How to Address Smart Homes with a Social Robot?
A Multi-modal Corpus of User Interactions with an Intelligent Environment

Patrick Holthaus, Christian Leichsenring, Jasmin Bernotat, Viktor Richter,
Marian Pohling, Birte Carlmeyer, Norman Köster, Sebastian Meyer zu Borgsen,
René Zorn, Birte Schiffhauer, Kai Frederic Engelmann, Florian Lier, Simon Schulz,
Philipp Cimiano, Friederike Eyssel, Thomas Hermann, Franz Kummert,
David Schlangen, Sven Wachsmuth, Petra Wagner, Britta Wrede, Sebastian Wrede

Cluster of Excellence Cognitive Interaction Technology (CITEC),
Bielefeld University, Inspiration 1, 33619 Bielefeld, Germany
patrick.holthaus@uni-bielefeld.de

Abstract

In order to explore intuitive verbal and non-verbal interfaces in smart environments we recorded user interactions with an intelligent apartment. Besides offering various interactive capabilities itself, the apartment is also inhabited by a social robot that is available as a humanoid interface. This paper presents a multi-modal corpus that contains goal-directed actions of naive users in attempts to solve a number of predefined tasks. Alongside audio and video recordings, our data-set consists of large amount of temporally aligned sensory data and system behavior provided by the environment and its interactive components. Non-verbal system responses such as changes in light or display contents, as well as robot and apartment utterances and gestures serve as a rich basis for later in-depth analysis. Manual annotations provide further information about meta data like the current course of study and user behavior including the incorporated modality, all literal utterances, language features, emotional expressions, foci of attention, and addressees.

Keywords: interaction corpus, smart home, social robot

1. Introduction

With smart home technologies becoming increasingly widespread, for example in elderly care (Morris et al., 2013; Hendrich et al., 2014; Cavallo et al., 2014; Aubergé et al., 2014), new opportunities for the collection of interaction data arise. Such environments thereby promise a dense web of functionalities and services that are so seamlessly integrated into living spaces that artifacts to control, query or program them all but vanish. In fact, if every component possessed its own interface, the environment would be utterly cluttered and lose its familiarity and calmness. The question arises how to bridge the interface gap and how to control these systems in an easy, intuitive, or – better – natural way. One approach to create novel device interaction strategies is to start bottom-up – i. e. from observing how inexperienced users would do it if they could do whatever they like to initiate a desired function (e. g. Valdes et al. (2014)).

In this work, we present a multi-modal corpus on goal attainment strategies in a number of experimental situations, carefully selected to avoid a specific bias regarding the modalities to use for interactions. The data accordingly contains free-form interaction sequences with the environment that are not restricted in the way given tasks are performed by the participants.

To this end, individual users trigger a set of specific device actions and responses in our *cognitive service-robotics apartment* (CSRA)[1], a smart apartment embedded in our research building that offers services both from a variety of smart ambient components (sensors and actuators) and via a bi-manual mobile robot. Beyond offering versatile in-
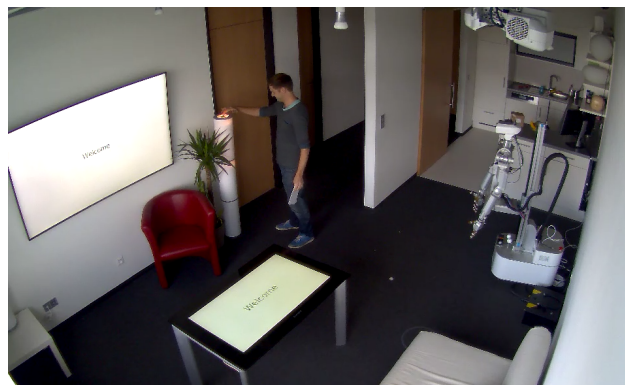

Figure 1: Overview shot of a study participant attempting to alter the brightness of a floor light in the living room of the *cognitive service robotics apartment* (CSRA) by touching its edge. The scene is part of the presented corpus.

teraction means, the environment at the same time delivers synchronized multi-modal data repositories for interaction analysis. With the acquired corpus we take first steps towards user-centered interfaces of high expectability. Careful examination of the gathered data will refine our ongoing design paradigms for interfaces to be established for 24/7 readiness for use of the intelligent environment.

We find in our data that participants explore both verbal and non-verbal interaction modalities to address their immediate surroundings in order to achieve certain goals such as switching off lights or turning on the radio. Please refer to Figure 1 for an exemplary depiction of a study participant interacting in the apartment. The Wizard-of-Oz setup (cf. Kelley (1984)) enables participants to maintain the illusion that their mode of invoking functions is actually correctly understood by the

---

[1] http://cit-ec.de/en/content/cognitive-service-robotics-apartment-ambient-host

smart environment. The corpus itself consists of audio and video material as well as robot and apartment reactions plus a variety of additional sensor and actuator information.

This paper describes the interactive setup and the resulting data-set in more detail, how it has been obtained, and what opportunities it holds for other researchers of various disciplines. We value the data as useful for further in-depth analyses of people's interactions with devices, ambient intelligence, and robots in everyday environments. In particular, Section 2. briefly introduces the scientific aspiration for creating such an interaction corpus and gives an insight on the experimental setup as well as participant instructions. Details on the apartment's infrastructure and the technical recording pipeline are discussed in Section 3. The multimodal content of the corpus is presented in Section 4. along with a preliminary annotation scheme. The paper concludes with a short summary of the presented data-set and its possibilities for further evaluation in Section 5.

2. Goal and Research Question

Typically, a smart home can solve a certain number of tasks that are delegated by the inhabitant, e. g. altering the temperature or dimming the light. A major requirement thereby is that such functionality can be accessed in an intuitive way (cf. Green et al. (2004)). To be able to provide such intuitive interaction capabilities, we first need to know the types of interfaces people expect an intelligent apartment to have available.

This is something where researcher bias can get in the way and it is therefore important to try and minimize any influence on the participants as to whether they use verbal or gestural commands, and whether interaction with the robot would be preferred over a disembodied addressee in the form of the apartment as a whole. As a consequence, participants of our study are not explicitly briefed about the sensory capabilities of the apartment and robot beforehand, i. e. neither gesture nor speech have been teased as possible interaction methods.

Our experimental setup thus targets at exploring which modalities users intuitively consider when given a task in a smart home inhabited by a robot. With the chosen parameters and tasks we encourage participants to try out goal-directed communicative acts towards the apartment or robot. Accordingly, people unfamiliar with the CSRA and its interactive potential have been invited to participate in the following course of action inside our apartment:

A given participant enters the apartment together with an experimenter who first gives a short explanation about the environment and introduces the robot. Some lights inside the apartment are briefly illuminated and the robot waves its arm in order to hint at their communicative abilities. Participants are then confronted with the following sequence of everyday tasks they need to solve as soon as the experimenter leaves the room: (i) *turn on the light* in the hallway, (ii) *turn it off* again from another room, (iii) *listen to music* and query about (iv) the *current time*, (v) whether a *call* or (vi) *delivery* has been missed, and (vii) alter the *brightness* of a light. The selected configuration enables participants to find an intuitive solution to their tasks that is however not the result of habits acquired from past technological limitations.

To further encourage this, light switches in the apartment are disabled during the study, and no radio or amplifier is available featuring a volume knob. In the same vein, the apartment does not contain any clock or telephone and the use of the participants' own mobile phones or watches was prohibited. As to also promote nonverbal interactive strategies, only in half of the trials, speech has been employed by the apartment or robot in order to give textual information like the current time. In the other trials, the same information is given via screens or solely via ambient cues in the case of the parcel task.

In a remote-controlled "Wizard-of-Oz" (cf. Kelley (1984)) setup, a human operator observes the participants attempts at solving the current task. On detecting an attempt at solving the task, the operator controls the environment so that the participants believe that the apartment actually understands and executes their commands. Only obviously goal-directed actions are regarded as valid attempts, e. g. a gesture towards the light that should be switched, an utterance directed to either robot or apartment, clapping, direct object contact. All actions are recorded via four cameras and two microphones, system events are registered with the help of our integrated sensor and actuator infrastructure.

3. System Overview: The Cognitive Service Robotics Apartment (CSRA)

Our smart environment consists of a living room, a sports area, a bathroom, and a kitchen, all of which are interconnected by a short hallway (cf. Figure 2). The kitchen is fully furnished and equipped with a fridge, a stove, and an oven. In the living room, there is a sofa, two chairs, and a table. A television device is mounted at the wall. Also, the bathroom is fully operational and even containing a shower. The exercise room consists of a large multi-purpose area with another television device. It employs a large variety of sensors to record the ways a user interacts with the devices or the robot and various modes to give feedback to the user. For the contributed data-set, only the kitchen, living room, and hallway are considered because participants were only present in these areas.

To visually detect and record the users themselves[2], there are 12 Asus Xtion Pro Live RGB-D cameras[3] in the ceiling used for person tracking, looking straight down to cover the whole apartment area approximately at waist level, 4 high-quality Sony FCB-EH6300 1080p CMOS cameras[4], and 4 lower quality Basler BIP2-D1920c-dn cameras[5] streaming 960×720 H.264 video. The presented corpus solely contains recordings from the lower quality cameras because the Sony cameras produce such vast amounts of data that real-time encoding and recording is impractical. Instead, they are reserved for special recognition tasks where compression artifacts would be a major hindrance.

---

[2]The bathroom is excluded from any visual surveillance for privacy reasons.

[3]http://asus.com/3D-Sensor/Xtion_PRO_LIVE

[4]http://sony.co.uk/pro/product/industrial-cameras-seh/fcb-eh6300

[5]http://baslerweb.com/en/products/cameras/network-cameras/ip-fixed-dome-cameras/bip2-d1920c-dn-indoor
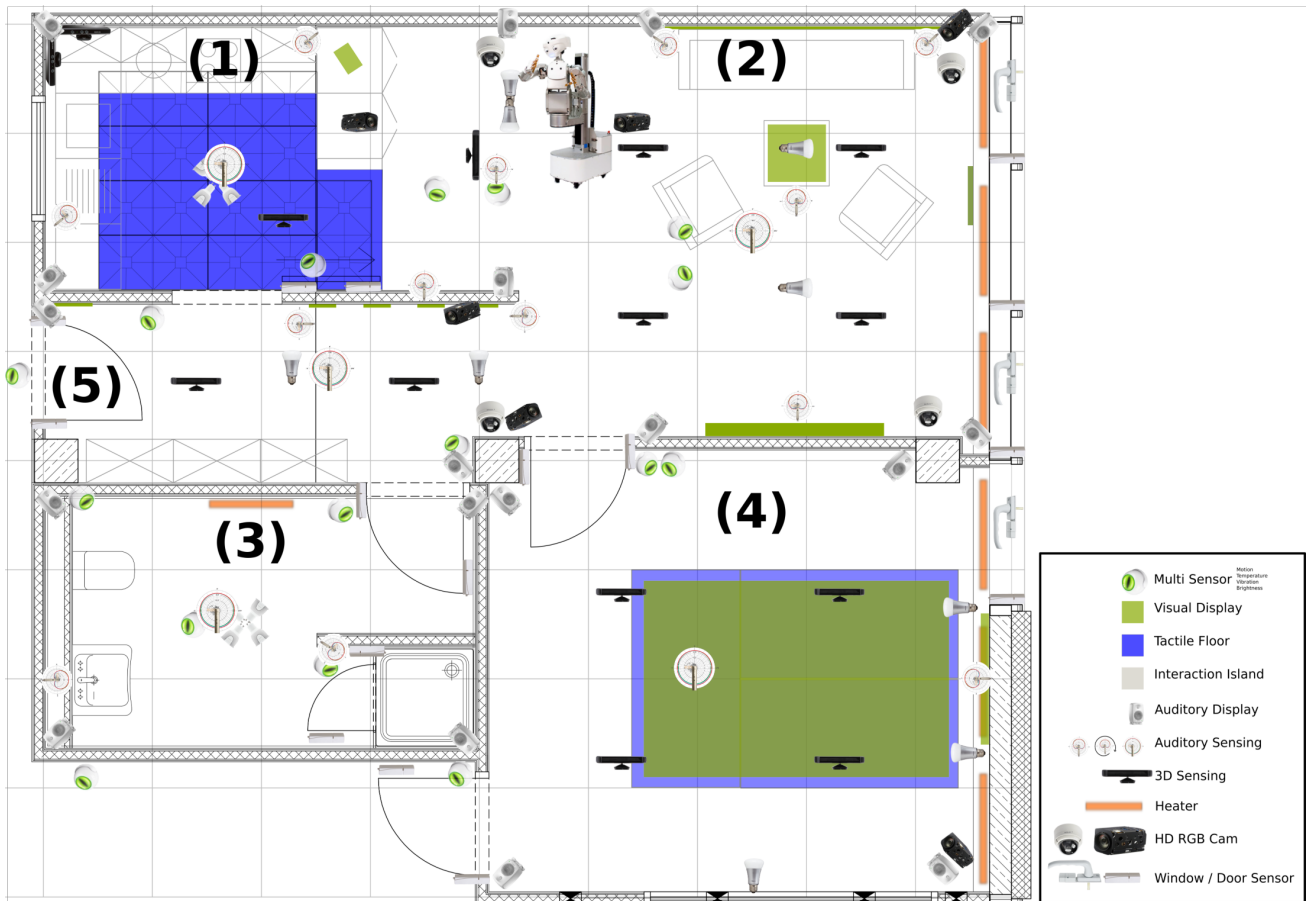
Figure 2: Overview of the *cognitive service robotics apartment* (CSRA) and its infrastructure. In between the kitchen area (1) and the living room (2) a humanoid robot is positioned. The bathroom (3) and sports area (4) are located at the other side of the apartment. The main entrance is located at the end of the hallway (5).

There are also 16 Fibaro motion sensors[6] to detect motion in an area without much pre-processing but also a low temporal and spatial resolution. In addition, there are two areas where a Future Shape SensFloor capacitive floor[7] senses the presence of people. To get a picture how the visual sensors and sensitive floor play together in determining user positions, please refer to Figure 3.

For audio recording and speech recognition, there are 12 Rode NT3 cardioid microphones[8] and 5 Rode NT55 omnidirectional microphones[9] installed. With the help of a KNX system[10], it can be sensed which devices are running and how much power they consume. The apartment's windows and doors are equipped with wireless HomeMatic reed switches[11] to indicate whether they are open, tilted or closed. Also, reed switches for reading the state of cupboard doors and drawers are installed in the kitchen.

Where output is concerned, the apartment can use a range of displays and projectors (see dark green areas in Figure 2), 18 Genelec 8000 series loudspeakers[12] plus one 7050B subwoofer and more than 50 Philips Hue LED light bulbs[13] with configurable hue and intensity, as well as smart drawer and cupboard handles that can show different light colors and patterns for targeted attention control. Most home automation components, i. e. light bulbs and motion sensors, are integrated via openHAB[14] which is used for hardware integration but not for controlling purposes.

For verbal communication we use a combination of incremental speech processing (cf. Baumann and Schlangen (2012)) and dialog manager (cf. Peltason and Wrede (2010)) by Carlmeyer et al. (2014). Speech recognition is thereby realized with the Sphinx (Huang et al., 1993) framework while MaryTTS (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, 2015) is used as a speech synthesizer. Dialog and speech are configured to use the German language (instead of English which is also available) in order to provide participants the most familiar interface as all of them are German native speakers.

---

[6] http://fibaro.com/us/the-fibaro-system/motion-sensor

[7] http://future-shape.com/en/technologies/23/sensfloor-large-area-sensor-system

[8] http://en.rode.com/microphones/nt3

[9] http://en.rode.com/microphones/nt55

[10] http://knx.org

[11] http://eq-3.de/sensoren-detail/items/hm-sec-sc.html

[12] http://genelec.com/studio-monitors/8000-series-studio-monitors/8020c-studio-monitor

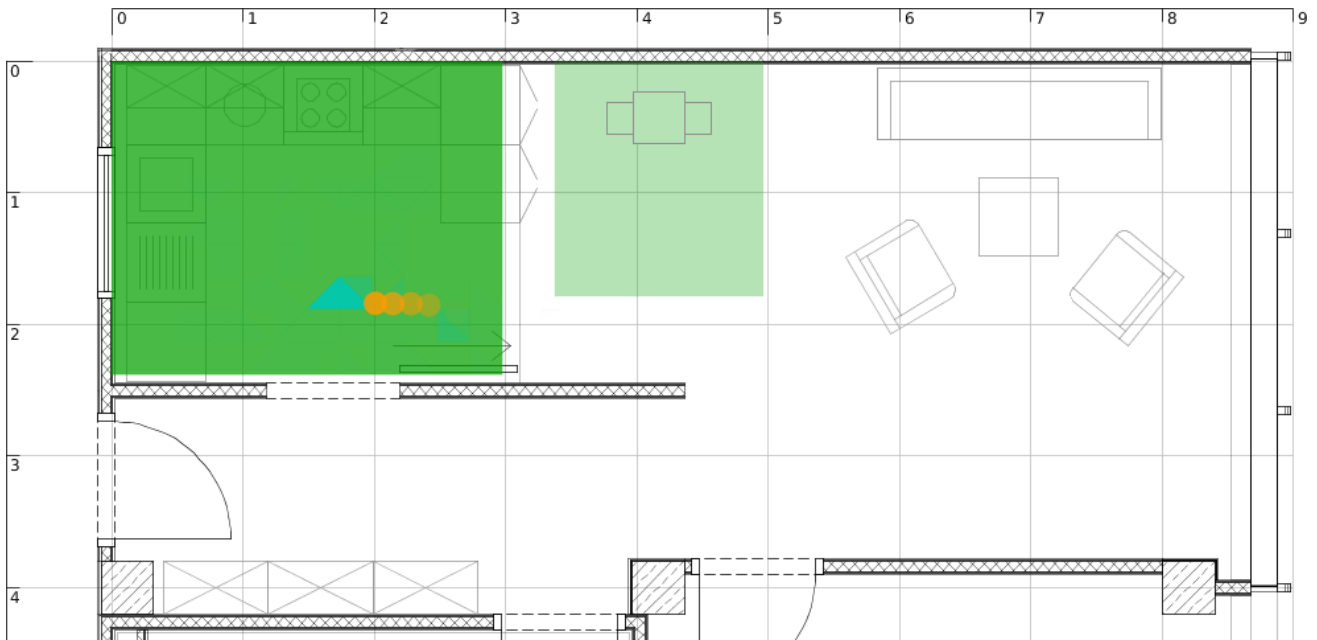[13] http://meethue.com

[14] http://openhab.org

Figure 3: Live visualization of person percepts: Motion sensors give a rough estimation of presence (green areas), while the sensitive floor (cyan triangles) and camera tracking (yellow dots) allow for more accurate positioning. The capacitive floor is only available in the kitchen whereas the whole apartment is covered with depth sensors.

The humanoid robot in the apartment (cf. Meyer zu Borgsen et al. (in press)) features multiple cameras, a laser range finder and microphones allow to gather information from the environment. A real-time-enabled computer controls the compliant force controlled actuators with four-fingered hands. In total, the robot is equipped with 37 motor-powered joints. It has 7 per arm, 5 per hand, 2 in the head, 2 in the torso and 9 joints actuate the base including a z-lift. For interaction, it incorporates the same speech processing pipeline as the apartment. The robot is also able to nod if an utterance has been understood and it can exhibit pointing gestures towards various items inside the apartment.

All these devices produce a lot of data (at least 1300 MiB/s) that needs to be distributed, processed, and at least partially stored. With the Robotics Service Bus (RSB) and the Robotics Systems Types Repository (RST) (Wienke and Wrede, 2011) we established a homogeneous architecture that integrates most of the devices used in the apartment. As a middleware, we use a partitioned RSB network with Spread (Amir et al., 2004) and socket transports in order to handle the bandwidth and latency requirements. This also lets us store all the sensor data in a unified format that provides time-stamps for every data packet in so-called channels separately for each device or software component.

For annotation purposes, we merge the videos from all four cameras perspectives into one video file and add a single audio track from one of the videos. Afterwards, we generate ELAN (Max Planck Institute for Psycholinguistics, 2015; Wittenburg et al., 2006) files that show the video and audio data along with the annotations collected automatically by the smart environment (Moringen et al., 2013). Such a combined file contains only data tiers for a channel that was actually present in a particular recording. RSB time-stamps are used to synchronize all the tracks and crop them to the part where they overlap. Alternatively, exports in JSON format can be generated as well for use with other annotation software.

4. Corpus content

In total, the corpus consists of approximately 7:38 hours of video and audio material of 62 trials with 32 male and 30 female participants and an average age of 26 years. A typical trial after the introduction by the experimenter takes around five minutes. The first half of the videos (31) has been captured from three different angles that cover each interactive location of the apartment. In the remaining recordings, a fourth camera providing an overview has been utilized additionally. Please refer to Figure 1 for an exemplary depiction of the recorded scenes.

Besides visual and acoustic material from kitchen, hallway and living room, the whole system behavior is available in separate channels with timing information for each event. Most importantly, all actuated items are temporally aligned with the videos and accessible as annotations. In particular, the following events are included:

- *Wizard action:* Points in time where the wizard triggers actions in the apartment or robot as a response to user behavior. The annotation typically consists of nine to twelve events: At least one for each of the seven tasks plus welcoming routines emitted by both robot and apartment and clearing the screens (cf. Figure 4).

- *Utterances of robot and apartment:* Times and durations of all responses towards participant-initiated tasks that have been communicated verbally, i. e. the time of day, missed calls, and newly arrived parcels. This tier only occurs in verbal trials (28).
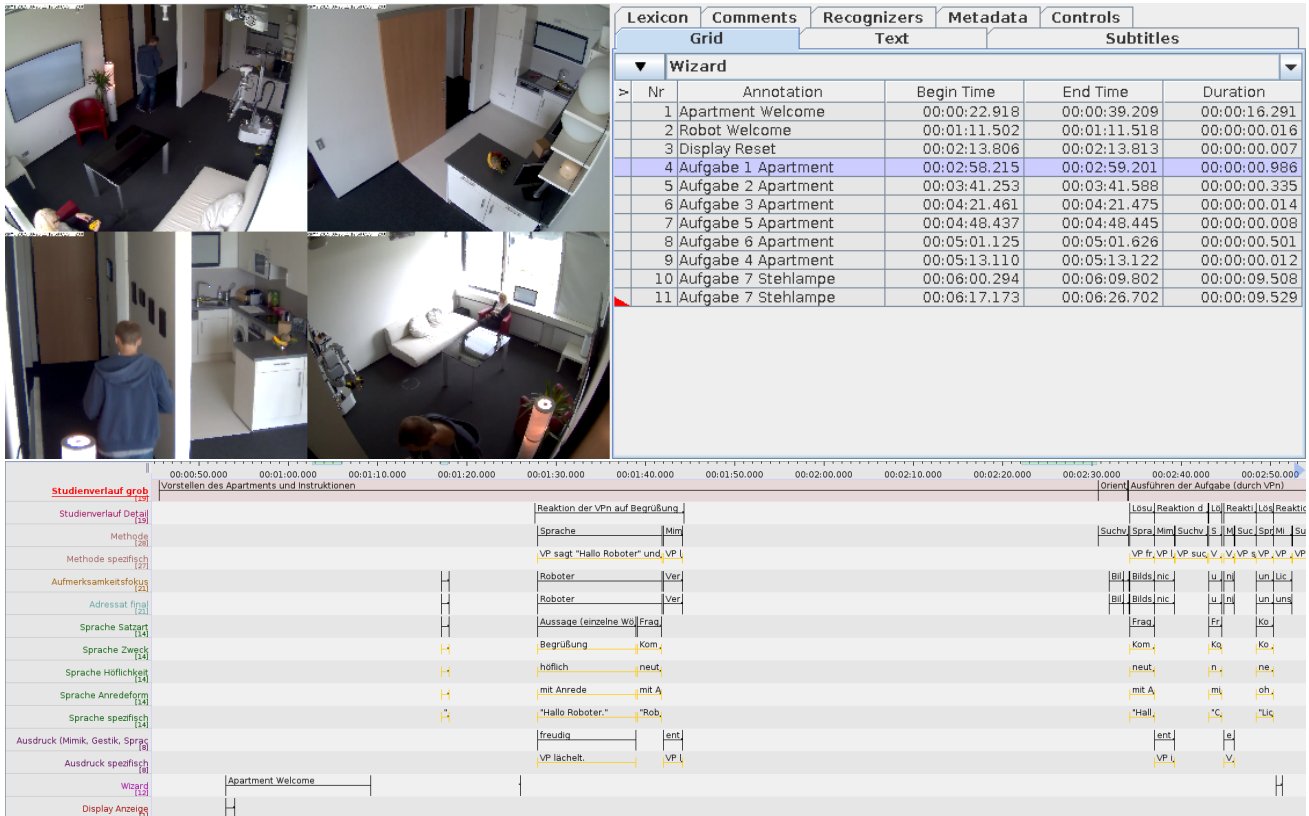
Figure 4: Snapshot of the ELAN program featuring a live interaction. The scene is displayed from four camera angles (upper left) and exemplary annotation tiers are displayed below (German). The upper right part gives details on the annotation tier that contains actions performed by the wizard (also German).

- *Robot gesture:* Greeting gestures, nodding, and pointing gestures towards other output devices, i.e. screens and the kitchen cupboard.

- *Display contents:* Text notifications on the five displays, containing the current time, missed calls, and parcel delivery. Display contents are only present in nonverbal data-sets (34).

- *Radio:* Time and duration of music being played.

- *Cupboard state:* Whether the kitchen cupboard is open or closed, and times when handle indicates a delivered parcel with a notification beep and blue color.

- *Entrance door state:* Contains information on whether the front door, operated by any of the experimenter or participant, is opened or closed.

Furthermore, sensory and pre-processed information from kitchen, hallway, and living room (cf. Figure 2) is also available in the corpus. The motion sensors are recorded to provide a rough estimate on human activity in each room as well as in front of the entrance door. Data from the kitchen's touch-sensitive floor is contained in the corpus as well. For a more accurate representation of user movements, person tracking data of the depth cameras in the relevant areas (kitchen, living room, hallway) is also included. Audio from two microphones is present as individual channels in the data-set (one from the center of the living area and the other from the hallway microphone). Additionally, although

button functions have been disabled, each attempt at manipulation has been recorded as well.

The apartment's hardware status is part of the corpus with information about power consumption at 16 outlets, color and brightness of 27 (ambient) lights, eight temperature sensors, and the power state of each screen. For data reconstruction and alignment, video and system recording times and information about active software components and hosts is also part of the gathered data.

Besides automatically collected data, the corpus also contains annotations of the video material so that the data-set includes extensive information about participant behaviors. As the study has been carried out with German native speakers, participant behaviors are annotated in German as well – particularly their literal utterances. In detail, the following information has been entered manually:

- *Course of study – overview:* Description of rough sections of the study depending on participants' progress and behavior, i.e. introduction of the apartment, the robot and the tasks, stage of reading, stage of orientation and reflection (when a participant has read the task and thinks about how to solve it) and stage of solving the task.

- *Course of study – in detail:* Description of participants' behavior and progress during the study in more detail, i.e. participants' reaction to the apartment's/robot's welcoming or to the wizard's action, participants' reaction when they can (not) solve the task successfully

or when there is a technical response the participant does not understand, interruption (when a participant has a question and wants to talk to the experimenter) and notice in case a participant cannot solve the task.

- *Method:* Assessment of the participants' method to (try to) solve the task, i. e. speech, gesture, eye movements, facial expression, touch, motion, conventional (i. e. using light switches) or a combination of multiple methods (i. e. gesture and speech). An additional tier (method – specific) verbally describes the participant's actions in the annotator's own word.

- *Language features:* Separate tiers each give information about the addressee, politeness, and intention of an utterance. In the addressee tier, it is annotated whether participants explicitly address an entity (e. g. apartment, robot, light) or not. Whether they use politeness phrases such as "please" or are particularly harsh is given in the politeness tier. What the intention of the utterance might be, i. e. to greet or to interact with an entity, is also given. A further tier (language – specific) gives information about the literal utterances of the participant.

- *Emotional expression:* Manual classification of what emotion participants express, i. e. pleasure, surprise, fear, or neutral. An additional tier (emotional expression – specific) describes the participants' emotional expression verbally.

- *Focus of attention:* Definition of what participants address shortly before solving the task, i. e. robot, apartment, screens, furnishings, doors, windows, light switches, experimenter, themselves, unspecific (when addressing something, but unclear what exactly) and not discernible.

- *Final addressee:* Annotation of what participants address in order to solve the task successfully (same options as focus of attention). Please note that the final addressee and the focus of attention might differ from each other – for example if someone gazes towards a monitor but verbally addresses the robot.

The quality and richness of the recorded multi-modal data allows for further in-depth analysis of user interactions with intelligent environments and robots. We are currently investigating gestures and utterances as well as their respective targets in order to assess which modalities are preferred and how people expect to interact with the environment to solve a given task. First findings suggest that participants use both verbal and nonverbal strategies to solve their task but verbal attempts can be observed more frequently in tasks that involve information retrieval. Participants also apply the strategy that apparently worked in the first tasks to the later ones and seem to be irritated if they are not transferable.

Our next evaluation steps include a statistical analysis of idle and exploration phases in order to determine people's occupancy with their task or potential confusion. In addition, timings, frequencies, and sequences of behavioral strategies offer ample opportunities for further examination.

5. Conclusion

With this paper, we present a multi-modal corpus of user interactions inside a versatile intelligent environment – the *cognitive service robotics apartment*. The structured data-set consists of video streams from four perspectives containing audio, temporally aligned system behavior such as carried out actions, gestures, and utterances. Moreover, also user behavior has been annotated with regard to interaction attempts, i. e. modality, language features, attention, addressee, and emotional expression. With the presented data, gained in a Wizard-of-Oz setting, we take first steps towards user-centered interfaces of high expectancy. By providing the data set on-demand to other researchers, we offer an opportunity for better understanding human expectations and interaction types to control smart environments.

Acknowledgements

References

Amir, Y., Danilov, C., Miskin-Amir, M., Schultz, J., and Stanton, J. (2004). The Spread Toolkit: Architecture and Performance. Technical report, Johns Hopkins University, Center for Networking and Distributed Systems (CNDS).

Aubergé, V., Sasa, Y., Bonnefond, N., Meillon, B., Robert, T., Rey-Gorrez, J., Schwartz, A., Batista Antunes, L., De Biasi, G., Caffiau, S., and Nebout, F. (2014). The EEE Corpus - Socio-Affective "Glue" Cues in Elderly-Robot Interactions in a Smart Home with the EmOz Platform. In *5th International Workshop on Emotion, Social Signals, Sentiment and Linked Open Data*.

Baumann, T. and Schlangen, D. (2012). The InproTK 2012 Release. In *NAACL-HLT Workshop on Future Directions and Needs in the Spoken Dialog Community: Tools and Data*, pages 29–32.

Carlmeyer, B., Schlangen, D., and Wrede, B. (2014). Towards Closed Feedback Loops in HRI: Integrating InproTK and PaMini. In *Proceedings of the 2014 Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction*, MMRWHRI '14, pages 1–6. ACM.

Cavallo, F., Aquilano, M., Bonaccorsi, M., Limosani, R., Manzi, A., Carrozza, M., and Dario, P. (2014). Improving Domiciliary Robotic Services by Integrating the ASTRO Robot in an AmI Infrastructure. In Florian Röhrbein, et al., editors, *Gearing Up and Accelerating Cross-Fertilization between Academic and Industrial Robotics Research in Europe*, volume 94 of *Springer Tracts in Advanced Robotics*, pages 267–282. Springer International Publishing.

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH, (2015). *The MARY Text-to-Speech System (MaryTTS)*. http://mary.dfki.de/.

Green, W., Gyi, D., Kalawsky, R., and Atkins, D. (2004). Capturing User Requirements for an Integrated Home Environment. In *Proceedings of the Third Nordic Conference on Human-computer Interaction*, NordiCHI '04, pages 255–258. ACM.

Hendrich, N., Bistry, H., and Zhang, J. (2014). PEIS, MIRA, and ROS: Three frameworks, one service robot - A tale of integration. In *2014 IEEE International Conference on Robotics and Biomimetics*, ROBIO, pages 1749–1756. IEEE.

Huang, X., Alleva, F., Hon, H.-W., Hwang, M.-Y., Lee, K.-F., and Rosenfeld, R. (1993). The SPHINX-II Speech Recognition System: An Overview. *Computer Speech & Language*, 7(2):137–148.

Kelley, J. F. (1984). An iterative design methodology for user-friendly natural language office information applications. *Transactions on Information Systems*, 2(1):26–41.

Max Planck Institute for Psycholinguistics, (2015). *ELAN*. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. http://tla.mpi.nl/tools/tla-tools/elan/.

Meyer zu Borgsen, S., Korthals, T., and Wachsmuth, S. (in press). ToBI-Team of Bielefeld The Human-Robot Interaction System for RoboCup@ Home 2016.

Moringen, J., Nordmann, A., and Wrede, S. (2013). A Cross-Platform Data Acquisition and Transformation Approach for Whole-Systems Experimentation – Status and Challenges. In *ERF2013 Working Session on Infrastructure for Robot Analysis and Benchmarking*.

Morris, M. E., Adair, B., Miller, K., Ozanne, E., Hansen, R., Pearce, A. J., Santamaria, N., Viega, L., Long, M., and Said, C. M. (2013). Smart-Home Technologies to Assist Older People to Live Well at Home. *Journal of Aging Science*, 1(1):1–9.

Peltason, J. and Wrede, B. (2010). Pamini: A Framework for Assembling Mixed-Initiative Human-Robot Interaction from Generic Interaction Patterns. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '10, pages 229–232.

Valdes, C., Eastman, D., Grote, C., Thatte, S., Shaer, O., Mazalek, A., Ullmer, B., and Konkel, M. K. (2014). Exploring the design space of gestural interaction with active tokens through user-defined gestures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 4107–4116. ACM.

Wienke, J. and Wrede, S. (2011). A middleware for collaborative research in experimental robotics. In *2011 IEEE/SICE International Symposium on System Integration*, SII, pages 1183–1190. IEEE.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). ELAN: A Professional Framework for Multimodality Research. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, LREC, pages 1556–1559. ELRA.