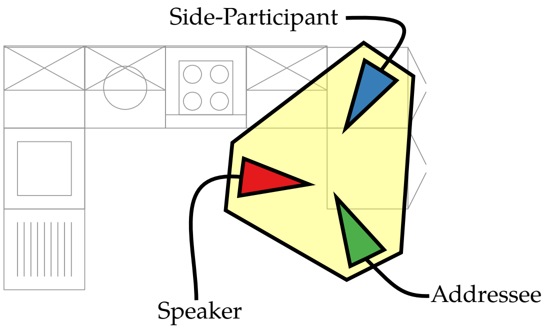
VIKTOR RICHTER

ADDRESSING IN SMART ENVIRONMENTS

VIKTOR RICHTER

ADDRESSING IN SMART ENVIRONMENTS

An Investigation of Human Conversational Behaviours Towards Devices and Autonomous Agents in a Smart Environment



Doctoral thesis presented for the degree of
Doctor of Engineering (Dr.-Ing.) at

Faculty of Technology
Bielefeld University
Applied Informatics
Inspiration 1
33619 Bielefeld
Germany

REVIEWERS
Prof. Dr.-Ing. Franz Kummert
Prof. Dr. Ing. Katrin Solveig Lohan

BOARD
Prof. Dr.-Ing. Ulrich Rückert
Dr. Basil Ell

DEFENDED AND APPROVED
April 24, 2020

Digital version for online publication.

ACKNOWLEDGMENTS

I never would have been able to submit this thesis without the funding, chances, help, and advice from others. First of all, I owe thanks to Britta Wrede, Franz Kummert, and Gerhard Sagerer for giving me the opportunity and funding to work in the Applied Informatics Group and the Cluster of Excellence Cognitive Interaction Technology "CITEC" (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG). Thanks to my supervisor Franz Kummert, in particular, for guiding me during my research and taking the time for regular and fruitful discussions. Moreover, I want to thank Prof. Lohan for accepting the invitation to review this thesis, and Prof. Rückert and Dr. Ell for taking part in my examination board. As this research is not build up from nothing, I owe thanks to the research community and open source community in general for laying the foundations. The members of the AI Group and CITEC deserve thanks for helpful feedback, discussions and support. In particular, I my colleagues in the CSRA project who worked with me overtime and into weekends to create and maintain the apartment and conduct studies in it. Furthermore, the people who agreed to participate in these studies deserve special thanks too. I want to thank my family and friends for the support they give me. For being there for me and for making my life worth living. However, the biggest thanks of all deserves Birte. My schoolmate, fellow student, colleague in the CSRA and AI, reviewer, proofreader, family, friend, and partner. She superbly juggles all of these roles and more and deserves thanks for each of them.

ABSTRACT

Human communication is complex, dynamic and implicit. People know when others want to interact with them. They know when they are addressed, whether they need to react, and to whom. This understanding is learnt early and refined throughout the whole life. Artificial agents, in contrast, do not grow up. They are not exposed to great amounts of high quality training in interaction as humans are. Nevertheless, if we want to interact with artificial agents in as we do with humans, we need them to understand our communication. They need to recognize the states we are in, the intentions we pursue, and the behaviours we display to achieve this. In this thesis, I investigate which human behaviours can be observed to infer the conversational state and intentions of humans in interactions with artificial agents in a smart environment. After a detailed review of literature on the principles of human interaction and the efforts to transfer these to artificial agents and smart environments, I investigate human conversational cues in interactions with different kinds of agents. With these investigations I show that (1) although addressing in unconstrained interactions of single users with devices and agents is diverse, the addressed entity can be recognized to a high degree from audio-visual cues, (2) a robot in a human-robot conversational group can utilize facial information of its interlocutors to decide whether it is addressed or not, and (3) the conversational group and role of a virtual agent can be recognized by observing the motion and facial expressions of the people in its vicinity. The insights from these investigations and the corresponding models allow an automatic interpretation of human conversational behaviour in interactions with artificial agents. This can be used to create agents which better understand and utilize human communication, to make interaction more natural and effective.

CONTENTS

III GROUPS & ROLES IN COPRESENCE

LIST OF FIGURES

LIST OF TABLES

NOTATION

ATTRIBUTION OF AUTHORSHIP

I speak of myself using *I* in case of work originally done by myself alone. In case the results of a collaboration with others are presented, I use *we*. The respective collaborators are indicated by the co-authors of the corresponding publication.

MARGIN NOTES

✎ Definitions are highlighted with this icon.

FOOTNOTES

I use numbered footnotes to give additional, technical information. Furthermore, I repeat research questions and claims in footnotes wherever applicable. These repetitions are marked with a ↻.

NOTATION OF BAYESIAN NETWORKS

When describing Bayesian Networks, I use arrows to mark conditional dependencies. $A \rightarrow B \leftarrow C$ means that B conditionally depends on A and C. Furthermore, I use curly brackets to represent dependencies on multiple variables. So, $A \rightarrow B \leftarrow C$ can be written as $\{A, C\} \rightarrow B$.

CONFIDENCE INTERVALS

Wherever possible, I calculate 95% confidence intervals for measurements and visualize them as error bars.

QUALITY MEASURES

To assess the quality of classifiers, I calculate the usual measures. For binary classifications this is done as follows: From the classification results (predictions) and known, correct results, cases of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) can be calculated and shown in the:

$$\textit{confusion matrix} = \begin{bmatrix} \mathrm{TP} & \mathrm{FP} \\ \mathrm{TN} & \mathrm{FN} \end{bmatrix}$$

✎ confusion matrix

The numbers of positive and negative observations in the data and positive and negative predictions can be directly extracted from this matrix:

$$\text{condition positive (CP)} = \text{TP} + \text{FN}$$

$$\text{condition negative (CN)} = \text{FP} + \text{TN}$$

$$\text{predicted positive (PP)} = \text{TP} + \text{FP}$$

$$\text{predicted negative (PN)} = \text{FN} + \text{TN}$$

The proportion of positive observations in the data:

prevalence ✎

$$\mathit{prevalence} = \frac{\text{CP}}{\text{CP} + \text{CN}}$$

The proportion of overall correct classifications:

accuracy ✎

$$\mathit{accuracy} = \frac{\text{TP} + \text{TN}}{\text{CP} + \text{CN}}$$

Other measures of classification performance are calculated as follows:

recall ✎

$$\text{true positive rate (TPR)} = \mathit{recall} = \text{sensitivity} = \frac{\text{TP}}{\text{CP}}$$

fall-out ✎

$$\text{false positive rate (FPR)} = \mathit{fall\text{-}out} = \frac{\text{FP}}{\text{CN}}$$

$$\text{false negative rate (FNR)} = \frac{FN}{CP}$$

$$\text{true negative rate (TNR)} = \mathit{specificity} = \mathit{selectivity} = \frac{TN}{CN}$$

precision ✎

$$\text{positive prediction value (PPV)} = \mathit{precision} = \frac{TP}{PP}$$

$$\text{false omission rate (FOR)} = \frac{FN}{PN}$$

$$\text{false discovery rate (FDR)} = \frac{FP}{PP}$$

$$\text{negative prediction value (NPV)} = \frac{TN}{PN}$$

F1-score ✎

$$\mathit{F1\text{-}score} = 2 \cdot \frac{\mathit{precision} \cdot \mathit{recall}}{\mathit{precision} + \mathit{recall}}$$

ALTERNATIVE QUALITY MEASURES

For biased data, when the prevalence differs strongly from 0.5, quality measures that are less sensitive to bias can be calculated. The value of diagnostic odds ratio (DOR) can be interpreted as the odds of correctly classifying divided by the odds of falsely rejecting [Gla+03].

$$\text{positive likelihood ratio (LR+)} = \frac{TPR}{FPR}$$

$$\text{negative likelihood ratio (LR-)} = \frac{FNR}{TNR}$$

$$\textit{diagnostic odds ratio (DOR)} = \frac{LR+}{LR-}$$

✎ DOR

Furthermore, markedness and informedness can be used as prevalence-free precision and recall alternatives [Pow11]. Markedness indicates how trustworthy the decision of a model is. A value of zero signifies that the decision is random, one means that it is fully trustworthy—all classifications are correct.

$$\textit{markedness} = PPV + NPV - 1$$

✎ markedness

Informedness indicates how informed the classifier is about the classes in the data. A value of zero signifies that the model is uninformed about the data, one means that it is fully informed—positive and negative cases can both be correctly retrieved.

$$\textit{informedness} = TPR + TNR - 1$$

✎ informedness

Part I

RESEARCH TOPIC

In this part of the thesis, I introduce the overarching goal, deduce the research questions that guide this work, and give an overview of the remaining document. Furthermore, I present the relevant literature from social and computer sciences to create an understanding of human behaviour in focused and unfocused interaction with humans, artificial agents, and interactive devices. In doing so, I additionally introduce required terms and concepts.

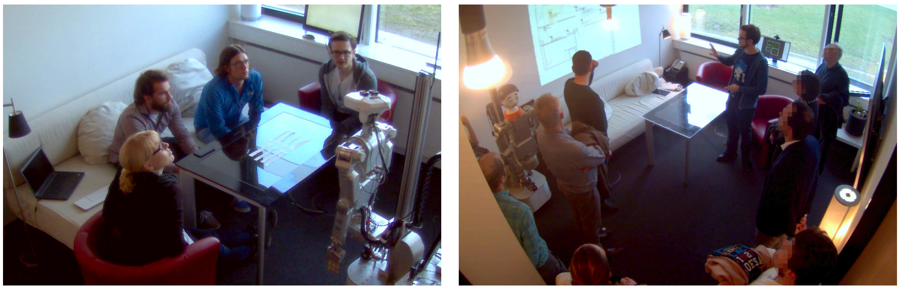
1 INTRODUCTION

In this chapter I present the context and overall aim of this thesis. After a short introduction to the subject, I highlight the problems that can evolve in human interaction with artificial agents in a smart environment. I formulate the overarching goal of this thesis and derive four research questions to approach this goal. Furthermore, I introduce a smart environment which allows this research and in which the relevant experiments take place. Finally, I give an overview of the structure of this thesis.

1.1 INTERACTION IN SMART ENVIRONMENTS

In recent years, smart home technologies and robots not only receive a lot of interest in research communities but get increasingly widespread in daily living. Private homes can be equipped with many sensors and actuators. Movement detectors, temperature sensors, heating systems, doors, lights, shutters, dishwashers, and more can be connected to automate insignificant and tedious tasks. While general purpose social robots, which do chores and serve guests at parties are neither affordable nor available yet, many specialized solutions emerge. Fully automatic vacuum-cleaners, litter-boxes, pet-feeding machines, and cooking machines perform their task sufficiently good for people to buy and use them in their private homes. Simultaneously, Intelligent Personal Assistants (IPAs)—like the Mycroft AI, Amazon Alexa, Microsoft Cortana, Google Assistant, and Apple's Siri—in smart speakers, phones or TVs allow us to retrieve information, or control lights and media playback with simple verbal commands.

When people interact in a smart environment, they do not always interact with an artificial agent in the form of a robot, virtual assistant, or smart speaker. They usually interact with other people or do not want to communicate at all (example situations can be seen in Figure 1.1). Additionally, there may be multiple groups of conversing people at the same time, each of them trying to interact with an agent or not. However, gestures and sounds are perceived by everyone in the vicinity, not only by the addressed persons or agents. A permanently present agent is required to cope with these problems. It needs to respond when a person expects it to do so and ignore communication which is not directed at it. Furthermore, according to The Media Equation [RN96], people treat systems that show characteristics associated with humans in a human like manner. This applies to Human-Computer-Interaction (HCI) [NST94] and IPAs [LW18]. The interaction with such systems



(a) A group interacting with a robot

(b) People interacting with each other

Figure 1.1: Two situations as they happen in a smart environment. The first picture (1.1a) shows a group of people interacting with a robot. The second picture (1.1b) shows a scene where people speak with each other while the robot stands in the background. It is excluded from their interaction.

elicits social responses in people. When agents have an embodiment that supports this effect, it gets even stronger. A virtual agent presented by Cassell et al. [CT99] uses turn taking signals to enhance the perceived efficiency of the interaction. According to Mutlu et al. [Mut+09], a robot can be used to manipulate the conversational roles in an interaction. The robot *Kismet*[1] can express synthetic emotions to persuade people into nurturing it [BV99]. Furthermore, in a tutoring scenario with a robot presented by Lohan et al. [Loh+12], the human tutors elicit more social communication when the robot shows contingent gaze behaviour. Consequently, Kerstin Dautenhahn et al. show that people want robots to communicate in a human-like way and argue for a *robotiquette* to make Human-Robot-Interaction (HRI) acceptable and comfortable to human interaction partners [Dau+05; Dau07]. However, this robotiquette does not just require robots to react when addressed. It poses a set of requirements with increasingly complex challenges from *not standing in the way*, to *showing interest* and *waiting for an adequate moment to talk* to *being considerate and polite*. To have a chance at fulfilling these requirements, agents need a detailed understanding of human behaviour and expectations towards them.

robotiquette

Therefore, the goal of this thesis is to investigate how the sensors of a smart environment and contained agents can be used to analyse the conversational state and expectations of inhabitants. In the following section, I formulate this goal in detail and derive research questions to guide the contributions of this thesis.

1.2 RESEARCH QUESTIONS

People are likely to elicit social behaviours when confronted with a device that shows human like characteristics [RN96]. However, the level of this effect greatly depends on the persons expectations and the

[1] http://www.ai.mit.edu/projects/humanoid-robotics-group/kismet/

appearance and behaviour of the artificial agent [Heg+08]. Therefore, a varying amount of elicited conversational cues can be expected for different kinds of agents. However, as different agents and devices have different sensors, they often do not have the sensory means to recognize such cues. A smart environment though, has access to much broader sensing capabilities than its individual agents and the possibility to combine its sensors with the agent's. Therefore, the overall goal of this dissertation can be stated as follows:

GOAL *Use the perception of a smart environment and its agents to recognize the conversational state and expectations of inhabitants towards different kinds of artificial agents.*

To approach this goal, I investigate the following research questions:

RQ 1 (*INDIVIDUAL ADDRESSING*) *Which behaviours in naïve human interaction with a smart environment can be observed to distinguish which agent is addressed with a deliberate communicational act?*

Traditionally, lights and multi-media devices are controlled by pushing switches and buttons on the device or a remote. The addressee of the touch, thereby, is never ambiguous. This does not apply to gestures and speech. In multi-modal interactions the addressee of a communicative act is inherently ambiguous and needs to be resolved. As speech and gestures can be observed by everyone in the vicinity, people need to indicate the addressee of their communicative acts. The same problem arises when people multi-modally control the functionality of a smart environment. Therefore, a set of cues must exist which people use to indicate the addressee in such a situation.

RQ 2 (*ADDRESSING IN GROUPS*) *How can an artificial agent visually recognize whether it was addressed by a person within its conversational group or not?*

An interaction of a group of people with a robot, naturally does not only contain communication directed towards the latter. While addressees in verbal interactions are sometimes explicitly stated, this is usually not needed. To know who is addressed, people utilize their knowledge about the interaction and the behaviour of others. Therefore, it is necessary for an artificial agent to know who is speaking and monitor the conversational cues of its interlocutors to be able to distinguish if a statement was addressed at it or not.

RQ 3 (*CONVERSATIONAL GROUPS*) *How can focused interactions of people with artificial agents be automatically recognized in a smart environment?*

Artificial agents are not always part of an interaction. People dynamically create and change conversational groups in which the agent may or may not participate. Depending on whether the agent is part of an interaction, it can have different options and duties. On the one hand, it should be open for interactions but not intrusive when nobody wants to interact. On the other hand, it should actively participate and support the interaction when this is desired. To be able to fulfil these conflicting requirements, the agent needs to know whether someone (and who) intends to interact with it.

RQ 4 (CONVERSATIONAL ROLES) How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

Speaker and addressee are not the only roles participants of a conversation can assume. Furthermore, these roles do not only exist when one person stops to talk and another begins. In a conversation, all participants assume a role at any given time. When an artificial agent can recognize its conversational role, it not only knows when it needs to listen and when to speak. It can use conversational cues to influence the distribution of roles in a more informed manner. Furthermore, it can compare its recognition to its expectations. It can detect deviations between them, to start repair strategies.

1.3 RESEARCH ENVIRONMENT

smart environment

To specify what a *smart environment* is, I use the definition of Diane Cook and Sajal Das:

> A smart environment is a small world where all kinds of smart devices are continuously working to make inhabitants' lives more comfortable. [...] [It] is able to acquire and apply knowledge about an environment and also to adapt to its inhabitants in order to improve their experience in that environment. [CD04, p. 3]

Cognitive Service Robotics Apartment as Ambient Host (CSRA)

apartment

smart home

To be able to investigate the presented research questions, a smart environment is needed that not only allows the observation of human interactions with devices, but also with virtual agents, and robots. Furthermore, the execution and recording of corresponding interaction studies should be supported. The *Cognitive Service Robotics Apartment as Ambient Host (CSRA)*[2] [Wre+17] is a laboratory in the Cluster of Excellence Cognitive Interaction Technology (CITEC)[3] at Bielefeld University which meets these requirements. If not stated differently, I refer to the CSRA whenever I use the term *apartment*. It is furnished as a flat with the capabilities of a smart environment—so it is a *smart home*—and additionally has the observation and recording capabilities of an interaction laboratory. Photographs of the apartment, its agents and a layout plan be seen in Figures 1.2 and 1.3. It is a suitable environment

[2] https://www.cit-ec.de/csra

[3] https://cit-ec.de

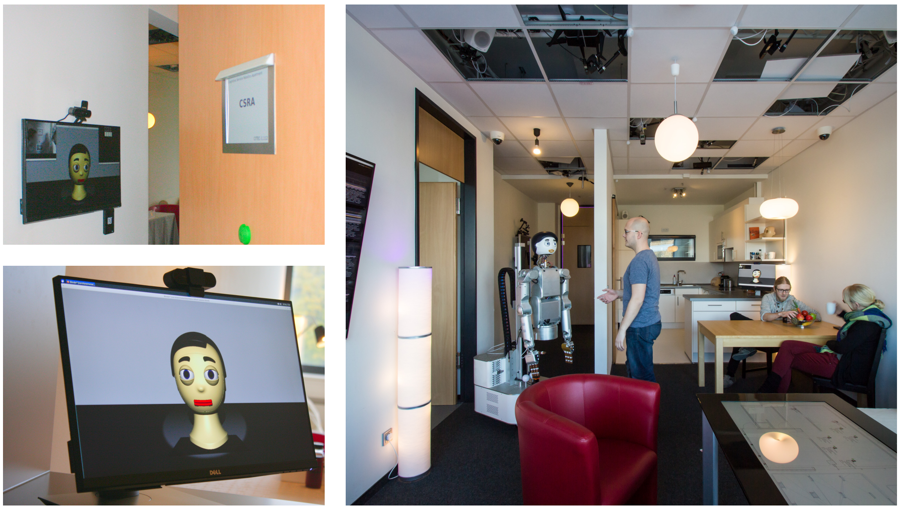


Figure 1.2: Photographs of the Flobi agents on the left (Flobi Entrance behind the half open door at the top, Flobi Assistance from within the kitchen at the bottom). The right picture shows the apartment from the outer right end of the living room. In this image, two persons are chatting at the table while another is interacting with the Floka robot.

to investigate human behaviour and interaction with humans, robots, virtual agents and computers for the following reasons:

ARTIFICIAL AGENTS & INTERACTIVE APPLIANCES: The apartment has multiple interactive devices, virtual agents and a robot (shown in Figures 1.2 and 1.3). The two virtual Flobi heads [Lüt+10; LSW14]—one in the corridor (Flobi Entrance) and one in the kitchen (Flobi Assistance)—function as the apartment's hosts to welcome and introduce people. The mobile robot Floka [Wac+17] is based on the *MekaBot M1*.[4] It features an anthropomorphic upper body with manipulation capabilities. Moreover, for its head between the original sensor head and an adapted version of the Flobi head can be chosen [SBW19]. Furthermore, the apartment contains lights, speakers, screens, door-handles, a pan-tilt beamer, and an interactive plant which can be used to unobtrusively inform people, interact with them or guide their attention.

VARIETY OF SITUATIONS: As a fully integrated smart flat on a university campus, it is used in diverse ways. Therefore, various kinds of interactions can be observed in it. In the first place, it constitutes a workspace for the involved staff and students who develop, integrate and test new functionalities and interaction metaphors. However, demonstrations introduce people to the apartment who are not familiar with it and its possibilities. Meetings and socializing events take place regularly, and finally, it is used to conduct studies. These can range from human interaction with the smart

[4] https://robots.ieee.org/robots/m1/

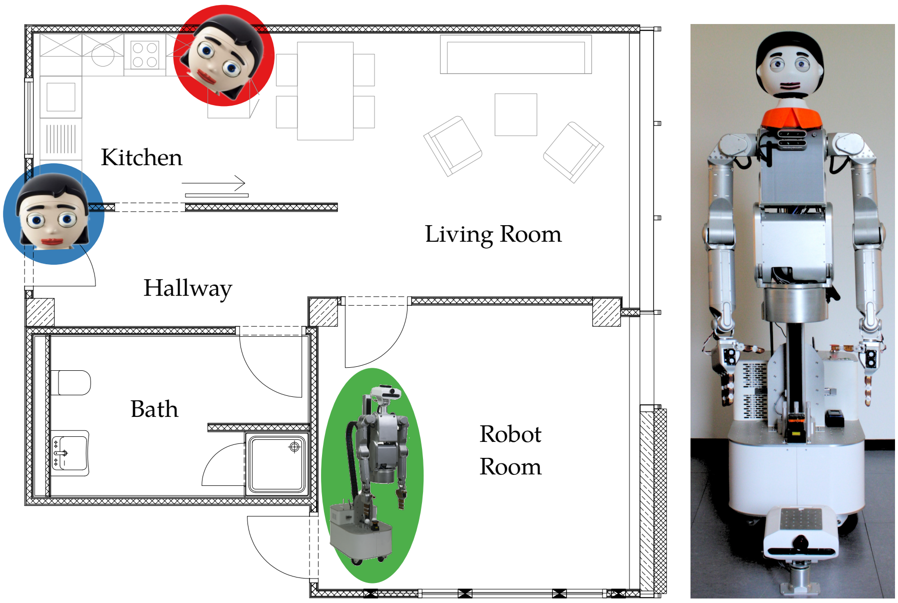


Figure 1.3: On the left, the layout of the apartment can be seen. The robot Floka is highlighted in green, Flobi Entrance in blue, and Flobi Assistance in red. The right picture shows the robot Floka mounted with it's adapted Flobi head. The sensor head is placed on the floor in front of the robot.

environment to research that is not concerned with smart environments but utilizes the recording and analysis facilities.

SENSORS & INTROSPECTION: The apartment features various sensing and recording capabilities. Four cameras in the corners of the apartment provide an overview of the whole situation. Eight RGBD-Cameras capture the apartment from a top-down perspective. Three Web-cameras provide high resolution video captures from the Flobis' viewpoints and for a screen in the living room. Furthermore, a sensitive floor in the kitchen can detect peoples positions to enrich the apartment's person sensing capabilities. Microphones capture the global sound and interactions at designated interaction zones in particular. Movement detectors, sensors for temperature, light, the opening of doors, windows, cupboards, and drawers can be used to follow human physical interaction with the apartment.

AVAILABILITY & RECORDING The apartment is operational 24/7 and can automatically record interactions as they occur. High level information about the apartment's state, the contained agents, and people are available through the *Robotics Service Bus* (*RSB*) by Sebastian Wrede et al. [RSB]. Compressed video and audio streams are created and accessible via *GStreamer* by Wim Taymans et al. [gstreamer] and the RTP protocol [Sch+03]. All this data can be recorded on demand for studies [Hol+16] or started and stopped automatically based on usage in a 24/7 operation [RK18].

1.4 DOCUMENT OVERVIEW

The remaining document is composed as follows. In the next chapter, I introduce the topic of human conversational behaviour in focused and unfocused interaction from the viewpoint of social sciences. While at it, I define the relevant terms and models. Furthermore, I give an overview of the topic of human interaction and Human-Agent-Interaction (HAI) from the viewpoint of computer sciences. To this end, I additionally establish a taxonomy for the distinction of interactive entities that is applies throughout this thesis.

The central two parts of this thesis, are concerned with the four research questions stated in Section 1.2. In Part II, I investigate human addressing in smart environments. To this end, I present a study of naïve interactions of people with a smart environment (Chapter 3). I examine RQ 1 by performing an in-depth investigation of the resulting corpus and creating a model for human addressing behaviour towards artificial agents in smart environments. In Chapter 4, I present a mixed HRI scenario with the anthropomorphic robot Floka. On this basis, I investigate RQ 2. In Part III, I aim at a more global understanding of human behaviour in copresence with artificial agents. To this end, I present a scenario and interaction study that allows the analysis of free, dynamically changing conversations of humans with artificial agents (Chapter 5). On this basis, I explore RQ 3 in Chapter 6 by creating and evaluating a detection framework for mixed human-agent conversational groups in a smart environment. In Chapter 7, I use the models resulting from Chapter 4 and Chapter 6 to investigate the recognition of conversational roles and assess RQ 4.

In the final part (Part IV), I summarize the contributions and impact I make with this thesis to research on human interaction with smart environments and artificial agents. Furthermore, I discuss the limitations of this work, present possibilities for improvement, and give ideas for applications and future research.

2 PRINCIPLES OF HUMAN INTERACTION

Humans are social beings. We communicate our thoughts and ideas through speech and coordinate our actions and behaviour to accomplish more than a single individual can achieve. We share our knowledge and strengthen our social bonds through multi-modal interaction. To make these interactions fluent and successful, people coordinate through behavioural cues. This coordination is fundamental for human communication. Therefore, and according to The Media Equation [RN96], it can be assumed that people will produce such cues when interacting with robots, virtual agents presented on a screen or smart speakers. People understand whether others in their vicinity are open for communication, and who they address with their speech, by observing their behaviour. Similarly, an artificial agent needs to observe the communicative cues, directed at it—or others—to better understand and utilize human behaviour and expectations. In their book on The Media Equation, Byron Reeves and Clifford Nass already point out the necessity of politeness for computers [RN96]. They argue that systems need to greet, use eye contact and match the users' modality. This is especially important in long-term and in the wild interactions, where the agent is a potential interaction partner, but one of many. For an agent to be acceptable to humans not only within a single interaction but in the long-term it needs to behave socially appropriate. This robotiquette [Dau07] means that an agent needs to behave appropriately even when it is not interacted with. In such a situation it—for example—may leave the interaction and orient somewhere else to show civil inattention [Gof63].

To be able to better satisfy human expectations, and simplify interactions in smart environments for naïve persons, knowing these expectations is crucial. Knowledge about human interactive behaviour towards one another can be used to reason about the motives and causes of communication signals in human behaviour. Additionally, this allows understanding which signals and behaviours would be naturally comprehensible for humans and therefore can effectively be used by artificial agents in interaction. Therefore, I start the literature review by outlining how people interact with each other. Furthermore, interaction can not always be focused. An artificial agent that is in copresence with humans for an extended period needs to handle both focused interaction and unfocused interaction. To account for the different requirements of these two types of interaction, I divide the literature on human interaction with other humans, artificial agents, and smart environments along this distinction.

In the first part of this chapter, I investigate everyday interactions between humans. I illustrate the difference between unfocused and focused interaction, and how people behave in such situations. Furthermore, I analyse focused interactions with regard to conversations and the roles that can be taken by the participants in this process. In the second part, I present a taxonomy of interactive entities that I use throughout this work. With a literature review on human interaction with artificial agents and smart homes, I investigate how far human interaction principles can be, or already are, applied in such scenarios. Finally, I assess to what extent the presented effects and patterns in human interaction are transferable to people with different cultural backgrounds.

2.1 INTERACTION BETWEEN HUMANS

As of today, there is an ever growing set of ways for people to interact with each other, even without being physically collocated. Emails, telephones, and other technical solutions allow communication across great distances or even distributed through time. However, depending on the applied solution, such interactions are restricted. The restriction can be in the available modalities, as in communication via telephone. Moreover, a distribution over longer time periods—as when using text-messaging—can be an intended property or just a side effect of the technical solution. Although these ways of interaction and the way people utilize them are interesting in themselves, they often strongly differ from direct interaction. For example, people speaking on the telephone introduce themselves and exhibit a set of other verbal interactions which would be clumsy, or otherwise redundant for people who have unobstructed physical access to each other [Aue17a]. According to Schegloff [Sch68] these additional interactions are introduced to gather information which otherwise would be transmitted multi-modally or known from the situational context. These additional interactions show which information the participants need to be able to successfully communicate with each other. As I focus on collocated communication in this thesis, I only consult other types when they highlight crucial parts of the interaction.

The kind of interaction in focus of this work happens when people have direct, physical access to each other without any obstructions. Erving Goffman termed this *copresence*:

copresence ✎

> [To be copresent] persons must sense that they are close enough to be perceived in whatever they are doing, including their experiencing of others, and close enough to be perceived in this sensing of being perceived. [Gof63, p. 17]

Goffman further states: ‘Copresence renders persons uniquely accessible, available, and subject to one another.’ [Gof63, p. 22]. In copresence people can interact in a focused or unfocused manner but they can not avoid interaction. Additionally, people perceive the presence of others differently depending on their distance. The subdivision of the space which people claim around them and concede to others is investigated in the field of proxemics and has an effect on both focused and unfocused interaction.

2.1.1 *Proxemics*

Proxemics are important in both focused and unfocused interaction between people. Edward T. Hall coined the term proxemics while investigating peoples usage of public space [Hal69]. *Proxemics* define four circularly extending distances—intimate distance, personal distance, social distance and public distance— around a person with different physiological and interactional implications (see Table 2.1). Within *intimate distance* physical contact is probable. It may only be entered by partners. The used voice is low or whispered. People that know each other well can discuss personal topics within *personal distance*. Here the other is within arm's reach but not touched and the voice level is moderate. The close phase of the *social distance* is used for less personal interactions between colleagues or during social gatherings. The voice level is normal and the distance can be interpreted as a hint to the participants' involvement. In the far phase of the social distance people can easily engage and disengage without being rude. The voice level becomes louder. At *public distance* the voice gets loud and the phrasing more formal. Non-verbal communication is performed through gestures and bodily stance since facial expressions cannot be perceived [Hal69, p.117-125]. The distances are not exact but vary depending on the culture, relationship, context, age, gender and size of the participants.

proxemics

intimate distance

personal distance

social distance

public distance

Distance	Close Phase	Far Phase
Intimate distance	≤ 0.15 m	≤ 0.46 m
Personal distance	≤ 0.76 m	≤ 1.22 m
Social distance	≤ 2.13 m	≤ 3.66 m
Public distance	≤ 7.62 m	7.62 m and more

Table 2.1: The different proxemic radii around a person and their properties according to Hall [Hal69]. Each distance is further divided into an inner and outer phase. The exact distances vary depending on culture, age, gender and other characteristics of the person and situation.

2.1.2 *Unfocused Interaction*

Being physically copresent not only allows people to multi-modally sense others but obliges them to communicate and interact. Each action that is performed in copresence can, and will be perceived by others and interpreted in the situational and social context. As Adam Kendon notes: '[A]ll aspects of behaviour in a situation of co-presence must be considered at least, potentially, to have a role in the communication process.' [Ken90, p. 27] The purpose of *unfocused interaction* is the management of copresence. To this end, people adjust their movements, gestures, sound level, and overall displayed involvement in the situation at hand [Gof63].

unfocused interaction

2.1.2.1 *Coordination and Social Communication*

When people in copresence do not actively participate in a mutual activity, they still interact with each other—albeit unfocused. Two persons that pass each other on the pavement in opposite directions, are communicating. They at least require some coordination to not bump into each other. Simultaneously, there is more communication happening between them than necessary for this task alone. As they approach each other, they traverse through different proxemic distances, where different behaviours are perceived as adequate (see Section 2.1.1). While being in the public distance, both may look at each other freely but will—according to Goffman [Gof63]—avert their gaze at around 2.4 m. This is near the boundary of the close phase of the social distance (see Table 2.1). A similar pattern is observed in an investigation of human greeting behaviour at a social event by Kendon: '[W]e note that people do not usually look at one another continuously as they approach one another, and they may often look sharply away just prior to the close salutation.' [Ken90, p. 163]

2.1.2.2 *Civil inattention*

In many situations, not engaging in a focused interaction becomes an active process in itself. Being copresent in a lift or in facing seats on a bus causes people to actively divert their attention. They stare out of the window, accurately examine their fingernails or play with their mobile phones. This actively displayed non-perception of others, is often referred to as *civil inattention* [Gof63, p. 84]. It can be used to maintain the unfocused nature of an interaction although, e.g. the distance or orientation of the participants favours focused interaction (see Sections 2.1.1 and 2.1.3).

civil inattention

2.1.2.3 *Initiation of Focused Interaction*

As the actions of people in copresence are subject to observation by others, individuals can use the situation to signal their intent to transition into a focused interaction. To this end, they need to establish a situation in which both are ready and willing to perform this transition: 'An encounter is initiated by someone making an opening move, typically by means of a special expression of the eye but sometimes by a statement or a special tone of voice at the beginning of a statement.' [Gof63, p. 91] It can even be much more tentative according to Kendon [Ken90] who observes two individuals *p* and *q* prior to a greeting:

> [W]e see *p* orienting to *q*, but not approaching him until *q* has oriented his eyes to *p*. *p* by his orientation to *q* may be said to announce his intention to approach, but he does not do so until *q* has given his "clearance". [Ken90, p. 170]

Similarly, an investigation of customer-bartender interactions shows that customers fully orient towards the bartender—using their gaze and body orientation—to show their intention to initiate an interaction [GHG12]. When ever people are in copresence with others their behaviour is observed and interpreted. To behave in a socially accepted manner, they need to recognize the intentions of the people in their vicinity and recognizably show their intentions.

2.1.3 *Focused Interaction*

The area in front of a person is best suited for focused interaction. This is where their capabilities of manipulation and reception can be applied most efficiently and parsimoniously. Furthermore, it is the area of which the person has the most control. Whenever people interact with their environment in a focused way—e.g. when they read a book or look into a shop window—they naturally create a distinctive space between them and the object of interest. This space is called the *transactional segment* [CK80, p. 240]. Other people in the situation will try to avoid crossing this area to not disrupt the interaction. The transactional segment of people can vary with their size and activity, but is always a narrow region in front of them (examples can be seen in Figure 2.1).

✎ transactional segment

2.1.3.1 *Face Engagements*

While people can interact with many things in a focused manner, the interaction with other persons has a special significance. Goffman uses the term *face engagement* or encounter for instances of focused interaction between people. He defines them as follows:

✎ face engagement

> Face engagements comprise all those instances of two or more participants in a situation joining each other openly in maintaining a single focus of cognitive and visual attention—what is sensed as a single *mutual activity*, entailing preferential communication rights. [Gof63, p. 89]

When talking about both face engagements and single, unengaged persons, the term *participation unit* can be used [Gof63, p.91]. The definition of face engagement addresses many properties of focused interactions

✎ participation unit

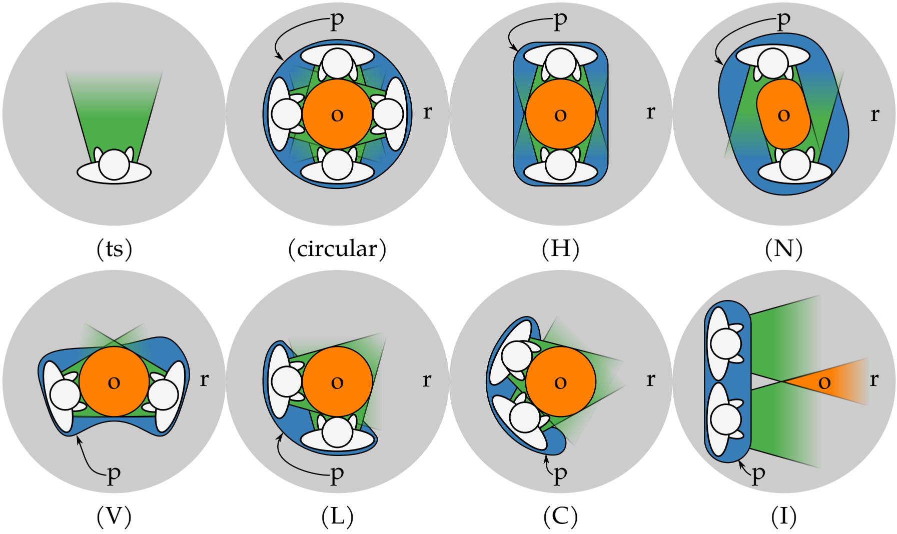


Figure 2.1: Transactional segments, F-Formations and o-spaces, p-spaces and r-spaces as known from Ciolek et al. [CK80]. The transactional segments of persons are shown as green cones in front of them (individually presented in *ts*). Usual F-Formations are shown for groups (*circular*) and dyadic interactions (*H*)–(*I*). The formations (*H*)–(*V*) are also known as *vis-à-vis* and (*L*)–(*I*) as *side-by-side*. The o-spaces, p-spaces and r-spaces of each formation are highlighted in orange, blue and brown and marked with letters (*o*, *p* and *r*) respectively.

between humans. First of all, it naturally requires two or more persons to participate. These persons cooperate on a common task, which requires them to focus their cognitive and visual attention to the same target. Because all participants need to share the same, single focus of attention, one person can not be in multiple participation units at the same time. Finally, this cooperation is exhibited openly. Therefore, other people in copresence can observe and distinguish participation units and to which of them each person in the situation belongs. If not stated differently, I use the term focused interaction as a synonym to face engagement.

2.1.3.2 *Conversational Groups*

The arrangements that people take over when forming a face engagement, depend on factors such as the task at hand, the spacial structure, and the crowdedness of the environment. It is more probable to find face engagements in places that already are spatially distinct or identifiably separated from the rest of the environment. Furthermore, when in an open space—as on a pavement or a campus—the participants of a focused interaction may move closer together than in a confined private room.

When people want to communicate efficiently over a prolonged time, the communication itself is the focus of attention. Therefore, participants need to optimize their mutual reception of each other. They enter an *F-Formation*: 'An F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access.' [Ken90, p. 209] The more cognitively stressful or critical the conversation gets, the higher the preference to establish such a configuration [Aue17a, p. 10]. Being in an F-Formation entails, that the participants orient in a way that their transactional segments overlap and create a 'joint interaction space'. They cooperate and dynamically adapt their position and orientation to maintain this space. They ensure that members have equal access to it and restrict the access of non-members. This separates the environment into three actively maintained spaces (exemplary visualizations of these spaces can be seen in Figure 2.1). The *o-space* is the joint interaction space of the F-Formation with equal access by all participants. The *p-space* is the space occupied by the participants bodies, limbs and belongings. It functions as a barrier that shields the joint interaction space from the environment. Finally, the group is surrounded by the *r-space*. The r-space functions as an extra buffer between the group and other people or groups in copresence. Other participation units avoid this space or—if they need to cross it—show significant avoidance behaviours [CK80, p. 241–260]. When non-participants do not avoid the r-space of an F-Formation, this has a situational meaning. On the one hand, people can enter the r-space to announce their intention to participate [Ken90, p. 231]. On the other hand, people can stay in the r-space as a ratified associate of the focused interaction. In this role they have no direct access to the o-space and can not play an active role—they still are non-participants. They can passively follow the interaction [Ken90, p. 233].

✎ F-Formation

✎ o-space

✎ p-space

✎ r-space

F-Formations can manifest in different forms (see Figure 2.1). When a group of more than two persons enters an F-Formation, they arrange in a *circular* or *semi-circular* manner. This ensures, that each member has equal access to the o-space and excludes external persons to the highest possible degree. Deviations from the circular arrangement and from a uniform distribution of participants within the p-space have a situational meaning [Ken90, p. 216]. Such a deviation is often observed when one or more participants of the group have a special role in the interaction—as e.g. a tour-guide. In addition to the *circular* arrangement, a set of dyadic F-Formations, as they are known from Ciolek et al. [CK80], are shown in Figure 2.1. It can be seen that the *H*-configuration the maximally closed dyadic case. This way the participants achieve the best exclusion of others and have the most control over their o-space. By gradually changing the configuration through *N*, *V*, *L* and *C* the interaction becomes more open to the environment. The interaction occupies less of the visual attention of the participants and allows them to display more attention towards the environment. This allows more control over environment and simplifies joining for non-participants [Ken90, p. 226]. The *I-* or *side-by-side*-configuration represents the most open F-Formation. The participants look into the same direction while standing close enough that their transactional segments overlap. While an o-space is still maintained, the participants loose the ability to observe many of the others non-verbal communication. Instead, they gain the ability to look at the same part of the environment—creating a joint view of the world [Ken90, p. 251].

2.1.3.3 *Conversational Roles*

conversation ✎

speaker ✎

addressee ✎

side-participant ✎

non-participant ✎

This work focuses on *conversations*, verbal interactions that are performed in focused interaction. Conversations are highly organized interactions, in which a set of roles is negotiated, assumed and switched by the participants. Every conversation has a *speaker*. This is the only person—at a moment—who has the right to speak. The roles of the other participants and associates of the conversation depend on the speaker. The speaker in a conversation produces speech acts, which are directed toward *addressees*. In a dyadic conversation the person that is not the speaker automatically assumes the role of addressee [SSJ78]. In a multi-party conversation—when more than two persons converse—the addressee is less obvious. Consequently, the speaker can select the participant or participants that ought to be addressee by naming, pointing, gazing or a multi-modal combination of cues—e.g. by using second personal pronouns and gaze [Aue17b]. Nevertheless, often only one participant is mainly addressed with a speech act. The other participants of the conversational group are *side-participants* (ratified listeners). They are an active part of the conversation and share the responsibility for its success. Although the speech is not directly addressed at them it is produced with them in mind. Finally, people that are not part of the conversational group have a role regarding the group too. Ratified associates can reside in the r-space of the conversational group. In this role they can attend the conversation and may be considered by the speaker during speech production but can neither become speaker nor addressee without first fully entering the conversation [Tra04]. Other people, that are not ratified by the speaker but can hear what is said in the conversation can be called overhearer. Therefore, every person in copresence, that is not part of the conversation or a ratified associate, is a potential overhearer. In this work, I use the role *non-participant* for both ratified associates and overhearers. An exemplary scene with conversing people and their assumed roles can be seen in Figure 2.2.

2.1.3.4 *Turn-Taking System*

turn ✎

The right to speak is a resource of the conversation. This resource is often called the *turn* or conversational floor [Hay88]. It is not owned by one participant throughout the conversation but taken, yielded, and competed for [SSJ78]. The transition of the turn from one participant of

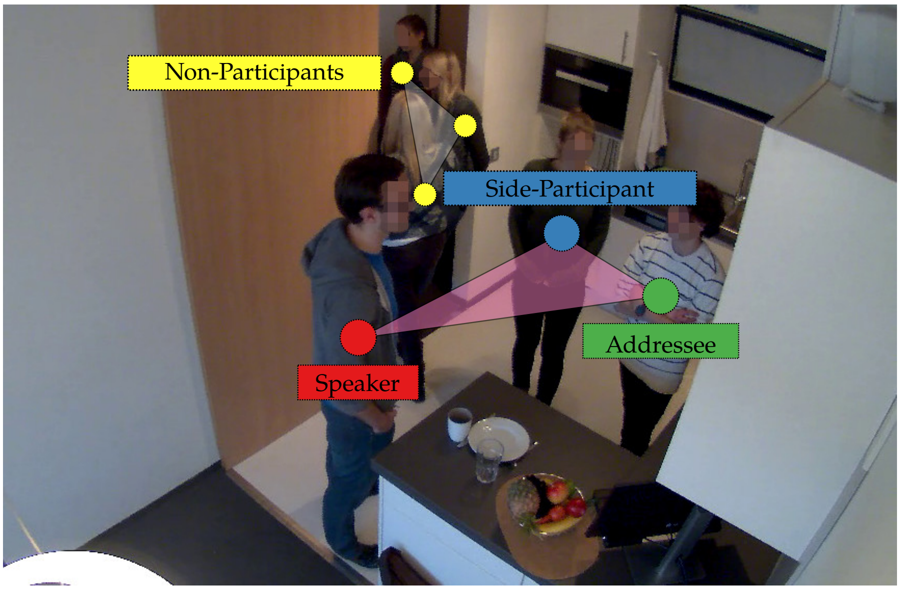


Figure 2.2: An exemplary conversational group in the front (red triangle) and a second group in the back (white triangle). The group in the front consists of a speaker (red dot), addressee (green dot), and a side-participant (blue dot). Other copresent persons are non-participants regarding this group.

turn taking system

the conversational group to another is organized through the *turn taking system*. According to Sacks et al. [SSJ78], it is a system that is 'interactionally managed'. Hence, it is the responsibility of the participants of a conversation to coordinate and ensure its effective application. Furthermore, it considers the transition between the current and next speaker but not the previous and next turn. Therefore, it is a 'local system'. As a result of the turn taking system, only one person can be the speaker at a point of time. Exceptions, occur in case of errors (e.g. miscommunication of the turn transition between the participants), deliberate violations of the system (e.g. interruptions) or termination sequences of an interaction (where people change into simultaneous speech and rhythmical alignment) [Aue17a, p. 3]. In case of an error, a set of repair mechanisms can be applied to resolve the situation—e.g. one of the simultaneous speakers will prematurely stop speaking and leave the turn to the other. The next speaker is negotiated between the participants during the current turn. While current speakers can use their role to select the next speaker—this is often the addressee—and release the turn, all other participants can compete for the next turn. They can apply turn-allocation techniques to communicate their intent to speak or use other means to show their preference for next speaker. This can effectively constraint the current turn by shortening or lengthening the conceded time. Finally, as participants of a conversation are highly trained in applying the turn taking system, the gaps between turns can get as small as 200 ms and even smaller. This suggests that the next speaker is not only established before the end of a turn but can estimate the time of transition and prepare an own contribution to minimize the gap [HE10].

2.1.3.5 *Role of Gaze in Conversation*

Gaze is an important conversational cue. It has multiple, often competing functions in conversation. It can be used to monitor the non-verbal behaviour of co-participants, express own attitudes and feelings, or regulate the conversational floor [Ken67]. As a result, gaze can be used as a predictor for conversational attention and addressing behaviours of speakers. Vertegaal et al. [Ver+01] show that speakers look more often at the addressee of their utterance than at other participants of the interaction. This still applies when the speaker is addressing multiple persons. In this case—although the speaker's attention gets divided between multiple addressees—the speaker's overall amount of attention towards addressees increases and each addressee gets more attention than people who are not addressed. Similarly, there is a high chance of 88% that a listener will look at the speaker instead of a ratified participant according to Vertegaal et al. [Ver+01]. Listeners in a conversation show a set of further interesting behaviours. They use their gaze to show their level of participation [Aue17b]. It can be observed that listeners shift their visual attention to the next speaker around 50 ms prior to the end of a turn, according to Holler et al. [HK15]. The author argues that his shows their ability to predict the time of the transition and the next speaker in advance. As already discussed in 2.1.3.4, such changes of gaze can have further reasons, as gaze does not only indicate a person's attention but additionally is an important cue in the negotiation of the next speaker. This is even more important for speakers, as their influence on the negotiation of the next speaker is higher. A speaker, who looks at a participant at the end of the turn grants privileged access to the counterpart. This privilege can even become an urge if the gaze remains on the participant and no one else self-selects [Aue17b].

2.2 INTERACTION WITH ARTIFICIAL AGENTS

Before I expand on the current state of research on human interaction with different kinds of non-human entities, I want to establish which kinds of entities may be interesting for human interaction in smart environments and which terms I use for the different groups throughout this work. For a visualization see Figure 2.3.

In Section 2.1, I present properties of human interaction with other humans in copresence. The non-human entities that a person can interact with in a smart environment, and which are in the focus of this work, can be categorized into devices, robots, and virtual agents. With *interactive entity*, I indicate any entity that can perceive the actions of another entity and change its internal state on that account. This includes

interactive entity

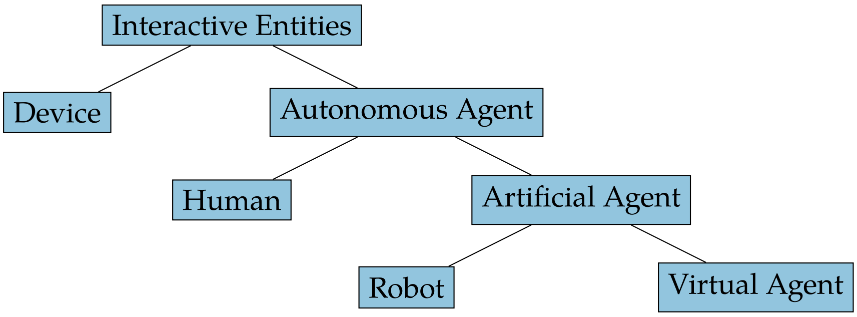


Figure 2.3: The taxonomy of interactive entity groups as used in this document. Subsets are represented as sub-trees and leaves.

devices, robots, virtual agents, and humans. In case of *devices*, the action can be as simple as pressing a button, performing a fixed gesture or speaking a command. The resulting state change can be a change of the lighting if the device is a lamp or the volume if the device is an amplifier. To distinct *autonomous agents* from other kinds of interactive entities, the definition of Dautenhahn [Dau98] can be used: 'Autonomous agents are entities inhabiting our world, being able to react and interact with the environment they are located in and with other agents of the same and different kinds.' [Dau98]. Furthermore, an autonomous agent has its own believes, and goals and the interaction is a way to achieve these goals. Although, many *artificial agent* are not designed in a way that explicitly models their goals and how to approach them, they are always designed by humans for a specific purpose which is served trough interaction. A *robot* is an artificial agent associated with an embodiment that occupies physical space. Robots may manipulate things, navigate through space, or reconfigure themselves—thus, altering the availability of space. A *virtual agent* does not occupy physical space as such—thus, its movements do not necessarily change the availability of space. Nevertheless, virtual agents can have an embodiment that is visualized in some way—e.g. on a screen, as a spot of light, or in form of a loudspeaker—and where attention can be directed to. According to this taxonomy, IPAs are virtual agents. For simplicity, the term *agent* is used synonymously with artificial agent throughout this work. If not explicitly stated, the term Human-Agent-Interaction (HAI) is used for interactions with both robots and virtual agents while Human-Robot-Interaction (HRI) is used for interactions with robots only. Whether the smart environment as such can act as a single device, as an agent, or a combination of both strongly depends on its interface and the intended way of usage. It may be perceived as a single device with many functionalities or just as a casing for other devices and agents. Similarly, it can be perceived as one big artificial agent or as a device that is controlled by an agent.

✎ device

✎ autonomous agent

✎ artificial agent

✎ robot

✎ virtual agent

2.2.1 *Agents in Unfocused Interaction*

Humans communicate mainly with other humans. Therefore, it feels natural for us to model robots and their communication skills according to our understanding and expectations of communication. Social spaces and unfocused interaction as known from Human-Human-Interaction (HHI) can—to some extent—be similarly applied to human interactions with robots or virtual agents. The following works illustrate that the copresence of artificial agents has an impact on human behaviour and show ways to handle such situations.

People apply proxemic rules to artificial agents. It has been shown that people concede a personal distance to robots and virtual agents and react to intrusions of artificial agents into their own personal distance similarly as one could expect it in Human-Human-Interaction (HHI). Takayama et al. [TP09] evaluated how mutual gaze influences the distance people feel comfortable when approaching a robot and when a robot approaches them. Their results support the idea that the proxemic rules regarding the size of the personal distance people employ in interaction with robots are similar to the rules in HHI. In a comparison of reactions to approaching humans and robots, Sardar et al. [Sar+12] show that people tend to show even more compensatory behaviour with robots than with other humans when they enter their personal distance. A similar investigation of human interaction in a virtual reality is done by Bailenson et al. [Bai+01]. They compare the proxemic behaviour of people in interaction with a humanoid virtual agent and with a geometric object (a pylon). They show that people grant the agent more personal distance than they grant the pylon. Furthermore, this personal space grows when the agent's gaze behaviour gets more realistic. This shows that the copresence, form, and actions of artificial agents in unfocused interaction affect human behaviour.

Like people, robots can use the public space to initiate conversations. Holthaus [Hol14] investigated how a robot can behave towards people in different personal and social distances to support a dyadic interaction. He shows that the user interaction can be enhanced by employing strategies such as gradually increasing attention towards approaching people and proactively greeting at appropriate distances. Shi et al. [Shi+11] present a model for the initiation of a conversation by a robot. To this end, they observe a scenario in HHI, where a shop owner welcomes customers and presents a product. They use these observations to develop a set of positioning rules that can be utilised in different phases of the interaction. Their robot's positioning is rated better, when it behaves according to the found rules in contrast to always greeting and always standing by the product. Satake et al. [Sat+09] present a robot that approaches people in a mall to give recommendations. By approaching people from the front at public distance, the robot shows its presence and intention to interact. It prompts the initiation of conversation non-verbally by facing them directly at social distance. Verbal interaction is only initiated when people stop by the robot. The authors can show that this approach results in more successful interactions than approaching on the shortest path and directly starting to talk at social distance. With their behaviour in copresence, robots can prompt people to initiate a focused interaction. Therefore, it is important for robots in unfocused interaction to behave in a manner that signals their intention.

Even when not intending to communicate with humans, an artificial agent in copresence needs to understand human communication to behave in a socially acceptable manner. This is actively investigated in the field of social navigation. To this end, Lindner et al. [LE11] created a taxonomy of social spaces by defining five types of spaces. In this taxonomy, Proxemic space[1] describes the zones used in Hall's proxemics. The activity space corresponds to the activity of one or more agents. The affordance space corresponds to a potential activity. Territory space—e.g. a fenced area or closed room—may not be entered without permission. Furthermore, the space affected by an activity—e.g. by noise or odour—is described as the penetrated space. These spaces may overlap but do not necessarily need to fully contain each other in contrast to proxemics. Knowledge about these spaces is used to navigate while better respecting the personal space and activities of other agents. Similarly, Rios-Martinez et al. [Rio+12] generate navigation plans by considering the personal distance, information process space and o-space of people to reduce discomfort. An overview of different notions of social spaces and robot navigation in copresence with people is presented by Rios-Martinez et al. [RSL15]. These works intend to enhance the acceptability of the way robots navigate based on interactional, conversational, and social aspects of the usage of space in copresence.

2.2.2 *Agents in Focused Interactions*

It is widely accepted, that the principles of human conversation (Section 2.1.3) can be transferred to human interactions with artificial agents. Spexard et al. emphasize that '[t]he user should be able to communicate with the system by, e.g., natural speech, ideally without reading an instruction manual in advance' [SHS07]. Similarly, Dautenhahn et al. [Dau+05] found that 71 % of the participants of their study wished that a robot companion would communicate in a 'human-like' manner. She further discusses:

> The fact that subjects wanted a robot companion to have humanlike communication was not a surprising one, as it is a natural human instinct to want to communicate using

[1] Original term *personal space* changed to proxemic space to avoid confusion.

> speech and gestures that are recognisable by humans. [Dau+05]

In the following, I show that the effect of an agent's behaviour on the course of a focused interaction can not be ignored. Agents need to understand the social signals humans show and how their own behaviour is perceived by humans. Furthermore, I present literature on the problems of addressee recognition, turn taking behaviour generation, and the detection and utilization of conversational groups from the perspective of computer sciences.

2.2.2.1 *Impact on the Perception of Interaction and Human Behaviour*

As in HHI, in focused interactions in HAI, a common goal must be approached in a collaborative manner—otherwise there would be no need for interaction. Simultaneously, because people are highly trained to observe their interaction partners and assess their beliefs and intentions, every property of the agent's behaviour is under evaluation and can affect the interaction. When an agent does not perceive its interaction partner, its behaviour still has implications.

In a comparison of affective behaviour generation in story-telling performed by a humanoid robot, Rosenthal-von der Pütten et al. [RKH18] find that human-like—and in parts robot-specific—non-verbal behaviour can increase the perceived animacy of the robot and the participant's willingness for self-disclosure. Furthermore, when an agent can observe its interaction partner it can dynamically adjust its behaviour. Kopp et al. [Kop+18], present a virtual agent that recognizes human backchannel signals and produces multi-modal conversational cues previously extracted from human interactions. In a user study, the authors show that the approach allows participants to successfully apply repair strategies. A robot that observes the participants gaze and—anticipating a choice—reaches for objects in a collaborative ordering task is created by Huang et al. [HM16]. The authors show that such a behaviour conveys the impression that the robot is aware of the users choice. It is apparent that the behaviour of agents during a focused interaction with humans can have a strong influence on the development of the interaction and the human perception of the agent. As gaze plays an important role in conversation (Section 2.1.3.5), the generation of eye-gaze has a high impact on the interaction too. It can be used to generate turn taking-cues, greatly enhancing the efficiency and perceived quality of the interaction with a virtual agent [CT99]. Andrist et al. [And+14] present a humanoid robot that shows gaze aversions. A gaze aversion at the beginning of an utterance is thereby ought to display internal processing. By displaying aversions between two utterances the agent tries to hold the conversational floor. In a corresponding interaction study, the authors show that such a behaviour not only increases the perceived thoughtfulness of the robot but also allows the robot to keep the turn longer before getting interrupted. An in-depth overview on generation of eye-gaze for virtual agents and its effects is performed by Ruhland et al. [Ruh+15]. All of these works investigate dyadic interactions between an artificial agent and a person. Therefore, their need for the distinction of addressees and conversational roles is limited.

However, in multi-party interactions, agents can have different conversational roles and a strong influence on their distribution. Mutlu et al. [Mut+09] conducted an experiment in which a robot leads two persons through a travel consultation. The robot leads the participants through the interaction by talking and occasionally asking questions about their preferences. While doing so, it communicates different conversational roles to the participants by applying gaze behaviour known from HHI. As a result, the participants of the study nearly always adhere to the imposed roles. As a side note, Mutlu et al. [Mut+09] observe that in some cases repeatedly addressed participants pass their turn to the side-participant. In these cases the authors assume that the addressees feel uncomfortable with the other participant being ignored and try to involve them in the interaction. In this set-up a wizard recognizes speech and activates the robots reactions. The system can not automatically recognize the situation to generate appropriate robotic behaviour. A similar effect on the conversational role distribution can be achieved with agents in virtual reality as shown by Pejsa et al. [PGM17]. In this set-up, a participant is interacting with two virtual agents (all three have fixed positions). By manipulating the agents orientation and gaze behaviour the conversational role of the human participant is affected. This manifests in adapting amounts of the participants total speaking time. In this set-up a speech recognition system is used. The agents take turns with the participant based on recognitions of single speech acts and always show the same (inclusive or exclusive) behaviour. As the agents ask questions and wait for an answer, no distinction of the participants addressee is performed. Matsuyama et al. [Mat+15] introduce a robot-mediator to a conversation that equalizes this distribution. Their robot takes part in a conversation as a fourth participant. When one person withdraws from the conversation it acquires the turn and then draws the person back into the discussion through addressing. They evaluate this behaviour by showing a recording of such an interaction and letting people rate the robot's behaviour. Observers rate the robot's behaviour as more acceptable and with a higher level of groupness, when it acquires the turn and waits for approval before addressing the other participant. While the generated behaviour is rated effective, the model for the detection of withdrawing participants and appropriate moment for intervention are only evaluated on a synthetic data.

People prefer robots that comply with group arrangements which are known from HHI (as presented in Section 2.1.3.2). In a cross cultural study Joosse et al. [Joo+14] confronted people with images of a family in circular F-Formation with a robot positioned in changing distances to the groups centre. While they found different preferences in people from China, USA, and Argentina the overall predominant preference was for configurations where the robot stayed out of the groups o-space, and somewhere within the p-space. This is in agreement with how humans create F-Formations (Section 2.1.3.2). Similarly, Hüttenrauch et al. [HTS09] perform a study in which people guide a robot through their home and present different objects and rooms to it. An analysis of the recorded interactions reveals that the participants prefer to assume *vis-à-vis* formations and an o-space-size as known from HHI. Finally, adaptivity of humans when it comes to maintaining F-Formations, is utilized by a museum-guide robot presented by Kuzuoka et al. [Kuz+10]. By changing its bodily orientation, the robot can change its conversational group from a *vis-à-vis* formation into an *L-shaped* formation. Thereby the groups focus can be directed to the exhibit that the robot is talking about. While these works show that people form and maintain conversational groups with artificial agents similar to interactions with other humans, none tries to detect conversational groups.

The presented works show that many of the observations drawn from HHI are transferable to HAI. They show that it is possible for artificial agents to influence the interaction according to their goals. However, while showing the potential of HAI, the recognition of human conversational cues is out of the focus of most of these works. To use these effects autonomously, artificial agents first need to understand the behaviour of their human interlocutors.

2.2.2.2 *Automated Addressee Recognition*

Addressee recognition is an important sub-problem in the recognition of conversational roles and often investigated separately. As RQs 1 and 2 mainly focus on this problem, I present literature that specifically focuses on addressee recognition. In HHI, the identification of addressees of an utterance is important for automatic conversation analysis. Jovanovic et al. [JAN06] present a Bayesian Network classifier that predicts the addressee of an utterance in a four-party meeting. They incorporate lexical features of the utterance, information about the previous turn, the type of meeting and how often participants look at each other. Using all these features they achieve an accuracy of ≈ 82%. They can achieve ≈ 73% accuracy with context information—the speaker, addressee, and dialogue act of the previous utterance—alone. Furthermore, they note that gaze is not an important addressee indicator in their set-up. This may stem from the circular seating arrangement in their corpus and the existence of external gaze targets as whiteboards and notes. Akker et al. [AT09] compare this approach with other recognition methods using the same corpus and an adapted feature set. The authors show that a simple, rule based addressee recognition can achieve a similar accuracy as the Bayesian Network approach—≈ 65% vs ≈ 62% using their feature set. Furthermore, when using only information about the gaze distribution of the speaker during the utterance, they already achieve an accuracy of ≈ 57%. Both approaches use manually annotated information as input for their addressee recognition systems. An automatic addressee detection system for three-party HHI is presented by Takemae et al. [TO06]. The authors use a gaze prediction from the participants head rotations to calculate relative amounts and frequencies of looking at people and mutual gaze between them. These measures are then combined using Naïve Bayes to estimate whether a single person was addressed or the whole group. In case of a single addressee, the person with the largest amount of gaze from the speaker is chosen. After manually removing back-channel utterances, they achieve an accuracy of ≈ 74% on their dataset. These works concentrate on HHI, so no non-human entities can be addressed. Additionally, the participants are equipped with microphones or placed at specific positions, so the set-ups are rigid. Furthermore, only Takemae et al. [TO06] does not require pre-annotated information to decide, to whom an utterance is addressed.

Research on interaction with virtual agents focuses more on fully automatic recognition systems that potentially can be used in interactions. A mixed multi-party interaction between two humans and a computer screen is presented by Turnhout et al. [Tur+05]. In this scenario the two persons discuss an excursion through the Netherlands and fill a corresponding form through verbal interaction with a computer screen. The system is controlled by a wizard and the interactions annotated afterwards to investigate addressee recognition in such a scenario. By using a Naïve Bayes classifier to combine information about looking behaviour, utterance features and the systems dialogue state, the authors achieve an Area Under the Curve (AUC) result of ≈ 0.81 for the decision whether the agent was addressed or not. They further observe, that adding the screen into the interaction strongly biases the distribution of the participants' gaze towards the monitor. When speakers address the monitor they focus it in ≈ 95% of the observations. When they address the other participant, they look at them only in 42% of the cases. A system that does not rely on pre annotated features and therefore can work autonomously is presented by Huang et al. [HBN11]. The used scenario is a travel planning set-up with a humanoid virtual agent. During the study, the agent is controlled by a wizard. The system approximates the participants Visual Focus of Attention (VFoA) by detecting head rotations and uses duration and focus-change features in combination with prosodic information. By applying a Support Vector Machine (SVM), they achieve an accuracy of around 80% on pre-segmented utterances. Both systems perform addressee recognition for each recognized utterance. An approach that decides on the agent's conversational role while the person speaks and acts accordingly is proposed by Vertegaal et al. [Ver+01]. The authors transfer HHI-principles to HAI by investigating three-party HHI-discussions and extracting addressing behaviours for their virtual agents. They use their insights to create a multi-agent conversational system—two faces on a screen—for interaction with single persons. Gaze tracking and utterance analysis are used to decide the conversational role for each agent and generate corresponding attentional behaviours (looking at the speaker when addressed or otherwise at the addressee). In this set-up the addressee is the agent that the speaker looks at during the utterance and the other is the side-participant. While this system can both recognize which agent is addressed and generate corresponding behaviour, it is created to always interact with a single person. A fully autonomous virtual agent, that can interact with groups of people, is presented by Bohus et al. [BH11]. In this set-up the speaker of an utterance is detected through sound source localization and the addressee is defined as the speakers VFoA. Furthermore, long utterances and non-understandings are assumed as not addressed towards the agent. They suggest to combine their sound source localization with visual cues to enhance their addressee detection.

Similar approaches are made in HRI research. A simple interaction between two persons and robot—consisting of a camera and a microphone—is presented by Katzenmaier et al. [KSS04]. In this scenario one person presents the robot to the other, directs commands toward it and discusses its advantages and disadvantages. In the analysis of their recordings, the authors observe a similar effect as Turnhout et al. [Tur+05]. When the presenter looks at the robot, the robot is addressed in ≈ 65% of the observations. In the remaining observations, the other participant is addressed. When the other participant is looked at, the robot is almost never addressed. They combine an addressee predictor that uses the presenters head orientation with one that used acoustic features to achieve an accuracy of ≈ 92%. An interaction with a humanoid robot can be found in the work of Jayagopi et al. [Jay+13] and Sheikhi [She14]. This is a Wizard-Of-Oz (WoZ) scenario, in which the robot performs a quiz with two persons. In this scenario, VFoA of all participants and the robots dialogue-context information are used to classify the addressees of utterances. With this feature combination, an accuracy of ≈ 82% can be achieved, when the possible VFoAs are reduced to only the participants of the interaction. Both systems do not recognize the addressee during the interaction. A multi-modal approach is presented by Lang et al. [Lan+03] where potential interaction partners in the vicinity of the robot are tracked. Using sound source localization, the robot directs its attention towards a person that has a high probability of talking. If the robot is additionally faced by the person, it is considered addressee. A similar approach with a robotic head is used by Skantze et al. [SJB15] in a three-party card game scenario. They distinguish speakers by using close talk microphones and their addressees from the speakers VFoA. This interaction if further investigated by Johansson et al. [JS15] to include different information into the decision. From information about head poses, part of speech, card movements, prosody and the robots dialogue state they train a Multilayer Perceptron (MLP) that decides whether the robot should react to an utterance or not. These systems use acoustic information to detect the current speaker and information about this participant to infer the addressee.

Most HAI-research in which addressee recognition is performed, does that on the basis of utterances. The systems incorporate multiple sources to decide whether a recognized utterance was addressed at the agent or not. If people want to participate in such an interaction, the need to equip themselves with a microphone or reside in the agents Field of View (FOV). The agent, either assumes all observable people to be part of the conversation or performs dyadic interaction with the most prominent person.

2.2.2.3 *Turn-taking behaviour generation*

As seen in Sections 2.1.3.3 and 2.1.3.4 conversational roles and turn taking are intertwined problems. An agent that interacts with another agent necessarily conducts turn taking. Nevertheless, most systems do not explicitly model this behaviour as such.

A few investigations concentrate on how an agent can generate behaviour to actively shape the course of the interaction. Skantze et al. [SJB15] generate turn taking behaviours with their robot during a card playing game and show that they can be applied effectively. They can show that focusing on a participant after asking a question increases the probability that this participant will take the turn. They further show that—by generating filled pauses or smiling and looking away—the robot can successfully obtain a yielded turn and prevent others from taking it. Finally, they observe that the robots gaze has an impact on the next-speaker selection even if the robot is a side-participant. While this system actively uses turn taking, it only does that after getting addressed and to overcome its processing times.

Other systems investigate the correct detection of turn changes. To better recognize the end of an utterance Bilac et al. [BCL17] apply a multi-modal approach too. They recognize filled pauses and gaze aversions produced by human interaction partners to better distinguish between the end of a turn and hesitations. With this approach they can reduce the number of interruptions made by the robot, allowing their participants to produce longer utterances. A similar target is pursued by Lala et al. [LIK18]. In this work, the authors use different types of dyadic conversations with a remote controlled humanoid robot. They train models to predict when a turn is transferred from the participant to the robot based on acoustic and linguistic features. These systems do not model turn taking.

A multi-party interaction system with a turn manager is presented by Żarkowski [Żar19]. In this scenario, a robotic head plays a trivia game with two persons. Instead of always reacting to speech after a short silence, the turn manager ensures that the robot has the conversational floor before talking. It assumes to have the turn when the last speaker looks at the robot or when silence persists. Furthermore, the robot pays visual attention to both interaction partners and grants them more discussion time after asking a question. With these changes, the authors can increase the percentage of correct turn exchanges from 51.5% to 80.5%. An explicitly modelled turn taking system for an artificial agent is presented by Bohus et al. [BH11]. The agent classifies peoples gaze and speech as turn management actions. The entity in the speaker's VFoA is perceived as addressed and obliged to take the turn at the end of the utterance. If the addressee rejects the turn, the other participant can take it—when the agent was not addressed but no one else answered for a specific time duration it still can take the turn. The agent accepts interruptions by yielding the turn. It looks at the speaker when listening, at the intended addressee when speaking and in between turns at the participant that ought to be the next speaker. This system implicitly models the conversational roles speaker, addressee, and side-participant. The act of taking a turn that is not claimed or yielding in case of interruptions, can potentially repair misunderstandings in the turn management. While this approach tackles multiple requirements of conversations, it is only applied at the end of utterances, when the conversational floor is transferred from one participant to another. To generate behaviour during a turn, an explicit model and recognition of conversational roles is additionally required.

2.2.2.4 *Conversational Group Detection*

Conversational groups and their detection in computer sciences are investigated from the side of HHI-conversation analysis as F-Formations. Most approaches use the positions and orientations of copresent people to estimate their probable affiliation to groups. Evaluations are performed on corpora that show people freely communicate, create, and change conversational groups in an open space—e.g. at coffee breaks or poster presentations. Cristani et al. [Cri+11a] inject uncertainty into their pose estimations and use a voting algorithm on a regular grid to find o-spaces and the corresponding participants. They evaluate using a synthetic and two realistic datasets to achieve a precision of 0.75 and a recall of 0.86. They simultaneously optimize these estimations to achieve a better results than other state-of-the-art approaches. Setti et al. [Set+15] formulate the problem of detecting F-Formations as a graph-cuts problem and present an extensive comparison of methods on multiple datasets. Furthermore, their approach achieves the best performance in an initial evaluation of recognizers on a new dataset [Ala+16]. In the work of Zhang et al. [ZH16], the notion of F-Formation is extended to consider *ratified associates* (Section 2.1.3.2). By detecting persons that are not full participants of the conversation they can enhance the overall performance of F-Formation detectors. A multi-modal approach using accelerometer data and speech activity information from a worn sensor-device is presented by Hung et al. [HEC14]. They use the data to recognize dyadic conversational groups. Ricci et al. [Ric+15] and Varadarajan et al. [Var+18] argue that the recognition of persons orientations and affiliation to conversational groups can both benefit from their high interdependency. They use model this interdependency to simultaneously enhance both predictions. Similarly, Alameda-Pineda et al. [ARS18] exploit the inherent coupling of the human head and body pose together with their temporal consistency and multi-modal data. They enhance the prediction of persons head and body orientations by formulating it as a matrix completion problem and show that this can further increase the quality of F-Formation detectors. Furthermore, they confirm that F-Formation detection works better on the basis of body orientations than head orientations. As Alameda-Pineda et al. [ARS18] propose, the results of F-Formation detection can be used for further, higher level analyses of social interactions. Cristani et al. [Cri+11b] correlate the physical distances of people in an interaction with their social relations. They can show that there is a high correlation only when one considers the geometric constraints that F-Formation impose on the situation. None of these works considers the presence of artificial agents in such an interaction.

2.2.2.5 *Utilizing Conversational Groups*

In the context of HAI, the detection of F-Formation is not the focus of analysis. Therefore, the problem of recognizing conversational groups is not reported on (this applies to the majority of work presented in Section 2.2.2.2). Nevertheless, there is some work that tries to utilize the properties of F-Formation to enhance the acceptability of robots. Rios-Martinez et al. [RSL11] enable a robot to consider conversational groups in its navigation. To this end they formulate the crossing of personal distance and o-space as a navigational risk (the rink of acting socially inappropriate). This allows their robot to successfully avoid disturbing conversational groups. As a side effect, when the navigational goal is set to the centre of an o-space, the robot approaches the group and positions itself automatically within the groups p-space. A robot that actively seeks to join conversational groups and blend in is presented by Althaus et al. [Alt+04]. It follows three simple rules: (1) approach a person, (2) detect other persons and their orientation in vicinity, and (3) maintain orientation and distance to the centre of the group. A similar effect is achieved by Repiso et al. [RGS18]. In this work a force model is used to navigate a robot, accompanied by a person, to a second person to form a group. The robot stops at a specific distance and rotates towards the centre of the group to optimize engagement. The authors use proxemics as their quality measure. A transfer from HHI to HAI of how an F-Formation can be actively assumed is performed by Yamaoka et al. [Yam+10] and Shi et al. [Shi+15]. The authors analyse how presenters position themselves when explaining an object to someone else. They implement a behaviour that positions the robot in a way that the robot and listener are in each other's and the object in both FOV. It furthermore, maintains a specific distance to the listener and object which results in an L-shaped (see Figure 2.1) F-Formation. The presented works utilize properties of F-Formation to enhance robotic navigation or allow them to approach a group of persons. However, they model these groups implicitly and do not detect F-Formations, analyse their properties, or use them to distinguish interlocutors of the robot from copresent people.

When it comes to interactions in virtual environments the generation of conversational behaviour between multiple agents and their influence on humans is of high interest. Rehm et al. [RAN05] present a virtual environment in which virtual agents can wander around and create conversational groups. In a user study they can show that people prefer joining open groups over closed formations (see Figure 2.1 for a visualization of both kinds of groups). Cafaro et al. [CRA16] generate conversational behaviour of virtual agents and let persons join the group or navigate to a goal behind it. In this scenario, the agents stand in a circular F-Formation with differing distances while showing different signals of friendliness within the group and towards the participant. The authors of these works do not elaborate on how the detection of conversational groups in their scenarios could be performed.

2.2.3 *Summary*

Artificial agents affect the behaviour of people in copresence. On the one hand, people grand them a form of proxemics by assuming similar distances as in interactions with other people. Although this effect depends on the embodiment of the agent, it can be measured with a wide range of embodiments. On the other hand, agents can actively influence the interaction with a person. They can signal that they want to enter or leave a focused interaction. They can assume and change F-Formations. Within a group, they can regulate the distribution of conversational roles through their gaze or by acquiring and yielding conversational turns. Furthermore, human conversational groups can be detected and the roles an agent assumes in conversations too. The effects are measurable and the systems can achieve acceptable results in their respective scenarios. However, there is little work on agents that combine all of these possibilities into one system. In the following section I investigate whether such observations can be transferred to interactions with devices and smart environments and how human interaction with them may look like.

2.3 INTERACTION WITH DEVICES AND SMART ENVIRONMENTS

Making the distinction between focused interactions and unfocused interactions in human interaction with devices and control of smart environments is not as obvious as it is in HAI. As devices usually only react to inputs, they are acted upon but can not actively participate in a focused interaction. However, smart environments can utilize the presence of inhabitants and use their actuators to engage in ways that may or may not urge people to actively engage with them. To point out this difference, I start this section with some examples of unfocused interaction with smart devices. In the main part of this section, however, I present how focused interaction with devices and smart environments can be performed using different modalities and what implications this has for addressee recognition.

2.3.1 *Unfocused Interaction with Devices & Smart Environments*

In contrast to HAI, it is not obvious how human unfocused interaction with devices and smart environments manifests. However, as the Media Equation applies to interactive and communicative devices, they can use notions of space. Greenberg et al. [Gre+11] present a variant of proxemics for human interaction in ubiquitous computing (ubicomp). In contrast to Hall's distances between people, this theory aims at interactions between all kinds of entities. The authors distinguish different dimensions—distance, orientation, movement, identity and location. An architecture for the creation of interfaces that use this notion of proxemics is presented by Marquardt et al. [Mar+11]. The authors present exemplary use cases which analyse peoples relative positions and posture to adapt the way information is presented on an advertisement display or for the control of music and games. The idea of an interactive advertising display is further explored by Wang et al. [WBG12]. Their system tracks people in front of it and presents products by trying to attract and keep the peoples' attention. The used strategies change with the persons relative position and orientation. Sørensen et al. [Sør+13] present a multi-room music system. Based on peoples positions and movements, the music can be automatically adapted to *follow* them through their home. The control interface—on a smart phone—additionally adapts according to their location to allow control of the music in their direct vicinity. These systems adapt to the behaviour of people in copresence without actively engaging.

Additionally, one can take advantage of peoples receptiveness to their surroundings to design information displays. Cha et al. [CLN16] present a lamp that unobtrusively shows the inhabitants their amount of physical activity by changing the transparency of its shades. Leichsenring et al. [Lei+16] aim to raise people awareness of their of water consumption. To this end, the sound that is produced by the flowing water is captured, amplified, and played back. Similarly, Groß-Vogt et al. [Gro+18] artificially increase the reverberation of a kitchen to make the average electricity consumption perceptible to inhabitants. A matrix of led lamps is used to signal recommendations for inhabitants to change light, door and blinds states in a work presented by Domaszewicz et al. [Dom+16]. This way the information can be perceived, but people are not forced to actively engage with it.

2.3.2 *Focused Interaction with Devices & Smart Environments*

There is a lot of current research on human interaction with smart homes, some of which contain artificial agents. However, the recognition of addressing behaviour and differentiation of addressed entities is often out of focus. In the following, I give an overview of different ways an inhabitant can interact with devices in a smart environment in a focused way using different modalities.

2.3.2.1 *Touch & Gui*

Touch, especially in the form of switches and buttons, is the prevalent way of controlling technical devices. It promotes a strong coupling between the area that is touched and the controlled functionality—one switch controls one set of lights. When this coupling is dissolved, e.g. by using remote controls and Graphical User Interfaces (GUIs), new metaphors are required to communicate the addressee and functionality. An interesting approach to smart home control with a single, yet simple remote is presented by Sandnes et al. [San+17]. In this proposal, the remote control—a disk-shaped device with dial and click possibilities—recognizes the spatially closest device and provides a set of manipulation possibilities within this context. Therefore, a person can switch a lamp by moving close to it and clicking, or change the radio-volume by moving there and using the dial. This is a good example for the trade-off that has to be made between simplicity and functionality in the design of an interaction. The complexity of device selection—the addressing—is traded for the possibility to control devices from a remote location. The simplicity of the remote, at the same time, only allows to control a small set of functionalities of a device.

This trade-off can be addressed by using GUIs. A GUI can ease the device selection through different metaphors and afterwards present a dedicated control interface for the device. The selection can be achieved by providing a menu, a 2D representation [PLH19],[2] or a 3D representation of the premises [BRT02]. Borodulkin et al. postulate that the 3D view design is realistic and the interaction is intuitive. Augmented reality can be used to further widen these interaction possibilities. Seifried et al. [Sei+09] present a system, in which the user sees a live image of the current room, augmented with control menus on a couch table-display. Thus, touching a device on the image allows access to its controls and the live video provides a direct feedback of the results of the interaction. A similar, but mobile approach is realized by Pohling et al. [PLH19].[3] In this approach a person can freely move around. A smartphone application shows a live view from the phone's camera, augmenting the video stream with menus at the locations of controllable devices. GUIs are controlled by touch interaction. Either directly on a screen or through pointing devices such as a computer mouse. Furthermore, the interaction is not performed directly with the controlled device, but with its representation in the GUI.

By projecting a GUI into the environment simple objects can be augmented with further functionalities. Such a system that supports people in preparing a meal is presented by Neumann et al. [Neu+17]. To this end, a projector is used to highlight ingredients on the kitchen counter. Cookware is augmented with information about as the current and required level of filling or temperature. The interaction is done implicitly by performing the required cooking steps or explicitly by via touch and gestures. A projected GUI for smart home control is presented by Pizzagalli et al. [Piz+18]. In this case the projection allows the control of basic functionalities like light and temperature via touch.

2.3.2.2 *Gestures*

When an interaction is desired, that is independent from the persons position but not performed with a remote control, gestures can be used. To switch lights and open or close curtains, Kim et al. [KK06] propose eight distinct gestures. In this case, a gesture encodes the addressee and desired action simultaneously. However, combining the selection of the addressee and the action into one single gesture results in a steep combinatoric growth, requiring a distinct gesture for each combination of device and functionality. Therefore, most systems keep the addressee selection and control action separated. Mayer et al. [MS14] extract the addressed device from the inhabitants gaze using smart glasses. Similarly, Budde et al. [Bud+13] detect pointing gestures using a Kinect. Both suggest using a smartwatch or smartphone to further control the selected device. Although they rely on a GUI, they allow selecting the addressee with a different modality. The GUI can automatically show the interface of the addressed entity. Carrino et al. [Car+11] use a

[2] *BCozy - A Location Based Smart Home User Interface* by Marian Pohling et al. [BCozy]

[3] *Augmented Reality Smart Home Control App* by Julian Daberkow et al. [BComfy]

dedicated pointing device to specify the addressee, which then can be interacted with using gestures or speech. They argue that communication through deictic gestures, symbolic gestures and speech commands is natural. Inversely, Kühnel et al. [Küh+11] select the addressee via a smartphone screen and control the functionality through gestures. This way they can use the same gesture for lowering the blinds as for lowering the volume. An interesting alternative—presented by Verweij et al. [Ver+17]—is to visualize unique movements, which a person can follow with a gesture to activate a specific functionality. The addressee is found by correlating the gesture with the displayed movement. In this case, gestures do not need to be learnt beforehand and work with all devices that can display some kind of motion.

To efficiently communicate using gestures, a vocabulary of gestures is needed. Most of the presented systems, therefore, combine gestures with other modalities to reduce the size of the required vocabulary. Nevertheless, for most people this is not the primary way of communicating—neither with devices nor with other people. Therefore, such interactions create the need for prior training. Furthermore, as the addressee is not necessarily inherent to the gesture, the problem of communicating the addressee gains additional importance.

2.3.2.3 *Speech*

For verbal communication, people already have a huge vocabulary and know how to use it to convey their intentions. Although, arbitrarily complex information can be transmitted via speech, the addressee in human conversations is often not included in the words but displayed with other modalities or inferred from the context (Section 2.1.3.3). Portet et al. [Por+13] perform a WoZ study, in which the participants need to trigger some smart home functionalities during an interview. To solve this task, most participants use indirect speech acts like 'it's time to lower the blinds' or direct commands like 'lower blinds'. While they say they would prefer directly stating their intent 'to the home' instead of speaking with a device or robot, it is not further investigated how they would address the home. Modern artificial agents can not cope with the complexity of free human speech. They need to narrow down the possible interactions and simplify the problem. Therefore, most IPAs are activated using a keyword. Although this keyword is not always a name, it is a direct, verbal addressing from the agents point of view. The speech that is following the keyword can be interpreted as a command or question directed at the agent. For human interaction with a smart home, Potamitis et al. [Pot+03] recognize combinations of `<agent>` and `<command>` which can be surrounded and interleaved by arbitrary words. The recognized `<command>` is then passed to the chosen `<agent>`. The authors argue that interaction via speech is user-friendly and that calling the agent by name is robust. Authors often do not further elaborate on how the distinction of addressees can be performed in robot inhabited smart homes. In the work of Park et al. [Par+07] and Park et al. [Par+08], different agents can be addressed to control a smart home by 'naming or pointing'. Similarly, Seung-Ho Baeg et al. [Seu+07] and Gross et al. [Gro+12] present HRI scenarios in smart homes but give no information on how they decide who is addressed. Other scenarios present GUIs for smart home control that can be used with verbal commands [VKH13; ZZC16] but do not distinguish between different addressees. A study that investigates the addressing behaviour of naïve users in a smart robotic flat is presented by Bernotat et al. [Ber+16]. The study shows empirically which modalities and interfaces people prefer to use for a set of daily tasks. It is not further investigated in which ways the participants convey the addressee of their communication.

If one does not want to have a remote, gesture, or name for each function of a smart environment, the coupling between addressee and command needs to be eliminated. Therefore, command and addressee need to be encoded separately. In contrast to touch, gestures and speech in general do not contain the addressee. Therefore, the addressee needs to be actively encoded in the command, localized using additional modalities or inferable from the situation. While different ways of displaying the addressee in communication with a smart environment are presented, none of these works explores which metaphors naïve users conceive in such a situation.

2.4 CROSS-CULTURAL APPLICABILITY

Most of the HHI-behaviours presented in this chapter are drawn from observations of interactions between people from central European or North American countries. Their applicability to interactions from different countries or cultures is not necessarily given. Gaze behaviours and formations of conversational groups can vary strongly. Two good examples for such cultural variances are compiled by Rossano et al. [RBL09]: In question-answer interactions between native speakers in Tzeltal—a Mayan language spoken in Tenejapa, a region in Mexico—the participants sit side by side and almost never exchange gazes. On the contrary, in similar interactions in Yèlî Dnye—spoken on Rossel Island, in eastern Papua New Guinea—mutual gaze can be sustained even during silence and speaker changes [RBL09]. This difference can be observed in the way the participants of an interaction form conversational groups. While people from Tenejapa prefer sitting side-by-side, people from Rossel Island prefer sitting face-to-face. Furthermore, this entails different ways of showing recipiency. As Tzeltal speakers do not see the others face, the addressee regularly produces phrasal back-channel acts, repeating whole parts of the previously said. Conversations in Yèlî Dnye use an inventory of visual, facially performed feedback [RBL09]. According to Hayashi [Hay88], the rules for turn management can be different too. Japanese speakers often deliberately talk simultaneously for a duration of multiple sentences. They do this in a coordinated and rhythmically synchronized manner. This behaviour is recognized as supportive and emphasizing the harmony of the interaction. American speakers on the contrary, use simultaneous talk when competing for the conversational floor. Such conflicting conventions can result in misunderstandings, repair, and eventual adaptation of behaviour between people. The turn taking system as presented in Section 2.1.3.4 on page 18, results from culture specific assumptions and must be adapted when assumptions change. Nevertheless, it stays a coherent system with simple rules. As Meyer writes: '[T]he resulting conversational organization is by no means "chaotic" or "anarchic". To the contrary, it is not less well ordered and comprehensible than the classical pattern.' [Mey18, p. 304]. An artificial agent, designed with a specific cultural background in mind and faced with unpredicted conventions would make similar mistakes as people do in such a situation. In the best case scenario the agent would need to recognize such mistakes and adapt its behaviour—as people do. Although conversation depends on the participants cultural background, similarities and patterns can be found between the presented literature on HHI and HAI. In this work I focus on the observability of addressing behaviour in smart environments, the formation of conversational groups in HAI, and the conversational roles an artificial agent can assume in such a situation. I aim to create models that can work or be adapted for interactions with people from different backgrounds. Nevertheless, the models presented in this work are biased towards the characteristics of interaction between central European adults, as they are the main group of subjects I have access to.

2.5 SUMMARY

The aim of this chapter is to create an idea of how people interact with others and what their expectations towards an interaction with an artificial agent, smart environment or device may be. I first give an overview of how humans interact with other humans in Section 2.1. To this end, I introduce what copresence is and the notion of proxemics (Section 2.1.1) as a generally accepted partitioning of space around a person with corresponding types of social interaction. In Section 2.1.2, I collect observations on what copresence means beyond proxemics and which kinds of interaction take place in unfocused interactions. I give an overview of focused interactions in Section 2.1.3. After a definition of face engagements (Section 2.1.3.1), I present an inspection of human interaction in conversations. This comprises how people form and maintain conversational groups (Section 2.1.3.2), which roles they assume within and regarding such groups (Section 2.1.3.3) and the rules by which they negotiate these roles (Section 2.1.3.4). Finally, I discuss gaze as a prominent conversational cue in Section 2.1.3.5. In the second part of this chapter, I present research on human interaction with artificial agents (Section 2.2). To this end, I first establish a taxonomy of the agents that are relevant for this thesis (Figure 2.3). I show that effects, known from unfocused HHI, can be reproduced or similarly observed in human interaction with artificial agents (Section 2.2.1). In the context of focused interaction, I show the importance of the behaviour of artificial agents on the overall and perceived quality of HAI (Section 2.2.2.1). Subsequently, I present how addressee recognition is performed (Section 2.2.2.2) and how turn taking behaviour generation can be modelled (Section 2.2.2.3). After summarizing how human conversational groups can be automatically detected (Section 2.2.2.4), I show how conversational groups are harnessed in recent research on human interaction with artificial agents (Section 2.2.2.5). In the third section of this chapter, I deal with human interaction in present smart environments (Section 2.3). I show how proxemics can be used for more situated and automatic adaptation of device functionalities and interfaces (Section 2.3.1). In Section 2.3.2, I present work on focused human interaction with smart environments (Section 2.3.2) and how different modalities affect the problem of determining the addressee of communication. I conclude this chapter by expanding on the cross-cultural applicability of the presented works, observations, and models and establishing the bounds of generalizability of this thesis (Section 2.4).

Part II

ADDRESSEE IN COMMUNICATIVE ACTS

In this part, I investigate human addressing behaviour in interactions with smart environments and robots. To this end, I evaluate the importance of different features for addressee recognition in interactions of naïve people with a robot inhabited smart home. Furthermore, I present and evaluate a visual approach for speaker detection and addressee recognition for a robot in interaction with a group of people.

3 ADDRESSING BEHAVIOUR IN SMART ENVIRONMENTS

> We tend not to recall the spacial organization of the event, how we decided when it was our turn to speak, how we organized ourselves when we did so and how the others showed that they did, or did not understand what we said.
> [Ken90, p. 1]

In this chapter, I investigate RQ 1: 'Which behaviours in naïve human interaction with a smart environment can be observed to distinguish which agent is addressed with a deliberate communicational act?'. To this end, I describe a study of naïve user interactions in a smart home that was carried out jointly by the contributors of the CSRA project [Ber+16]. The resulting corpus is published by Holthaus et al. [Hol+16]. On the basis of this corpus, I investigate how the available information and different modalities correlate with the addressee for different mundane tasks. I use the collected insights to create and evaluate an initial addressee recognition model for multi-modal single user interactions in a smart environment. Finally, I discuss the relevance of the obtained insights for the research question of this chapter.

3.1 INTRODUCTION

As discussed in Section 2.1.3, a conversation is a dynamic yet highly organized process between the participants of a focused interaction. People use multiple modalities, conversational cues and information from the overall situational context to distinguish the addressee of a conversational act. As people communicate with each other on a regular basis, they have an elaborate understanding of which information needs to be acquired, processed or inferred to reliably know the addressee of speech. Speakers know which information they must provide for others to understand who their addressee is. If visitors want to change the illumination settings of a smart environment, they are confronted with a problem. They do not know how to control the environment. They can only draw on their experiences and choose the control metaphors that they think are the most appropriate. For people to be able to solve such a problem, the control metaphor needs to be reasonably inferable. A smart environment therefore, needs to provide control metaphors that are obvious and memorable. A smart environment or artificial agent needs to be able to interact efficiently with people who did not undergo a special training to be able to understand and control it. By recognizing human conversational cues, they can better understand the expectations of human interaction partners. If a smart environment can correctly recognize the content and addressee of naïve inhabitants communication, it can better fulfil their expectations. It can work as expected.

In the literature about human interaction with smart environments (summarized in Section 2.3) different ways are proposed, how a person can select the agent to be addressed in a smart environment. While the presented research evaluate the selection accuracy or task completion rate, none investigates whether the chosen approach is one that a naïve user spontaneously would have chosen. How naïve people convey the addressee of their commands is not investigated. Furthermore, as suggested in the introductory quote of this chapter, people find it hard to explain which cues they use to understand and control an interaction.

Therefore, it is important to investigate which behaviours can be observed in naïve human interaction with a smart environment to distinguish which agent is addressed with a deliberate communicational act. We performed an initial attempt to such an analysis, with manually extracted observations and a simple model [RK16]. In this chapter I present a fully reproducible way of extracting observations of interactions from a corpus, and a detailed analysis of the resulting data and derived models.

3.2 INTERACTION CORPUS

addressing study

We collected a corpus of multi-modal interactions of naïve users with a robot-inhabited smart flat [Hol+16]. It was compiled in conjunction with a user study in the CSRA [Ber+16]. In the following, I call this the *addressing study*. The aim of the addressing studys was to observe how naïve users would solve everyday tasks in a smart environment and whom they would address for that. This corpus constitutes a unique basis for investigating how people convey their addressee in such a situation. Therefore, in this section I present the user study and original corpus, and how I extract the data needed for the following investigation. For further information regarding the original corpus and study, please refer to the corresponding publications [Hol+16; Ber+16].

3.2.1 *Experimental Set-up*

The study presented by us [Ber+16] was created to investigate how naïve people would address a smart robotic apartment when solving everyday tasks. In particular, the question was which entity (robot, light, apartment) would be addressed and which modality (speech, gesture, touch) used thereby. It was conducted in the CSRA in a collaborative effort by the contributors of the corresponding project. The layout of the apartment during the study can be seen in Figure 3.1. The 47 study-

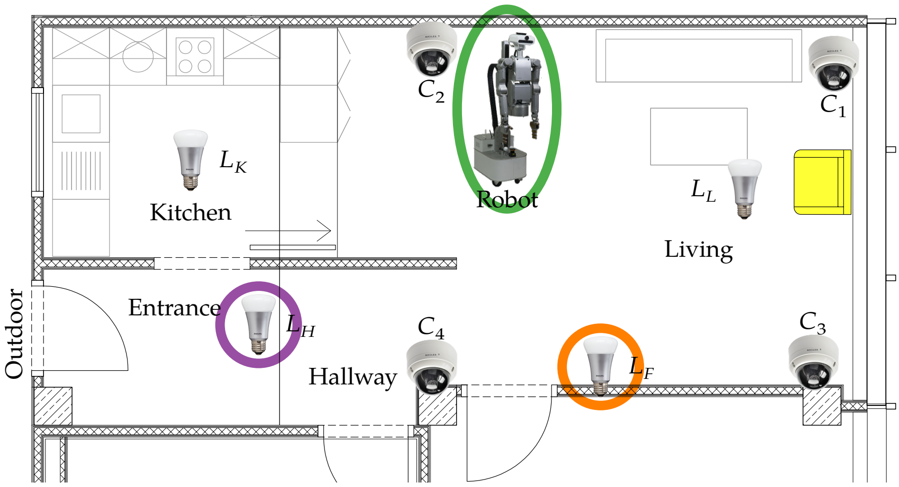


Figure 3.1: The layout of the CSRAduring the study. The robot (green) stayed at its position in the living room. The safety person sat in the armchair (yellow). The lamp from the first two tasks is L_H (purple) and the lamp from the seventh task is L_F (orange).

participants (25 women, 22 men, $18 \leq age \leq 50$, $\mu_{age} = 25.26$, $\sigma_{age} = 5.69$) were recruited from the campus of Bielefeld University, gave consent to the recording of video and audio material and received 6 € compensation for their attendance.

3.2.1.1 *Study Procedure*

An experimenter brought the participants into the apartment, introduced them to the task, and left the apartment for the duration of the task. After completion, the same experimenter escorted the participant to the post-trial procedure, which encompassed a questionnaire, the possibility to freely ask questions, and the monetary compensation. A second experimenter (WoZ) observed the participants' behaviour during the trial from an adjoining room and executed reactions of the apartment or the robot as required (Section 3.2.1.4). Due to safety reasons a third experimenter needed to stay in the apartment during the trials. This person monitored the robot, was introduced as such, and did not further interfere with the experiment.

3.2.1.2 *Briefing*

The experimenter escorted each participant through the entrance, hallway, and kitchen into the living room (Figure 3.2), while introducing the apartment (a map can be seen in Figure 3.1). The participants were told they are in a smart flat, which rooms are the hallway, kitchen and living room, showed the robot and explained the role of the security person. Whenever a room was mentioned by the experimenter, the ceiling light of this room was turned on by the wizard for a short time.

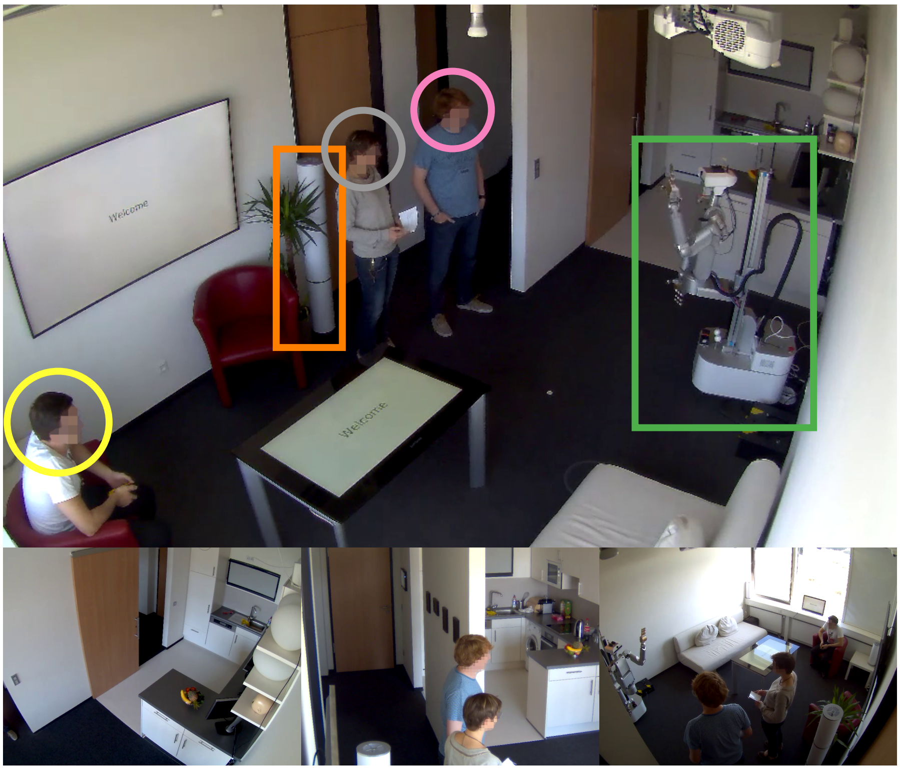


Figure 3.2: A scene in the living room, during the introduction of a participant from the perspective of camera C_1 (top) and the cameras C_2, C_3 and C_4 (bottom left to right). The camera positions can be found in the apartment map (Figure 3.1). From left to right in C_1 can be seen: the security person (yellow) with the emergency shut-down, the lamp from the seventh task L_F (orange), the experimenter (grey) introducing the robot, the participant (pink) of the trial, the robot (green) waving at them. The screens display the text 'Welcome'.

When the robot was introduced, it raised its left arm and waved. Seven mundane tasks, written on a set of cards, were handed to the participant. The participant was told to solve the tasks intuitively and in the order given by the cards. It was forbidden to contact the security person, or to use light switches and remote controls. Then the experimenter escorted the participant back into the hallway, left the apartment and waited outside to be approachable in case of a problem.

3.2.1.3 *Participant's Tasks*

The tasks of the addressing studys are specifically designed to meet a set of requirements. (1) They had to be reasonably simple and (2) required in a home environment on a regular basis. At the same time (3) they needed to be diverse enough to allow participants to consider different approaches and addressees for their solution. To further increase the variability of the results, some solutions were discouraged. To this end, light switches were non-functional, and there were no visible clocks and radios. Furthermore, the participants were prohibited to use their own clocks or phones. After the introduction, the participants started in the hallway (Figure 3.1), holding their task-cards in their hands (a list of the tasks and their order can be found in Table 3.1).

Id	A	V	Task
1	1		Turn on the light in the hallway, then go to the kitchen.
2	2		Turn off the light in the hallway.
3	3		Listen to music.
4	6	*	Find out if a parcel was delivered.
5	4	*	Find out if there was a phone call.
6	5	*	Find out the current time.
7	7		Alter the brightness of the floor lamp in the living room without talking.

Table 3.1: The tasks that participants needed to solve in the study. Id: shows the original order of the tasks. A: shows the alternative order of the tasks that was used for randomization. V: marks the tasks that elicited a verbal response of the apartment or robot in the verbal condition with a '*'.

The first two tasks, switching light in the current and adjoining room, allow insight on how people address daily appliances from different distances. The third task, listening to music, requires the participants to control an entity that has no visible embodiment. Tasks (4-6) require the retrieval of information, which is expected to encourage verbal interaction. The last task (7) allows for a continuous control of the result through a closed-loop interaction while enforcing a non-verbal solution. Within the information retrieval tasks, the order of was altered (4, 5, 6 vs. 5, 6, 4). The order of the remaining tasks fixed to prevent the information retrieval tasks from biasing the solution of the first three tasks towards verbal interaction.

3.2.1.4 *Task solution*

The wizard audio-visually observed the participants' actions and activated the corresponding actions. A task solving action was triggered, when the wizard observed an action of the participant and recognized it as dedicated to the solution of a task. Additionally, the wizard chose whether the robot or apartment should react to the action. The reaction was split into a verbal and a non-verbal condition for the information retrieval tasks (4-6). In the verbal condition, the robot or apartment verbally answered the questions. In the non-verbal condition, the apartment printed information on the screens (seen Figure 3.2) and the robot used deictic gestures.

3.2.2 *Recording & Annotation*

The trials were observed by the wizard through four parallel video streams (Figure 3.2) and an audio stream, which all were recorded for later analyses. Additionally, system events were recorded using the communication middleware [RSB]. System recordings contain all control instructions from the wizard, determining when the wizard decided that a task was solved, which addressee (robot or apartment) they selected and which functionality was executed subsequently. Furthermore, they contain two additional audio streams (hallway and living room), motion sensor observations, and power consumption data. Finally, recordings inform about when doors, cupboards, drawers and windows were opened and closed.

For the manual annotation of the corpus, the *ELAN Linguistic Annotator* by Birgit Hellwig [ELAN] was used. To this end, we created overview videos by combining the four camera perspectives (Figure 3.2) and the audio stream into a single video file. Furthermore, we created annotation templates in coordination with the final annotators and pre-filled them using a subset of the recorded system events. An overview of the kinds of generated and manually annotated tiers of interest for this chapter can be found in Table 3.2. A detailed description of the recording and annotation process is presented by Holthaus et al. [Hol+16].

Tier	Type	Annotated
Addressee final	C	*
Focus of attention	C	*
Expression (facial, gestural, verbal)	C	*
Expression specific	F	*
Method	C	*
Method specific	F	*
Speech form of address	C	*
Speech politeness	C	*
Speech type of sentence	C	*
Speech specific	F	*
Speech intention	C	*
Study progress coarse	C	*
Study progress fine	C	*
Wizard	C	

Table 3.2: A selection of the tiers, available in the annotations [ELAN]. Type depicts the kind of annotation: categorical (C), or free-text (F); Source depicts whether the tiers were manually annotated (*) or extracted from system events. A detailed table with all tiers can be seen in Table A.1 on page 139.

3.3 ANALYSIS OF ADDRESSING BEHAVIOUR

Our original aim with the study was to investigate how naïve users would intuitively interact with a robot inhabited smart home to solve simple daily tasks [Ber+16]. To this end, interactions were recorded, annotated and analysed. The study showed that the participants preferred to solve the given tasks using speech when allowed. The method used to solve the tasks $1-6$ was speech in more than 50 % of the cases in each task separately. In case of information requests and control of the radio—which lacks an embodiment—the proportion of verbal solutions was much higher (88 %). When an appliance had to be controlled, and it had a distinct physical location and extent, people more often addressed it directly than through some other entity (56 % in tasks 1 and 2 and 89 % in task 7). On the other hand, the robot was addressed around 30 % of the time in the information request tasks, while only 10.6 % in the lighting tasks and 12.8 % in the radio task. Finally, the addressee was *unspecific* in a high proportion of the task solutions (37 % in tasks 4 and 6 and 52 % in tasks 3 and 5). In the original publication [Ber+16] we analyse which entities of a smart home people address to solve different kinds of daily tasks and which modalities they use thereby. In the following, I use the resulting corpus to investigate *how* people address these entities.

3.3.1 *Observations of Addressing Behaviour*

The corpus annotations are a good starting point for the investigation of how people display the addressee of their deliberate communication. To this end, I further examine the participants' behaviour at the moment of addressing. This moment can be determined from the tiers *Wizard* and *Study progress fine* (Table 3.2). *Study progress fine* shows when the participants attempted to communicate with their environment. Such time periods are tagged as *Attempt to a solution*. The annotation tier *Wizard* is automatically generated from the reactions of the Wizard during the trial. It depicts the point in time, where the *Wizard* activated the solution of a task. Furthermore, it tells which the entity—robot or apartment—was responsible for its realization. The action implies that the wizard has the necessary information to understand that a task solution is attempted and which addressee is more appropriate. Furthermore, it means that the wizard perceived the action of the participant as in so far complete, that a reaction is advisable. This makes the time of the Wizard's action the best moment for the smart environment to react to the communication of an inhabitant from an observers point of view. To extract observations of interaction in a fully automatic and reproducible way, I apply the following approach: For each *Wizard* action the corresponding *Attempt to a solution* annotation is searched and the associated manual annotations extracted. To correspond with a solution attempt, a wizard action needs to overlap with it in time or happen not later than 2 s after its end. The maximum of 2 s was chosen on the basis of a sighting of the recordings, which revealed that annotations with a higher difference occurred due to misunderstandings. Repeated wizard actions, that correspond to the same solution are ignored for the same reason. A visualization of the matching process can be seen in Figure 3.3.

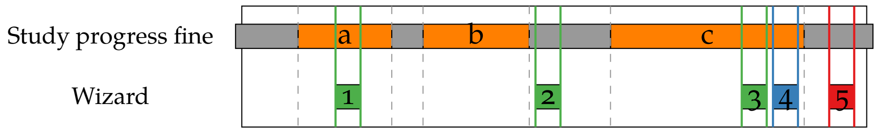


Figure 3.3: Matching between actions in the *Wizard* tier and instances of *Attempt to solution* (highlighted in orange) in the *Study progress fine* tier. Green time periods highlight wizard actions that can be assigned to a task solution (1 matches a, 2 matches b, and 3 matches c). Blue wizard actions (4) are not assigned because the task solution is already observed (3). Red wizard actions (5) are not assigned because they are more than 2 s after a solution attempt. Vertical lines show the boundaries of time periods for better comparability.

The resulting annotations are further filtered: Entries where *Addressee final* is *not discernible* (the annotator could not see the person) are unusable for this investigation and therefore removed. Furthermore, in case of redundant entries, only one attempt is kept. Redundant entries arise from trials that were annotated by multiple raters or from repeated solutions to the same task by the same participant. This results in a set of 307 annotations of peoples addressing attempts.

In contrast to this approach, the initial evaluation [RK16] was performed based on the information at the moment of the wizards actions for all actions of the wizard. Missing annotations of addressee and attention were subsequently manually annotated. This resulted in a different set of observations. I use the approach presented in this chapter to achieve reproducible and fully automatic results.

To assess the quality of the annotations, I calculate the inter-rater agreement for the extracted observations. Seven trials contain annotations of both raters and therefore, can be used for this analysis. As inter-rater agreement, I calculate Cohen's Kappa [Coh60] using the categorical annotations. Free form annotations were not sufficiently formalized beforehand and therefore cannot be compared objectively. Furthermore, the tiers *Study progress coarse*, *Study progress fine*, and *Speech intention* are constant after the data extraction. They are excluded from this calculation. The annotators achieve a good [Alt91] Kappa of $\kappa = 0.78$.

3.3.1.1 *Content of Observations*

In the following, I describe which information is extracted from the addressing observations in the original data to form the addressing corpus. *Annotations* and *tiers* in the original corpus will be called *values* and *variables* in the extracted corpus to clarify the distinction.

First of all, information from the tiers *Study progress coarse*, and *Study progress fine* can not be used in the observations because their values are always the same during task solution attempts. Similarly, *Speech intention* is always *Communication attempt* or not applicable in the observation. It is subsumed by *Method*. *Expression specific* is empty in most cases and therefore ignored too. The following variables (abbreviations in square brackets) are derived from the manually annotated tiers:

ADDRESSEE FINAL REDUCED [AR]: Is created from *Addressee final* by combining parts of the apartment into a single group *Parts of the apartment*. This encompasses furniture, switches, and screens but not the task relevant lights. Furthermore the entities *self* (addressed four times) and *not discernible* (never addressed) are included in *Unspecific*. The mapping can be seen in Table A.2. This is a reasonable pooling that allows quantitative analyses by greatly reducing the amount of possible addressees. The resulting variable can assume five different values: *Unspecific* [*U*], *Parts of the Apartment* [*Ap*], *Robot* [*R*], *Light in the hallway* [*LH*], or *Floor lamp* [*LF*].

FOCUS OF ATTENTION REDUCED [FR]: Is created from *Focus of attention*. Targets are grouped with the same mapping as in *Addressee final reduced* (*Ar*). The entity *self* is focused eight times and *not discernible* is focused once.

ADDRESSEE EQUALS FOCUS [AEF]: Is created by checking whether the annotations in *Addressee final* and *Focus of attention* have the same value.

EXPRESSION REDUCED [ER]: Is created from *Expression* (*facial*, *gestural*, *verbal*) by clustering emotions into *negative*, *neutral*, and *positive*.

METHOD [M]: Encodes the used modality—*speech*, *gesture* or *touch*.

METHOD SPECIFIC REDUCED [MSR]: Is created by extracting the usage of gestures (*clap*, *wave*, *wipe*, and *point*) from the textual descriptions in *Method specific*.

SPEECH FORM OF ADDRESS [SF]: Tells whether the entity is named or not, when speech is used.

SPEECH POLITENESS [SP]: Tells whether the phrasing is polite or not, when speech is used.

SPEECH TYPE OF SENTENCE REDUCED [STR] Is extracted from the tier *Speech type of sentence*. It can take the values *Command*, *Question*, or *Statement*.

SPEECH PHRASING [SPH]: Is extracted from the tier *Speech type of sentence*. It tells whether a full *Sentence* is said or single *Words*.

SPEECH SPECIFIC REDUCED [SSR]: Is drawn from the speech of the participants, encoded in *Speech specific*, by detecting the first appearance of addressing terms. It can take the values *you*, *light*, *robot*, and *none*.

Furthermore, the following variables are extracted from annotations in the *Wizard* tier and meta information about the trial:

WIZARD ADDRESSEE [AW]: Encodes which entity is chosen by the wizard to react to a communication attempt. It can take the values *Apartment*, *Floor lamp*, or *Robot*.

WIZARD TASK [T]: Tells which task is solved in a specific observation.

CONDITION [C]: Encodes whether the participant is in the verbal or non-verbal condition.

ORDER [O]: tells whether the tasks were be solved in normal or alternative order.

PARTICIPANT ID [PID]: Numerically identifies the participant.

addressing corpus

The resulting 16 dimensional set of observations of human interactions with a smart environment is the *addressing corpus* that is used in the following analyses.

3.3.2 *Predictability of Addressee*

In this chapter, I want to find out how naïve people narrow down the addressee of their communicative actions. With the newly generated corpus of addressing-behaviour observations (addressing corpus), I can perform this investigation. The addressee in each observation in this corpus is encoded in the *Addressee final reduced*-variable. This is the dependent variable that needs to be predicted. Other variables encode observable behaviour of the participants or information that is not part of the displayed behaviour. Both types of information can correlate with the choice of addressed entities. In the following sections, I investigate the connections between the variables of the addressing corpus with a special focus on *Addressee final reduced*.

3.3.2.1 *Correlations between Variables*

For a better understanding of the addressing corpus, I test the variables for statistical independence. To this end—for each combination of two variables—a contingency table is created. Applying the null hypothesis that the rows and columns of the table are independent, the variables are tested for independence and a level of significance is calculated. The significance tests are performed using Pearson's chi-square test with Monte Carlo simulation.[1] The Monte Carlo simulation is performed to account for small expected cell counts for some variable combinations. The resulting p-values are binned into intervals and visualized in Figure 3.4. To better understand the impact of correlations, I additionally examine the effect sizes (association) between the variables. To this end Cramér's $\tilde{V}$[2] is visualized in Figure 3.5. The following observations can be made from correlations and effect sizes between the variables:

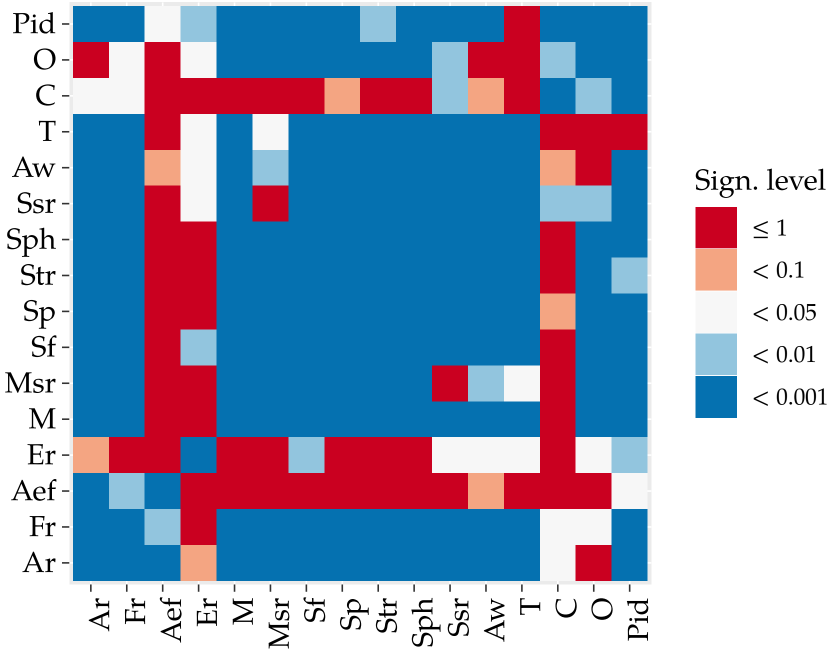


Figure 3.4: The p-values obtained from Pearson's chi-square test with Monte Carlo simulation and 1e+05 replicates, binned into intervals. Small p-values (blue and white colours) for a combination of variables suggest that there is a correlation between them. The variables are sorted and abbreviated as presented in Section 3.3.1.1.

ADDRESSEE FINAL REDUCED [AR]: The dependent variable shows correlations with most other variables in the corpus. The only exceptions are *Order* [*O*] and *Expression reduced* [*Er*]. The strongest effect size can be found in combination with *Wizard addressee* [*Aw*]

[1] using chisq.test from the stats package (v3.5.1) in R [stats] with 1e+05 replicates.
[2] Cramér's V with bias correction (Cramér's $\tilde{V}$) is used to prevent overestimation.

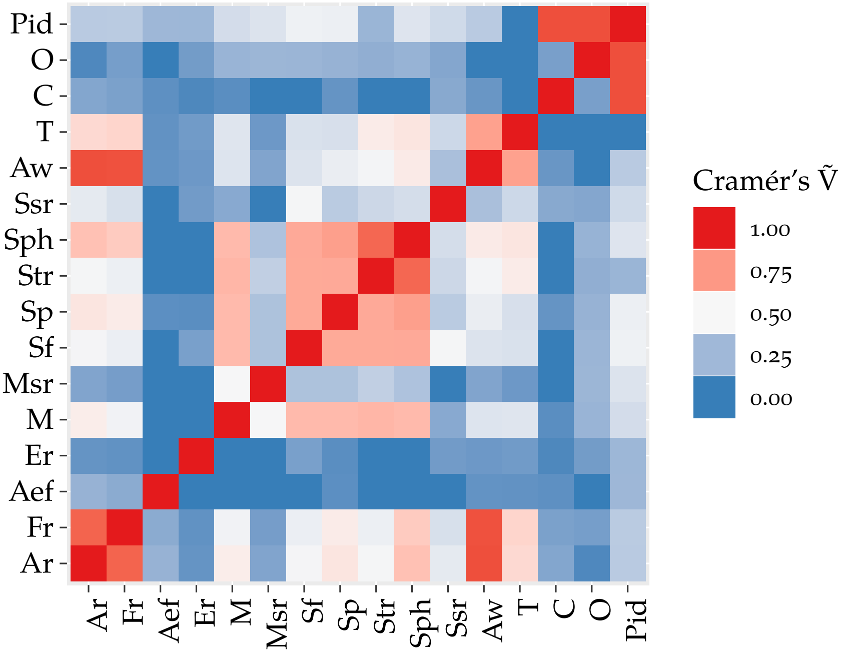


Figure 3.5: The Cramér's $\tilde{V}$ for each combination of variables in the addressing corpus. High values (red) represent high association (strong effect size), low values (blue) represent low association (weak effect size). The matrix is symmetric and has a value of 1 on the diagonal as $\tilde{V}(A, B) = \tilde{V}(B, A)$ and $\tilde{V}(A, A) = 1$. The variables are sorted and abbreviated as presented in Section 3.3.1.1.

and *Focus of attention reduced* [*Fr*]. Furthermore, *Wizard task* [*T*] shows a considerable effect size. This confirms that the wizard and annotators recognize the participants' addressee in a similar way and that the task at hand has an influence on the addressee—as expected in the study design. The *Participant-Id* [*Pid*] is a strong cue for the recognition of the addressed entity too, showing that participants had different preferences. Furthermore, the strong correlations and effect sizes with *Method* [*M*] and the speech related variables [*Sf, Sp, Str, Sph, Ssr*] show that these can be used as predictors too.

FOCUS OF ATTENTION REDUCED [FR]: This shows correlations and effect sizes that are similar to *Addressee final reduced* [*Ar*]. This observation supports the expectation that they are strongly correlated.

ADDRESSEE EQUALS FOCUS [AEF]: Knowing this value together with the participants' focus of attention is informative for addressee inference. However, it depends on the addressee and therefore can not be directly observed. The correlations with *Focus of attention reduced* [*Fr*] and *Participant id* [*Pid*] indicate that whether an addressed entity is looked at strongly depends on the entities involved in an interaction. The independence between *C* and *Aef* is in disagreement with the results in the initial evaluation [RK16]. This has mainly two reasons. (1) The original analysis was done based on a reduced set of addressees (similar to *Ar*). (2) The automatic corpus creation process 3.3.1 additionally produces not exactly the same observations as the manually annotated observations [RK16].

METHOD & SPEECH [MSR, SF, SP, STR, SPH, SSR]: All the method and speech related variables show similar correlations. Nevertheless, differences in correlations and effect sizes are still present. Furthermore, a compound effect from the choice of the modality is possible. This suspicion is confirmed by the correlations between *Method specific reduced* [*Msr*]—which encodes gestures—and the speech specific variables. Nevertheless, as the effect sizes between these variables and the addressee vary, they still can provide information that is not encoded in the chosen modality. A short inspection by only considering the verbal part of the dataset shows that the speech based variables still strongly correlate with addressee after removing the influence of *Method* (not visualized). This means that the type of gesture, the chosen sentence, the politeness, and the form of addressing all inform about which entity is addressed.

PARTICIPANT ID [PID]: correlates with all variables except *Wizard task* [*T*]. The non-correlation with the task and the strong correlations with *Order* and *Condition* result from the study design. The other correlations suggest that the participant's preferences can have a strong influence on the interaction.

OTHER VARIABLES [AW, T, C, O]: The variables *Wizard addressee* [*Aw*], *Wizard task* [*T*], *Condition* [*C*], and *Order* [*O*] are inherent to the study and can not normally be used directly for addressee recognition. Nevertheless, they show some interesting correlations. The strong correlations of *Wizard Task* show that the task at hand is important for the way participants approach an interaction. The correlation of *Condition* [*C*] with the addressee shows that the participants adapt to the capabilities of their environment.

To sum up, this analysis shows that there are multiple cues pointing at the addressee of naïve users communication attempts in a smart environment. While the task at hand influences which entity is addressed, the participant's focus of attention and chosen modality are strong hints for its recognition. Furthermore, multiple other cues that can be considered to further narrow down the addressee.

3.3.2.2 *Addressee and Attention*

In the previous section, a high of correlation and the strongest effect size can be seen between the addressee and focus of attention. In this section, I further investigate the distributions of these variables. The frequencies of *Addressee final reduced* and *Focus of attention reduced* are visualized in Figure 3.6. It can be seen that the entities are not addressed equally often.

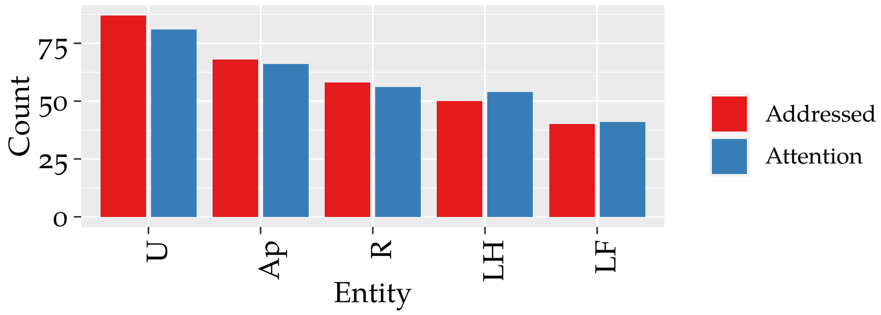
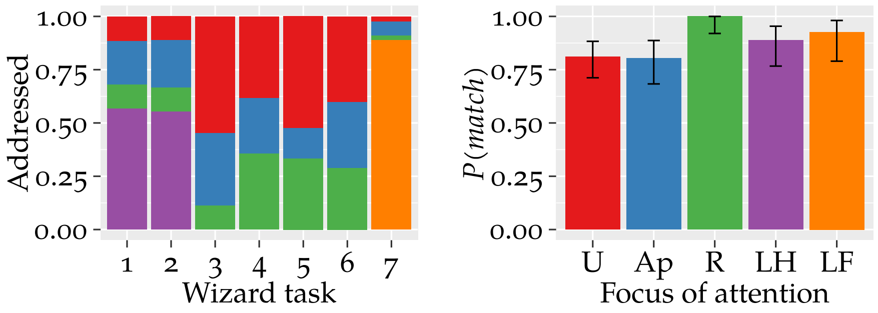


Figure 3.6: The reduced set of entities, as they can be observed in the addressing corpus, for the variables *Addressee final reduced* [*Addressed*] and *Focus of attention reduced* [*Attention*]. Addressees are distributed on the x-axis (*Unspecific* [*U*], *Parts of the apartment* [*Ap*], *Robot* [*R*], *Light in the hallway* [*LH*], and *Floor lamp* [*LF*]).

Although, *Parts of the apartment* [*AP*] combines different addressees, in more than 76.55% of the interactions, the addressed entity is *Unspecific* [*U*], *Robot* [*R*], *Light in the hallway* L_H [*LH*], or *Floor lamp* L_F [*LF*]. Additionally, in Figure 3.7a it can be seen that the distribution of addressees is different for the task sets 1–2, 3–6, and 7. We [Ber+16] suggest that there are multiple reasons for this distribution. (i) The study design requires the participants to control embodied entities in tasks 1, 2, and 7. In such cases people tend to directly address the entity that needs to be controlled, which results in the high proportion of addressed *Light in the hallway* and *Floor lamp*. Furthermore, (ii) when participants do not address the controlled device directly they address something that resembles a control interface (e.g. screens and switches) or an entity that may be able to control the device for them (like the robot). The same applies to cases where no embodiment for a functionality can be spotted as in task 3. (iii) If the participants need to retrieve information (as in tasks 4–6) they prefer addressing an entity that may be able to provide information. One option in such cases is the robot. However, especially in the non-verbal condition, where the answers are presented on the screens, people often interacted with the screens. Additionally, the addressee is often *Unspecific* [*U*] in the tasks 3–6. The participants addressed something, often verbally, but it is hard to tell what because they spoke *into the room* or *towards the ceiling*. This suggests that they addressed the apartment as a single entity or a non-embodied *agent*, which they expected to exist and be able to control the apartment.

In addition to the distribution of addressed entities, Figure 3.6 shows how often they were the participant's focus of attention during interactions. The similarity of the distributions confirms the observed correlation between the variables. Calculating the overall proportion of observations with matching addressee and attention (87.3%) further verifies this observation. This is in sync with the literature, as people look at their counterpart when they interact (Section 2.1). The equality between addressee and attention is represented by the variable *Addressee equals focus* [*Aef*] in the corpus. A visualization of the distributions of *Addressee equals focus* [*Aef*] for different *Focus of attention reduced* [*Fr*] with the corresponding confidence intervals can be found in Figure 3.7b. This visualization suggests that the addressee in the observed interactions is predominately equal to the focus of attention for all types of addressees. A difference can be found between interactions with *Robot*, and *Unspecific* or *Parts of the apartment*. The robot is always addressed when looked at in this corpus. The difference may be caused by the diverse embodiment of the anthropomorphic robot and the devices, switches, and screens that are combined in *Parts of the apartment*. These different embodiments result in differently strong social reactions. The lower equality of addressee and attention for *Unspecific* may additionally be caused by the inherent difficulty of recognizing *Unspecific* as addressee or focus of attention. Differences between the other entities are within the confidence interval.



(a) Addressees in tasks.

(b) Match in focus of attention.

Figure 3.7: (a) shows proportions of *Addressee final* for different *Wizard task*. The colour codes for the different entities can be found in 3.7b. (b) shows the probability of matching *Addressee final* and *Focus of attention* given *Focus of attention reduced* for the addressees *Unspecific* [*U*], *Parts of the apartment* [*Ap*], *Robot* [*R*], *Light in the hallway* [*LH*], and *Floor lamp* [*LF*]. The bars are augmented with 95% confidence intervals.

3.3.2.3 *Summary*

In this section I investigated the interdependences of the variables of the addressing corpus showing observations of human interactions with changing entities in a smart environment. I have shown that focus of attention not only correlates with addressee with a strong effect size but is predominantly equal to the addressee in many interactions. Knowledge about peoples attention is therefore informative of their addressee. Another information that can indicate the addressee is the modality (*Method* [*M*]) that the person uses to conduct this interaction. Naturally, the content of speech—if applied—informs about the interaction and therefore the addressee too. On the one hand, the addressee can directly be stated. On the other hand, properties like the type of the used sentence, the politeness, or the form of address can narrow the amount of probable addressees. When people interact using gestures, the type of gesture can be informative—to a small amount. However, it can not be assumed that the variables of the corpus independently contribute to addressee recognition. Naturally, the *Method* that is applied by a participant always limits the options for method-specific—speech and gesture related—variables. Furthermore, dependencies between variables can be caused by deeper relations rooted in interaction or other properties of the interaction that are not encoded in this corpus. How the variables can be utilized for addressee recognition is investigated in the following section.

3.4 ADDRESSEE MODELLING & RECOGNITION

In this section, I evaluate the recognizability of the addressee in interactions of the addressing corpus. To this end, I create models for addressee recognition and evaluate them using subsets of the corpus variables. The chosen subsets represent different capabilities of an automatic recognizer for the variables. Because of the high dimensionality of *Addressee final* and *Focus of attention*, their reduced versions *Addressee final reduced* (*Ar*) and *Focus of attention reduced* (*Fr*) are used in this section.

3.4.1 *Modelling Addressing Behaviour*

I use three different Bayesian Network structures to evaluate the recognizability and deepen the understanding of the interdependencies of the corpus variables. Bayesian Networks are especially suitable for this purpose. On the one hand, their structure can be used to impose or interpret the reasoning behind recognition results. On the other hand, they can cope with missing or uncertain input data and, therefore, be used in changing environments.

The most simple network is based on the observation that addressee and attention are strongly correlated. In this baseline model (*BF*), *Focus of attention reduced* [*Fr*] is conditionally dependent on *Addressee final reduced* [*Ar*] ($Ar \rightarrow Fr$). A graphical representation of the model can be seen in Figure 3.8. Because of the distribution of addressees and the high probability of equality between addressee and focus, it can be expected that this network always returns the participants' focus of attention as the most probable addressee. If *Focus of attention reduced* [*Fr*] can not be observed, the resulting addressee is the overall most probable addressee. As can be seen in Figure 3.6, this is *Unspecific* [*U*].

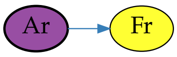


Figure 3.8: Simple Bayesian Network structure (*BF*) that uses only *Fr* to infer addressee. The nodes depict variables in the corpus, the arrows show dependency relationships. The colour of the arrow (blue) matches the network colour in further analyses. The output variable *Ar* is highlighted in purple.

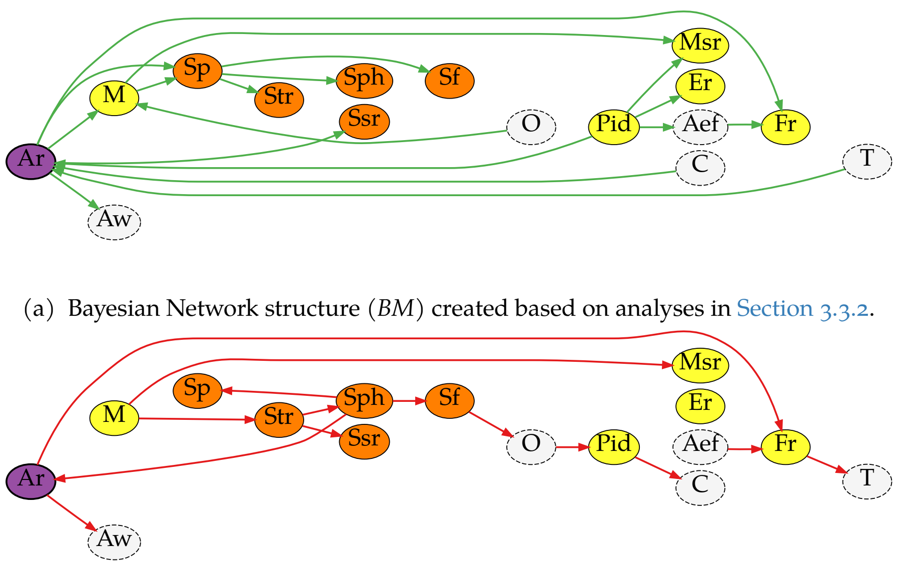
For the second and third Bayesian Network, I use all variables of the corpus. Based on expert knowledge about the study and the observations made during the analysis of correlations (Section 3.3.2), I manually craft the structure of the *BN* network. A detailed reasoning for the chosen structure is presented in Figure A.1. The resulting structure (shown in Figure 3.9a) is intuitive and in agreement with the observed correlations (see Section 3.3.2). Not all correlations are represented directly in this network as this would create circular dependencies within the graph and amplify the effect of correlations within the input variables—e.g. between *Method* and the speech variables. However, other plausible configurations can be created.

The third network structure—*BA*—is automatically extracted from the corpus data using the a *Hill Climbing* approach.[3] The structure of *BA* can be seen in Figure 3.9b. It is interesting to examine the automatically extracted structure and compare it to the manually created model and intuition. Both networks have a common core structure:

$$Aef \rightarrow Fr \leftarrow Ar \rightarrow Aw$$

Furthermore, they both show the connection between modality and gesture ($M \rightarrow Msr$). This suggests that these connections are characteristic for the scenario and data used in this chapter. The observed strong interdependence between the speech related variables can be found in the auto generated network too, although in a different structure. In *BA* the correlation between addressee and the speech related variables is represented by $Sph \rightarrow Ar$. The only speech related information considered for addressee recognition is, therefore, whether the participants

[3] Using `bnlearn::hc` from the `bnlearn` package (v4.4) in R [bnlearn] with 1000 restarts and 1000 perturbations.



(a) Bayesian Network structure (*BM*) created based on analyses in Section 3.3.2.

(b) Bayesian Network structure automatically extracted from corpus data (*BA*).

Figure 3.9: Manual (*BM*) and auto-generated (*BA*) Bayesian Network structures. The nodes depict variables in the corpus, the arrows show dependency relationships. The colours of the arrows match the network colour in further analyses (green for *BM*, red for *BA*). The node positions are fixed for better comparability between the networks. Node styles depict their type: The output variable *Ar* is purple, *Speech* related nodes orange, *Visual* nodes yellow, and non-observable nodes are gray with dashed outlines.

use full sentences, single words, or no verbal interaction at all. When all input variables are observed, the addressee is inferred from {*Aef, Fr, Aw, Sph*} only. *Method* only contributes when *Str* and *Sph* are unknown. Furthermore, the connection

$$Msr \leftarrow M \rightarrow Str \rightarrow Sph \rightarrow Ar$$

suggests that the only information in method that hints at the addressee is whether speech is used or not. Similarly,

$$C \leftarrow Pid \leftarrow O \leftarrow Sf \leftarrow Sph \rightarrow Ar$$

are chained in a way that an observation can provide information for addressee recognition only if all other variables between it and *Ar* are unknown. The participants' expressions (*Er*) are independent from other observations in the corpus according to *BA*.

For comparison, I additionally create a Random Forests[4] based classification model—*RF*.

[4] Using `randomForest` package (v4.6-14) [randomForest] in combination with the `mlr` package (v2.13) [mlr] in R.

3.4.2 *Evaluation Procedure*

For the analysis of recognizability of addressee *Addressee final reduced* [*Ar*] is used as the target variable. I compile four sets of input variables to represent different capabilities of the underlying system. The first set *Speech* represents data that can be deduced when auditory sensor data is available. It contains the speech specific variables *Sf*, *Sp*, *Str*, and *Ssr*. The second set *Visual* represents data that can be deduced visually. It contains the variables *Fr*, *M*, *Msr*, *Pid*, and *Er*. By combining *Speech* and *Visual*, the *Observable* set is created. It used data that can be observed visually or auditory. The final set *All* uses all information available from the corpus. These sets are used to evaluate the proposed models.

The evaluation is based on a leave-one-out Cross-Validation (CV). The parameter tuning for the *RF* model is performed in each iteration of the CV, within the training set. To this end, an additional sub-sampling with 100 iterations is performed in the training of the *RF* model to optimize the parameters `ntree` $\in [1, 2000]$, `mtry` $\in [1, numVar]$, `nodesize` $\in [1, 100]$, and `maxnodes` $\in [2, 100]$—with *numVar* being the amount of input variables. The predictions of *Addressee final reduced* for each iteration of the CV are compared to their ground truth annotations to estimate the performance of the models.

3.4.3 *Results & Discussion*

Using the predictions of the created models (*BF*, *BM*, *BA*, and *RF*) with the presented variable sets (conditions *Speech*, *Visual*, *Observable*, and *All*) in the CV, the model performances with corresponding confidence intervals can be calculated. For this evaluation a confidence interval of 95% is used. I use the following notation: BN_A refers to the *BN* model in the *All* condition. The other conditions are abbreviated with *S* (*Speech*), *V* (*Visual*), and *O* (*Observable*). The results can be seen in Table 3.3 and are visualized in Figure 3.10.

	Speech	Visual	Observable	All
BF	0.30 ± 0.10	0.89 ± 0.07	0.89 ± 0.07	0.89 ± 0.07
BM	0.55 ± 0.11	0.88 ± 0.07	0.87 ± 0.08	0.96 ± 0.04
BA	0.42 ± 0.11	0.89 ± 0.07	0.88 ± 0.07	0.94 ± 0.05
RF	0.37 ± 0.11	0.90 ± 0.07	0.89 ± 0.07	0.95 ± 0.05

Table 3.3: Accuracy with 95% confidence interval of the Bayesian Network based *BF* (focus only/baseline), *BM* (manual), *BA* (auto generated) models, and the Random Forests based classifier *RF*. The performance is measured using CV with the four sets of input variables (*Speech*, *Visual*, *Observable*, and *All*).

In the *Speech* condition, *BF* performs worse than the other Bayesian Networks and *BM* performs better than all other models. In the *All*

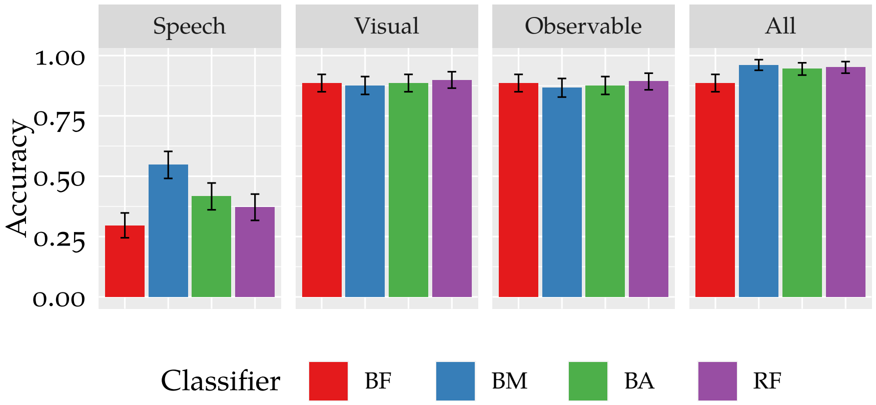


Figure 3.10: Accuracy of the Bayesian Network based *BF* (focus only/baseline), *BM* (manual), *BA* (auto generated) models, and the Random Forests based classifier *RF*. The performance is measured using CV with the four sets of input variables (*Speech*, *Visual*, *Observable*, and *All*). Confidence intervals (95%) are shown at each bar.

condition, *BM* and *RF* perform better than *BF*. When using the *Visual* or *Observable* sets of variables, no differences in the recognition results of different models can be found. When only *Speech* variables are observed, all models perform worse than in the other configurations. The *BM* and *RF* networks show better results in *All* than in the other conditions.

This evaluation reveals which properties of the investigated interaction are especially relevant for addressee recognition. The strong increase in recognition quality that is introduced with the variables in the *Visual* condition shows how informative vision is for such a task. There is no difference between the results in *Visual* and *Observable*—neither within the conditions nor in between. This entails, that when *Visual* information is known *Speech* does not provide enough information to strongly enhance the recognition model in this scenario. Furthermore, *BF* produces results that are as good as the other models in *Visual* and *Observable* and can compete with *BA* in the *All* condition. Therefore, it can be assumed that the results in these configurations are primarily based on the values of *Focus of attention reduced*. The interaction between the results of the Bayesian Networks in *All* and *Speech* reveals an interesting difference between the models. On the one hand, *BM* performs better than *BF* in both the *Speech* and *All* condition. On the other hand, *BM* is better than *BA* in the *Speech* condition but not in the *All* condition. This means that there is an effect of the *Speech* variables on *BM* which is not strong enough to outperform *BA* in the *All* condition but sufficient in the *Speech* condition. This difference indicates that *BM* can draw additional information compared to *BA* from the *Speech* variables. As these differences in the results can already be observed in the *Speech* condition, it is probable that the interaction between the addressee and the *Speech* variables produces this difference. However, it is hard to say how this enhancement is achieved without an in depth analysis of the structures of the networks, the resulting conditional independences given the sets of variables, and the influences of changes in the structure. The results of the Random Forests model do not strongly differ from the results of *BA*. This means that for the task of addressee recognition, both automatically tuned models are equally good. Nevertheless, the Bayesian Network approach can produce better results when created by a domain expert as in *BM*. The low overall performance of the addressee recognition models in the *Speech* condition confirms the importance of visual information for this task. Nevertheless, the Bayesian Network models are better than the baseline (*BF*) and can infer the addressee from speech information alone in 34%–60% of the observations.

3.5 summary

In this chapter I investigated RQ 1[C]. To this end, I presented a study and a corpus of unconstrained human interactions with entities in a smart home. The corpus was especially suitable because it contained interactions with all types of non-living entities including a robot and did not limit the way in which participants may approach the interaction. From this data, I extracted an addressing corpus with 307 observations of successful interactions, consisting of sixteen categorical variables. An in depth analysis of the mutual covariances between the variables of the corpus, showed that the participants' focus of attention is the most informative cue for addressee recognition. This is followed by the used modality and the content of the speech—when speech is used. Subsequently, an inspection of the values of addressee and focus of attention revealed that people predominately focus the addressed entity. This was observed for all types of addressees but especially true for interactions with the robot. Using the gained knowledge, I manually created a Bayesian Network structure for addressee recognition. I evaluated the model's recognition performance and compared it with a model based only on focus of attention and two data-driven recognition models. The evaluation was performed on four sets of input variables which represent different levels of capabilities of a smart environment. A manually created Bayesian Network structure performed equally well or better than the other models for all presented combinations of input variables. The performance of the automatically learnt models was always on par. While the content of speech informed about the addressee of an interaction, its effect was not strong enough to create differences when focus of attention was likewise observed. A recognition performance, that is better than a model that always returns the

[C] RQ 1: Which behaviours in naïve human interaction with a smart environment can be observed to distinguish which agent is addressed with a deliberate communicational act?

focus of attention as addressee, only was achieved with the manually created Bayesian Network structure or Random Forests based approach when using all available variables.

The results of this chapter provide some answers to the underlying research question (RQ 1). People that were not trained to interact with a specific smart environment construct interactions that emerge from their own background and previous experiences. In doing so, they exhibit a set of multi-modal cues that can be used to infer who is addressed. A human observer naturally interprets these cues and distinguishes the addressees. However, to provide an artificial system with such a capability an in-depth understanding of these cues and their interaction is needed. In this chapter I have shown that visual observation of inhabitants and especially their focus of attention is an important and strong cue for addressee recognition. Whether verbal, gestural or touch interaction with lamps, switches, robots, or the smart environment as an integrated, interactive entity—the attention often can be used to infer who is addressed. If the environment can not observe a person's attention—e.g. because of blind spots, occlusions or privacy concerns—it is still, to some degree, possible to recognize the addressee using only speech information. Furthermore, if the amount of fully annotated data for learning is small, as in the presented corpus, a recognition model that is manually tuned by an expert in human interaction outperforms automatically learnt models. This is a reasonable result as an expert can provide the model with background knowledge that cannot be extracted from scarce data.

4 ADDRESSING IN HUMAN-ROBOT CONVERSATIONAL GROUPS

> From *p*'s point of view then, *p* may be said to be 'offering' *q* the floor, for in looking steadily at him he indicates that he is now 'open' to his actions, whatever they may be.
> [Ken67, p. 36]

In this chapter I investigate RQ 2[C]. To this end, I present a scenario in which a robot participates in a conversational group with multiple people in the CSRA to solve tasks verbally directed at it. The collected observations are used to evaluate an approach to addressee recognition in multi-party Human-Robot-Interaction. Finally, I discuss the results of the evaluation and the implications for the research question.

4.1 INTRODUCTION

In dyadic HRI scenarios, it can by definition be assumed that only one person interacts with the robot. Therefore, robots can assume to always be the addressee of speech [Hol14; CSW14; HM16]. For instance, a poll questioning robot—presented by Bruce et al. [BNS02]—pays attention to a person from the moment its *area of interest* is entered and until the end of the interaction. All other persons are ignored during that time. However, the assumptions that conversations are always dyadic and that all utterances are produced to cause a verbal response from the conversational partner do not hold in most interaction scenarios. When several persons are present, they can not only speak to a robot. They dynamically create and change conversational groups and converse with each other (see: Section 2.1.3.1). Even when the robot knows that it is in a conversational group and with whom,[1] it still needs to actively participate in this group. At least, it needs to know when to react to an utterance and when not. As presented in Section 2.2.2.2, different approaches to automatic addressee recognition are possible. Most of the approaches used in HRI employ information about the participants VFoA and acoustic information to decide whether the robot is addressed by a speaker or not. To distinguish which person speaks at a particular time, techniques for sound source localization or close-talk microphones

[1] the problem of conversational group detection with artificial agents is investigated in Chapter 6

[C] RQ 2: How can an artificial agent visually recognize whether it was addressed by a person within its conversational group or not?

are often applied [Lan+03; SJB15]. In this chapter, I present a multi-party, Human-Robot-Interaction scenario, designed to confront the robot with the problems of participation in a conversational group and autonomous decision whether to react to an utterance within the group or not. We designed this scenario, implemented it on the Floka, and applied it in an experimental study in the CSRA [Ric+16]. While the study design and execution was collaboratively perfomed by the authors of the paper, I was responsible for the addressee recognition that was applied during the study and is evaluated in this chapter. This scenario is particularly suitable for the evaluation of addressee recognition approaches for artificial agents. A description of the study set-up, recording and annotation, used platform, implemented attention and dialogue management systems, as well as an initial evaluation of the used addressing recognition can be found in the corresponding publication [Ric+16]. On the basis of the collected data, I investigate RQ 2 by examining the following claims:

CLAIM 4.1 (SPEAKER DETECTION) *By visually observing movements of lips, an agent can recognize if a person has the role of speaker in a conversational group.*

CLAIM 4.2 (NEXT SPEAKER DETECTION) *Recognizing mutual gaze with a participant of the conversational group at the end of an utterance, can be interpreted by an agent as a prompt to take the next turn.*

CLAIM 4.3 (ADDRESSEE DETECTION) *Mouth movement and gaze information about a participant of a conversational group can be combined and put into context to recognize if a robot is addressed with an utterance or not.*

I investigate RQ 2 by incorporating these claims into an addressee recognition system and evaluating its applicability in a multi-party HRI scenario.

4.2 HUMAN-ROBOT ADDRESSING CORPUS

The evaluation of addressee recognition approaches in HRI poses multiple challenges to the design of the interaction. Addressee recognition is the distinction of utterances addressed towards the agent from utterances exchanged between other people in the situation. Therefore, a scenario needs to be chosen that encourages both: interactions with the agent and with other people. Furthermore, the interaction should encourage that both the speaker and the addressee change on a regular basis and both addressing and non-addressing of the robot can be regularly observed. In this section, I present a HRI scenario and implemented robot behaviour that satisfies these requirements. Subsequently, I describe the corresponding experiment and annotation procedure. The resulting corpus is presented at the end of this section.

4.2.1 *Multi-Party Interaction Scenario*

We designed a multi-party HRI scenario in the CSRA in which human participants ask a robot questions and make it control the environment. To this end, the participants are equipped with a set of notes, containing tasks for the robot. To enforce a better balance between addressing of the robot and other participants, a two-step communication of the tasks is performed. The procedure is as follows:

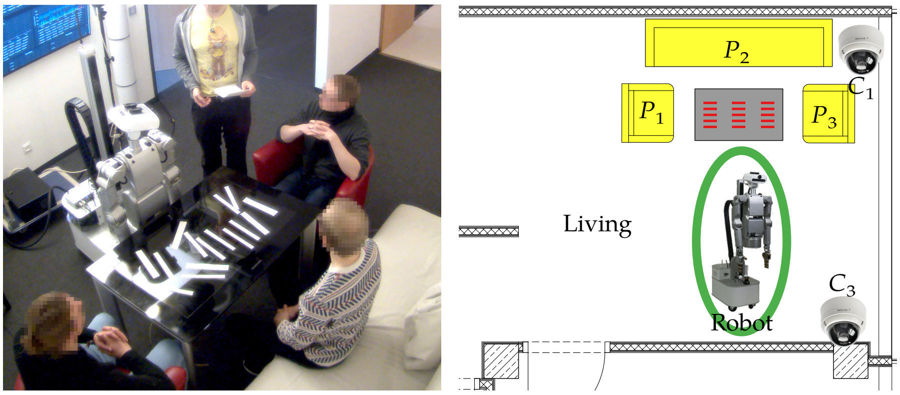
1. Participant P_a uncovers a note.
2. P_a communicates the task to a second participant P_b.
3. P_b takes the turn and communicates the task to the robot R.
4. The robot R accepts and solves the task.
5. P_b takes the role of P_a.[2]
6. This is repeated until all notes are uncovered.

The communication of the task from one participant to another not only enforces interaction between the participants. It additionally confronts the robot with utterances that verbally match its expectations but address someone else. To further support a uniform distribution of the conversational roles in the interaction, we recommend a closed, circular conversational group configuration. To this end, we chose interactions with three human participants and one robot, distributed around a table in the CSRA's living room (see Figure 4.1). This allows the participants to interact with each other and with the robot without the need to rearrange between turns. Therefore, there is no need for changes in the conversational group arrangement. When the interactants position themselves at the edges of the table, their resulting distribution in the p-space of the conversational group is uniform. Therefore, all participants have the same access to the conversational group and no particular distribution of roles is imposed on the interaction (see Section 2.1.3.2). With this approach we create a suitable scenario for the evaluation of robotic addressee recognition with (1) a clear motivation for the participants, (2) a fixed conversational group, and (3) a good ratio between addressing of the robot and other participants.

4.2.2 *System Set-Up*

Literature suggests that the behaviour of an agent influences the behaviour of participants (see Section 2.2.2.1). Therefore, it is important to consider which behaviours of the robot are desirable and how these may influence the course of the interaction. We used the anthropomorphic

[2] Participants naturally ensure a uniform distribution of participation in the group in such a tash. This can be expected based on the literature and is confirmed in this study.



(a) Scene during the briefing.

(b) Map of the study set-up.

Figure 4.1: (a) The left image shows a scene during the briefing of participants, recorded by the camera C_1. The participants and robot are already correctly placed for the study. The experimenter is standing next to the robot and describing the study procedure. (b) The right image shows a map of the apartment's living room during the study. The participants P_1, P_2, and P_3 are seated on the armchairs and sofa (yellow). They and the robot (green) surround a table (gray) on which the task notes (red) are spread.

robot Floka with its sensor head (as presented in Figure 1.3 on page 8). The human-like upper body of the robot allows a clear recognizability of its front, and therefore allows people to estimate its transactional segment. Furthermore, by turning its head—pan and tilt—the robot can show attention.

To allow Floka to actively participate in the interaction, we provided it with an attention management system. It integrates multi-modal sensor information from the robots camera and microphones to direct its visual attention to salient areas. This allows Floka to (1) focus on participants of the interaction by looking at their faces and (2) shift its attention towards a speaker by turning towards sound sources. The robot continuously exhibits this attentive behaviour to provoke the impression that it is following the interaction. Furthermore, this should allow the robot to focus on the current speaker of the conversational group. The implementation of the attention management system is presented in the corresponding publication [Ric+16]. The dialogue management system, presented by Carlmeyer et al. [CSW14], utilizes the results of the addressee recognition component to only processes utterances when the robot considers itself addressed. In this case the robot implicitly takes the turn, produces a verbal response to the recognized task, and finally yields its turn by ending speech production and continuing with its gazing behaviour. The addressee recognition component is presented in detail in the following section.

4.2.3 *Addressee Recognition*

To investigate the Claims 4.1 to 4.3 I implemented a system for addressee recognition that incorporates the detection of mouth movements and mutual gaze to distinguish utterances addressed towards the robot from other utterances. It combines two kinds of information (Claim 4.1) whether a person is speaking, and (Claim 4.2) whether this person maintains mutual gaze with the robot. To this end, I extended the gaze detector created by Schillingmann et al. [SN15] to make its gaze recognition results and the results of the facial landmark detection [Sag+13] available within the CSRA. The used features are visualized in Figure 4.2. The gaze detection provides the horizontal and vertical angle between the participants gaze direction and the image centre. Because the robot's camera is located at the centre of its head, the person is looking directly into the robots face when both angles are zero. Therefore, to decide whether a person p maintains mutual gaze with the robot, I compare the magnitude of both their horizontal (α_p) and vertical (β_p) gaze angles to a threshold θ.

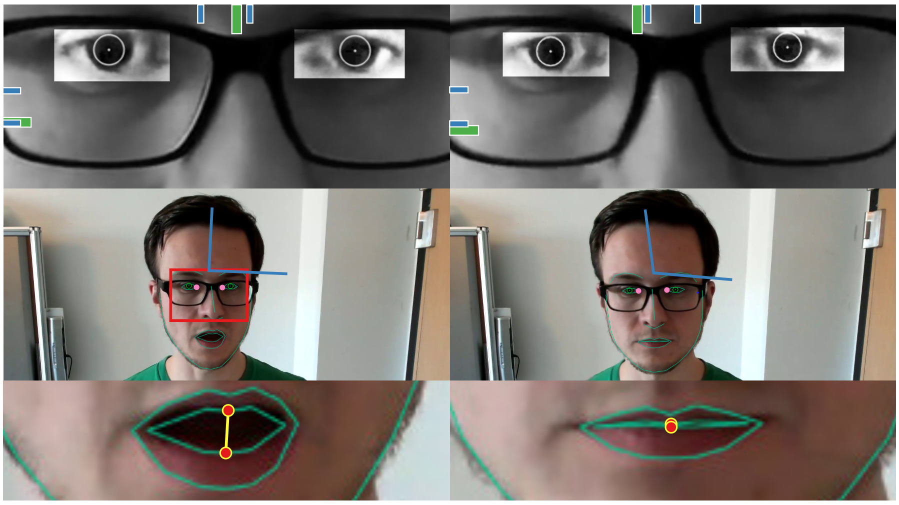


Figure 4.2: Visualization of the features used in the addressee recognition. The person on the left maintains mutual gaze and is speaking, the person on the right does neither of it. The upper images show the results of the gaze recognition system—horizontal and vertical angle—from Schillingmann et al. [SN15], augmented with the default thresholds for mutual gaze (blue bars) and the current estimation (green bars). The centred images show visualizations of the estimated head orientation (blue), and a red rectangle in case of mutual gaze. The inner corner-points of the eyes are highlighted in pink. The lower images show the facial landmarks (cyan lines) in the region of the mouth and highlight the central mouth landmarks (red dots) and corresponding distances (yellow lines).

$$G(p) = |\alpha_p| < \theta \wedge |\beta_p| < \theta \tag{4.1}$$

For the recognition of mouth movements, distances between the central points of the inner mouth (see Figure 4.2) are observed during a sliding time window Δ_t. This results in a set of distances $\mathbf{D_p}$ for each person p. A person is recognized as speaking if the variance of these distances exceeds a threshold d.

$$S(p) = \mathrm{Var}(\mathbf{D_p}) > d \tag{4.2}$$

To decide whether a person p is addressing the robot at a particular time, the results of both Equation (4.1) and Equation (4.2) are then combined.

$$A(p) = G(p) \land S(p) \tag{4.3}$$

The thresholds $\theta = 12°$, $d = 1.5$, and the time window $\Delta_t = 600\,\mathrm{ms}$ were selected based on a pre-study with the same set-up.

4.2.4 *Study Procedure*

After the participants gave their consent to the recording, they were accompanied into the apartment's living room by the experimenter and asked to take a seat in the armchairs and sofa. The robot was already waiting at the table. The following tasks were laid out on the table: (i) *Turn the light on.* (ii) *Turn the light off.* (iii) *What time is it?* (iv) *Has a parcel arrived?* (v) *Has anyone called?* (vi) *Which data is recorded?* (vii) *What exhibits are there?* (viii) *What's with the garden?* All tasks except *Turn the light off* existed twice. This added up to 15 tasks to allow each participant to solve five tasks. After the participants took a seat, the experimenter explained the task by giving an example of an iteration of the interaction (see Section 4.2.1). Additionally, the participants received two hints: (1) They need to acquire the robot's attention when they want to talk to it. (2) They do not need to repeat a task more than three times when the robot does not solve it. When the participants had no further questions, the experimenter left the room, activated the robot, and monitored the interaction from an adjoining room. When all tasks were solved, the experimenter entered the room for an informal debriefing. The typical duration of an interaction was around 10 min, resulting in around 15 min for the whole trial with briefing and debriefing.

The study was performed in German, with groups of three participants at the CSRA. The 15 study participants (2 female, 13 male) were all native German speakers from the CITEC. The participants gave consent to the recording of audio and video material and received confectionery as compensation for their attendance.

4.2.5 *Contents of the Corpus*

The recordings of the study sum up to 53 min of multi-party HRI. They contain overview video recordings from the camera perspectives of C_1 and C_4 (see Figure 4.1b), and Floka's perspective using its head camera. Audio data was recorded for the apartment's living room and hallway. Furthermore, system events of the apartment, the communication between the apartment and robot, and the internal, high-level states of the robot were recorded.

The robot's speech, mouth-movement and mutual-gaze recognition results were extracted from the recordings into *ELAN Linguistic Annotator* [ELAN] tiers [Ber+16]. This strongly simplified the ground-truth annotation of the robot's addressee recognition. In sum, the robot recognized 841 dialogue acts. A large proportion of these acts consists of confirmations and negations which were irrelevant for the task at hand and ignored by the robot. Therefore, the manual annotations and remaining analyses focus on the dialogue acts that had the potential of causing a response from the robot. These interactions were annotated with the annotation tool [ELAN] to manually distinguish whether (i) the robot was looked at by the focused person, (ii) the focused person was speaking, (iii) the robot was addressed, and (iv) the robot was looking at the wrong person (not the speaker) at the moment the speech act was recognized by the robot. The resulting corpus consists of 176 dialogue acts that were recognized by the robot as task relevant. The sum of utterances addressed at the robot can be expected to be at least the number of trials times the number of repetitions of the task. Additionally, recognized tasks can stem from communication between the participants, repetitions, and mis-classifications of tasks. Similarly, mis-classifications can result in tasks being recognized less often than expected. The overall distribution of these tasks, the expected number of addressing, and the proportion of tasks addressed to the robot can be seen in Figure 4.3. As apparent in Figure 4.3, the amount of recognized *light* and *time* tasks is much higher than expected while *exhibits* and *data* requests happened less frequent than expected. This is mainly because these utterances often were wrongfully understood as *light* and *time* requests.

The corpus provides observations of HRI and a ground truth annotation of whether a focused person was speaking, looking at the robot, and whether the robot was addressed at the time of a recognized dialogue act. In the following sections I use this data to assess the claims stated for this chapter.

4.3 VISUAL SPEAKER DETECTION

In Claim 4.1, I suggest that a person can be visually identified as having the role of speaker by observing mouth movements. The corresponding

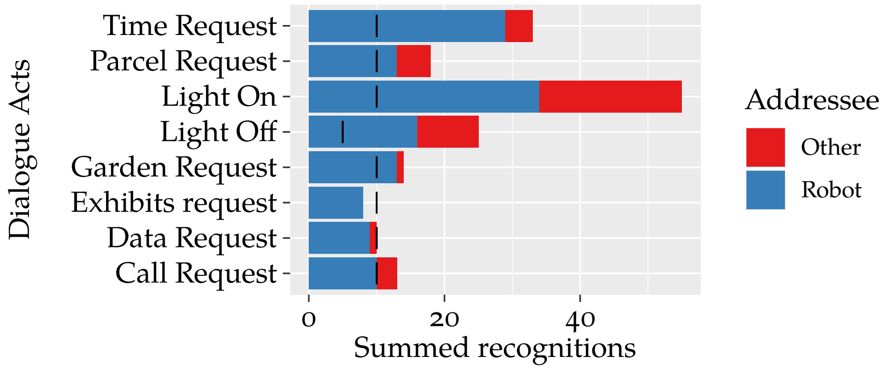


Figure 4.3: This plot shows how often each relevant dialogue act was recognized by the robot during the study over all trials. The utterances that were addressed at the robot are highlighted in blue, the other in red. A black, horizontal line shows the amount of tasks that were expected to be addressed at the robot during the trials.

model is defined in Equation (4.2), and applied during the user study. For the evaluation of visual speaker detection, I call this model the *study-model*. Additionally, I create an *accept-all-model* that always assumes that there is a speaking person in front of the robot. This is a reasonable approach because a proportion of 0.699 (prevalence) of the observations shows a speaking focused person. On the basis of the relevant utterances, recognized during the study, and the created ground truth annotations, I assess the models' performances. To this end, I calculate their precision, recall, and accuracy with 95% confidence intervals according to Clopper et al. [CP34][3] and F1-score as commonly used measurements for classifier performance. To account for the prevalence of the data, I additionally calculate the measurements markedness, informedness, and DOR. The results are visualized in Figure 4.4.

The precision and accuracy of the *accept-all-model* are both 0.699, and its *recall* is 1. This results directly from the prevalence of the data and the model design, which classifies all observations as speaking. The *study-model*, in contrast, achieves a higher precision of 0.87 and a lower recall of 0.85. Although the *study-model* achieves a higher proportion of correct classifications (accuracy: 0.81), the difference is small. In the F1-score—the harmonic mean between precision and recall, the models show a similar performance too. The markedness of the *accept-all-model* can not be determined because it never rejects. Furthermore, this is not an informed decision. Therefore, it's informedness is zero. The *study-model's* markedness and informedness are 0.54 and 0.55. This means that it makes informed decisions that can be trusted to be correct. Nevertheless, there is still room for improvement on both sides. Finally, the DOR of the *accept-all-model* is undefined because it does not reject. The *study-model's* DOR means that the odds of correct classifications of mouth movements are 13.49 higher than the odds of false rejections.

[3] Using `binom.test` from the `stats` package (v3.5.1) in R [stats].

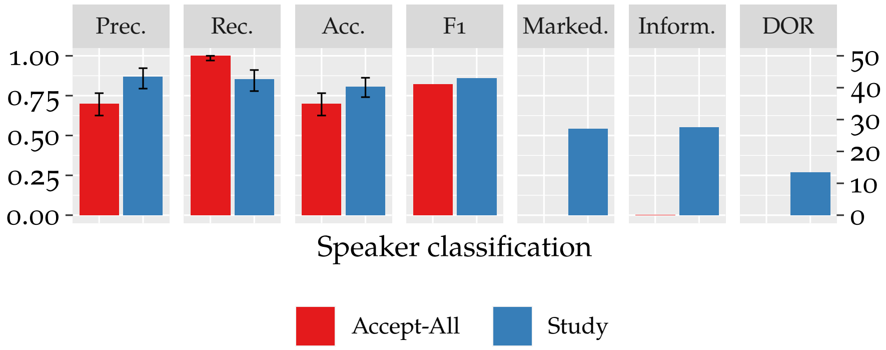


Figure 4.4: Precision, recall, accuracy, F1-score, markedness, informedness and DOR for the classification of speaking persons during the study. 95% confidence intervals are shown for precision, recall, and accuracy. markedness, and DOR are undefined for the *accept-all-model* because it does not reject. Its informedness is zero. The scale on the right side corresponds to DOR. Red bars present the results of the *accept-all-model*. Blue bars show the results of the model that was used during the study (Equation (4.2)).

4.3.1 *Discussion*

The presented observations show that, because of the prevalence of the data, a simple *accept-all-model* already achieves high measures. A threshold based model, as used during the study (*study-model*), increases the precision of the model at the cost of a decreased recall. However, the markedness, informedness, and DOR measurements show that this model can decide whether a focused person is currently speaking or not in a trustworthy and informed way. Therefore, it can be confirmed that, by observing movements of a persons lips in a conversational group, it can be visually recognized whether this person is currently speaking or not (Claim 4.1).

4.4 TURN-RELEASE DETECTION

In the presented scenario the robot is addressed with task related utterances and in anticipation of a response. A prompt to take the next turn in combination with a task-related utterance should, therefore, identify the robot as the addressee of the utterance. In Claim 4.2, I suggest that detecting mutual gaze can be interpreted as such a prompt. In this section I use the mutual gaze recognition results and ground truth annotations of the corpus to investigate (1) whether the mutual gaze recognition performs successfully and (2) whether mutual gaze at the end of an utterance is a good cue for addressee recognition during the study.

4.4.1 *Mutual Gaze Detection*

The model that was used for mutual gaze recognition during the study is defined in Equation (4.1). Therefore, I call it the *study-model* in this evaluation. Its performance is visualized in Figure 4.5 using the same metrics as in the speaker detection and with an *accept-all-model* which always assumes mutual gaze for comparison.

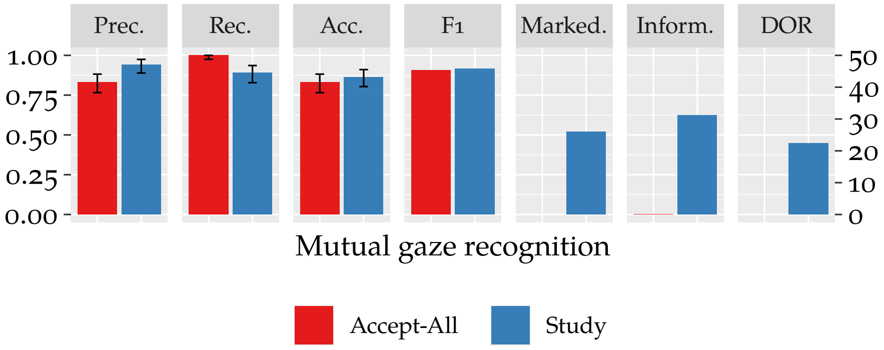


Figure 4.5: Precision, recall, accuracy, F1-score, markedness, informedness and DOR for the classification of mutual-gaze during the study. 95% confidence intervals are shown for precision, recall, and accuracy. Markedness, and DOR are undefined for the *accept-all-model* because it does not reject. Its informedness is zero. The scale on the right side corresponds to DOR. Red bars present the results of the *accept-all-model*. Blue bars show the results of the model used during the study.

The precision and accuracy of the *accept-all-model* are both 0.83, and its recall is 1. Like in the mouth movement detection, this results from the prevalence of the data and the model design (always assume mutual gaze). The *study-model* for mutual gaze recognition achieves a higher precision (0.94) and a lower recall (0.89) than the *accept-all-model*. The accuracy and F1-score measure show negligible differences. The quality measures that are not biased by prevalence—markedness (0.52), informedness (0.62) and DOR (22.34)—indicate that the mutual gaze detection performs better than the speaker classification. The strong bias for mutual gaze in the corpus does not leave much room for enhancements in accuracy and F1-score. Nevertheless, the models results are more *precise*. It is *trustworthy* and *informed*. From the observations, it can be concluded that the model for mutual gaze recognition, which was used during the study, can decide whether a focused person maintains mutual gaze with the robot or not.

4.4.2 *Addressee Deduction from Mutual Gaze*

To investigate whether mutual gaze at the end of an utterance is a good cue for addressee recognition during the study, I test the annotations (*annotation-model*) and classifications (*recognition-model*) of mutual gaze as an indicator of whether the robot is addressed. The *recognition-model* is defined in Equation (4.1). The performances of these approaches for addressee recognition can be seen in Figure 4.6.

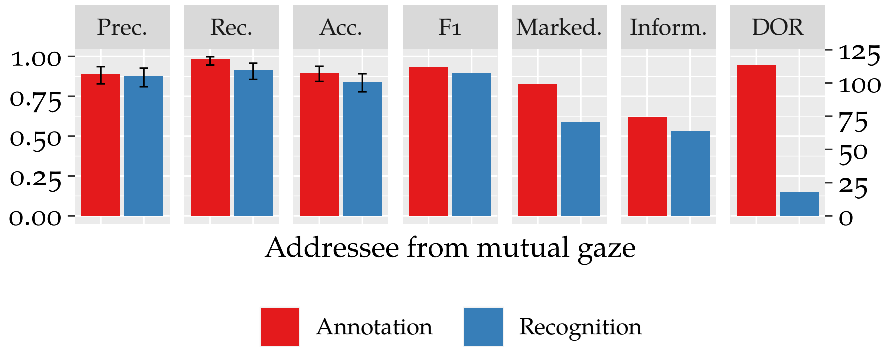


Figure 4.6: Precision, recall, accuracy, F1-score, markedness, informedness and DOR for the classification of addressee from mutual gaze. 95% confidence intervals are shown for precision, recall, and accuracy. The scale on the right side corresponds to DOR. Red bars present the results based on annotations of mutual gaze. Blue bars show the results of the classifier used during the study (Figure 4.5).

I first investigate the performance of the *annotation-model* to evaluate the applicability of having mutual gaze as a marker for being addressed in the corpus. The model's precision reveals that when the robot is looked at, it is addressed in 89% of observations. In the 17 remaining cases the robot is the speaker (65%), looking at the wrong participant (30%), or both (5%). There is a single observation in which the participant is oriented towards the robot and speaking but not addressing it. The high recall (0.98) shows that whenever the robot is addressed, it was additionally looked at by the person in almost all cases. There are two observations where the robot is addressed but not looked at by the person it is focusing. In both cases the robot focused a side-participant who in turn was looking at the actual speaker. Therefore, mutual gaze is a strong cue indicating the addressee in the interaction. Accuracy, measuring the proportion of correct classifications, is a direct measure for the consistency between mutual gaze and addressing. The high value (0.9) confirms that mutual gaze can be used as a predictor. The F1-score (0.94) does not provide additional information in this case. The measurements that are not biased by prevalence support the conclusions that were drawn from the other measurements. The markedness (0.82) shows that most predictions are correct. The informedness (0.62) of the model is lower. This is a result of the cases where the robot is looked at but not addressed. Finally, the DOR shows that it is 113.75 times more probable that the robot is addressed when it is looked at by the focused person than when not. Although being looked at by a person does not always mean that the robot is addressed, it is an indicator that strongly facilitates the decision.

The *recognition-model*, which uses the results of the automatic mutual gaze recognition to deduce whether the robot is addressed or not, shows a slightly worse performance. In the basic measures—precision, recall, accuracy, and F1-score—the results are similar to the *annotation-model*. It has a lower markedness (0.59). This means that its classifications are less trustworthy. As the precision is equally high, this means that it is more probable that a classification as *not-addressed* is wrong than in the other model. Similarly, the DOR tells that it is 17.47 times more probable that the robot is addressed when the model classifies mutual gaze than when not. Both the lower markedness and DOR originate in the higher amount of false negative and lower recall of the mutual gaze recognition (comp. Figure 4.5).

4.4.3 *Discussion*

In summary, the data shows that the robot is looked at when addressed in most interactions of the presented scenario. There is a high correlation between the robot being looked at and being addressed when it recognizes a task related utterance. Therefore, it is helpful to detect mutual gaze as a prompt for the robot to take the turn. Furthermore, it is possible to automatically detect mutual gaze with the focused person during the study, and the results are a good cue for addressee recognition. However, looking at the robot serves multiple purposes. It is not always a prompt to take the turn. This can be seen in the precision of the *Annotation* based prediction. Nevertheless, on the basis of the data, Claim 4.2 can be confirmed. Recognizing mutual gaze at the end of an utterance can be interpreted as a prompt to take the next turn.

4.5 BAYESIAN ADDRESSEE RECOGNITION

In Claim 4.3, I propose to combine information about the interlocutors mouth movements and gaze, and context information to predict if the robot was addressed with an utterance or not. To investigate this proposition, I use the classifications of mouth movements (*Mouth*) and mutual gaze (*Gaze*) as recognized by the robot during the study. Additionally, I inspect if the robot was speaking at the moment of task recognition (*Speaking*), and which task was recognized by the robot (*Task*). I use these features to predict if the robot was addressed (*Addressee*)—for which the ground truth can be drawn from the manual annotations in the corpus. For the evaluation, I combine these five variables into Bayesian Networks.

4.5.1 *Bayesian Models*

The networks use *Addressee* as a parent node which influences the outcome of other variables. To evaluate the individual influences of *Mouth* and *Gaze*, I create corresponding models which only use these variables:

$$(\mathit{Mouth}) \quad \mathit{Addressee} \rightarrow \mathit{Mouth}$$

$$(\mathit{Gaze}) \quad \mathit{Addressee} \rightarrow \mathit{Gaze}$$

To assess their combined performance, I create a Bayesian Network with both variables:

$$(\mathit{Both}) \quad \mathit{Addressee} \rightarrow \{\mathit{Mouth}, \mathit{Gaze}\}$$

Finally, to investigate the impact of contextual information I extend this network with knowledge about the robots inner state and recognized task:

$$(\mathit{Both{+}Self}) \quad \mathit{Addressee} \rightarrow \{\mathit{Mouth}, \mathit{Gaze}, \mathit{Speaking}\}$$

$$(\mathit{All}) \quad \mathit{Addressee} \rightarrow \{\mathit{Mouth}, \mathit{Gaze}, \mathit{Speaking}, \mathit{Task}\}$$

The networks are visualized in Figure 4.7.

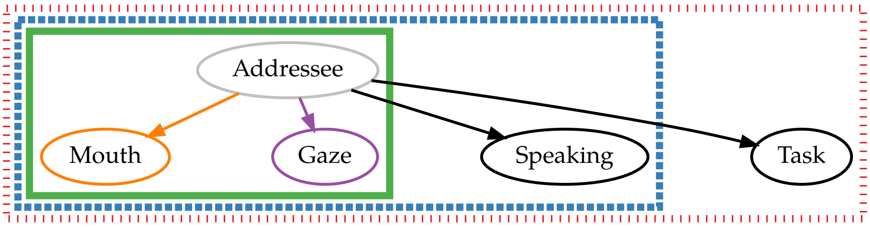


Figure 4.7: Bayesian Network structure used in the evaluation. The nodes depict variables in the corpus, the arrows show dependency relationships. This visualization encodes five different networks: (*Mouth*) orange arrow, (*Gaze*) violet arrow, (*Both*) green-lined box, (*Both+Self*) blue-dashed box, and (*All*) red-dotted box.

To assess if these networks can predict whether the robot is addressed or not, I perform a CV. To this end, the corpus is split according to the five trials, trained using four trials and tested on observations from the remaining trial. Furthermore, to assess the influence of misclassifications of the used mouth movement and mutual gaze detection systems, the same procedure is repeated with annotated inputs. The strength of the model's belief for each observation and model is used to create receiver operating characteristic (ROC) curves and calculate the corresponding AUC (Figure 4.8). Furthermore, the visualizations are augmented with the performance of the models that were available

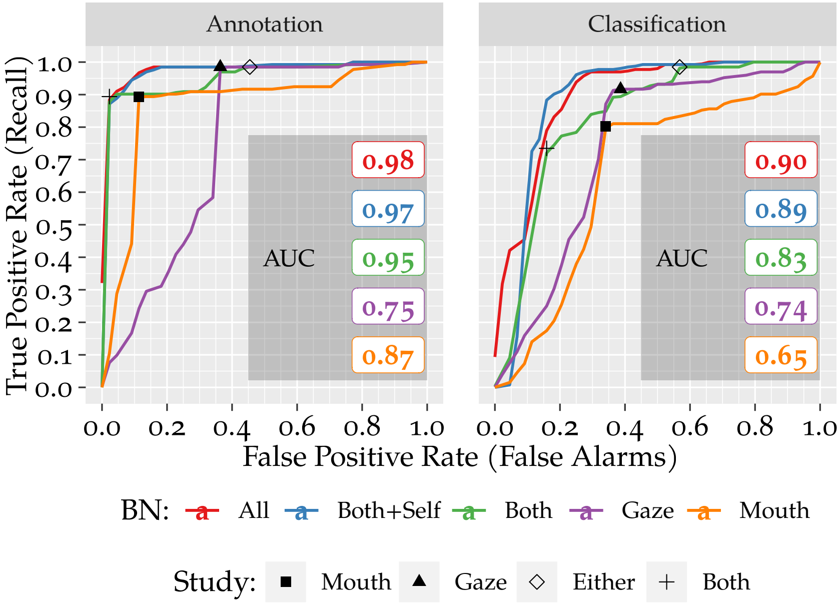


Figure 4.8: Performance visualizations of addressee recognition models in the ROC space. Mutual gaze and mouth movement information is taken from *Annotation* (left) or *Classification* (right). The ROC curves (lines) and corresponding AUC (labels in the gray box on the lower right side) are visualized for the Bayesian Network models BN: *All*, *Both+Self*, *Both*, *Gaze*, and *Mouth*. Additionally, the results of the models that were available during the study—*Study*: *Mouth*, *Gaze*, *Either*, and *Both*—are illustrated as shapes.

during the study (*Study-Models*). These are *Mouth* (addressed if mouth movement was detected), *Gaze* (addressed if mutual gaze was detected), *Either* (addressed if either of them was detected, non-exclusive), and *Both* (addressed if both of them were detected). These can not be shown as ROC curves because they provide only a fixed result. As the observations in the corpus are unbalanced, additionally precision-recall curves, corresponding AUC and *Study-Models* results are shown in Figure 4.9.

4.5.2 *ROC Performance*

When looking at the ROC curves of the Bayesian Networks with perfect inputs (*Annotation* in Figure 4.8) multiple observations can be made. First of all, mutual gaze alone is not a good predictor for addressing. The *Gaze* model achieves an overall AUC of 0.75 and a recall of > 80% only with ≈ 35% false alarms. At this point, there is a steep increase in recall, allowing the model to achieve ≈ 0.97 recall at 36% false positives. The *Mouth* model achieves a recall of ≈ 0.9 at only ≈ 10% false alarms. Its recall remains under 0.95 until a high false positive rate (FPR) of ≈ 0.75. The overall better performance is reflected in its AUC (0.87). By combining both mutual gaze and mouth movement information, the addressee prediction can utilize the strengths of both. The *Both* model achieves a recall of ≈ 0.9 at $\approx 5\%$ false alarms and shows an increase to ≈ 0.97 at 36% false alarms. Therefore, it can provide the high recall of the *Gaze* model and simultaneously improve upon the *Mouth* models low FPR. By taking the robots inner state into account, the *Both+Self* model shows further recall enhancements in the lower range of the FPR. It achieves a recall of > 0.95 at 10% false alarms and an AUC of 0.97. The *All* model that additionally uses the type of the recognized task achieves a slightly better AUC (0.97) but does not show salient differences from *Both+Self* in its curve.

The Bayesian Networks, which utilize results of the models from the study (*Classification* in Figure 4.8), all achieve lower AUC results than the models with perfect inputs. *Gaze* shows a similar trajectory but an AUC of 0.74. It has a recall of ≈ 0.9 at 35% false alarms and only small improvements with a growing FPR. The performance of the *Mouth* model is much worse on the automatically classified data. It only achieves an AUC of 0.74 and performs worse than the *Gaze* based model. The remaining models—*Both*, *Both+Self*, and *All*—use their additional information to gain further improvements. Nevertheless, the noise in the classifications has a strong effect on the overall addressee recognition.

The positions of the *Study-Models* in both the *Annotation* and the *Classification* set-up are located in the vicinity of the *BN* models optimal results for equal false positive and false negative costs. As both have the same input information and the Bayesian Network optimizes for accuracy, this is a consequence of the models. By choosing the threshold for the classifications of the Bayesian Networks a trade-off can be chosen between recall and FPR.

4.5.3 *Precision-Recall Performance*

To get a better insight in the created models, precision-recall curves, corresponding AUC, and *Study-Models* results are shown in Figure 4.9. With this visualization it is easier to take the prevalence of the corpus into consideration. The precision of the models for low $recall < 0.4$ is noisy, and not of great interest because these configurations reject the majority of interactions. Therefore, I do not further elaborate on this range. The interesting area of the visualizations is on the upper left quarter, where a trade-off is made between the two dimensions.

By looking at the models' performances in case of perfect input data (*Annotation*), some properties of the models can be observed. For $recall < 0.85$ the precision of all models is nearly constant with growing recall. This allows optimizing for recall without loosing precision. The models show a similar development in the precision-recall space as in the ROC curves. The *Gaze* model performs worse than *Mouth* in case of perfect data until a recall of ≈ 0.9. At this point the *Mouth* model exhibits a strong drop in precision while the *Gaze* model keeps

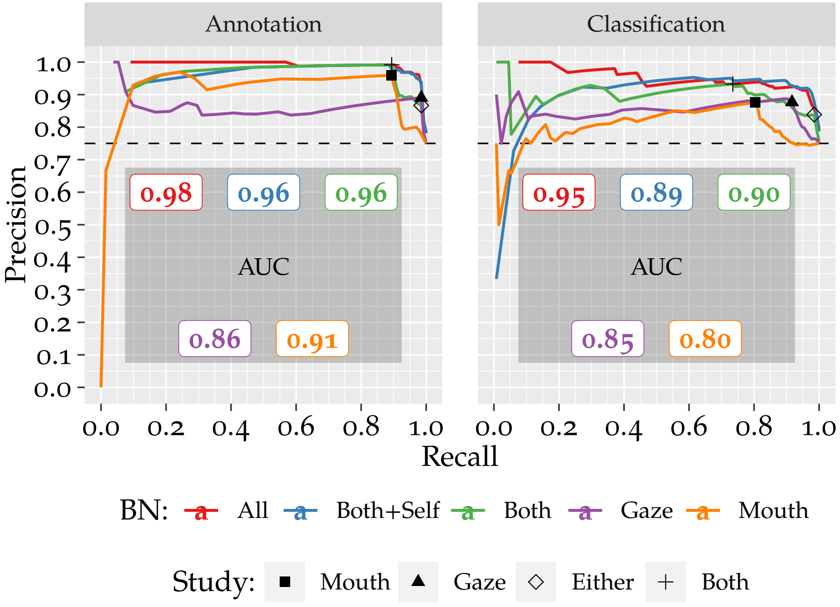


Figure 4.9: Performance visualizations of addressee recognition models in the precision-recall space. Mutual gaze and mouth movement information is taken from *Annotation* (left) or *Classification* (right). The precision-recall curves (lines) and corresponding AUC (labels in the gray box on the lower half) are visualized for the Bayesian Network models *BN*: *All*, *Both+Self*, *Both*, *Gaze*, and *Mouth*. Additionally, the results of the models that were available during the study—*Study*: *Mouth*, *Gaze*, *Either*, and *Both*—are illustrated as shapes. The dashed line represents the results of a baseline model (precision equals prevalence).

its—overall lower—precision until a recall of ≈ 0.98. The *Both* model joins information of mutual gaze and mouth movements to achieve an overall better precision and compensate for the drops in quality of the individual models. As a side effect, its results show a drop in precision at around *recall* ≈ 0.95. Therefore, for a high recall, the results of the *Both* model have a lower precision than the *Gaze* model. The models with additional information, *Both+Self* and *All*, outperform the other models throughout the whole range. Their loss in precision for a recall between 0.95 and 0.99 is only gradual, allowing to accurately choose a suitable trade-off between these measurements for a specific situation.

The precision-recall curves for the *Classification* based analysis show more noise than in the *Annotated* analysis but an otherwise similar shape. The *Mouth* model performs worse than the *Gaze* model with classified input. Its precision is lower in the majority of observations, achieves ≈ 0.87 at recall ≈ 0.8 and decreases for higher values. The *Gaze* model achieves ≈ 0.88 at a recall of ≈ 0.9. In the region around this point ($0.86 < \mathit{precision} < 0.92$) it performs better than the *Both* model which is otherwise always better. The *Both+Self* and *All* models achieve a precision of ≈ 0.9 at a recall of ≈ 0.95 and show only a gradual decrease in precision for higher recall values.

The positions of the *Study-Models* in the *Annotation* and the *Classification* set-up are located at points with a high *recall*, right before a strong drop in *precision*. Therefore, they are optimal for an equal weighting of precision and *recall*. The *Study-Either-Model* optimizes for better recall and the *Study-Both-Model* for better precision on the curve of the Bayesian Network based *Both* model.

4.5.4 *Discussion*

The ROC and precision-recall curve analyses show that information about the gaze and mouth movements of an interlocutor can be used as a predictor for addressee recognition. The features have different properties. Mouth movement information, on the one hand, can achieve a high recall with few false alarms and an overall high precision. Mutual gaze information, on the other hand, generates a higher amount of false alarms but can achieve a higher recall. Its has a lower overall precision but can keep it at that level for much higher recall values. By taking both features into account, a model can be created that combines their advantages to produce overall better results. This model can be further enhanced by taking other contextual information into account. Furthermore, by choosing an appropriate threshold for the model's belief, a trade off between recall, and precision or fall-out can be made. Misclassifications in mouth movement or mutual gaze detection directly result in a degradation of the addressee recognition. This observed effect is stronger in case of wrong mouth movement detections. However, in summary it can be said that information about mouth movements and the gaze of a participant in a conversational group can be combined with context information to recognize if a robot is addressed with an utterance or not. Therefore, Claim 4.3 can be confirmed.

4.6 SUMMARY

In this chapter, I investigated RQ 2[C]. To this end, I compiled the claims that in a mixed human-robot conversational group: mouth movements assert that a person is the current speaker of the group (Claim 4.1), mutual gaze at the end of an utterance can be interpreted as a prompt to take the next turn (Claim 4.2), and these informations can be combined with context information to create an addressee recognition model for an artificial agent (Claim 4.3). I presented a HRI scenario, which was specifically designed to challenge the robot's addressee recognition skills. This scenario was conceived and implemented in a study in a

[C] RQ 2: How can an artificial agent visually recognize whether it was addressed by a person within its conversational group or not?

joint effort by me and my colleagues in the CSRA [Ric+16]. I created an addressee recognition model, based on the stated claims, and applied it during this study. Furthermore, I augmented the resulting corpus with ground truth annotations about the robots interlocutors gaze and speaking state. On the basis of the resulting corpus I was able to test the presented claims. In Section 4.3 I assessed the recognition performance of the mouth movement detection. I confirmed Claim 4.1 by showing that the proposed mouth movement detection model can be used to distinguish speaking from non-speaking interlocutors. In Section 4.4 I assessed the recognition performance of the mutual gaze detector and its applicability as a cue to take the next turn. I confirmed Claim 4.2 by showing that the proposed model can recognize situations in which the robot is looked at and that this information can be used to predict if the robot needs to take the next turn. In Section 4.5 I created multiple Bayesian Networks and evaluated their performance on annotated and automatically classified mouth movement and mutual gaze detections and with additional contextual information. I assessed the individual power of mouth movement detection and mutual gaze detection for addressee recognition and the improvements that can be achieved by combining these features. Furthermore, I examined the effect of errors in the recognition of these features. I confirmed Claim 4.3 by showing that a model that combines information from mouth movements and mutual gaze performs better than models using only one of these features and adding contextual information further improves the model's performance.

The scenario and corpus that was investigated in this chapter was specifically designed to challenge robotic addressee recognition. Therefore, the robot had to continuously take part in an conversation with multiple people and was confronted with utterances that—while having exactly the same content—may have been addressed towards anyone within the group. This allowed to investigate the possibilities of visual addressee detection. Nevertheless, the presented scenario reflects human addressing behaviour in a narrow field, within a fixed conversational group, and an explicit task. Addressing behaviour in other scenarios—e.g. when the participants have no means to establish a conversational group, or when they converse without having a fixed task—can depend on other forms of establishing the addressee of an utterance and be much more dynamic. Furthermore, the participants of the study were all students of a German university and most of them had a technical background. Observations of people from different age-groups or with different cultural backgrounds have the possibility to enrich the obtained insights.

The observations in this chapter, pose some answers to the investigated RQ 2. When a robot in a conversational group recognizes an utterance, it can visually observe multiple cues of its interlocutors to decide whether this utterance was addressed at it or not. It can monitor its interlocutors mouth movements to decide which participant of the group speaks, and it can observe their gaze to predict the next speaker. By combining this information, the robot can make an informed decision and balance the costs of falsely assuming and falsely ignoring human speech. Considering contextual information can further enhance the quality of this decision.

Part III

GROUPS & ROLES IN COPRESENCE

In this part, I broaden the social skills of artificial agents by investigating how mixed human-agent conversational groups can be detected and how the conversational roles of the agent within such a group can be recognized. To this end, I present an appropriate scenario and create a corresponding corpus. On this basis, I evaluate F-Formation detection as an approach to detecting conversational groups in an unconstrained interaction between a group of people and two virtual agents with changing constellations. Furthermore, I evaluate different approaches to detecting the conversational role of the agent in such constellations.

5 HUMAN-AGENT INTERACTION CORPUS

> By definition, an accessible engagement does not exhaust the situation; [. . .] What we find instead is some obligation and some effort on the part of both participants and bystanders to act as if the engagement were physically cut off from the rest of the situation. [Gof63, p. 156]

To investigate RQ 3 and RQ 4, a corpus of open, multi-centric interactions between people and artificial agents is needed. In this chapter, I discuss the requirements of such a corpus and present a suitable scenario. On this basis, I create a corpus of interactions between a group of people and the Flobi agents in the CSRA. This corpus is utilized in the following chapters to examine the recognition of conversational groups (Chapter 6) and conversational roles (Chapter 7) of artificial agents.

5.1 INTRODUCTION

As discussed in Section 2.1, people who are copresent, always interact with each other in some way. In a focused interaction they turn towards each other, reduce their distance, and increase the frequency and duration of mutual gazes (Section 2.1.3). Moreover, they can have different conversational roles which are dynamically negotiated using the turn taking system. In an unfocused interaction, people display that they are part of the situation. They acknowledge the presence of others but direct their attention somewhere else to show civil inattention. They reduce mutual gazes, increase their distance and turn away from each other (Section 2.1.2). It is necessary for people to be able to distinguish and display whether and with whom they are in a focused or unfocused interaction. Otherwise, they are not able to act appropriately and may be perceived as offensive [Gof63, p. 29, p. 157].

Artificial agents—e.g. in form of robots or virtual agents—are potential interaction partners. In a smart environment that contains interactive artificial agents, people are always in their copresence. To be able to behave in a manner that is appropriate to the situation, an agent first needs to understand it. As presented in Section 2.2.2.4, there already exists work on the detection of human-only conversational groups. However,

RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?

RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

work on the detection of mixed conversational groups in HAI is sparse. The possible effects of specific formations and the relevance of human conversational groups for socially acceptable robotic navigation are more in focus of HAI research (Section 2.2.2.5). As a sub-problem of conversational role recognition, approaches to the recognition of utterances addressed towards the agent are presented in Section 2.2.2.2. However, the presented interactions are restricted. The agent is controlled by a wizard, its conversational group is fixed or the participants are augmented with close-talk microphones. To investigate, how dynamically changing conversational groups with people and artificial agents can be detected and how the conversational roles within such groups can be recognized, a fitting interaction corpus is required.

In the conversational group detection community there are multiple, frequently used datasets for the evaluation of recognition models. The most widely adopted are *Synthetic* and *Coffee Break* from Cristani et al. [Cri+11a], *IDIAP Poster Data* from Hung et al. [HK11], *Cocktail Party* form Setti et al. [Set+13], and *GDet* from Loris et al. [Lor+13]. They are used for the evaluation of vision based conversational group detection models as performed by Setti et al. [Set+15] and Vascon et al. [Vas+16]. Furthermore, there are multi-modal datasets like *SALSA* from Alameda-Pineda et al. [Ala+16] and *MatchNMingle* from Cabrera-Quiros et al. [Cab+18]. The authors use these datasets to evaluate multi-modal approaches to conversational group detection and subsequent analyses of group properties. All these datasets show similar situations: A crowded place where people stand and interact at a poster presentation, coffee break, speed-dating or similar socializing event. These scenarios are intentionally chosen to allow people to act as natural as possible while simultaneously allowing the observation of many interactions and changing groups. None of these scenarios feature artificial agents.

Datasets for the analysis of conversational roles, often use fixed groups of people without artificial agents [JAN06; Akh+17; MTH18]. If artificial agents are present in such interactions, they are controlled by a wizard [Tur+05; Jay+13], or only distinguish utterances addressed at them from others [BH11; SJB15]. Furthermore, these systems consider conversational groups only as a distinction between interacting and non-interacting people, using simple heuristics. Additionally, if people without experience in HAI find themselves in copresence with an artificial agent its novelty may lead to a higher amount of attention and overall amplified measures. Therefore, a scenario is needed in which a group of people can freely interact with each other and with artificial agents for an extended time. This way they can get accustomed to the situation and act more natural.

5.2 SCENARIO

To fulfil the presented requirements, I collect a new corpus in the CSRA. For a visualization of the recorded videos see Figure 5.1. The corpus contains the following scenario: A group is invited into the apartment for a demonstration composed of three parts.

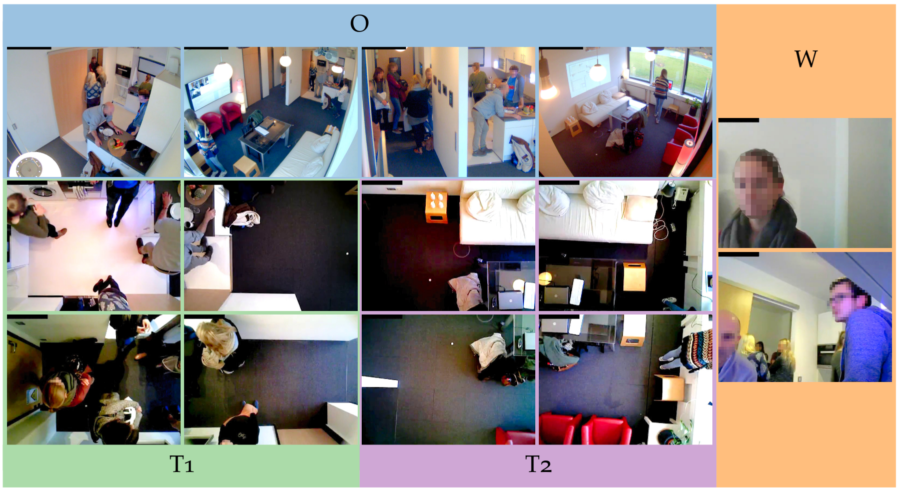


Figure 5.1: The 14 perspectives from which videos are recorded during the study. The overview perspectives are in the top row (*O*, blue background), the web-cam perspectives are in the right column (*W*, orange background), and the remaining images show top-down perspectives (*T1:* kitchen and hallway with green background and *T2:* living-room with violet background).

BRIEFING In the first part, the presenter—a person that is acquainted with the environment and realizes the demonstration—and the guests gather in front of the apartment. The participants get an explanation about what the apartment is, what is being recorded during the study, and their rights regarding data privacy. After the participants gave their consent, the main part of the demonstration begins.

PRESENTATION The presenter enters the apartment together with the participants and guides them through the hallway, living room and kitchen. In each room the group stops and the presenter gives information about the apartment or shows interaction possibilities. In the hallway and kitchen, respectively an interaction with the virtual agent Flobi—the host of the apartment—is performed. The living room is used to give information about the apartment's actuation and introspection capabilities.

FREE INTERACTION In the third part of the demonstration, the participants are allowed to freely chat, test the different control metaphors of the apartment or interact with the virtual agents. During this period, the presenter remains in the apartment to answer further questions.

The scenario is recorded using the apartment's study facilities. The recording is started before the apartment is entered and stopped after the last participant left.

5.3 RECORDING

With this approach I perform a long demonstration for a group of students and their lecturers. The recording contains (1) four overview videos from the apartment's corners, (2) two web-cam videos from the agents' perspective, (3) eight top-down videos, (4) two audio streams, and (5) the communication between software components of the CSRA. To see the video perspectives, refer to Figure 5.1. This resulting 57 min of unconstrained, mixed human and human-agent interaction are composed of 20 min presentation and 37 min free interaction. The participants are 8 women and 3 men—including the presenter. This dataset is a good basis for the analysis of mixed, human-agent interactions. Before it can be used to evaluate automatic conversational group detection and conversational role recognition in mixed human-agent interactions, some processing and annotation is required. A visualisation of the annotations can be seen in Figure 5.2.

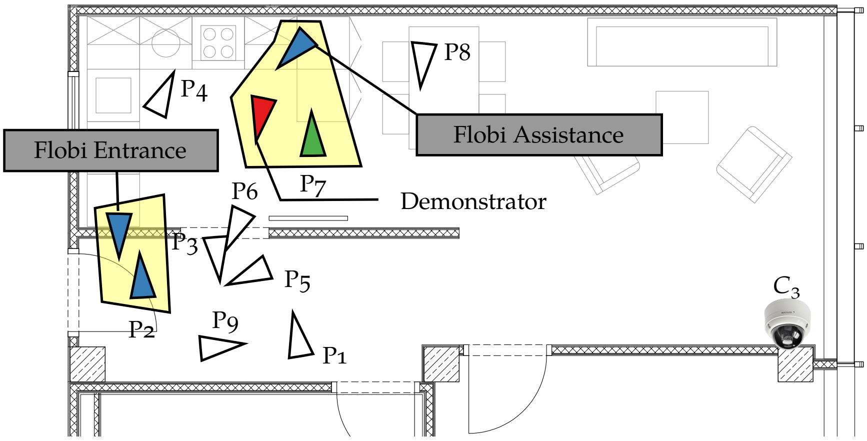


Figure 5.2: Annotations of the scene from Figure 5.1. The participant's and agent's poses are shown as triangles with annotated *participant-id*. Conversational groups are highlighted in yellow. Conversational roles are shown in red (speaker), green (addressee), blue (side-participant), and white (non-participant).

5.4 ANNOTATION

The study recording contains multiple video streams, audio streams and communication between software components of the CSRA. To be able to examine RQs 3 and 4, a set of ground truth annotations is required. A visualization of the annotations for the scene in Figure 5.1 can be seen in Figure 5.2. To this end, the required information is annotated on the basis of the top-down videos of the hallway and kitchen *T1*. The remaining areas (*T2*) of the apartment are not annotated because they are not relevant for interactions with the Flobis. The following information is manually annotated: (1) Pose (position and rotation) of all participants and Flobis, (2) participant ids, (3) all conversational groups the Flobis participate in, and (4) conversational roles for all annotated conversational groups. Annotations are done whenever a change in the participants pose, a group, or a role can be observed. In comparison to annotating fixed time intervals, this reduces the amount of needed annotations and allows an arbitrary sampling of annotations. Participants poses can be interpolated between annotations and groups and roles are static between changes. Furthermore, poses of the same person, when annotated from multiple viewpoints, can be averaged to further enhance estimations of the participants pose. The final dataset is created by sampling these annotations at a fixed rate of 15 Hz. The overall distribution of group-sizes for the two agents at this sampling rate can be seen in Figure 5.3.

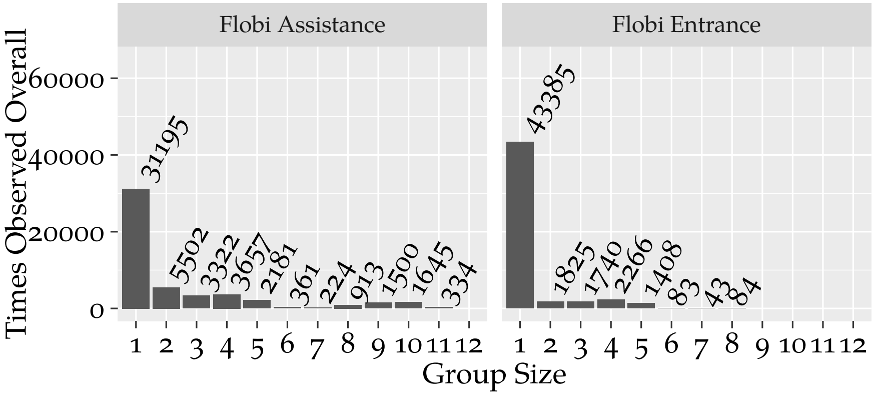


Figure 5.3: The distribution of observed group sizes for the two agents in the group annotations of the corpus. The counts are additionally shown on top of the bars. A group size of 1 means that the agent is alone—not in a group with others.

5.5 AUTOMATIC DATA EXTRACTION

To assess the possibilities of fully automatic recognition of conversational groups and conversational roles, the following features are extracted from the system communication recordings: (1) Positions of persons in the apartment (at 30 Hz) from the person tracking system. Furthermore, in the FOV of both Flobis (*W* perspectives in Figure 5.1): (2) the Region of Interest (ROI) of each detected face, (3) the recognized gaze-directions, and (4) the detected facial landmarks (A visualization of these features can be seen in Figure 4.2 on page 69). Finally, for each agent (5) whether it is speaking.

As the person tracking of the CSRA does not provide rotational information, I additionally apply the *OpenPose: Real-Time Multi-Person Keypoint Detection Library for Body, Face, Hands, and Foot Estimation* by Gines et al. [OpenPose] to the overview recordings (*O* in Figure 5.1) to detect 2D key-points of participants in video coordinates (see Figure 5.4). On the basis of these key-points I create additional person

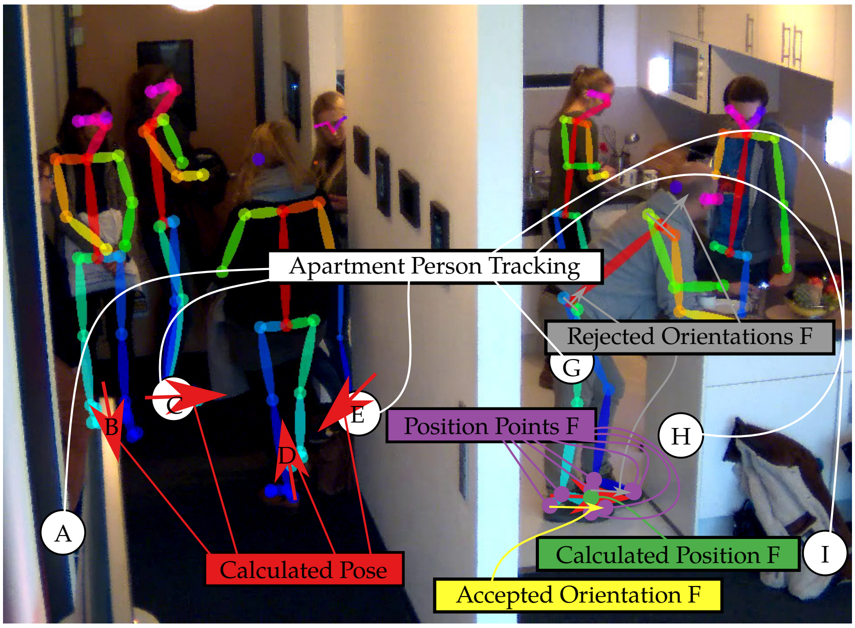


Figure 5.4: Visualization of the automatic pose extraction. The image shows the scene in Figure 5.1, from the perspective of C_3 (see Figure 5.2), augmented with person key-points as they are detected by Gines et al. [OpenPose]. White circles (*A*, *C*, *E*, *G*, *H*, and the *false-positive I*) depict positions as they were detected by the apartment's person tracking system. Red arrows (*B*, *C*, *D*, *E*, and *F*) depict poses—position and rotation—as they were derived from the results of Gines et al. [OpenPose]. The information used to derive poses is highlighted for *F*. In this case, violet points depict points of the feet that are used to calculate the position (green point), gray arrows show rejected orientation hypotheses, and the yellow arrow depicts which orientation hypothesis was accepted.

hypotheses as follows: For each detected person, the mean of all detected feet-key-points is calculated as the position in image coordinates. The person's orientation is estimated from the following orientation-candidate vectors:

SHOULDERS perpendicular to left shoulder → right shoulder

HIP perpendicular to left hip → right hip

LEFT FOOT along left heel → left big toe

RIGHT FOOT along right heel → right big toe

The longest of these vectors is chosen as the most reliable cue and its direction accepted as the person's orientation in image coordinates. Transforming this position and orientation into the floor-plane of the apartment, results in an additional source for automatic tracking of persons in the apartment.

To fuse detections from multiple camera perspectives and the person tracking, and select the corresponding annotations, the optimal assignment is calculated using the *C++ Implementation of the Hungarian Algorithm* by Justin Buchanan et al. [hun]. This identification of person percepts allows the evaluation of the fully automatic detections. A visualization of the annotated and detected overall movements of the participants in the corpus can be seen in Figure 5.5.

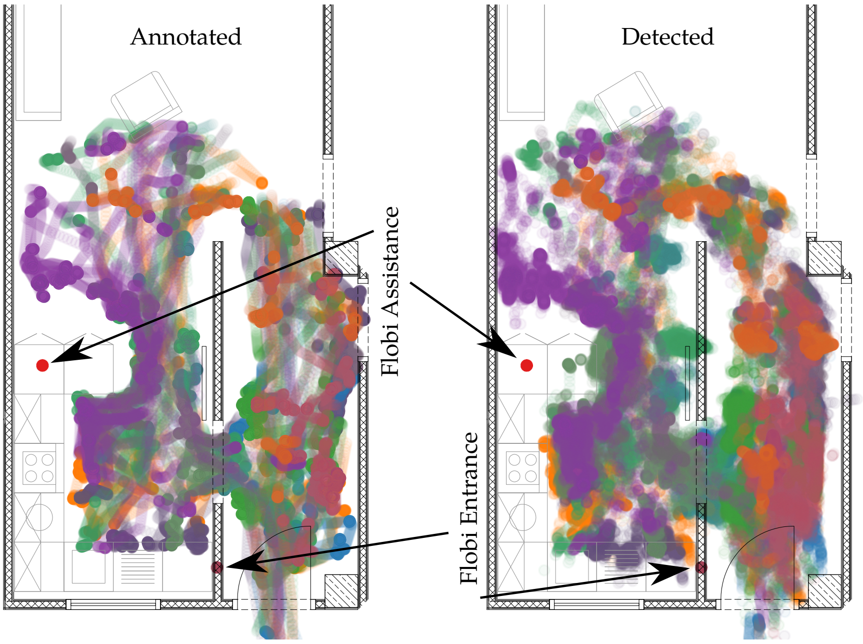


Figure 5.5: All positions of participants and agents in the *T1*-region during the study. Observations are sampled at 15 Hz from the annotations (left) and automatic detections (right). Automatic detections are created by fusing results of the apartment's person tracking and the person detection based on Gines et al. [OpenPose] keypoint detection. Each side shows ≈ $7 \cdot 10^5$ observations as points with 95% transparency. Different colours represent the positions of different persons. Flobi positions are constant.

The annotated and automatically extracted poses and annotations of conversational groups can be used to examine RQ 3ᶜ. The results of the apartment's face detection and gaze recognition can be used for further analyses. Furthermore, RQ 4ᶜ can be investigated by combining

ᶜ RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?

ᶜ RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

the results of the conversational group detection with other features. These investigations are performed in Chapters 6 and 7.

5.6 SUMMARY

In this chapter I created the foundation for the investigation of RQ 3[↺] and RQ 4[↺]. To this end, I presented the requirements a corpus for the investigation of conversational groups and conversational roles in dynamically changing interactions of humans with artificial agents needs to meet. After showing that available corpora do not meet the requirements, I presented a new HAI scenario in the CSRA. I recorded an extended interaction and created an approach for fully automatic detection of person positions and orientations. Finally, I created ground truth annotations of conversational groups and conversational roles, which can be used for the investigation of the presented research questions.

[↺] RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?

[↺] RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

6 CONVERSATIONAL GROUP DETECTION

The literature presented in Section 2.1.2 suggests that people, who perceive an agent as copresent, show similar behaviours and have similar expectations towards it as in interactions with humans. To be accepted in long term interactions, agents therefore need to understand these behaviours and the implied expectations. They need to know when to interact and when to show civil inattention (Section 2.2.2.1). Without distinguishing participants of the agents conversational group from non-participants, it can not treat them accordingly. The detection of conversational groups in the presence of humans and artificial agent is therefore an important basis for conversational role recognition and behaviour recognition. Therefore, I investigate RQ 3[C] in this chapter. To this end, I use the corpus of unconstrained, mixed human-agent interactions presented in Chapter 5. Because of the discussed similarities between peoples behaviour towards humans and artificial agents, I investigate this research question by examining the following claim:

CLAIM 6.1 (GROUP FROM F-FORMATION) *Mixed conversational groups of people and artificial agents can be detected using F-Formations as known from HHI.*

It is often sufficient to know if the agent is in a group or not without identifying the participants of the group. Furthermore, the detection of F-Formations requires a good understanding of the distribution of persons in the scene which can be computationally expensive and not always feasible. Therefore, I propose the following simplifications for comparison:

CLAIM 6.2 (GROUP FROM MUTUAL GAZE) *The detection of peoples gaze direction in the agents field of view can sufficiently inform about whether it is in a conversational group or not.*

CLAIM 6.3 (GROUP FROM FACES) *The detection of faces in the agents field of view can sufficiently inform about whether it is in a conversational group or not.*

In this chapter I evaluate the detection of conversational groups for virtual agents in a smart environment. To this end, I first evaluate Claim 6.1—the portability of F-Formation detection as presented by Setti et al. [Set+15] from human groups to mixed human-agent groups—on automatically extracted data (see Section 6.1). Subsequently, I evaluate

[C] RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?

Claims 6.2 and 6.3 by testing if the agents' participation in a conversational group can be deduced from gaze directions and detected face sizes (see Section 6.2).

6.1 F-FORMATION DETECTION

The participants of a conversational group need to optimize their mutual access to the joint interaction space (o-space). This is achieved by overlapping their transactional segments (Section 2.1.3). One approach for F-Formation-detection in HHI, is presented by Setti et al. [Set+15]. It uses 2D positions $[x, y]$ and orientations θ of persons in an open space to estimate the centres of their transactional segments (TS). This segment is assumed to be in front of the person with its centre at a fixed distance called stride (S). Based on the distance between the TS and the o-space of a potential group and its visibility for a person, a cost function can be created. This cost-function should be zero for a perfect overlap of transactional segment and o-space without obstacles, and grow when they move apart or the o-space is occluded. A good assignment of persons to conversational groups can then be found by optimizing the overall costs for a scene. In Figure 6.1 an exemplary scene can be seen with visualizations of the relevant variables. The used variables are explained in the next subsection.

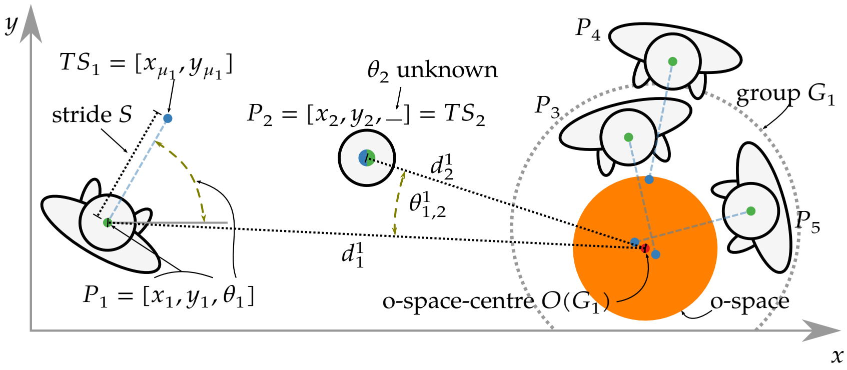


Figure 6.1: A visualization of a scene in the F-Formation detection. People (P_i) are defined as positions $[x_i, y_i]$ (green points) with an optional orientation θ_i. P_2 is shown as a circle because there is no known orientation. Centres of transactional segments $(TS_i = [x_{\mu_i}, y_{\mu_i}]$, blue points) are at the distance S (stride) in front of persons or—when orientation is unknown—at their position $(TS_2 = P_2)$. P_3 and P_5 form the conversational group G_1 (gray dashed circle) with the o-space centre $O(G_1)$ (red dot). P_4 is not part of G_1 because the o-space is occluded by P_3. The transactional segments of P_1 and P_2 are too far away from $O(G_1)$ for them to be part of the group. The distances d_1^1 and d_2^1 represent the distance between the o-space centre $O(G_1)$ (superscript index) and the position of P_1 or P_2 (subscript index) respectively. The angle $\theta_{1,2}^1$ is the angle between P_1 and P_2 (subscript index) regarding the o-space centre $O(G_1)$ (superscript index).

To investigate if this approach can be used to detect conversational groups with an artificial agent, I create the open-source group detection framework *fformation* [ffm]. It models observations of persons and the cost of assigning them to conversational groups according to the cost function presented by Setti et al. [Set+15] with an adaptation to cover observations with unknown orientation.

6.1.1 *Assignment Costs & Detectors*

The cost function is drawn from Setti et al. [Set+15]. The calculation is done with the following definitions in mind: An observation is defined as a set of persons $P = \{P_1, \dots, P_n\}$ with 2D positions and orientations—i.e. $P_i = [x_i, y_i, \theta_i]$. From a person's pose, the transactional segment TS_i can be computed using the stride (S), which encodes the expected distance between a persons position and transactional segment. If the orientation of a person is not known, the TS_i can not be computed. Therefore, I extend the TS_i formula to return the original position of the person when the orientation is not known. This equals to the mean TS_i over all possible orientations. The calculation of transactional segments is visualized in Figure 6.1 and formalized as follows:

$$TS_i = [x_{\mu_i}, y_{\mu_i}] = \begin{cases} [x_i + S\cos(\theta_i)], y_i + S\sin\theta_i], & \text{if } \theta_i \text{ is known} \\ [x_i, y_i], & \text{otherwise} \end{cases}$$

This allows calculating $TS = \{TS_1, \dots, TS_n\}$ from $P = \{P_1, \dots, P_n\}$.

An assignment of persons to groups is defined as the set of groups $G = \{G_1, \dots, G_m\}$ with each person assigned to exactly one group. Groups with $|G_k| = 1$ are allowed and represent persons that do not participate in a conversational group. Consequently, the group of person P_i is unambiguous and can be defined as $g(P_i)$. The o-space-centre $O(G_k)$ of a group G_k is calculated from the transactional segments of its participants.

$$O(G_k) = [u_{G_k}, v_{G_k}] = \frac{\sum_{i \in G_k} TS_i}{|G_k|}$$

The o-space-centre that corresponds to the group of P_i is therefore:

$$O(g(P_i)) = [u_{g(P_i)}, v_{g(P_i)}]$$

assignment cost The overall cost of an assignment G (Equation (6.1)) is calculated as the sum of o-space-distance costs (Equation (6.2)), o-space-visibility costs (Equation (6.3)) and an additional Minimum Description Length (MDL)-prior (Equation (6.4)):

$$\underbrace{C(G)}_{\text{assignment cost}} = \underbrace{D(G)}_{\text{distance cost}} + \underbrace{V(G|P)}_{\text{visibility cost}} + \underbrace{P(G)}_{\text{MDL-prior}} \tag{6.1}$$

The different parts of the formula are explained in the following. A detailed motivation and in-depth analysis of this cost function can be looked up from Setti et al. [Set+15].

DISTANCE COST The *distance cost* $D(G)$ penalizes deviations of transactional segment centres from the groups' o-space. It therefore is zero when the transactional segments of the participants and the o-space perfectly overlap and rises when the distance between transactional segment and o-space centre grows. A high distance cost, means that the o-space is too far away from the person to be maintained effectively. The distance cost is calculated as follows:

$$D(G) = \sum_{i \in P} (u_{g(P_i)} - x_{\mu_i})^2 + (v_{g(P_i)} - y_{\mu_i})^2 \tag{6.2}$$

VISIBILITY COST The *visibility cost* ensures that the o-space $O(g(P_i))$ of a person P_i's conversational group is not occluded by any other person $P_{j \neq i}$. Because the visibility can be occluded by any other person in the scene, the overall visibility cost for an assignment is the sum over the visibility constraint for all two-person permutations from P:

$$V(G) = \sum_{i,j \in P, i \neq j} \underbrace{R_{i,j}(g(P_i))}_{\text{visibility constraint}} \tag{6.3}$$

The *visibility constraint* can be calculated for a combination of two persons P_i, P_j, and a group centre $O(G_k)$. To not disrupt P_i's visibility to the o-space-centre, P_j must either stand farther away than P_i or on a different side of it. The first can be ensured by comparing their distances to $O(G_k)$, and the second by considering the angle between them regarding the same $O(G_k)$. To this end, d_i^k is defined as the distance between the position of P_i and $O(G_k)$, and $\theta_{i,j}^k$ is the angle between P_i and P_j regarding $O(G_k)$. An exemplary visualization of these measurements can be seen in Figure 6.1. The relevance of occlusions can be adjusted with a constant factor K for angles between 0 and a cut-off $\hat{\theta}$. As applied in the evaluations by Setti et al. [Set+15], these are chosen as $\hat{\theta} = 0.75$ and $K = 100$ for the evaluations in this chapter. On this basis, the visibility constraint for any combination of P_i, P_j, and G_k calculates as follows:

$$R_{i,j}(G_k) = \begin{cases} 0, & \text{if } \theta_{i,j}^k > \hat{\theta} \text{ or } d_i^k < d_j^k \\ \exp(K\cos(\theta_{i,j}^k))^{\frac{d_i^k - d_j^k}{d_j^k}}, & \text{otherwise} \end{cases}$$

MDL-PRIOR Finally, Equations (6.2) and (6.3) both result in zero assignment costs for $|G_k| = 1$. Therefore, an additional term is required to penalize small groups. This is achieved by adding an *MDL-prior* over the amount of groups, which can be adapted with the parameter M:

$$P(G) = M|G| \tag{6.4}$$

An assignment of persons in a scene to conversational groups can be created by optimizing the cost function in Equation (6.1). To this end, I implement three different detectors: (1) Detector *Gco* from *fformation-gco* [ffm-gco]: uses *GCoptimization - Software for Energy Minimization with Graph Cuts Version 3.0* by Nuno Subtil et al. [gco-v3.0] to implement the approach from Setti et al. [Set+15]. For comparison, the original implementation can be found in *Graph-Cuts for F-Formation* by Francesco Setti [GCFF]. The detectors (2) *Shrink* and (3) *Grow* from *fformation* [ffm] use *k-means* to find the best assignment for a fixed number of groups. To find the best number of groups, *shrink* increases it starting from one—thereby shrinking the group size—as long as the cost decreases. *Grow* starts with one group for each person and reduces the amount of groups as long as the cost decreases. Finally, two dummy detectors are available: *None* always returns a group for each person and *One* always assigns all persons to a single group [ffm]. By applying these detectors to the created corpus, I can investigate Claim 6.1^C.

6.1.2 *Evaluation*

To evaluate Claim 6.1, I use the ground truth annotations of group assignments and person positions from both the annotations and the automatic detections. An evaluation of the F-Formation detection for HHI is already done by Setti et al. [Set+15]. Because the evaluation in this section is concerned with the quality of detected conversational groups of artificial agents, a quality metric is required that can be applied to a single agent in the scene separately. To this end, I use the definition of a tolerant match by Setti et al. [Set+15]:

DEFINITION 1 (TOLERANT MATCH) *With a threshold $T \in [0,1]$ (tolerance threshold), a predicted group G_k is a* tolerant match *if at least $T|G_k|$ participants of the group are correctly assigned and less than $1 - T|G_k|$ participants are falsely assigned to it [Set+15].*

The choice of T determines how exact the detected group must match the ground truth to be correct.

COROLLARY 1 *Tolerant matches for $T < \frac{1}{2}$ classify groups with more than 50% wrong or missing persons as correct.*

With this definition, a confusion matrix can be calculated. In contrast to Setti et al. [Set+15], the confusion matrices in this section are cal-

C Claim 6.1: Mixed conversational groups of people and artificial agents can be detected using F-Formations as known from HHI.

culated for each agent separately. The matrix is calculated as sums of observations where:

TP: the agent is correctly assigned to a group and the group matches the annotation according to the tolerant match.

FP: the agent is falsely assigned to any group.

TN: the agent is correctly classified as not in a group.

FN: the agent is falsely classified as not in a group or assigned to the wrong group according to the tolerant match.[1]

By definition, the agent is always a participant of it's own group: $P_i \in g(P_i)$. Therefore, it is always correctly assigned and the match is always greater than 0.

COROLLARY 2 *With* $T = \frac{1}{|P|+1}$ *the confusion matrix measures the detection of the agent being in a group or not, regardless the other participants of the group.*

On the basis of these measures, four detector configurations are chosen from a grid search on MDL, stride, and algorithm. For the range of $T \in [0.5, 1]$ (Corollary 1), the usual quality measurements are calculated over the whole corpus and visualized for each agent on annotated and detected person percepts. The detected percepts are the fusion results from the apartment's person tracking and Gines et al. [OpenPose] based detections. The resulting plot can be seen in Figure 6.2 and is analysed in the following.

6.1.3 *Results*

A set of observations and conclusions can be drawn from the performance visualization in Figure 6.2: (i) Group detections on the basis of automatically detected person percepts show worse performance for all configurations and tolerance thresholds in all quality metrics—except for recall in case of Flobi Entrance and a small T. The strongest performance decrease can be observed for Flobi Entrance in the precision and markedness measurements. Only 20%-40% of the groups detected for the Flobi Entrance correctly match the annotation in case of automatic person detections while 65%-80% based on the annotations. The effect is much smaller for recall, and informedness at Flobi Entrance and for all metrics in case of Flobi Assistance. It is consequential that the group detection results are worse with automatically detected person percepts than with manually annotated person positions. Especially

[1] Counting an assignment to the wrong group when the agent is in a group as false negative may seem counter-intuitive. However, false positive implies that the agent is not in a group. This case can be interpreted as: falsely not assigning to the correct group (false negative).

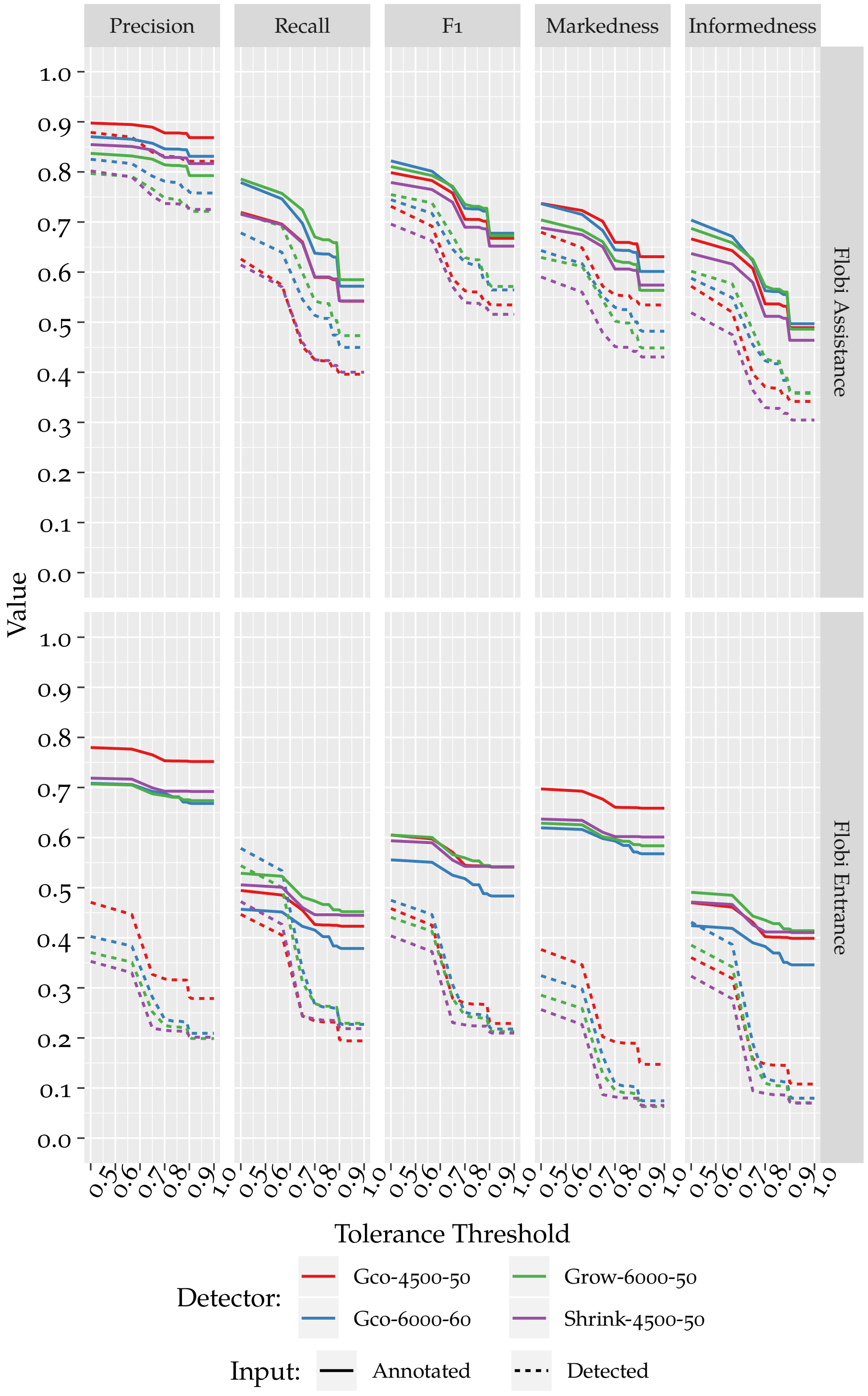


Figure 6.2: The values of performance metrics precision, recall, F1-score, markedness, and informedness, plotted over different choices of T (Threshold). The results for Flobi Assistance are shown in the upper row, Flobi Entrance can be seen in the lower row. Solid lines represent the results of group detections based on annotated person positions. Dashed lines show group detection results from automatically detected person percepts. The detector names consist of the name of the algorithm, the used MDL and the stride.

the strong, decrease in case of Flobi Entrance suggests problems in the detection of persons. Indeed, both person detection approaches are challenged by the used corpus. For the Gines et al. [OpenPose] based approach, the positions of the agents and camera perspectives (O perspectives in Figure 5.1) are problematic. The feet of people interacting with Flobi Assistance can not be detected as they stand behind the kitchen unit. This is not a problem in case of Flobi Entrance, but the corridor is crowded, which results in occlusions between the persons. The person detection of the apartment, works on the basis of top-down perspectives ($T1$ and $T2$ in Figure 5.1). Therefore, people neither can occlude each other nor be occluded by furniture. Nevertheless, this system does not perform well in crowded situations, because it often can not separate the percepts of people who are standing close to each other. Especially problematic is the hallway, which is narrow and crowded. (ii) The overall performance in case of Flobi Entrance is worse than in case of Flobi Assistance. This can have multiple reasons. First of all, the prevalence for being in a group for Flobi Entrance (0.15) is much lower than for Flobi Assistance (0.39). Therefore, with the same classifier-quality, a lower precision can be expected. This is confirmed by the fact that the classifiers' markedness—which is less prone to the prevalence—is on a similar level for both agents. For recall and informedness this is different. They are both worse for Flobi Entrance. Therefore, the classifiers are less qualified in correctly distinguishing group and non-group configurations for Flobi Entrance. (iii) While the precision decreases slightly for a growing tolerance threshold, the impact is much stronger for recall, markedness, and informedness in case of Flobi Assistance. This is rooted in the definitions of false positive and false negative for these evaluations. Correctly detecting that an agent is in a group, but assigning it to the wrong group is counted as false negative. Therefore, these cases have no impact on the classifier's precision. The markedness metric accounts for false negative and therefore shows a decline. A much smaller decline can be seen in case of Flobi Entrance. This is rooted in the overall smaller group sizes observed for the hallway. (iv) *Gco-4500-50* achieves the best precision, and markedness for both agents. This means that the classifier configuration can be chosen in an agent-agnostic way. Furthermore, it shows that the results of the *graph cuts* based approach are the most trustworthy. (v) The recall and informedness results are best in case of *Grow-6000-50*. The *k-means* based approach can correctly find more of the annotated groups and non-groups than the other configurations when using a higher MDL. This underlines the trade-off that has to be made between precision and recall. (vi) Finally, the plots of precision and recall show only small differences to markedness and informedness. This means that the probabilities of false omissions (FOR) and false alarms (FPR) are small. The only exception here is precision and markedness for Flobi Assistance witch indicates $\approx 20\%$ false omissions. The results can not be directly compared to the performance of F-Formation detections in HHI [Set+15; Vas+16] because the quality measures in this chapter had to be calculated for each agent separately. Nevertheless, the results show that by detecting F-Formations with the approach presented by Setti et al. [Set+15], conversational groups with artificial agents can be detected.

6.1.4 *Discussion*

In this section, I Investigated Claim 6.1[C]. To this end, I implemented a system for F-Formation-detection according to Setti et al. [Set+15] and evaluated its applicability for the detection of conversational groups with artificial agents using the corpus presented in Chapter 5. The evaluation shows that the quality of the conversational group detection depends on and the crowdedness of the environment and the participants incentive to interact with the agents. Errors in the detection of participants can strongly interfere with the quality of the group detections. Furthermore, it gets increasingly hard to assign each participant to the correct group while the group size grows. Nevertheless, it is evidently possible to utilize the approach to correctly find the participants of the conversational groups of artificial agents in the majority of the observations, which confirms Claim 6.1.

6.2 IN/OUT OF GROUP DISTINCTION

As suggested in the beginning of this chapter, it is not always necessary for an artificial agent to correctly identify all participants of its conversational group to be able to solve its task in a socially acceptable manner. For example, to exhibit civil inattention it is fully sufficient to know whether the robot is in a conversational group or not. Therefore, in such situations a much simpler classifier may be sufficient. In this section, I investigate whether the detection of faces in the agents field of view (Claim 6.3[C]), or mutual gaze (Claim 6.2[C]), can sufficiently inform about whether the agent is in a conversational group or not (*in-group*-detection). To this end, I compare the applicability of face detections, and gaze detections with the results of the conversational group detection based on F-Formation from Section 6.1.

[C] Claim 6.1: Mixed conversational groups of people and artificial agents can be detected using F-Formations as known from HHI.

[C] Claim 6.3: The detection of faces in the agents field of view can sufficiently inform about whether it is in a conversational group or not.

[C] Claim 6.2: The detection of peoples gaze direction in the agents field of view can sufficiently inform about whether it is in a conversational group or not.

6.2.1 *Detectors*

To better understand, if face detection, gaze detection, and F-Formation detection can be applied to distinct situations where an agent is in a group from situations where it is not, I create a scalar feature from each of these inputs. This is done as follows:

FACE: To get a scalar feature for the face detection I calculate the amount of the agents field of view that is occupied by the face of the nearest person. As an approximation, the size of the ROI from the face detection can be used. If multiple faces are detected simultaneously, the biggest face (ROI) is chosen as the nearest and therefore most informative. The resulting feature is zero when no face can be detected and grows when someone gets closer to the agent.

GAZE: The gaze feature is calculated from the horizontal and vertical gaze angle of the face of the nearest person. To this end, the angle of the combined rotation (yaw and pitch) is calculated and used as a scalar feature vector. It is zero when the agent is directly looked at and π when the person looks in the opposite direction. Because the gaze detection model is based on frontal views of faces, this angle does not exceed $\frac{\pi}{3}$ in this evaluation. In case no face and therefore no gaze direction can be detected, the angle is set to $\frac{\pi}{2}$ as an upper bound.

GCO-AGENT: From the F-Formation-detection models as presented in Section 6.1, I use the *graph-cuts* based detection with $\text{MDL} = 4500$ and $\text{stride} = 50$ (*Gco-4500-50*). This configuration has an overall good performance in the F-Formation detection. The automatically detected person percepts are used as input information to allow a fully automatic recognition. To get a scalar feature, I calculate the distance and visibility cost (Equations (6.2) and (6.3)) for the agent. Cases where the agent is detected as not in a group ($|G_k| = 1$) are set to the maximum observed cost. The resulting feature approaches zero for group assignments with low costs for the agent, and grows when it becomes difficult to access a group.

6.2.2 *Evaluation*

To evaluate the applicability of face information for *in-group* detection, the face detection observations need to be linked to the group annotations. To this end, the face detection data—which is produced with 15 Hz—is assumed to be constant between observations and sampled using the timestamps from the group annotations. This results in 50834 observations for each agent and feature. By varying a threshold between the smallest and largest observed value of each feature, ROC and precision-recall curves can be created. It should be possible to create a classifier that combines these features to achieve an overall better performance. However, the performance of such a classifier is not relevant for the investigated claims (Claims 6.2 and 6.3) and therefore not inspected. The performance of the three features is visualized in Figures 6.3 and 6.4 and further analysed in the following.

6.2.3 *Results*

The visualization of the ROC curves and AUC of the proposed detectors *Face*, *Gaze*, and *Gco-Agent* in Figure 6.3 allows some insights into their applicability for *in-group* detection. (i) The recall is strongly increasing for small values of FPR and there are no observations of FPR in the range [0.09, 0.97]. This strong increase in the FPR happens for the face detection based features when observations without a face detection are accepted as *in-group* and for the F-Formation based feature when no group was detected. At this point all observations are accepted as *in-group* and the detectors lack diagnostic power. (ii) The F-Formation based feature (*Gco-Agent*) produces the lower bound in recall and the upper bound of *False Alarms* for Flobi Entrance. Although a perfect F-Formation based group detection would result in an optimal ROC curve, this is the worst performing classifier. The lower quality of this feature for Flobi Entrance, therefore, must be rooted in (1) the overall worse performance of F-Formation detection for this agent and (2) the

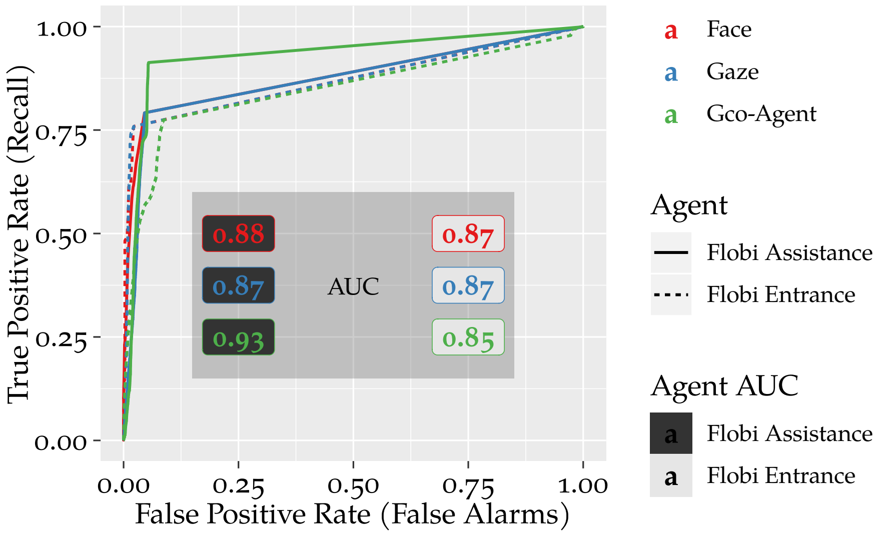


Figure 6.3: Performance visualizations of *in-group* detection from the *Face* (red), *Gaze* (blue), and *Gco-Agent* (green) models in the ROC space. The corresponding AUC values are shown in the gray box on the lower centre. The results are shown separately for the agents Flobi Assistance (solid lines, dark-filled AUC), and Flobi Entrance (dashed lines, light filled AUC).

higher noise in the person detection in the vicinity of this agent. (iii) For Flobi Assistance, this feature achieves the upper bound in recall while showing a slightly worse FPR. It furthermore achieves the highest AUC. This shows that the feature has a much higher potential in correctly deciding whether an agent is in a group or not. (iv) Finally, the face detection based detectors result in similar curves for each agent and feature. While they achieve a slightly better recall for Flobi Assistance, their FPR is lower for Flobi Entrance. Nevertheless, the overall difference is negligible, as can be seen in their AUC. This property indicates that information about the size of the detected face or gaze direction—at least in this scenario—does not provide much information to *in-group* detection. The detectability of a face as such is much more important.

Because the proportion of observations in which the agent is in a group are low in the used data (0.39 for Flobi Assistance and 0.15 for Flobi Entrance), the FPR value is an optimistic measure. Therefore, it is interesting to additionally investigate the applicability of the features in the precision-recall space. The curves can be seen in Figure 6.4 and give further insights. (i) The breakdown in diagnostic power of the features when all observations are accepted as *in-group* can be seen in this visualization too (for high values of recall). (ii) The differences of the features' applicability is more apparent in this visualization. It can be seen in the curves and the AUC values of the classifiers. (iii) The *Gco-Agent* based classifier for Flobi Entrance is gradually loosing precision when recall

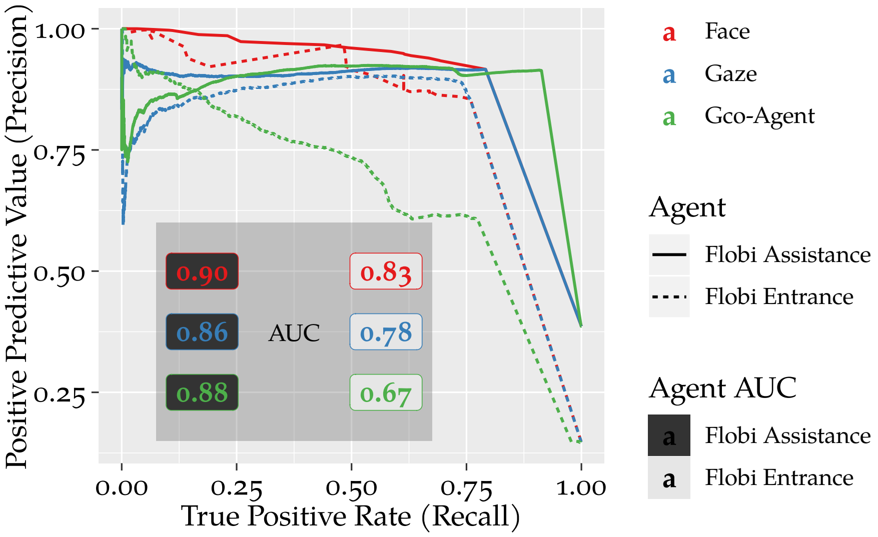


Figure 6.4: Performance visualizations of *in-group* detection from the *Face* (red), *Gaze* (blue), and *Gco-Agent* (green) models in the precision-recall space. The corresponding AUC values are shown in a gray box on the lower left. The results are shown separately for the agents Flobi Assistance (solid lines, dark-filled AUC), and Flobi Entrance (dashed lines, light filled AUC).

is increased and shows an overall low AUC of ≈ 0.67. (iv) This does not apply for Flobi Assistance, for which the feature produces a continuously high precision of $\approx 90\%$ even for high recall values. (v) *Gco-Agent* for Flobi Assistance and *Gaze* for Flobi Entrance show a high variance in precision for low recall values. While configurations with low recall are not of great interest, this property lowers the overall AUC of these detectors. (vi) The best AUC is achieved by the *Face* feature for Flobi Assistance which shows a reliably high precision for recall $< 80\%$.

6.2.4 *Discussion*

In this section, I created scalar features from the results of a face detection and a gaze detection approach. I used these features to test their applicability as *in-group* detectors and compared them to a feature based on an F-Formation group assignment cost to investigate Claim 6.2[C] and Claim 6.3[C]. On the one hand, this investigation shows that a feature based on the detection of conversational groups can outperform approaches that only utilize informations from the detection of faces in the agents field of view. This is especially true when a high *recall* is required. On the other hand, this only works if the conversational group detection itself produces reliable results. If the reliability of the group detection approach is low, an approach based on the detection of gazes produces better results. This feature achieves a recall of $\approx 75\% - 80\%$ with a precision of $\approx 90\%$ and produces overall better results for Flobi Entrance than the group detection based approach. Therefore, it can be said that the detection of peoples gaze can inform about whether the agent is in a group or not sufficiently to outperform a group detection based approach. This confirms Claim 6.2. Although, this is only the case when the reliability of the group detection is low—as with Flobi Entrance. The feature based on the size of the face produces comparable results. This confirms Claim 6.3. Furthermore, it is apparent that both *face-detection* based features show only small changes in the investigated quality measures for varying thresholds as long as observations without a detected face can be distinguished from observations with a face. It can be concluded from this observation that the most important information for the *in-group* detection is not the gaze angle or face size but whether a face can be detected in the first place or not.

6.3 summary

By detecting arrangements of verbally interacting people as conversational groups, an agent enriches its understanding of the social situation.

[C] Claim 6.2: The detection of peoples gaze direction in the agents field of view can sufficiently inform about whether it is in a conversational group or not.

[C] Claim 6.3: The detection of faces in the agents field of view can sufficiently inform about whether it is in a conversational group or not.

This allows it to distinguish focused from unfocused interactions, and behave more socially adequate (see Sections 2.1.2 and 2.1.3). Therefore, I investigated RQ 3[C] in this chapter. To this end, I showed that a conversational group with an artificial agent can be detected using F-Formations as known from HHI (Claim 6.1). Furthermore, I examined whether simpler approaches can be used as a substitute for the group detection in cases where the participants of the group are of no interest. To this end, I showed that the detection of peoples gaze direction in the agents field of view is sufficient for *in-group* detection (Claim 6.2), and that this can be further simplified by using face detection results (Claim 6.3).

Considering RQ 3 the collected observations show that conversational groups with artificial agents can be detected as F-Formations using the same approach as known from HHI research. When the agent only needs to know whether it is in a conversational group, and an identification of the groups participants is not required, the F-Formation based group detection can achieve better results than approaches based on face and gaze detection. However, this investigation shows that noise in the person detection has a strong negative impact on this group detection. In case of noisy person detection, better *in-group* detection results can be achieved by using the result of a face detector instead of detecting person positions and calculating assignment costs.

[C] RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?

7 CONVERSATIONAL ROLE RECOGNITION

With conversational group detection, an agent can behave towards people depending on whether they are part of its conversational group or not. It can show civil inattention towards some while optimizing the mutual observability with others. However, to correctly interact with the people within the group, further information is required. The agent needs to guide its attention towards the speaker and identify utterances addressed towards it. Simultaneously, it should support the interaction by acting according to its role, or may apply turn taking to influence the roles within the group. However, to efficiently use them, the agent needs to know its current role (Sections 2.2.2.2 and 2.2.2.3). Therefore, I stated RQ 4[C]. To create an approach to conversational role recognition, I elaborate on the following arguments and corresponding claims: On the one hand, conversational roles can often be directly inferred from the state of the scene. Someone who speaks, for example, automatically becomes the accepted speaker or stops speaking. A *Non-Participant*, by definition, lacks a conversational group. Therefore, I claim:

CLAIM 7.1 (HIGH-LEVEL FEATURES) *Given a set of high-level features, the conversational role of an agent can be recognized using simple models.*

On the other hand, the signals used to negotiate conversational roles in human interaction can often be minimal and hard to determine:

CLAIM 7.2 (LOW-LEVEL FEATURES) *By learning from lower level features, the recognition of conversational roles can be further enhanced.*

Furthermore, a conversation is a dynamic process that unfolds in time and in which the roles are negotiated through the system of turn taking, one turn at a time. Therefore, the problem should be treated as a time-dependent recognition problem:

CLAIM 7.3 (TIME-BASED APPROACH) *By observing how the interaction unfolds in time, the conversational role can be better recognized than from the latest observation alone.*

To investigate RQ 4 and the presented claims, I create and evaluate conversational role recognition models for virtual agents in a smart environment. To this end, I use high-level features and simple models on the one hand (Claim 7.1), and artificial neural networks with varying feature sets (Claim 7.2) and time sequence information (Claim 7.3) on the other hand. Like in Chapter 6, this investigation is performed

[C] RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

using the corpus presented in Chapter 5. It contains 57 min of mixed human-agent interaction with the virtual agents Flobi Entrance and Flobi Assistance. Observations are sampled with a regular rate of 15 Hz for each agent separately. This results in a total of 101668 observations, uniformly distributed over the time and agents. The observations are spread between the roles as follows: *Speaker* (3.35%), *Addressee* (5.60%), *Side-Participant* (18.06%), and *Non-Participant* (73.00%). This imbalance between the classes needs to be considered during evaluation.

7.1 HIGH-LEVEL ROLE FEATURES

For the evaluation of Claim 7.1[C], I create a set of high-level features based on the observations made in the previous chapters. These features are then combined using a set of rules and, for comparison, a set of Bayesian Networks to create simple classifiers for the recognition of conversational roles.

7.1.1 *Feature Selection*

To allow the classification of conversational roles, a set of high-level features can be compiled. Based on the observations made in the previous chapters, the following features can be automatically detected or are directly accessible to the agent during the interaction:

Speaking: Whenever an agent is speaking, its conversational role must be speaker or will become speaker over time (Section 2.1.3.3). The agent's *speech-production activity* (binary) is therefore an important information in the recognition of the speaker role. As this is part of the agent's inner state, it does not need to be detected. The information can be extracted directly from the corpus (see Section 5.5).

Addressed: Whether an agent has the role of addressee is determined by the current speaker (Section 2.1.3.3). In Chapter 4, I present and evaluate a model to decide whether the agent is addressed. From this evaluation, I select a model for addressee recognition that uses information about the interlocutors *mouth movements* and *mutual gaze* with the agent.[1] I train this model using the full corpus that is presented in Section 4.2. By applying this model to observations in this evaluation, a binary prediction can be made whether the agent is *addressed* ($P(A|G, M) > 50\%$) or not.

In-group: One reason for the importance of conversational group detection for conversational role recognition is that it subdivides

[1] the Bayesian Network named *Both* in Figure 4.7 on page 77

[C] Claim 7.1: Given a set of high-level features, the conversational role of an agent can be recognized using simple models.

the set possible roles (Section 2.1.3.2). Whenever the agent is in a group, it can not assume the role of *Non-Participant*. In Chapter 6, I present an approach to F-Formation detection that can be applied here too. As in Section 6.2.2, I use the *graph-cuts* based optimization with $M = 4500$ and $S = 50$ in this investigation. By applying the F-Formation detection to an observation, a binary *in-group* detection ($|g(P_{agent})| > 1$) can be created for the agent.

Mutual gaze: Gaze is an important information in conversations. The distribution of the participants gaze in a conversational group can be used as an indicator of the distribution of conversational roles. Furthermore, gaze is used in the negotiation of the next turn (Section 2.1.3.3). In Chapter 4, I present a model for a binary *mutual gaze* detection which is based on the (continuous) angle of the interlocutors *gaze*.

Mouth movements: Like the *speaking* feature for the agent, the detection of *mouth movements* can identify its interlocutor as speaker. When a different participant of the conversation is the current speaker, the agent can not assume this role simultaneously. Furthermore, whether the interlocutor is speaking or not has implications for the interpretation of *mutual gaze* (Section 2.1.3.3). In Chapter 4, I present a model for (binary) *mouth movement* detection based on the (continuous) variance of the distances between the interlocutors upper and lower lip (*lip variance*) (Section 4.2.3).

This set of high level, binary classification results are used in the following to create simple models for conversational role recognition.

7.1.2 *Rule Based Model*

For the evaluation of Claim 7.1, I create a simple, rule based conversational role recognition model. It performs the following decisions:

In-Group: Whenever the agent is not in a conversational group, its role must be *Non-Participant*. Therefore, the agent is *Non-Participant* when $|g(P_a)| = 1$. When it is part of a conversational group ($|g(P_a)| > 1$), it can assume one of the remaining conversational roles.

Speaking: The literature suggests, that only one participant can have the role of speaker at a time during conversations. This implies that, if one agent does not *have the floor* but starts and continues to speak, the current speaker will eventually yield and the agent become speaker (see Section 2.1.3.3 on page 18). Therefore, the agent is classified as speaker, whenever it is part of a conversational group and speaking. This information is drawn directly from the agent's *speech-production activity*.

Addressed: When the agent is in a conversational group but not the speaker, it can assume the role *Addressee* or *Side-Participant*. To distinguish these two roles, the addressee recognition model as presented in Section 4.5 is used. The model is trained using the full corpus presented in Section 4.2.5 and predicts the probability of the agent being *addressed* given *mutual gaze* and the interlocutors *mouth movements*. As this is a binary decision, the agent is assumed to be *Addressee* when the probability $P(A|G, M)$ is higher than 50%. Finally, when the agent is neither speaker nor *Addressee* but still a participant of the conversational group, it must have the role of a *Side-Participant*.

By applying these three binary decisions in succession, a rule-based addressee classifier can be created. A visualization of the resulting model can be seen in Figure 7.1. This simple model can be used to assess the applicability of approaches to conversational role recognition based on high level features (Claim 7.1).

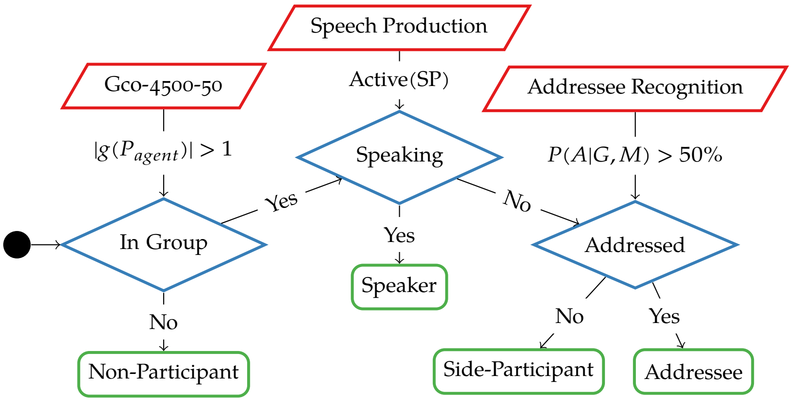


Figure 7.1: The decision tree for rule based role recognition. The root node is shown as a black circle, decision nodes as blue diamonds, input nodes as red parallelograms, and role recognition results as green rectangles. Decision nodes perform a binary (yes/no) distinction based on data provided by the input nodes.

7.1.3 *Bayesian Network*

For comparison, I create Bayesian Networks that incorporate the high level features of the rule model. To this end, the agent's role is assumed to be dependent on these features. Furthermore, this model uses *mutual gaze* and *mouth movement* information directly (in contrast to the rule based model which uses the results of the addressee recognition) to allow a better adaptation to the addressing behaviour in this corpus. The resulting model (*BnM*) can be seen in Figure 7.2. Additionally, I apply structure learning[2] to automatically extract Bayesian Network

[2] Using `bnlearn::hc` from the `bnlearn` package (v4.4) in R[bnlearn] with 1000 restarts and 1000 perturbations.

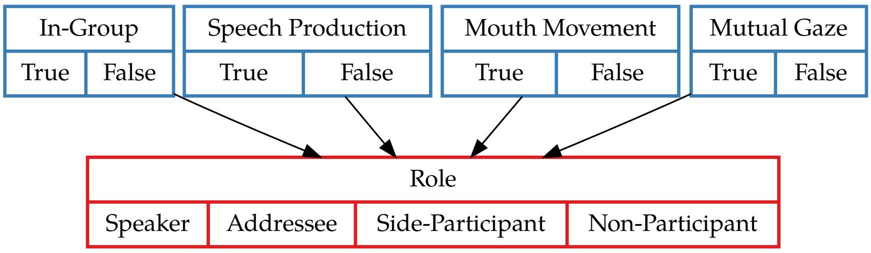


Figure 7.2: Manually created Bayesian Network structure (*BnM*) that uses the high level features *in-group*, *speech production*, *mouth movement*, and *mutual-gaze* for conversational role recognition. Blue rectangles represent binary input nodes. The red rectangle represents the conversational role of the agent with the four possible outcomes *Speaker*, *Addressee*, *Side Participant*, and *Non-Participant*. The arrows represent conditional dependency—the role depends on all other nodes.

structures. In the following, models created via structure learning are called *BnA*. Furthermore, as Bayesian Networks are sensitive to class-imbalance, I additionally perform under-sampling in the training of these models. To prevent ties in the Bayesian Network results, the under-sampling uses slightly more examples for more common roles. This is done by taking all n observations of the rarest role and randomly sampling $n + 1$ of the second rarest, $n + 2$ of the second most common, and $n+3$ of the most common role observations from within the training data. Models trained with under-sampling are subscripted with 'u'. The resulting models BnM_u and BnA_u are trained with *nearly equal* amounts of observations of all roles. In Table 7.1 a classification of the resulting models is shown.

	Full Training Set	Under-sampling
Manual Structure	*BnM*	BnM_u
Automatic Structure	*BnA*	BnA_u

Table 7.1: The configurations of Bayesian Networks as evaluated in this chapter. Manual networks use the structure shown in Figure 7.2. The structure of automatic networks is learnt from the training data. The 'u' subscript highlights models that are trained with under-sampling.

In the following section, I evaluate the classification performance of the rule-based classifier and compare it to the results of the Bayesian Network based approaches.

7.1.4 *Evaluation*

For the evaluation, I split the corpus into 5 min long slices on a regular scale. This results in 12 subsets which can be used to perform a 12-fold CV, by holding one subset out for validation and training with the remaining 11. Structure learning and under-sampling are performed on the respective training-dataset within the fold. The rule based model (*Rule*) does not need to be trained. In total, five models (*Rule*, *BnM*, BnM_u, *BnA*, and BnA_u), are evaluated in a 12-fold cross-validation. To get an impression of the models' performances for each role, I first visualize the F1-score, markedness, and informedness for each class separately. The resulting plot can be seen in Figure 7.3. Furthermore, to allow the comparison of the overall quality of the models, I calculate two further performance measures. The accuracy of the models can be calculated as the fraction of correct classifications in the population. However, because all observations are considered individually, accuracy does not account for the class imbalance. Therefore, I additionally calculate the macro average of the F1-score over the possible roles:

$$F1_\mu = \frac{\sum_{r\in\text{Roles}} F1_r}{|\text{Roles}|}$$

This way, each role is considered equally important for the model's performance, regardless of the prevalence. The resulting plots of $F1_\mu$ and accuracy for the presented models can be seen in Figure 7.4. These visualizations are used to investigate and discuss the performance of conversational role recognition with high-level features and simple models.

7.1.5 *Results*

Looking at the classification performances of the models for each class (see Figure 7.3) the following observations can be made:

F1: The *Rule*, *BnM*, and BnM_u models achieve similar F1-scores for *Speaker* and *Addressee*, slightly better scores for *Side-Participant* and good results for *Non-Participant*. In case of *Side-Participant*, *BnM* beats the other models by more than 0.15. The Bayesian Network with automatically deduced structure (*BnA*) can not compete for *Speaker* ($\Delta > 0.1$), and fails to classify *Side-Participant* and *Addressee* (both < 0.2). When trained with under-sampling, the same approach (BnA_u) produces a competitive F1-scores for *Speaker*, *Addressee*, and *Non-Participant* but entirely ignores *Side-Participant* (no orange bars for *P* because always rejecting results in division by zero).

Markedness: The *Rule* based model achieves a high markedness for *Speaker* and *Non-Participant*, while the values for *Addressee* and *Side-Participant* are much lower (half as good). *BnM* behaves similar but can achieve better results for *Side-Participant* ($\Delta > 0.05$). BnM_u has a much lower markedness for *Speaker* than the other models but behaves otherwise similar to *Role*. The *BnA* model achieves a high markedness of > 0.9 for *Speaker*, performs similar to the other models for *Addressee* and *Side-Participant*, and worse

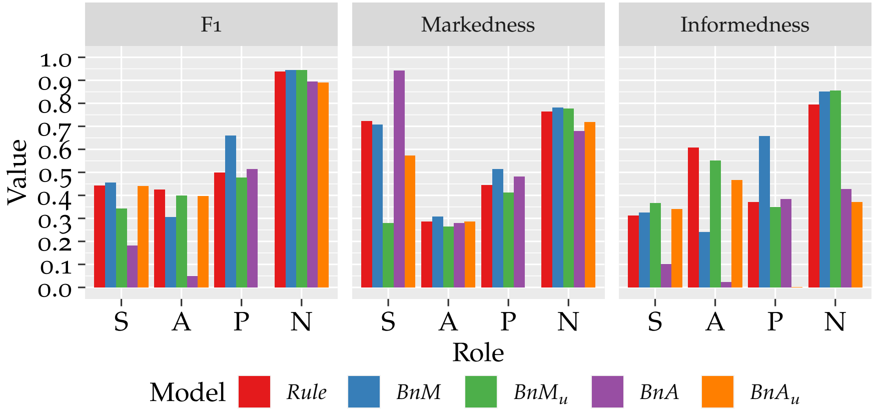


Figure 7.3: Performance measures F1-score, markedness, and informedness for the rule based model and the Bayesian Network based models calculated for each role separately. The roles are (*S*)*peaker*, (*A*)*ddressee*, *Side*-(*P*)*articipant*, and (*N*)*on-Participant*. The results of the different models—from left to right—are coloured in red (*Rule*), blue (*BnM*), green (BnM_u), violet (*BnA*), and orange (BnA_u). The results of BnA_u for *P* can not be calculated as it does not predict *Side-Participants* (division by zero).

than the other models for *Non-Participant*. BnA_u performs similar to the other models for *Addressee* but is slightly worse for *Speaker* and *Non-Participant*. *Side-Participant* is never predicted by this model.

Informedness: The informedness for *Speaker* of all models except BnM_u is much worse than the markedness. BnM_u has a better informedness for *Speaker* and *Addressee* than markedness. While *Rule* and BnM_u achieve a much higher informedness for *Addressee* than *BnM* ($\Delta > 0.25$), the difference is inverse for *Side-Participant*. The informedness for *Non-Participant* is high for *Role*, *BnM*, and BnM_u but low in comparison to the others for *BnA* and BnA_u. Finally, BnA_u shows an overall low informedness.

To get a better impression of the models overall performances, Figure 7.4 can be consulted. The following observations can be drawn from the visualizations of accuracy and $F1_\mu$:

Accuracy: The accuracy of the evaluated models ranges between 0.76 (*BnA*) and 0.84 (*BnM*). While *BnM* achieves the highest accuracy, the *Rule* model (0.81) is second best. The models *Rule*, BnM_u, and *BnA* are in the mean third of the range. BnA_u performs worst with an accuracy of 0.76 which is only slightly higher than the prevalence of the *Non-Participant* role.

$F1_\mu$: The $F1_\mu$ measures show a similar but less encouraging image. *BnM*, *Rule*, and BnM_u perform best and in the same order as for

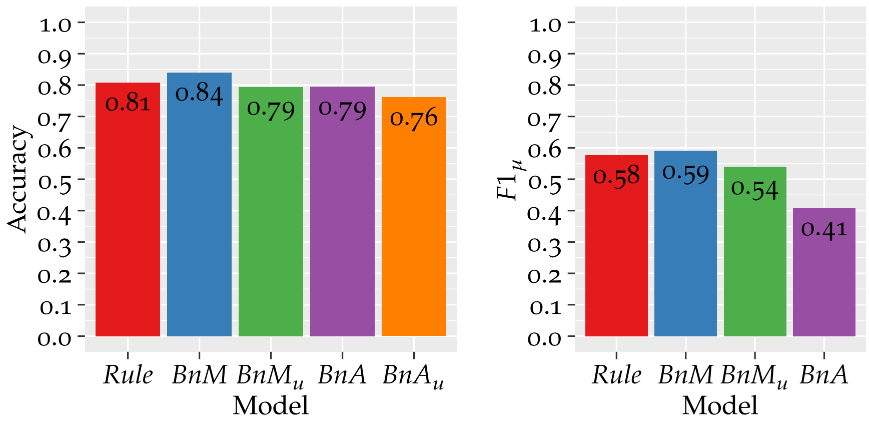


Figure 7.4: The overall accuracy and $F1_\mu$ of the evaluated models. The bars represent different models with the same colour coding as in Figure 7.3. While accuracy is shown in the left plot, the models $F1_\mu$ can be seen in the right plot. $F1_\mu$ for BnA_u can not be calculated as it does not predict *Side-Participants* (division by zero).

accuracy. They achieve an $F1_\mu$ between 0.54 and 0.59. The *BnA* model achieves a low mean F1-score of 0.41. Because $F1_\mu$ requires the model to be able to predict all roles and BnA_u does not predict *Side-Participant*, there is no result for BnA_u.

7.1.6 *Discussion*

The best results in this evaluation are achieved by the Bayesian Network based model *BnM*. Under-sampling (BnM_u) does not improve it's F1-score or markedness but it's informedness for *Speaker* and *Addressee*. These observations are plausible considering the structure of the model (Figure 7.2). With all features being parent-nodes of the *Role* node, the model can learn the probability distribution of roles for each combination of the feature values separately. This is possible because of the small amount of binary features (the resulting conditional probability table for *Role* has 64 cells). Given enough examples, and based on these features, this model always performs optimal from a statistical point of view. Additionally, it is not affected by an imbalance in the observations it learns from. This is confirmed by the lower performance of BnM_u which has the same structure but trains with fewer observations.

In contrast to the *BnM* structure which is optimized for *Role* prediction, the structure learning optimizes for a global model of the data. Therefore, the *BnA* models can not compete with the other models in this set-up. This approach is better suited for bigger feature sets—which makes it impractical to set *Role* a child-node of all features—or when the prediction of varying nodes in the network is required from incomplete observations.

With an accuracy of 0.81 and an averaged F1-score of 0.58, the *Rule* based conversational role classification approach achieves the second best results. The *in group* feature from the F-Formation detection allows predicting the *Non-Participant* role with a high reliability. As the F1-score for the remaining roles is between 0.4 and 0.5, it is evident that they are less easy to distinguish based on the provided features. The much higher markedness than informedness for *Speaker* is rooted in the *speech production activity* feature not accounting for situations in which the agent holds *the floor* but does not speak or speaks without holding the floor. For *Addressee* the difference between markedness and informedness is inverse. This imbalance means that the role of *Addressee* is found correctly in the data but with a high amount of false positives.

While the reliability of the *Rule* based conversational role classification differs with the role that needs to be detected, its overall results are good for a multi-class problem. Given the high-level features used in this evaluation, an optimized, statistical model achieves slightly better results. Therefore, it can be concluded that simple models can be used to predict the conversational role of an artificial agent when high-level information about the situation is available (Claim 7.1[C]).

7.2 low-level & time-based features

For the investigation of Claim 7.2[C] and Claim 7.3[C], a classifier is required that predicts the conversational role of the agent from lower-level features and based on sequences of observations. Artificial neural networks have the potential to extract the relevant information from low-level features. Therefore, I utilize artificial neural networks with simple, fully connected layers to test Claim 7.2. Furthermore, they can be extended to learn models on sequences of observations. To test Claim 7.3, I create models with layers of Long Short-Term Memory (LSTM) units, which are specifically designed to process time-series [HS97].

7.2.1 *Feature Selection*

For the investigation of Claim 7.1, a set of high-level features is used. In the following, I expand on the set of possible inputs and create different feature vectors for the evaluation of Claim 7.2 and Claim 7.3.

rule: The *rule* feature vector contains the four binary features used in the *Rule* and Bayesian Network based models in the evaluation

[C] Claim 7.1: Given a set of high-level features, the conversational role of an agent can be recognized using simple models.

[C] Claim 7.2: By learning from lower level features, the recognition of conversational roles can be further enhanced.

[C] Claim 7.3: By observing how the interaction unfolds in time, the conversational role can be better recognized than from the latest observation alone.

of Claim 7.1 (*in group*, *speech production activity*, *mouth movement*, and *mutual gaze*).

$rule_{raw}$: This feature vector contains the continuous features that are used in the calculation of the *rule* features as presented in Section 7.1.1. The *speech production activity* is used *as is*. For *mouth movement* the continuous feature *lip variance*, and for *mutual gaze* the continuous feature *gaze angle* are used. Instead of *in group*, the agents F-Formation *participation costs* C_p are used. These are calculated as presented in the cost function of the F-Formation detection (Section 6.1.1) but only regarding the agent P_a's distance and visibility costs without the MDL prior.

$$C_p(P_a) = \overbrace{(u_{g(P_a)} - x_{\mu_a})^2 + (v_{g(P_a)} - y_{\mu_a})^2}^{\text{distance cost}} + \underbrace{\sum_{j \in P, j \neq a} R_{a,j}(g(P_a))}_{\text{visibility cost}}$$

This results in a 4D vector of continuous features.

$full$: For this feature vector, I use lower-level information from the system data, face detection, and group detection. The list of information used in the feature vector with descriptions and dimensionality can be seen in Table 7.2. I do not use the time-integrated features *lip variance* and *mouth movements*, and the other deduced features *addressed*, *in-group*, and *mutual gaze*. Hence, the time-integration feature abstraction has to be done in the model. The resulting feature vector has 148 dimensions.

All these features can be automatically detected during an interaction and are used in the following to train and evaluate different artificial neural networks. To support the training, the feature vectors *rule* (4D), $rule_{raw}$ (4D), and $full$ (148D) are centred and scaled. This is done for each dimension separately by first subtracting the mean and then dividing the results by the standard deviation[3] (in the following, if not stated explicitly, a preceding centring is always implied when scaling is performed).

7.2.2 *Neural Network Models*

For this evaluation, I use the *kerasR: R Interface to the Keras Deep Learning Library* by Andrie de Vries et al. [kerasR] (`v0.6.1`) with *Keras: The Python Deep Learning library* by François Chollet et al. [Keras] (`v2.2.4`) and *TensorFlow: A System for Large-Scale Machine Learning* by Yong Tang et al. [TensorFlow] (`v1.14`) as back-end. I create two types of models, one that uses only the most recent observation to recognize the agents

[3] Performed with the `scale` function from the `base` package (v3.5.1) in R [base].

Description	Dim	Source
For which agent the conversational role recognition is performed.	1D	System Data
Whether the agent is currently speaking or not.	1D	System Data
The number of faces detected in the agent's FOV.	1D	Face Detection
The size of the agent's interlocutors face.	1D	Face Detection
Facial landmark positions of the agent's interlocutor.	136D	Face Detection
The gaze angle of the agents interlocutor.	1D	Gaze Detection
The number of participants in the agents conversational group.	1D	Group Detection
The position of the o-space centre of the agents conversational group.	2D	Group Detection
Assignment and visibility costs for the agent and its conversational group.	4D	Group Detection

Table 7.2: The information that is encoded in the $full$ feature set for conversational role recognition with artificial neural networks. Each row describes a portion of the feature vector with dimensionality (Dim) and source. For the final feature all portions are stacked into a 148D vector.

conversational role (*dense*) and one that trains the recognition on time-series (*lstm*).

dense: The model that classifies only on the basis of the most recent observation uses *fully connected* layers to model the input features and a final *dense* output layer with 4 units and *soft-max* activation. It's results are interpreted as a probability distribution over the possible conversational roles and the most probable role is chosen as the result.

lstm: The time-based model uses layers of LSTM units [HS97] to learn a temporal representation of the input features. It is trained with observation sequences of length $n = 15$, which encode the most recent 1 sec. Given that the negotiation of conversational roles is performed in close collaboration this is a good trade-off between model size and history information available to the model. The LSTM layers are followed by a *time distributed*, *dense* (fully connected) output layer with 4 units and a *soft-max* activation. This results in a sequence of 15 four-dimensional vectors interpreted as role probabilities over the last 1 sec. The first 14 encode the progress (history) of the agent's conversational role and the final vector identifies the current role of the agent. Further processing of the history will not improve the results as this is the task of the LSTM layers. Therefore, for the evaluation I use the final vector (current role).

To prevent over-fitting, the non-output layers of the models are followed by a *dropout* layer with $p = 0.5$.[4] The amount of units $u \in \{128, 512\}$ in the inner layers of the models and the number of layers $n \in \{1, 4\}$ are varied during the evaluation. A visualization of the model structure can be seen in Figure 7.5. With the three variations of input features (*role*, $role_{raw}$, and $full$) and the parametrization of the two model types (*lstm* and *dense* with $u \in \{128, 512\}, n \in \{1, 4\}$), there are 24 artificial neural network models to evaluate. The model-names encode the following information separated by dots: the type of the model ($D = dense, L = lstm$), the feature set ($R = rule, W = rule_{raw}, F = full$), the number of units in each hidden layer $\{128, 512\}$, and the number of layers $\{1, 4\}$. Therefore, *L.F.128.4* represents an *lstm* typed model, trained using the *full* feature vector with 128 units and 4 layers.

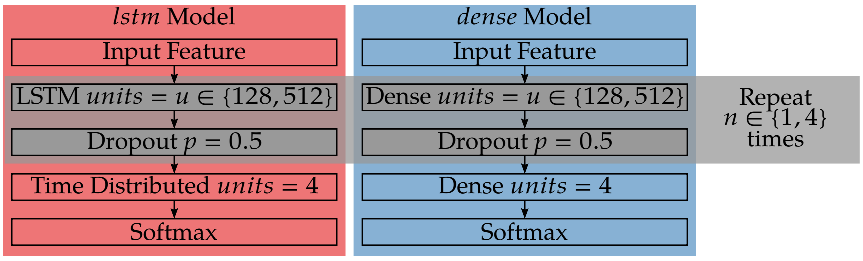


Figure 7.5: The artificial neural network models *lstm* (left, red) and *dense* (right, blue) as created for this investigation. They consist of $n \in \{1, 4\}$ layers (highlighted in grey) with $u \in \{128, 512\}$ units, followed by a *dense* layer with *soft-max* activation. While the *dense* model considers only the most recent information, the *lstm* model is trained from observation sequences and has a *time distributed*, *dense* output-layer.

7.2.3 *Evaluation*

To investigate the performance of the presented models, I perform a 12-fold cross-validation with folds of 5 min length. An equal scaling (see Section 7.2.1) of the feature vectors in the training and test sets is ensured by determining the scaling parameters (mean and standard deviation in each input dimension) from the training set and considering them a part of the model. Thereafter, the test set is scaled using the same parameters. Furthermore, as the models are randomly initialized and prone to find different local minima, I create multiple instances (8 *seeds*) of each model configuration. Thus, a confidence interval can be calculated for

[4] For for an in depth motivation and analysis of *dropout* see Srivastava et al. [Sri+14].

the models' performances. As in Section 7.1.4, I investigate the accuracy, $F1_\mu$, and class-wise F1-score of the presented models.

7.2.4 *Results*

In the following, I present and discuss the results of the *dense* and *lstm* model configured with one hidden layer and 128 units in each layer after 50 epochs of training. Increasing the number of layers, units, or epochs has only a small impact on the results (this can be seen in Figures B.1 to B.3 in the appendix).

The accuracy and $F1_\mu$ of the models, in combination with the results of the *Rule* and *BnM* models, can be seen in Figure 7.6. The following observations can be made on the basis of this visualization:

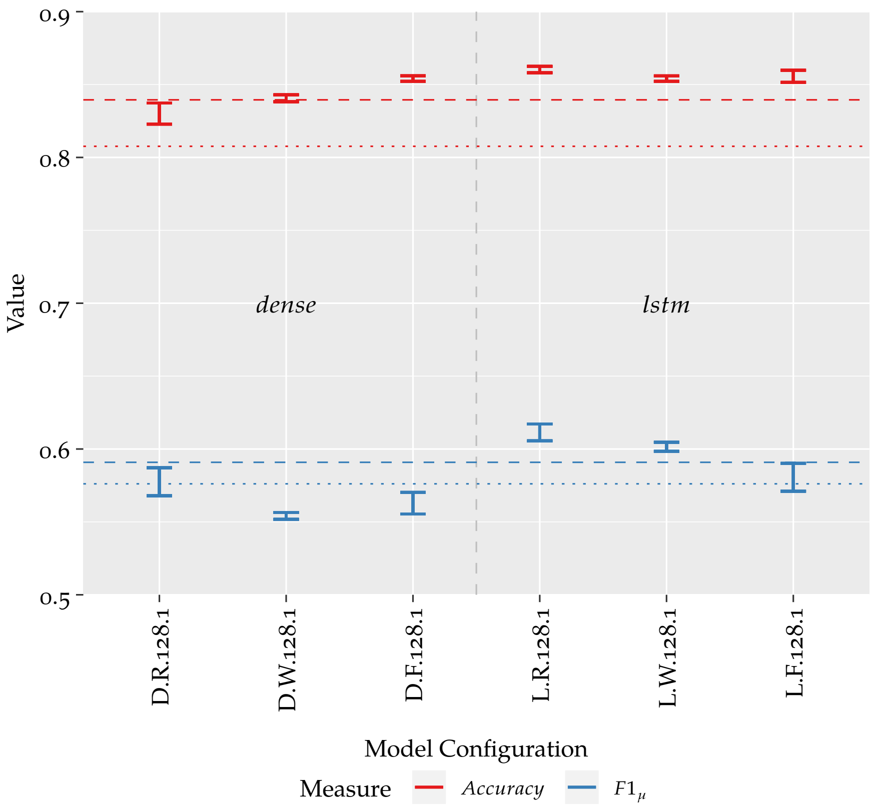


Figure 7.6: Accuracy (red) and $F1_\mu$ values (blue)—both on the same vertical axis–for the artificial neural network based models (horizontal axis). The visualized value range is reduced from $[0, 1]$ to $[0.5, 0.9]$ for better visibility. Horizontal lines represent the accuracy and $F1_\mu$ of the *rule* (dotted) and *BnM* (dashed) model.

Both artificial neural network models achieve a better accuracy than the *Rule* based model. For the *dense* model, an increase in the accuracy and a decrease in the size of the confidence interval can be observed with growing feature complexity. When using the *full* feature vector, the *dense* model can achieve better accuracy than the *BnM* model. This does not happen with the $rule_{raw}$ and *rule* feature for which the *dense* models perform similar to *BnM*. The accuracy of the *lstm* based models is above the results of the *BnM* model. Between the different feature sets, only small changes in the accuracy can be observed for the *lstm* model.

Considering the $F1_{\mu}$ measurements, the models show different, partially contradictory, results. From the perspective of $F1_{\mu}$, the *dense* models perform for all feature sets equally or worse than both the *Rule* and the *BnM* model. They never perform better. Additionally, they show a high variability in combination with the *rule* and *full* feature. The $rule_{raw}$ feature allows the *dense* models to achieve results which are more stable but not necessarily better. *Lstm* based models achieve better $F1_{\mu}$ than *Rule* and *BnM* when used with the *rule* and $rule_{raw}$ features. In case of the $full$ feature, the $F1_{\mu}$ of the LSTM models shows a degradation in comparison to the other features. The $F1_{\mu}$ results of this configuration are in between the *Rule* and *BnM* model.

Although, most configurations of the artificial neural network models achieve a better accuracy than the *Rule* model, a majority of them can not achieve better $F1_{\mu}$. To further analyse this discrepancy, the F1-score for each class can be investigated. A visualization of these measurements is shown in Figure 7.7. This representation of the model performance allows the following observations: The F1-score of *Non-Participant* is high for all models including the *Rule* and *BnM* based models. Artificial neural network based models achieve better results—by a small margin—when using LSTM layers with high-level features or the *full* feature set and the *dense* model. Predictions of *Side-Participant* are better than the *Rule* based predictions in all configurations. The strongest improvement in comparison to the *Rule* model can be observed for this class. To achieve a prediction of *Side-Participant* that is comparable to *BnM*, the *dense* model needs to be trained with the *full* feature set and the *lstm* model with the $rule_{raw}$ features. Better results can be achieved when using the *rule* feature with the *lstm* model. In the prediction of *Speaker*, the *dense* model achieves results similar to the *BnM* model with the *rule* feature, similar to the *rule* model with the $rule_{raw}$ feature, and worse with the $full$ feature set. The *lstm* models achieve similar results for *Speaker* with the *rule* and $rule_{raw}$ features but show a strong drop in F1-score for the $full$ feature. *Addressee* is the only role for which the *BnM* model can not compete with the *Rule* model. This observation is even stronger for the artificial neural network models. The F1-scores of the *dense* models for *Addressee* is equal to or worse than the *BnM* models. While the *lstm* model can achieve results between the *Rule* and *BnM* models, it's best recognition of *Addressee* is achieved with the $full$ feature set.

By looking at the confusion matrices of the best performing *dense* and *lstm* models in comparison to the *rule* and *BnM* model (Figure 7.8)

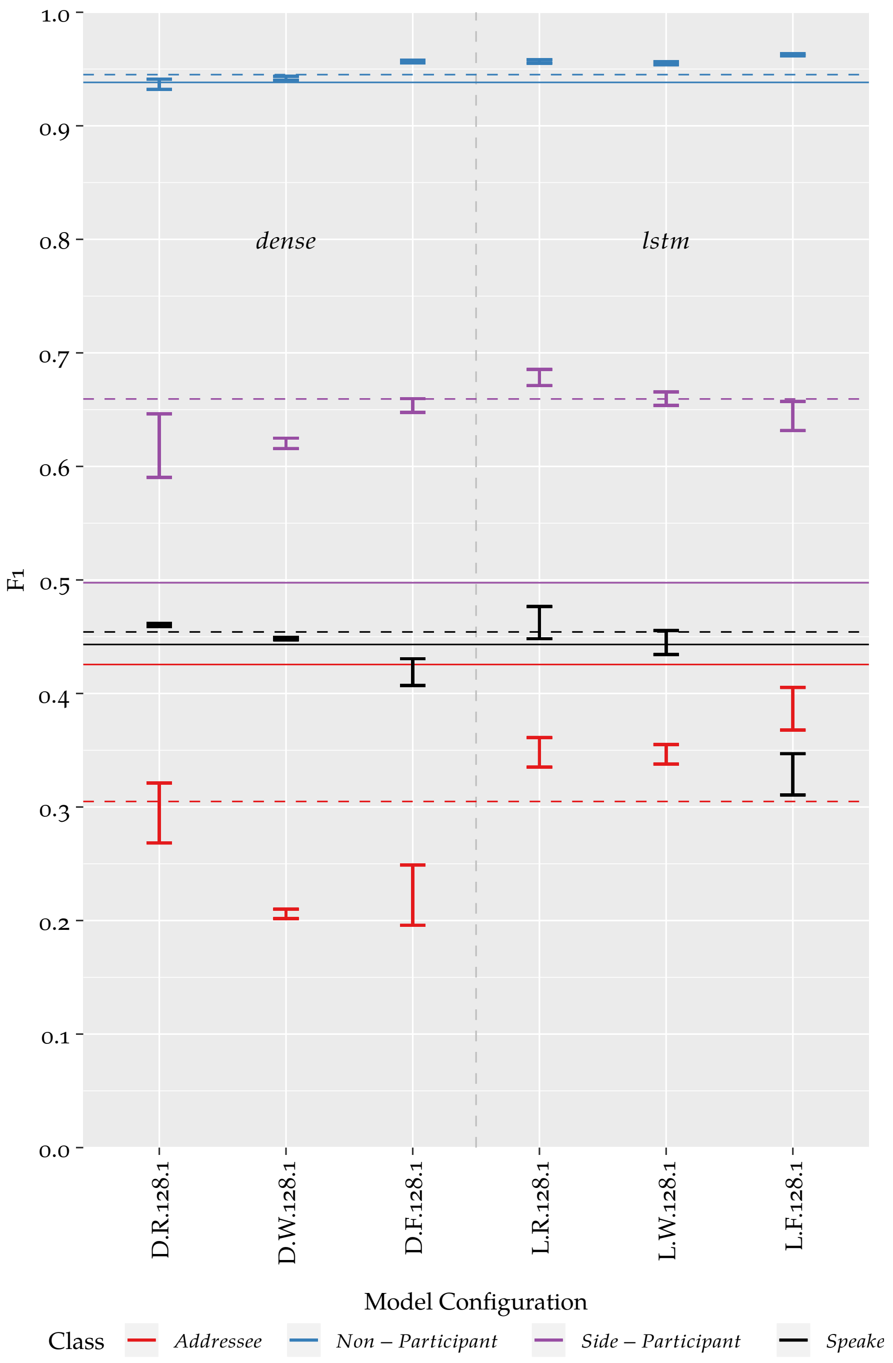


Figure 7.7: F1-scores (vertical axis) for the artificial neural network based models (horizontal axis) for each class separately. Colours of the error bars encode the classes *Speaker* (black), *Addressee* (red), *Side-Participant* (violet), and *Non-Participant* (blue). Horizontal, lines represent the F1-score of the *rule* (solid) and *BnM* (dashed) model, using the same colour coding.

further observations can be made. The recall of *Non-Participant* is high in all models. However, when the agent is part of a conversational group the *rule* model shows a bias towards predicting *Addressee* and the other models towards *Side-Participant*. The higher accuracy of the *dense* model is rooted in its focus on *Non-Participant*. While it has the best result for this role, all other predictions are worse than the *BnM* model's. Only the *lstm* model trained with the *rule* feature set can outperform the *BnM* model for most conversational roles.

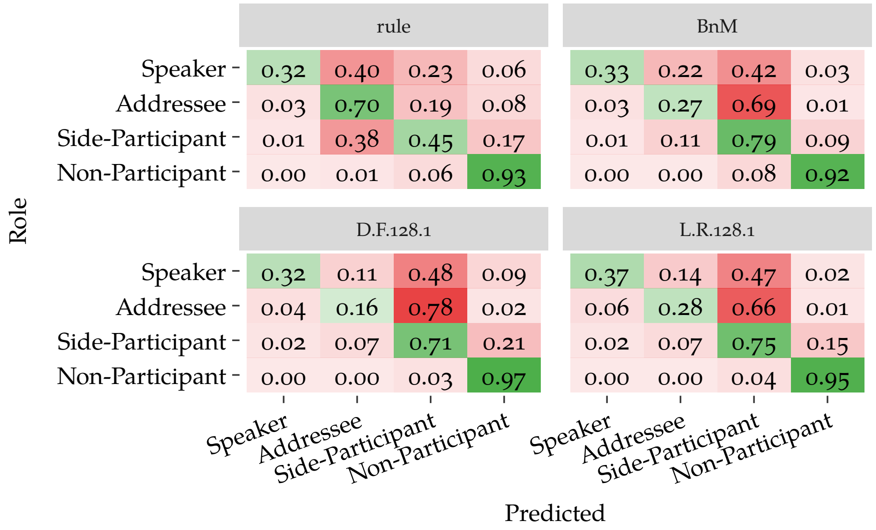


Figure 7.8: Confusion Matrices of the *rule*, *BnM* and best performing *dense* and *lstm* models. The results are normalized by the amount of observations of each role in the corpus. The green tiles, represent recall and the red tiles can be interpreted as the chance to misclassify into the corresponding role.

7.2.5 *Discussion*

By utilizing artificial neural networks for the recognition of conversational roles, the overall accuracy can be improved in comparison to the *Rule* model. The accuracy of the *BnM* model can only be surpassed when using high-dimensional data ($full$ feature) with the *dense* model or time sequences of low-dimensional, high-level features (*lstm* model). The $F1_\mu$ and class-wise F1-scores measurements present a more nuanced picture. In case of the *rule* feature, and without time information, the *BnM* model can not be outperformed. It, by definition, produces statistically optimal results. This is confirmed by the observation that achieving higher accuracy is only possible by providing the models with more information. By using the $full$ feature set, the *dense* models can achieve an accuracy of $\approx$ 85% which is better than the accuracy of the *BnM* model. A comparison with the F1-scores suggests that the increased accuracy is achieved by further enhancing the recognition of *Non-Participant*. This can be confirmed by investigating the corresponding confusion matrices. For *Side-Participant* the recognition remains unchanged and gets worse for the other roles. By observing the situation over time, the models that use LSTM layers can outperform the *Rule* and *BnM* models. In combination with the *rule* or $rule_{raw}$ feature set, these models achieve the overall best accuracy and $F1_{\mu}$. This is the only configuration that can achieve a higher F1-score than *BnM* for *Addressee* without loosing performance for the *Speaker* and *Side-Participant* roles. The results of the LSTM models can not be enhanced by increasing their complexity—the number of layers or units in each layer (see Figures B.1 and B.2). This suggests that the less complex models already have enough capacity to represent the informative properties of the data. Over-fitting is reliably reduced by the drop-out layers, allowing the more complex models to achieve similar results. The resilience of the models against long training (Figure B.3), further confirms this. Increasing the complexity of the input by using the $full$ feature set does not help to further enhance the *lstm* models' performance. While the F1-score for *Non-Participant* further improves, the results for the other classes get worse. The strongest loss in F1-score can be observed for *Speaker*, the role with the least amount of examples. This suggests that the 148 dimensional $full$ feature over a sequence of 15 observations introduces too much noise to be handled with the available amount of data. Like the Bayesian Networks, the artificial neural network models give preference to the more common role *Side-Participant* before the *Addressee* role. Therefore, the *Rule* based model may be considered preferable when the recognition of the *Addressee* role is more important than of the *Side-Participant*. However, as the Bayesian Networks and artificial neural networks produce a probability distribution over the roles, their results can be further processed to account for any trade-off between the classes. Furthermore, the *Rule* model performance is fixed. The performance of the *BnM* model can not get much better because of its fixed structure and input. By using artificial neural networks with low-level features, the accuracy of the system can be enhanced in comparison to the other models. Simultaneously, the $F1_{\mu}$ of the system goes down. It is possible that these results will improve when trained with more data. Nevertheless, Claim 7.2[C] can not be confirmed on the basis of the observations in this chapter. Using sequences of observations, the *lstm* models achieve better overall results than the other models when using the high-level features. Provided with more training data, they have the potential to further improve—especially for the less common roles and low-level features. Therefore, it can be confirmed that models

[C] Claim 7.2: By learning from lower level features, the recognition of conversational roles can be further enhanced.

using time information achieve better results than models that only use the newest observation Claim 7.3↻.

7.3 SUMMARY

Each participant of a conversational group can assume different roles within the conversation which dynamically change over time. Acting in accordance to its conversational role allows an artificial agent to better meet the expectations of its interaction partners and raise the quality of the interaction (see Sections 2.1.3.3 and 2.2.2.1). Furthermore, this knowledge is a basic requirement if the agent is supposed to influence its role or the roles of its interlocutors in an informed and autonomous manner. Therefore, I stated RQ 4↻ and investigated it in this chapter. I showed that, on the basis of the high-level features that were developed in the previous chapters, the conversational role of an artificial agent can be recognized using simple rule or Bayesian Network based models (Claim 7.1↻). Furthermore, I examined whether a better classification can be achieved with different, lower-level features or by observing the situation over time. To this end, I created artificial neural networks that use only the most recent observation and such that use all observations of the preceding second to predict the agent's role. An evaluation of these models revealed that observing the high-level features over time allows achieving better classification results than possible from only a single observation (Claim 7.3). Using all available, raw features does not further improve the model performance. Therefore, Claim 7.2 could not be confirmed. Nevertheless, the increase in the accuracy of the model suggests, that further improvements are possible with more data.

Considering RQ 4, the collected observations show that the conversational roles of artificial agents can be recognized on the basis of simple models when high-level information about the interaction is available. This can be further enhanced by observing the situation over time with more complex models.

↻ Claim 7.3: By observing how the interaction unfolds in time, the conversational role can be better recognized than from the latest observation alone.

↻ RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?

↻ Claim 7.1: Given a set of high-level features, the conversational role of an agent can be recognized using simple models.

Part IV

PERSPECTIVES

In this part, I summarize the contributions I make in this thesis. Furthermore, I present possible enhancements, extensions, applications, and future work.

8 RECAPITULATION OF CONTRIBUTIONS

In this thesis, I investigate human interaction with arbitrary devices, virtual agents, and robots in a smart environment. With my investigations, I focus on the problems of addressee recognition in interactions with changing types of devices and mixed human-agent groups, and the more fundamental problems of conversational group detection and conversational role recognition in unconstrained HAI.

8.1 RESEARCH TOPIC

The overarching goal of this thesis is to use the perceptive capabilities of the environment and contained agents to better recognize the conversational expectations of inhabitants towards different kinds of artificial agents. To approach this goal from different directions, I articulate four research questions. Two of them consider the distinction of the addressee of communicational acts in interaction with changing agents on the one hand: 'RQ 1: Which behaviours in naïve human interaction with a smart environment can be observed to distinguish which agent is addressed with a deliberate communicational act?' and in interactions in a fixed, conversational human-robot group on the other hand: 'RQ 2: How can an artificial agent visually recognize whether it was addressed by a person within its conversational group or not?'. The second two research questions aim at a more global recognition of the conversational participation of artificial agents. By investigating the detection of conversational groups: 'RQ 3: How can focused interactions of people with artificial agents be automatically recognized in a smart environment?' and the recognition of conversational roles of the agent: 'RQ 4: How to determine conversational roles of artificial agents in dynamically changing interactions in a smart environment?'. In Chapter 2, I present the scientific foundation of these investigations. By performing a literature review on human interaction from the perspective of social sciences, I create an overview of how people behave in copresence. I contrast these findings with a literature review from the perspective of computer sciences. In this review, I show that behaviours and expectations from HHI interaction can be similarly observed in human interactions with artificial agents. Furthermore, I show how they can be automatically recognized and how they are utilized in interaction scenarios with artificial agents and smart environments in the literature. The compiled overview of research on human interaction from both social sciences and computer science helps to understand human expectations and behaviours in copresence with artificial agents.

8.2 ADDRESSEE IN COMMUNICATIVE ACTS

In Part II, I investigate human addressing behaviour and how it can be interpreted in interactions with devices, robots, and smart environments. Artificial agents and devices are products of human imagination and design. To make them usable for people without prior training, they need to match the user's intuition. The literature on smart environments, suggests and evaluates multiple ways of communicating the addressee and task in interaction. However, approaches in which people can freely choose their own way of interacting are rare (Section 2.3.2). The contribution I make with RQ 1 in Chapter 3, is an investigation of human addressing behaviour in a smart, robot inhabited apartment with naïve users and deliberately unconstrained interaction possibilities. The investigation shows that the visual focus of the user is the most important cue for the distinction of addressed entities, regardless of their form. If the addressee has a humanoid form, like the robot Floka, this effect is stronger. The used modality and its form encode further important features. In contrast to interactions proposed in the literature, direct, verbal addressing with terms such as *robot* or *apartment* is rare. The used gestures are general—e.g. waving, pointing, clapping—and only intelligible in combination with other modalities such as gaze or speech.

When an artificial agent participates in a conversational group with multiple persons, the problem of distinguishing utterances addressed towards it from utterances exchanged between the other participants is evident. Most systems that explicitly deal with this distinction, use close talk microphones or sound source localization to find the current speaker and derive the addressee from the speaker's visual focus of attention. The usefulness of the sound source localization to detect the speaker is plausible and the speaker's gaze behaviour is repeatedly investigated in HHI and HAI. In the investigation of RQ 2 in Chapter 4, I make a contribution to addressee recognition within mixed human-robot conversational groups by presenting and evaluating a visual mouth-movement detection as an alternative source of information. A person—focused using sound source localization—can be verified or rejected as the current speaker with this approach. On this basis, I create and evaluate an addressee recognition model that fuses information from speaker detection, mutual gaze detection, speech recognition, and the robots state using Bayesian Networks. The resulting model can strongly reduce the amount of unwanted responses in human interactions with artificial agents at the cost of a small amount of false rejections.

8.3 GROUPS & ROLES IN COPRESENCE

To behave socially appropriate in copresence with humans over longer periods of time, artificial agents need to understand their role in the situation and act accordingly. In Part III, I investigate how artificial agents can better understand the situation and the flow of interaction in copresence with humans. To be able to investigate RQ 3 and RQ 4, I present a scenario and corresponding corpus in which people freely interact with each other and two virtual agents over an extended time period (Chapter 5).

Conversational groups are a fundamental building block for human interactions. By creating a conversational group, people not only optimize the efficiency of the communication within the group. They impose specific behaviours and responses on all agents in copresence (Sections 2.1.2 and 2.1.3.2). To be able act in conformance to these social norms, an artificial agent needs to have an understanding of conversational groups. While F-Formations are utilized in some HRI scenarios, their effects on people are mainly investigated in virtual environments and the detection performed on human-only groups (Section 2.2.2.5). To investigate RQ 3 in Chapter 6, I present a fully autonomous approach for the detection of mixed human-agent conversational groups and dedicated in-group detections for agents that lack a robust person tracking. As a contribution, I show that conversational groups containing a combination of humans and artificial agents can be detected using F-Formations, as it is done in the analysis of human interactions. I further show, that the distinction between in group and out-of group can already be performed on the basis of face detections in the agents field of view with acceptable results.

The distinction of utterances that are addressed towards an artificial agent from other utterances is a basic requirement for most modern conversing systems. In the literature, as in Chapter 4, this is done on the basis of utterances (Section 2.2.2.2). While this allows the agent to respond to people's statements, it is not enough to generate role specific behaviour or to notice deviations between the course of the interaction and the robots inner representation (Section 2.2.2.3). For the investigation of RQ 4 in Chapter 7, I harness the results and models made while working on RQs 1–3. I contribute to the advancement of the capabilities of artificial agents by showing that their conversational roles can be recognized continuously. With simple rule based and Bayesian Network models, I show that the features developed in the previous chapters allow an automatic recognition of an agent's conversational role. Furthermore, I show how time-sequence information and lower level features can be harnessed to further enhance the recognition quality with artificial neural networks. These results constitute a basis for further research on the automatic recognition and utilization of conversational roles in HAI.

9 OUTLOOK

The contributions made in this thesis are based on the current state of research on human interaction with artificial agents and further extend on it. Although, the presented models leave room for further improvements, they already solve existing, practical problems. In doing this, they also lay the foundation for further advances in the state of the art.

9.1 POSSIBILITIES FOR IMPROVEMENT

The analyses in Chapter 3 are performed on manually annotated data. To create a better understanding of human addressing of agents in smart environments, a fully automatic approach for the addressee detection is required. Only with an automatic recognition, these findings can be practically applied to aid further studies. To this end, it is interesting to see if approaches to conversational group detection and conversational role recognition are transferable to human interactions with devices. Furthermore, all objects in the CSRA were possible addressees in the addressing study and were annotated by an uninvolved person. Further research needs be undertaken to better understand which entity the participants *intended* to address and if some objects only functioned as proxies for expected, invisible agents.

The addressee recognition model used in Chapter 4 can be further enhanced by reducing the noise in the robots face detections model or creating a dedicated artificial neural network to better distinct speaking from other types of mouth movements. Furthermore, the model currently only considers visual information of the focused person. Acoustic information is used only indirectly through the attention management of the robot. By maintaining a model of each participant and fusing information from visual and acoustic information, the speaker detection can be further enhanced.

The group detection model presented in Chapter 6 is sensitive to the quality of the underlying person tracking. Reducing the noise in the person tracking and eliminating blind spots will enhance the overall results of the group detection. Furthermore, a detailed investigation of the cases that pose problems in the evaluation can reveal possible improvements to the model. Because the model for conversational group detection is based on an approach to F-Formation detection from HHI research, it is best suited for interaction of standing people in open areas. For the application in a smart flat, other situations need additional consideration. To deal with seated people or people who are occupied with a different task, further adaptations of the model may be required.

In the role recognition presented in Chapter 7, the classification of the *speaker* and *addressee* roles needs further improvements. Because these are the roles with the least amount of observations, extending the corpus with observations of more interactions is likely to improve the results To this end, observations with a low certainty in the artificial neural network model results could be specifically and automatically collected. Furthermore, the role recognition currently only uses the assignment costs and o-space centre of the conversational group. By considering all the individuals that are part of the group and maintaining models of their state, further improvements of the conversational role recognition are possible.

Finally, all studies are performed in the CSRA with the corresponding artificial agents. The participants are native German speakers, recruited from the campus of the Bielefeld University. Further research with participants with different backgrounds and capabilities, and with different agents will broaden the applicability of the results.

9.2 APPLICATIONS & POSSIBILITIES FOR FURTHER RESEARCH

Despite the possibilities for further improvement, the results of this thesis already allow practical applications and pave the way for further research.

The investigations on addressing behaviour present a strong argument for the consideration of the users attention and other modalities in human interaction with smart environments and artificial agents. Further research on the detection of attention in smart environments should be done to allow the utilization of this important information while simultaneously respecting other requirements of smart environments such as privacy, data safety, and energy efficiency.

The proposed model for addressee recognition in multi-party interaction can be used in intelligent devices, IPAs, or artificial agents to enhance their interaction capabilities. In the CSRA, this model is used to reduce unintended responses from the virtual agents. The visual detection of speakers is especially helpful for agents that can not perform sound source localization. Because the approach is fully automatic, it can be applied in long-term studies on human interactions in smart environments.

The detection of mixed human-agent conversational groups is also continuously active in the CSRA and forms a basis of the presented conversational role recognition models. Through it's fully automatic nature, it is a basic building block for the creation of group analysis and group behaviour generation models for autonomous agents. On the one hand, this allows the transfer of research from group based human interaction (Section 2.2.2.4) to mixed human-agent interaction scenarios. This entails, the analysis of the social and interactive meaning of group formations and distances. On the other hand, research on the generation and effects of artificial agent behaviour (Section 2.2.2.5) can be performed without having to keep fixed group configurations. Furthermore, this allows the analysis of group formations and peoples preferences in HAI over longer periods of time.

The recognition of conversational roles allows generating role specific agent behaviour. While some possible behaviours are already explored in the literature (Section 2.2.2.3), their application is rather sparse. The robots in these investigations use conversational signals when taking or releasing the turn or to show their attention when not speaking. By recognizing the role continuously, the behaviour of the agent can be further adapted during a turn. An agent should not only be able to react when addressed or interrupted, but also when attention shifts and role changes happen during a turn. By comparing the recognized role of the agent with the expected role, problems in the mutual understanding of the situation can be detected and repair strategies initiated.

To be accepted in long term interactions, artificial agents do not only need to understand commands. They need a model of the interaction they are in. They need to know to whom to talk and whom to listen to. To people who want to converse, they need to direct special attention, and they need to respect that others do not want attention. With the investigations in this thesis, I intend to make this more achievable by broadening the recognition-possibilities of artificial agents in smart environments.

Part V

APPENDIX

A

ADDRESSING BEHAVIOUR IN SMART ENVIRONMENTS

Tier	T	A	V	N	L
Addressee final	C	*			18
Expression (facial, gestural, verbal)	C	*			6
Expression specific	F	*			21
Focus of attention	C	*			19
Method	C	*			3
Method specific	F	*			282
Speech form of address	C	*			3
Speech politeness	C	*			3
Speech type of sentence	C	*			7
Speech specific	F	*			149
Speech intention	C	*			4
Study progress coarse	C	*			4
Study progress fine	C	*			8
Wizard	C				21
Radio	B				2
Robot gesture	C				4
Robot speech	F		*		11
Displayed text	F			*	27
Apartment Call	B		*		2
Apartment Parcel	B		*		2
Apartment Time	B		*		2
Cupboard handle light	B				2
Cupboard handle sound	B				2
Cupboard door state	B				2
Apartment door state	B				2

Table A.1: The tiers, available in the *ELAN Linguistic Annotator* [ELAN] annotations of the addressing study. The column T (Type) depicts the kind of annotation: (C)ategorical, (F)ree-text, or (B)inary; The column A (Annotated) depicts whether the tiers were manually annotated (*) or extracted from system events. The columns V (Verbal) and N (Non-Verbal) highlight tiers which are only relevant in the verbal/non-verbal condition. The column L (Levels) tell how many different values the tier assumed in the annotation. *Addressee final* and *Focus of attention* have different levels because not all entities were addressed/focused during the study.

Entity	Mapped to
Floor lamp L_F	Floor lamp L_F
Light in the hallway L_H	Light in the hallway L_H
Robot	Robot
Unspecific	Unspecific
Not discernible	Unspecific
Self	Unspecific
Furniture of apartment	Parts of the Apartment
Loudspeaker (by the fridge)	Parts of the Apartment
Screen (entrance)	Parts of the Apartment
Screen (kitchen on worktop)	Parts of the Apartment
Screen (living room table)	Parts of the Apartment
Screen (living room wall)	Parts of the Apartment
Screen (living room window)	Parts of the Apartment
Sliding door (btw. hallway & kitchen)	Parts of the Apartment
Switch (living room)	Parts of the Apartment
Switches (entrance)	Parts of the Apartment
Switches (kitchen by the fridge)	Parts of the Apartment
Switches (living room by the kitchen)	Parts of the Apartment
Switches (living room lamp),	Parts of the Apartment
Parts of the apartment	Parts of the Apartment

Table A.2: This shows the mapping of entities that could be addressed and focused to the reduced set that is used during the analyses in Chapter 3 on page 43.

{*T, PID*} → *AR*: The addressee is chosen according to the task at hand and the participants personal preferences.

C → *AR*: *Condition* correlates with the addressee. It can understood as an external factor that influences the participants preferences in interaction with the environment.

AR → {*FR, M*}: The selection of the addressee influences the participants attention and which method is applied for interaction.

AR → {*SP, SSR*}: When speech is used, it additionally influences the politeness and whether the entity is verbally named.

O → *M*: *Order* correlates with the applied method. Therefore, *Order* acts as an external factor that influences the participants preferences.

PID → *AEF* → *FR*: The participants preferences influence whether addressee and focus must be equal which in return affects the focus. The connection *Aef* → *Fr* ← *Ar* suits the intuition, that *Addressee equals focus* can only inform about the addressee when the focus is also known.

SP → {*SF, STR, SPH*}: Politeness affects the appropriateness of forms of addressing, types of sentences, and phrasing.

PID → *ER*: Whether participants show strong emotions is determined by their character.

M → {*SP, MSR*}: Whether speech or gestures are used at all, in implied by the applied method.

PID → *MSR*: The choice of the appropriate gesture is influenced by the participants preferences and expectations.

AR → *AW*: The wizards choice of addressee only depends on the wizards estimate of the participants addressee.

Figure A.1: The reasoning behind the manually created Bayesian Network structure in Section 3.4 on page 58.

B CONVERSATIONAL ROLE RECOGNITION

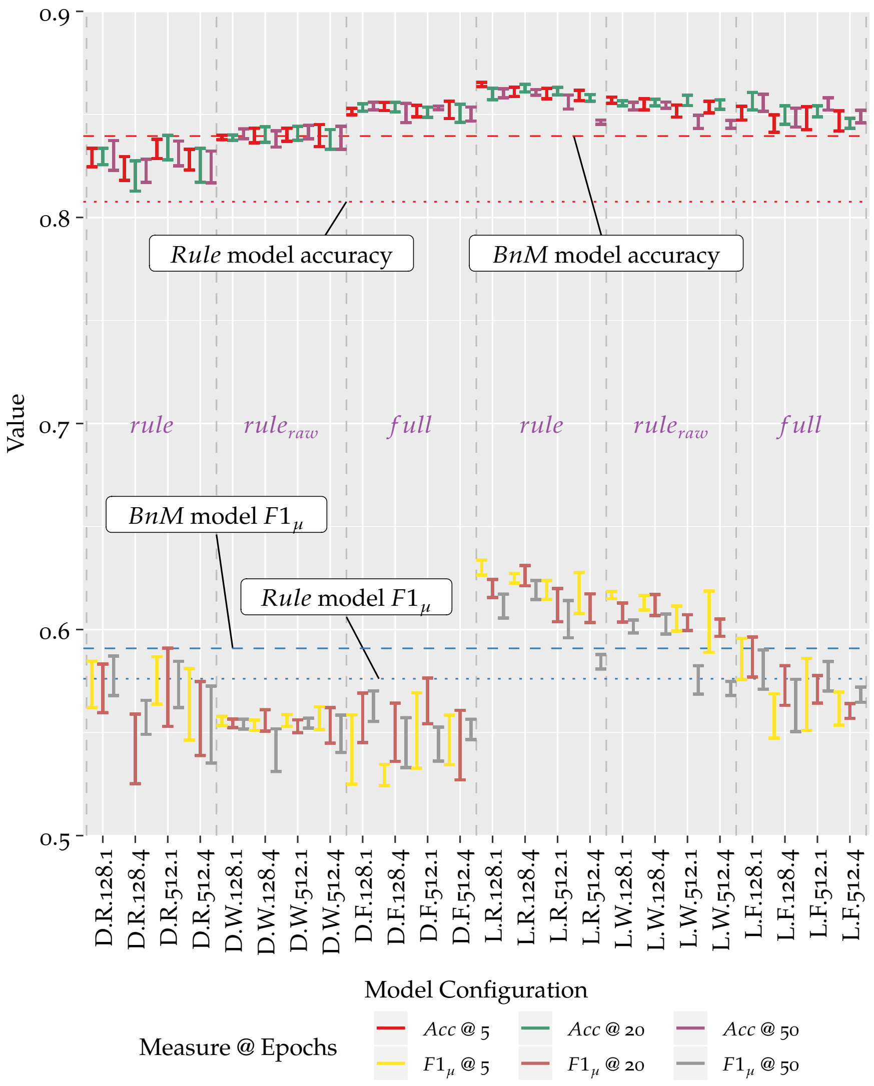


Figure B.1: Accuracy and $F1_\mu$ values (both on the same vertical axis) for the artificial neural network based models (horizontal axis). The visualized value range is reduced from $[0, 1]$ to $[0.5, 0.9]$ for better visibility. Colours of the error bars encode the type of measurement and number epochs the models were trained. Horizontal lines represent the accuracy (dashed) and $F1_\mu$ (dotted) of the *rule* (red) and *BnM* (blue) model. The different feature sets are separated by vertical dashed lines and labelled in violet. The *dense* models can be seen on the left, the *lstm* models on the right side.

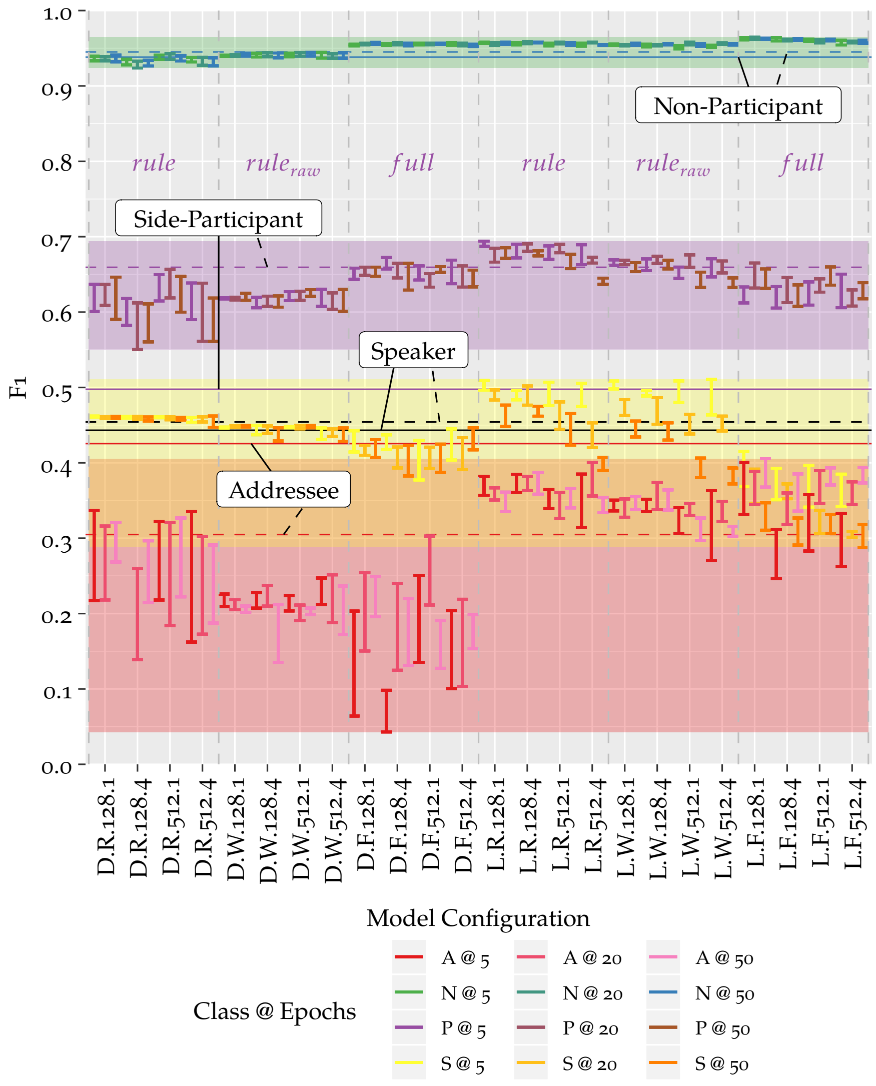


Figure B.2: F1-scores (vertical axis) for the artificial neural network based models (horizontal axis) for each class separately. Colours of the error bars encode the class and number of epochs the models were trained (*S* = *Speaker*, *A* = *Addressee*, *P* = *Side-Participant*, and *N* = *Non-Participant*). Coloured, translucent background-strips highlight the value ranges of the classes. Horizontal, lines represent the F1-score of the *Rule* (solid) and *BnM* model (dashed) for the classes, coloured in red (*Addressee*), black (*Speaker*), violet (*Side-Participant*), and blue (*Non-Participant*). The different feature sets are separated by vertical dashed lines and labelled in violet. The *dense* models can be seen on the left, the *lstm* models on the right side.

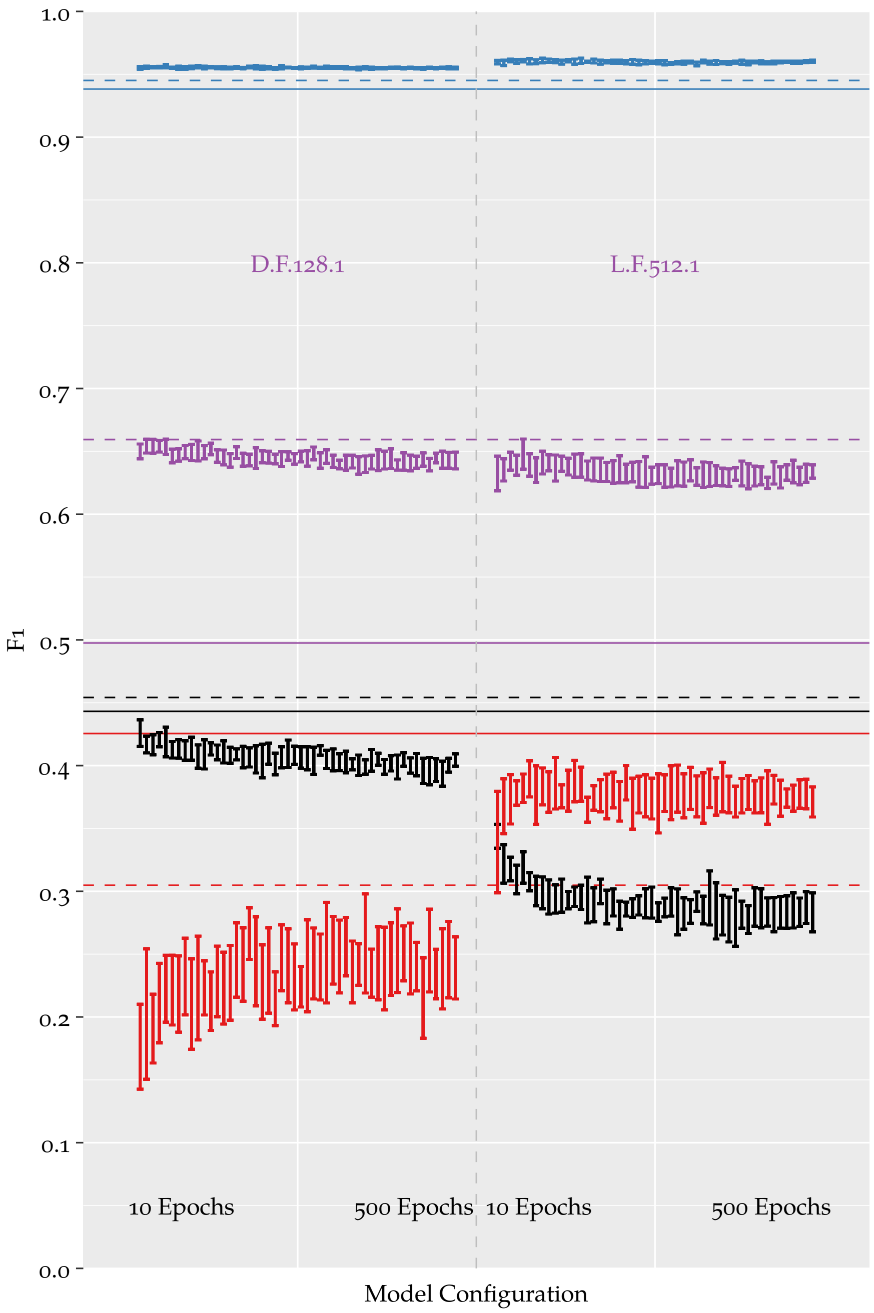


Figure B.3: F1-scores (vertical axis) for a *dense* model (left) and an *lstm* model configuration after 10 to 500 epochs of training (left to right for either side). Colours of the error bars encode the class *Speaker* (black), *Addressee* (red), *Side-Participant* (violet), and *Non-Participant* (blue). Horizontal, lines represent the F1-score of the *rule* (solid) and *BnM* model (dashed). The models are separated by a dashed lines and labelled in violet.

ACRONYMS

A

AUC

Area Under The Curve. *used on: pp. 27, 77–80, 105–107*

C

CITEC

Cluster of Excellence Cognitive Interaction Technology. *used on: pp. 6, 70*

CN

Condition Negative. *used on: p. xvi*

CP

Condition Positive. *used on: p. xvi*

CSRA

Cognitive Service Robotics Apartment As Ambient Host. *used on: pp. 6, 43–45, 65–67, 69, 70, 82, 87, 89, 90, 92, 94, 133, 134*

CV

Cross-Validation. *used on: pp. 61, 62, 77, 113*

D

DOR

Diagnostic Odds Ratio. *used on: pp. xvii, 72–76*

F

FDR

False Discovery Rate. *used on: p. xvi*

FN

False Negative. *used on: pp. xv, xvi, 76, 79, 100, 102*

FNR

False Negative Rate. *used on: pp. xvi, xvii*

FOR

False Omission Rate. *used on: p. xvi, 102*

FOV

Field of View. *used on: pp. 29, 32, 91, 119*

FP

False Positive. *used on: pp. xv, xvi, 79, 100, 102, 117*

FPR

False Positive Rate. *used on: pp. xvi, xvii, 78, 79, 102, 105, 106*

G

GUI

Graphical User Interface. *used on: pp. 34, 35, 37*

H

HAI
: Human-Agent-Interaction. *used on: pp. 9, 21, 24, 26, 28, 29, 31–33, 38, 39, 88, 94, 129–131, 135*

HCI
: Human-Computer-Interaction. *used on: p. 3*

HHI
: Human-Human-Interaction. *used on: pp. 22, 24–28, 30, 32, 37–39, 95, 96, 99, 103, 108, 129, 130, 133*

HRI
: Human-Robot-Interaction. *used on: pp. 4, 9, 21, 28, 37, 65–67, 71, 81, 131*

I

IPA
: Intelligent Personal Assistant. *used on: pp. 3, 21, 36, 134*

L

LR-
: Negative Likelihood Ratio. *used on: p. xvii*

LR+
: Positive Likelihood Ratio. *used on: p. xvii*

LSTM
: Long Short-Term Memory. *used on: pp. 117, 119, 120, 122, 125*

M

MDL
: Minimum Description Length. *used on: pp. 97, 99–102, 104, 118*

MLP
: Multilayer Perceptron. *used on: p. 29*

N

NPV
: Negative Prediction Value. *used on: pp. xvi, xvii*

P

PN
: Predicted Negative. *used on: p. xvi*

PP
: Predicted Positive. *used on: p. xvi*

PPV
: Positive Prediction Value. *used on: pp. xvi, xvii*

R

ROC
: Receiver Operating Characteristic. *used on: pp. 77–79, 81, 104, 105*

ROI
: Region of Interest. *used on: pp. 91, 104*

S

SVM

Support Vector Machine. *used on: p. 27*

T

TN

True Negative. *used on: pp. xv, xvi, 100*

TNR

True Negative Rate. *used on: pp. xvi, xvii*

TP

True Positive. *used on: pp. xv, xvi, 100*

TPR

True Positive Rate. *used on: pp. xvi, xvii*

V

VFoA

Visual Focus of Attention. *used on: pp. 27–30, 65*

W

WoZ

Wizard-Of-Oz. *used on: pp. 28, 36, 45*

GLOSSARY

A

accuracy

$$\text{accuracy} = \frac{TP + TN}{CP + CN} = \frac{|\text{correct classifications}|}{|\text{all classifications}|}$$

The amount of correct classifications (correctly accepted and correctly rejected) in comparison to the total sample size. This measurement can be generalized for problems with more than two classes as the sum of correct classifications divided by the size of the population. *used on: pp. xvi, 26–28, 44, 72–76, 79, 114–117, 121, 122, 124–126, 143*

activity space

The space that is occupied by the activity of an agent. Entering it may cause discomfort in the agent [LE11]. *used on: p. 23*

addressee

The addressee is the participant a speaker directs the speech to. While multiple or all participants of a conversation can be addressed at the same time, one person usually can be considered the main addressee. *used on: pp. 5, 6, 18–20, 24–30, 33–37, 39, 41, 43, 44, 46, 48–69, 71, 73–76, 78–82, 90, 110, 112, 129, 130, 133, 134, 141*

addressing corpus

The addressing corpus is the corpus I automatically extract in Section 3.3 from the corpus presented in [Hol+16]. *used on: pp. 51–54, 56, 58, 63*

addressing study

A study of interactions of naïve people in the CSRA, in which participants needed to solve a set of mundane tasks. The study is presented in [Ber+16] and the corresponding corpus in [Hol+16]. *used on: pp. 44, 46, 133, 139*

affordance space

A space of a potential activity. Being there may prevent other agents from performing that activity [LE11]. *used on: p. 23*

apartment

Also known as CSRA *used on: pp. 6–8, 44–49, 51, 57, 68, 70, 71, 89–93, 100, 102, 130*

Area Under the Curve (AUC)

The area under the ROC curve can be calculated as a measure of classifier performance over a set of possible parametrisations. *used on: pp. 27, 77–80, 105–107*

artificial agent
: In contrast to living beings, artificial agents are a result of human engineering. This encompasses robots and virtual agents. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. vii, 1, 3, 5, 6, 9, 11, 12, 21–23, 25, 26, 30, 31, 34, 36, 38, 39, 43, 65, 66, 81, 85, 87, 88, 93–95, 97, 99, 103, 108, 109, 117, 126, 129–131, 133–135*

artificial neural network
: Inspired by the mechanics of biological neural networks, these networks can be used for machine learning tasks. They receive an activation in their input layer and propagate it through the network to produce an activation at their output layer. *used on: pp. 109, 117–126, 131, 133, 134, 143, 144*

autonomous agent
: Autonomous agents are interactive entities that can react and interact with their environment and with other entities in it. They have believes about the world and goals which they pursue through interaction. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. 21, 134*

B

Bayesian Network
: A probabilistic, directed, acyclic graph that models dependent probabilities. *used on: pp. 26, 27, 58–64, 76–82, 110, 112–117, 125, 126, 130, 131, 141*

C

civil inattention
: In unfocused interaction, people need to display that they acknowledge the others presence and can potentially be approached. At the same time they may want to prevent the impression that they are trying to enter, disturb or eavesdrop on a focused interaction. *used on: pp. 11, 14, 87, 95, 103, 109*

close phase
: The inner region of a proxemic distance as defined by [Hal69]. *used on: pp. 13, 14*

Cognitive Service Robotics Apartment as Ambient Host (*CSRA*)
: The CSRA is a robot inhabited, smart home-laboratory in the CITEC at Bielefeld University. A discription of the CSRA can be found in Section 1.3, [Wre+17] and `https://www.cit-ec.de/csra`. *used on: pp. 6, 43–45, 65–67, 69, 70, 82, 87, 89, 90, 92, 94, 133, 134*

condition negative (*CN*)
: The sum of negative elements in the sample of a confusion matrix. *used on: p. xvi*

condition positive (CP)

The sum of positive elements in the sample of a confusion matrix. *used on: p. xvi*

confusion matrix

A matrix that contrasts classifications to ground truth data to visualize the performance of a classifier.

$$\text{confusion matrix} = \begin{bmatrix} TP & FP \\ TN & FN \end{bmatrix}$$

used on: pp. xv, 99, 100

conversation

A focused interaction with the purpose of communication between a group of people in copresence. Although, conversations can have different forms—e.g. manually coded or text based—this work focuses on direct, verbal conversations. *used on: pp. vii, 4–6, 9, 12, 17–20, 22–26, 28–32, 36, 38, 39, 43, 44, 65, 82, 109, 111, 126, 129, 135*

conversational floor

Often interchangeably used for the turn in a conversation [Hay88]. *used on: pp. 18, 20, 24, 30, 38*

conversational group

A group of two or more persons that conduct a conversation (see Section 2.1.3.2). Conversational groups often assume F-Formations. *used on: pp. vii, 5, 9, 18, 19, 24, 26, 30–32, 37–39, 65–68, 73, 81, 82, 85, 87, 88, 90, 91, 93–99, 103, 107–112, 119, 124, 126, 129–131, 133, 134*

conversational role

The roles people can assume in respect to a conversation. These can be speaker, addressee, or different types of side-participants (see Section 2.1.3.3) *used on: pp. 4, 6, 9, 25–30, 32, 38, 67, 85, 87, 88, 90, 91, 93–95, 109–114, 117, 119, 124–126, 129, 131, 133–135*

copresence

People are copresent, when they sense that they are close enough to mutually perceive each other and their mutual sensing of perceiving and being perceived. *used on: pp. 9, 11–20, 22, 23, 30, 32, 33, 38, 87, 88, 95, 129, 131*

D

device

A device is any interactive entity that is not an autonomous agent. It has an inner state that can be changed, but neither believes nor goals. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. vii, 1, 4–7, 20, 21, 31, 33–36, 38, 39, 56, 57, 129, 130, 133, 134* *diagnostic odds ratio* (*DOR*)

$$DOR = \frac{LR+}{LR-}$$

The diagnostic odds ratio is an indicator for test quality which is independent from the prevalence of the test set. It can be read as 'The odds of correcly accepting is x times higher than the odds of falsely rejecting'. Therefore, tests with discriminative power have a DOR > 1 [Gla+03]. *used on: pp. xvii, 72–76*

E

encounter

An alternative term for face engagement [Gof63]. *used on: p. 15*

F

F1-score

The harmonic mean between precision and recall:

$$F_1 = 2\frac{PPV \cdot TPR}{PPV + TPR}$$

used on: pp. xvi, 72–76, 101, 114–117, 121–125, 144, 145

face engagement

Goffman uses the term face engagement for instances of focused interaction between people [Gof63, p. 91]. *used on: pp. 15, 16, 38*

fall-out

Also known as FPR *used on: pp. xvi, 81, 155*

false discovery rate (*FDR*)

$$FDR = \frac{FP}{PP}$$

used on: p. xvi

false negative (*FN*)

Wrongly rejected elements in a confusion matrix (*Type II Error*). *used on: pp. xv, xvi, 76, 79, 100, 102*

false negative rate (*FNR*)

$$FNR = \frac{FN}{CP}$$

Miss rate. *used on: pp. xvi, xvii*

false omission rate (*FOR*)

$$FOR = \frac{FN}{PN}$$

used on: p. xvi, 102

false positive (*FP*)

Wrongly accepted elements in a confusion matrix (*Type I Error*). *used on: pp. xv, xvi, 79, 100, 102, 117*

false positive rate (*FPR*)

$$FPR = \frac{FP}{CN}$$

Probability of false alarm, also known as fall-out. *used on: pp. xvi, xvii, 78, 79, 102, 105, 106*

far phase

The outer region of a proxemic distance as defined by [Hal69]. *used on: p. 13*

F-Formation

A spacial and orientational arrangement entered and maintained by a group of people. The space between them is the o-space, to which all participants have equal and exclusive access [Ken90, p. 209]. *used on: pp. 16–18, 25, 26, 30–32, 85, 95, 96, 99, 103–105, 107, 108, 111, 117, 118, 131, 133*

Flobi

Flobi is an anthropomorphic robot head designed at Bielefeld University specifically for HRI applications. It can actuate its eyes, lids, brows, mouth and neck to show emotions, attention, mouth movements during speech [Lüt+10]. A corresponding simulation is a virtual agent with similar capabilities which can be used interchangeably with the robot head [LSW14]. The adapted robot head, that was created for the Floka is presented in [SBW19]. *used on: pp. 7, 8, 87, 89, 91, 93*

Flobi Assistance

Flobi Assistance is an instance of the simulation of the anthropomorphic robot head Flobi. It is located in the kitchen of the CSRA. A photograph of it can be seen in Figure 1.2 on page 7. *used on: pp. 7, 8, 100–102, 105–107, 110*

Flobi Entrance

Flobi Entrance is an instance of the simulation of the anthropomorphic robot head Flobi. It is located in the hallway of the CSRA. A photograph of it can be seen in Figure 1.2 on page 7. *used on: pp. 7, 8, 100–102, 105–107, 110*

Floka

Floka is an anthropomorphic robot based on the *MekaBot M1*. It has an omni-directional drive, can lift it's upper body up and down and two arms that end in hand-like manipulators. It's head can be chosen between the original sensor head of the *MekaBot M1* and a version of the Flobi head that was adapted for this particular case [SBW19]. The robot is presented in [Wac+17] *used on: pp. 7–9, 66, 68, 71, 130*

focused interaction

In a focused interaction, people come together and actively cooperate to maintain a joint focus of attention (Section 2.1.3). *used on: pp. 5, 11, 12, 14–18, 23, 24, 32, 33, 38, 39, 43, 87, 93–95, 108, 129*

I

in the wild

This term is used to emphasize research or situations that are performed outside of the controlled laboratory environment. Studies that place a robot on a public square to interact with whoever passes by are in the wild. *used on: p. 11*

information process space

The space in front of pedestrians in which they observe other pedestrians and obstacles [KF10]. *used on: p. 23*

informedness

$$\text{informedness} = TPR + TNR - 1$$

An alternative measure for recall which is not biased by the prevalence of the sample [Pow11]. It tells whether the model can detect positive and negative observations. 1 means all observations will be correctly retrieved, -1 means all will be wrongfully retrieved. *used on: pp. xvii, 72–75, 100–102, 114–117*

Intelligent Personal Assistant (IPA)

Speech activated and verbally interacting artificial agents which can be embedded in loudspeakers, televisions, smart phones, etc.. In contrast to other virtual agents they do not have a specific embodiment. *used on: pp. 3, 21, 36, 134*

interactive entity

Interactive entities are all entities with which a person can interact and which as a result change their internal state. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. 9, 12, 20, 21, 64*

intimate distance

The area $\leq 0.46\,\text{m}$ around a person where physical contact is probable and only partners and good friends may enter without discomfort [Hal69]. *used on: p. 13*

K

Kinect

The Microsoft Kinect is a sensor that can provide a coloured video stream, a depth—distance measurement—stream and an audio stream. *used on: p. 35*

L

Long Short-Term Memory (*LSTM*)

Recurrent neural networks with memory cells and gate units which allows efficient learning of long term time dependencies. They are introduced in [HS97]. *used on: pp. 117, 119, 120, 122, 125*

M

markedness

$$\text{markedness} = PPV + NPV - 1$$

An alternative measure for precision which is not biased by the prevalence of the sample [Pow11]. It tells how trustworthy the models predictions are. 1 means all predictions are correct, -1 means all predictions are wrong. *used on: pp. xvii, 72–76, 100–102, 114–117*

N

Naïve Bayes

A Bayesian Network with strong independence assumptions. *used on: p. 27*

negative likelihood ratio (*LR-*)

$$LR- = \frac{FNR}{TNR}$$

used on: p. xvii

negative prediction value (*NPV*)

$$NPV = \frac{TN}{PN}$$

used on: pp. xvi, xvii

non-participant

All people in copresence that do not participate a conversation can be considered non-participants of regarding this conversation. *used on: pp. 17–19, 90, 95*

O

o-space

The space between the participants of a focused interaction. The participants orient their upper body towards its center and coordinate themselves to maintain equal accessibility for participants and non-accessibility for non-participants [CK80, p. 243]. *used on: pp. 16–18, 23, 26, 30, 31, 96–98, 119, 134*

P

participation unit

Participation unit is a collective term for face engagements and single, unengaged persons [Gof63, p. 91]. *used on: pp. 15–17*

penetrated space

The space that is affected by an activity—e.g. through noise or odour. Its form may be different from the activity space [LE11]. *used on: p. 23*

personal distance

The area $\leq$ 1.22 m around a person personal topics can be discussed by people who know each other. It is at the edge of 'arms reach' [Hal69]. *used on: pp. 13, 22, 23, 31*

positive likelihood ratio (*LR*+)

$$LR+ = \frac{TPR}{FPR}$$

used on: p. xvii

positive prediction value (*PPV*)

$$PPV = \frac{TP}{PP}$$

Also known as precision. *used on: pp. xvi, xvii*

precision

Also known as PPV *used on: pp. xvi, xvii, 72–76, 78–81, 100–102, 104, 106, 107, 154, 157, 158*

predicted negative (*PN*)

The sum of elements rejected by a model. *used on: p. xvi*

predicted positive (*PP*)

The sum of elements accepted by a model. *used on: p. xvi*

prevalence

$$\text{prevalence} = \frac{CP}{CP + CN}$$

The proportion of true elements in the sample. *used on: pp. xvi, xvii, 72–75, 79, 80, 102, 114, 115*

proxemic space

The set of spaces from Hall's proxemics as used in a taxonomy of social spaces by [LE11]. The original term in the taxonomy is *personal space*. Proxemic space is used in this dissertation to prevent confusion. *used on: p. 23*

proxemics
: Proxemics investigate how people perceive and use interpersonal distances in different situations and social contexts. *used on: pp. 13, 14, 22, 23, 32, 33, 38, 39*

p-space
: A narrow zone in F-Formations where the bodies and personal belongings of the participants are located [CK80, p. 259]. *used on: pp. 16, 17, 26, 31, 67*

public distance
: The area ≤ 7.62 m around a person where the voice level needs to get loud, the phrasing more format and facial expressions get replaced by gestures. It is more appropriate for speeches and presentations than for conversations [Hal69]. *used on: pp. 13, 14, 22*

R

Random Forests
: A learner that uses ensembles of decision trees and voting for classification [Bre01]. *used on: pp. 60–64*

recall
: Also known as TPR *used on: pp. xvi, xvii, 72–76, 78–81, 100–102, 104–107, 124, 154, 156, 161*

receiver operating characteristic (ROC)
: The receiver operating characteristic (ROC) Curve visualizes the TPR of a classifier against its FPR for a collection of thresholds. This allows a better assessment of the trade-off between the probabilities of detection and false alarm. *used on: pp. 77–79, 81, 104, 105*

robot
: Robots are artificial agents with an embodiment that occupies physical space. They may be able to navigate, reconfigure themselves, or manipulate objects but not without changing the availability of space in doing so. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. vii, 3–9, 11, 20–26, 28–32, 36, 37, 41, 44–49, 56, 57, 63–77, 79, 81–83, 87, 88, 103, 129–131, 133, 135*

robotiquette
: Kerstin Dautenhahn proposed that a robot needs to behave in a manner that is socially acceptable to humans. This is often referred to as robotiquette [Dau07]. *used on: pp. 4, 11*

r-space
: The space outside of and between F-Formations. In this space associates of the F-Formation are usually located and it is avoided by other participation units [CK80, p. 260]. *used on: pp. 16–18*

S

selectivity

Also known as TNR *used on: p. xvi, 161*

sensitivity

Also known as TPR *used on: p. xvi, 161*

side-participant

A participant of a conversation who is neither speaker not addressee has the role of a side-participant. *used on: pp. 18, 19, 25, 28–30, 75, 90*

smart environment

A smart environment is composed of interconnected devices with the capability of sensing and actuating. It can observe inhabitants and adapt to improve their experience or simplify their tasks [CD04]. *used on: pp. vii, 3–9, 11, 20, 21, 33, 34, 37–39, 41, 43, 44, 49, 52, 55, 58, 63, 64, 87, 93–95, 108, 109, 126, 129, 130, 133–135*

smart home

A smart home is a home with the capabilities of a smart environment. *used on: pp. 3, 6, 12, 34–37, 41, 43, 49, 63*

social distance

The area $\leq$ 3.66 m around a person where less personal interactions—e.g. with colleagues or on social gatherings—can be carried out. This distance allows simple engagement and disengagement [Hal69]. *used on: pp. 13, 14, 22, 23*

speaker

The speaker is the participant of a conversation who has the right to speak. While the speakers change during the conversation, only one person can be speaker at a time. *used on: pp. 18–20, 26–30, 37, 38, 41, 65, 66, 68, 70, 71, 74, 75, 81, 83, 90, 109–112, 130, 133, 134*

specificity

Also known as TNR *used on: p. xvi, 161*

stride

The distance between a persons position and the center of its transactional segment as presented in [Set+15]. *used on: pp. 96, 97, 100, 101, 104*

T

territory space

The space that is claimed and accordingly marked by an agent or a group—e.g. a garden or room [LE11]. *used on: p. 23*

The Media Equation

According to [RN96], humans treat computers, artificial agents, and media in general similar to other humans. They show politeness, ascribe gender stereotypes, and physically react to visualized motion. *used on: pp. 3, 11, 33* *tolerance threshold*

The threshold used in a tolerant match (see Definition 1). *used on: pp. 99, 100, 102*

tolerant match

A definition from [Set+15] of how to compare a detected group of persons with a ground truth annotation to decide whether they match or not. The definition can be looked up in Definition 1. *used on: pp. 99, 100*

transactional segment

Is the area in front of a person. In this area people perform most of their activities, have the best perception of and highest degree of control over their environment [CK80, p. 240]. *used on: pp. 15–18, 68, 96–98*

true negative (*TN*)

Correctly rejected elements in a confusion matrix. *used on: pp. xv, xvi, 100*

true negative rate (*TNR*)

$$TNR = \frac{TN}{CN}$$

Also known as specificity or selectivity. *used on: pp. xvi, xvii*

true positive (*TP*)

Correctly accepted elements in a confusion matrix. *used on: pp. xv, xvi, 100*

true positive rate (*TPR*)

$$TPR = \frac{TP}{CP}$$

Probability of detection, also known as recall or sensitivity. *used on: pp. xvi, xvii*

turn

The right to speak in a conversation. Only one participant of said conversation can own the turn. *used on: pp. 18–20, 25, 26, 29, 30, 32, 37, 43, 66–68, 73, 75, 76, 81, 82, 87, 109, 111, 135*

turn taking

The act of acquiring or releasing a turn in a conversation. Often used as synonym for the turn taking system *used on: pp. 4, 24, 29, 30, 38, 39, 109*

turn taking system

A set of rules and behaviours by which the transition of a turn between participants of a conversation is negotiated (see Section 2.1.3.4). *used on: pp. 19, 30, 87*

U

ubicomp

In (), technology and interfaces blend with their environment and therefore can not be distinguished from it [Gre+11]. *used on: p. 33*

unfocused interaction

When people are not in a focused interaction but to manage their copresence they are in an unfocused interaction. *used on: pp. 9, 11–14, 22, 23, 33, 38, 87, 108*

V

virtual agent

Virtual agents are artificial agents which are not robots. They may have an embodiment—e.g. visualized on a screen—but do not change the availability of space when they act. IPAs can be counted as virtual agents too. For the taxonomy of interactive entities see Figure 2.3. *used on: pp. vii, 4, 6, 7, 11, 20–22, 24, 25, 27, 28, 32, 85, 87, 89, 95, 109, 110, 129, 131, 134*

W

wizard

The person that controls the behaviour of an agent in a WoZ study. *used on: pp. 25, 27, 45, 47–50, 52, 54, 88, 141*

Wizard-Of-Oz (WoZ)

A study set-up, in which participants interact with a seemingly interactive interlocutor that is secretly controlled by a hidden experimenter. *used on: pp. 28, 36, 45*

BIBLIOGRAPHY

OWN PUBLICATIONS

[Ber+16] Jasmin Bernotat, Birte Schiffhauer, Friederike Eyssel, Patrick Holthaus, Christian Leichsenring, Viktor Richter et al. 'Welcome to the Future – How Naïve Users Intuitively Address an Intelligent Robotics Apartment'. In: *International Conference on Social Robotics* (*ICSR*). Springer International Publishing, 2016, pp. 982–992. DOI: 10.1007/978-3-319-47437-3_96. *used on: pp.* 37, 43, 44, 49, 56, 71, 151

[Hol+16] Patrick Holthaus, Christian Leichsenring, Jasmin Bernotat, Viktor Richter, Marian Pohling, Birte Carlmeyer et al. 'How to Address Smart Homes with a Social Robot? A Multimodal Corpus of User Interactions with an Intelligent Environment'. In: *International Conference on Language Resources and Evaluation* (*LREC*). Portorož, Slovenia: ELRA, 2016, pp. 3440–3446. *used on: pp.* 9, 43, 44, 48, 151

[Ric+16] Viktor Richter, Birte Carlmeyer, Florian Lier, Sebastian Meyer zu Borgsen, David Schlangen, Franz Kummert et al. 'Are You Talking to Me? Improving the Robustness of Dialogue Systems in a Multi Party HRI Scenario by Incorporating Gaze Direction and Lip Movement of Attendees'. In: *International Conference on Human Agent Interaction* (*HAI*). Singapore: ACM Press, 2016, pp. 43–50. DOI: 10.1145/2974804.2974823. *used on: pp.* 66, 68, 82

[RK16] Viktor Richter and Franz Kummert. 'Towards Addressee Recognition in Smart Robotic Environments'. In: *Workshop on Embodied Interaction with Smart Environments* (*EISE*). New York, New York, USA: ACM Press, 2016, pp. 1–6. DOI: 10.1145/3008028.3008030. *used on: pp.* 44, 50, 55

[RK18] Viktor Richter and Franz Kummert. 'Continuous Interaction Data Acquisition and Evaluation'. In: *Companion of ACM/IEEE International Conference on Human-Robot Interaction* (*HRI*). New York, New York, USA: ACM Press, 2018, pp. 217–218. DOI: 10.1145/3173386.3177005. *used on: p.* 9

GENERAL

[Aba+15] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. 2015. URL: http://download.tensorflow.org/paper/whitepaper2015.pdf. *used on: p.* 181

[Aba+16] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean et al. 'TensorFlow: A System for Large-Scale Machine Learning'. In: *Usenix Symposium on Operating Systems Design and Implementation (OSDI)*. Savannah, GA: Usenix Association, 2016-11, pp. 265–283. *used on: p. 181*

[Akh+17] Oleg Akhtiamov, Maxim Sidorov, Alexey A. Karpov and Wolfgang Minker. 'Speech and Text Analysis for Multimodal Addressee Detection in Human-Human-Computer Interaction'. In: *Conference of the International Speech Communication Association (Interspeech)*. ISCA, 2017-08, pp. 2521–2525. DOI: 10.21437/Interspeech.2017-501. *used on: p. 88*

[Ala+16] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri et al. 'Salsa: A Novel Dataset for Multimodal Group Behavior Analysis'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.8 (2016-08), pp. 1707–1720. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2496269. arXiv: 1506.06882. *used on: pp. 31, 88*

[Alt+04] Philipp Althaus, Hiroshi Ishiguro, Takayuki Kanda, Takahiro Miyashita and Henrik I. Christensen. 'Navigation for Human-Robot Interaction Tasks'. In: *International Conference on Robotics and Automation (ICRA)* (2004), pp. 1894–1900. ISSN: 1050-4729. DOI: 10.1109/ROBOT.2004.1308100. *used on: p. 31*

[Alt91] Douglag G. Altman. *Practical Statistics for Medical Research*. 1991. ISBN: 978-0-412-27630-9. *used on: p. 50*

[And+14] Sean Andrist, Xiang Zhi Tan, Michael Gleicher and Bilge Mutlu. 'Conversational Gaze Aversion for Humanlike Robots'. In: *International Conference on Human-Robot Interaction (HRI)*. New York, New York, USA: ACM Press, 2014, pp. 25–32. DOI: 10.1145/2559636.2559666. *used on: p. 24*

[ARS18] Xavier Alameda-Pineda, Elisa Ricci and Nicu Sebe. 'Multimodal Analysis of Free-Standing Conversational Groups'. In: *Frontiers of Multimedia Research*. Ed. by Shih-Fu Chang. New York, NY, USA, 2018, pp. 51–74. ISBN: 978-1-970001-07-5. DOI: 10.1145/3122865.3122869. *used on: p. 31*

[AT09] Rieks op den Akker and David Traum. 'A Comparison of Addressee Detection Methods for Multiparty Conversations'. In: *DiaHolmia Workshop on the Semantics and Pragmatics of Dialogue*. Stockholm, Sweden, 2009, pp. 99–106. *used on: p. 26*

[Aue17a] Peter Auer. 'Anfang und Ende Fokussierter Interaktion: Eine Einführung'. In: *InLiSt - Interaction and Linguistic Structures* 59 (2017). *used on: pp. 12, 17, 19*

[Aue17b] Peter Auer. 'Gaze, Addressee Selection and Turn-Taking in Three-Party Interaction'. In: *InLiSt - Interaction and Linguistic Structures* 60 (2017). *used on: pp. 18, 20*

[Bai+01] Jeremy N. Bailenson, Jim Blascovich, Andrew C. Beall and Jack M. Loomis. 'Equilibrium Theory Revisited: Mutual Gaze and Personal Space in Virtual Environments'. In: *Presence: Teleoperators and Virtual Environments* 10.6 (2001), pp. 583–598. ISSN: 10547460. DOI: 10.1162/105474601753272844. arXiv: arXiv:1011.1669v3. *used on: p. 22*

[BCL17] Miriam Bilac, Marine Chamoux and Angelica Lim. 'Gaze and Filled Pause Detection for Smooth Human-Robot Conversations'. In: *International Conference on Humanoid Robots (Humanoids)*. IEEE-RAS, 2017-11, pp. 297–304. DOI: 10.1109/HUMANOIDS.2017.8246889. *used on: p. 29*

[BH11] Dan Bohus and Eric Horvitz. 'Multiparty Turn Taking in Situated Dialog: Study, Lessons, and Directions'. In: *Special Interest Group on Discourse and Dialogue (SIGDIAL)* 1974 (2011), pp. 98–109. *used on: pp. 28, 30, 88*

[BK04] Y. Boykov and V. Kolmogorov. 'An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.9 (2004-09), pp. 1124–1137. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.60. *used on: p. 180*

[BNS02] Allison Bruce, Illah Nourbakhsh and Reid Simmons. 'The Role of Expressiveness and Attention in Human-Robot Interaction'. In: *International Conference on Robotics and Automation (ICRA)*. IEEE, 2002, pp. 4138–4142. DOI: 10.1109/ROBOT.2002.1014396. *used on: p. 65*

[BR09] Hennie Brugman and Albert Russel. 'Annotating Multi-Media / Multi-Modal Resources with ELAN'. In: *International Conference on Language Resources and Language Evaluation (LREC)*. October. 2009, pp. 2065–2068. *used on: p. 180*

[Bre01] Leo Breiman. 'Random Forests'. In: *Machine Learning* 45.1 (2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. *used on: p. 159*

[Bri92] Robert Bringhurst. *The Elements of Typographic Style*. 3.2. Hartley & Marks, 1992. ISBN: 978-0-88179-206-5. *used on: p. 185*

[BRT02] Leonid Borodulkin, Heinrich Ruser and Hans-Rolf Trankler. '3D virtual "smart home" user interface'. In: *International Symposium on Virtual and Intelligent Measurement Systems* (2002). May. IEEE, 2002, pp. 111–115. DOI: 10.1109/VIMS.2002.1009367. *used on: p. 35*

[Bud+13] Matthias Budde, Matthias Berning, Christopher Baumgärtner, Florian Kinn, Timo Kopf, Sven Ochs et al. 'Point & Control—Interaction in Smart Environments'. In: *Conference on Pervasive and Ubiquitous Computing Adjunct Publication (UbiComp Adjunct)*. New York, NY, USA: ACM Press, 2013, pp. 303–306. DOI: 10.1145/2494091.2494184. *used on: p. 35*

[BV99] Cynthia Breazeal and J Velasquez. 'Robot in Society: Friend or Appliance'. In: *Autonomous Agents Workshop on Emotion-Based Agent Architectures*. 1999, pp. 18–26. *used on: p. 4*

[BVZ01] Yuri Boykov, Olga Veksler and Ramin Zabih. 'Fast Approximate Energy Minimization via Graph Cuts'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23.11 (2001-11), pp. 1222–1239. ISSN: 0162-8828. DOI: 10.1109/34.969114. *used on: p. 180*

[Cab+18] Laura Cabrera-Quiros, Andrew Demetriou, Ekin Gedik, Leander van der Meij and Hayley Hung. 'The MatchNMingle Dataset: a Novel Multi-Sensor Resource for the Analysis of Social Interactions and Group Dynamics In-The-Wild during Free-Standing Conversations and Speed Dates'. In: *IEEE Transactions on Affective Computing* PP.c (2018), pp. 1–1. ISSN: 1949-3045. DOI: 10.1109/TAFFC.2018.2848914. *used on: p. 88*

[Cao+17] Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh. 'Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields'. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017. *used on: p. 180*

[Car+11] Stefano Carrino, Alexandre Péclat, Elena Mugellini, Omar Abou Khaled and Rolf Ingold. 'Humans and Smart Environments: A Novel Multimodal Interaction Approach'. In: *International Conference on Multimodal Interfaces (ICMI)*. New York, New York, USA: ACM Press, 2011, p. 105. DOI: 10.1145/2070481.2070501. *used on: p. 35*

[CD04] Diane Cook and Sajal Das. *Smart Environments - Technology, Protocols, and Applications*. Ed. by Diane J. Cook and Sajal K. Das. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2004-09. ISBN: 978-0-471-68659-0. DOI: 10.1002/047168659X. *used on: pp. 6, 160*

[CK80] T. Matthew Ciolek and Adam Kendon. 'Environment and the Spatial Arrangement of Conversational Encounters'. In: *Sociological Inquiry* 50.3-4 (1980-07), pp. 237–271. ISSN: 0038-0245. DOI: 10.1111/j.1475-682X.1980.tb00022.x. *used on: pp. 15–17, 157, 159, 161*

[CLN16] Seijin Cha, Moon-Hwan Lee and Tek-Jin Nam. 'Gleamy: An Ambient Display Lamp with a Transparency-Controllable Shade'. In: *International Conference on Tangible, Embedded, and Embodied Interaction (TEI)*. New York, New York, USA: ACM Press, 2016, pp. 304–307. DOI: 10.1145/2839462.2839501. *used on: p. 34*

[Coh60] Jacob Cohen. 'A Coefficient of Agreement for Nominal Scales'. In: *Educational and Psychological Measurement* 20.1 (1960-04), pp. 37–46. ISSN: 0013-1644. DOI: 10.1177/001316446002000104. *used on: p. 50*

[CP34] C. J. Clopper and E. S. Pearson. 'The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial'. In: *Biometrika* 26.4 (1934-12), p. 404. ISSN: 00063444. DOI: 10.2307/2331986. *used on: p. 72*

[CRA16] Angelo Cafaro, Brian Ravenet and Vilhj Almsson. 'The Effects of Interpersonal Attitude of a Group of Agents on User's Presence and Proxemics Behavior'. In: *ACM Transactions on Interactive Intelligent Systems* (*TiiS*) 6.2 (2016). *used on: p. 32*

[Cri+11a] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue et al. 'Social Interaction Discovery by Statistical Analysis of F-Formations'. In: *Procedings of the British Machine Vision Conference*. British Machine Vision Association, 2011, pp. 23.1–23.12. DOI: 10.5244/C.25.23. *used on: pp. 30, 88*

[Cri+11b] Marco Cristani, Giulia Paggetti, Alessandro Vinciarelli, Loris Bazzani, Gloria Menegaz and Vittorio Murino. 'Towards Computational Proxemics: Inferring Social Relations from Interpersonal Distances'. In: *International Conference on Privacy, Security, Risk and Trust / International Conference on Social Computing*. IEEE, 2011-10, pp. 290–297. DOI: 10.1109/PASSAT/SocialCom.2011.32. *used on: p. 31*

[CSW14] Birte Carlmeyer, David Schlangen and Britta Wrede. 'Towards Closed Feedback Loops in HRI'. In: *Workshop on Multimodal, Multi-Party, Real-World Human-Robot Interaction* (*MMRWHRI*). New York, New York, USA: ACM Press, 2014, pp. 1–6. DOI: 10.1145/2666499.2666500. *used on: pp. 65, 68*

[CT99] Justine Cassell and Kristinn R. Thórisson. 'The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents'. In: *Applied Artificial Intelligence* 13.4-5 (1999), pp. 519–538. ISSN: 10876545. DOI: 10.1080/088395199117360. *used on: pp. 4, 24*

[Dau+05] Kerstin Dautenhahn, Sarah Woods, Christina Kaouri, Michael L. Walters, Kheng Lee Koay and Iain Werry. 'What is a Robot Companion - Friend, Assistant or Butler?' In: *International Conference on Intelligent Robots and Systems* (*IROS*) (2005), pp. 1488–1493. DOI: 10.1109/IROS.2005.1545189. *used on: pp. 4, 23, 24*

[Dau07] Kerstin Dautenhahn. 'Socially Intelligent Robots: Dimensions of Human–Robot Interaction'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 362.1480 (2007-04), pp. 679–704. ISSN: 0962-8436. DOI: 10.1098/rstb.2006.2004. *used on: pp. 4, 11, 159*

[Dau98] Kerstin Dautenhahn. 'The Art of Designing Socially Intelligent Agents: Science, Fiction, and the Human in the Loop'. In: *Applied Artificial Intelligence* 12.7-8 (1998-10), pp. 573–617. ISSN: 0883-9514. DOI: 10.1080/088395198117550. *used on: p. 21*

[Del+12] Andrew Delong, Anton Osokin, Hossam N. Isack and Yuri Boykov. 'Fast Approximate Energy Minimization with Label Costs'. In: *International Journal of Computer Vision* 96.1 (2012-01), pp. 1–27. ISSN: 0920-5691. DOI: 10.1007/s11263-011-0437-z. *used on: p. 180*

[Dom+16] Jaroslaw Domaszewicz, Spyros Lalis, Aleksander Pruszkowski, Manos Koutsoubelias, Tomasz Tajmajer, Nasos Grigoropoulos et al. 'Soft Actuation: Smart Home and Office with Human-in-the-Loop'. In: *IEEE Pervasive Computing* 15.1 (2016-01), pp. 48–56. ISSN: 1536-1268. DOI: 10.1109/MPRV.2016.5. *used on: p. 34*

[GHG12] Andre Gaschler, Kerstin Huth and M Giuliani. 'Modelling State of Interaction from Head Poses for Social Human-Robot Interaction'. In: *International Conference on Human-Robot Interaction (HRI) Gaze in Human-Robot Interaction Workshop*. Boston, MA: ACM/IEEE, 2012. *used on: p. 15*

[Gla+03] Afina S. Glas, Jeroen G. Lijmer, Martin H. Prins, Gouke J. Bonsel and Patrick M.M. Bossuyt. 'The Diagnostic Odds Ratio: a Single Indicator of Test Performance'. In: *Journal of Clinical Epidemiology* 56.11 (2003-11), pp. 1129–1135. ISSN: 08954356. DOI: 10.1016/S0895-4356(03)00177-X. *used on: pp. xvii, 154*

[Gof63] Erving Goffman. *Behavior in Public Places*. First Free. New York, NY, USA: Free Press, 1963, p. 248. ISBN: 978-0-02-911940-2. *used on: pp. 11, 12, 14, 15, 87, 154, 158*

[Gre+11] Saul Greenberg, Nicolai Marquardt, Rob Diaz-marino and Miaosen Wang. 'Proxemic Interactions: The New Ubicomp?' In: *Interactions* 18.1 (2011), pp. 42–50. *used on: pp. 33, 162*

[Gro+12] H.-M. Gross, Ch Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley et al. 'Further Progress Towards a Home Robot Companion for People with Mild Cognitive Impairment'. In: *International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2012-10, pp. 637–644. DOI: 10.1109/ICSMC.2012.6377798. *used on: p. 37*

[Gro+18] Katharina Groß-Vogt, Marian Weger, Robert Höldrich, Thomas Hermann, Till Bovermann and Stefan Reichmann. 'Augmentation of an Institute's Kitchen: an Ambient Auditory Display of Electric Power Consumption'. In: *International Conference on Auditory Display (ICAD)*. 2018, pp. 105–112. *used on: p. 34*

[Hal69] Edward T. Hall. *The Hidden Dimension: Man's Use of Space in Public and Private*. Bodley Head, 1969. ISBN: 978-0-370-01308-4. *used on: pp. 13, 23, 33, 152, 155, 156, 158–160*

[Hay88] Reiko Hayashi. 'Simultaneous Talk? From the Perspective of Floor Management of English and Japanese Speakers'. In: *World Englishes* 7.3 (1988-11), pp. 269–288. ISSN: 0883-2919. DOI: 10.1111/j.1467-971X.1988.tb00237.x. *used on: pp. 18, 37, 153*

[HBN11] Hung-Hsuan Huang, Naoya Baba and Yukiko Nakano. 'Making Virtual Conversational Agent Aware of the Addressee of Users' Utterances in Multi-User Conversation using Nonverbal Information'. In: *International Conference on Multimodal Interfaces (ICMI)*. New York, New York, USA: ACM Press, 2011, p. 401. DOI: 10.1145/2070481.2070557. *used on: p. 27*

[HE10] Mattias Heldner and Jens Edlund. 'Pauses, Gaps and Overlaps in Conversations'. In: *Journal of Phonetics* 38.4 (2010-10), pp. 555–568. ISSN: 00954470. DOI: 10.1016/j.wocn.2010.08.002. arXiv: arXiv:1011.1669v3. *used on: p. 20*

[HEC14] Hayley Hung, Gwenn Englebienne and Laura Cabrera Quiros. 'Detecting Conversing Groups With a Single Worn Accelerometer'. In: *International Conference on Multimodal Interaction (ICMI)*. New York, New York, USA: ACM Press, 2014, pp. 84–91. DOI: 10.1145/2663204.2663228. *used on: p. 31*

[Heg+08] Frank Hegel, Soren Krach, Tilo Kircher, Britta Wrede and Gerhard Sagerer. 'Understanding Social Robots: A User Study on Anthropomorphism'. In: *International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2008-08, pp. 574–579. DOI: 10.1109/ROMAN.2008.4600728. *used on: p. 5*

[HK11] Hayley Hung and Ben Kröse. 'Detecting F-Formations as Dominant Sets'. In: *International Conference on Multimodal Interfaces (ICMI)*. New York, New York, USA: ACM Press, 2011, p. 231. DOI: 10.1145/2070481.2070525. *used on: p. 88*

[HK15] Judith Holler and Kobin H. Kendrick. 'Unaddressed Participants' Gaze in Multi-Person Interaction: Optimizing Recipiency'. In: *Frontiers in Psychology* 6.FEB (2015-02), pp. 1–14. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2015.00098. *used on: p. 20*

[HM16] Chien-Ming Huang and Bilge Mutlu. 'Anticipatory Robot Control for Efficient Human-Robot Collaboration'. In: *International Conference on Human-Robot Interaction (HRI)*. Section V. ACM/IEEE, 2016-03, pp. 83–90. DOI: 10.1109/HRI.2016.7451737. *used on: pp. 24, 65*

[Hol14] Patrick Holthaus. 'Approaching Human-Like Spatial Awareness in Social Robotics - An Investigation of Spatial Interaction Strategies with a Receptionist Robot'. Doctoral dissertation. Bielefeld University, 2014. *used on: pp. 22, 65, 185*

[HS97] Sepp Hochreiter and Jürgen Schmidhuber. 'Long Short-Term Memory'. In: *Neural Computation* 9.8 (1997-11), pp. 1735–1780. ISSN: 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. *used on: pp. 117, 119, 157*

[HTS09] Helge Hüttenrauch, Elin A. Topp and Kerstin Severinson-Eklundh. 'The Art of Gate-Crashing: Bringing HRI Into Users' Homes'. In: *Interaction Studies* 10.3 (2009-12), pp. 274–297. ISSN: 1572-0373. DOI: 10.1075/is.10.3.02hut. *used on: p. 26*

[JAN06] Natasa Jovanovic, Rieks op den Akker and Anton Nijholt. 'Addressee Identification in Face-to-Face Meetings'. In: *Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (2006), pp. 169–176. *used on: pp. 26, 88*

[Jay+13] Dinesh Babu Jayagopi, Samira Sheiki, David Klotz, Johannes Wienke, Jean-Marc Odobez, Sebastien Wrede et al. 'The Vernissage Corpus: A Conversational Human-Robot-Interaction Dataset'. In: *International Conference on Human-Robot Interaction (HRI)*. ACM/IEEE, 2013-03, pp. 149–150. DOI: 10.1109/HRI.2013.6483545. *used on: pp. 28, 88*

[Joo+14] Michiel P. Joosse, Ronald W. Poppe, Manja Lohse and Vanessa Evers. 'Cultural Differences in How an Engagement-Seeking Robot Should Approach a Group of People'. In: *International Conference on Collaboration Across Boundaries: Culture, Distance & Technology (CABS)*. New York, New York, USA: ACM Press, 2014, pp. 121–130. DOI: 10.1145/2631488.2631499. *used on: p. 25*

[JS15] Martin Johansson and Gabriel Skantze. 'Opportunities and Obligations to Take Turns in Collaborative Multi-Party Human-Robot Interaction'. In: *Annual Meeting of the Special Interest Group on Discourse and Dialogue*. September. 2015, pp. 305–314. *used on: p. 29*

[Ken67] Adam Kendon. 'Some Functions of Gaze-Direction in Social Interaction'. In: *Acta Psychologica* 26 (1967), pp. 22–63. ISSN: 00016918. DOI: 10.1016/0001-6918(67)90005-4. *used on: pp. 20, 65*

[Ken90] Adam Kendon. *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Ed. by John J. Gumperz. Studies in. Studies in Interactional Socio. Cambridge: Press Syndicate of the University of Cambridge, 1990. ISBN: 978-0-521-38938-9. *used on: pp. 14, 15, 17, 18, 43, 155*

[KF10] Kay Kitazawa and Taku Fujiyama. 'Pedestrian Vision and Collision Avoidance Behavior: Investigation of the Information Process Space of Pedestrians Using an Eye Tracker'. In: *Pedestrian and Evacuation Dynamics*. Ed. by Wolfram W. F. Klingsch, Christian Rogsch, Andreas Schadschneider and Michael Schreckenberg. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 95–108. ISBN: 978-3-642-04503-5. DOI: 10.1007/978-3-642-04504-2_7. *used on: p. 156*

[KK06] Daehwan Kim and Daijin Kim. 'An Intelligent Smart Home Control Using Body Gestures'. In: *International Conference on Hybrid Information Technology*. IEEE, 2006-11, pp. 439–446. DOI: 10.1109/ICHIT.2006.253644. *used on: p. 35*

[Kop+18] Stefan Kopp, Katharina Cyra, Franz Kummert, Lars Schillingmann, Mara Brandt, Farina Freigang et al. 'Conversational Assistants for Elderly Users – The Importance of Socially Cooperative Dialogue'. In: *AAMAS Workshop on Intelligent Conversation Agents in Home and Geriatric Care Applications co-located with the Federated AI Meeting*. Stockholm, Sweden, 2018, pp. 10–17. *used on: p. 24*

[KSS04] Michael Katzenmaier, Rainer Stiefelhagen and Tanja Schultz. 'Identifying the Addressee in Human-Human-Robot Interactions Based on Head Pose and Speech'. In: *International Conference on Multimodal Interfaces (ICMI)*. New

York, New York, USA: ACM Press, 2004, p. 144. DOI: 10.1145/1027933.1027959. *used on: p. 28*

[Küh+11] Christine Kühnel, Tilo Westermann, Fabian Hemmert, Sven Kratz, Alexander Müller and Sebastian Möller. 'I'm Home: Defining and Evaluating a Gesture Set for Smart-Home Control'. In: *International Journal of Human-Computer Studies* 69.11 (2011-10), pp. 693–704. ISSN: 10715819. DOI: 10.1016/j.ijhcs.2011.04.005. *used on: p. 36*

[Kuz+10] Hideaki Kuzuoka, Yuya Suzuki, Jun Yamashita and Keiichi Yamazaki. 'Reconfiguring Spatial Formation Arrangement by Robot Body Orientation'. In: *International Conference on Human-Robot Interaction* (*HRI*). New York, New York, USA: ACM Press, 2010, p. 285. DOI: 10.1145/1734454.1734557. *used on: p. 26*

[KZ04] V. Kolmogorov and R. Zabih. 'What Energy Functions Can be Minimized via Graph Cuts?' In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2 (2004-02), pp. 147–159. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2004.1262177. *used on: p. 180*

[Lan+03] Sebastian Lang, Marcus Kleinehagenbrock, Sascha Hohenner, Jannik Fritsch, Gernot a Fink and Gerhard Sagerer. 'Providing the Basis for Human-Robot-Interaction: A Multi-Modal Attention System for a Mobile Robot'. In: *International Conference on Multimodal Interfaces* (*ICMI*). New York, New York, USA: ACM Press, 2003, p. 28. DOI: 10.1145/958432.958441. *used on: pp. 28, 66*

[LE11] Felix Lindner and Carola Eschenbach. 'Towards a Formalization of Social Spaces for Socially Aware Robots'. In: *Lecture Notes in Computer Science*. Ed. by Max Egenhofer, Nicholas Giudice, Reinhard Moratz and Michael Worboys. 2011, pp. 283–303. ISBN: 978-3-642-24263-2. DOI: 10.1007/978-3-642-23196-4_16. arXiv: 9780201398298. *used on: pp. 23, 151, 158, 160*

[Lei+16] Christian Leichsenring, Jiajun Yang, Jan Hammerschmidt and Thomas Hermann. 'Challenges for Smart Environments in Bathroom Contexts'. In: *Workshop on Embodied Interaction with Smart Environments* (*EISE*). Tokyo, Japan: ACM Press, 2016, pp. 1–7. DOI: 10.1145/3008028.3008033. *used on: p. 34*

[LIK18] Divesh Lala, Koji Inoue and Tatsuya Kawahara. 'Evaluation of Real-time Deep Learning Turn-Taking Models for Multiple Dialogue Scenarios'. In: *International Conference on Multimodal Interaction* (*ICMI*). New York, New York, USA: ACM Press, 2018, pp. 78–86. DOI: 10.1145/3242969.3242994. *used on: p. 29*

[Loh+12] Katrin S. Lohan, Katharina J. Rohlfing, Karola Pitsch, Joe Saunders, Hagen Lehmann, Chrystopher L. Nehaniv et al. 'Tutor Spotter: Proposing a Feature Set and Evaluating It in a Robotic System'. In: *International Journal of Social Robotics*

4.2 (2012-04), pp. 131–146. ISSN: 1875-4791. DOI: 10.1007/s12369-011-0125-8. *used on: p. 4*

[Lor+13] Bazzani Loris, Cristani Marco, Tosato Diego, Farenzena Michela, Paggetti Giulia, Menegaz Gloria et al. 'Social Interactions by Visual Focus of Attention in a Three-Dimensional Environment'. In: *Expert Systems* 30.2 (2013-05), pp. 115–127. ISSN: 02664720. DOI: 10.1111/j.1468-0394.2012.00622.x. *used on: p. 88*

[LSW14] Florian Lier, Simon Schulz and Sven Wachsmuth. 'Reality check! - A Physical Robot Versus its Simulation'. In: *International Conference on Human-Robot Interaction (HRI)*. New York, New York, USA: ACM Press, 2014, pp. 331–331. DOI: 10.1145/2559636.2559787. *used on: pp. 7, 155*

[Lüt+10] Ingo Lütkebohle, Frank Hegel, Simon Schulz, Matthias Hackel, Britta Wrede, Sven Wachsmuth et al. 'The Bielefeld Anthropomorphic Robot Head Flobi'. In: *International Conference on Robotics and Automation (ICRA)*. IEEE, 2010-05, pp. 3384–3391. DOI: 10.1109/ROBOT.2010.5509173. *used on: pp. 7, 155*

[LW18] Irene Lopatovska and Harriet Williams. 'Personification of the Amazon Alexa'. In: *Conference on Human Information Interaction & Retrieval (CHIIR)*. New York, New York, USA: ACM Press, 2018, pp. 265–268. DOI: 10.1145/3176349.3176868. *used on: p. 3*

[Mar+11] Nicolai Marquardt, Robert Diaz-Marino, Sebastian Boring and Saul Greenberg. 'The Proximity Toolkit: Prototyping Proxemic Interactions in Ubiquitous Computing Ecologies'. In: *ACM Symposium on User Interface Software and Technology (UIST)*. New York, New York, USA: ACM Press, 2011, p. 315. DOI: 10.1145/2047196.2047238. *used on: p. 33*

[Mat+15] Yoichi Matsuyama, Iwao Akiba, Shinya Fujie and Tetsunori Kobayashi. 'Four-Participant Group Conversation: A Facilitation Robot Controlling Engagement Density as the Fourth Participant'. In: *Computer Speech & Language* 33.1 (2015-09), pp. 1–24. ISSN: 08852308. DOI: 10.1016/j.csl.2014.12.001. *used on: p. 25*

[Mey18] Christian Meyer. *Culture, Practice, and the Body*. Stuttgart: J.B. Metzler, 2018. ISBN: 978-3-476-04605-5. DOI: 10.1007/978-3-476-04606-2. *used on: p. 38*

[MS14] Simon Mayer and Gabor Soros. 'User Interface Beaming – Seamless Interaction with Smart Things Using Personal Wearable Computers'. In: *International Conference on Wearable and Implantable Body Sensor Networks Workshops*. IEEE, 2014-06, pp. 46–49. DOI: 10.1109/BSN.Workshops.2014.17. *used on: p. 35*

[MTH18] Takashi Makino, Yoshinari Takegawa and Keiji Hirata. 'Predicting Turn Taking from Gaze Transition Patterns Considering Participation Status in Multi-Party Conversation'. In: *International Conference on Intelligent Automation and Robotics (ICIAR)*. 2018. *used on: p. 88*

[Mut+09] Bilge Mutlu, Toshiyuki Shiwa, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. 'Footing in Human-Robot Conversations'. In: *International Conference on Human-Robot Interaction (HRI)*. 1. New York, New York, USA: ACM Press, 2009, p. 61. DOI: 10.1145/1514095.1514109. *used on: pp. 4, 25*

[Neu+17] Alexander Neumann, Christof Elbrechter, Nadine Pfeiffer-Leßmann, Risto Kõiva, Birte Carlmeyer, Stefan Rüther et al. 'KogniChef: A Cognitive Cooking Assistant'. In: *Künstliche Intelligenz (KI)* 31.3 (2017-08), pp. 273–281. ISSN: 0933-1875. DOI: 10.1007/s13218-017-0488-6. *used on: p. 35*

[NST94] Clifford Nass, Jonathan Steuer and Ellen R. Tauber. 'Computers are Social Actors'. In: *Companion of Human Actors in Computing Systems (CHI)*. June 2014. New York, New York, USA: ACM Press, 1994, p. 204. DOI: 10.1145/259963.260288. arXiv: 1607.05174. *used on: p. 3*

[Par+07] Kwang-Hyun Park, Zeungnam Bien, Ju-Jang Lee, Byung Kook Kim, Jong-Tae Lim, Jin-Oh Kim et al. 'Robotic Smart House to Assist People with Movement Disabilities'. In: *Autonomous Robots* 22.2 (2007-01), pp. 183–198. ISSN: 0929-5593. DOI: 10.1007/s10514-006-9012-9. *used on: p. 37*

[Par+08] Kwang-hyun Park, Hyong-euk Lee, Youngmin Kim and Z. Zenn Bien. 'A Steward Robot for Human-Friendly Human-Machine Interaction in a Smart House Environment'. In: *IEEE Transactions on Automation Science and Engineering* 5.1 (2008-01), pp. 21–25. ISSN: 1545-5955. DOI: 10.1109/TASE.2007.911674. *used on: p. 37*

[PGM17] Tomislav Pejsa, Michael Gleicher and Bilge Mutlu. 'Who, Me? How Virtual Agents Can Shape Conversational Footing in Virtual Reality'. In: *International Conference on Intelligent Virtual Agents*. 2017, pp. 347–359. DOI: 10.1007/978-3-319-67401-8_45. *used on: p. 25*

[Piz+18] Simone Pizzagalli, Daniele Spoladore, Sara Arlati, Marco Sacco and Luca Greci. 'HIC: An Interactive and Ubiquitous Home Controller System for the Smart Home'. In: *International Conference on Serious Games and Applications for Health (SeGAH)*. IEEE, 2018-05, pp. 1–6. DOI: 10.1109/SeGAH.2018.8401374. *used on: p. 35*

[PLH19] Marian Pohling, Christian Leichsenring and Thomas Hermann. 'Base Cube One: A Location-Addressable Service-Oriented Smart Environment Framework'. In: *Journal of Ambient Intelligence and Smart Environments* 11.5 (2019-09), pp. 373–401. ISSN: 18761372. DOI: 10.3233/AIS-190533. *used on: p. 35*

[Por+13] François Portet, Michel Vacher, Caroline Golanski, Camille Roux and Brigitte Meillon. 'Design and Evaluation of a Smart Home Voice Interface for the Elderly: Acceptability and Objection Aspects'. In: *Personal and Ubiquitous Computing* 17.1 (2013-01), pp. 127–144. ISSN: 1617-4909. DOI: 10.1007/s00779-011-0470-5. *used on: p. 36*

[Pot+03] Ilyas Potamitis, K. Georgila, Nikos Fakotakis and George Kokkinakis. 'An Integrated System for Smart-Home Control of Appliances Based on Remote Speech Interaction'. In: *European Conference on Speech Communication and Technology* (2003). *used on: p. 36*

[Pow11] David Powers. 'Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation'. In: *Journal of Machine Learning Technologies* 2.1 (2011), pp. 37–63. *used on: pp. xvii, 156, 157*

[RAN05] Matthias Rehm, Elisabeth André and Michael Nischt. 'Lets Come Together – Social Navigation Behaviors of Virtual and Real Humans'. In: *Intelligent Technologies for Interactive Entertainment*. Ed. by Mark Maybury, Oliviero Stock and Wolfgang Wahlster. Vol. 3814 LNAI. 2005, pp. 336–336. ISBN: 978-3-540-30509-5. DOI: 10.1007/11590323_47. *used on: p. 32*

[RBL09] Federico Rossano, Penelope Brown and Stephen C. Levinson. 'Gaze, Questioning, and Culture'. In: *Conversation Analysis*. Ed. by Jack Sidnell. Rossano. Cambridge: Cambridge University Press, 2009, pp. 187–249. DOI: 10.1017/CBO9780511635670.008. *used on: p. 37*

[RGS18] Ely Repiso, Anais Garrell and Alberto Sanfeliu. 'Robot Approaching and Engaging People in a Human-Robot Companion Framework'. In: *International Conference on Intelligent Robots and Systems (IROS)*. IEEE/RSJ, 2018-10, pp. 8200–8205. DOI: 10.1109/IROS.2018.8594149. *used on: p. 32*

[Ric+15] Elisa Ricci, Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulo, Narendra Ahuja and Oswald Lanz. 'Uncovering Interactions and Interactors: Joint Estimation of Head, Body Orientation and F-Formations from Surveillance Videos'. In: *International Conference on Computer Vision (ICCV)*. IEEE, 2015-12, pp. 4660–4668. DOI: 10.1109/ICCV.2015.529. *used on: p. 31*

[Rio+12] Jorge Rios-Martinez, Alessandro Renzaglia, Anne Spalanzani, Agostino Martinelli and Christian Laugier. 'Navigating Between People: A Atochastic Optimization Approach'. In: *International Conference on Robotics and Automation (ICRA)*. IEEE, 2012-05, pp. 2880–2885. DOI: 10.1109/ICRA.2012.6224934. *used on: p. 23*

[RKH18] Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer and Jonathan Herrmann. 'The Effects of Humanlike and Robot-Specific Affective Nonverbal Behavior on Perception, Emotion, and Behavior'. In: *International Journal of Social Robotics* 10.5 (2018-11), pp. 569–582. ISSN: 1875-4791. DOI: 10.1007/s12369-018-0466-7. *used on: p. 24*

[RN96] Byron Reeves and Clifford Nass. *The Media Equation - How People Treat Computers, Television, and New Media Like Real People and Places*. New York, NY, US: Cambridge University Press, 1996. ISBN: 978-1-57586-053-4. *used on: pp. 3, 4, 11, 160*

[RSL11] Jorge Rios-Martinez, Anne Spalanzani and Christian Laugier. 'Understanding Human Interaction for Probabilistic Autonomous Navigation using Risk-RRT Approach'. In: *International Conference on Intelligent Robots and Systems* (*IROS*). IEEE/RSJ, 2011-09, pp. 2014–2019. DOI: 10.1109/IROS.2011.6048137. *used on: p. 31*

[RSL15] Jorge Rios-Martinez, Anne Spalanzani and Christian Laugier. 'From Proxemics Theory to Socially-Aware Navigation: A Survey'. In: *International Journal of Social Robotics* 7.2 (2015-04), pp. 137–153. ISSN: 1875-4791. DOI: 10.1007/s12369-014-0251-1. *used on: p. 23*

[Ruh+15] Kerstin Ruhland, Christopher E. Peters, Sean Andrist, Jeremy B. Badler, Norman Ira Badler, Michael L. Gleicher et al. 'A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception'. In: *Computer Graphics Forum* 34.6 (2015-09), pp. 299–326. ISSN: 01677055. DOI: 10.1111/cgf.12603. *used on: p. 25*

[Sag+13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou and Maja Pantic. '300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge'. In: *International Conference on Computer Vision Workshops* (*ICCV*). IEEE, 2013-12, pp. 397–403. DOI: 10.1109/ICCVW.2013.59. *used on: p. 69*

[San+17] Frode Eika Sandnes, Jo Herstad, Andrea Marie Stangeland and Fausto Orsi Medola. 'UbiWheel: A Simple Context-Aware Universal Control Concept for Smart Home Appliances that Encourages Active Living'. In: *SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation* (*SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI*). c. IEEE, 2017-08, pp. 1–6. DOI: 10.1109/UIC-ATC.2017.8397460. *used on: p. 34*

[Sar+12] Aziez Sardar, Michiel Joosse, Astrid Weiss and Vanessa Evers. 'Don't Stand so Close to Me: Users' Attitudinal and Behavioral Responses to Personal Space Invasion by Robots'. In: *International Conference on Human-Robot Interaction* (*HRI*). January. New York, NY, USA: ACM Press, 2012, p. 229. DOI: 10.1145/2157689.2157769. *used on: p. 22*

[Sat+09] Satoru Satake, Takayuki Kanda, Dylan F. Glas, Michita Imai, Hiroshi Ishiguro and Norihiro Hagita. 'How to Approach Humans? – Strategies for Social Robots to Initiate Interaction'. In: *International Conference on Human Robot Interaction* (*HRI*). New York, New York, USA: ACM Press, 2009, p. 109. DOI: 10.1145/1514095.1514117. *used on: p. 22*

[SBW19] Simon Schulz, Sebastian Meyer Zu Borgsen and Sven Wachsmuth. 'See and Be Seen – Rapid and Likeable High-Definition Camera-Eye for Anthropomorphic Robots'. In: *International Conference on Robotics and Automation* (*ICRA*).

IEEE, 2019-05, pp. 2524–2530. DOI: 10.1109/ICRA.2019.8794319. *used on: pp. 7, 155, 156*

[Sch+03] H. Schulzrinne, S. Casner, R. Frederick and V. Jacobson. *RTP: A Transport Protocol for Real-Time Applications*. STD 64. RFC Editor, 2003. URL: http://www.rfc-editor.org/rfc/rfc3550.txt. *used on: p. 8*

[Sch68] Emanuel A. Schegloff. 'Sequencing in Conversational Openings'. In: *American Anthropologist* 70.6 (1968-12), pp. 1075–1095. ISSN: 0002-7294. DOI: 10.1525/aa.1968.70.6.02a00030. *used on: p. 12*

[Sei+09] Thomas Seifried, Michael Haller, Stacey D. Scott, Florian Perteneder, Christian Rendl, Daisuke Sakamoto et al. 'CRISTAL: A Collaborative Home Media and Device Controller Based on a Multi-touch Display'. In: *ACM International Conference on Interactive Tabletops and Surfaces (ITS)*. New York, New York, USA: ACM Press, 2009, p. 33. DOI: 10.1145/1731903.1731911. *used on: p. 35*

[Set+13] Francesco Setti, Oswald Lanz, Roberta Ferrario, Vittorio Murino and Marco Cristani. 'Multi-Scale F-Formation Discovery for Group Detection'. In: *International Conference on Image Processing*. IEEE, 2013-09, pp. 3547–3551. DOI: 10.1109/ICIP.2013.6738732. *used on: p. 88*

[Set+15] Francesco Setti, Chris Russell, Chiara Bassetti and Marco Cristani. 'F-Formation Detection: Individuating Free-Standing Conversational Groups in Images'. In: *PLOS ONE* 10.5 (2015-05). Ed. by Rongrong Ji, e0123783. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0123783. arXiv: arXiv:1409.2702v1. *used on: pp. 30, 88, 95–99, 103, 160, 161*

[Seu+07] Seung-Ho Baeg, Jae-Han Park, Jaehan Koh, Kyung-Wook Park and Moon-Hong Baeg. 'Building a Smart Home Environment for Service Robots Based on RFID and Sensor Networks'. In: *International Conference on Control, Automation and Systems*. IEEE, 2007, pp. 1078–1082. DOI: 10.1109/ICCAS.2007.4407059. *used on: p. 37*

[She14] Samira Sheikhi. 'Inferring Visual Attention and Addressee in Human Robot Interaction'. Doctoral dissertation. École Polytechnique Fédérale de Lausanne (EPFL), 2014, p. 132. *used on: p. 28*

[Shi+11] Chao Shi, Michihiro Shimada, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. 'Spatial Formation Model for Initiating Conversation'. In: *Robotics: Science and Systems VII*. May 2014. Robotics: Science and Systems Foundation, 2011-06. DOI: 10.15607/RSS.2011.VII.039. *used on: p. 22*

[Shi+15] Chao Shi, Masahiro Shiomi, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. 'Measuring Communication Participation to Initiate Conversation in Human–Robot Interaction'. In: *International Journal of Social Robotics* 7.5 (2015-11), pp. 889–910. ISSN: 1875-4791. DOI: 10.1007/s12369-015-0285-z. *used on: p. 32*

[SHS07] Thorsten P. Spexard, Marc Hanheide and Gerhard Sagerer. 'Human-Oriented Interaction With an Anthropomorphic Robot'. In: *IEEE Transactions on Robotics* 23.5 (2007-10), pp. 852–862. ISSN: 1552-3098. DOI: 10.1109/TRO.2007.904903. *used on: p. 23*

[SJB15] Gabriel Skantze, Martin Johansson and Jonas Beskow. 'Exploring Turn-taking Cues in Multi-party Human-Robot Discussions about Objects'. In: *International Conference on Multimodal Interaction (ICMI)*. New York, New York, USA: ACM Press, 2015, pp. 67–74. DOI: 10.1145/2818346.2820749. *used on: pp. 28, 29, 66, 88*

[SN15] Lars Schillingmann and Yukie Nagai. 'Yet Another Gaze Detector: An Embodied Calibration Free System for the iCub Robot'. In: *International Conference on Humanoid Robots (Humanoids)*. IEEE-RAS, 2015-11, pp. 8–13. DOI: 10.1109/HUMANOIDS.2015.7363515. *used on: p. 69*

[Sør+13] Henrik Sørensen, Mathies G. Kristensen, Jesper Kjeldskov and Mikael B. Skov. 'Proxemic Interaction in a Multi-Room Music System'. In: *Australian Computer-Human Interaction Conference on Augmentation, Application, Innovation, Collaboration (OzCHI)*. New York, New York, USA: ACM Press, 2013, pp. 153–162. DOI: 10.1145/2541016.2541046. *used on: p. 33*

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever and Ruslan Salakhutdinov. 'Dropout: A Simple Way to Prevent Neural Networks from Overfitting'. In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. *used on: p. 120*

[SSJ78] Harvey Sacks, Emanuel A. Schegloff and Gail Jefferson. 'A Simplest Systematics for the Organization of Turn Taking for Conversation'. In: *Studies in the Organization of Conversational Interaction*. Ed. by Jim Schenkein. Vol. 84. 3. Elsevier, 1978, pp. 7–55. ISBN: 978-0-16-048774-3. DOI: 10.1016/B978-0-12-623550-0.50008-2. *used on: pp. 18, 19*

[TO06] Yoshinao Takemae and Shinji Ozawa. 'Automatic Addressee Identification Based on Participants' Head Orientation and Utterances for Multiparty Conversations'. In: *International Conference on Multimedia and Expo*. IEEE, 2006-07, pp. 1285–1288. DOI: 10.1109/ICME.2006.262773. *used on: p. 27*

[TP09] Leila Takayama and Caroline Pantofaru. 'Influences on Proxemic Behaviors in Human-Robot Interaction'. In: *International Conference on Intelligent Robots and Systems (IROS)*. NOVEMBER 2009. IEEE/RSJ, 2009-10, pp. 5495–5502. DOI: 10.1109/IROS.2009.5354145. *used on: p. 22*

[Tra04] David Traum. 'Issues in Multiparty Dialogues'. In: *Workshop on Agent Communication Languages (ACL)*. Ed. by Frank Dignum. 2004, pp. 201–211. ISBN: 978-3-540-20769-6. DOI: 10.1007/978-3-540-24608-4_12. *used on: p. 18*

[Tur+05] Koen van Turnhout, Jacques Terken, Ilse Bakx and Berry Eggen. 'Identifying the Intended Addressee in Mixed Human-Human and Human-Computer Interaction from Non-Verbal Features'. In: *International Conference on Multimodal Interfaces* (*ICMI*). May 2014. New York, New York, USA: ACM Press, 2005, p. 175. DOI: 10.1145/1088463.1088495. *used on: pp. 27, 28, 88*

[Var+18] Jagannadan Varadarajan, Ramanathan Subramanian, Samuel Rota Bulò, Narendra Ahuja, Oswald Lanz and Elisa Ricci. 'Joint Estimation of Human Pose and Conversational Groups from Social Scenes'. In: *International Journal of Computer Vision* 126.2-4 (2018-04), pp. 410–429. ISSN: 0920-5691. DOI: 10.1007/s11263-017-1026-6. *used on: p. 31*

[Vas+16] Sebastiano Vascon, Eyasu Z. Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo and Vittorio Murino. 'Detecting Conversational Groups in Images and Sequences: A Robust Game-Theoretic Approach'. In: *Computer Vision and Image Understanding* 143 (2016-02), pp. 11–24. ISSN: 10773142. DOI: 10.1016/j.cviu.2015.09.012. *used on: pp. 88, 103*

[Ver+01] Roel Vertegaal, Robert Slagter, Gerrit van der Veer and Anton Nijholt. 'Eye Gaze Patterns in Conversations: There is More to Conversational Ggents than Meets the Eyes'. In: *Conference on Human Gactors in Computing Systems* (*CHI*). New York, New York, USA: ACM Press, 2001, pp. 301–308. DOI: 10.1145/365024.365119. arXiv: 01/0003 [1-58113-327-8]. *used on: pp. 20, 28*

[Ver+17] David Verweij, Augusto Esteves, Vassilis-Javed Khan and Saskia Bakker. 'Smart Home Control using Motion Matching and Smart Watches'. In: *International Conference on Interactive Surfaces and Spaces* (*ISS*). New York, New York, USA: ACM Press, 2017, pp. 466–468. DOI: 10.1145/3132272.3132283. *used on: p. 36*

[VKH13] Jan Vanus, Jiri Koziorek and Radim Hercik. 'The Design of the Voice Communication in Smart Home Care'. In: *International Conference on Telecommunications and Signal Processing* (*TSP*). IEEE, 2013-07, pp. 561–564. DOI: 10.1109/TSP.2013.6613996. *used on: p. 37*

[Wac+17] Sven Wachsmuth, Florian Lier, Sebastian Meyer zu Borgsen, Johannes Kummert, Luca Lach and Dominik Sixt. 'ToBI - Team of Bielefeld The Human-Robot Interaction System for RoboCup@Home 2017'. In: *RoboCup 2017*. Nagoya, 2017. *used on: pp. 7, 156*

[WBG12] Miaosen Wang, Sebastian Boring and Saul Greenberg. 'Proxemic Peddlerr: A Public Advertising Display that Captures and Preserves the Attention of a Passerby'. In: *International Symposium on Pervasive Displays* (*PerDis*). 3. New York, New York, USA: ACM Press, 2012, pp. 1–6. DOI: 10.1145/2307798.2307801. *used on: p. 33*

[Wei+16] Shih-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh. 'Convolutional Pose Machines'. In: *Computer Vision and Pattern Recognition (CVPR)*. 2016. *used on: p. 180*

[Wie18] Johannes Wienke. 'Framework-Level Resource Awareness In Robotics and Intelligent Systems'. Doctoral dissertation. Bielefeld University, 2018. DOI: 10.4119/unibi/2932136. *used on: p. 185*

[Wre+17] Sebastian Wrede, Christian Leichsenring, Patrick Holthaus, Thomas Hermann and Sven Wachsmuth. 'The Cognitive Service Robotics Apartment'. In: *Künstliche Intelligenz (KI)* 31.3 (2017-08), pp. 299–304. ISSN: 0933-1875. DOI: 10.1007/s13218-017-0492-x. *used on: pp. 6, 152*

[WW11] Johannes Wienke and Sebastian Wrede. 'A Middleware for Collaborative Research in Experimental Robotics'. In: *International Symposium on System Integration (SII)*. IEEE, 2011-12, pp. 1183–1190. DOI: 10.1109/SII.2011.6147617. *used on: p. 181*

[Yam+10] Fumitaka Yamaoka, Takayuki Kanda, Hiroshi Ishiguro and Norihiro Hagita. 'A Model of Proximity Control for Information-Presenting Robots'. In: *IEEE Transactions on Robotics* 26.1 (2010-02), pp. 187–195. ISSN: 1552-3098. DOI: 10.1109/TRO.2009.2035747. *used on: p. 32*

[Żar19] Mateusz Żarkowski. 'Multi-party Turn-Taking in Repeated Human–Robot Interactions: An Interdisciplinary Evaluation'. In: *International Journal of Social Robotics* (2019-11). ISSN: 1875-4791. DOI: 10.1007/s12369-019-00603-1. *used on: p. 30*

[ZH16] Lu Zhang and Hayley Hung. 'Beyond F-Formations: Determining Social Involvement in Free Standing Conversing Groups from Static Images'. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016-06, pp. 1086–1095. DOI: 10.1109/CVPR.2016.123. *used on: p. 31*

[ZZC16] Yu Zhao, Xihui Zhang and John Crabtree. 'Human-Computer Interaction and User Experience in Smart Home Research: A Critical Analysis'. In: *Issues in Information Systems* 17.3 (2016), pp. 11–19. *used on: p. 37*

SOFTWARE PACKAGES

[base] R Core Team and contributors worldwide. *base*. URL: https://stat.ethz.ch/R-manual/R-devel/library/base/html/00Index.html (visited on 2019-12-11). *used on: p. 118*

[BComfy] Julian Daberkow and Marian Pohling. *Augmented Reality Smart Home Control App*. URL: https://github.com/openbase/bco.bcomfy (visited on 2019-11-06). *used on: p. 35*

[BCozy] Marian Pohling, Julian Daberkow, Lili Schroeder, Hendrick Oestreich, Andreas Gatting, Timo Michalski et al. *BCozy - A Location Based Smart Home User Interface*. URL: http://bcozy.org (visited on 2019-11-06). *used on: p. 35* [bnlearn] Marco Scutari and Robert Ness. *bnlearn: Bayesian Network Structure Learning, Parameter Learning and Inference*. URL: https://cran.r-project.org/package=bnlearn (visited on 2019-12-11). *used on: pp. 59, 112*

[classicthesis] André Miede. *classicthesis*. URL: https://ctan.org/pkg/classicthesis (visited on 2019-12-05). *used on: p. 185*

[ELAN] Birgit Hellwig. *ELAN Linguistic Annotator*. URL: https://tla.mpi.nl/tools/tla-tools/elan/ (visited on 2019-07-04). Presented in [BR09]. *used on: pp. 48, 71, 139*

[ffm] Viktor Richter. *fformation*. URL: https://github.com/vrichter/fformation (visited on 2019-07-04). *used on: pp. 97, 99*

[ffm-gco] Viktor Richter. *fformation-gco*. URL: https://github.com/vrichter/fformation-gco (visited on 2019-07-04). *used on: p. 99*

[GCFF] Francesco Setti. *Graph-Cuts for F-Formation*. URL: https://github.com/franzsetti/GCFF (visited on 2019-07-04). *used on: p. 99*

[gco-v3.0] Nuno Subtil and Viktor Richter. *GCoptimization - Software for Energy Minimization with Graph Cuts Version 3.0*. URL: https://github.com/vrichter/gco-v3.0 (visited on 2019-07-04). Presented in [BVZ01; KZ04; BK04; Del+12]. *used on: p. 99*

[gstreamer] Wim Taymans, @thomasvs, Tim-Philipp Müller, Sebastian Dröge, Stefan Sauer, David Schleef et al. *GStreamer*. URL: https://gstreamer.freedesktop.org/ (visited on 2019-11-11). *used on: p. 8*

[hun] Justin Buchanan and Viktor Richter. *C++ Implementation of the Hungarian Algorithm*. URL: https://github.com/vrichter/hungarian (visited on 2019-07-04). *used on: p. 93*

[Keras] François Chollet, Fariz Rahman, Taehoon Lee, Gabriel de Marmiesse, Oleg Zabluda, Max Pumperla et al. *Keras: The Python Deep Learning library*. URL: https://keras.io/ (visited on 2019-10-07). *used on: p. 118*

[kerasR] Andrie de Vries and Taylor Arnold. *kerasR: R Interface to the Keras Deep Learning Library*. URL: https://github.com/statsmaths/kerasR (visited on 2019-10-07). *used on: p. 118*

[mlr] Bernd Bischl, Michel Lang, Lars Kotthoff, Julia Schiffner, Jakob Richter, Zachary Jones et al. *mlr: Machine Learning in R*. URL: https://cran.r-project.org/package=mlr (visited on 2019-12-11). *used on: p. 60*

[OpenPose] Gines, Raaj, Bikramjot Hanzra, Zhe Cao, Tianyi Zhao, Lim Xiang Yann et al. *OpenPose: Real-Time Multi-Person Keypoint Detection Library for Body, Face, Hands, and Foot Estimation*. URL: https://github.com/CMU-Perceptual-Computing-Lab/openpose (visited on 2019-07-04). Presented in [Wei+16; Cao+17]. *used on: pp. 92, 93, 100, 102* [randomForest] Leo Breiman, Adele Cutler, Andy Liaw and Matthew Wiener. *randomForest: Breiman and Cutler's Random Forests for Classification and Regression*. URL: `https://cran.r-project.org/package=randomForest` (visited on 2019-12-11). *used on: p. 60*

[RSB] Sebastian Wrede, Jan Moringen and Johannes Wienke. *Robotics Service Bus (RSB)*. URL: `https://code.cor-lab.org/projects/rsb` (visited on 2019-11-11). Presented in [WW11]. *used on: pp. 8, 48*

[stats] R Core Team and contributors worldwide. *stats*. URL: `https://stat.ethz.ch/R-manual/R-devel/library/stats/html/00Index.html` (visited on 2019-12-11). *used on: pp. 53, 72*

[TensorFlow] Yong Tang, Derek Murray, Gunhan Gulsoy, Benoit Steiner, River Riddle and Sanjoy Das. *TensorFlow: A System for Large-Scale Machine Learning*. URL: `https://www.tensorflow.org/` (visited on 2019-10-07). Presented in [Aba+15; Aba+16]. *used on: p. 118*

DECLARATION OF AUTHORSHIP

According to the Bielefeld University's doctoral degree regulations §8(1)g: I hereby declare to acknowledge the current doctoral degree regulations of the Faculty of Technology at Bielefeld University. Furthermore, I certify that this thesis has been composed by me and is based on my own work, unless stated otherwise. Third parties have neither directly nor indirectly received any monetary advantages in relation to mediation advises or activities regarding the content of this thesis. Also, no other person's work has been used without due acknowledgment. All references and verbatim extracts have been quoted, and all sources of information, including graphs and data sets, have been specifically acknowledged. This thesis or parts of it have neither been submitted for any other degree at this university nor elsewhere.

Viktor Richter

Place, Date

COLOPHON

This thesis uses a template created by Johannes Wienke and is typeset using the [classicthesis] typographical look-and-feel developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography '*The Elements of Typographic Style*' [Bri92]. The structure of this thesis is inspired by [Hol14; Wie18].