, pages 644–650, 2002.

[71] M. Valera and S. Velastin. Intelligent distributed surveillance systems: a review. *IEEE Proceedings Vision, Image and Signal Processing*, 152(2):192–204, 2005.

[72] C. Vesper, A. Soutschek, and A. Schubö. Motion coordination affects movement parameters in a joint pick-and-place task. *The Quarterly Journal of Experimental Psychology*, 62(1):1–15, 2009.

[73] C. Vesper, S. Stork, and A. Schubö. Movement Times in Inter-and Intrapersonal Human Coordination. In *Proceedings of the 2008 ECSIS Symposium on Learning and Adaptive Behaviors for Robotic Systems (LAB-RS)*, pages 17–22. IEEE, 2008.

[74] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*, volume 1, pages 511–518, Kauai, Hawaii, 2001.

[75] M. Wimmer, F. Stulp, S. Pietzsch, and B. Radig. Learning local objective functions for robust face model fitting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 30(8):1357–1370, 2008.

[76] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys (CSUR)*, 38(4):13, 2006.