Inferring Visual Attention and Addressee in Human Robot Interaction

THIS IS A TEMPORARY TITLE PAGE
It will be replaced for the final print by a version
provided by the service academique.

Thèse n. 1234 2014
présenté le 11 July 2014
à la Faculté des Sciences et Techniques de l'Ingénieur
laboratoire Idiap Research Institute
programme doctoral en Génie Électrique
École Polytechnique Fédérale de Lausanne

pour l'obtention du grade de Docteur ès Sciences
par

Samira Sheikhi

acceptée sur proposition du jury:

Prof. David Atienza Alonso, président du jury
Dr. Jean-Marc Odobez, directeur de thèse
Dr. Daniel Gatica-Perez, rapporteur
Prof. Britta Wrede, rapporteur
Prof. Frederic Lerasle, rapporteur

Lausanne, EPFL, 2014

To my parents...

Acknowledgements

There are so many people I would like to acknowledge for their help, support and friendship during the long journey I had towards this PhD.

First of all I would like to thank my supervisor Jean-Marc. His consistent guidance, useful advice and critical comments helped me to overcome the scientific and technical difficulties during my PhD. I am also thankful for having prof. David Attienza, Dr. Daniel Gatica Perez, Prof. Britta Wrede and Dr. Frédéric Lerasle as members of my thesis jury.

I would like to acknowledge the European project HUMAVIPS for supporting my research. I would like to thank Britta Wrede, Sebastian Wrede, Radu Horaud, Vaclav Hlavac and many others involved in this project for the fruitful experience. Thanks to Johannes Wienke and David Klotz for the smooth collaboration during HUMAVIPS. Special thanks to Vasil, Dinesh and Salim. Working with you has been a great pleasure. Also thanks Rémi for helping me to get through the difficulties at the beginning of the project.

During these four years I have been lucky to meet many great people at Idiap. Gelareh, thanks for being an amazing friend and as well for your help and tips for living in Switzerland. It has always been joy and relief to have you around. Kate, thanks for the good times we spent together and also for being so nice and caring. Rémi, you have been a great friend and teacher. Thanks for the enjoyable moments and activities as well as the tricks and advice. Kenneth and Alex, You are truly amazing friends. We shared lots of good times as well as difficult moments during the last four years. I am grateful for your company and support along the way. Maryam, I enjoyed a lot your friendship and the time we spent together. Other activities and lunch and dinner gatherings with you, Hassan, Gelareh, Peyman and the others were also very enjoyable. Thanks to people in the Perception group and in office 308 I have had a great atmosphere for work, discussion and learning. I have also learnt a lot from and enjoyed the company of many other friends at Idiap. Thanks to Afsaneh, Ramya, Ivana, CC, Sara, Marzieh, Mohammad Javad, Leo, Marc, Nesli, Pierre-Edouard, Rui, Gulcan, Cijo, James, Adolfo, Laurent, Joan, Laurent, Elie, Marco, Raphael, Harsha, Claire, Amiel, Jagan, Anindya, Gokul, Dinesh, Vasil, Niklas and many others. I also enjoyed a lot the lunch time Yoga sessions. Special thanks to Marc and Norbert for organizing it and to our Yoga teacher and the others for the nice experience. Thanks as

well to Nadine and Sylvie, for tirelessly helping me with all sort of problems during these years related to living in Switzerland. Thanks to everyone from the administrative and system staff at Idiap. Thanks to Corinne, Chantal and Fabienne and everyone at EPFL who was involved in the organization of my defense.

I have been privileged to have great friends in Lausanne who made it a very nice place to live. Thanks to Zahra and Sohrab who generously shared lots of enjoyable moments with me and helped me in many ways. Special thanks to Zahra. You have been a very supportive friend. Mitra and Arash, thanks for your friendship over these years. I really enjoyed the time spent with you both as well as the coffee breaks with Mitra which helped me to stay attached to my younger self. Elham and Zahra, I will never forget your hospitality during my stay at your place and I appreciate your friendship. Many thanks to Sara. You are a great friend and I loved the running, biking and dinings with you. Also thanks to Masoud, Milad, Mohammad and Jan for the good times and for your caring friendship.

Finally a big thank you to my family for their love and support and for staying with me in the ups and downs. Sara and Saman, you are amazing siblings. Thanks for your support, encouragement, positive energy and kindness. Mom and dad, I cannot express in words how thankful I am to you. This thesis is dedicated to you.

Lausanne, 18 August 2014 Samira Sheikhi

Abstract

With the recent advances in the field of robotics, now it is the time to have robots with the capability to interact with humans in natural ways similar to them. One of the essential aspects for a robot to be capable of performing such an interaction is perceiving humans along with their states, behaviors and actions. In this direction, gaze has a key role as a nonverbal behavior since it reveals important information about people's interests and intentions. It shows to whom or what the person's attention is directed and to whom somebody is speaking. It also helps in communicative tasks such as ground management and turn taking, and helps the robot to know whether people are interested to continue the conversation and to estimate their involvement. Given the importance of gaze, it is necessary to provide algorithms for its recognition in human robot interaction settings.

Eye gaze estimation with commercial robots is often impossible to achieve given the unconstrained conditions of people motion and the available video sensors. Therefore, most systems currently rely on head pose as an approximation of gaze, or to recognize its discrete version the Visual Focus of Attention (VFOA), defined as whom or what a person is looking at. However, using head poses creates ambiguities since the same head pose can be used to look at different VFOA targets. To address this challenge, we proposed a dynamic Bayesian model for the VFOA recognition from head pose, where we make two main contributions. First, taking inspiration from behavioral models describing the relationships between the body, head and gaze orientations involved in gaze shifts, we proposed several novel gaze models that dynamically and more accurately predict the expected head orientation used for looking in a given gaze target direction. Obtaining the expected head pose for looking at different directions is a neglected aspect of previous works but essential for recognition in conditions where setting the parameters manually or from the training data is not applicable. Secondly, we proposed to exploit contextual information from the robot conversational state (when he speaks, people he addresses, and objects to which he refers) in the recognition framework to set appropriate priors on candidate VFOA targets and reduce the inherent VFOA ambiguities.

As another contribution of this thesis, we investigated the recognition of the addressee of people's speech (defined as to whom they speak), in our human robot interaction setting, which is another important communication cue. As it is well known that addressee can primarily be derived from the speaker's VFOA, we proposed a method for estimating addressee using automatically extracted VFOA from head pose. Moreover, we investigated the role of

conversational context in improving the recognition by using it either directly as a side cue in the addressee classifier, or indirectly by improving the VFOA recognition. Finally, from a computational perspective, we studied which VFOA features and normalizations are better for addressee estimation. Particularly we addressed whether it matters for the VFOA recognition module to only monitor when a person looks at potential addressee targets (the robot, people) or if it is better to consider all objects of interest in the environment (paintings in our case) as additional VFOA targets.

Experiments were conducted on three datasets, including our public Vernissage dataset where the humanoid robot NAO plays the role of an art guide and quiz master. For VFOA recognition, they demonstrate the benefit and complementarity of the two contributions we propose for improving the recognition. Experiments on the second part of the recordings in Vernissage data, where the humanoid Nao robot offers a quiz to two participants show that reducing VFOA confusion (either through context, or by ignoring VFOA targets) improves addressee recognition.

Key words: gaze, VFOA, visual focus of attention, head pose, addressee estimation, human robot interaction, HRI, robot context, conversation context, Bayesian models.

Résumé

Les récentes avancées dans le domaine de la robotique permettent d'envisager des robots capables d'interagir avec les humains de façon naturelle. Un des aspects essentiels pour atteindre cet objectif est la mise au point d'algorithmes de perception des personnes, ainsi que leurs états, comportements et actions.

Le regard joue un rôle essentiel dans cette perspective, car il fournit des informations importantes quant à l'intérêt et l'attention des personnes. Il montre à qui s'adresse une personne ou vers quoi son attention est orientée. Dans les situations de communication, il permet à un robot de savoir si une personne est intéressée par la conversation ou d'estimer son degré de participation. Etant donnée l'importance du regard, il est nécessaire d'élaborer des algorithmes capables de le reconnaître dans le contexte des interactions avec un robot.

Le regard en tant qu'orientation des yeux est difficile voire impossible à estimer avec les robots commerciaux actuels dans des conditions non contraintes et les capteurs vidéos disponibles. C'est pourquoi le plupart des systèmes actuels s'appuient sur l'orientation de la tête, ou pose, comme approximation du regard. Cependant, celle-ci peut être ambiguë, car une même orientation peut permettre par exemple de regarder deux cibles différentes.

Nous proposons dans cette thèse d'utiliser un modèle bayésien dynamique pour la reconnaissance du centre d'attention visuel à partir de la pose de la tête, et faisons deux contributions dans ce cadre. La première s'inspire des modèles de comportement décrivant les relations entre le corps, la tête et le regard et propose plusieurs nouveaux modèles du regard qui prédisent de façon dynamique et plus précise l'orientation de la tête utilisée pour regarder une cible donnée. Prédire cette orientation pour différentes cibles est un aspect peu traité des méthodes existantes, bien qu'il soit essentiel à la reconnaissance dans des conditions où l'ajustement des paramètres à partir de données d'entraînement est impossible. Comme seconde contribution, nous proposons d'exploiter l'information contextuelle de l'état conversationnel du robot (quand il parle, à qui il s'adresse, à quels objets il fait référence) comme a priori sur les cibles candidates pour le regard, afin de diminuer les ambiguïtés.

Dans un autre partie de la thèse, nous nous sommes intéressés à la reconnaissance du destinataire d'un tour de parole, qui, dans le contexte d'une interaction homme-robot, est une autre importante manifestation de communication. Etant donné l'importance du regard dans ce processus, nous avons tout d'abord proposé une méthode permettant de reconnaître qui est le destinataire de la parole en utilisant l'attention visuelle estimée à partir de la pose de la tête. De plus, nous nous sommes intéressés au rôle du contexte de la conversation pour améliorer cette reconnaissance, en l'utilisant soit directement dans la classification ou indirectement

pour améliorer l'estimation de l'attention visuelle. Finalement, nous avons étudié quelles représentations de l'attention visuelle et normalisations étaient les meilleures pour déterminer le destinataire de la parole. Plus particulièrement, nous avons déterminé s'il était nécessaire de modéliser seulement les cibles de regard potentielles de la conversation (robot, personnes) ou s'il fallait considérer tous les objets environnants (tableaux de musée dans notre cas) lors de la reconnaissance de l'attention visuelle.
Nous avons effectué des expériences sur trois bases de données, incluant le corpus de données public Vernissage auquel nous avons contribué et dans lequel le robot humanoïde NAO joue le rôle d'un guide de musée ou pose les questions d'un quiz. Ces expériences démontrent la validité et le bénéfice de nos contributions pour améliorer l'estimation de l'attention visuelle. Elles montrent aussi qu'il est possible d'améliorer la reconnaissance du destinataire en réduisant de différentes façons la confusion de l'estimation du regard.

Mots-clés : interaction homme-robot, communication nonverbale, regard, pose, estimation du destinataire, comportements, contexte, modèles bayésiens.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Recent advances in the fields of robotics or embodied conversational agents (ECAs) open the doors for having agents advanced enough to interact with users in natural human like manners. The ultimate goal would be to have robots endowed with advanced social skills to interact with humans in open world situations as we experience in our daily lives without presumed constraints and enforced conditions. Such robots or ECAs could provide different kind of functionalities or services, like delivering information, helping to maintain social bonds or entertaining humans. Figure 1.1 shows a few examples from all the possible applications for these robots. Robots are currently used as assistants for senior or dependent people by monitoring their health and facilitating their contact to their caregivers or family members. They also serve as Foreign language teachers, being remotely controlled by human teachers who are possibly native speakers living in other countries. They could also be used as shopping assistants, either by helping an individual person in finding and filling their shopping baskets or by being in shopping centers, detecting people who are looking for direction or information and guiding them.

Given the technological advances, robots are often connected to internet or home appliances and thus have access to information that can be useful for humans. ECAs or humanoid robots could then be the ultimate interface: their role could consist of understanding the user's request, retrieving information from the internet and presenting and delivering it in an appropriate fashion to them. Therefore, having robots which can interact with humans in natural manners and take the role of an engaging and interactive companion, makes humans independent of a keyboard, computer or mouse to communicate with technology. Talking naturally to our robots would be all we need to do, and the robot would respond and deliver that information using our familiar human like media.

Robots with capabilities to interact with humans using natural human-specific means rely on different components, devoted to the perception of the scene and surrounding people, to the analysis of the information, decision making and finally to the synthesis of the be-

Figure 1.1 – Different applications of embodied conversational agents. a) Shows robot Kompaii as an elderly assistant to monitor their health and contact their caregiver if needed - from [technocrazed.com, 2013]. b) Shows a robot used for teaching English language to students, being remotely controlled by a real teacher - from [koreaittimes.com, 2010]. c) A robot named Robovie-II moves around a grocery store during an assisted shopping experiment. This little robot greets shoppers at the entrance of a grocery store and then follows them while holding a grocery basket. It can also remind people of items on a shopping list - from [bits.blogs.nytimes.com, 2010]. d) Relying on data from surveillance sensors, Robovie spots people who look disoriented, approaches them and asks, “Are you lost?” If so, the robot provides simple directions - from [engineering.curiouscatblog.net, 2008].

haviors and actions in a way which looks appealing to the users. A sample architecture for human-robot-interaction is illustrated in Figure 1.2 showing how different components are integrated to provide an example of such systems. Still lots of advances are needed in different areas including speech recognition and synthesis, multimodal sensing and fusion, dialog and interaction modeling, and conversational scene analysis. It is very important to realize that endowing such systems with social skill capabilities requires the design of reliable human behavior perception and understanding algorithms. These algorithms should go beyond the more studied tasks of people localization and determination of their speaking status and move towards understanding their activities and intentions.

In daily life situations, an important capability for the robot is the ability to carry on a meaningful dialog with two or more participants. It should keep track of the conversation flow and in particular know when it is each participant’s turn to speak. This capability requires the

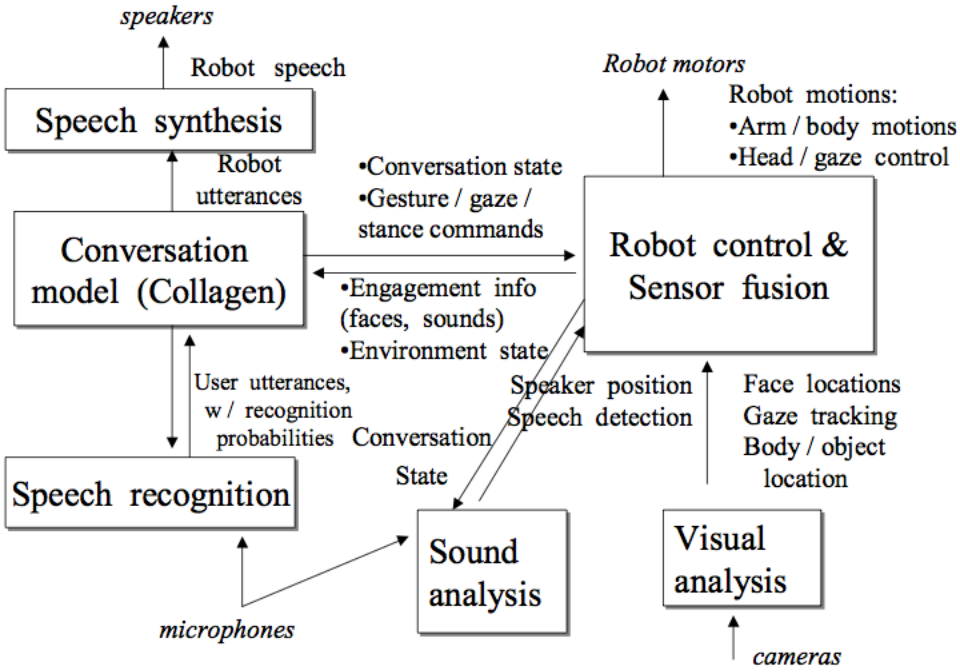


Figure 1.2 – An example for human robot interaction architecture - from [Sidner2004]. The left side of the image shows the dialog part of the system. This part gets input from the microphones and performs speech recognition. Then by integrating engagement information and environment state from the control module decides on the utterances robot should speak in addition to the gesture and gaze behaviors which should be synthesized with them. The right side is more focused on sensing, and data fusion together with decision making which is performed in a robot control module. The control module gives feedback to the other modules and decides on the robot movements.

understanding of several communication information: from essential ones like who is where and who is speaking to more complex ones such as to whom or what a participant's attention is directed at, addressee detection (to whom somebody is speaking, and in particular, when is the robot addressed) or finding when is a relevant time to speak. At a higher level, these communication information also help the robot to know whether people are interested to continue the conversation and how much everyone is involved. Accordingly the robot can frame its dialog and conversational acts to hold a suitable and pleasant conversation.

For a broader view on the motivations and different tasks related to this problem, we refer the reader to the HUMAVIPS project[1] which is the parent project of this thesis. In HUMAVIPS, we had the main goal of providing a robot which is capable of performing natural interaction with a group of people. The project robot, Nao[2], could use its input sensory data consisting of audio and video channels to get an understanding of its surrounding, and show appropriate behavior by mean of its speech, gestures and movements. In HUMAVIPS we tried to address

[1] http://humavips.inrialpes.fr
[2] http://www.aldebaran.com/en/humanoid-robot/nao-robot

this goal by using the expertise of different European project partners in the areas of sound and audio processing, computer vision, dialog and social interaction, multimodal data integration, signal processing and pattern recognition.

Different groups of tasks were studied during the HUMAVIPS project and resulted in useful algorithms implemented on Nao. In a first group, perception of humans and their nonverbal behaviors was addressed by targeting a number of tasks. Nao should be able to detect and track people and identify them even if it looses them for a short time, and should recognize faces and determine their gender and age. Moreover, it should be able to estimate their head pose, the addressee of their speech, and detect when they nod. The second group was related to audio localization and audio-visual association. This was important in order to let Nao know the direction the sound is coming from and the person or object which is the source of it. Modeling group interaction was also studied in the project. Nao should know what kind of group it is interacting with; like the size of the group or the age range of people. It should have engagement strategies to decide when to wait for people to speak and when to propose to give some explanations. Moreover, it should estimate how much people are following its explanations and are interested. Having this information helps Nao to adapt its speech and behavior depending on the context and situation. Furthermore, localization was also addressed to help Nao know the location of the important objects and be able to localize itself with respect to them. Finally, providing a cognitive robotics architecture was necessary to integrate all the information from different modules (audio, localization, video, dialog and his own actions) and maintain a coherent representation of the world. At every moment, given this representation, Nao should be able to decide what to do.

As can be seen from the previous paragraphs, perceiving humans along with their state, behavior, and action is essential for robot to interact naturally with humans. In the next Section we will emphasize on the role of gaze as a nonverbal behavior which reveals important information about people's interests and intentions. Given the importance of gaze, it is necessary to provide algorithms for its recognition and interpretation in terms of Visual Focus of Attention (VFOA), defined as whom or what a person is looking at in HRI or ECA settings. This is the main topic addressed in this thesis.

1.2 Objectives and Challenges

In this Section we will introduce the main objectives of this thesis which are the recognition of visual focus of attention and addressee in HRI scenarios. In addition we will describe challenges for addressing these problems in order to provide a better insight and emphasize on the importance of studying them.

Gaze is amongst one of the most important behaviors exhibited during interactions. In particular, it shows visual attention which is a close substitute for attention and is important to track while carrying on a conversation. It is a good indicator of the addressee (to whom a person is speaking) of an utterance, which is an important information to know for robots

or ECAs interacting with multiple people. Due to this important role, gaze has been used for turn-taking management, and at a higher level to monitor people engagement and intention or recognize user's predefined states of interaction in HRI application. This nonverbal cue has many other functions like establishing relationships, or exercising social control; and Chapter2 will analyze them more thoroughly.

Due to this importance, the first objective of this thesis is to investigate algorithms for its recognition and interpretation in terms of VFOA in HRI scenarios. Several conditions defined in the HUMAVIPS project frame the approaches investigated in this thesis. First of all, in these scenarios, people are not constrained on the way they interact with the robot: they can freely move and take arbitrary distances and poses with respect to the robot. Second, a focus of the project was on commercially available robots (Nao in the experiments). Therefore, only consumer sensors are assumed to be available rather than high-end sensors. Furthermore, the goal is to provide methods which are useful for open and dynamic environments.

Measuring and interpreting the gaze of people is a difficult task in general. Eye tracking devices can be used but are usually expensive, considered as intrusive, and usually not applicable for natural interactions. Nevertheless, benefiting from advances in computer vision tracking systems, researchers have mainly considered the head pose as an approximation of the gaze, a trend that should increase with the new Kinect camera and API that directly delivers this information. Not having the eyes information makes it impossible to extract the exact gaze, but still providing some approximations of gaze or its interpretation in terms of VFOA should be possible.

However, while interpreting the head pose as looking at VFOA targets is supported by both behavioral modeling and empirical evidence, it has its own challenges. First, the robot needs to know where the people and VFOA targets are and what is their position with respect to each other. To obtain this information, the robot should be equipped with good localization and tracking modules. Dealing with moving people makes this problem specially more challenging. Furthermore, associating head poses with VFOA targets remains ambiguous since in realistic scenarios, the same pose can be used to look at different targets depending on the situation.

The second objective of the thesis is to investigate the recognition of the addressee of people speech, which is another important communication cue. For instance, in one part of the main scenario of the HUMAVIPS project the robot makes a quiz and when people are speaking, they might be discussing the answer among themselves or responding to the robot. In this case it is very important for the robot to recognize whether it is the addressee of their speech to respond appropriately or not.

There are different challenges for detecting the addressee. The robot first needs to recognize that someone is speaking, then identify or localize the speaker and be able to extract the speech utterance. After having obtained that, the robot should then use audio and visual cues to predict the addressee.

The speaker's gaze or VFOA has been shown to be a very informative cue of addresseehood since people mostly look at the person they are addressing rather than the others. However, even accurate gaze information is not sufficient for addressee estimation. As a result, researchers have investigated other cues to provide context (e.g. lexical, prosodic cues) and improve performance. These contextual cues could be derived from the other participants activities or the interaction happening between the robot and the users. Therefore, it is interesting to study how gaze information could be used for estimating the addressee while integrated with other contextual cues, and which contextual cues could be useful for this goal.

1.3 Contributions

In this Section we provide an overview of the contributions made in this thesis, which are related to the recognition of the VFOA from head pose and addressee estimation. Moreover, we participated in recording a HRI dataset appropriate for investigating the above mentioned problems.

A central issue when trying to decode the sequence of VFOA targets given the head pose sequence is the following: what is the expected head pose of a person that looks at a given VFOA target? In gazing behaviors, the difference between a gaze direction and the head pose used to look in that direction, which is due to the missing eye information, can not be considered as a random noise with zero mean. Rather, it is often biased, and the bias depends on several factors related to the body, head and eye dynamics. In spite of the importance of the mentioned problem, this issue has seldom been addressed in the past. Some methods like [Foster et al., 2012] use training data to directly infer VFOA from head pose without defining gaze as an intermediate step. Learned parameters, however, are then specific to the geometric configuration between the sensor (robot), person, and VFOA targets, and thus such an approach is not suitable for robot dealing with moving people.

To address this problem one of our main contributions was to study and validate different gaze to head pose behavior models by taking inspiration of the results on human gazing behavior and head-eye dynamics involved in saccadic gaze shifts which study gaze models relating the head pose, gaze direction, and body orientation. Similar ideas had been used in the past to provide simplified algorithms in static scenarios where people are seated and limited in terms of their body movements. However, these models had not been successfully applied for more complicated and dynamic scenarios with freely moving people as we would face in situated human robot interaction. A main contribution in this thesis was to explore this direction and find effective methods which are appropriate for dynamic conditions.

A second contribution of our work has been to use the interaction and dialog context as given by the robot to solve the head pose interpretation ambiguity. Indeed it has been shown that one way to resolve ambiguities in nonverbal behavior understanding was to use other social cues, leveraging on the fact that some behaviors provide context to others. In human-human interaction, examples for VFOA recognition include speaker information or higher

conversational states, that can be complemented with group activity. While in the above cases the social cues used as context have to be inferred from the data and might be noisy, in the robotic or ECA cases, the agent is fully aware of its own conversational acts, allowing them to be conveniently exploited to better interpret the nonverbal cues performed by interacting people. The robot's conversational acts could even be viewed as important causes which affect where people look. However, to our knowledge, while estimating the VFOA is considered by several systems, the use of the robot dialog context to improve the recognition of a user attention (VFOA) has not been explored in the past. Therefore, we investigated leveraging different types of conversational context from the robot to improve the VFOA recognition. This solution removes some of the ambiguities introduced by using head poses as the only input for the recognition.

As the third contribution, addressee estimation, one of the applications of gaze or VFOA in higher level tasks in human behavior understanding, was also studied in this thesis. We were interested in studying how well addressee could be estimated in this kind of scenario using estimated VFOA information. Several research questions arise there, like how should we feed the VFOA information into the addressee estimation algorithms, and whether contextual cues could be used in addition to the most commonly used gaze cue from the participant to improve recognition. VFOA estimations obtained from the approached developed in the thesis were used as inputs for the higher level task of addressee estimation. We experimented with two main different conditions. In the first one, VFOA was estimated assuming that the robot is aware of all of the targets and considers them in its computations. In the second case, in contrast to the first one, the robot was assumed to consider only the targets which are probable addressees (i.e. the other participants and the robot). Moreover, we experimented the effect of adding contextual information to the gaze cues and studied different combinations.

Finally, as another contribution, we participated in the designing of the scenario and recording of the Vernissage dataset. In order to study the perceptual tasks which are important in human robot interaction as described in section 1.1, computer scientists require a corpus of relevant data. This data should allow to understand the behavior of people interacting with a humanoid robot, to create appropriate models of the interaction and algorithms decoding the current situation, and also benchmark the performance of these algorithms. This is important for designing algorithms which enable the audio-visual perception of the scene and control the generation of appropriate behaviors for the robot. The Vernissage dataset provided us with realistic and natural human robot interaction data which was not publicly available before. This dataset was recorded in collaboration with our HUMAVIPS partner from the Bielefeld University and using their infrastructure. We were mostly involved in the design of the scenario and behaviors, the implementation of the wizard of oz, and annotations of the data, while Bielefeld University provided important infrastructure for the recordings and managed all of the technical parts. This dataset is publicly available as the Vernissage Corpus, and contains challenging behaviors where the robot is an active partner in the interactions. Eventually we used this dataset as the main data for different tasks of VFOA recognition and addressee estimation since it includes rich contextual data to experiment with.

The above mentioned contributions have also resulted in several publications as listed below:

- Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions, S. Sheikhi, J-M. Odobez, Submitted to the Special Issue on Human Computer Interaction, Pattern Recognition Letter, 2014.
- Context Aware Addressee Estimation for Human Robot Interaction, S. Sheikhi, D.B. Jayagopy, V. Khalidov and J-M. Odobez, in: the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction, ICMI, 2013.
- Leveraging the Robot Dialog State for Visual Focus of Attention Recognition, S. Sheikhi, V. Khalidov, D. Klotz, B. Wrede and J-M. Odobez, in: Int Conf. on Multimodal Interaction (ICMI), 2013.
- The Vernissage Corpus: A Conversational Human-Robot-Intercation Dataset, D.B. Jayagopy, S. Sheikhi, D. Klotz, J. Wienke, J-M. Odobez, S. Wrede, V. Khalidov, L. Nguyen, B. Wrede, D. Gatica-Perez, in Proceedings of the 8th ACM/IEEE international conference on Human-Robot interaction, 2013.
- Investigating the Midline Effect for Visual Focus of Attention Recognition, Samira Sheikhi and Jean-Marc Odobez, in: Int Conf. on Multimodal Interaction (ICMI), 2012.
- The Vernissage Corpus: A Multimodal Human-Robot-Interaction Dataset, D.B. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J-M. Odobez, S. Wrede, V. Khalidov, L. S. Nguyen, B. Wrede and D. Gatica-Perez, Idiap-RR-33-2012.
- Recognizing the Visual Focus of Attention for Human Robot Interaction, S. Sheikhi, V. Khalidov and J-M. Odobez, in: IEEE International Conference on Intelligent Robots and Systems (IROS) - Human Behavior Understanding Workshop (IROS-HBU), 2012.

1.4 Thesis plan

The thesis is outlined as follows. In **Chapter 2** we provide background information and a review on the related work. In **Chapter 3** we describe the datasets used in the thesis. **Chapter 4** provides the gaze models and VFOA recognition algorithms. In **Chapter 5**, experiments on the VFOA recognition task are explained and results are provided. The addressee estimation method is presented in **Chapter 6** and **Chapter 7** gives a conclusion on the thesis.

2 Background and Related Works

This thesis is focused on the design of principled methods for the estimation of gaze or of its discrete variation the visual focus of attention (VFOA) in the context of social interaction between a robot and a group of people. This includes the influence of higher level interaction cues from the VFOA. Therefore, in this section, we will focus on related works pertinent in these areas.

To start with, we will first review in Section 2.1 the related background on the social and communicative roles of gaze in human interactions. In order to perceive humans gaze it could be very helpful to understand the mechanism underlying gaze behaviors. In Section 2.2 we will remind the different types of gaze motions performed by humans, with an emphasize on gaze shifts that represent the main activity for changing focus. Humans perform those shifts by incorporating eye, head and their body motion, and in Section 2.2 we will thus go through the existing models which describe the coordinations between these different parts during the gaze shifts.

The social and communicative roles discussed in Section 2.1 could also exist when considering human like natural interactions between people and computers, embodied agents or robots. Thus in Section 2.3, we introduce and discuss the gaze usages in these areas with regards to both the perception of human gazing behavior and the synthesizing of similar behavior on embodied agents and robots.

The estimation of gaze and recognition of the VFOA constitutes the main part of this thesis. Thus in Section 2.4 we will review the main classes of methods commonly used to address these problems. This will be followed in Section 2.5 with an overview of the literature on the exploitation of context in human behavior understanding and more specifically VFOA recognition.

Finally in Section 2.6 we will give a summary on the literature for addressee estimation problem as one of the tasks which could be studied after the perception of visual focus of attention.

2.1 Communicative and social roles of human gaze behavior

Gaze is used widely is our daily tasks and activities. It plays basic roles in exploring scenes, reacting to sudden changes, manipulation objects in coordination with other body parts and of course, in interactions. Gaze direction is an important nonverbal cue to express visual attention [Langton et al., 2000] and as such has many functionalities in conversation and more generally in human human interaction. It fulfills functions such as establishing relationship (through mutual gaze), regulating the course of interaction, expressing intimacy [Argyle and Dean, 1965], and exercising social control [Langton et al., 2000].

In dyadic interactions, Kendon [Kendon, 1967] suggested that we can distinguish between at least four functions of gaze:

- to provide visual feedback,
- to regulate the flow of conversation,
- to communicate emotions and relationships,
- to improve concentration by restriction of visual input.

In these interactions, speakers tend to look away at the beginning of an utterance and turn their attention towards the conversational partner at the end [Kendon, 1967]. Moreover, Argyle and Cook [Argyle and Cook, 1976] showed that while listening, people look nearly twice as much (75%) than while speaking (41%).

Considering multiparty interactions, [Vertegaal et al., 2001] studied 4 people conversations. Here again listeners gaze more than speakers. A speaker gazed 3.3 times more at an addressed individual who is the target of their speech (39.7%), than at others (11.9%). Reversely, listeners gazed approximately 7.3 times more at the individual who was speaking (62.4%), than at others (8.5%). Thus people use the other's gaze to determine when they were addressed or expected to speak, which is important specially at transition points. Moreover, a speaker's gaze often correlates with the gaze of his addressees, especially at a sentence end where the gaze can be interpreted as a request of back-channel or an invitation to grab the floor [Jovanović and op den Akker, 2004].

Beyond these statistics and of high importance in multiparty interactions, gaze direction serves as an important cue in floor management or turn taking [Duncan, 1972, Goodwin, 1980, Schegloff, 1968]. As defined in [Bohus and Horvitz, 2011], at any time instance, each participant in an interaction can take one of four following floor management actions: a hold action for maintaining the floor (or the speaking turn); a release action for giving the floor to the other participants; a take action to try to acquire the floor; and finally, a null action which indicates that a participant is not making any floor claims. The floor shifts from one participant to another can then be considered as the result of the joint and cooperative floor management actions taken by the participants. Speakers and listeners use their gaze for this regards. For instance, speakers might look away from their addressees to indicate that they are in the process of constructing their speech and do not want to give away their

turn, and look at their addressees to signal the end of a remark and pass the floor to another participant [Schegloff, 1968]. In this context, the participant at whom a speaker looks at the end of a remark would be more likely to take the role of the next speaker [Kendon, 1967]. Shifting of roles might be delayed when remarks do not end with the speaker gazing at another participant [Kendon, 1967, Vertegaal and van der Veer, 2000]. On the other hand for a listener, monitoring his own gaze in concordance with the speaker's gaze is a way to find appropriate time windows for taking the floor [Duncan, 1972, Novick et al., 1996].

We are also very sensitive to the gaze of others when it is directed towards objects of interest within or even outside our field of view [Pourtois et al., 2004]. Since people look at the objects related to what they listen to [Cooper, 1974], gaze can also be used to monitor whether people follow the conversation or evaluate their level of interest. Therefore, in group conversations where artifacts exist and play role in tasks and activities of people, they can affect people's gazing behaviors. In the meetings for instance [Ba and Odobez, 2008], slides on the screen and papers on the table are shown to effective on the participants gaze. In a museum, art works affect the visitors gaze and thus, monitoring their gaze pattern can be used to reveal their interest about different paintings. Gaze behaviors could be understood as a mixture of all these effects, which renders its analysis and interpretation quite difficult.

Gaze is also shown to be greatly involved in higher level cognition processes that underly human interactions, like engagement and grounding. According to Sidner and colleagues [Sidner et al., 2004] engagement "is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". In this view, while the listener employs gaze to indicate that s/he is paying attention to the speaker, the speaker monitors the listener's gaze to find out whether s/he is still interested in continuing the conversation [Rehm and Andre, 2005]. These functions of gaze are thus playing an important role in displaying engagement in a dialogue. In fact, lacking or failing (on purpose or not) to properly display these expected communication behaviors can be interpreted by others as a disengagement within the interaction. As another process which goes on during interaction, we refer to grounding which is the process of updating mutual knowledge, mutual assumptions and mutual beliefs during the interaction. To minimize the chance of errors during this process due to misunderstandings humans extend their verbal utterances with gaze and pointing gestures and social robots will have to rely on that modalities as well [Häring et al., 2012].

Beyond being useful at monitoring interactions at different levels, gaze behaviors can also be related to more fundamental social construct, and to the personality traits. Examples are the relevance of gaze patterns to identify dominance and social status [Jayagopi et al., 2008]. Dominant and high status people often receive more visual attention and look at others more often [Cook and mith, 1975]. Automatic estimation of these two constructs could be performed using gaze in addition to other nonverbal cues [Jayagopi et al., 2008, Hung et al., 2008]. Furthermore, detecting unconventional gaze behavior in social interactions could be useful for the diagnosis of disorders like Autism. Indeed autistic children display distinct

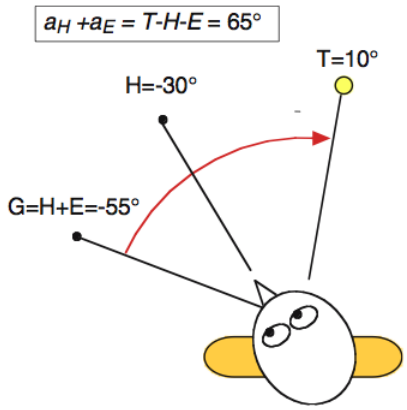


Figure 2.1 – Coordinates for an upcoming gaze shift. The physical parameters needed for the upcoming gaze shift are spanned by the target position T, the initial head position H^0 and the initial eye position E^0 (with respect to the head). The initial gaze position is $G^0 = H^0 + E^0$. For an accurate gaze shift, the movement of the gaze vector must be equal to the gaze error $T - G^0 = T - H^0 - E^0$ - from [Hanes and McCollum, 2006] – I should change them to H^0 and E^0.

behaviors from a very young age. They look at others less often and make less and shorter eye contacts. In addition they have problems in using joint attention: for example they may look at a pointing hand instead of the pointed-at object. Accordingly, this cue is considered for developing an automated diagnosis system for Autism which relies on analyzing interactions between children and caregivers [Rehg, 2013].

2.2 Eye-head coordination during rapid eye/head gaze shifts

Human's eye movement consists of different patterns that can be classified into three main categories of fixations, saccades, and smooth pursuits. A fixation occurs when the gaze rests for some minimum amount of time on a small area. Saccades are fast rotations of the eye (and potentially the head) that occur between fixations to two distinct areas, with the aim of bringing the objects of interest into the visual field. Smooth pursuit movements describe the eye when following a moving visual target. In this Section we concentrate on saccades (also called gaze shifts) because they are of more interest and relevance to our work.

Under natural conditions, humans perform gaze shifts by incorporating eye, head, torso and even foot movements. Therefore an accurate gaze shift can potentially be achieved through many alternative combinations of these motions. Considering stationary foot and torso, head and eyes would contribute to the gaze shift as illustrated in Figure 2.1. For instance a person who aims to look at a target 30° to the right can make an eye saccade of 30° or can rotate her head 20° and her eyes 10° in the head.

One of the main questions here is to understand how the brain translates a desired movement of the gaze direction into particular movements of relevant body segments [Hanes and Mc-

Collum, 2006]. In [Freedman and Sparks, 1997] this problem was studied by analyzing gaze shifts made by trained rhesus monkeys with completely unrestrained heads. They found that coordinated eye-head movements are characterized by a set of lawful relationships, and that the initial position of the eyes in the orbits and the direction of the gaze shift are two factors that influence these relationships.

In [Wang and Jin, 2001], the authors tried to reformulate the existing relationships and provide predictions for the displacement of the eye and head components of the gaze shift based on the previous experimental data [Freedman and Sparks, 1997]. They suggested that the movements of the eyes and head during the unrestrained-head gaze shift follow a tight linear relationship linking head contribution and gaze amplitude. In the following we first describe the prediction model suggested in [Wang and Jin, 2001] and then introduce another model suggesting the importance of the initial head pose on the predicted contributions for head and eye gaze.

2.2.1 Horizontal gaze shift given gaze displacement and previous eye orientation

For simplicity, horizontal gaze shifts are considered in the following. The gaze shift is the sum of the eye-in-head movement and head-in-space movement and to consider the effect of the initial position of the eye in the orbit, it is considered to be formed as follows:

$$G = E(E^0) + H(E^0) \tag{2.1}$$

where G is the gaze displacement, E^0 is the initial position of the eye, and E and H are the eye displacement and head movement contributions. Then the purpose of these models is to predict the eye displacement $E(E^0)$ and the head movement contribution $H(E^0)$ if the gaze displacement G and the initial position E^0 of the eyes in the orbits are known. The models are inspired from the experimental data of [Freedman and Sparks, 1997] that are illustrated in Figure 2.2(a).

Eyes centered in the orbit: When the eyes are initially centered in their orbit, the relationship between the eye amplitude and the gaze amplitude is characterized by two linear functions, shown as in Figure 2.2(b). The empirical rules observed in [Freedman and Sparks, 1997] can be used to specify the parameters of these piecewise linear functions. Specifically it is considered that for $G < 25°$, the eye moves without head contribution. Then, for $G > 25°$, the head begins to contribute to the gaze shift. Assuming that the subject has a maximal visual target offset of 80°, the head contribution increases linearly for increasing gaze amplitude for movements between 25° and 80°. When the gaze amplitude G reaches its maximum 80°, the eye movement amplitude saturates at amplitude 35°.

Eyes not centered in the orbit: When the eyes are not initially centered in orbit and are at 10° to 30° contralateral to the direction of the gaze shift, the relationship in the experimental data of [Freedman and Sparks, 1997] showed that the relationship between the eye amplitude and

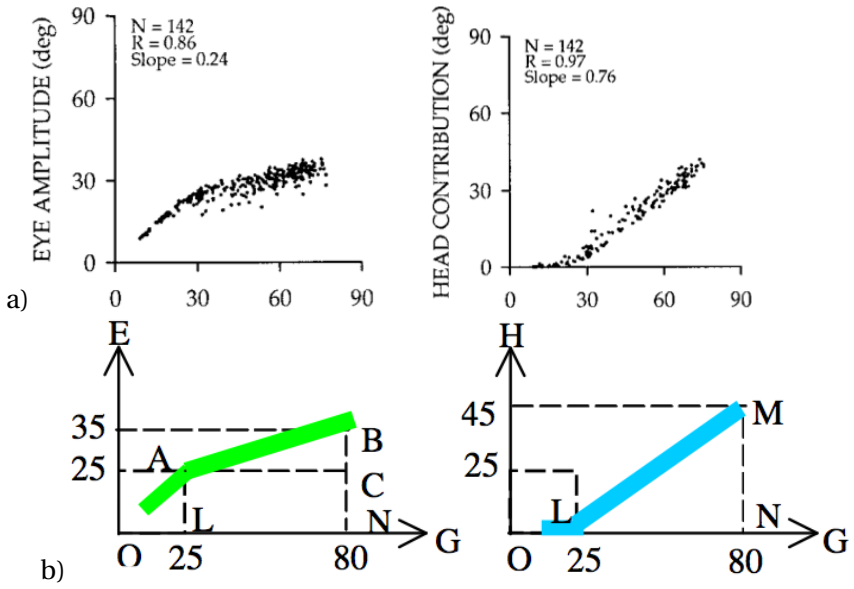


Figure 2.2 – (a) Eye (left plot) and head (right plot) contributions to the amplitude of the gaze shifts when the eyes are initially centered in the orbit [Freedman and Sparks, 1997]. (b) Linear functions proposed by [Wang and Jin, 2001] describing the underlying relationships illustrated in (a).

gaze amplitude could be characterized with similar functions as above.

2.2.2 Horizontal gaze shift considering previous head pose and midline

The previous approach only incorporated the gaze displacement $G = T - G^0$ and initial eye position (E^0) with respect to the head for estimating the head and eye contributions.

However, in [Hanes and McCollum, 2006], which is meta-study analysing the previous published data and modules and that proposed an axiomatic approach to represent the properties of gaze shift models, a bigger variable set is considered as necessary for determining the head and eye contributions. More precisely, the target position T, the initial head position H^0 (with respect to the stationary shoulders), and the initial eye position E^0 (with respect to the eyes) should be considered simultaneously. These elements are illustrated in Figure 2.1. Subsequently the initial gaze position $G^0 = H^0 + E^0$ is also assumed to be known.

An important feature of this study was to more explicitely consider and account for the effect of the torso or shoulder orientation, The authors introduced a specific term, the **midline**, to refer to the vector normal to the torso. They claimed that the relative position of the head with respect to the midline should be considered when determining the contribution of the different parts to gaze shift, and denoted this influence as the midline effect.

In the following three subsections we will first consider the case where both head and eyes are initially centered at the midline and provide the general characteristics of the gaze shift model

for this condition. Then we will describe the effect of midline on gaze shift behavior when we do not have the previous centered condition. Finally considering the first two points, we will provide an axiomatic model for the general gaze shifts.

Gaze shifts from the midline

They defined the basic case to be when both head and gaze are initially centered at the midline (i.e. $H^0 = E^0 = 0$). $F_c(T)$ denotes the centered function specifying the amplitude of the head movement that occurs for a gaze shift to a visual target at position T in this basic case.

From the analysis of several sources of data, they extracted several interesting properties regarding this centered function $F_c(T)$. Among them, this function (representing the head movement) is weakly increasing and for targets close enough to the initial gaze position, eye-only saccades are used to shift the gaze. The head movement is always towards the target but never overshoots. Moreover, as an implication of the properties of $F_c(T)$, they also mention that the eye movement is weakly increasing as a function of target position and there is a limit on allowable eccentricity of the eyes in the head. Note that the model of Section 2.2.1 verify these properties.

The midline effect

The midline effect addresses the case of non-centered initial head positions in gaze shifts and predicts the overall gaze shift behavior in these cases. For instance, for a gaze shift for which the target direction is at the midline, the head is initially in any orientation, and the eyes are initially centered in the head, the head will be moved to the midline. As an important case, if T and H^0 are on the same side of the midline, the eyes are initially centered in the head, and $|H^0| > |T|$ (i.e. the target appears between the head direction and the midline), then the amplitude of the head movement is $T - H^0$ meaning that the final head position is in the same direction than the target (as illustrated in Figure 2.3).

The consideration of this effect has always been overlooked in the literature probably in large part due to the fact that most data on gaze shifts are for the shifts from the midline.

Axiomatic gaze model

The authors of [Hanes and McCollum, 2006] proposed an axiomatic model to summarize the different gaze behaviors. Since the initial eye position has only a small effect, we only provide the details for a function proposed relying on the variables T and H^0 for the cases where the eyes are initially centered and neglect their possible variations. The following are the axioms specifying this model:

Axiom 1 : To each gaze shift target position corresponds an interval of potential head positions. For a gaze shift in which the eyes are initially centered, the final position of the head is the

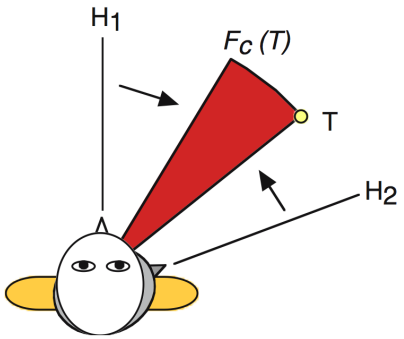


Figure 2.3 – The interval of head positions $[Fc(T), T]$ corresponding to a gaze shift to the target at position T. When the gaze is moved to T from initial position H_1, the head is moved to $F_c(T)$. When the gaze shift is centripetal from H_2 to T, the head is moved to T. For initial head positions between $F_c(T)$ and T, an eye-only saccade to T is made.- from [Hanes and McCollum, 2006]

point of this interval that is nearest to its initial position.

The endpoints of the abovementioned interval for targets T can be deduced from the following two axioms:

Axiom 2 : The direction of the head movement is not opposite to the initial head motor error $T - H^0$ (although there may be no head movement); when the eyes are initially centered in the head, the magnitude of the head movement is not greater than that of the head motor error.

Axiom 3 : When the eyes are initially centered in the head, the final head position is not eccentric to the target position.

Considering the case $H^0 = 0$, where the head is moved to the position $F_c(T)$, we would conclude that the endpoint of the interval nearest the midline must be $F_c(T)$. On the other hand, combining axioms 2 and 3 shows that when the head is initially directed eccentrically to the target, its final position must be in line with the target. Thus T itself is the endpoint of the interval further from the midline.

As a conclusion, for rapid eye/head gaze shifts in which the eyes are initially centered in the head, the endpoint of the associated head movement, (if any) is the point in the interval $[Fc(T), T]$ that is nearest to the initial head position H^0 as illustrated in Figure 2.3.

2.2.3 Discussion on gaze shift models

The models in sections 2.2.1 and 2.2.2 could both be used for predicting the head and eye contributions in a gaze shift. While the model described in section 2.2.1 neglects the effect of the midline on the gaze dynamics, it could still be used in conditions where we assume that gaze shifts begin with the head being close to the midline.

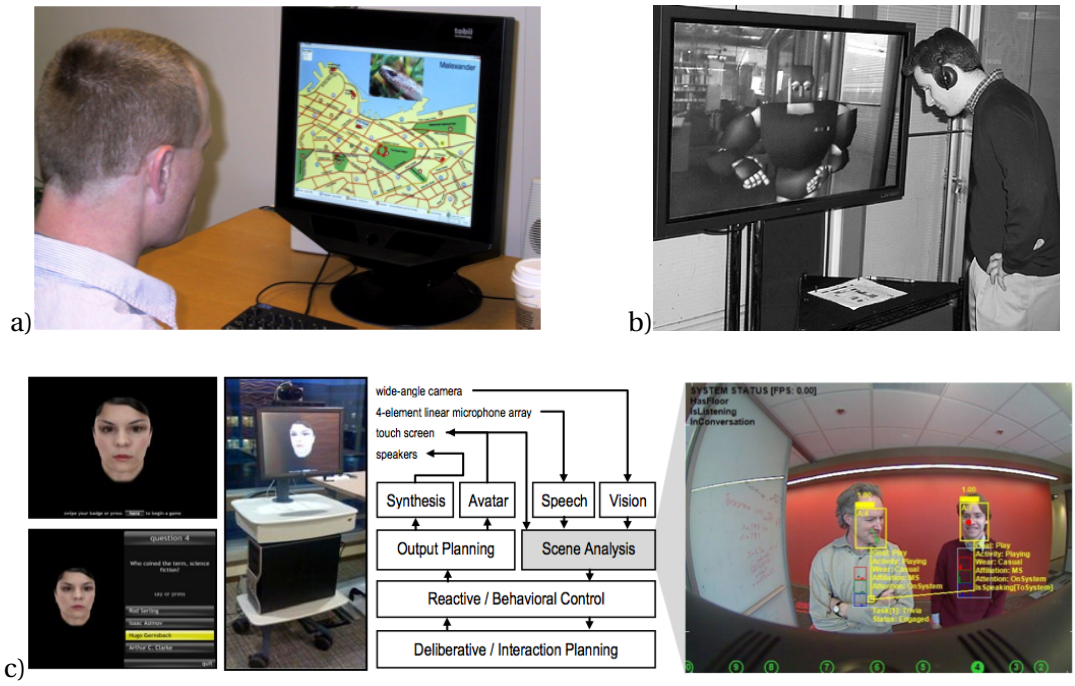


Figure 2.4 – Gaze used in HCI and ECA domains. (a) City trip planning application in [Qvarfordt and Zhai, 2005], tracks the user's gaze fr ground management, (b) Gaze is used in [Nakano et al., 2003] for establishing common ground and updating the dialog in the context of direction giving, (c) the multimodal interactive kiosk in [Bohus and Horvitz, 2009a] performs interaction with multiple users.

Moreover, we can take this model in Section 2.2.2 to define the function $F_c(T)$ introduced in section 2.2.2 and therefore use it as the starting endpoint of the interval in the axiomatic gaze shift model proposed in section 2.2.2. These models will be explained in Chapter 4 as the basis for providing gaze to head pose mappings.

2.3 Gaze in HCI, ECA and HRI

The same communicative functions that we assume for gaze in human-human-interaction as introduced in Section 2.1, can (and should) also hold true in interaction between humans and computers, embodied agents or robots. However, in order to have agents which can leverage these functionalities in the same way, they should also be equipped with two different kinds of capabilities as humans are. The first one is percieving and understanding the users' gaze, and the second one is synthesizing human like gaze behavior. In the following, we briefly discuss how these two directions are addressed in the related works.

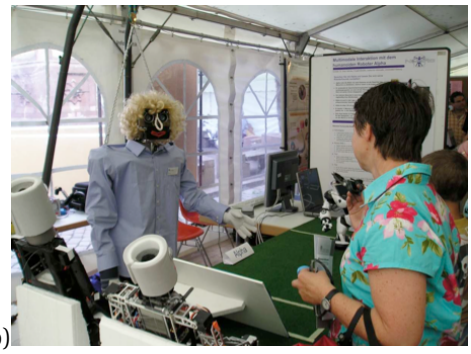
a)  b)  s

Figure 2.5 – Gaze used in HRI. (a) The robot James acts as a bartender and the visual attention to recognize different states of interaction [Foster et al., 2012], (b) The robot Alpha interacts with people as a museum guide and gives different importance to people according to their gaze [Bennewitz et al., 2007].

2.3.1 Perceiving human gaze for interaction with computers and embodied agents

As expected visual attention is extensively used in designing HCI and ECA systems for all purposes including estimating user's attention, facilitating engagement and ground management. For instance, in [Qvarfordt and Zhai, 2005] it is used in city trip planning application relying on a computer. The user gaze moves relate to the spatial contents on the maps, and the system uses this information for managing the dialog. In a similar direction giving task, [Nakano et al., 2003] proposed an ECA which relies on verbal and nonverbal signals including eye gaze to establish common ground and update the dialog. The agent uses eye gaze and other cues for the context of direction-giving task. In other works, head is used as an estimation of visual attention. For instance, in [Huang et al., 2011] this information is used to make a virtual agent aware of the addressee of the utterances.

In one of the very advanced systems [Bohus and Horvitz, 2009a], head pose is used extensively in a multimodal interactive kiosk capable of handling intention recognition and turn-taking (see Figure 2.4). In addition, it can perform multi-party engagement for the times new users arrive while it is interacting with others, to check if they need something and either respond or aks them to wait. Moreover, in [Bohus and Horvitz, 2009b] gaze is used to predict an intention of engagement from people who come in proximity of the dialog system and allows the system to anticipate and start the interaction.

The user's visual attention is also used for facilitating interaction in the HRI domain. For instance, it is used for monitoring the engagement with a robot in [Michalowski, 2006]. In [Bennewitz et al., 2007] it is used as one of the cues to assign importance to different people and decide how the robot acting as a museum guide would share its attention on them. In a more recent work [Foster et al., 2012, Gaschler et al., 2012] the robot James acts as a bartender and uses visual focus of attention as one of the cues to recognize different states of interaction such as "attention to bartender" and "attention to another guest" which are expected in such an interaction. Knowing these states helps the robot to plan the appropriate actions.

2.3.2 Gaze synthesis for embodied conversational agents:

In order to create effective conversational human-computer or human-agent interfaces, it is desirable to have systems which not only can sense a user's gaze and infer appropriate conversational cues but also display them. Embodied conversational agents, either in robotic form or implemented as virtual avatars, have the ability to demonstrate conversational gestures through eye gaze and body gesture [Morency and Darrell, 2008]. Natural gaze behavior is critical to the realism and believability of an animated character. An ECA should employ social gaze for interpersonal interaction and also possess human attention attributes so that its eyes and facial expression convey appropriate distraction and attending behaviors. As discussed in Section 2.1 there are many eye-related communicative functions which should be considered such as eye contact, mutual gaze, gaze aversion, line of regard, and fixation [Gu and Badler, 2006].

In this direction, [Traum and Rickel, 2002] presented an agent which accepts speech input from human, and produce both speech and gestural output. For real-time verbal communication [Colburn et al., 2000] utilizes behavior models of eye gaze patterns based on the psychological literature. They model when the avatar should be looking at the speaker, at the non-speaker or look away. Their simple computational model for eye gaze has been shown to be effective at simulating what people do in conversations and their experiments [Colburn et al., 2000] have indicated that having a natural eye gaze model on an avatar elicits changes in the viewers' eye gaze patterns. In their study, looking at different targets is done by only turning the head.

In [Gu and Badler, 2006] different functions for controlling eye motion are defined by saccade, fixation, smooth pursuit, squint and blink. Parameters to describe these movements include gaze direction, magnitude, velocity, duration, the degree of eye open, blink, and so on and attention models mainly consider all or a subset of these parameters to define suitable effect on the eye. [Gu and Badler, 2006] develops a computational model to predict visual attention behavior for an embodied conversational agent in a dynamic environment and observes its behavior and consequences under varying environmental distractions, conversation workload, and participant engagement.

2.4 Gaze and visual focus of attention estimation

Estimation of the gaze of people or its interpretation is a difficult task. Available eye tracking devices can be used in HCI applications [Qvarfordt and Zhai, 2005], but are usually expensive, considered as intrusive, and usually not applicable for natural interactions. Two streams of work exist for tracking eye gaze. Active sensing based methodologies based on infrared light are used very often to measure the eye gaze. Computer vision techniques based on natural light on the other hand use perceived information from gaze, head and body posture for measuring the gaze. The accuracy of these methods is highly dependent on having high definition images of the eyes. Since this is not always available, people have studied the discrete substitute of the gaze which is the visual focus of attention (VFOA).

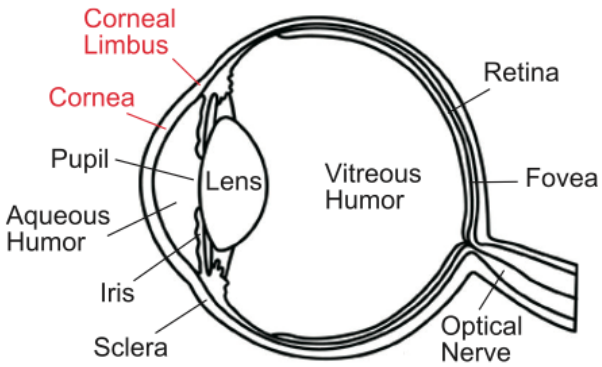


Figure 2.6 – Structure of human eye. Important parts of the eye like pupil, iris, cornea and limbus are shown -from [Nakazawa and Nitschke, 2012]

In the following subsections we give a brief overview on the approaches used for estimating gaze as well as VFOA.

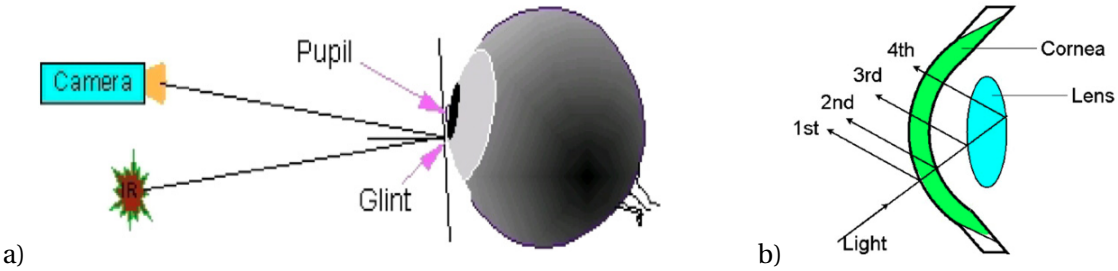
2.4.1 Gaze Estimation

The recent survey by Hansen [Hansen and Ji, 2010] provides a comprehensible overview of computer vision methods for estimating the gaze defined as the direction or the point of regard. Below we summarize the main approaches and issues that need to be addressed.

Gaze tracking based on reflected infrared light (active infrared lighting)

The reflected light eye gaze estimation techniques rely on the amount and direction of the infrared light reflected by specific parts of the eye such as the limbus displayed in Figure 2.6, the pupil, the corneal. When it comes to remote and non-intrusive eye tracking, the most commonly used technique is the pupil-center corneal-reflection (PCCR) technique [Morimoto and Mimica, 2005, Guestrin and Eizenman, 2006] used in many systems (Tobii[1], SMI[2] and EyeGaze[3]). The basic concept is to use a light source to illuminate the eye causing highly visible reflections, and a camera to capture an image of the eye showing these reflections. The image captured by the camera is then used to detect the reflections of the light source on the cornea (glint) and in the pupil center as illustrated in Figure 2.7(a). It is then possible to calculate a vector formed by the angle between the cornea and pupil reflections – the direction of this vector, combined with other geometrical features of the reflections, will then be used to calculate the gaze direction. Furthermore, the reflections from the different components of the eyes create the Purkinje images, shown in Figure 2.7(b), that can be used to estimate

[1] http://www.tobii.com
[2] http://www.eyegaze.com
[3] http://www.eyegaze.com



a) b)

Figure 2.7 – (a) pupil-center corneal-reflection (PCCR) technique, which uses the location of pupil center and glint (reflection of the light source on the cornea) to calculate the gaze direction. (b) the reflections from the different components of the eye create 1st, 2nd, 3rd, and the 4th Purkinje images- from [Morimoto and Mimica, 2005]

the eye gaze in other techniques [Pelz et al., 2000, Ohno and Mukawa, 2004, Babcock and Pelz, 2004].

These eye trackers can be remote and nonintrusive, but still their dependence on specific light sources makes limitations on their usage and on the users movements. Therefore natural light vision techniques are used for less constrained settings.

Natural light computer-vision geometric methods

Natural light based methods that are applicable to more general conditions have also been studied, many of which also make use of geometric models like the ones used in infrared methods.

These methods rely on explicit models of the geometry of the human eye and related parameters, i.e. eyeball location and radius, iris radius, etc. They require to extract geometric features from the eyes, which can be an ellipse fitted to the pupil [Li et al., 2005], or more complex shapes [Yuille et al., 1992]. However, the lower contrast (as compared to IR illuminated cases) makes these extractions more difficult. The main advantages of the geometric explicit modeling are that the required quantity of training samples can be reduced in comparison to appearance based methods (see below) and gaze inference is not based on interpolation. On the other hand, the requirement for geometric features calls in general for high contrast or high resolution image data, captured either from a head mounted camera or from cameras with limited field of views that restrict user mobility.

Natural light computer-vision appearance-based methods

To avoid local features fitting and tracking, there has been an increased interest on appearance based methods [Funes Mora and Odobez, 2012, Lu et al., 2011b, Lu et al., 2011a, Sugano et al., 2008] that learn a direct mapping from the eye image to gaze parameters. Such approaches

potentially allow gaze estimation under low-resolution imaging by relying on machine learning techniques [Baluja and Pomerleau, 1994, Williams et al., 2006] or on user and session specific appearance models [Funes Mora and Odobez, 2012, Lu et al., 2011b]. Altogether, however, appearance based methods suffer from generalization problems. Either they require large amounts of training data [Baluja and Pomerleau, 1994, Williams et al., 2006, Sugano et al., 2008] to handle variabilities due to eye shape, pose, illumination conditions, or they are trained from session dependent samples [Lu et al., 2011b, Lu et al., 2011a, Funes Mora and Odobez, 2012] to be used for interpolation. In both cases, the absence of an explicit geometric model makes them rather inappropriate for adaptation to users or ambient conditions, or extrapolation in the 3D space, which is problematic when training from a few points on a screen and estimating gaze for different head poses.

2.4.2 Estimating the visual focus of attention

In situations where head pose and eye gaze can be achieved accurately, recognizing the VFOA of people sould be straightforward assuming of course that we are able to monitor the environment and to know where are the visual targets together with their directions with respect to people's heads. However, we know that active sensing based methodologies based on infrared light are quite invasive and restrictive [Babcock and Pelz, 2004]. Computer vision gaze estimation techniques on the other hand are highly dependent on having high definition images of the eyes and therefore using them usually restricts the mobility of the subject because cameras with narrow field-of-views should be used. Due to the abovementioned reason, reliable eye gaze information is often not available for HRI applications like our scenario.

As an alternative to the eye gaze, researchers have considered head pose information as main gaze cues [Gaschler et al., 2012, Gorga and Otsuka, 2010, Yücel and Salah, 2009, Stiefelhagen, 2002, Ba and Odobez, 2006, Otsuka et al., 2005]. This idea is supported by the fact that turns of the head are very informative cues in recognizing where the subjects are looking at [Langton et al., 2000]. The models described previously in Section 2.2.2 specifically emphasize on the importance of the head pose for looking at the visual targets and provide models for showing how it contributes to gaze shifts. Moreover, experimental work with adults, children and non-human primates has suggested that the orientation of the head makes a large contribution to the understanding of another's direction of attention [Langton et al., 2000] (even wothout seeing the eyes). Stiefelhagen studied this hypothesis in [Stiefelhagen, 2002], when he conducted an experiment in which he recorded the head and eye orientations of participants in a meeting using special tracking equipment. The results demonstrate that head orientation was a sufficient indicator of the subjects' VFOA in 89% of the time.

However, head poses are ambiguous cues for recognizing VFOA: in realistic scenarios, depending on the dynamics of the head different poses might be used by one person for looking at the same target and the same pose can be used to look at different targets. Thus, a central

question is, what head pose is used for looking at a target placed at a given direction. The difference between a gaze direction and the head pose used to look in that direction depends on several factors related to the body, head and eye dynamics. Therefore, the first part to improve for removing the ambiguities is the prediction of the gaze direction from the head pose, and allow for a better association of a head pose with looking at a given target given the ongoing head and gaze dynamics.

Several works explored Dynamic Bayesian Networks (DBN) for decoding VFOA states from the head pose sequence [Stiefelhagen, 2002], [Ba and Odobez, 2009], and [Otsuka et al., 2005] and relied on Gaussians to model the distribution of head pose for looking at a given target. Again an critical issue with this approach is how to set the expected head pose of an observer that looks at a given target? In other words, how to define a mapping from the gaze target direction to the corresponding head pose. These expected head poses for different targets are the means of the Gaussians and important parameters of the model.

In order to set these parameters and make the association between head pose and visual targets, some works rely on manual setting, potentially followed by adaptation [Otsuka et al., 2005]. This requires a static configuration of the targets with respect to the person's position and is not accessible to all applications. Data driven approaches [Gaschler et al., 2012] use training data to directly infer VFOA from head pose without defining gaze as an intermediate step. Learned parameters, however, are then again specific to the geometric configuration between the sensor (robot), the person, and VFOA targets. While this might be suitable in fixed settings [Gaschler et al., 2012], it is not adapted for a mobile robot dealing with moving people.

One of the few works addressing the headpose-to-gaze correspondence problem is [Ba and Odobez, 2009]. Exploiting results on human gazing behavior and head-eye dynamics involved in saccadic gaze shifts [Langton et al., 2000, Hanes and McCollum, 2006], [Hanes and McCollum, 2006, Freedman and Sparks, 1997, Hayhoe and Ballard, 2005], they introduced a simple linear gaze model based on what we described as the first gaze shift model in section 2.2.2 to the head pose, gaze direction, and head reference (coined gaze midline in section 2.2.2) for gaze shifts. Despite its simplicity, the method worked when applied to meetings with static configurations. However, it is not very efficient in dynamic scenarios and suffers from a major drawback: the reference direction, which corresponds to the direction perpendicular to the shoulder, was assumed to be fixed and set according to the setup. This assumption does not hold true in realistic situations with dynamic settings. For instance in HRI with multiple people where the robot is not always the main focus, or more generally in scenarios involving people free to move and re-orient themselves.

One study has tried to address the problem of the previous approach with dynamic setups [Voit and Stiefelhagen, 2008]. However, they do not consider changing the static reference which is the main reason the model does not apply properly to dynamic situations. Alternatively they propose to use a discrete set of different head-to-gaze ratios and choose the most likely ratio over time based on the existing gaze dynamics.

2.5 Context in behavior understanding and VFOA recognition

Considering the fact that gaze and head pose are imprecise, when the number of targets increases, the chance of erroneous decisions increases as well. One solution to remove some of these errors is to know which are the 'active' targets at a given instant. In order to obtain this kind of information about the targets, other contextual cues can be very useful.

Context is extensively used for different tasks in related domains. In computer vision for instance, context has been used to improve the recognition of individual objects given the current overall scene category [Torralba et al., 2003]. For speech recognition, in [Sarma and Palmer, 2004] context learnt from the lexical co-occurences of the words in a large corpus of the outputs of an ASR system is used for improving the output of automatic speech recognition systems.

One possibility to further improve the VFOA recognition is to use other social and behavioral cues, leveraging on the fact that the recognition of nonverbal cues should not be done in isolation, but jointly, as some behaviors provide context to the others. This could be useful both for compensating the problems introduced by the model's limitations and behavioral variations and also to improve the recognition in presence of noisy measurements. In human-human interaction, examples of these additional cues include speaker information [Stiefelhagen et al., 2002, Ba and Odobez, 2008] or higher conversational states [Gorga and Otsuka, 2010, Otsuka et al., 2005], that can be complemented with group activity [Ba and Odobez, 2008]. Similar behavior co-occurrences have also been used for instance in a head gesture recognition task [Morency, 2009].

While in the above cases the social cues used as context have to be inferred from the data and might be noisy, in the robotic or ECA cases, the agent is fully aware of its own conversational acts, allowing them to be conveniently exploited to better interpret the nonverbal cues performed by interacting people. For instance, in [Lemon et al., 2002], the grammar of the speech recognizer changes depending on the agent's previous action or utterance which makes improvement on the speech recognition output. As another example, in [Morency et al., 2005], different types of features (lexical, timing, gesture displayed) performed by an ECA are exploited within a supervised learning framework to predict head nods and head shakes in combination with a vision-based head gesture recognizer.

Several facts about human behavior during their interactions with other humans or agents strongly support the use of the agent's conversational acts as context for VFOA recognition. For instance, the fact that people look nearly twice as much while listening than when they speak [Argyle and Cook, 1976] supports leveraging the speaking status of robot. This is specially important since comparisons conducted between human-human and human-robot interactions in [Rehm and Andre, 2005], revealed that people spend more time looking at an agent that is addressing them than if it is a human speaker. The second example is the fact that people look at objects relevant to what they listen to [Cooper, 1974], which makes it reasonable to use the context obtained from the robot's conversation to extract the relevant objects.

2.6 Addressee Estimation

Addressee estimation which is to recognize to whom a spoken utterance is intended can be performed by using verbal and nonverbal cues simultaneously. This problem has not received much attention in the HRI literature (except [Katzenmaier et al., 2004]) as compared to Human Computer Interaction (HCI) / Virtual Avatar [van Turnhout et al., 2005, Bohus and Horvitz, 2010, Huang et al., 2011, Siracusa et al., 2003] or Human-human interaction literature [Takemae and Ozawa, 2006, Jovanović et al., 2006]. In all cases it has been shown that from the nonverbal cues, eye gaze serves an important role in guiding the conversation, and also is an important cue for determining the addressee.

In human-human interaction context, [Jovanović et al., 2006] authors have studied the addressee identification in face-to-face meetings. They used gaze, conversational context and utterance features and added an additional feature above these which specifies the meeting context as being monologue, discussion, presentation or white-board. Their result showed that speaker's gaze is the most predictive cue and performs better in combination with other cues. In contrast to the statistical classifier used in [Jovanović et al., 2006], another study [op den and Traum, 2009] provided a rule based addressee detection method for face-to-face meetings. They used speakers gaze, dialogue history, usage of addressee terms and the type of the dialogue act as features. A rule based method is more transparent than the statistical classifiers synthesizing empirical findings of addressing behavior in conversations. They have analyzed their methods on the same multi-modal AMI meeting corpus which has been previously used for developing statistical addressee classifiers in [Jovanović, 2007]. Their reliability analysis has shown that in specific situations this rule-based method outperforms the statistical methods. For instance, when the speaker uses "you", or when the speaker performs an initiating act, supported by visual attention directed to the addressed partner, the method outperforms the statistical methods.

In [van Turnhout et al., 2005] the problem of determining addressee of an utterance in the context of multimodal mixed human-human and human-computer interaction is studied. They indicated that in this context eye gaze behavior cannot directly be used as a cue for determining addressee. The reason is that a computer or robot with has a central role in the task that participants are performing, highly attracts their visual attention and changes the normal gaze behaviors. Katzenmaier et al. [Katzenmaier et al., 2004] used a Bayesian scheme to combine speech features and head pose to solve the task for human-human-robot interactions. They could identify the correct addressee in 93% of time. Commands towards the robot could be detected with a recall of 0.8 and a precision of 0.6, resulting in an f-measure of 0.7. Alternatively In [Huang et al., 2011] prosodic features of the user's speech are used in addition to the head pose (as a proxy for gaze).

Considering this kind of setting, users tend to speak differently to systems which is potentially useful for detecting when the robot is addressed. Moreover, what robot is doing or has just done can structure the conversation or evoke specific reactions from participants. The fact that

we have access to the robot's state motivates using these kind of cues for human-human-robot interaction scenarios.

2.7 Conclusion

In this Chapter we provided some background emphasizing the importance of gaze in human interaction and in the same way in interactions with robots, other conversational agents, or computers. Moreover, since humans use a combination of body, head and eyes dynamics to gaze at visual targets surrounding them, we summarized some studies aiming to model different contributions envolved in gaze mechanism. These results open the doors for leveraging head poses for estimating the persons gaze direction.

Moreover, we reviewed the main classes of methods commonly used to address the estimation of gaze and recognition of VFOA. We specially summarized what has been done for the situation where percieving the eye gaze is not possible from the high defintion natural images or infrared sensors, which is suitable for our scenario. The shortcomes of the previous works for dealing with dynamic scenarios in HRI domain motivates our contributions in designing the VFOA recognition algorithms. Given the ambiguities present in recognizing VFOA from head poses, we provided an overview of the literature where context is used in human behavior understanding and more specifically VFOA recognition. Considering the previous works and the possibility and benefits of using the robot context, we would consider different kinds of robot conversational context for improving the VFOA recognition.

Finally for addressee estimation, previous works suggest gaze as the most informative cue. However, since in human-agent interactions scene elements referred to in the conversation can attract the users gaze and changes their behavior, using additional cues as context for addressee estimation can potentially be useful. In this thesis we would study how well addressee could be estimated using VFOA information and whether or not adding other cues, possibly from the robot's state, could improve the estimation.

3 Datasets

3.1 Introduction

In this chapter we will present different datasets used in this thesis for the experiments since at the beginning we did not have a suitable dataset for VFOA recognition compatible with the goals of the HUMAVIPS project, we started performing experiments and testing our VFOA recognition algorithms on the IDIAP Head Pose dataset[1]. This dataset consists of recordings featuring 4 people meeting around a table and can therefore be used for human human interaction studies. Afterwards we recorded a small dataset NaoD, in order to work with similar video specifications (image size, quality, frame rate) than those we would expect to have with Nao in the project. This involved a simple scenario where people are seating in front of the robot and discussing some information about it. However, while the robot can be the focus of attention, it is mainly used as a passive sensor.

Considering our requirements for natural human robot interaction in unconstrained conditions, these datasets were not fully appropriate. Specially one condition was to include people that could move more freely in terms of their position with respect to the robot. This condition was not met in the seated scenarios of the two previous datasets. More generally, we needed a dataset recorded with the robot's sensors for studying communication cues like VFOA and addressee and provides us with natural and rich behavior. In order to reach this goal we designed a scenario where the robot performs as a quiz master. It engages with a group of people, addresses individual persons and asks them questions one at a time, and waits for the answer after they discuss it among themselves, which results in a mixture of human-human and human-robot interaction.

People can enter the robot's field of view, either they try to engage or the robot engages them, and they start the interaction with the robot while being quite dynamic. They can talk to each other or to the robot, and this makes it possible to study the addressee estimation task.

On the other hand, one objective and requirement of the HUMAVIPS project, was to design

[1] https://www.idiap.ch/dataset/headpose

a robot (Nao) able to interact and manage a group. The scenario chosen to this end was to consider Nao as an art guide explaining artworks for groups of people in a museum. Eventually we included the benefits of both scenarios by mixing them into a unified dataset and recorded the Vernissage corpus where the robot starts by giving explanations about the artworks as in the original art guide scenario and ends by giving a quiz to the participants. We contributed in designing the scenario, and the script for Nao, implementing the wizard of Oz for controlling the robot, and participated in recording the dataset which was done at Bielefeld University using their effort and expertise and infrastructures in recording multimodal data with robots.The details of this corpus are previously published in a technical report [Jayagopi et al., 2012] and a conference paper [Jayagopi et al., 2013] in collaboration with Dinesh Babu Jayagopi and other colleaques from Idiap Research Institute and Bielefeld University.

In this Chapter we will provide a description for these 3 datasets. First we will describe our main project dataset, Vernissage Corpus which is used for most of the experiments in this thesis in Section 3.2. Then we will present the second dataset used for our VFOA experiments, NaoD data, in Section 3.3. In Section 3.4 we will describe the Meeting dataset, the initial dataset we used for VFOA recognition and we make a conclusion on this Chapter in Section 3.5.

3.2 The Venissage dataset

One of the fundamental challenges in HRI is providing humanoid robots with the audio-visual perception capabilities to interact with multiple human partners [Fong et al., 2003]. Towards this goal, realistic interaction scenarios need to be studied, in which the humanoid robot actually performs nonverbal behaviors which as a result will induce natural nonverbal behavior of the humans, e.g. looking towards a picture when the robot indicates this or looking at the human partner when discussing a painting. In this view to be a realistic interaction partner, the humanoid robot needs to perform appropriate actions, for example nodding, or gaze changes to point to paintings, or look at specific participants performed in the current case by head rotations. Although these actions are desirable from an interaction perspective such behavior severely degrades the sensing quality, as the sensor is moved and motor noise is added. In addition, as sensing, computing, and communication capabilities on the robot are limited and constrain each other, such scenarios become much more difficult when it comes to study and address the perceptual tasks.

A first step towards this goal is thus to obtain suitable datasets for these scenarios which allow the study of those perceptual tasks in interaction with a robot. In this regard in collaboration with Bielefeld University we aimed to provide a dataset "Vernissage Corpus"[2] [Jayagopi et al., 2013, Jayagopi et al., 2012] recorded by the robot Nao shown in Figure 3.1. According to our knowledge, none of the existing HRI datasets which have focused on audio-visual perception tasks [Mohammad et al., 2008, Arnaud et al., 2008, Alameda-Pineda et al., 2013] in a conversational scenario have all the advantages of our dataset: an interesting scenario,

[2] Vernissage is the French word for the opening of an art exhibition

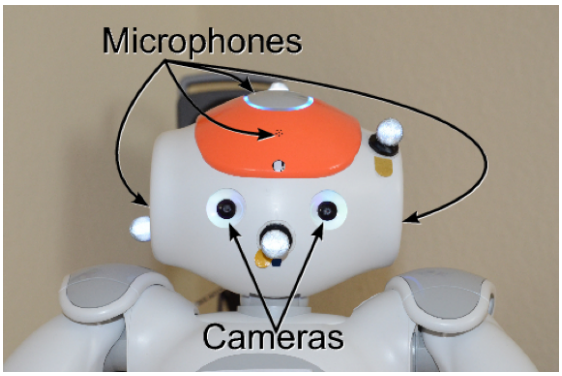


Figure 3.1 – The humanoid robot NAO and its primary sensors used for the recordings. VICON markers (silver balls) for motion capturing are visible.

more than one interaction partner, a commercially available robot (with consumer sensors rather than high-end sensors), extensive annotations, and public availability. The dataset we have produced comes with rich annotations and ground-truth from the external sensors and robot internal states. This allows researchers in multimodal perception community to investigate interaction behavior cues at a low level (such as 'who is speaking', 'who is looking at whom', 'nodding') as well as at a higher level (such as 'who is being addressed', turn-taking or conversational behavior).

3.2.1 Scenario and recordings

In order to capture a dataset that can be used for HRI analysis as well as to test various audio-visual perception techniques, we decided to choose the Vernissage scenario, where the robot serves as a conversational partner in a reasonably realistic application setting: as an art guide and a quiz master. As an art guide it would be involved in managing the group as a whole and less in individual communication with people, whereas as a quiz master it would be more engaged in multiparty interactions which needs communication exchanges between the individuals. This scenario offers sufficient flexibility as well as control over the human-robot interaction. The first part of the scenario was inspired by a recent work that has studied and documented human interaction experiences with NAO as an art guide in a German art museum [Pitsch et al., 2011]. In this scenario, as the robot is stationary (except for head rotations and nods and hand gestures), the complexity involved in adapting and extending existing perception methods is reasonable, but still challenging.

We recorded 13 sessions (10 main sessions with naive participants and 3 test sessions with project collaborators) of the humanoid robot, NAO, interacting with two persons. A wizard-of-oz was used to manage the dialog as well as the robot's gaze and nodding. The behavior of the human partners was unconstrained. Each interaction lasted around 11 minutes. Figure 3.2 gives an overview of the corpus.

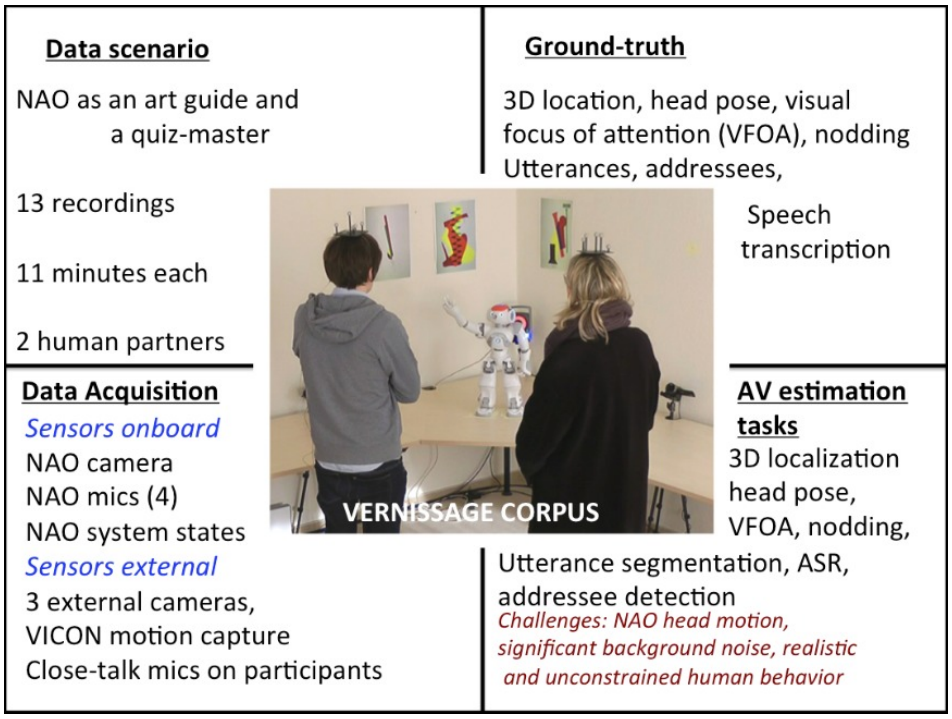


Figure 3.2 – Overview of the Vernissage corpus: scenario, various modalities, annotations, and possible audio-visual perception tasks.

In the Vernissage corpus, the scenario unfolds as follows:

- The visitors arrived in pairs and were greeted by the robot when they entered within a normal interacting distance. After this greeting, the robot offered some explanations about the paintings present in the Vernissage.
- When the visitors agreed to this[3], the robot started explaining three different groups of paintings using speech and matching gestures. These explanations included pauses intended to elicit comments by the visitors and also gave them the chance to tell the robot if they wanted to hear another explanation at specific points.
- When the explanations were finished, the robot asked the visitors if they were interested in participating in a quiz. After they agreed to this, NAO introduced itself and asked each participant to give their name and to introduce themselves.
- The robot then explained the general quiz rules which included that the visitors should discuss among themselves before giving the answers. The robot then proceeded to ask several questions about the paintings and more general topics and also judged the answers given by the participants.

[3]Although the robot was asking the participants whether they want to continue, as they were expected to finish the interaction all of them proceeded with the full scenario.

- After the quiz was finished, the robot asked each visitor to decide on a favorite painting and afterwards told the participants to discuss and choose one common favorite and also to propose a new fitting name for that painting.

The participants spoke in English and they were mostly non-native speakers recruited in a university environment. Figure 3.3 shows the images of 20 participants in 10 main sequences of the dataset taken from the Nao's video sensor.

Wizard-of-Oz. To govern the behavior of the robot in a repeatable fashion, we used a "Wizard-of-Oz" (WOz) approach [Dahlbäck et al., 1993]. This means that the robot was not acting autonomously, but instead was controlled by human operators. For our recordings, we mostly used two operators (or "wizards"), which worked in a separate room hidden from the participant's view. One operator controlled the utterances and associated gestures of the robot by choosing them from a predefined set of buttons. Limiting the set of possible robot utterances like this was meant to reduce the gap towards an autonomous system with a real dialog engine. The second operator controlled the viewing direction of the robot by choosing points in the live streamed camera image, causing the robot to turn its head in that direction. In addition to these specialized interfaces, both operators also had access to the sound coming from the microphones the participants wore and the live image of an external camera providing an overview of the interaction. To facilitate later analysis, the button clicks from the wizard interface were also logged as part of the corpus.

Set-up and sensors. NAO video data is mono at VGA resolution and audio data comes from four microphones. In order to have ground-truth information for all the audio-visual processing tasks, 3 close-field external cameras, Vicon[4] motion capturing system and close-talk microphones on the human interaction partners were also deployed.

The final dataset comprises of a synchronized multimodal corpus, with multiple auditory, visual, and robotic system information channels. Details on data acquisition and synchronization is provided in [Jayagopi et al., 2012].

Data acquisition was inspired by the SInA method [Lohse et al., 2009] which focuses on synchronizing internal logging data with external manually annotated data in order to analyze specific issues of HRI.

3.2.2 Annotation and measurements

Two types of annotations or ground truth (GT) are made available with the dataset. First, the GT data derived from the recorded measurements, which comprises:

- the participant's 3D head location and poses (in Vicon reference system). In order to be compatible (comparable) with the output of visual processing systems, we used a

[4] http://www.VICON.com

Figure 3.3 – Image samples from 20 annotated people from the Vernissage corpus. As it can be seen, people locate themselves freely inside the room and they are free to rotate towards different pictures. VICON markers are visible on participants heads.

Figure 3.4 – Visualization of annotations: head-pose (pan above and tilt in right of the head bounding box), VFOA (in yellow), addressee in blue (displayed when speaking), and nodding in green. When speaking, bounding box color is partially blue.

simple software to transform the 3D head pose Vicon data into head pose measurements defined in the local coordinate system of the time-dependent Nao camera view.

- the directions of looking at different targets with respect to the participants. In order to obtain the directions of targets with respect to the participants, their 3D head locations as well as the robot and the target locations were used from the Vicon data. This information made it possible to calculate the direction of any of the targets with respect to the participants' heads. These directions were then transformed into Nao's head local coordinate system. In our work, we considered the direction of Nao to be $(0,0,0)$ and measured the other angles according to this reference.

- NAO system data, including its dialog and gesture information.

The second group of ground truth corresponds to the manual annotations which include several important cues to study the HRI process and analyze the verbal and nonverbal behavior patterns. These include 2D head location, speech/non speech, head nodding, Visual Focus of Attention (VFOA), utterances, and addressees as illustrated in Figure 3.4. In the following, we describe our annotation process for the cues which are used in this thesis.

- **Head Location.** We annotated the 2D image location of people in the recording. Exploiting the VICON 3D location data was not possible since it did not localize the 2D head bounding box in the image captured by NAO as NAO's camera is constantly changing its orientation and position. We thus resorted to simple manual annotation of the visibility, ID, and position (bounding box) of each person. Annotation was done at 1 frame per second. Interpolation was automatically generated, and manually revised (i.e. interme-

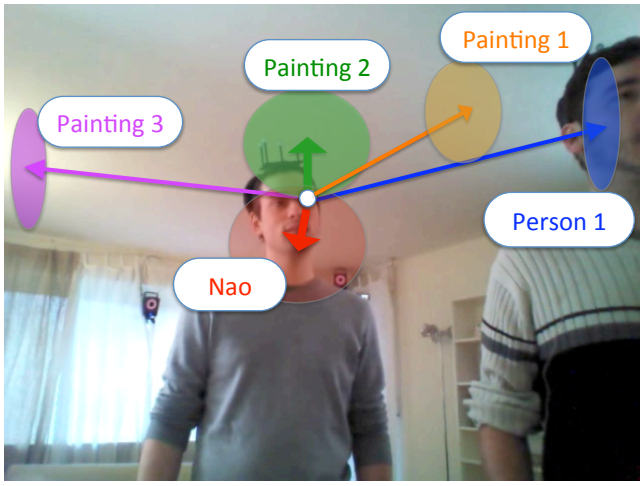


Figure 3.5 – Rough illustration of the configuration of VFOA targets in the scene. In the annotations we also have labels OT, DK and NV in addition to these to these main targets.

diate frames were annotated), whenever these interpolations deviated too much from the true head location, to have sufficient accuracy at important transition points.

For this thesis, since tracking results will be used in the experiments, it is necessary to have the head pose ground truth to validate the tracker quality.

- **Visual Focus of Attention (VFOA).** Given the scenario, 5 main VFOA targets have been identified and considered as labels. They are: NAO, OP (the other participant), and the three paintings Pai1, Pai2, and Pai3. In addition, we defined a label OT (others) to denote a person looking at any other place in the room, and a label DK (don't know) when there is too much ambiguity between several VFOA targets and making a decision for the annotation is not possible, and NV (not visible) when the person is not in the robot's field of view. Figure 3.5 illustrates the approximate configuration of different targets in the scenario. Annotation was performed by several annotators. Each annotator performed the labeling using an interface displaying the images of the video acquired from NAO (i.e. taking the robot perspective). Annotation was done with a precision of 150 ms on the average. VFOA statistics for this dataset and the following two datasets are provided in Chapter 5.

  Reliability. We carried out secondary annotation on 2 minutes of data for 15 randomly chosen people among the total 26 participants. Table 3.1 contains the confusion matrix for our two annotation sets with the five main labels of interest. As seen from the table, the confusion between NAO and Painting 2 was high as the painting was right above NAO as seen in Figure 3.2 Apart from this, annotations are very reliable.

- **Utterance.** An utterance is the basic speech unit and following the literature on addressee detection, we defined it as 'a speech turn followed by silence more than 0.5 seconds' (e.g. [Takemae and Ozawa, 2006]). We decided to also include a 'Laughter' label to differentiate actual speech turns and laughter, so our three labels were. were *Speech*,

Label	NAO	OP	Pai1	Pai2	Pai3
NAO	21221	22	15	1502	46
OP	6	3812	132	2	33
Pai1	36	1	4617	110	48
Pai2	894	5	29	5177	47
Pai3	22	415	0	44	2576

Table 3.1 – Confusion matrix of VFOA annotations. Primary versus secondary for VFOA annotation. The table shows, in number of frames, the aggreement and disaggrements of the primary and secondary annotations.

Label	NAO	OPerson	Group	NoLabel
NAO	238	3	0	0
OPerson	11	242	0	0
Group	12	3	40	0
NoLabel	0	0	0	67

Table 3.2 – Confusion matrix - Primary vs Secondary for addressee annotation

Silence, Laughter. As the task of manual segmentation and then assigning a label is quite cumbersome, we used a semi-automatic approach. We started with an automatic method (speech activity detection by cross-talk suppression) to obtain the speech/silence segmentation. Then an annotator revisited and adjusted the segmentation and labels. This process was carried out using the ELAN graphical interface. Each recording has an average 60 utterances. The average duration of an utterance being 1.3 seconds.

- **Addressee.** Addressee is the person or group of people to whom 'a speech utterance is intended to'. Given the scenario, we are interested in labeling the addressee of the utterances from the two human participants. We assigned the following labels: {NAO, PRight, PLeft, Group, NO LABEL}. PRight and PLeft are the persons to the right and left of NAO. Group label corresponds to the situation where one participant addresses jointly NAO and the other participant. We assign NO LABEL if the current utterance has no addressee or if it is a speech act like 'Laughter'. The labeling of each utterance was done by one annotator having full access to the audio-visual recording. The GROUP label mainly occurred during the self-introduction phase. 13 interactions were used to compute the statistics.

  Reliability. A secondary coder performed the annotation for 4 out of 13 interactions (i.e. 30% overlap). The results show that Cohen's Kappa, the interannotator agreement, was 0.93, meaning they are infact quite reliable.

Figure 3.6 – NaoD dataset sample images from the first and second phases of the recording. P1 and P3 are seating on the left and right sides in the left image. P2 is seating on the left side in the right image.

3.3 NaoD Dataset

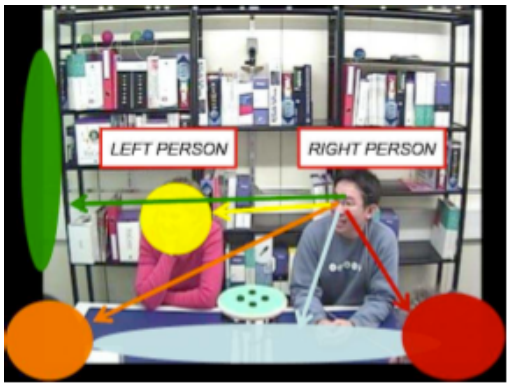
The second dataset (NaoD) is a small dataset recorded with Nao sensors before the Vernissage corpus. Here, in contrast to the previous dataset, Nao is not actively interacting with people and is used as a passive sensor. However, given the scenario which is explained below, the robot is one of the main visual targets. This makes the data close to what we get in HRI scenarios with respect to the attentional behaviour of people.

3.3.1 Scenario

The scenario considered for this recording is the following: there are two participants seating on a couch in front of the robot as shown in Figure 3.6. One of them introduces Nao to the second one and talks about the robot's features and capabilities. The second participant tries to remember these information and asks questions if neccessary. He is supposed to give the same introduction to a third participant who joins later (Figure 3.6 on the right). At the second phase of the recording, the third participant replaces the first one and gets the information from the second participant. Participants show things on the robot and point to it during the introduction.

3.3.2 Recording

We made one recording of the described scenario resulting in 18 minutes of data in total. The three participants P1, P2, P3 involved in this recording are shown in Figure 3.6. Nao video data is mono at VGA resolution. In this dataset there is no audio information since at that time of the recording, joint audio-video acquisition was technically not possible.

a) 

b) 

Figure 3.7 – (a) Meeting data setup, showing the two participant who are seating in front of the organizers, together with their possible VFOA targets. (b) Meeting data sample image.

3.3.3 Annotations

For this dataset we only annotated the visual focus of attention states. Each of the participants have 3 visual targets: the other participant, Nao and a booklet which they refer to during the recording. Like with the previous dataset, labelling was performed using an interface displaying the images of the video acquired from NAO (i.e. taking the robot perspective).

3.4 Meeting Dataset

The third dataset we used is the meeting room recordings of the IDIAP Head Pose Database (IHPD) which was used for VFOA analysis in the previous thesis [Ba, 2007]. We use it in this work to evaluate our VFOA recognition methods on other datasets in addition to our data recorded by Nao.

3.4.1 Scenario and recordings

In this dataset we have the recordings of 8 meeting sessions with a total duration of 145 minutes. All of the meetings are recorded under the same condition and with similar configuration as shown in Figure 3.7, with 4 people (Person left Pl and Person right Pr seen on the image, and two organizers O1 and O2 seating in front of them) discussing statements displayed on slides. We perform our study on the two persons on the seats in front of the camera.

During the recordings the participants were first asked to write their name on a sheet of paper on the table and discuss statements displayed on the projection screen. The scenario gives full freedom to the participants about their head motion, pose and gestures. People were acting naturally as in real meeting situations. The meeting lengths vary between 7.6 to 14 minutes, thus studying the visual focus of attention in these recording is interesting. The recordings are long enough to exhibit a wide range of gazing behaviors.

3.4.2 Annotation and measurements

For this dataset ground truth head poses are available and captured from flock of bird (FOB) sensors.

Annotations for the visual focus of attention(VFOA) of the participants are also available with the dataset. Each of the participants has 5 possible gaze targets: 3 other persons, the slide screen and the table.

3.5 Conclusion

In this chapter we presented three datasets used for the experiments in this thesis. The meeting dataset was available from the previous studies on human-human behavior analysis in the meeting context. The NaoD dataset was recorded in the early stage of our research to provide data with the same specifications we would expect to get from robot Nao sensors. However, considering that Nao is only an interesting object and not engaged in interation with the participants, this dataset does not contain similar behavior that people would show while engaged in a multiparty interaction where the robot has a key role in the conversation. Consequently to obtain a more suitable dataset the Vernissage data was recorded in collaboration with our project partners in Bielefeld University thanks to their recording infrastructure.

The senario used in the Vernissage corpus has two different parts. In the first part, the robot starts by giving some explainations about the paintings surrounding it to the participants, and in the second part, it gives a quiz and individually addresses one person at a time, asks him/her a question and expects him to answer. Given the scenario, participants' unconstrained behavior, and the sensors, there are lots of challenges for addressing the perceptual tasks. For VFOA analysis, challenges are due to the Nao sensors, and different body poses and gaze behaviors of people.

4 VFOA recognition models

4.1 Introduction

In this chapter, we will address the recognition of VFOA in HRI or Embodied Conversational Agent (ECA) settings. Measuring and interpreting the gaze of people is a difficult task as mentioned in Chapter 2. Especially in HRI applications, eye tracking devices are not a suitable options and high definition images are not accessible because people are not necessarily close to the robots. Therefore in the absence of eye gaze information we will mainly rely on the head pose input. Considering the advances in computer vision tracking systems, other researchers have largely considered head pose as an approximation of the gaze [Nakano et al., 2003, Foster et al., 2012]. However, interpreting the head pose as looking at VFOA targets remains ambiguous since in realistic scenarios, the same pose can be used to look at different targets depending on the situation. In this chapter, we will propose our novel Input-Output HMM (IO-HMM) combining two complementary approaches to improve VFOA recognition as described below.

As for the first approach, we explored the head pose-gaze correspondence. As mentioned in chapter 2, one of the few works addressing this problem without being dependent on the specific given configuration, is by [Ba and Odobez, 2009]. Exploiting results on human head-eye dynamics involved in saccadic gaze shifts [Langton et al., 2000, Freedman and Sparks, 1997, Wang and Jin, 2001], as described in the first model in Section 2.2.1, they introduced a gaze model relating the head pose, gaze direction, and body orientation. In this thesis we followed this approach and proceeded by addressing its two main drawbacks for targeting dynamic settings. The first drawback is that the body orientation was assumed to be fixed and set according to the setup. This approach is not feasible in more dynamic settings where people are free to move and re-orient themselves, as illustrated in Figure 4.4. The second drawback, pointed out in several psycho-visual works, is that the mapping not only depends on the the gaze direction and body orientation, but also on the head or gaze direction before the shift, resulting in different head poses for looking at different targets even for the same head reference direction. We propose models relying on a time-varying and implicit estimation of

the body orientation to implement dynamic gaze-to-head mapping and gaze shift models inspired by [Hanes and McCollum, 2006] to address the above mentioned problems. As shown in our HRI scenario, these models considerably improve quantitatively the accuracy of the predicted head pose used to look at VFOA targets, and VFOA recognition as a consequence. These models have been previously published in a workshop [Sheikhi et al., 2012] and a conference paper [Sheikhi and Odobez, 2012] in collaboration with a post doctoral researcher, Vasil Khalidov.

As for the second approach, we considered using other conversational cues considering the fact that some behaviors provide context to the others. Since social cues have to be inferred from the data and might be noisy, the robot context could be considered as a better option for contextual information. Given that in the robotic or ECA cases, the agent is fully aware of its own conversational acts, allows them to be conveniently exploited as context to better interpret the non-verbal cues performed by interacting people. We propose to benefit from the HRI context by exploiting two types of robot dialog acts that can influence VFOA expectations: communicative acts (people look more at speakers, including the robot) and verbal acts (references to scene objects). This second approach has been published in [Sheikhi et al., 2013b], with contribution from V. Khalidov and our partners at Bielefeld university (D. Klotz and B. Wrede) w.r.t. to the automatic extraction of dialog acts. The combination of both the dynamic gaze model and contextual approach is submitted as a journal paper [Sheikhi and Odobez, 2014].

The chapter is organized as follows. Section 4.2 provides an overview of the approach, while the Sections that follow describe the baseline algorithm (Section 4.3), the novel gaze dynamical mapping (Section 4.4), and the contextual model (Section 4.5). Section 4.6 concludes the chapter.

4.2 Approach Overview

Our objective is to monitor the visual attention of people in a given environment relying on head pose since eye gaze is not directly accessible in our intended scenario. To address this problem, we assume to have a specific set of visual targets $\mathbb{F}$ which are of interest in our given context. We would like to recognize which of these targets a given person is looking at.

As a reminder, the main robotic setup which we have considered is based on the Vernissage scenario and database shown in Figure 4.1 (a), and that we have described in Chapter 3, where a robot acts as an art exhibition guide, providing explanations about artworks placed around it, and in a second phase, giving a quiz. Recognizing what or whom people are looking at in this context gives useful information about their attention to the robot and whether they follow the explanation or not which could be used to decide how to proceed in the conversation.

Figure 4.1 (b) shows an illustration of the Vernissage setting with the robot, participants and

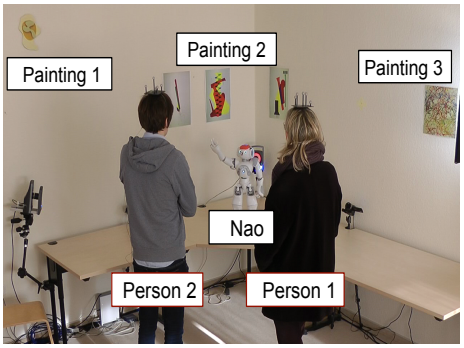
a)  b) 

Figure 4.1 – a) Our Vernissage scenario: the robot is explaining artworks as an exhibition guide. b) Vernissage data considered for evaluation: in the recordings the robot explains 3 groups of paintings to participants, and then gives them a quiz. Our task is to monitor people attention, i.e. recognize whether they look at Nao, the other person, the paintings, or elsewhere.

paintings. In this case we define $\mathbb{F}$ as:

$$F = \{Nao, partner, pai_1, pai_2, pai_3, other\} \tag{4.1}$$

where Nao refers to the robot, pai_j refers to painting number j and $other$ stands for VFOA that is not attributed to any other label.

The recognition approach is illustrated in Figure 4.2. Broadly speaking, the middle part (box) shows the main recognition process, which consists of an HMM allowing the decoding of the sequence of head poses in terms of VFOA states $F_t \in \mathbb{F}$. The head pose $H_t \in \mathbb{R}^2$ at a given time can be represented by three angles pan, tilt and roll. The pan rotation is a left-right rotation, the tilt rotation is an up-down rotation, and the roll rotation is a head-on-shoulder roll as illustrated in Figure 4.3 (a). In this work we only use the pan and tilt angles, thus $H_t = (H_t^{pan}, H_t^{tilt}) \in \mathbb{R}^2$. Then this process is affected in two ways. First, by the gaze-head mapping model shown at the bottom part, whose goal is to dynamically predict at each instant t the expected head pose $\mu_t^h = (\mu_t^{h,pan}, \mu_t^{h,tilt})$ used to look at each VFOA target, as addressed in Sections 4.3 and 4.4. It is designed to reflect the findings from studies on human gazing behavior related to the coordination of the body, head and eyes in gaze shift. Secondly, as shown at the top part of Fig. 4.2, by leveraging contextual information aiming to remove the ambiguities introduced by relying on noisy head poses measurements rather than gaze. Given our robotics setup, contextual cues are extracted from the robot's conversational acts as discussed in Section 4.5.

4.3 Baseline: HMM with Geometrical Mapping

We build our VFOA recognition model based on a Hidden Markov model (HMM) which is illustrated in the middle part of Figure 4.2, without exploiting the context at this stage. In

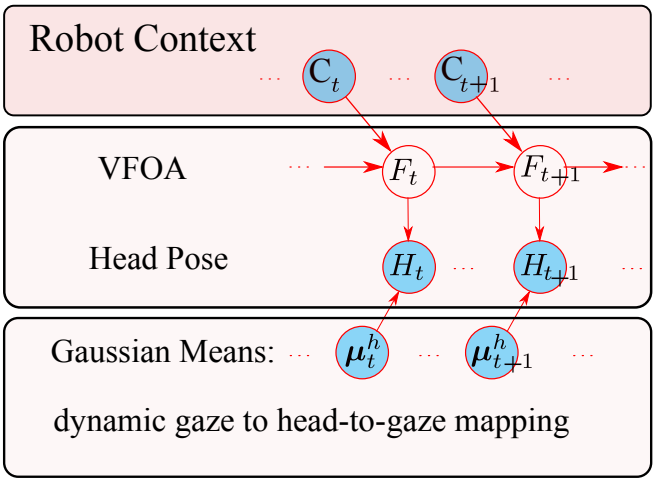


Figure 4.2 – VFOA recognition from head pose. The robot conversation context C_t appears as an input observation and provides expectations about which VFOA should be observed. At the bottom, a gaze-head mapping module dynamically monitors the expected head pose associated with each VFOA target.

this model, the distribution of head poses associated to a given VFOA target is represented by a Gaussian distribution, whereas transitions between VFOA targets are represented by a transition matrix A. More specifically, let H_t and F_t indicate the head pose and focus values at time t, and $\mu_t^h(f) \in \mathbb{R}^2$ and $\Sigma_H(f) \in \mathbb{R}^4$ denote the mean and covariance of the Gaussian associated with target f. The HMM equations can be written as:

$$P(H_t|F_t = f, \mu_t^h) = \mathcal{N}(H_t|\mu_t^h(f), \Sigma_H(f)) \tag{4.2}$$

$$P(F_t = f|F_{t-1} = \hat{f}) = A_{f\hat{f}} \tag{4.3}$$

Parameter setting is a major issue in this approach. Following previous works, the covariance of Gaussians can be set to reflect the target sizes and the head pose estimation variability. In absence of other information, the temporal prior $p(F_t|F_{t-1})$ modeled by the transition matrix A can be used to perform temporal smoothing by setting large probabilities to stay in the same state and equal low probability to transit to other states:

$$A_{f\hat{f}} = \begin{cases} a, & \text{if } f = \hat{f}, \\ (1-a)/(n-1), & \text{otherwise.} \end{cases} \tag{4.4}$$

where a is a constant denoting the probability of staying in the same state, and n is the number of targets. This satisfies our expectation of preserving the VFOA continuity in the sequence.

However, although they play the most important role in the model, setting the Gaussians

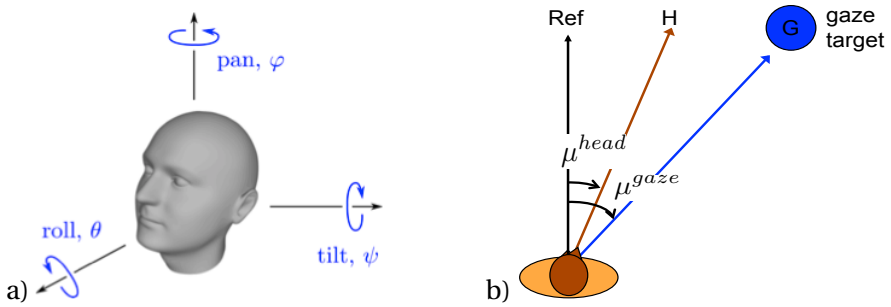


Figure 4.3 – a) Head pose specified by pan, tilt and roll angles. b) Geometrical Gaze Model (for the pan angle). The person is assumed to be looking at the reference direction, or midline (body orientation). Then, looking at a gaze target is accomplished by rotating both the eyes and head, with the head part being a fixed fraction of the full gaze rotation. In the picture μ^{gaze} corresponds to μ_t in equation 4.5.

means μ_t^h is not possible in an easy way. As discussed in the introduction using training data is not an option since annotation needs to be gathered for each configuration of the observer, targets and settings. This is especially problematic if people are free to move.

A solution to overcome the above difficulty is to use **gaze models** derived from the findings about human's gazing behavior as explained in Section 2.2. According to the model explained in 2.2.1, gazing at a target is accomplished by rotating both the eyes ('eye-in-head' rotation) and the head as illustrated in Figure 4.3 (b). More precisely, we assume that the fraction of head rotation as compared to the total gaze rotation is a constant, independent of the amplitude of the gaze rotation and of the current context. Then, as a first approximation, $\mu_t^{hb}(f)$, the mean of the Gaussian[1] to look at target f can be set as a fixed linear combination of the gaze and *head reference* directions:

$$\begin{aligned} &\mu_t^{hb}(f) - R_0 = \alpha \star (\mu_t(f) - R_0) \\ \Rightarrow &\mu_t^{hb}(f) = \alpha \star \mu_t(f) + (1_2 - \alpha) \star R_0 \end{aligned} \tag{4.5}$$

where $\star$ denotes the component wise product, $1_2 = (1,1)$, $\alpha = (\alpha^{pan}, \alpha^{tilt})$, $R_0 \in \mathbb{R}^2$ denotes the reference direction assumed to be constant (independent of time), and $\mu_t \in (\mathbb{R}^2)^K$ denote the gaze angles specifying the directions of the given K targets which are assumed to be known. Note that these gaze angles μ_t, which are the pan and tilt angles of the unit vector going from the person's head to each of the K targets can vary over time as a consequence of the person movement, or of the targets movements. The head-to-gaze ratio for the pan, α^{pan}, is usually set between 0.5 and 0.7, and between 0.3 and 0.5 for the tilt. Equation 4.5 can be used to set the head pose mean for looking at the target f in our HMM model. Our baseline thus consists of the above HMM model with the reference R_0 set to a constant value in order to set the means. Note that this model has been used in the context of meeting data in [Ba and Odobez, 2009].

[1] μ^{hx} indicates the way the mean μ^h is set using algorithm x.

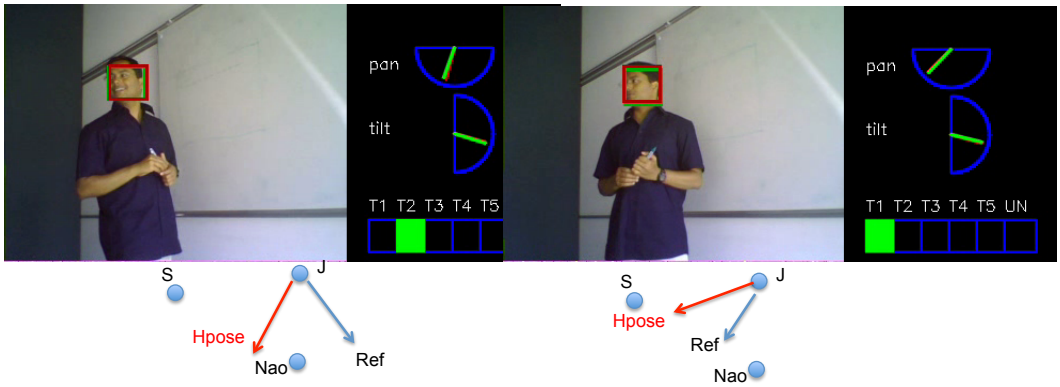


Figure 4.4 – Different reference directions (shoulder orientations) lead to different poses for looking at the same target. In both images, person J looks at person S. These images illustrate that the geometric model is holding true: the head orientation is approximately half-way between the reference direction and the gaze direction. Note that for the image on the left, using looking at Nao as reference direction R in Equation 4.5 would most probably lead to a wrong interpretation of the head pose as looking at Nao rather than at the person S.

4.4 Gaze to head dynamical mapping

The introduced baseline geometrical model has been shown to be useful in static scenarios. In meetings, since people are mostly seated and do not move their bodies extensively setting the reference on the middle of the targets has been a good solution [Ba and Odobez, 2009]. When the participants upper body and shoulders exhibit more dynamics, the baseline model becomes inaccurate since having a static body reference becomes an unrealistic assumption. Furthermore, the geometrical head-gaze mapping model was originally designed for discrete gaze shifts when the person moves his head from the midline and intentionally for looking at a given direction [Langton et al., 2000]. Therefore, it might not be sufficient for mapping the head poses to the gaze directions when the user is continuously moving his head for looking at different targets. In this context, using the evolution of head poses or gaze directions in the past could be useful for obtaining dynamic and more precise predictions of the head poses used to look at a given target at the current instant. In the following subsections we explore and introduce three models that can potentially address the mentioned shortcomings of the baseline geometric model.

4.4.1 Model G1: Dynamical Head Reference

Setting the Gaussians means using the geometrical model requires the knowledge of R_0 and of the target directions. Equation 4.5 shows the importance of the reference: using a wrong value for R_0 shifts the mean values $\mu_t^h(f)$ for all targets f simultaneously, which can have dramatic effects for head pose interpretations. The importance of knowing the head reference (shoulder orientation) for interpreting and associating a head pose with the corresponding VFOA target is also illustrated in Figure 4.4. Thus, unless the reference direction is constrained

by the setting (e.g. when people are seated), using a constant reference can be problematic. More general and natural interactions will result in more variations and shifts in the reference as people are free to move. This motivates the need for setting the reference dynamically.

Our goal is to derive a gaze model that accounts for a dynamically estimated reference. To address this point in absence of shoulder orientation information, we rely on the following intuition. A person tends to orient himself towards the set of gaze targets he/she spends time looking at. Such a body position makes it more comfortable and less energy consuming to rotate his head towards different gaze targets. As a corollary, this means that his average head pose over time is a good indicator of his reference direction, and can be used as an estimation for it. Therefore we propose to set the reference value at frame t denoted by R_t as the head pose average computed over the temporal window of duration W^R preceding the instant t:

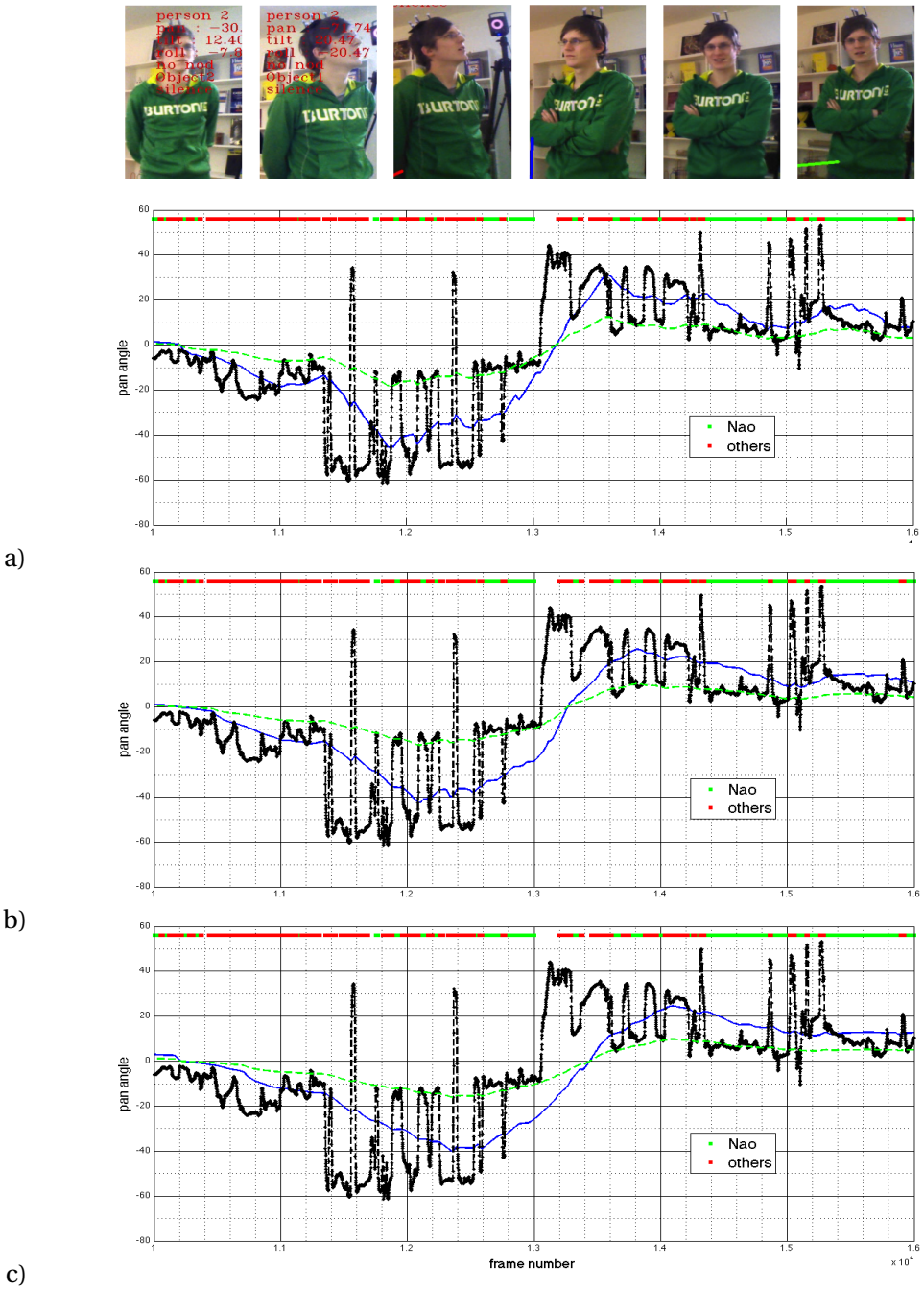
$$R_t = \frac{1}{W^R} \sum_{i=t-W^R}^{t} H_i \tag{4.6}$$

Figure 4.5 shows the evolution of the pan angle of this reference on a segment taken from the Vernissage dataset. The three plots (a), (b) and (c) show the reference pan angle for different window sizes (20, 30 or 40 seconds). Images shown on top correspond to sample frames taken from the same sequence. It can be seen that using this implicit reference is a suitable approximation: in the first three images, the person's body is oriented to the right, which corresponds to a negative body orientation well reflected in the blue curves; while in the last three images, the body is oriented to the right (positive pan angle). Note as well that the body is more frontal in the first and two last images, a fact that can also be observed in our body reference approximation (see the blue curve values at frames 10500, 1450 and 15500 compared to 11500, 12500 and 13500). Choosing different window sizes W^R produces different reference sequences as shown in the three rows. With bigger window sizes, the reference evolves more smoothly and is less affected by relatively long side head pose of the person (that should not be taken as the body orientation). However, bigger window result in latency in cases of important body shifts (compare the blue curves on frame segment 13000 to 13400 in the third and first plots). The choice of a suitable window size thus becomes important.

This reference value can then be substituted for the static reference in the baseline model of Equation 4.5, leading to the definition:

$$\mu_t^{hg1}(f) = \alpha \star \mu_t(f) + (1_2 - \alpha) \star R_t \tag{4.7}$$

We will denote this gaze model by G1. The improvement in recognition accuracy provided by this model over the baseline can be observed in Figure 4.5 by following the green curves which show the expected head pose for looking at Nao using the above model (and $\alpha^{pan} = 0.6$). Consider the VFOA ground truth bar shown on the top part, and segments [12800, 13000], [14400, 14800] or [15300, 15600] and compare the two models. Using the baseline with the static reference set as the robot's direction $(0,0)$, the expected pan angle for looking at Nao



a)

b)

c)

Figure 4.5 – Dynamical reference (pan angle) estimated from the head pose averages. The first row shows sample images corresponding to frame numbers 10500, 11500, 12500, 13500, 14500 and 15500 of the sequence. On each of the three plots (a), (b), and (c) the following elements are displayed: the head pose pan angle of the person as given by the Vicon data (black curve); the reference direction (blue curve) computed from Eq. 4.6 as the average head pose over a window of 20 (plot a), 30 (plot b) and 40 seconds (plot c); and the expected pan angle for looking at Nao (green curve) predicted according to Eq. 4.7. The green and red bar on top of each of the plots shows VFOA the ground truth (Nao or other). The estimated reference varies along with the head pose and body variations. Different window sizes change the smoothness and delay factors of the reference.

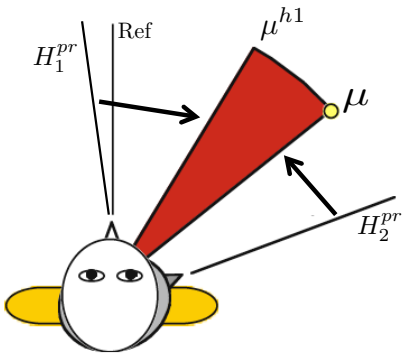


Figure 4.6 – Gaze Model with Midline Effect [Hanes and McCollum, 2006]. The target direction for the shift is denoted by μ. When the gaze is moved to μ from the initial head pose H_1^{pr}, the head is rotated to μ^{h1} according to the geometrical model. The head position at the end of the shift is thus independent of the initial head position. However, when the gaze shift is centripetal from H_2^{pr} to μ, the head is moved to μ. For initial head positions between μ^{h1} and μ (red zone), an eye-only saccade to μ is made (the head position remains the same).

would be 0 which is not very close to the head pose pan values on these segments and makes it difficult for the baseline model to recognize the VFOA correctly as looking at Nao. However, it is evident that using model G1 better expected pan angle (the green curve) is estimated for looking at Nao (instead of 0), as they are closer to the observed head pose values and thus will help in correctly recognizing Nao as the visual target.

4.4.2 Model G2: Midline Effect

The previous model was derived based on the assumption of a gaze shift from the reference to the gaze direction as explained in Section 2.2.2. Thus, the gaze model defined by Equation 4.5 is not valid for gaze shifts with different initial gaze directions.

Indeed, the study on gaze shift [Hanes and McCollum, 2006] behaviors summarized in Section 2.2.2 shows that how much of a gaze shift is accomplished by the head or by the eye depends significantly on the position of the head at the start of the gaze (which in general is not aligned with the reference), and whether the shift goes through the reference or not[2]. From the analysis of the gaze behavior literature, these authors derived the gaze model illustrated in Figure 4.6 that we investigated in this thesis.

Note that two main conditions happen considering the relation of the previous head pose and the reference. In the first one, gaze shifts towards the side, then the final head pose is independent of the initial head pose value since what is important is the final head pose, not how much the eye or head rotated to reach there. This is important, as it validates the model

[2] In [Hanes and McCollum, 2006], the reference is called midline. Note that as the model was only studied for the pan variable, in the G2 model (and G3 as well), the tilt gaussian means were set using the G1 model.

G1 as a way to define the expected head pose to look at a target. In the second case, where gaze is coming back from the side and towards the reference, the initial head pose becomes important. In this condition, head does not go farther than the target direction to be aligned with the expected head pose predicted by model G1.

To implement this midline effect we need to know what was the value of the head pose before the gaze shift occurs. To this end, we introduced the variable $H_t^{pr,pan}$ defining the head pose pan angle prior to a shift and used as estimate of this variable the average of the head poses (pan angles) computed over a window of size W^p separated by a gap Δ^p from the current instant:

$$H_t^{pr,pan} = \frac{1}{W^p} \sum_{i=t-W^p-\Delta^p}^{t-\Delta^p} H_i^{pan} \tag{4.8}$$

The G2 gaze model was then implemented by setting the head pose mean $\mu_t^{hg2,pan}(f)$ of the head pose pan angle[2] of target f using the following rules. For $\mu^{pan}(f) > 0$:

$$\mu_t^{hg2,pan}(f) = \begin{cases} \mu_t^{hg1,pan}(f), & \text{if } H_t^{pr} < \mu_t^{hg1,pan}(f)\,, \\ min\left(\mu_t^{pan}(f), \mu_t^{hg1,pan}(f) + \alpha_H(H_t^{pr,pan} - \mu_t^{hg1,pan}(f))\right) & \text{otherwise.} \end{cases} \tag{4.9}$$

and for $\mu^{pan}(f) \leq 0$:

$$\mu_t^{hg2,pan}(f) = \begin{cases} \mu_t^{hg1,pan}(f), & \text{if } H_t^{pr} > \mu_t^{hg1,pan}(f)\,, \\ max\left(\mu_t^{pan}(f), \mu_t^{hg1,pan}(f) + \alpha_H(H_t^{pr,pan} - \mu_t^{hg1,pan}(f))\right) & \text{otherwise.} \end{cases} \tag{4.10}$$

Figure 4.7a) shows the resulting probabilistic graphical model G2. The factor α_H indicates how much we take into account the previous head pose in the estimate. When $\alpha_H = 0$, we always have $\mu_t^{hg2,pan} = \mu_t^{hg1,pan}$, which means that the head pose means are set using the standard geometric model (but using a dynamically set reference). When $\alpha_H = 1$, the implemented model is exactly the axiomatic model described in Chapter 2.2.2. Note that in this case when $\mu^{hg1,pan}(f) < H^{pr,pan} < \mu^{pan}(f)$, or similarly $\mu^{hg1,pan}(f) > H^{pr,pan} > \mu^{pan}(f)$, the expected head pose is given by the previous head pose. Since the estimation for the previous head pose is not necessarily very accurate, setting the head pose mean in this way could be overconfident. Therefore, having $0 \leq \alpha_H \leq 1$ might be more appropriate.

4.4.3 Model G3: implementing gaze shifts

When implementing the midline effect, the previous model has one drawback: at each time step, a gaze shift is assumed. In other words, even if the person is focusing on target f, the previous head pose $H_t^{pr,pan}$, estimated through recursion over a short temporal window,

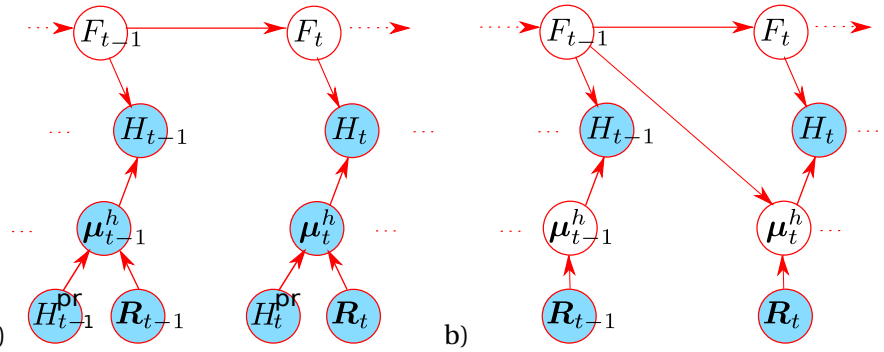


Figure 4.7 – Probabilistic graphical models. (a) Model G2. The head reference direction R_t and the mean head pose of the Gaussians μ^h_t are time dependent variables, and the recent head pose H^{pr}_t can be exploited. (b) Model G3. The mean head pose for looking at a target (μ^h_t) depends on the gaze target at the previous time step (F_{t-1}). Shaded nodes indicate that the corresponding random variables are set directly from observation, whereas unshaded nodes denote hidden variables that need to be inferred.

evolves and as a consequence it may introduce an evolution of what the head pose for looking at target f should be, especially when $H^{pr,pan}_t$ is close to the expected head pose.

As alternative to the model G2, we define the gaze situation prior to the visual attention shift by the actual gaze direction defined by the (discrete) VFOA at the previous instant. We then propose to define the mean of the head pan angle[2] to look at target f at time t, given the previous focus $F_{t-1} = \hat{f}$, by:

$$\mu^{hg3,pan}_t(f) = \alpha_1 \mu^{pan}_t(f) + \alpha_2 \mu^{pan}_t(\hat{f}) + (1 - \alpha_1 - \alpha_2) R^{pan}_t \tag{4.11}$$

Thus, in absence of gaze shift ($F_{t-1} = F_t = f$), the head pose mean is simply set using the geometrical model with $\alpha^{pan} = \alpha_1 + \alpha_2$ and therefore the problematic pose evolution during fixation described above does not exist. In case of a gaze shift ($F_{t-1} \neq f$) the head pose pan angle is not only affected by the reference and new gaze direction $\mu^{pan}_t(f)$ as in G1, but also by the direction towards the VFOA target at previous instant (the head will be closer to direction of the previous VFOA target than what would be predicted by the model G1).

Figure 4.7b) shows the new graphical model G3. Note that here the effect of the previous head pose is not considered by exploiting the midline effect as in G2. The link between the hidden states F_{t-1} and μ^h_t renders the inference more complex than in a standard HMM. In practice, we conducted the inference sequentially, using the estimated focus at time $t-1$ to estimate the optimal focus at time t.

4.4.4 Model inference

Considering the IOHMM structure of our models, at each given instant t given the model parameters and the sequence of observations, we would like to find the distribution over the

last hidden variable at the end of the sequence, i.e. to compute $p(F_t|H_{1:t})$. This problem can be handled efficiently using the forward algorithm to perform the calculations recursively as follows:

$$p(F_t, H_{1:t}) = \sum_{F_{t-1}} p(F_t, F_{t-1}, H_{1:t}) \tag{4.12}$$

Using the chain rule:

$$p(F_t, H_{1:t}) = \sum_{F_{t-1}} p(H_t|F_t, F_{t-1}, H_{1:t-1}) p(F_t|F_{t-1}, H_{1:t-1}) p(F_{t-1}, H_{1:t-1}) \tag{4.13}$$

H_t is conditionally independent of everything but F_t, and F_t is conditionally independent of everything but F_{t-1}, thus this simplifies to

$$p(F_t, H_{1:t}) = p(H_t|F_t) \sum_{F_{t-1}} p(F_t|F_{t-1}) p(F_{t-1}, H_{1:t-1}) \tag{4.14}$$

Since $p(H_t|F_t)$ and $p(F_t|F_{t-1})$ are given by the model's emission distributions and transition probabilities, one can recursively and quickly calculate $p(F_t, H_{1:t})$ from $p(F_{t-1}, H_{1:t-1})$. Ultimately, since $p(F_t|H_{1:t}) = p(F_t, H_{1:t})/P(H_{1:t})$, and $P(H_{1:t})$ is independent of the state values, we can obtain our target posterior distribution, and use as recognized VFOA the target F maximizing this posterior.

Note that for G1 and G2 graphical models, at each time instant t the values of R_t, H^{pr} and μ^h (either μ^{hg1} or μ^{hg2}) can be calculated through the equations. Thus, we also used the standard HMM filtering to infer the VFOA for these models.

As shown in Figure 4.7, in model G3 there is a link between the hidden states F_{t-1} and μ^h_t which renders the inference more complex than in a standard HMM. In practice, however, we used the following optimization scheme. We determined at time $t-1$ the state $\hat{F}_{t-1}$ maximizing the posterior $p(F_{t-1}|H_{1:t-1})$, and used this state value to compute the expected means. Given these means, the posterior $p(F_t|H_{1:t})$ was computed recursively using the method described above.

4.5 Context Modeling

In Section 4.4, we proposed three models for dynamic mapping of gaze to head pose which helps in decoding the VFOA states from the head pose sequence. As illustrated in Figure 4.2, the second way for improving the VFOA recognition is by using contextual information to remove the ambiguities introduced by relying only head pose information for estimating the gaze.

In this Section, we aim to leverage context cues to improve VFOA recognition from head pose. Contextual information could potentially help in removing some of the ambiguities due to

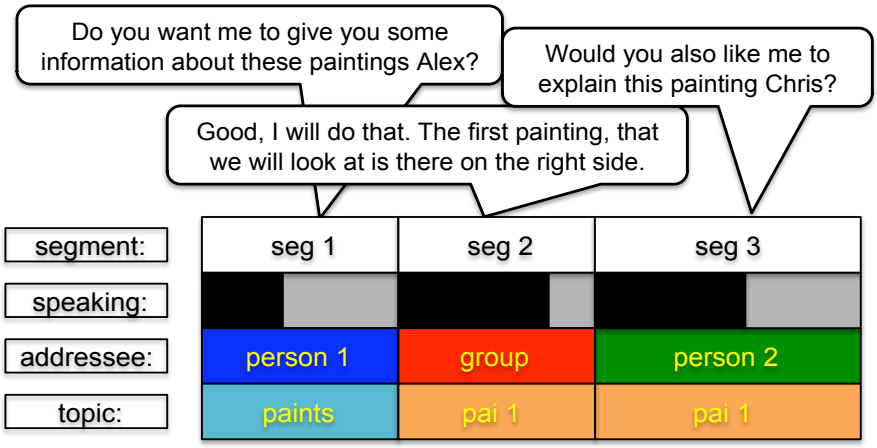


Figure 4.8 – Illustration of the context assignments. Each segment corresponds to one of the robot's speech turns and the pause after it (during this robot speaking pause, participants may answer a robot's question or talk together, etc.) and thus composed of two subsegments with different speaking status ($s = 1$ and $s = 0$). Depending on the robot's speech, addressee and topic states are assigned to each of these segments.

the limitations of our head pose based VFOA recognition models and to compensate for noisy estimations of the head poses. The main idea is that when interacting with a robot, its actions influence what people do in certain situations. Therefore, this information, which the robot is aware of, can be used to predict and better interpret people behavior.

In the following subsections, prior to describing more precisely the recognition model, we will first introduce the features that we have exploited as context for VFOA recognition in 4.5.1 and discuss how these cues could be extracted from the robot's system in 4.5.2. In 4.5.3 we will describe our conversational aware VFOA recognition method and in 4.5.4 explain how context tables are trained for this method.

4.5.1 Robot Conversation Context

Given our task, the question is which of the robot actions affect people VFOA, and how? In interactions, these mainly relate to the communication functions of gaze and their relationships with speaking turns [Kendon, 1967]. However, it is also known that objects that play a central role in the conversation may attract the attention whereby overruling the communication patterns observed in natural conversation [Van Turnhout et al., 2005]. In our art guide scenario this corresponds to physical locations in the room and particularly paintings. We thus defined the robot interaction context, illustrated in Figure 4.8, as described below.

Speaking context.

Listeners are known to gaze more at speakers than at non-speakers to show attention towards

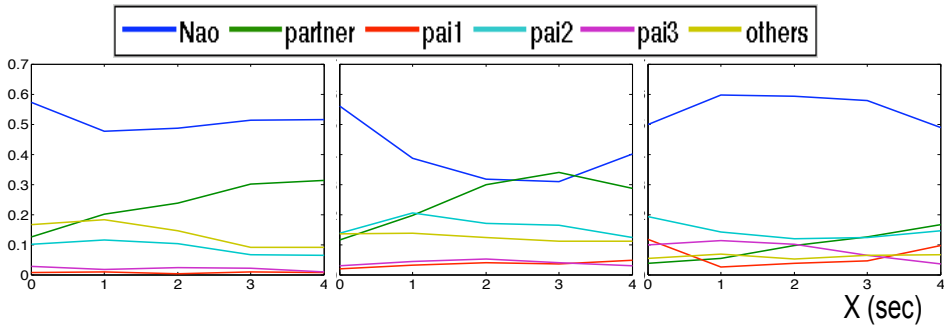


Figure 4.9 – Participants VFOA statistics given the robot's different addressee states. (left) shows the VFOA statistics for the participant who is individually addressed by the robot, (middle) shows the statistics for the non-addressed participant, and (right) shows the statistics when both participants are addressed. The x axis denotes the time since the end of the robot utterance. The statistics for $x = 0$ are collected during the robot's utterance. Different curves correspond to different visual targets.

them. Thus we defined a speaking context state variable $s_t \in SC = \{0, 1\}$ as whether Nao speaks or not at time t.

It could be important to consider how long Nao has been speaking or if it is possibly close to finish its speech turn. However, modeling all of these factors is a complicated task. In order to keep the model simple and avoid overfitting, at this step we only considered a binary variable for the speaking context which has the same effect regardless of the distance to the beginning and end of the speaking segments.

Addressee context.

It is known that speakers monitor their addressees' attention by gazing at them, and expect gaze in return [Kendon, 1967]. Considering this effect, we thus defined the addressee context $a \in AC = \{pers_1, pers_2, group\}$ of a speech segment as the situation when the robot addresses the first person, the second person, or both.

As with the speaking status, one may wonder whether the addressee context may impact the VFOA of people differently depending on the timing (during the utterance, after the utterance). To study this effect the VFOA statistics depending on the addressee status are shown in Figure 4.9, either during the robot speech (displayed for $x = 0$) or x seconds after the end of the speech.

Comparing the three plots in Figure 4.9 for the individually addressed person (left plot), the non-addressed person (middle plot) when the robot addresses a single person, and the case where both participants are addressed, we can see some general differences. In spite of the noise, we can notice that addressed people tend to stay more in visual contact with the robot, while non-addressed people disengage quicker to look at the other person or elsewhere.

Moreover, when both people are addressed they tend to stay in visual contact with the robot even longer and less with each other. These differences motivates using this type of context as a cue to improve VFOA recognition.

Considering the temporal variation of VFOA probabilities (after the utterance), we can notice that specifically for the non-addressed person there are some differences during different stages as in Figure 4.9(middle). During the robot's utterance and right after the utterance, such a person seems in visual contact with the robot, then looks more at the partner and at the end engages again with the robot while in general the addressed person might answer to the robot or releases his turn.

However, in spite of these observations, to avoid overfitting and keep the models simple, we implemented a constant model for $x > 0$ and found it to be also reasonable. We thus defined the addressee context state a_t at t as the addressee context derived from the current (if $s_t = 1$) or preceding (if $s_t = 0$) robot utterance.

Topic context.

Given our Vernissage scenario and dataset, the topic context is considered to show whether the robot informs or refers to a specific painting, to two or all paintings, or none of them. The topic context set is then defined as $OC = \{pai_1, pai_2, pai_3, paintings, none\}$ corresponding to the mentioned states. The topic context state $o_t \in OC$ at time t is thus defined as the topic context of the robot utterance that precedes t.

We could expect that the exact moment when the robot explicitly points to a picture or mentions its name affects the participants behavior only right after this specific moment, and that timing plays some role on this effect. However, extracting the exact timing information of references to the objects from the robot system and modeling their effect in a very accurate level would bring many complications into the approach. Moreover, in some cases (i.e. explanation part in the Vernissage scenario), people may look at the paintings during the whole utterance and explanations, whereas it might be more punctual in other cases (i.e. when robots refers to a painting to remind it during the quiz). Distinguishing between these cases might also be difficult for an automatic system. In order to avoid these complications, we consider that the topic context will have the same effect on VFOA all over the segment and do not assume any temporal variations.

Overall conversational context C_t. As a summary, at each instant t, the different context states s_t, a_t and o_t are automatically assigned according to the spoken utterances and temporal segments, as illustrated in Figure 4.8. The final context state C_t is then defined as the Cartesian product of all contexts, ie $C_t = (s_t, a_t, o_t)$, with $C_t \in OC = SC \times AC \times OC$ and will influence the VFOA recognition as explained in the next Section.

4.5.2 Conversation Context Derivation from the Robot System

In this work, we assumed that the robot is aware of all conversational context types defined in Section 4.5.1. In this part we explain how this information was derived in our scenario and how it could be accessed in other platforms.

In our scenario, the context is automatically derived from the robot system data. The robot system data includes wizard commands and internal events for speech and gesture production. This data is both available for online use and is also recorded along with our Vernissage corpus. Speaking status is automatically derived from the internal events. Wizard commands sent to the dialog system implicitly contain the information needed to derive the addressee and topic states. Therefore, the dialog system is aware of who is addressed (either a person, or a group) along with the way to address them, which in our set-up was accomplished for a given individual by naming him and turning the head towards him, or by directing the head in between participants when both persons were addressed. In the same way, the dialog system is aware of the current topic the robot has spoken about and also the way to show it to the participants, accomplished by mentioning the name of the painting or its painter, turning the head towards it or using hand gestures to point at it.

Alternatively, instead of being dependent on a wizard, in a more realistic case the robot would rely on an autonomous dialog system. In this case the dialog manager would decide which speech and dialog acts the robot should make during its interactions. Considering the information from users behavior and requests they make during the interaction, the dialog manager would decide what the robot should say (contains the information about the topic) and who it should address. Therefore, in the same kind of scenario with multiple users and objects of interest which could be the topic of the conversation, The VFOA recognition module would be able to receive similar information from the dialog manager.

4.5.3 Conversation Aware VFOA Recognition

To address VFOA recognition using head pose and context information, we use the IOHMM graphical model of Fig. 4.2. In this model, the VFOA is inferred by maximizing the posterior probability of the sequence of VFOA states $F_{1:t}$ given all observed variables: head pose $H_t \in \mathbb{R}^2$ and context C_t. The posterior for the graphical model of Fig. 4.2 is expressed as:

$$p(F_{1:t}|H_{1:t}, C_{1:t}, \mu^h_{1:t}, R_{1:t}) \propto \prod_{t=1:t} p(H_t|F_t, \mu^h_t)\, p(F_t|F_{t-1}, C_t)$$

with

$$p(H_t|F_t = f, \mu^h_t) = \mathcal{N}(H_t|\mu^h_t(f), \Sigma_H(f)), \tag{4.15}$$

$$p(F_t|F_{t-1}, C_t) \propto p(F_t|F_{t-1})\, p(F_t|C_t) \tag{4.16}$$

where the different terms are explained below.

Data likelihood.

The term in Equation 4.15 represents the likelihood of an observed head pose for a given focus, and is modeled as in Section 4.3, with a Gaussian distribution per focus. Note however that here we will rely on dynamic means μ_t^h set according to the different models in Section 4.4, and which play a crucial role for VFOA recognition.

Contextual prior.

Equation 4.16 denotes the prior on the focus, which we assumed can be decomposed in two parts. The first one is the temporal prior $p(F_t|F_{t-1})$ modeled by a transition matrix A set as in Section 4.3 to allow temporal smoothing. The second one, which denotes the prior on the VFOA according to the Robot context is modelled using a multinomial distribution parametrized by the vector $B_c = (B_{ci})_{i \in F}$.More precisely, we have:

$$p(F_t = f|C_t = c) = B_{cf}$$

This term affects the VFOA recognition by altering the expectations about what people might be looking at depending on the context. It is parameterized by the probability tables $B = \{B_c\}$ as explained below.

4.5.4 Learning the context tables

There are several ways to set the tables, depending on goals and assumptions. Here, we use a learning approach, with smoothing to handle the lack of data for some contexts, and further modeling assumptions to avoid data overfitting and better capture the model generalization capabilities.

More precisely, given a training dataset, we gather the VFOA data $\mathscr{D} = \{D_c, c \in OC\}$ where $D_c = \{f_t|c_t = c\}$ contains VFOA data observed under each given context c. Then the goal will be to learn the vector of parameters B_c of the multinomial for each context c. For learning the parameters we use a Maximum A Posteriori approach to maximize

$$p(B_{c\cdot}|D_c) \propto p(D_c|B_{c\cdot})Dir(B_{c\cdot}|\alpha) \tag{4.17}$$

where $Dir(B_{c\cdot}|\alpha)$ denotes a conjugate Dirichlet prior on the parameter B_c. According to this model, the optimal parameters are given by:

$$B_{cf} = \frac{n_f + \alpha_f}{\Sigma_{f\prime}(n_{f\prime} + \alpha_{f\prime})} \tag{4.18}$$

where n_f denotes the number of occurrences of the focus f in D_c, and α_f denotes the Dirichlet

Table 4.1 – Sample context VFOA count table(using only the topic context)

Context	Nao	partner	pai_1	pai_2	pai_3	others
pai_1	7847	692	10851	3704	101	771
pai_2	4860	247	253	12629	72	519
pai_3	8404	798	57	985	10414	618
paints	13704	6113	4447	12411	3312	2489
none	103186	31045	1938	15233	4264	22086

prior parameters for each focus f. These priors were set as:

$$\alpha_f = 0.1 N_f / (K \times N_C) \tag{4.19}$$

where N_f, K and N_C denote the number of observation in the whole training set, the number of VFOA targets, and the number of contexts, respectively. In other words, the prior corresponded to the addition of virtual observations equally spread amongst table entries and amounting to 10% of the total number of real observations.

Priors learned using the above scheme might overfit the specific setup. In particular, the painting positions or the duration of references and explanations about each of them lead to the gathering of different statistics for each painting. Therefore, to be more general, we applied parameter tying, enforcing that all table entries involving paintings which play the same role should be the same. In order to do that, after counting the observations (video frames) with each specific context state $c \in OC$ and VFOA target $F \in \{Nao, partner, pai_1, pai_2, pai_3, other\}$ in the training data to obtain the table of the raw counts, we make further simplifications as follows:

- For the context types c where the topic cue is not pai_1, pai_2 or pai_3, we take a same number of occurrences for looking at the paintings by taking the average of their occurrences.
- When the topic cue of c is pai_1, pai_2 or pai_3, hence the robot is referring to a specific paintings, we differentiate between the painting which is being referred and the two other paintings. Averaging is done separately for the referred painting and the two other ones. Therefore, we will obtain different occurrence numbers for looking at the referred painting versus the others.

Finally the probabilities are obtained by normalizing the rows of this table. As an example, when we consider a context set only consisting of the topic contexts ($C = TC$) Table 4.1 illustrates the raw observation occurrences, and the final context probability priors obtained after parameter tying, normalization and smoothing are shown in Table 4.2.

Table 4.2 – Sample context probability priors (using only the topic context) showing parameter tyings.

Topic Context	Nao	partner	pai_1	pai_2	pai_3	others
pai_1	0.33	0.03	0.53	0.04	0.04	0.03
pai_2	0.33	0.03	0.04	0.53	0.04	0.03
pai_3	0.33	0.03	0.04	0.04	0.53	0.03
paints	0.32	0.14	0.16	0.16	0.16	0.06
none	0.58	0.17	0.04	0.04	0.04	0.12

4.6 Conclusion

In this chapter we proposed two different but complimentary approaches towards recognizing the visual focus of attention based on head pose input for human robot interaction application.

The first approach focuses on the fact that in unconstrained conditions, dynamic gaze-to-head mappings with more accurate gaze shift models are needed. Therefore, we proposed several models for improving the head to gaze pose correspondence. In the first model, we incorporated an implicit estimate of the body reference into the model which varies over time. In the second one, we implemented the midline effect which considers the importance of the previous head pose, when the gaze shift occurs from the side towards the head reference. Finally in our third model, we consider the effect of the previous gaze only when a gaze shift happens by incorporating the previous gaze target.

In our second approach, we aimed at leveraging contextual information which is specially obtainable from the robot's conversational states. To this end we introduced different contextual cues from the robot's conversation which we expect to have effect on the user's behavior and a methodology on how to exploit them in the VFOA recognition process.

Our experimental studies in the following will demonstrate the utility of these approaches for improving VFOA recognition as well as their complimentary benefits.

5 VFOA recognition experiments

5.1 Introduction

In Chapter 4, we presented different models for VFOA recognition approach which uses head pose-gaze mapping and leverages robot contextual context for inferring the VFOA labels from the head pose sequence. In this chapter, we will evaluate and validate different components and contributions, in particular different head pose-gaze mappings and the effect of the robot's conversational context. The content of this chapter was partially presented in workshop and conference papers [Sheikhi et al., 2012, Sheikhi and Odobez, 2012] and is currently submitted as a journal paper [Sheikhi and Odobez, 2014].

Before looking at the results, we first summarize how head pose tracking estimates were obtained in Section 5.2. Then in Section 5.3 we will remind the parameters involved in different models, provide their default values and explain the general strategy to learn them. In addition we will explain our evaluation protocol.

In Sections 5.4 and 5.5 we will remind the datasets and provide the experiment results. Since Meeting and NaoD data have no robot context, we will present their results separately in Section 5.4 and present those for the Vernissage dataset in Section 5.5. In Section 5.6 we will provide discussion and conclusion on the experiments.

5.2 Head pose tracking methodology

In this section we describe the head pose tracking method used for extracting the head poses for our experiments. We used on the head pose tracking algorithm explained in [Ba and Odobez, 2005] for extracting the head pose estimates for NaoD dataset. Then we relied on an extension of this approach [Khalidov and Odobez, 2013] which was developed for Humavips. Here we provide a summary on this monocular visual head pose tracking method.

The method has considered the difficulties existing in HRI applications such as having moving persons, moving robot and unconstraint environment such as different distances of people

and different lighting conditions. Therefore, a tracker should have different properties to be suitable for these applications. First of all it should be robust against lighting variations, appearance changes, human motion, robot motion and occlusion. Second, the initialization and destruction of the trackers should be accurate and timely, and finally it should guarantee real-time performance and be able to work with video stream sampled irregularly. In order to provide these properties, tracking by detection approach is adopted. Face detectors for different face orientations (frontal and profile) are employed and the method makes use of trained prior models of color and texture features for various head poses to track in variable lighting conditions and with appearance changes. Moreover, tracker management techniques are employed following [Duffner and Odobez, 2013] to help for creating and removing tracks and handle occlusions.

Problem Formulation: Assumt that the observation (image) at time t is denoted by o_t and the individual tracker state representing the head pose configuration is denoted by s_t. Following the Bayesian formulation of the tracking problem, the objective is to estimate the distribution $p(s_t|o_{1:t})$ where $o_{1:t}$ denotes the sequence of observations up to time t. This distribution could be written as:

$$p(s_t|o_{1:t}) \propto p(o_t|s_t) \int_{s_{t-1}} p(s_t|s_{t-1}) p(s_{t-1}|o_{1:t-1}) ds_{t-1} \tag{5.1}$$

Tracker state s is defined as $s = \{u, v, h, e, \alpha^{pan}, \alpha^{tilt}\}$ where (u, v) is head location on the 2D image plane, h and e are *scale* and *eccentricity* parameters that determine the 2D bounding box where measurements are made and α^{pan} and α^{tilt} are *pan* and *tilt* head rotation angles denoting the pose angles.

Since the above equation cannot be solved analytically, approximations should be used. More precisely, in this method a particle filter approach is used in which a set of weighted particles is exploited to approximate the optimal filtering density. In more detail, a sequential importance sampling strategy is adopted to sample the new particles according to an importance function $q(s_t|s_{t-1}, o_t)$ and update the weights using standard formulation. One specifity of the tracker is to define the proposal q as a mixture of the state dynamic model $p(s_t|s_{t-1})$ and image-based proposal distribution. In this way both dynamics and observed data are used in order to approximate the optimal proposal distribution. These two component are explained below:

- **Dynamic model:** Given the importance of dynamics used to explore the state space and on the other hand difficulties for predicting people's motion, it is defined as a mixture of two elements. The first one is a random search around the previous state which accounts for the situations where the person is not moving, a likely case when in our HRI context. Whereas the second one is an order one state-based velocity model which accounts for constant speed motion of the person obtained from previous state estimates.

- **Image-based proposal distributions:** The previous dynamic models are only based on state evolution (constant position or velocity) and random search. However, there are

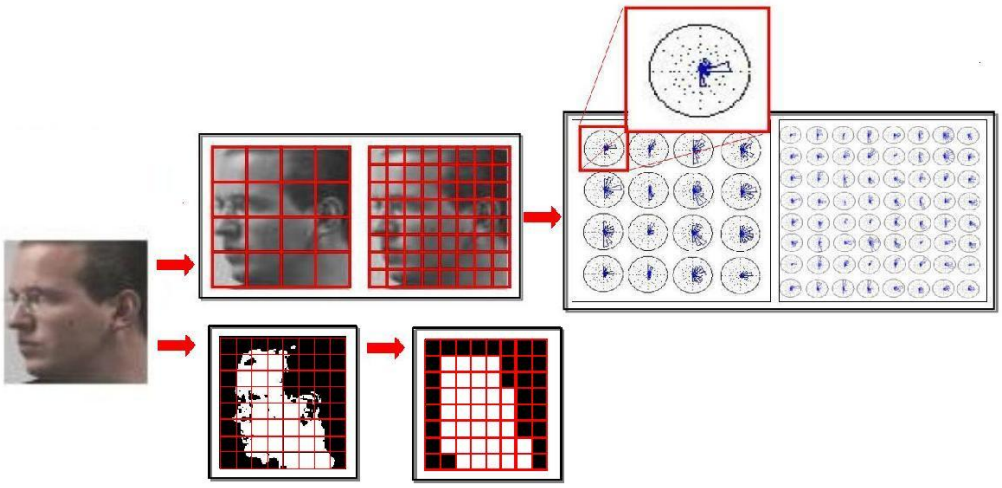


Figure 5.1 – Texture and color features.

also abrupt speed changes which are difficult to predict based only on past information, and these are the situations that often lead to failure. Therefore, in these cases it is better to directly exploit the information contained in the images, which are of two different natures: instantaneous observations reflecting the presence of the object (as produced by a face detector) and sequential observations reflecting observed image-based motion between frames. Thus, in this proposal, particle states are sampled either around face detections close to the state and obtained from a frontal and profile detection, or around states predicted using a robust motion estimator.

Likelihood distributions: a classification based approach is adopted to head pose estimation based on feature templates where the likelihood is defined as

$$p(o^f|s) \propto exp\{-\lambda d(o^f, o^f(\alpha^{pan}, \alpha^{tilt}))\} \tag{5.2}$$

with o^f denoting a set of features (f = texture or skin) and $o^f(\alpha^{pan}, \alpha^{tilt})$ denoting the template built from the POINTING head pose database containing images of 15 persons taken at different discretized pan and tilt angles [Gourier et al., 2004]. The following two kinds of features are used to characterize the tracker state and are illustrated in Figure 5.1.

- **Texture likelihood:** Texture features are computed using multiscale descriptors based on histograms of oriented diagrams (HOG) [Ricci and Odobez, 2009].
- **Skin likelihood:** Features based on skin color are also used to characterize image patches. Color models trained on frontal face images are used to clasify skin and non-skin pixels and then extract a skin binary mask.

Single and multiple person track management: Finally to achieve proper tracker manage-

Table 5.1 – Summary of the main parameters for different dynamical models introduced in Chapter 4. The optimal parameters were estimated mostly through cross-validation on the training set. .

	Models Parameters
Σ_H	Gaussian variance - set given the size of the targets and the expected noise.
A	HMM transision matrix - self-loop value a is set by cross-validation.
$unfoc$	probability threshold for determinging the unfocused state (other) - set by hand.
α^{pan}	gaze direction factor for μ^h pan angle - set by cross-validation
α^{tilt}	gaze direction factor for μ^h tilt angle - set by hand
W^R	window size for the dynamic reference - set by cross-validation
W^p	window size for the previous head pose for G2 - set by cross-validation
Δ^p	gap for the previous head pose for G2 - set by cross validation
α_H	previous head pose factor for μ^h in G2 - set by cross-validation
α_1	current gaze direction factor for μ^h pan angle in G3 - set by cross-validation
α_2	previous gaze direction factor for μ^h pan angle in G3 - set by cross-validation

ment and handle multiple trackers, the long term tracking framework of [Duffner and Odobez, 2013] is employed. The main ideas are a) to run the face detector and wait for several detection firings (unless it is in a region where previous tracks have been observed) to initialize tracks and make the initialization more robust to false detections, b) to run a filter on the tracker statistics (position, speed, likelihood, variance estimation) to quickly identify tracking failures and remove tracks, c) to manage the cases where persons occlude each other to avoid the situation where two different trackers end up tracking the same person.

5.3 General parameter setting and evaluation protocol

5.3.1 Model parameters

Different parameters are involved in the VFOA models described in Chapter 4. A summary of all parameters involved in the gaze to head dynamic mapping models described in section 4.4 is displayed in Table 5.1. Most of the parameters are set through cross-validation for the Meeting and Vernissage datasets. For the NaoD dataset, since it only consisted of three people, we used parameters obtained from the Meeting dataset. More details on setting the parameters will be provided separately for each of the datasets.

The parameter α^{tilt} is set to 0.5 by hand for all models and datasets and the value for the parameter 'unfoc', i.e. the threshold for deciding the other state is set to $\frac{1}{180\times180}$. With this threshold when the Gaussian's standard deviation for looking at a target is set to $(10,10)$ (we use roughly this standard deviation for most of the visual targets in the experiments with the Meeting and NaoD datasets), a point in the 1 standard deviation distance from the Gaussian

center would have a higher probability for belonging to the Gaussian than being unfocused whereas a point at the distance of 1.5 standard deviation would belong to the unfocused area and obtain the 'other' label.

For the context modelling, the context tables described in section 4.5.4 are also set through cross-validation.

5.3.2 Performance measure

Different measures used for comparing the algorithms and evaluate their performances. In particular we used Frame based Recognition Rate as our main evalusation measure. However, we also used other measurements to gain better insight to compare different models. Our measurements are defined as:

- Frame based Recognition Rate (FRR): the percentage of frames during which the VFOA has been correctly recognized. This is the main performance measure used in these experiments.
- Confusion matrix: the information about actual and predicted classifications.
- Head pose prediction error per VFOA class: the mean of the errors in degrees, between the head pose actually used to look at the target, and the prediction made by head pose-gaze correspondence models.
- Average recognition per VFOA class: the average percentage of correct recognitions for each target.

5.3.3 Statistical significance test

As in [Ba and Odobez, 2011], we used a variant of the McNemar test to evaluate whether the difference between the recognition results of two algorithms is statistically significant. The McNemar test looks only at the samples where the two algorithms give different results. It checks whether an algorithm provides almost systematically the same or a better answer than the other one. Following [Ba and Odobez, 2011], to ensure independence between VFOA samples, we extracted data chunks of 5 minutes separated by 1 minute intervals. On these chunks we performed a variant of the McNemar test that can account for correlated data in the clusters [Durkalski et al., 2003]. In this approach, it is assumed that there could be some correlation between data inside each cluster but different clusters are independent from each other.

In this approach the paired responces $(Y_{ijk}, Y_{\prime ijk})$ of the two algorithms to be compared are used as input, where Y_{ijk} is the binary outcome (correct or wrong result), i is the algorithm $(i = 1,2)$, j is the unit within the cluster $(1,2,\ldots,n_k)$, k is the cluster $(k = 1,2,\ldots,K)$. N denotes the total number of units $(\sum_{k=1}^{K} n_k)$ over all clusters, and K is the total number of clusters used in the study. For each cluster k, the data can be presented in a 2×2 contingency table for matched-pair data with the frequencies a_k, b_k, c_k and d_k of the concordant and discordant

	algorithm 1 true	algorithm 1 false	row total
algorithm 2 true	a_k	b_k	$a_k + b_k$
algorithm 2 false	c_k	d_k	$c_k + d_k$
column total	$a_k + c_k$	$b_k + d_k$	n_k

Table 5.2 – 2 × 2 contingency table for cluster k.

pair types inside the cluster as shown in Figure 5.2.

Having these frequencies, the proposed test statistic of [Durkalski et al., 2003] is

$$\chi^2_V = \frac{(\sum_{k=1}^{K} \frac{1}{n_k}(b_k - c_k))^2}{\sum_{k=1}^{K} [\frac{b_k - c_k}{n_k}]^2} \tag{5.3}$$

which is assumed to be asymptotically distributed as a chi-square with one degree of freedom for large number of clusters.

Given this test statistic χ^2_V and assuming a chi-square distribution for it, we can compute the p-value which is the probability of obtaining the test statistic result, assuming that the null hypothesis (in our case, that the two algorithms perform similarly) is true. If the p-value is less than the significance level 0.05, we conclude that the difference between the algorithms is statistically significant. Using this approach we computed the test results for the Vernissage dataset.

5.4 VFOA recognition results on Meeting and NaoD datasets

In this Section we present the results for the Meeting and NaoD datasets, since they both do not contain robot contextual data and experiments are limited to different head pose-gaze correspondence models.

5.4.1 Meeting Dataset

The Meeting dataset as described in 3.4 and shown in Figure 5.2 was the first data we used for the experiments with VFOA models. We used all 8 sequences and performed our study on the two persons on the seats in front of the camera, for a total of 16 persons. Although we aim to provide algorithms applicable to usual human robot interaction scenarios with more dynamics and different setting, using such dataset helps to test and verify that our models are also valid and can be used as well in less dynamic setting with different behavior.

VFOA targets and statistics: As mentioned in Chapter 3.4, VFOA annotations are available for all sequences of this dataset. Each of the participants has 5 possible gaze targets: one other participant, two meeting organizers (observers), the slide screen and the table defined. These targets in addition to 'other' for looking elsewhere define the VFOA target set by F =

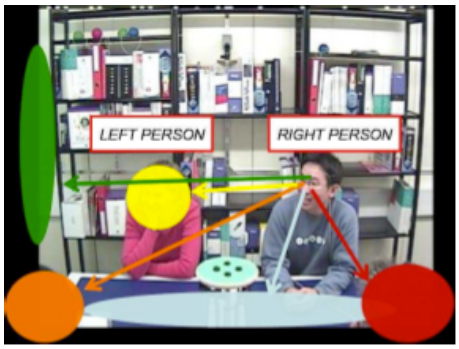
a)  b) 

Figure 5.2 – Meeting Data set. (a) A view from the meeting room and settings. Where two organizers O_1 and O_2 are seating on the left side of the table(O_1 is the one closer to the slide screen) and two participants are seating on the right side. (b) Data set view, with VFOA targets for the participant seating on the right.

Label	OP	O_1	O_2	TB	SS	$other$
Frame Frequency	0.15	0.35	0.05	0.17	0.25	0.02
Event Frequency	0.15	0.36	0.12	0.16	0.17	0.03
Average Duration (frames)	58.2	58.6	26.1	62.6	84.8	40.4
Average Duration (seconds)	2.3	2.3	1.0	2.5	3.4	1.6

Table 5.3 – VFOA frequency for the meeting dataset in percentage of frames, events frequency, and average event duration in number of frames and in seconds.

$\{OP, O_1, O_2, SS, TB, other\}$.

Table 5.3 provides the VFOA statistics for this recording from three dfferent participants. As can be seen, the first important target that participants look at is the first organizer and the second one is the slide screen which is used to show and explain the materials of the meeting. In terms of durations, events are generally short. However, people make longer gazes on the slide screen as compared to the organizer and another targets. This is also due to the importance of the slide-screen in this scenario and the fact that following the slides is less interactive than group conversations and therefore there are less interruptions there.

Head pose inputs: For this dataset we performed the experiments using ground truth head poses, captured from flock of bird sensors.

Gaze directions: Position of objects and people (and their head) were assumed to be known and fixed for each recording (thus neglecting people's motion), and therefore defined mainly from the geometrical setting.

Parameter setting: We followed the work of [Ba and Odobez, 2009] for setting the parameters for this dataset. We set the variances of the Gaussians according to the size of the targets and the expected noise and uncertainty. For the meeting data we use the same values as

Table 5.4 – Performance on the Meeting data

Person	Training	Baseline	Model G1	Model G2	Model G3
Person on left	same seat	63.8	64.9	66.3	**68.5**
Person on left	other seat	63.2	65.1	65.2	**67.8**
Person on right	same seat	56.7	58.8	58.6	**60.0**
Person on right	other seat	44.7	59.1	59.5	**60.1**

in which are $\sigma_{pan}(O_1, O_2, OP) = 12$, $\sigma_{pan}(SS) = 25$, and $\sigma_{pan}(TB) = 20$ for the pan, and $\sigma_{tilt}(O_1, O_2, OP) = 12$, $\sigma_{tilt}(SS, TB) = 15$ for the tilt. Here, σ_{pan} and σ_{tilt} denote the standard deviations for pan and tilt angles.

The initial value for the reference direction is particularly important for the baseline where it remains the same over time, but less important for the other models as the reference value evolves and quite rapidly becomes the average over head pose values. For the baseline, we experimented with setting the reference as the middle of the gaze target directions, which was shown to work the best in previous works [Ba and Odobez, 2009].

VFOA recognition results: Table 5.4 shows the results of the three models. The first model outperforms the baseline, particularly in more mismatched conditions, when parameters are learned from the other seat, exhibiting therefore a better adaptation capacity. The main (mismatched) parameters leading to the degradation is the parameter α^{pan} of the gaze model (see Eq. 4.5) that directly impacts the prediction of the head poses: for the 'person left', the optimal parameter is around 0.8, which can be understood as the person has to rotate more the head to look at the different targets (with the slide screen completely on his right, and the 'person left' completely on his left). For the 'person right', however, the optimal value for α^{pan} is around 0.5. The different values of α^{pan} obtained from two different seats could be due to the fixed choice of the reference (in the middle of the targets; that is i.e. 0 degrees for person left and 45 degrees for person right) and this fixed body reference should be used for a larger visual domain. This effect does not exist for the first model G1 and the chosen parameters through cross-validation are consistent with an optimal value for both seats around 0.7 for α^{pan}.

On the other hand, we can see that G3 performs better than the G1 in most cases. We observed that the improvement happened mainly during consecutive gaze shifts involving stable head pose changes. This is due to the fact that in this scenario people anticipate that they may go back to the previous gaze in a discussion. Therefore, previous gaze plays an explicit role here.

Figure 5.3 – NaoD dataset sample image.

Label	OP	Nao	$booklet$	$other$
Frame Frequency	0.49	0.34	0.13	0.05
Event Frequency	0.38	0.39	0.13	0.11
Average Duration (frames)	26.2	17.3	19.9	9.4
Average Duration (seconds)	2.9	1.9	2.2	1.0

Table 5.5 – VFOA frequency for NaoD dataset in percentage of frames, events frecuency, and average event duration in number of frames (9 Fps) and in seconds.

5.4.2 NaoD dataset

The second dataset for our experiments explained in Section 3.3 is a video recorded by our robot Nao envolving three participants (two at any given instant) as shown in Figure 5.3 . The total duration of this video is 22 minutes. In this case there are two participants seating in front of Nao as shown in 5.3.

VFOA targets and statistics: As mentioned in Chapter 3.3, we have annotated VFOA for this dataset. Each of the participants have 3 visual targets: the other participant, Nao and a booklet which they refer to during the recording and is placed on the coffee table in front of the participants. By adding 'other' for looking elsewhere, the total set of VFOA targets is defined by $F = \{OP, Nao, booklet, other\}$.

Table 5.5 provides the VFOA statistics for this recording from the three different participants. As can be seen, as a consequence of the scenario, looking at the other partner is clearly dominating and after that is looking at Nao as the main topic of discussion. In terms of durations, people make longer gazes on the other person (2.9 sec) as compared to the robot (1.9 sec) or the booklet (2.2 sec).

Head pose inputs: For this dataset, we do not have ground truth head poses. Therefore, our experiments are done using the head poses obtained from our previous tracker (section 5.2) which does joint tracking and head pose estimation.

Gaze directions: The participants were seating and were more or less static with respect to

Table 5.6 – Performance on NaoD data

Person	Baseline	Model G1	Model G2	Model G3
Person1	46.3	62.6	62.6	**63.4**
Person2	89.6	94.4	94.3	**94.7**
Person3	63.8	63.4	63.5	**64.6**

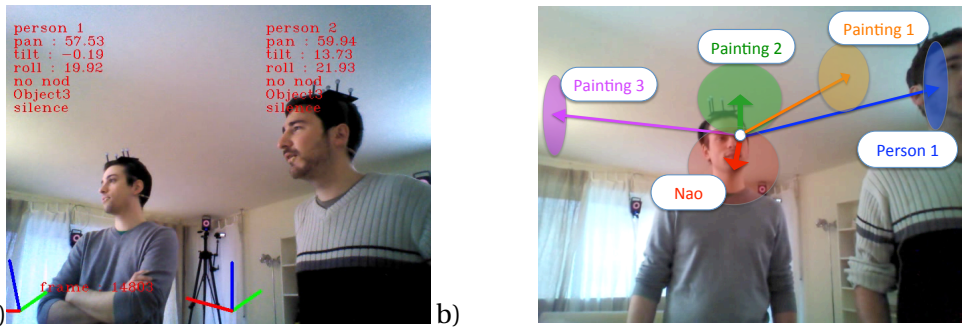
the camera Nao's head. Therefore, gaze directions to look at different targets were asumed to be fixed and people's slight motions were neglected. These directions were defined from the geometrical setting.

Parameter setting: The initial value for the reference direction is considered to be Nao's direction (i.e. 0 for pan and tilt angles) which is a reasonable choice in human robot interaction scenario when people face the robot. We set the standard deviations of the Gaussians in the likelihood model of the targets to 10 for the pan and 8 for the tilt angle.

For the rest of the parameters, since there are only a few number of people in this dataset with different gazing behaviors cross-validation will not produce reliable parameters. To choose the parameters we consider the meeting data as the training set and use parameters obtained from that data for running our algorithms on Nao's data. Note however, that the resulted α^{pan} value from meeting data is 0.7. In this Nao dataset this ratio is big considering the fact that we use tracked head poses which are a little underestimated. Therefore we do our experiments with a smaller value of 0.65.

VFOA recognition results: The results with this dataset are summarized in Table 5.6. Despite the quite different setting, the conclusions are similar to the meeting data. However, model G1 outperforms the baseline with a larger difference. This is particularly true for the first person, who was sitting on the edge of the sofa, and being more dynamic during the interaction, shifted her body orientation towards both the robot and the other participants, wheras the two other people remained more firmly seated in the back of the sofa and thus remained oriented towards Nao, which better matches the looking at Nao assumption of the baseline. Also, model G3 performs better than model G1 for all of the sequences.

Note that the results for person 2 are in general much higher than the other two participants. Person 2 is the guest who comes in at the second part of the recording. His role is to listen to the explanations of the other person. Therefore, he sits very calmly, does not use much body dynamics and mostly looks only at the robot and the partner and not the bookflet. Therefore, the recognition task becomes much easier for him than the other two participants.



a) b)

Figure 5.4 – Vernissage dataset. a) Potential ambiguity between looking at painting 3 or the partner, for the person on the right. b) VFOA targets.

5.5 VFOA recognition results on Vernissage dataset

5.5.1 Dataset properties

We used the Vernissage dataset explained in section 3.2 and illustrated again in Figure 5.4 to conduct most of our experiments. As mentioned in section 3.2, it involves two participants standing in front of Nao and free to walk around and look at different objects. In each recording (10 minutes on average), Nao first engages with the two participants and explains them three paintings. Then, in the second part, he gives them a quiz in which participants could discuss before the person to whom a question was addressed gave the answer. Both parts are approximately of equal duration. Some of the questions (4 out of 10) referred to paintings in the room. We denote the person on the left side of the robot by 'person 1' and the one on the right side of the robot by 'person 2'.

VFOA targets and statistics:

Given the scenario, 5 main VFOA targets have been identified and shown in Figure 5.4 (a). They are: Nao, $partner(ptr)$ (the other participant), and the three paintings pai_1, pai_2, and pai_3. In addition, we defined a label $other$ to denote a person looking at any other place in the room. In the dataset there are two additional labels, DK (don't know) is used when there is too much ambiguity between several VFOA targets and making a decision for the annotation is not possible, and NV (not visible) is used when the person is not in the robot's field of view. Data with VFOA labels DK and NV are not used in the evaluations and the VFOA target set is defined as $\{Nao, partner, pai_1, pai_2, pai_3, other\}$.

Table 5.7 provides the annotation statistics from eight of the recordings. As can be seen, as a consequence of the scenario, looking at Nao is clearly dominating, especially in terms of durations and is characterized by long gazes (average duration of 2.6 s). Among the paintings, pai_2 is more important because it is right above the robot's head and easier for the participants to look at. However, note that the occurrence frequencies are not distributed evenly during

Label	NAO	ptr	pai_1	pai_2	pai_3	$other$	NV	DK
Frame Frequency	0.43	0.11	0.06	0.14	0.06	0.09	0.06	0.05
Event Frequency	0.25	0.10	0.06	0.14	0.06	0.17	0.04	0.19
Average Duration (frames)	78.9	54.1	44.3	47.2	48.4	25.7	67.6	11.9
Average Duration (seconds)	2.6	1.8	1.5	1.6	1.6	0.8	2.2	0.4

Table 5.7 – VFOA frequency for Vernissage dataset in percentage of frames, events frecuency, and average event duration in number of frames (30 Fps) and in seconds.

the sessions: in the first part (introduction to the paintings), looking at paintings obviously happen more often; during the quiz part, interacting with and looking at the other person is more frequent.

Head pose inputs:

As head poses, we used both measures derived from Vicon (a motion capturing system) data and estimates obtained from a computer vision algorithm. After inspection, the head pose Vicon measures of one sequence happened to be inconsistent in time (the head-bands attaching the Vicon markers to people head might have moved), and we dropped it.

Pose estimated from videos were obtained by applying the particle filter tracker framework described in section 5.2. In this dataset Nao is performing head gestures (pointing to paintings, rotating the head to address people, nodding) that greatly affects the video quality (with people disappearing from the field of view, lighting changes, etc). It thus happened that results were not very accurate. Since our goal is to evaluate VFOA performance under reasonable head pose estimation, the tracker output was filtered by keeping only track segments that matched the (sparse) ground truth location available in the dataset as mentioned in Section 3.2, and persons for whom the average pose error was too large or for whom the tracker recall was too low were removed. Ultimately, this resulted in a dataset of 14 persons, amounting to around 140 minutes of data for our experiments. On these sequences, the tracker could achieve an average recall (percentage of frames with an estimate) of 80.7%, (min: 48 and max:92), with average pose errors shown in Figure 5.5.

The overal tracker estimation distribution against the actual Vicon head poses is plotted as a quantile function in Figure 5.6. These curves suggests an underestimation of the pose. Moreover, for larger angles there is more underestimation and more ambiguity.

Gaze directions:

We also need to feed our algorithms with the gazing directions for different targets for each participant in terms of pan and tilt angles. Given that people are free to move in these recordings, these directions are not fixed and change over time. These values were obtained using the Vicon sensors which are placed on the Nao's head, participant's head and on each of the paintings as mentioned in Chapter 3.2. However, for a more general application, we

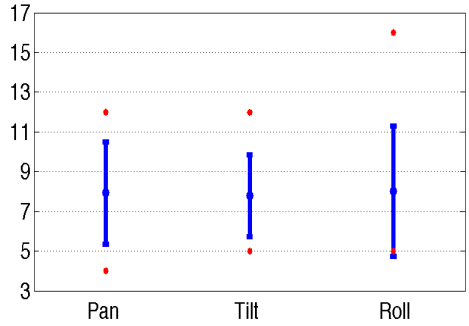


Figure 5.5 – Tracker head pose obtained from the 14 persons. Minimum, maximum, mean and standard deviations of the errors.

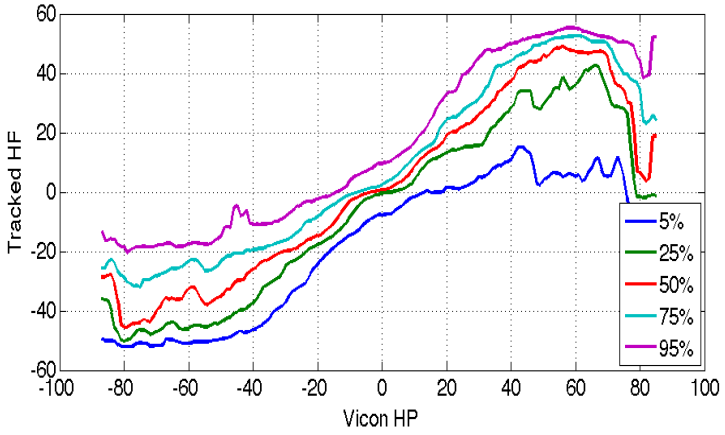


Figure 5.6 – Tracker vs Vicon head poses. Estimated head pose quantiles for the given head pose values. The tracker is relatively accurate up to 40 degrees, but with a tendency to underestimate the pose. This is accentuated for pose beyond 40 degrees.

assume that Nao knows the room's geometry and can localize itself in the room [Fojtu et al., 2012]. By tracking the participants and knowing its location regarding to the other objects in the room, it is capable of measuring these directions and using them for the recognition task.

Parameter Setting:

For both Vicon and tracked head pose data, the reference direction for the baseline was set as looking at Nao, which is a reasonable choice in our HRI scenario. Standard deviations of Gaussian were set to 20 and 10 for pan and tilt. The remaining parameters (including context tables) were adjusted by leave-one-out cross-validation separately for each of the models i.e. considering the rest of the all participants as the training set while testing on each participant.

Table 5.8 summarizes the parameters of the gaze-head pose mapping models described in

Table 5.8 – Parameters of the dynamical model obtained in majority through cross-validation on Vicon data. W^R, W^p and Δ^p are expressed in seconds.

Parameters	α^{pan}	W^R	W^p	Δ^p	α_H	α_1	α_2
Baseline	0.7	-	-	-	-	-	-
G1	0.6	20	-	-	-	-	-
G2	0.6	20	1	0.4	1	-	-
G3	0.7	20	-	-	-	0.22	0.07

Table 5.9 – Recognition rates of head-gaze mappings methods.

	Vicon head poses			Tracker head poses		
	Full	Explain	Quiz	Full	Explain	Quiz
Baseline	53.8	52.4	54.6	57.3	59.3	57.4
G1	65.5	68.8	64.2	59.1	61.7	58.7
G2	**66.6**	**69.9**	**65.3**	**59.8**	**62.3**	**59.3**
G3	64.3	66.7	63.3	56.7	60.2	56.0

Section 4.4 that were selected in majority for each of the dynamic model. We can notice that the selected value of α^{pan} (amongs values ranging from 0.4 to 0.9) corresponds to numbers reported in the literature. With respect to the size W^R of the window used to average the head poses and use it as an approximation of the body orientation, we can see that a rather short size of 20 second was selected (amongs values ranging from 20s to 50s). Indeed, while larger windows provide more stable results, they also introduce more lag to adapt to new situations in case of strong body shifts which occurs for instance when people look at painting pai_3 and then switch to looking at painting pai_1.

5.5.2 Results o head pose-gaze correspondence models

As first experiments, we evaluate and compare the different head pose-gaze dynamical mapping approaches (Baseline, G1, G2 and G3), leaving aside the context part. Experiments are done using both Vicon and tracker head poses. Table 5.9 summarizes the obtained results.

Vicon head poses

The baseline relying on the geometrical model to set the head pose means has only a 53.8% recognition accuracy. This is mainly due to the wrong predictions of the Gaussian means (head directions). In particular, as can be seen from typical confusion matrices. of the baseline (left matrices in Figure 5.7a) and 5.7b)) for a person located on the right (person 2) or left (person 1) in Figure 5.4a, the main source of confusion is between Nao and the painting pai_2. This is not surprising given their proximity in the gaze space, where they mainly differ in the tilt angular space. Similarly, as expected given the setup, confusion between looking at the third painting

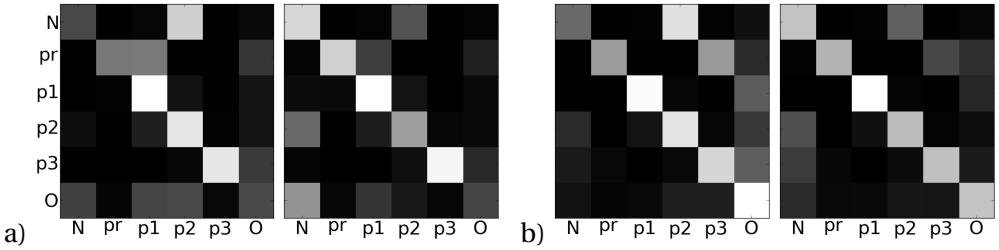


Figure 5.7 – Confusion matrices (rows are ground truth, columns denote the recognized labels) for (a) a person located in position 'person 1' in Figure. 5.4b) and (b) a person located in position 'person 2'. In (a) and (b), the matrices on the left are obtained from the Baseline model, whereas the matrices on the right are computed from the G2 results. For space reasons, VFOA targets in the legend of confusion matrices are denoted by *N* for Nao, pr for $partner$, pi for painting pai_i, and *O* for $other$. Notice how looking at Nao is often confused with looking at painting 2 (p2) and looking at the partner is confused with looking at painting 1 (for a 'person 1') or looking at painting 3 (for a 'person 2').

(pai_3) and $partner$ can be seen for the VFOA of person 2 (see Figure 5.4a) and between the first painting (pai_1) and $partner$ for person 1. Moreover, although the Gaussians standards deviations in the HMM are relatively large, several labels are wrongly recognized as looking at $other$.

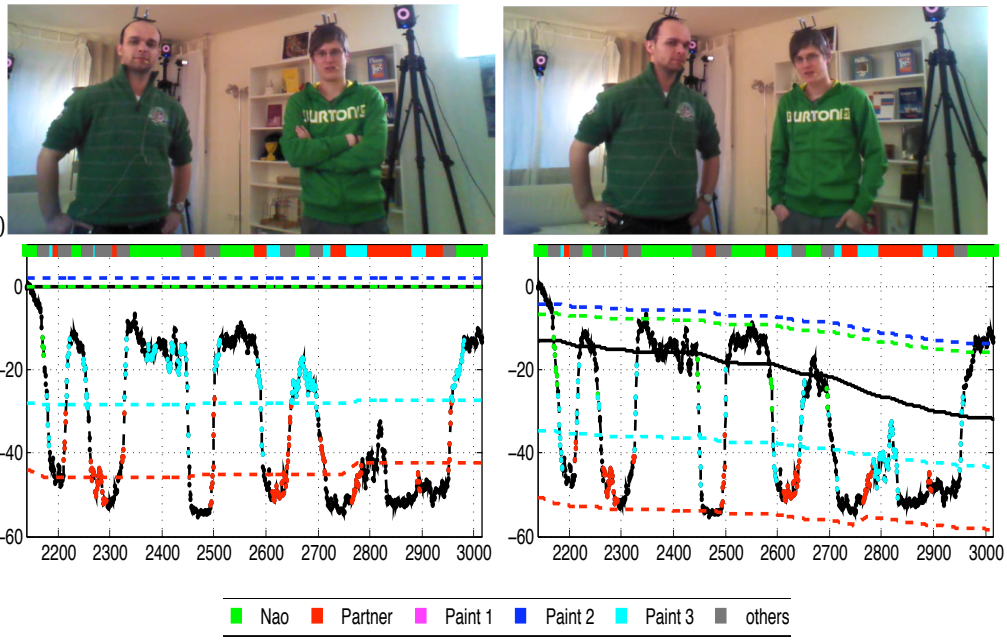
Among the different dynamic models, G2 which implements the midline model is the best, leading to an average gain of 13% over the baseline. Notice that the gain is more important in the explanation part (17.5%) where people do not face the robot all the time, but orient their bodies towards the paintings (see Figure 5.4a for instance), rather than in the quiz part (10.7%) where people mainly stay oriented towards the robot. The confusion matrices on the right of Figure 5.7a) and 5.7b) obtained with G2, compared to those from the baseline clearly show that the gain is due to a reduced confusion between Nao and painting pai_2, a reduction of the misclassifications between $partner$ and the confusing painting (either painting pai_3 for person 2, or painting pai_1 for person 1), and less recognition as $other$.

Table 5.10 – For each target, the table provides the means of the angular errors (in degrees) between the head pose actually used to look at the target, and the prediction made either by the baseline or the G2 models. Vicon head poses are used.

	Pan angle		Tilt angle	
Target	Baseline	Model G2	Baseline	Model G2
Nao	7.5	4.4	5.8	3.8
$partner$	10.6	9.9	5.0	5.0
pai_1	21.6	14.1	11.9	13.1
pai_2	38.6	30.3	7.2	4.6
pai_3	47.4	39.5	8.0	4.8

Since between the different models, the only elements that change are the setting of the mean

(a)



(b)

Figure 5.8 – (a) Left: during frames 1700-2200, Nao is the main speaker and participants tend to look straight at him. Right: afterwards (quiz part) participants discuss together, and alternatively look at the robot and the second person (amongst others). Their reference direction is thus different, and so are the poses for looking at Nao. (b) Vicon head pose (pan angle) of the person on the right in image (a). The ground truth VFOA is displayed in the top bar, with color codes displayed below the plot. The head pose pan data is displayed in the graph. It is black when the recognition is correct, and in the color of the wrongly recognized VFOA otherwise. Dashed lines indicate the pan pose mean for looking at each target for the baseline geometric model (left), or dynamic model G1 (right). In this later case, the black line shows the head reference R_t (computed on the average of head poses in previous frames). With the dynamic reference, head poses for looking at each of the target are better predicted, like for looking at Nao (despite its high variability: pan near 0 at frame 2150, near -17 at frame 2550).

head pose for looking at individual targets, the VFOA recognition improvement is clearly is due to a better prediction of the expected head pose for looking at the different targets. To quantify this prediction improvement, we performed the following experiments using the ground truth VFOA. We compared at any given instant the participants' head pose used for looking at the VFOA target with its predicted value as given by the baseline and the model G2. Ideally, these two measures should coinside. The resulting mean error computed over the 18 individuals are shown in Table 5.10, where the smaller the error, the better the predicted head pose is. As can be seen, the pan angle errors are smaller for all VFOA targets when the dynamic model G2 is used, and in all but one cases for the tilt angle. This is particularly important for Nao and pai_2 which differ only slightly in their tilt angle.

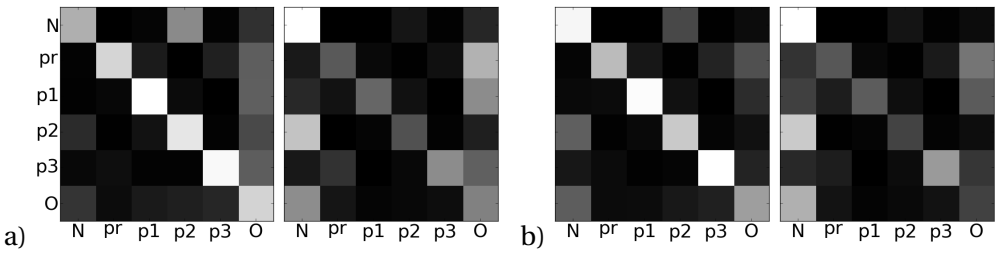


Figure 5.9 – Vicon vs Tracker data. (a) average confusion matrices obtained using either the Vicon (left) or tracker data (right). (b) confusion matrices for Vicon (left) vs tracker (right) data using the dynamic model G2.

Given the small gain obtained by G2 over G1, we can conclude that the dynamic mapping (through the estimated body orientation) is what contributes the most to the improvement. Qualitatively, its effect is illustrated in Fig. 5.8. Nevertheless, the midline effect (centripetal movement) is also useful as it provides better recognition results in 13 out of 14 sequences. However, since this effect is happening rarely in the data its effect on performance is also small.

Finally, we see that model G3 performs much better than the baseline, but a little worse than G1 and G2. Note however that applied to meeting and NaoD data, G3 was shown to outperform them, indicating that it might be more appropriate in presence of more frequent and shorter gaze shifts.

Head Pose tracker data

With these data, the main conclusions (ranking of the dynamical models) drawn using Vicon head poses hold. However, here the baseline already gives good recognitions as compared to the Vicon data, and the improvement is smaller (2.5%). This situation can be understood by looking at the average confusion matrices shown in Figure. 5.9 and comparing them with those of the Vicon data. As can be seen from the diagonal elements, the higher accuracy in the baseline is mainly due to a higher recognition for the Nao class, which, given its predominance in the data, results in a higher frame-recognition rate.

A potential explanation for the bias towards Nao can be understood by looking at the tracker estimation results in Figure 5.6 which displays estimated values given by tracker for ground truth head poses given by the Vicon. As mentioned earlier, these curves suggest an underestimation of the pose in general, with the effect of favoring the recognition of Nao as compared to painting 2 for instance (painting 2 is in fact a set of 3 paintings and is wider than Nao). In addition, the underestimation for larger poses leads to head poses that do not match well any of the predicted VFOA targets, and result in a higher recognition of the $other$ label (right column in confusion matrix). The dynamical model G2 (most right matrix of Figure 5.9) tends to reduce the later aspect in certain situations, and to increase the recognition of some targets

like painting pai_3, including looking at Nao.

Statistical significance

In order to check whether the difference of the algorithms we propose is significant we also ran the statistical test described in section 5.3.3.

When using Vicon head poses, the result of different models Baseline, G1, G2 and G3 were compared together. All algorithms differences in performance showed to be significant at the significance level of 0.05. When applying the algorithms on the tracked head poses, the differences between all pairs of algorithms were statistically significant except for Baseline and model G3. Importantly, the difference between our selected model G2 and the baseline is significant with both Vicon and tracked head poses.

5.5.3 Results of conversational dialog context

To evaluate the contribution of the different contexts, we considered different settings: No context, one single context cue (speaking, addressee, or topic), and all cues together. Furthermore, we experimented the use of the context with both the baseline (static geometrical model where the body orientation is assumed to be equal to 0, i.e. facing Nao) and the best dynamic gaze prediction models (G2) to investigate whether the context is still useful when more accurate gaze-to-head pose predictions are exploited. Tables 5.11 and 5.12 show the results when using Vicon and tracked head poses.

When using Vicon data and the baseline dynamical model, we see that the performance improves whatever individual cue we consider (with around 7, 7.5 and 9.5% improvement using speaking, addressee and topic cues). The increase is larger when we use the topic context. Altogether, the use of all context cues brings a considerable improvement of more than 10%. This improvement is valid for all of the 14 persons, and is illustrated through the confusion matrices of 2 persons in Figure 5.10. As suggested by the shown confusion matrices, the context improves the recognition of all targets simultaneously, and is particularly helpful for removing ambiguities between Nao and pai_2, $partner$ and pai_1 and $other$ for most cases.

Looking at the combination of context with the dynamical model G2, we can first notice that the context alone (i.e. with the geometric model and static body reference) does not reach the accuracy of the dynamical setting (64.2% with context vs 66.6% with G2). Still, the effects of both approaches are complementary, as the addition of context improves the results of G2 with a gain of 6% when using all cues, and further decreases the confusion between VFOA targets similarly to what is explained above (i.e. between Nao and pai_2, $partner$ and pai_1 or pai_3). The improvement due to context is observed for 12 out of 14 sequences, and the degradation for the other 2 sequences is very small (2.0% and 0.1%).

Interestingly, the results with individual cues exhibit different behaviors depending on the

Table 5.11 – Recognition rates using dialog act contexts - Vicon head poses

	Baseline Model			Model G2		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	53.8	52.4	54.6	66.6	69.9	65.3
Speak.	60.9	58.3	62.1	70.2	72.3	69.4
Addr.	61.4	59.8	62.2	70.8	73.1	69.9
Topic	63.4	62.2	64.0	72.1	75.3	70.9
All	**64.2**	**63.3**	**64.7**	**72.6**	**75.9**	**71.3**

Table 5.12 – Recognition rates using dialog act contexts - tracked head poses

	Baseline Model			Model G2		
Context	Full	Explain	Quiz	Full	Explain	Quiz
None	57.3	59.3	57.4	59.8	62.4	59.3
Speak.	59.1	61.5	59.1	61.0	63.1	60.9
Addr.	59.5	62.2	59.3	61.3	63.7	61.0
Topic	60.1	64.2	59.5	62.0	65.6	61.4
All	**60.6**	**65.4**	**59.8**	**62.4**	**66.4**	**61.7**

interaction phase. As can be seen, the communication cues (speaking, addressee) which emphasize Nao or people as VFOA prior make a bigger increase in performance during the quiz, which is more interactive, and lower increase during the painting explanations, whereas the topic context improves almost equally on both parts. Finally, using all cues, the performance is higher in all situations.

Considering the results on the tracked head poses, shown in Table 5.12, we can see that the main conclusions still hold. Individual cues are all useful, the topic cue is more beneficial especially on the explanation part. Combined with the baseline, the context and dynamical model lead to a total improvement of 5%, a gain that is smaller than with Vicon due less accurate head poses and thus more ambiguous situations.

Statistical significance. In order to check whether adding the context to our algorithms makes a significant dfference we ran the statistical test described in Subsection 5.3.3 with the significance level 0.05. We considered the addition of diffrent contexts ('none', 'speak', 'addr', 'topic', 'all') when Baseline or model G2 are used as the head pose-gaze correspondence models.

Considering the Baseline model, when using Vicon head poses, the performance differences between all pairs of models were statistically significant, except for when we were using either 'addr' or 'topic' contexts. When using Tracker head poses however, all pairs ofmodels were

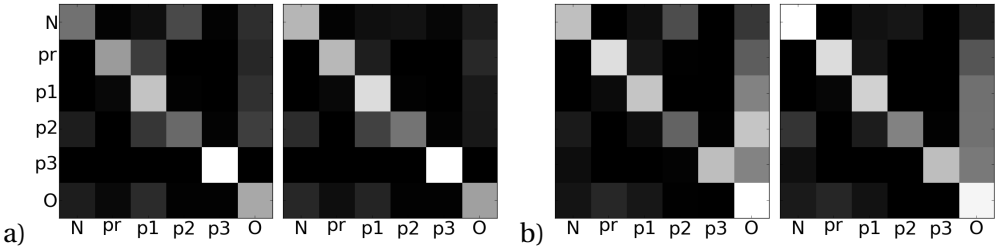


Figure 5.10 – Context effect (Vicon data) (a) the image on the left shows the confusion matrix for a given participant when context information is not used while the right one shows the matrix when using the context. (b) shows the same matrices for another participant.

significantly different except for the case of using either 'speak' or 'addr' contexts. Considering the selected dynamic model G2, with both Vicon and tracker head poses, all pairs of models were significantly different except for the case of using either 'speak' or 'addr' contexts.

What is specially important is that all the individual cues provide significant improvement compared to not using context at all. This, in addition to the fact that using a given individual cue does not always provide significant improvement over the other ones suggests that these cues can provide complementary improvements over eah other. In addition to that, the combination of them also provides significant improvement over the individual ones.

5.6 Conclusion

In this Chapter we addressed improving VFOA recognition from head poses in an HRI context using two different solutions. First, we proposed algorithms inspired from body, head and gaze behavioral models to improve the dynamic prediction of the head pose used to look at different VFOA targets. Our experiments on a challenging dataset showed that these models indeed generated more accurate predictions, improving head pose-gaze direction association for all VFOA targets, resulting in a performance increase of more than 10%. Secondly, we proposed a contextual VFOA recognition approach to exploit the robot's gaze-related conversational context (communicative cues, topical cues). It was shown to greatly improve results, and to be complementary to the head-pose dynamical model. Altogether, the combination of the two approaches led to an increase close to 20% in VFOA recognition.

The experiments also showed that obtaining unbiased and accurate head pose is important, as the improvement was smaller using head poses derived from our vision tracker than with the Vicon ones. Such pose estimation improvements come from advances in sensing, and in particular the use of RGB-Depth camera like Kinect. In practice, given the availability of real-time head pose tracking with such device[1], we expect our model to be directly usable by researchers and developers in the HRI or ECA field. Furthermore, the most effective part in our dynamical gaze-to-head prediction approach relies on the use of the body orientation.

[1] http://msdn.microsoft.com/en-us/library/jj130970.aspx

Hence it would be interesting in the future to test our method on a dataset with available RGB-D dataset that would provide a more direct and more accurate way of estimating it than what we propose. In another direction, with higher definition images, using image-based gaze directions [Gorga and Otsuka, 2010] would be beneficial, and could be combined with our approach. Our prediction model could provide priors on the gaze and be fused with actual image measurements even in noisy conditions.

On the context side, since the dialog act information required by the method is directly incorporated in the dialog system and used at runtime, the model can be exploited for any interactions and in any other scenarios implying objects with the robot is aware of. Finding more systematic ways of setting appropriate VFOA statistics is an avenue for future work, as well as the addition of timing information (how long is a dialog act active?) as well as the use of other cues that can affect the attention of interacting people, like the robot's gestures.

6 Addressee Estimation

6.1 Introduction

In human-human interactions, who is the current addressee of a spoken utterance, whether it is an individual or a group of people, plays an important role in regulating the conversation. The same holds true for natural and human-like interaction with robots. Most importantly knowing the addressee of a person's utterance is useful for the robot to decide automatically if he "should" or "should not" react and possibly respond to the utterance.

As discussed in Chapter 2, the gaze or VFOA of the speaker has been shown to be the most informative cue for recognizing the addresseehood (e.g. [Katzenmaier et al., 2004, Takemae and Ozawa, 2006]) since people mostly look at the person they are addressing rather than others. It has been shown as well that, even when we have accurate gaze information, this might not be sufficient for addressee estimation. As a result, researchers have investigated other cues (e.g. lexical, prosodic cues) to provide context and improve performance [Jovanović et al., 2006, Huang et al., 2011]. In this chapter, we also followed this path and investigated in our Vernissage scenario the effect of different contexts on the overall addressee estimation task where context can act at two different levels: directly for addressee recognition, or indirectly on VFOA recognition. We summarize below the main points we addressed.

First, context can be used directly for addressee classification. Thus, in addition to the speaker's gaze, we considered several other cues as inputs to the addressee classifier: the gaze cues from the other participants, whether the current speaker spoke the previous utterance or not, the subjective difficulty of the quiz question was also used. When relying on ground truth VFOA, these contextual cues were shown to provide a slight improvement for addressee estimation. However, this effect when using noisier automatically estimated VFOA had not been studied. Since the use of noisier VFOA input could be expected to degrade the addressee recognition performance, the hope in this situation is that the context would play a more important role and lead to a larger improvement, and we investigate this question in this Chapter.

Context could also be used to improve VFOA recognition, and thus indirectly addressee

recognition. In this work, we use the robot's conversational context as explained in Chapter 4. Although using context has been shown to improve VFOA recognition (Chapter 5), the question that arises is whether such improvement would benefit or not other tasks relying on VFOA such as addressee.

Finally, from a computational perspective, there are several issues that we investigated. Which features should we use as output from the VFOA recognition module? how should we normalize them? Importantly, in a given scenario the VFOA module may have to monitor gazing at other visual targets in the scene (for instance, the robot needs to monitor whether people look at the right paintings to check that they follow his explanations in our scenario) in addition to potential addressee targets. In this Chapter we show that accounting for such targets significantly affect the addressee recognition task which suggests that it might be better in a robotic system to exploit a VFOA module specifically devoted to addressee that only cares about potential addressee targets which are the robot and other participants. The material of this Chapter was published in [Sheikhi et al., 2013a] as a collaboration with two other postdoctoral researchers D. Babu Jayagopi and V. Khalidov.

In summary, in this Chapter we investigate the following. First, we investigate the use of context for addressee estimation, both at the addressee and VFOA levels. Secondly we study different computational issues that can significantly affect the addressee recognition task, as described above. Finally, we show that, when using automatically recognized VFOA (VFOA estimated using tracked head poses) for addressee estimation in our multiparty HRI setting, the performance does not drop much as compared to using ground-truth VFOA or VFOA estimated from ground-truth head-poses.

This Chapter is organized as follows. In section 6.2 we give an overview of our full addressee estimation system with its different parts for head pose tracking, VFOA recognition and addressee classification. Section 6.3 gives more details on the recognition approach used for estimating addressee. Section 6.4 contains our experimental protocol and results using ground truth VFOA and automatic VFOA (both from estimated head poses and ground truth head poses). Section 6.5 provides a conclusion for this Chapter.

6.2 Addressee detection: scenario and system overview

We propose a system for addressee detection that could be used in realistic robotic setup where a humanoid robot with significant nonverbal displays induces nonverbal behavior in human participants that interact with the robot in a natural way. More specifically, in this thesis, we considered the Vernissage scenario and used the data from our dataset for evaluation.

Data Considered:

We use 7 interactions from the Vernissage corpus (see overall description in Chapter 3). For the addressee analysis in this Chapter, we use only the quiz part, which consists of nine questions

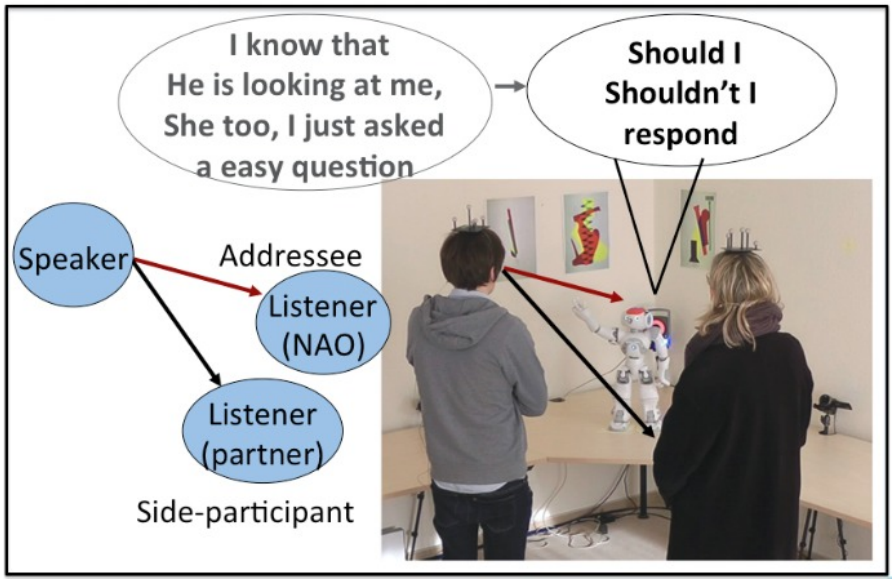


Figure 6.1 – Overview of the addressee estimation task. Based on different information (top left side: I know that he is looking at me, she too, and I just asked an easy question), the robot has to infer whether a person talked to him or not.

(or **quiz episodes**) in art and culture, which are the same across the participant set. The reason we only use the quiz part is the fact that in the explanations part of the recordings the robot is speaks most of the time and the participants only follow the explanations. Moreover, in the few cases where a participant speaks, he/she usually only provides short responses to the robot to show his/her agreement, disagreement or only provides backchannel which are not very interesting for the addressee estimation task. In the quiz parts, some of the questions are about a set of paintings that NAO introduced to the participants before the quiz in the explanation part of the scenario. In general, participants discuss among themselves before answering a question, but this is not always the case (e.g. when questions are 'easy'). Figure 6.1 illustrate the addressee estimation task in this setup. The robot has access to the speaking status of the participants, their gaze cues and additional cues from the interaction and given this information should detect if it is addressed to provide the appropriate response.

System.

Our system consists of three different modules: **head pose tracking**, **VFOA recognition** and **addressee estimation** as illustrated in Figure 6.2. The head pose tracking module described in Section 5.2 uses the video sequence captured by the robot to localize faces and extract head poses. These head poses are then used in the VFOA recognition module described in Chapter 4 to estimate the participants' gaze direction and recognize the visual target they are looking at. Finally the addressee estimation module uses participants' VFOA during the utterances to detect whether the robot or another participant is the addressee of the speaker's speech. Addressee estimation is studied at the utterance level in contrast to head pose tracking

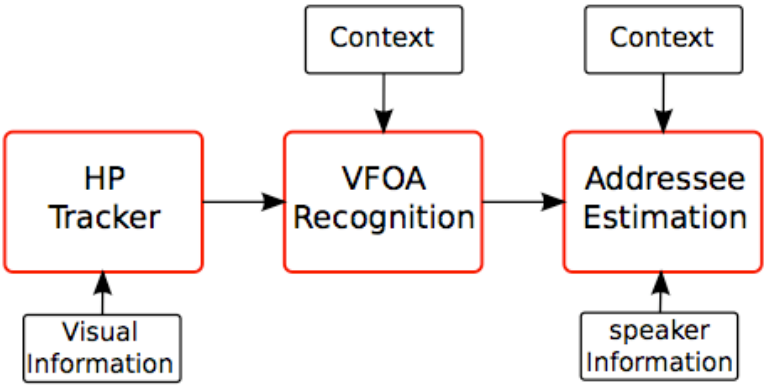


Figure 6.2 – Addressee detection system. It consists of head pose tracking, VFOA recognition and addressee estimation. Context can be used for both VFOA recognition and addressee estimation.

and VFOA recognition which are studied at frame level. As shown in the Figure, context information can be utilized in both the VFOA recognition and addressee estimation modules. Currently the context in the VFOA recognition module is provided by the robot's conversational state as defined in Section 4.5, whereas in the addressee estimation, it comes from the other participant's gaze information, the previous speaker and the difficulty of the quiz question. The addressee estimation module is explained in the following Section.

6.3 Contextual Addressee Estimation

We predict the addressees on semi-automatically estimated utterances using the speaker's VFOA and other contextual cues. Utterances are extracted as explained in Chapter 3). As contextual cues, we investigate not only the cues from the speaker, but also gaze cues from the side-participant, and contextual prior information about the current activity (here the quiz), and the current dialog context (previous speaker). In the following, we detail the features we used and present the recognizer used.

Features:

For every utterance, we defined the following features, summarized in Figure 6.3:

- SpkrL@NAO: denoted by SN, represents the proportion [%] of time when the speaker looked at NAO during the last one second of an utterance;
- SpkrL@Ptr: denoted by SP, represents the proportion [%] of time when the speaker looked at the partner;
- PtrL@NAO: denoted PN, represents the proportion [%] of time when the partner looked at NAO;
- PtrL@Spkr: denoted PS, represents the proportion [%] of time when the partner looked

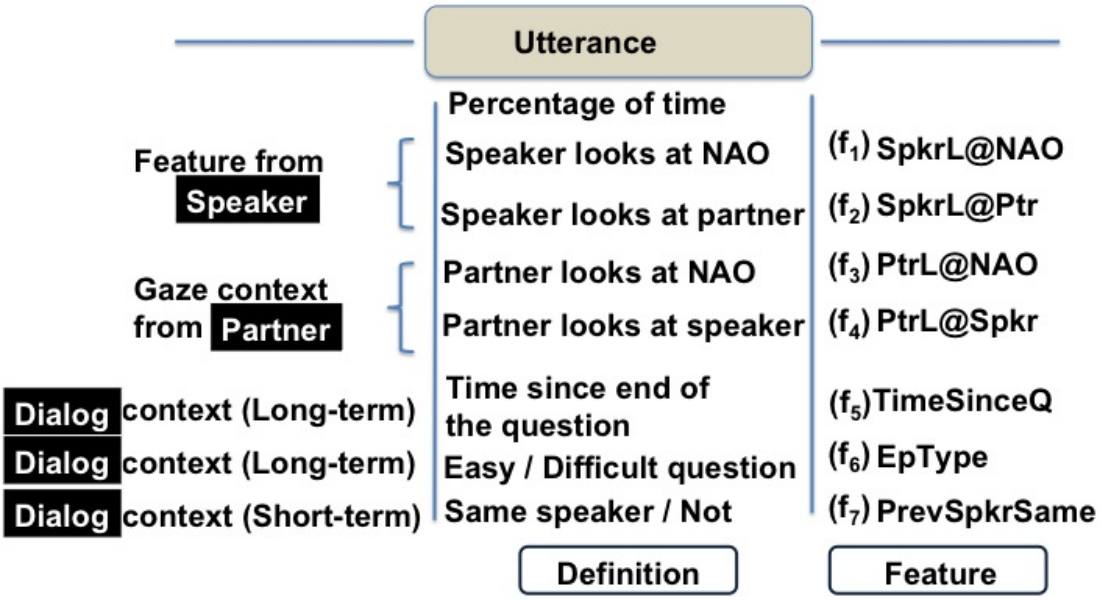


Figure 6.3 – Addressee estimation task. Tested features and their encoding.

at the speaker;

- TimeSinceQ: defined as the time difference between the end of a quiz question and the start of the utterance;
- PrevSpkrSame: defined as whether the previous speaker is the current speaker (coded as 2) or not (coded as 1);
- EpType, the episode type that roughly indicate the difficulty of the quiz question: 1 being easy and 2 being difficult.

In this work, we assigned the difficulty of the question manually, but it it could also be learned over multiple sessions i.e. with experience, as we implicitly propose towards the end of the addressee experiments. A question could be difficult because the listeners do not understand what the robot is saying or they follow the question but do not know the answer. Note that while PtrL@NAO and PtrL@Spkr are contextual cues from the side-participant, EpType is a task-related long term context, and PrevSpkrSame is a short-term context about the dialog.

Classifier:

We used a supervised and discriminative model to predict the addressee. This classifier is a Logistic Regression where the log-ratio of the probability of addressing the partner vs NAO is a linear function of the features:

$$\begin{aligned} LR_{LogR} &= \log\left(\frac{P(Add = NAO|f_{1:N})}{P(Add = Ptr|f_{1:N})}\right) \\ &= \beta_0 + \beta_1 f_1 + \beta_2 f_2 + ...\beta_N f_N \end{aligned} \tag{6.1}$$

the β parameters are estimated during training and indicate the relative importance of the features. When LR_{LogR} is greater than 0, then the estimated addressee is NAO.

6.4 Experiments

In this Section we will describe our experimental details and explain different experiments we conducted on the Vernissage data. We provide our experiment results and make the important conclusions made by the experiments.

6.4.1 Experimantal details

Ground truth data: In addition to the VFOA annotations which are provided for the Vernissage dataset, addressee is also manually annotated as mentioned in Chapter 3. In the 7 sequences that we used, which corresponds to data from 14 participants, there are 374 utterances of human participants in total, of which 176 were directed towards Nao, and 198 were directed towards a human partner (denoted Ptr henceforth).

VFOA recognition results: We have estimated the VFOA using the algorithms in Chapter 4 run on both the Vicon and tracked head poses. The best performing gaze-head dynamic model, G2, was used once without context and another time leveraging all contextual information.

Experimental protocol: All experiments were obtained using a leave-an interaction-and-quiz question-out evaluation. This means that when evaluating addressee recognition on the data of a question within an interaction, all data from the same interaction (same people) and of the given question (in the other interactions) were excluded from the training.

Statistical significant: In order to evaluate whether the improvements provided by different algorithms were statistically significant, we used the McNemar test. The exact McNemar test looks only at the samples where the two algorithms give different results. It checks whether an algorithm provides almost systematically the same or a better answer than the other one.

6.4.2 Results

We have performed three sets of addressee experiments to study our research questions. In the first one, we compare how the results vary depending on the VFOA data (ground truth data, estimations obtained from either Vicon head poses or from tracked head poses). In this experiment no context information is used. In the second experiment, we compare the results when context information is used at either the VFOA or addressee level.

In the third experiment we consider the situation when the robot is not aware of other visual targets in the scene and only considers itself and the other participant as possible visual targets. We would like to compare this condition with the initial one where the robot extracts

the full VFOA information consisting of all visual targets. This is important since for addressee estimation task it might be better to only consider looking at the targets which are also possible addressees. Considering that people can look at additional targets might distract the system from the information which really matters for detecting the addressee. For instance it might be more useful to assume that the speaker is looking at a participant even if this could be confused with looking at a painting.

The results are given in Figure 6.4 - 6.5. Note that predicting the majority class would provide 52.0% accuracy since we had 52.0% occurrences of Nao being addressed.

Analysis 1 - VFOA data input.
We first consider results obtained with VFOA features from the speaker (SP, SN, and both) computed without context information. We compare the addressee estimation accuracy obtained when using the ground-truth VFOA annotations, VFOA automatically derived from Vicon head poses or from tracker head poses. As can be seen, while when using the VFOA GT looking at the partner (SN) happened to be a better cue for recognition than looking at Nao (with an increase of 7%), the reverse happened with the automatically derived VFOA cases, especially when using the tracker poses. Furthermore, in general, the combination of both SN and SP did not increase performance (a fact that will stay true even when using the context for VFOA recognition). Surprisingly, the results without context, obtained using the noisier tracker head poses rather than the Vicon ones, produced much better results (76.3% vs 67.9%) -this might be due to the head pose bias towards Nao that has a positive effect for addressee. Still the best result with VFOA GT (83.4%) is 7% higher than the best result using estimated VFOA (76.3%). Given the natural scenario that included paintings as potential VFOA, this result is already very good. Errors were mainly due to participants addressing Nao's request while still looking at the partner, and vice-versa, making a side-comment to the partner while looking at Nao. Automatic VFOA produced additional errors due to wrong VFOA estimates.

Analysis 2 - Using Context.
Next we considered using the VFOA context (Nao speaking status and addressed person, topic context), and addressee context (that included the partner features e.g. PtrL@Spkr (PS) and dialog context -EpType EP and PrevSpkr). Results are shown in Figure 6.4. We observe that while the VFOA context helped in general across all addressee context conditions in the Vicon case, it does not impact the tracker results very much. Indeed, smaller improvement is expected since the context's effect on VFOA recognition is smaller with head pose tracker (see Table 5.12). Moreover, the VFOA context gives a higher expectation for recognizing Nao as VFOA, which cumulates with the head pose pan angles underestimation towards Nao. Therefore, although the context still improves the VFOA, this bias towards recognizing Nao when people are speaking may reduce its impact for addressee estimation.

Secondly, we can notice the results when using addressee context. In the tracking case and also when Vicon head poses are used to estimate VFOA, the addressee context helps to slightly improve the results. The feature combination of SpkrL@NAO, SpkrL@Ptr, PtrL@Spkr, and

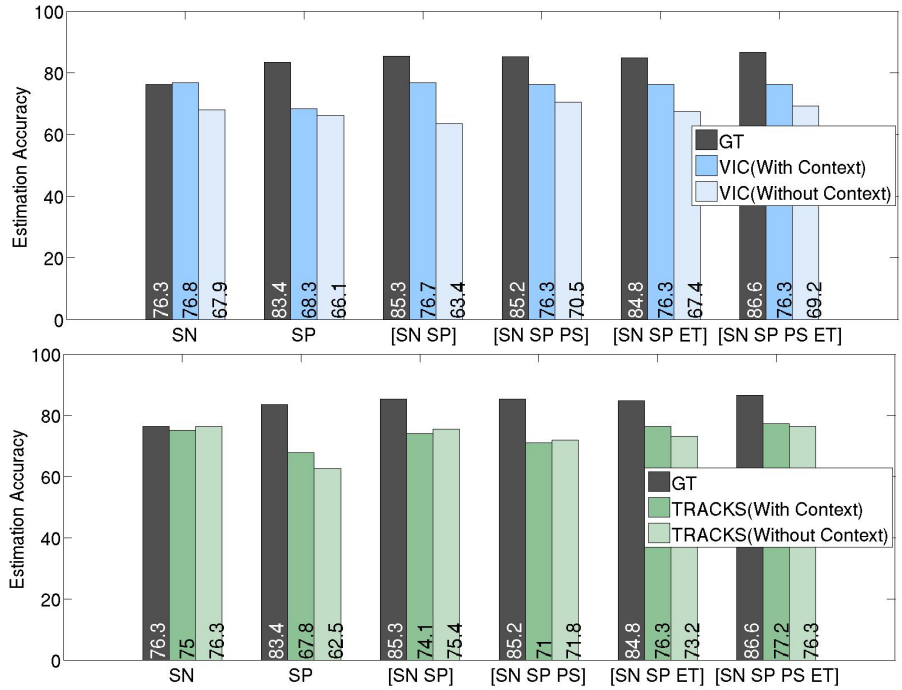


Figure 6.4 – Addressee recognition using GT VFOA and (Top) VFOA based on Vicon head pose (Bottom) VFOA based on tracker head pose.

EpType for example has the best accuracy of 77.2% in the tracking case. However, with accurate VFOA measures obtained from the ground truth, the gain of adding addressee context is higher. For the algorithms results on the ground truth VFOA we also ran statistical tests to see whether the difference between them is significant or not. As a result the difference between using the SN feature alone and using the feature sets { SN, SP }, { SN, SP, PS }, { SN, SP, ET } or { SN, SP, PS, ET } were statistically significant which suggest adding the contextual cues provides minor but significant improvement.

Experiment 3 - Generalization Capabilities

A change in the set of monitored VFOA targets may affect the VFOA measurements for addressee. More targets -e.g. more paintings in the room- may lead to more VFOA ambiguities, while using less targets than the number of real ones (e.g. if the robot is not aware of all paintings in the room) might lead to gaze erroneously assigned to a VFOA of the smaller set. In order to achieve robustness in terms of variation setting, we considered the following approaches:

E1. we considered normalized features sn and sp defined as the proportion of time gazed at a potential addressee compared to all the time gazed to any possible addressee: $sn = \frac{SN}{SN+SP}$ and $sp = \frac{SP}{SN+SP}$. Similarly, we defined pn and ps for the partner VFOA.

E2. we studied two alternative VFOA module settings. The "ALL TARGETS" setting is the one used in previous experiments, where the VFOA monitored looking at all targets including the paintings. In the "LIMITED TARGETS" condition, the VFOA module was configured to

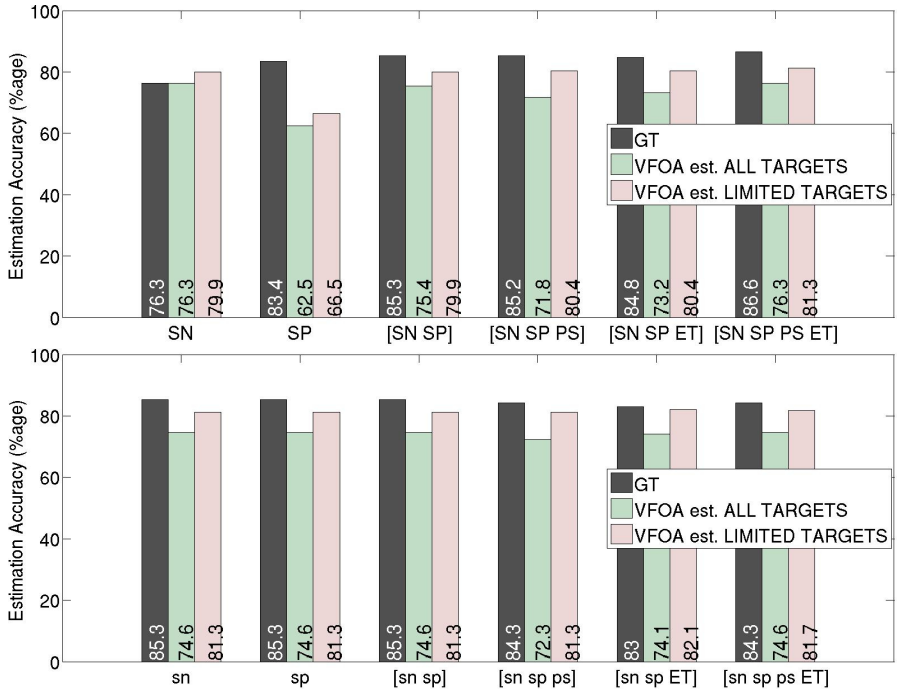


Figure 6.5 – Addressee recognition using the VFOA obtained from tracker results with no context, considering all VFOA targets or only the limited set of addressee targets and relying on un-normalized addressee features (Top) or normalized ones (Bottom).

only monitor looking at Nao, the Partner or Other as targets, thus reducing the possibility of confusion with the non-addressee VFOA. In both cases, no VFOA context was used since it did not help when using tracker pose estimates.

Results are shown in Figure 6.5. First, we can notice that normalization tended to produce the same result whatever the considered addressee feature set. Thus, while it did not improve the best results, it increased the performance that was not good in the un-normalized case (like SpkrL@Ptr, SP). Secondly, we can notice that the limited set condition systematically produced higher results by 7 to 10%. This is an interesting result, which shows that what matters for addressee is mainly to discriminate between VFOA addressee targets. While the VFOA context helped to reduce VFOA recognition ambiguities between such targets and others (by reducing the prior weights of the paintings when they are not the topic of discussion), it is not as drastic as removing them completely from the list of potential targets. In order to see whether the there is a significant difference between the approached which made this improvement, we also ran statistical significance tests to compare the approaches when all targets were monitored to the case of using only the relevant targets (green and pink bars in Figure 6.5). The difference was not significant for all feature combinations, but for the ones which show the highest improvement, { SN, SP, PS } and { SN, SP, ET } (improved the results from 71% to 80% and 73% to 80%) the difference was significant.

Finally, we observe that normalizing features and using only addressee VFOA targets gives the

best results. For instance, combining SP and SN into sp and sn along with ET has an accuracy of 82.1%.

6.4.3 Discussion

In order to compare our work with the previous addressee estimation studies, we discuss our results and methodology in contrast to two previous works [Katzenmaier et al., 2004] and [Jovanović et al., 2006].

In [Jovanović et al., 2006] addressee estimation is studied for human-human interaction in the meeting context. In their setup 4 people seated on a table have meetings during which one person could stand up and give a presentation or explain ideas on the white board. In their experiments different kinds of features like gaze, utterance and conversational context are used. Utterance features include lexical features (the occurance of special words such as we, you, etc.) relying on available speech transcription in addition to dialog act tags (e.g. agreement, question, agreement, etc.). Meeting context is also used which specifies the meeting actions categorized monologue, presentation, discussion and white-board.

In this work the features were not extracted automatically. However, using all features would be challenging for an automatic addresse estimation system considering the available speech recognition systems and the fact that determining the meeting action automatically is not obvious. Regardless of this fact, using all features together they obtain the addressee estimation accuracy of 83.7%. Without using the speech based cues and meeting context, they obtain the highest accutacy of 80%. These results are close to what we obtain in our setting, although their task has one more potential addressee.

In the other study [Katzenmaier et al., 2004], which is closer to our work in terms of the application, addressee estimation is studied for human-human-robot interaction. In their scenario one person -acting as the host- introduces to another person -acting as his/her guest- to the new household robot and gives the robot some commands in order to show its capabilities. The experiments are focused on recognizing whether the host addresses the robot or the guest and the addressee of the guest is not studied. Different acoustic and visual cues are used based on the speech transcription, statistical language models, context free grammars and visual focus of attention.

Since the interaction between the host and the robot is mainly giving commands to the robot and the robot is not acting as a conversation partner. Considering this fact, the features extracted from the speech take advantage of the differences in command and conversational sentences and are in this sense very specific to this scenario. Therefore, we only consider the results obtained from automatically estimated VFOA cues, in which they report the accuracy of 89%. This is higher than what we could obtain in our study using the esimated VFOA (addressee estimation of 82%). However, we should consider that in the Vernissage corpus the cooccurence of addressee and gaze target is less obvious which results from the higher number

of the visual targets and the fact that they attract the attention of the participants and bring additional complications. Furthermore, in this special scenario gaze is a very determining cue since for giving commands to the robots we normally tend to look at them (and use the word robot to call them).

6.5 Conclusion

In this chapter we provided a method for addressee estimation in our HRI setting. Our method is based on the gaze cues from the speaker and other contextual cues like the other participants' gaze, question difficulty and the previous speaker. We performed experiments under different conditions for three separate purposes. First, we wanted to see how effective could this method estimate addressee with noisy VFOA estimates extracted automatically and from inaccurate tracker head poses. Second, we wanted to study whether adding contextual cues at VFOA estimation or addressee classification steps could be useful for improving the estimation. Finally, we wanted to check what kind of VFOA output would be more beneficial for estimating the addressee. Specifically comparing the cases where the robot tracks all VFOA targets compared to the only ones which are potential addressees (the robot and participants).

We have reconfirmed in our setting that gaze cues from the speaker is the most important feature for addressee estimation. Our experiments showed that the tracker results are quite competitive, and the best results have an accuracy of 80%. The VFOA context did not improve the tracker results, as it did for the VICON results. The addressee context (gaze features from the fellow participant, short-term and long-term dialog-context features) helped, though only slightly. Our experiments show that it is better if the VFOA recognition module only monitors whether a person looks at potential addressee targets (the robot, people) rather than considering all objects of interest.

7 Conclusion

7.1 Conclusion

This thesis was conducted in the context of the Humanoids with Auditory and Visual Abilities In Populated Spaces (HUMAVIPS) project which aimed at providing a robot which is capable of performing natural interaction with a group of people. The humanoid robot Nao could use its input audio and video channels to get an understanding of its surrounding and show appropriate behavior. Different tasks were addressed in this project including human perception and behavior understanding, dialog management and robot localization.

Within this context, the goal of this thesis was to study the recognition of people's visual focus of attention (VFOA) from video and estimation of the speakers' addressee. We assumed that we could only rely on the video and audio captured by the robot. Moreover, images captured by the robot were not of high resolution and using eye gaze information was not accessible to apply eye gaze tracking methods. Moreover, people could freely move without constraints with respect to the robot, making the problem quite challenging and limiting the accuracy of previous approaches.

Therefore, from the visual data we relied only on head poses and provided methods for head pose-to-gaze mapping and recognizing the VFOA. Moreover, we exploited the robot's conversational context to improve the recognition. Estimated VFOA was then used in addition to the other contextual cues to detect the addressee of the speaker's utterance. The contributions of this thesis are summarized as follows:

VFOA and addressee database

In collaboration with our partners in the Bielefeld University and using their recording infrastructure, we built a publicly available dataset capturing interactions between the robot Nao and two participants. We designed the scenario so as to ensure interesting and natural behaviors and interaction patterns for human robot interaction studies. Data was recorded using different sensors: Nao camera and microphones, Vicon motion capturing system, close

talk microphones for participants, and additional HD cameras. We organized and collected different sets of annotations and specific to this thesis interest, VFOA and addressee annotations. This dataset is an important contribution in this project given the absence of realistic and unconstrained HRI datasets for studying non-verbal behaviors recognition methods.

VFOA recognition from head pose:

To address this problem, we proposed algorithms to leverage on models described in the literature for understanding human's body, head and gaze dynamics involved in shifting gaze at different directions, and derive head pose-to-gaze mapping models that can be used within an HMM framework to recognize looking in different visual targets directions. We provided three different head pose-gaze mapping models and implemented them in our VFOA recognition system. In the first method we added a dynamic head pose reference that played the role of the person's body orientation. In the second one we implemented the midline effect which was introduced in human behavior studies. Finally in the third proposition we tried to model gaze shift and consider the effect of the previous gaze direction. Our experiments on three datasets, showed that our dynamic models always provided higher recognition accuracy compared to the previous static model and this improvement is more evident on the Vernissage data where people have more freedom to move. The midline effect also improved the results, however, the improvement was not very high since this effect does not happen very frequently in the data. It is important to note that the models are general, and can be applied to any system able to capture head pose, as is the case for instance for the widely used Kinect sensor.

Leveraging robot conversational context for VFOA recognition:

As a second direction for improving the VFOA recognition and removing some of the ambiguities introduced by relying only on the head poses, we explored using contextual data. In contrast to the previous works which relied on group communication context, we extracted our contextual cues from the robot's conversational state. This has the benefit of having direct access to the robot's system state instead of relying on its potentially noisy scene analysis information. In this regard we defined three types of contextual cues which affect the participant's visual attention and integrated them into our VFOA recognition system. Experiments on the Vernissage data showed that the proposed method was very effective. In particular, when the robot is making reference to scene object, and when when more accurate head pose measurements are used.

Addressee estimation:

In this thesis we studied this problem in the context of our Vernissage scenario and more specifically during quiz sessions between the robot and the participants. To address this problem, in addition to the VFOA estimations obtained from the previous parts of the thesis,

we investigated the use of contextual cues at different levels for improving the recognition: at the VFOA recognition level (to improve VFOA and indirectly get better addressee estimates), and at the addressee estimation level. For the latter case, the other participant's gaze and short and long term dialog context features were studied. Our experiments confirmed the known conclusion that the speaker's gaze is the most important cue for addressee estimation. Overall we showed that the different contextual cues as proposed, improved the results, but the gain was relatively small. Interestingly, however, we showed that for the addressee task, it is better if the VFOA recognition module only monitors the potential addressee targets (robot, people) rather than considering all objects of interest.

7.2 Limitations and perspectives

There are several limitations to our work. Below we discuss several of them and how they could be addressed by proposing different research directions for VFOA recognition and addressee estimation.

Gaze shift indicator for VFOA recognition: beyond frame level VFOA recognition

As humans, even when we have partial information about visual targets that other people might be looking at, we are able to recognize when they have gaze shifts from one target to the other. We can also estimate roughly the position of the destination target with respect to the initial one. Obtaining an indicator for gaze shift and a measure for its direction and magnitude could thus be used for automatic VFOA recognition as well. Having a gaze shift indicator allows us to decompose the sequence of head poses into segments with constant VFOA. HMM would then work at this segmental level, and the state dynamics could be made more informative (currently only smooths the output). Gaze shift models could be refined according to the new requirements.

Temporal modeling of robot conversational context

In this thesis we assumed that the robot state contextual cues used to improve VFOA recognition have the same effect on the whole speech segment where they occur. However, for some of these cues, studies on human-human and human-robot interaction suggest that it would be helpful to consider this effect to be time dependent. For instance, when pointing gesture happens, it triggers looking at the pointed-at object immediately afterwards and not necessarily during the full utterance. As the second example, when someone speaks, people look more at him at the beginning and end of his utterance. Incorporating and modeling this timing effect could then give us a more accurate function for the effect of conversational cues on VFOA. This problem is closely related to dialog and the gesture synthesis design and obtaining such timing information, and studying their impact on people behavior could benefit from advances in this field.

Addressee estimation

For addressee estimation we did not use any kind of information from the speakers speech. However, it is known that people indeed speak very differently to the conversational agents as compared to the other humans. This difference could be revealed by the analysis of their prosodic cues. For instance people usually talk louder to the agents and try to pronounce words more clearly as their expectation of the agent's speech recognition capability is not as high as humans. Specially, in our quiz scenario, after they discuss the answer among themselves they announce it louder to the robot and use the words given in the possible list of answers to give the final answer as clear as possible. Therefore, using additional cues from the user's speech could be very useful to estimate the addressee more accurately.

Implementation level

Models were mainly tested on presented dataset where the robot was controlled by WOz approach. However, experiments using the real system is important and would definitely bring new challenges. For instance, the current VFOA module implemented on Nao is more limited by the accuracy of the head pose, and of the 3D position of people in the room. Using higher quality sensors or other sensors (RGB-D) could help resolving these problems.

Use of context priors

Also, related to the contextual priors, there are two main limitations. First of all, they could incorporate as well communication cues from participants in addition to robot state, which has not been addressed in this thesis. Second, using contextual priors as proposed, assumes that the interaction progresses smoothly and following the usual expectations. If this is not the case, they may lead to "hallucination", meaning that the robot may think that people are looking at some targets. Monitoring people's engagement and interest is thus important to know whether such priors can be reliably exploited. This means that a higher interpretation of different tasks is needed.

Data collection

The collection of Vernissage corpus has been very useful for studying perception tasks in human robot interaction. However, there are limitations in this corpus which could be addressed in the future datasets. Here we provide two suggestions to consider. First, in this corpus we only had two people interacting with the robot. However, in many applications the robot should interact with a variable number of users, from one to 3 or more. There are many research questions valuable to study in those applications as well. Therefore, it would be useful to provide similar datasets capturing the interaction between a robot and multiple users in different group sizes. Second, it would be interesting to use depth sensors In addition to the cameras. This would be useful for extracting additional information about the participants

body pose. For instance, shoulder orientation could be extracted and used in the proposed methods for VFOA recognition. Given the current technological advances, these depth sensors are widely used in different applications and it is reasonable to assume they would be available on the future robots.

Bibliography

[Alameda-Pineda et al., 2013] Alameda-Pineda, X., Sanchez-Riera, J., Wienke, J., Franc, V., Cech, J., Kulkarni, K., Deleforge, A., and Horaud, R. P. (2013). Ravel: An annotated corpus for training robots with audiovisual abilities. *Journal on Multimodal User Interfaces*, 7(1-2):79–91.

[Argyle and Cook, 1976] Argyle, M. and Cook, M. (1976). *Gaze and Mutual Gaze*. Cambridge University Press.

[Argyle and Dean, 1965] Argyle, M. and Dean, J. (1965). Eye-Contact, Distance and Affiliation. *Sociometry*, 28(3):289–304.

[Arnaud et al., 2008] Arnaud, E., Christensen, H., Lu, Y.-C., Barker, J., Khalidov, V., Hansard, M., Holveck, B., Mathieu, H., Narasimha, R., Taillant, E., Forbes, F., and Horaud, R. (2008). The cava corpus: Synchronised stereoscopic and binaural datasets with head movements. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 109–116, New York, NY, USA. ACM.

[Ba and Odobez, 2006] Ba, S. and Odobez, J. (2006). A study on visual focus of attention recognition from head pose in a meeting room. In *Proc. Workshop on Machine Learning for Multimodal Interaction (MLMI)*, Washington DC.

[Ba and Odobez, 2008] Ba, S. and Odobez, J.-M. (2008). Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*.

[Ba and Odobez, 2011] Ba, S. and Odobez, J.-M. (2011). Multi-person visual focus of attention from head pose and meeting contextual cues. *IEEE PAMI*, 33(1):101–116.

[Ba, 2007] Ba, S. O. (2007). *Joint head tracking and pose estimation for visual focus of attention recognition*. PhD thesis, Ecole Polytechnique.

[Ba and Odobez, 2005] Ba, S. O. and Odobez, J.-M. (2005). Evaluation of multiple cue head pose estimation algorithms in natural environments. In *Multimedia and Expo, 2005. ICME 2005.*, pages 1330–1333.

[Ba and Odobez, 2009] Ba, S. O. and Odobez, J.-M. (2009). Recognizing visual focus of attention from head pose in natural meetings. *Trans. Sys. Man Cyber. Part B*, 39:16–33.

[Babcock and Pelz, 2004] Babcock, J. S. and Pelz, J. B. (2004). Building a lightweight eyetracking headgear. In *Proceedings of the 2004 symposium on Eye tracking research & applications*, ETRA '04, pages 109–114. ACM.

[Baluja and Pomerleau, 1994] Baluja, S. and Pomerleau, D. (1994). Non-intrusive gaze tracking using artificial neural networks. Technical report, Pittsburgh, PA, USA.

[Bennewitz et al., 2007] Bennewitz, M., Faber, F., Joho, D., and Behnke, S. (2007). Fritz – a humanoid communication robot. In *Proc. of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 1072–1077.

[bits.blogs.nytimes.com, 2010] bits.blogs.nytimes.com (2010). The robots among us, http://bits.blogs.nytimes.com/2010/03/29/the-robots-among-us/?_php=true&_type=blogs&_r=0.

[Bohus and Horvitz, 2009a] Bohus, D. and Horvitz, E. (2009a). Dialog in the open world: platform and applications. In *Proceedings of the 11th International Conference on Multimodal Interfaces*, ICMI '09, pages 31–38.

[Bohus and Horvitz, 2009b] Bohus, D. and Horvitz, E. (2009b). Learning to predict engagement with a spoken dialog system in open-world settings. In *Proceedings of the SIGDIAL 2009 Conference: The 10th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL '09, pages 244–252.

[Bohus and Horvitz, 2010] Bohus, D. and Horvitz, E. (2010). On the challenges and opportunities of physically situated dialog. In *AAAI Fall Symposium: Dialog with Robots*, volume FS-10-05 of *AAAI Technical Report*. AAAI.

[Bohus and Horvitz, 2011] Bohus, D. and Horvitz, E. (2011). Multiparty turn taking in situated dialog: Study, lessons, and directions. In *Proceedings of the SIGDIAL 2011 Conference*, SIGDIAL '11, pages 98–109, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Colburn et al., 2000] Colburn, R. A., Cohen, M. F., and Drucker, S. M. (2000). The role of eye gaze in avatar mediated conversational interfaces. Technical report.

[Cook and mith, 1975] Cook, M. and mith, J. M. C. (1975). The role of gaze in impression formation. *British Journal of Social and Clinical Psychology*, 14(1):19–25.

[Cooper, 1974] Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, 6(1):84 – 107.

[Dahlbäck et al., 1993] Dahlbäck, N., Jönsson, A., and Ahrenberg, L. (1993). Wizard of oz studies: Why and how. In *Proceedings of the 1st International Conference on Intelligent User Interfaces*, IUI '93, pages 193–200, New York, NY, USA. ACM.

[Duffner and Odobez, 2013] Duffner, S. and Odobez, J. (2013). Track creation and deletion framework for long-term online multiface tracking. *Image Processing, IEEE Transactions on*, 22(1):272–285.

[Duncan, 1972] Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292.

[Durkalski et al., 2003] Durkalski, V. L., Palesch, Y. Y., Lipsitz, S. R., and Rust, P. F. (2003). Analysis of clustered matched-pair data. *Statistics in Medicine*, 22(15):2417–2428.

[engineering.curiouscatblog.net, 2008] engineering.curiouscatblog.net (2008). Robot finds lost shoppers and provides directions, http://engineering.curiouscatblog.net/2008/01/27/robot-finds-lost-shoppers-and-provides-directions.

[Fojtu et al., 2012] Fojtu, S., Havlena, M., and Pajdla, T. (2012). Nao robot localization and navigation using fusion of odometry and visual sensor data. In *Intelligent Robotics and Applications*, volume 7507, pages 427–438.

[Fong et al., 2003] Fong, T. W., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and Autonomous Systems*.

[Foster et al., 2012] Foster, M. E., Gaschler, A., Giuliani, M., Isard, A., Pateraki, M., and Petrick, R. P. (2012). Two people walk into a bar: dynamic multi-party social interaction with a robot agent. In *Proceedings of the 14th International Conference on Multimodal Interfaces*, ICMI '12. ACM.

[Freedman and Sparks, 1997] Freedman, E. G. and Sparks, D. L. (1997). Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77(5):2328–2348.

[Funes Mora and Odobez, 2012] Funes Mora, K. and Odobez, J. (2012). Gaze estimation from multimodal kinect data. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 25–30.

[Gaschler et al., 2012] Gaschler, A., Huth, K., Giuliani, M., Kessler, I., de Ruiter, J., and Knoll, A. (2012). Modelling state of interaction from head poses for social Human-Robot Interaction. In *Proceedings of the Gaze in Human-Robot Interaction Workshop held at the 7th ACM/IEEE International Conference on Human-Robot Interaction*.

[Goodwin, 1980] Goodwin, C. (1980). Restarts, pauses, and the achievement of a state of mutual gaze at turn-beginning. *Sociological Inquiry*, 50(3-4):272–302.

[Gorga and Otsuka, 2010] Gorga, S. and Otsuka, K. (2010). Conversation scene analysis based on dynamic bayesian network and image-based gaze detection. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '10, page 54.

[Gourier et al., 2004] Gourier, N., Hall, D., and Crowley, J. L. (2004). Estimating Face Orientation from Robust Detection of Salient Facial Features. In *Proceedings of Pointing 2004, ICPR, International Workshop on Visual Observation of Deictic Gestures*.

[Gu and Badler, 2006] Gu, E. and Badler, N. I. (2006). Visual attention and eye gaze during multiparty conversations with distractions. In Gratch, J., Young, M., Aylett, R., Ballin, D., and Olivier, P., editors, *IVA*, volume 4133 of *Lecture Notes in Computer Science*, pages 193–204. Springer.

[Guestrin and Eizenman, 2006] Guestrin, E. and Eizenman, E. (2006). General theory of remote gaze estimation using the pupil center and corneal reflections. *Biomedical Engineering, IEEE Transactions on*, 53(6):1124–1133.

[Hanes and McCollum, 2006] Hanes, D. A. and McCollum, G. (2006). Variables contributing to the coordination of rapid eye/head gaze shifts. *Biol. Cybern.*, 94:300–324.

[Hansen and Ji, 2010] Hansen, D. W. and Ji, Q. (2010). In the eye of the beholder: A survey of models for eyes and gaze. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(3):478–500.

[Häring et al., 2012] Häring, M., Eichberg, J., and André, E. (2012). Studies on grounding with gaze and pointing gestures in human-robot-interaction. In *Proceedings of the 4th International Conference on Social Robotics*, ICSR'12, pages 378–387, Berlin, Heidelberg. Springer-Verlag.

[Hayhoe and Ballard, 2005] Hayhoe, M. and Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences*, 9(4):188–194.

[Huang et al., 2011] Huang, H., Baba, N., and Nakano, Y. (2011). Making virtual conversational agent aware of the addressee of users' utterances in multi-user conversation using nonverbal information. In *Proceedings of the 13th ACM international conference on Multimodal interaction*, ICMI '11.

[Hung et al., 2008] Hung, H., Jayagopi, D. B., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008). Investigating automatic dominance estimation in groups from visual attention and speaking activity. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 233–236, New York, NY, USA. ACM.

[Jayagopi et al., 2013] Jayagopi, D., Sheikhi, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nguyen, L., Wrede, B., and Gatica-Perez, D. (2013). The vernissage corpus: A conversational human-robot interaction dataset. In *8th ACM/IEEE International Conference on Human-Robot Interaction*.

[Jayagopi et al., 2008] Jayagopi, D. B., Ba, S., Odobez, J.-M., and Gatica-Perez, D. (2008). Predicting two facets of social verticality in meetings from five-minute time slices and nonverbal cues. In *Proceedings of the 10th International Conference on Multimodal Interfaces*, ICMI '08, pages 45–52, New York, NY, USA. ACM.

[Jayagopi et al., 2012] Jayagopi, D. B., Sheikhi, S., Klotz, D., Wienke, J., Odobez, J.-M., Wrede, S., Khalidov, V., Nguyen, L. S., Wrede, B., and Gatica-Perez, D. (2012). The vernissage corpus: A multimodal human-robot-interaction dataset. Idiap-RR Idiap-RR-33-2012, Idiap.

[Jovanović, 2007] Jovanović, N. (2007). *To whom it may concern : adressee identification in face-to-face meetings*. PhD thesis, Enschede.

[Jovanović and op den Akker, 2004] Jovanović, N. and op den Akker, H. J. A. (2004). Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue, Boston, MA, USA*, pages 89–92, Pennsylvania, USA. Association for Computational Linguistics.

[Jovanović et al., 2006] Jovanović, N., op den Akker, H. J. A., and Nijholt, A. (2006). Addressee identification in face-to-face meetings. In McCarthy, D. and Wintner, S., editors, *Proceedings of 11th Conference of the European Chapter of the ACL (EACL), Trento, Italy*, pages 169–176, Pennsylvania, USA. Association for Computational Linguistics.

[Katzenmaier et al., 2004] Katzenmaier, M., Stiefelhagen, R., and Schultz, T. (2004). Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces*, ICMI '04, pages 144–151, New York, NY, USA. ACM.

[Kendon, 1967] Kendon, A. (1967). Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)*, 26(1):22–63.

[Khalidov and Odobez, 2013] Khalidov, V. and Odobez, J. (2013). Real-time multiple head tracking using texture and colour cues. In *Internal report, Idiap*.

[koreaittimes.com, 2010] koreaittimes.com (2010). English teaching robot comes with a cost, http://www.koreaittimes.com/story/8216/english-teaching-robot-comes-cost.

[Langton et al., 2000] Langton, S. R., Watt, R. J., and Bruce, I. (2000). Do the eyes have it? cues to the direction of social attention. *Trends Cogn Sci*, 4(2):50–59.

[Lemon et al., 2002] Lemon, O., Gruenstein, A., Gruenstein, E., Battle, A., and Peters, S. (2002). Multi-tasking and collaborative activities in dialogue systems. In *In Proceedings of 3rd SIGdial Workshop on Discourse and Dialogue, 113*, pages 113–124.

[Li et al., 2005] Li, D., Winfield, D., and Parkhurst, D. J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based approaches. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops - Volume 03*, CVPR '05, pages 79–, Washington, DC, USA. IEEE Computer Society.

[Lohse et al., 2009] Lohse, M., Hanheide, M., Rohlfing, K. J., and Sagerer, G. (2009). Systemic interaction analysis (sina) in hri. In *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, pages 93–100, New York, NY, USA. ACM.

[Lu et al., 2011a] Lu, F., Okabe, T., Sugano, Y., and Sato, Y. (2011a). A head pose-free approach for appearance-based gaze estimation. *Proceedings of the 22nd British Machine Vision Conference (BMVC 2011)*.

[Lu et al., 2011b] Lu, F., Sugano, Y., Okabe, T., and Sato, Y. (2011b). Inferring human gaze from appearance via adaptive linear regression. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 153–160.

[Michalowski, 2006] Michalowski, M. P. (2006). A spatial model of engagement for a social robot. In *In Proceedings of the 9th International Workshop on Advanced Motion Control (AMC 2006)*.

[Mohammad et al., 2008] Mohammad, Y., Xu, Y., Matsumura, K., and Nishida, T. (2008). The h3r explanation corpus human-human and base human-robot interaction dataset. In *Intelligent Sensors, Sensor Networks and Information Processing, 2008. ISSNIP 2008.*, pages 201–206.

[Morency, 2009] Morency, L.-P. (2009). Co-occurrence graphs: contextual representation for head gesture recognition during multi-party interactions. In *Proceedings of the Workshop on Use of Context in Vision Processing*.

[Morency and Darrell, 2008] Morency, L.-P. and Darrell, T. (2008). Conditional sequence model for context-based recognition of gaze aversion. In *Proceedings of the 4th international conference on Machine learning for multimodal interaction*, MLMI'07, pages 11–23, Berlin, Heidelberg. Springer-Verlag.

[Morency et al., 2005] Morency, L.-P., Sidner, C. L., Lee, C., and Darrell, T. (2005). Contextual recognition of head gestures. In *Proceedings of the 7th ACM international conference on Multimodal interaction*, ICMI '05, pages 18–24.

[Morimoto and Mimica, 2005] Morimoto, C. H. and Mimica, M. R. (2005). Eye gaze tracking techniques for interactive applications. *Computer Vision and Image Understanding*, 98(1):4 – 24. Special Issue on Eye Detection and Tracking.

[Nakano et al., 2003] Nakano, Y. I., Reinstein, G., Stocky, T., and Cassell, J. (2003). Towards a model of face-to-face grounding. In *Proc. of the Annual Meeting on Association for Computational Linguistics*, pages 553–561.

[Nakazawa and Nitschke, 2012] Nakazawa, A. and Nitschke, C. (2012). Point of gaze estimation through corneal surface reflection in an active illumination environment. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part II*, ECCV'12, pages 159–172, Berlin, Heidelberg. Springer-Verlag.

[Novick et al., 1996] Novick, D., Hansen, B., and Ward, K. (1996). Coordinating turn-taking with gaze. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, volume 3, pages 1888–1891 vol.3.

[Ohno and Mukawa, 2004] Ohno, T. and Mukawa, N. (2004). A free-head, simple calibration, gaze tracking system that enables gaze-based interaction. In *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications*, ETRA '04, pages 115–122, New York, NY, USA. ACM.

[op den and Traum, 2009] op den, R. A. and Traum, D. (2009). A comparison of addressee detection methods for multiparty conversations. In *DiaHolmia : 2009 Workshop on the Semantics and Pragmatics of Dialogue*, pages 99–106, Stockholm, Sweden. KTH Stockholm.

[Otsuka et al., 2005] Otsuka, K., Takemae, Y., and Yamato, J. (2005). A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In *Proceedings of the 7th international conference on Multimodal interfaces*, ICMI '05, pages 191–198. ACM.

[Pelz et al., 2000] Pelz, J. B., Canosa, R. L., Kucharczyk, D., Babcock, J. S., Silver, A., and Konno, D. (2000). Portable eyetracking: A study of natural eye movements. In *Electronic Imaging*, pages 566–582. International Society for Optics and Photonics.

[Pitsch et al., 2011] Pitsch, K., Wrede, S., Seele, J.-C., and Süssenbach, L. (2011). Attitude of german museum visitors towards an interactive art guide robot. In *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, pages 227–228, New York, NY, USA. ACM.

[Pourtois et al., 2004] Pourtois, G., Sander, D., Andres, M., Grandjean, D., Reveret, L., Olivier, E., and Vuilleurmier, P. (2004). Dissociable roles of the human somatosensory and superior temporal cortices for processing social face signals. *European Journal of Neuroscience*, 20(12):3507–3515.

[Qvarfordt and Zhai, 2005] Qvarfordt, P. and Zhai, S. (2005). Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 221–230. ACM.

[Rehg, 2013] Rehg, J. M. (2013). Behavior imaging and the study of autism. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 1–2, New York, NY, USA. ACM.

[Rehm and Andre, 2005] Rehm, M. and Andre, E. (2005). Where do they look? gaze behaviors of multiple users interacting with an embodied conversational agent. In *In: Proceedings of International Conference on Intelligent Agents (IVA-05), LNCS (LNAI)*, pages 241–252. Springer.

[Ricci and Odobez, 2009] Ricci, E. and Odobez, J.-M. (2009). Learning large margin likelihoods for realtime head pose tracking. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 2593–2596. IEEE.

[Sarma and Palmer, 2004] Sarma, A. and Palmer, D. D. (2004). Context-based speech recognition error detection and correction. In *Proceedings of HLT-NAACL 2004: Short Papers*,

HLT-NAACL-Short '04, pages 85–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Schegloff, 1968] Schegloff, E. A. (1968). Sequencing in conversational openings1. *American Anthropologist*, 70(6):1075–1095.

[Sheikhi et al., 2013a] Sheikhi, S., Babu Jayagopi, D., Khalidov, V., and Odobez, J.-M. (2013a). Context aware addressee estimation for human robot interaction. In *Proceedings of the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction: Gaze in Multimodal Interaction*, GazeIn '13, pages 1–6, New York, NY, USA. ACM.

[Sheikhi et al., 2013b] Sheikhi, S., Khalidov, V., Klotz, D., Wrede, B., and Odobez, J.-M. (2013b). Leveraging the robot dialog state for visual focus of attention recognition. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 107–110, New York, NY, USA. ACM.

[Sheikhi et al., 2012] Sheikhi, S., Khalidov, V., and Odobez, J.-M. (2012). Recognizing the visual focus of attention for human robot interaction. In *Human Behavior Understanding*, volume 7559 of *Lecture Notes in Computer Science*, pages 99–112. Springer Berlin Heidelberg.

[Sheikhi and Odobez, 2012] Sheikhi, S. and Odobez, J.-M. (2012). Investigating the midline effect for visual focus of attention recognition. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '12, pages 221–224, New York, NY, USA. ACM.

[Sheikhi and Odobez, 2014] Sheikhi, S. and Odobez, J.-M. (2014). Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. In *preprint*.

[Sidner et al., 2004] Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. (2004). Where to look: A study of human-robot engagement. In *Proceedings of the 9th International Conference on Intelligent User Interfaces*, IUI '04, pages 78–84, New York, NY, USA. ACM.

[Siracusa et al., 2003] Siracusa, M., Morency, L.-P., Wilson, K., Fisher, J., and Darrell, T. (2003). A multi-modal approach for determining speaker location and focus. In *Proceedings of the 5th international conference on Multimodal interfaces*, ICMI '03, pages 77–80, New York, NY, USA. ACM.

[Stiefelhagen, 2002] Stiefelhagen, R. (2002). Tracking focus of attention in meetings. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, ICMI '02, pages 273–. IEEE Computer Society.

[Stiefelhagen et al., 2002] Stiefelhagen, R., Yang, J., and Waibel, A. (2002). Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Trans. on Neural Networks*, 13(4):928–938.

[Sugano et al., 2008] Sugano, Y., Matsushita, Y., Sato, Y., and Koike, H. (2008). An incremental learning method for unconstrained gaze estimation. In *Proceedings of the 10th European*

Conference on Computer Vision: Part III, ECCV '08, pages 656–667, Berlin, Heidelberg. Springer-Verlag.

[Takemae and Ozawa, 2006] Takemae, Y. and Ozawa, S. (2006). Automatic addressee identification based on participants' head orientation and utterances for multiparty conversations. In *ICME*, pages 1285–1288. IEEE.

[technocrazed.com, 2013] technocrazed.com (2013). Kompai: A wonderful robot doctor for monitoring health of older people, http://www.technocrazed.com/kompai-a-wonderful-robot-doctor-for-monitoring-health-of-older-people-photo-gallery.

[Torralba et al., 2003] Torralba, A., Murphy, K., Freeman, W., and Rubin, M. (2003). Context-based vision system for place and object recognition. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 273–280 vol.1.

[Traum and Rickel, 2002] Traum, D. and Rickel, J. (2002). Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 2*, AAMAS '02, pages 766–773, New York, NY, USA. ACM.

[van Turnhout et al., 2005] van Turnhout, K., Terken, J., Bakx, I., and Eggen, B. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the 7th international conference on Multimodal interfaces*, ICMI '05, pages 175–182, New York, NY, USA. ACM.

[Van Turnhout et al., 2005] Van Turnhout et al., K. (2005). Identifying the intended addressee in mixed human-human and human-computer interaction from non-verbal features. In *Proceedings of the 7th ACM international conference on Multimodal interaction*, ICMI '05.

[Vertegaal et al., 2001] Vertegaal, R., Slagter, R., van der Veer, G., and Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '01, pages 301–308, New York, NY, USA. ACM.

[Vertegaal and van der Veer, 2000] Vertegaal, R. and van der Veer, G. (2000). Effects of gaze on multiparty mediated communication. In *In Proceedings of Graphics Interface*, pages 95–102.

[Voit and Stiefelhagen, 2008] Voit, M. and Stiefelhagen, R. (2008). Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, ICMI '08, pages 173–180. ACM.

[Wang and Jin, 2001] Wang, X. and Jin, J. (2001). A quantitative analysis for decomposing visual signal of the gaze displacement. In *Proceedings of the Pan-Sydney Area Workshop on Visual Information Processing - Volume 11*, VIP '01, pages 153–159, Darlinghurst, Australia. Australian Computer Society, Inc.

[Williams et al., 2006] Williams, O., Blake, A., and Cipolla, R. (2006). Sparse and semi-supervised visual mapping with the s3gp. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1*, CVPR '06, pages 230–237, Washington, DC, USA. IEEE Computer Society.

[Yücel and Salah, 2009] Yücel, Z. and Salah, A. (2009). Resolution of focus of attention using gaze direction estimation and saliency computation. In *Int. Conf. on Affective Computing and Intelligent Interfaces*.

[Yuille et al., 1992] Yuille, A. L., Hallinan, P. W., and Cohen, D. S. (1992). Feature extraction from faces using deformable templates. *Int. J. Comput. Vision*, 8(2):99–111.

SAMIRA SHEIKHI

CONTACT INFORMATION

Home Page: http://people.epfl.ch/samira.sheikhi
Email: samira.sheikhi(at)epfl.ch
Cell Phone: +41-775 010 988

RESEARCH INTERESTS

- Human Robot Interaction
- Human Behavior Understanding
- Machine Learning
- Computer Vision

EDUCATION

- **Ph.D. Student,** Department of Electrical Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Oct 2010 - present
- **M.Sc. in Computer Science**, Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran. 2007 - 2010
  Thesis: *"Learning of Shape Classes"*, Advisors: Prof. M. Shahshahani and Dr. M. R. Razvan.
- **B.Sc. in Computer Science**, Department of Mathematical Sciences, Sharif University of Technology, Tehran, Iran. 2003 - 2007

PUBLICATIONS

- **Context Aware Addressee Estimation for Human Robot Interaction**, *S. Sheikhi, D.B. Jayagopy, V. Khalidov and J-M. Odobez*, in: the 6th Workshop on Eye Gaze in Intelligent Human Machine Interaction, ICMI, 2013.
- **Leveraging the Robot Dialog State for Visual Focus of Attention Recognition**, *S. Sheikhi, V. Khalidov, D. Klotz, B. Wrede and J-M. Odobez*, in: Int Conf. on Multimodal Interaction (ICMI), 2013.
- **The Vernissage Corpus: A Conversational Human-Robot-Intercation Dataset**, *D.B. Jayagopy, S. Sheikhi, D. Klotz, J. Wienke, J-M. Odobez, S. Wrede, V. Khalidov, L. Nguyen, B. Wrede, D. Gatica-Perez*, in Proceedings of the 8th ACM/IEEE international conference on Human-Robot interaction, 2013.
- **Investigating the Midline Effect for Visual Focus of Attention Recognition**, *Samira Sheikhi* and *Jean-Marc Odobez*, in: Int Conf. on Multimodal Interaction (ICMI), 2012.
- **Recognizing the Visual Focus of Attention for Human Robot Interaction**, *S. Sheikhi, V. Khalidov* and *J-M. Odobez*, in: IEEE International Conference on Intelligent Robots and Systems (IROS) - Human Behavior Understanding Workshop (IROS-HBU), 2012.
- **The Vernissage Corpus: A Multimodal Human-Robot-Interaction Dataset**, *D.B. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J-M. Odobez, S. Wrede, V. Khalidov, L. S. Nguyen, B. Wrede* and *D. Gatica-Perez*, Idiap-RR-33-2012.
- **Learning of Shape Classes**, *Samira Sheikhi*, Masters Thesis in Computer Science, Sharif University of Technology, 2010.
- **Evaluation of Background Subtraction Methods**., Sorayya Panahi, Samira Sheikhi, Shahrzad Hadadan and Niloofar Gheissari, IEEE-DICTA 2008: 357-364.

SAMIRA SHEIKHI

HONORS AND AWARDS

- Ranked 2nd in Cumulative GPA, M.Sc. of Computer Science, Class of 2007, SUT.
- Ranked 4th in Cumulative GPA, B.Sc. of Computer Science, Class of 2003, SUT.
- Bronze Medalist in National Olympiad of Mathematics, Iran. Aug 2002.
- Studied in the National Organization of Exceptional Talents (NODET), Iran, 1996-2003.

SELECTED RESEARCH EXPERIENCE

- **Research Assistant** at *IDIAP Research Institute*, Switzerland Jun 2010 - present

  I am a research assistant at IDIAP and work on the Humavips project under the supervision of Dr. J-M. Odobez. The purpose of this project is to improve human-robot interaction capabilities especially for applications to social scenarios. The methods we employ are based on audio visual cues for recognizing visual focus of attention and addressee detection.

- **Student Researcher** at *IPM Computer Vision Group*, Iran Oct 2007 - May 2010

  I was involved in some projects at IPM under the supervision of Prof. M. Shahshahani (the group is now moved to MIVLAB at Sharif University of Technology). In this group I had the opportunity to gain a good background in computer vision, pattern recognition and machine learning.

  - **Background Modeling and Subtraction** Our work included the study and implementation of several background subtraction methods. We investigated their accuracy and efficiency in challenging circumstances which led to an evaluation paper in DICTA 2008.
  - **Shape Learning and Clustering** I worked on the problem of learning and clustering of shape classes using their contour information and took it as my M.Sc. thesis topic. I utilized structured descriptions of shapes extracted from skeletons and their graphical representations to solve the problem.

- **Researcher** at *Pars Khodro Co.* Research and Development Center Jan - Jun 2009

  I was involved in two research projects "Pedestrian Detection and Tracking" and "Lane Detection and Tracking" for vehicles. The projects were under the supervision of Dr. M. R. Razvan.

- **Advanced Graph Theory** Course Project Fall 2009

  I did a comparative study on "Spectral Clustering" methods and their application to computer vision under the supervision of Prof. A. Daneshgar for my Advanced Graph Theory course.

TEACHING EXPERIENCES

- Teaching Assistant in "*Combinatorics and Applications*" (Prof. E. S. Mahmoodian) 2008
- Teaching Assistant in "Linear Algebra" (Prof. S. R. Moghaddasi) 2006

110