Tracking and Modeling Focus of Attention in Meetings

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften
von der Fakultät für Informatik
der Universität Karlsruhe (Technische Hochschule)
genehmigte

Dissertation

von

Rainer Stiefelhagen
aus Stuttgart

Tag der mündlichen Prüfung:	5. Juli 2002
Erster Gutachter:	Prof. Dr. Alex Waibel
Zweiter Gutachter:	Prof. Dr. Matthew Turk

Abstract

This thesis addresses the problem of tracking the focus of attention of people. In particular, a system to track the focus of attention of participants in meetings is developed. Obtaining knowledge about a person's focus of attention is an important step towards a better understanding of what people do, how and with what or whom they interact or to what they refer. In meetings, focus of attention can be used to disambiguate the addressees of speech acts, to analyze interaction and for indexing of meeting transcripts. Tracking a user's focus of attention also greatly contributes to the improvement of human-computer interfaces since it can be used to build interfaces and environments that become aware of what the user is paying attention to or with what or whom he is interacting.

The direction in which people look; i.e., their gaze, is closely related to their focus of attention. In this thesis, we estimate a subject's focus of attention based on his or her head orientation. While the direction in which someone looks is determined by head orientation and eye gaze, relevant literature suggests that head orientation alone is a sufficient cue for the detection of someone's direction of attention during social interaction. We present experimental results from a user study and from several recorded meetings that support this hypothesis.

We have developed a Bayesian approach to model at whom or what someone is looking based on his or her head orientation. To estimate head orientations in meetings, the participants' faces are automatically tracked in the view of a panoramic camera and neural networks are used to estimate their head orientations from pre-processed images of their faces. Using this approach, the focus of attention target of subjects could be correctly identified during 73% of the time in a number of evaluation meetings with four participants.

In addition, we have investigated whether a person's focus of attention can be predicted from other cues. Our results show that focus of attention is correlated to who is speaking in a meeting and that it is possible to predict a person's focus of attention based on the information of who is talking or was talking before a given moment. We have trained neural networks to predict at whom a person is looking, based on information about who was speaking. Using this approach we were able to predict who is looking at whom with 63% accuracy on the evaluation meetings using only information about who was speaking. We show that by using both head orientation and speaker information to estimate a person's focus, the accuracy of focus detection can be improved compared to just using one of the modalities for focus estimation.

To demonstrate the generality of our approach, we have built a prototype system to

demonstrate focus-aware interaction with a household robot and other smart appliances in a room using the developed components for focus of attention tracking. In the demonstration environment, a subject could interact with a simulated household robot, a speech-enabled VCR or with other people in the room, and the recipient of the subject's speech was disambiguated based on the user's direction of attention.

Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der automatischen Bestimmung und Verfolgung des Aufmerksamkeitsfokus von Personen in Besprechungen.

Die Bestimmung des Aufmerksamkeitsfokus von Personen ist zum Verständnis und zur automatischen Auswertung von Besprechungsprotokollen sehr wichtig. So kann damit beispielsweise herausgefunden werden, wer zu einem bestimmten Zeitpunkt wen angesprochen hat beziehungsweise wer wem zugehört hat. Die automatische Bestimmung des Aufmerksamkeitsfokus kann desweiteren zur Verbesserung von Mensch-Maschine-Schnittstellen benutzt werden.

Ein wichtiger Hinweis auf die Richtung, in welche eine Person ihre Aufmerksamkeit richtet, ist die Kopfstellung der Person. Daher wurde ein Verfahren zur Bestimmung der Kopfstellungen von Personen entwickelt. Hierzu wurden künstliche neuronale Netze benutzt, welche als Eingaben vorverarbeitete Bilder des Kopfes einer Person erhalten, und als Ausgabe eine Schätzung der Kopfstellung berechnen. Mit den trainierten Netzen wurde auf Bilddaten neuer Personen, also Personen, deren Bilder nicht in der Trainingsmenge enthalten waren, ein mittlerer Fehler von neun bis zehn Grad für die Bestimmung der horizontalen und vertikalen Kopfstellung erreicht.

Desweiteren wird ein probabilistischer Ansatz zur Bestimmung von Aufmerksamkeitszielen vorgestellt. Es wird hierbei ein Bayes'scher Ansatzes verwendet um die A-posteriori Wahrscheinlichkeiten verschiedener Aufmerksamkteitsziele, gegeben beobachteter Kopfstellungen einer Person, zu bestimmen. Die entwickelten Ansätze wurden auf mehren Besprechungen mit vier bis fünf Teilnehmern evaluiert.

Ein weiterer Beitrag dieser Arbeit ist die Untersuchung, inwieweit sich die Blickrichtung der Besprechungsteilnehmer basierend darauf, wer gerade spricht, vorhersagen läßt. Es wurde ein Verfahren entwickelt um mit Hilfe von neuronalen Netzen den Fokus einer Person basierend auf einer kurzen Historie der Sprecherkonstellationen zu schätzen.

Wir zeigen, dass durch Kombination der bildbasierten und der sprecherbasierten Schätzung des Aufmerksamkeitsfokus eine deutliche verbesserte Schätzung erreicht werden kann.

Insgesamt wurde mit dieser Arbeit erstmals ein System vorgestellt um automatisch die Aufmerksamkeit von Personen in einem Besprechungsraum zu verfolgen.

Die entwickelten Ansätze und Methoden können auch zur Bestimmung der Aufmerksamkeit von Personen in anderen Bereichen, insbesondere zur Steuerung von computerisierten, interaktiven Umgebungen, verwendet werden. Dies wird an einer Beispielapplikation gezeigt.

Acknowledgments

This work was conducted at the *Interactive Systems Lab* (ISL) in the Institut für Logik, Komplexität und Deduktionssysteme. I would like to thank Prof. Alex Waibel, director of the ISL, for advising me during my doctoral studies and giving me the opportunity to work in such an inspiring research environment. I would also like to thank Prof. Matthew Turk, my co-advisor, for his thorough comments, which have greatly improved this dissertation.

I am very grateful to Dr. Jie Yang of the Carnegie Mellon University, who in recent years has provided a great deal of support and encouragement. I would like to thank him especially for many fruitful discussions both in person and via telephone.

Moreover, I would like to thank many current and former colleagues at the Interactive Systems Labs, both in Karlsruhe and at Carnegie Mellon, for their collaboration and their wonderful support: Michael Bett, Susi Burger, Eric Carraux, Matthias Denecke, Christian Fügen, Michael Finke, Jürgen Fritsch, Petra Geutner, Ralph Gross, Hermann Hild, Stefan Jäger, Thomas Kemp, Detlef Koll, Victoria MacLaren, Robert Malkin, Stefan Manke, John McDonough, Uwe Meier, Florian Metze, Céline Morel, Jürgen Reichert, Klaus Ries, Ivica Rogina, Thomas Schaaf, Tanja Schulz, Hagen Soltau, Bernhard Suhm, Martin Westphal, Monika Woszczyna and Jie Zhue.

Many thanks to Silke Dannenmaier for her cheerful assistance in all administrative matters. Thanks to Frank Dreilich, Martin Klein and Norbert Berger for having kept our machines and network running.

And thanks to everybody who participated in the numerous data collection sessions!

Also thanks to Susi Burger, Uwe Meier, Klaus Ries, Pankaj Rege and Ashish Sanil for their hospitality during my visits to Pittsburgh. Thanks to Thomas Schaaf and Jürgen Fritsch for taking care of my papyrus whenever I was away.

Special thanks to everybody who helped proofreading this thesis: Jürgen Fritsch, Hermann Hild, Günter Leypoldt, John McDonough, Klaus Ries and Jie Yang.

I also wish to thank my parents Hans-Dieter and Heidemarie and all the rest of my family for their support over the years.

Finally, I would like to thank Ines for her love and understanding.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Interaction with computers has for many years been dominated by the classical WIMP (Windows, Icons, Menus, Pointers) paradigm: users still typically interact with a desktop computer through graphical user interfaces, by pointing, typing and clicking. Moreover, interaction usually happens between one user and one computer at a time, and the user must *intentionally* perform various actions – such as pointing, typing – to accomplish the specific task that he has in mind.

In recent years many researchers have devoted substantial effort to investigating how computers can be used more efficiently to *support* users during various tasks and activities in their everyday lives without requiring them to attentively *control* specific computers or devices.

Black et al., for example, described their efforts to build a "digital office" [Black et al. '98]. Their goal was to remove the barrier between physical and electronic objects and to facilitate the interaction of humans with all kinds of documents in an office. They have thus augmented a physical office with cameras to scan documents on a desk and to capture a person's notes from a whiteboard. Cameras are also used to track a user's gestures and to enable gesture-driven interaction with a computer-supported whiteboard.

Mozer et al. have built an "adaptive house" which automatically adapts to its inhabitants' needs [Mozer '98]. They have equipped a real house with infrared sensors to detect the locations of the inhabitants and have also used various sensors to measure heating, whether doors are open or whether lights are switched on or off. The input from all sensors is then used to learn the inhabitants' preferences and to automatically adjust lighting and heating in the house accordingly. The house for example "learns" at what time the heating has to be at a comfortable level and during which

periods the temperature can be lowered because the inhabitants are usually at work. The house also automatically learns when to switch on the room lights.

Abowd et al. [Abowd et al. '96] have presented an "intelligent classroom" which provides technology and tools to support teachers and students during lectures and also facilitates retrieval of recorded lectures. Their classroom is equipped with electronic whiteboards, pen-based personal interfaces for the students and projectors to display the presenter's notes and related information from web pages. The room also has microphones and video cameras to record the lectures. During a lecture, the presenter's hand written notes, audio and video are automatically captured. In addition, each student is provided with a pen-based PC to record his own personal notes. Afterwards, the video documentation, the presenter's slides and his hand-written notes are automatically made available online as a multimedia document. Students can browse through the document using a special web-interface. They can search for certain topics, watch the corresponding slides and the presenter's handwritten annotations or they can look at the relevant parts in the video of the lecture.

At Microsoft Research, the "Easy Living Project" [Brumitt et al. 2000b] is concerned with the development of architectures and technologies for intelligent environments. These researchers have built a living room which automatically tracks the location of a user and provides the information or service that a user requests through appropriate devices in the proximity of the user. They have also investigated various multimodal interaction techniques to control the room lights, for example [Brumitt et al. 2000a].

In order to make such intelligent and interactive environments respond appropriately to users' needs, it is necessary to equip them with perceptive capabilities to capture as much relevant information about its users and the *context* in which they act as possible. Such information includes: detecting the number and locations of users in the room, their identities, facial expressions, body movements, the users' speech, their gaze direction and their focus of attention.

Obtaining knowledge about a person's *focus of attention* is a major step towards a better understanding of what users do, how and with what or whom they interact or to what they refer.

For instance, in a smart interactive environment, which incorporates appliances that respond to a user's speech, knowing where users look is essential to determine which appliance they address when talking. Even more importantly, such information can be used to detect whether any of the speech-enabled appliances is addressed at all, or whether a subject was talking to another person in the room, in which case none of the appliances need to react. We certainly don't want our VCRs to start recording or the room lights to react whenever we talk to other persons in our living room.

Focus of attention tracking could be especially useful in cars. Here, monitoring where drivers look could, for example, be used to determine whether they are aware of events on the street, whether they checked the rear-view mirrors before over-taking another car, or whether they have recently checked their speed indicator. In cases when a driver is apparently unaware of something important, an intelligent car could then notify the driver about these things.

1.1 Focus of Attention Tracking in Meetings

This thesis focuses on the problem of tracking focus of attention in meetings.

Having meetings is one of the most common activities in business. It is impossible, however, for people to attend all relevant meetings or to retain all the salient points raised in meetings they do attend. Protocols, notes and summaries of meetings are used to develop a corporate memory that overcomes these problems.

Hand recorded notes, however, have many drawbacks. Note-taking is time consuming, requires focus, and thus reduces one's attention to and participation in the ensuing discussions. For this reason notes tend to be fragmentary and partially summarized, leaving one unsure exactly what was resolved, and why. Recalling all details on certain topics of a meeting is therefore impossible from such hand-written notes. In addition, meeting notes tend to be biased by the minute taker's understanding of the meeting. And finally, non-verbal and social cues, which often are essential for a good understanding of the meeting, are often completely missing in such notes.

In order to provide really useful *meeting records*, as many of the modalities that humans use during interaction should be captured and analyzed as possible. These modalities include speech, gestures, emotions and body language.

Furthermore, the context in which the meeting took place has to be provided for a detailed meeting record. To fully understand the dynamics of a meeting, meeting records should provide details such as:

- Who participated in the meeting?
- When did the meeting take place?
- What were the topics of the meeting?
- What was said?
- Who said what and to whom?

- What was the social setting and the relationship of the participants? Was it a business meeting or an informal interaction between friends?

To analyze meetings with regard to these questions, a number of technologies are needed. These include speech recognition, dialogue processing, text summarization, person tracking and identification, and gesture recognition.

Visual communication cues, such as gesturing, looking at another person or monitoring his facial expressions, play an important role during face-to-face communication [Argyle '69, Goodwin '81]. Therefore, to fully understand an ongoing conversation, it is necessary to capture and analyze these visual cues in addition to spoken content. Knowing who is looking at whom at a certain moment in a meeting, can help us, for example, to understand to whom the participants are paying attention and whom a speaker does address.

At the Interactive Systems Lab at the Universität Karlsruhe and at Carnegie Mellon University in Pittsburgh, we are developing an **intelligent meeting room** to automatically capture and transcribe meetings.

The goal of this project is to develop a meeting room that eventually will identify when a meeting begins, start capturing and analyzing the meeting and create a meeting record with information about who was in the meeting, what was said, etc. Users should be able to use the meeting room as if it were a normal meeting room, and the technology used to capture meetings and extract information should be as non-intrusive to the users as possible.

To capture the necessary audio and video streams we have equipped our meeting room with a number of lapel microphones, pan-tilt-zoom cameras and an omni-directional camera on the meeting table. Figure 1.1 shows an image that was taken during a meeting in our meeting room at Carnegie Mellon University. On the table the omni-directional camera can be seen; in the back, one of the pan-tilt-zoom cameras is located.

In the project the following research issues are currently addressed.

Speech Recognition Speech recognition is needed to transcribe what the participants say during the meetings. Speech recognition in meetings is a particularly challenging task. Difficulties include the conversational speaking style that tends to be observed in meetings; the usually high specialization of topics, which make collection of appropriate training data difficult, and the degraded recording conditions, in which cross-talk and background-noise are prevalent [Waibel et al. 2001b].

Figure 1.1: Image taken in the meeting room.

Dialogue Analysis The idea of dialogue analysis in the meeting room context is to use features other than keywords for information access to spoken information. Features like speaking style or speaker dominance have proven to be helpful for information retrieval in meeting minutes and provide relevant information, which can easily be visualized when browsing through meeting transcripts [Waibel et al. 2001a, Ries & Waibel 2001].

Text Summarization This module provides a relevance ranked list of sentences from a given meeting. Thus, the most relevant passages of a meeting can be displayed, which gives the user a good quick overview of a meeting's content.

Detecting Emotions

Person Tracking Detecting and tracking participants and their faces in the room is a prerequisite for tasks such as face recognition, facial expression recognition, audio-visual speech recognition and focus of attention tracking.

Face- and speaker identification Identification of the participants based on their face and their voice is necessary in order to know who participated in the meeting and who said what.

Focus of attention tracking

The Meeting Browser An important part of meeting recognition is the ability to efficiently capture, manipulate and review all aspects of a meeting. The meeting

browser has been developed [Bett et al. 2000, Waibel et al. '98] to that end. It comprises the individual components to analyze meetings, facilitates rapid access to captured meetings, allows easy browsing through meeting transcripts and facilitates the retrieval of relevant parts of a meeting.

Figure 1.2 shows an image of the current meeting browser interface. The browser is implemented in Java. It is a powerful tool that allows one to review or summarize a meeting or search an existing meeting for a particular speaker or topic.

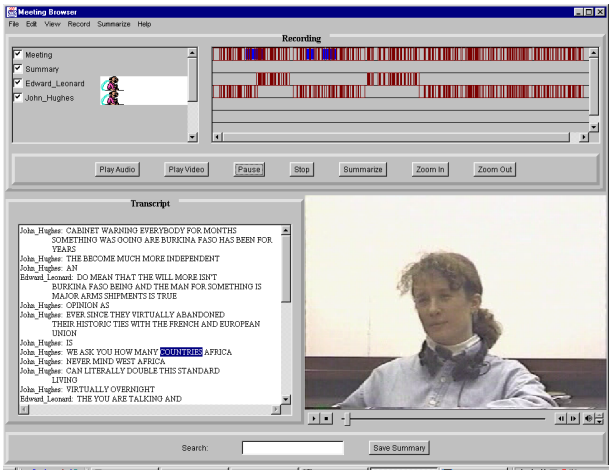


Figure 1.2: The main window of the meeting browser. It consists of three sections: an upper graphical display which shows the meeting over time, a lower left window that shows a transcript of the meeting and a lower right window which displays either a video of one of the participants or a dialogue summary.

The meeting browser interface displays meeting transcriptions, time-aligned to the corresponding audio and video files. Included in the meeting transcriptions are discourse features and emotions.

This thesis provides the components to the meeting room to track faces and persons and their locations around a meeting table as well as to determine each participant's focus of attention during the meeting.

Tracking **focus of attention** of the meeting participants can help us to better understand what was said in meetings and it might give us important additional information for the analysis of meeting transcripts.

For example, knowing which person was addressed by a speaker is essential for proper understanding of what was said. When a person says something like "You really did

a great job!" it is important to know which of the participants was addressed by the speaker.

Information about the focus of attention of participants can also be used for indexing and retrieval of parts of the meeting. Together with components for person and speaker identification [Yang et al. '99, Gross et al. 2000], queries such as "Show me all parts when John was telling Mary something about the multimedia project." become possible.

We believe that focus of attention tracking can furthermore be used to measure how actively certain participants followed ongoing discussions. This could for instance be done by counting how often they contributed to the discussion by saying something, by measuring how often the participants looked at the speakers or by monitoring how actively subjects looked at other participants in general. Counting how often a person was looked at, for instance, could provide us some idea about who was in the center of attention during different periods of the meeting.

1.2 Approach

A body of research literature suggests that humans are generally interested in what they look at. This has for example been demonstrated in Yarbus' classical experiments in which a subjects' eye movements were tracked when they watched a painting [Yarbus '67]. The close relationship between gaze and attention during social interaction has further been investigated by [Argyle '69, Argyle & Cook '76, Emery 2000]. User studies, in addition, recently reported strong evidence that people naturally look at objects or devices with which they are interacting [Maglio et al. 2000, Brumitt et al. 2000a]. This close relationship of gaze and attention will be discussed in more detail in Chapter 2.

A first step in determining someone's focus of attention, therefore is, to find out in which direction the person looks. There are two contributing factors in the formation of where a person looks: **head orientation** and **eye orientation**. In this study head orientation is considered as a sufficient cue to detect a person's direction of attention. Relevant psychological literature offers a number of convincing arguments for this approach (see Chapter 2) and the feasibility of this approach is demonstrated experimentally in this thesis.

We conducted an experiment which aimed at evaluating the potential of head orientation estimation in detecting who is looking at whom in meetings. In the experiment head orientation and eye gaze were captured using special high accuracy tracking equipment. The experimental results show that head orientation contributes 69% on

average to the overall gaze direction, and focus of attention estimation based on head orientation alone can achieve an average accuracy of 89% in a meeting application scenario with four participants.

A practical reason to use head orientation to estimate a person's focus of attention is that in scenarios such as those addressed in this thesis, head orientation can be estimated with non-intrusive methods while eye orientation can not. Although having people wear special equipment to track their eye gaze might be acceptable for user-studies or one-time occasions, it is certainly not acceptable during every day use in a meeting room. Certainly users would not want to wear head-mounted equipment, calibrate eye-gaze trackers each time they use the meeting room, or sit at fixed locations in front of the tracking hardware.

Detecting a person's head orientation, however, as this thesis will show, can be done with cameras and from a distance, even when participants are moving and when the camera resolution is low.

To map a person's head orientation onto the focused object in the scene, a model of the scene and the interesting objects in it is needed. In the case of a meeting scenario, clearly the participants around the table are likely targets of interest. Therefore, our *approach* to tracking at whom a participant is looking is the following:

1. Detect all participants in the scene
2. Estimate each participant's head orientation
3. Map each estimated head orientation to its likely targets using a probabilistic framework.

Compared to directly classifying a person's focus of attention target – based on images of the person's face, for example – our approach has the advantage that different numbers and positions of participants in the meeting can be handled. If the problem were treated as a multi-class classification problem, and a classifier such as a neural network were trained to directly learn the focus of attention target from the facial images of a user, then the number of possible focus targets would have to be known in advance. Furthermore, with such an approach it would be difficult to handle situations where participants sit at different locations than they were sitting during collection of the training data.

In our system, an *omni-directional camera* is used to capture the scene around a meeting table. Participants are detected and tracked in the panoramic image using a real-time **face tracker**. Furthermore, *neural networks* are used to compute head pose of each person simultaneously from the panoramic image.

A Bayesian approach is then used to estimate a person's focus of attention from the computed head orientation. With the proposed model, the a-posteriori probability that a person is looking at a certain target, given the observed head pose, is estimated. Using this approach, we have achieved an average accuracy of 73% in detecting the participants' focus of attention on several recorded meetings with four participants. In the experiments, each subject's focus of attention target could be one of the other three participants at the table.

Our approach to determine focus of attention is of course not perfect. Since eye gaze is neglected in our approach, a certain amount of error is introduced. The noisy estimation of head orientations from camera images introduces additional errors.

To improve the robustness of focus of attention tracking, we therefore would like to combine various sources of information. Attention is clearly influenced by external stimuli, such as noises, movements or speech of the other persons. Monitoring and using such cues might therefore help us to bias certain targets of interests against others.

Information about who is currently talking in a meeting clearly could be useful for the prediction of to whom people are attending. It seems intuitive that participants tend to look at the speaker. Argyle, for instance, pointed out that listeners use glances to signal continued attention, and that gaze patterns of speakers and listeners are closely linked to the words spoken [Argyle & Cook '76].

We have found that *focus of attention is correlated to who is speaking* in a meeting and that it is possible to estimate a person's focus of attention based on the information of who is talking at or before a given moment. To estimate where a person is looking, based on who is speaking, probability distributions of where participants are looking during certain "speaking constellations" are used. On recorded meetings with four participants we could achieve 56% accuracy in predicting the participants' focus of attention based on who is speaking.

The accuracy of sound-based prediction of focus of attention can furthermore significantly be improved by taking a history of speaker constellations into account. We have trained neural networks to predict focus of attention based on who was speaking during a short period of time. Using this approach, sound-based prediction could be increased from 56% to 66% accuracy on the recorded meetings.

Finally, the *head pose based* and the *sound-based* estimations are combined to obtain a multimodal estimate of the participants' focus of attention. By using both head pose and sound, we have achieved 76% accuracy in detecting the participants' focus of attention on the recorded meetings.

The system for focus of attention detection in meetings which is presented in this thesis has been successfully installed in both our labs at the Universität Karlsruhe, Germany and at Carnegie Mellon University in Pittsburgh, USA. A problem when *porting the system* to a new location is the need for appropriate training images for the neural network based approach to head orientation estimation. We therefore also investigated how much training data is necessary to port the system to a new location. We furthermore show how a network for head orientation estimation that was trained with images from one location (Karlsruhe) can be used in a new location (CMU) with new illumination conditions. This is done by adapting the network with a number of training images taken in the new location. In our experiments, new images from only four subjects were necessary for the adaptation of the neural network for and to achieve good focus of attention detection accuracy in the new location.

Focus of attention tracking could be also greatly beneficial for a number of other applications than analyzing meetings. To show how focus of attention tracking can be used for *multimodal context-aware interaction* in a smart environment, we have built a prototype system in which a subject can interact with a simulated household robot or a speech-enabled VCR in a room. In the demonstration system, we used the components developed for focus of attention tracking to determine whether a user was addressing the robot, the VCR or whether the user was just talking to other people in the room.

1.3 Outline

This thesis is organized as follows: In **Chapter 2** we discuss relevant literature concerning human attention, the relationship of gaze and attention and the perception of attention during social interaction. We also review state of the art techniques for eye-gaze tracking and head pose estimation.

Chapter 3 provides details about the skin-color based face tracking approach used in this work. We also describe how meeting participants can be simultaneously tracked using an omni-directional camera to capture the scene.

In **Chapter 4** we describe our approach to estimating head orientation with neural networks. We describe the data collection, the neural network architecture employed, different pre-processing methods that we investigated, and we provide experimental results.

In **Chapter 5** we present a probabilistic approach for determining at which target a person is looking, based on his or her head orientation. We discuss how the model

parameters can be adapted to different numbers and locations of meeting participants and provide experimental results on a number of recorded meetings.

In **Chapter 6** we present a user study investigating how reliably focus of attention can be estimated based on head orientation alone in meetings. Two questions were addressed in this experiment: 1) How much does head orientation contribute to gaze? 2) How reliably can we predict at whom the person was looking, based on his head orientation? To answer these questions, we have captured and analyzed gaze and head orientations of four people in meetings using special hardware equipment.

Chapter 7 suggests that focus of attention tracking could benefit from also tracking other relevant cues such as sound or movements. We specifically investigate whether focus of attention can be predicted based on who is speaking. We show that information about who is speaking is indeed a reliable cue for predicting the participant's focus. We present an approach to predict focus based on a sequence of audio-observations using a neural network. We also present experimental results indicating that the combination of sound-based focus of attention prediction and focus of attention estimation based on the subjects' head orientation leads to better results than using only one modality for focus estimation.

In **Chapter 8** we discuss how the presented system for focus of attention tracking can be installed in a new location. We explore how a neural network for head pan estimation can be *adapted* to work under new conditions by using some adaptation data collected in the new location. We examine how much adaptation data is necessary to obtain reasonable performance and compare the adaptation results to the results obtained with neural networks that are trained from scratch with the new data.

In **Chapter 9** a prototype system to demonstrate how focus of attention can be used to improve human-computer interaction is presented. In the demonstration environment, a subject can interact with a simulated household robot, a speech-enabled VCR or with other people in the room, and the recipient of the subject's speech is disambiguated using the focus of attention tracking components developed in this work.

Finally, **Chapter 10** summarizes the main contributions of this work and concludes with a discussion of limitations and future work.

Chapter 2

Background and Related Work

2.1 Human Attention

> "Every one knows what attention is. It is the taking possession by the mind, in a clear and vivid form of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others [...]." (William James)

This is how of William James, one of the most influential psychologists at the turn of the century defined attention in his major work, The principles of Psychology [James 1890/1981].

According to the Encyclopedia Britannica, attention can be defined as "the concentration of awareness on some phenomenon to the exclusion of other stimuli". It is the awareness of the here and now in a focal and perceptive way [Enc 2002].

While at first sight it might be expected that an individual is aware of all the events at a given moment, this is clearly not the case. Individuals focus upon – or attend to – a limited subset of the sensory information available at a given moment.

It is assumed that the reason for limited awareness is the limited processing capacity of the brain: we simply cannot consciously experience and process all the information available at a given time. In the primate visual system, for instance, the amount of information coming down the optical nerve is estimated to be on the order of 10^8 bits per second. This far exceeds what the brain is capable of fully processing and assimilating into conscious experience. Attention can be understood as a condition

of selective awareness. It is a strategy to deal with this processing bottleneck by only selecting portions of the input to be processed preferentially.

Attention has been a topic of study and scientific debate in experimental psychology for more than a hundred years. Psychologists began to emphasize attention in the late 19th century and early 20th century.

Wilhelm Wundt was among the first to point out the the distinction between the focal and the more general features of human awareness. He used the term "Blickfeld" to describe the wide field of awareness, within which lay the more limited focus of attention, the "Blickpunkt". He suggested that the range of the "Blickpunkt" was about six items [Enc 2002].

During the 20th century, several theories about the selective function of attention were developed.

In an influential work, Broadbent [Broadbent '58] postulated that the many signals entering the central nervous system are analyzed by the brain for certain features such as their location in space, their tonal quality, their size, their color, or other physical properties. These signals then pass through a filter that allows only those signals with appropriate, selected properties to proceed for further analysis.

Shiffrin and Schneider [Shiffrin & Schneider '77] (cf. [Enc 2002]) later formulated a "two-process" theory of attention. They distinguish between two modes of information processing: Controlled search and automatic detection. Controlled search demands high attentional capacity and is under the individual's control. By contrast, automatic detection comes into operation without active control or attention by the individual and it is difficult to suppress.

Most researchers now agree that the attention selection mechanism consists of two independent stages: an early preattentive stage, that operates without capacity limitation and in parallel across the entire visual field, followed by a later attentive stage, that can only deal with one to few items at a time [Theeuwes '93] (cf. [Glenstrup & Engell-Nielsen '95]).

The attentive selection process is however not a purely bottom-up process. Various studies indicate that visual attention can be controlled to focus on smaller areas of the visual field [LaBerge '83], [Eriksen & Yeh '85] (cf. [Glenstrup & Engell-Nielsen '95]). It is suggested that attention can be varied like a *spotlight* across the visual field, and that the spotlight "enhances the efficiency of detection of events within its beam" [Posner et al. '80] (cf. [Glenstrup & Engell-Nielsen '95]).

A good example for the willful (top-down) control of (audio-visual) attention is the "Cocktail-Party phenomenon". Cherry [Cherry '57] (cf. [Gopher '90]) described in a

series of experiment the perceived clarity and intensity for a person standing in one corner of the room, of a conversation taking place at another remote corner, but of high interest to him. This is the work of focused attention that seems to override the much louder vocalizations of surrounding parties and his own discussion partner.

Both modes of the selection process – bottom-up, preattentive selection of salient feature, and top-down control of visual attention – can happen at the same time. Visual stimuli can be willfully brought into the focus of attention, or they win the preattentive selection process [Itti & Koch 2000].

2.1.1 Computational models of attention

With the advance of computer technology and artificial intelligence, there is a growing interest in computational models of attention. Especially models of the visual attention system have been investigated.

Itti et al. [Itti et al. '98, Itti & Koch 2000] presented a bottom-up model for the control of visual attention based on saliency maps. A saliency map encodes early visual features such as color, intensity or orientation. In their model, the maximum in the saliency map is taken as the most salient stimulus and as a consequence, focus of attention is directed to this location. After inspection of one location, this location and its neighbors are "inhibited" in the saliency map and visual search proceeds to the next most salient point in the map. The idea of a saliency map to accomplish preattentive selection was first introduced by Koch and Ullman [Koch & Ullman '85]

Rao et al. [Rao et al. '95] proposed a model for saccadic targeting during a search task which combines bottom-up and top-down information. Their model uses iconic scene representations derived from spatial filters at various scales. Objects of interest to a search task are represented by a set of feature vectors, derived from the spatial filters. Visual search proceeds in a coarse-to-fine manner by finding the closest correspondence of the object of interest and the saliency image at each scale of the filters. And at each scale a saccade is directed to the closest match of the saliency map and the object representation. They report good agreement between eye movements predicted by their model and those recorded from human subjects.

Computational models of attention are also used to control gaze, visual search and orienting behaviours of robots.

Adams et al. [Adams et al. 2000], for instance, give an overview of the humanoid robot project COG, which aims at developing robots that can behave like and interact with humans. To control a robot's visual attention, they have implemented a model of visual search and attention, which was proposed by Wolfe [Wolfe '94]. The attentional

model combines color, motion and face detectors with a habituation function to produce an attention activation map. The attention process influences gaze control and the robot's internal behavioural state, which in turn influences the feature-map combination.

2.2 Where We Look Is Where We Attend To

Research literature suggest that humans are generally interested in what they look at.

Barber and Legge [Barber & Legge '76] (cf. [Glenstrup & Engell-Nielsen '95]), for example, carried out an experiment in which the they asked a group of subjects to tell what the most informative parts of pictures were. Then they tracked eye-gaze of another group of subjects regarding the same pictures. They concluded that there was good agreement between what was considered informative and what was looked at most often.

Similar conclusion can be drawn from Yarbus' classical experiments [Yarbus '67], in which eye movements are tracked as a subject responds to questions about a painted scene he watches. His experiments showed that the visual investigation of a complex scene involves complicated patterns of fixations, where the eye is held fairly still, and saccades, where the eyes move to foveate a new part of the scene, which is then attended to. The experiments also showed that the subject's eye movement patterns were highly dependent on the different tasks the subject tried to solve. It seems a reasonable explanation, that the different observed fixation patterns were due the the different information the subject was trying to find in the scene.

The human's eye has its highest acuity in the region of the fovea. This region approximately covers a visual angle of 2 degrees and is used by humans to make detailed observations of the world. The remaining part of the retina offers peripheral vision, which has only about 15-50% of the acuity of the retina, it is less color-sensitive but is more reactive to flashing objects and sudden movements [Jacob '95] (cf. [Glenstrup & Engell-Nielsen '95]). Since the fovea covers only such a small area, eye movements are necessary to capture details of our surroundings.

The movement of the eyes to a new location is performed by executing a saccade. Saccades are sudden rapid movement of the eyes, which are completed between 30-120ms after initiation. Saccades can be volutarily initiated, but are ballistic; i.e., once they are initiated, their path and target location cannot be changed. During saccades, processing of the visual image is suppressed. Thus, visual processing takes

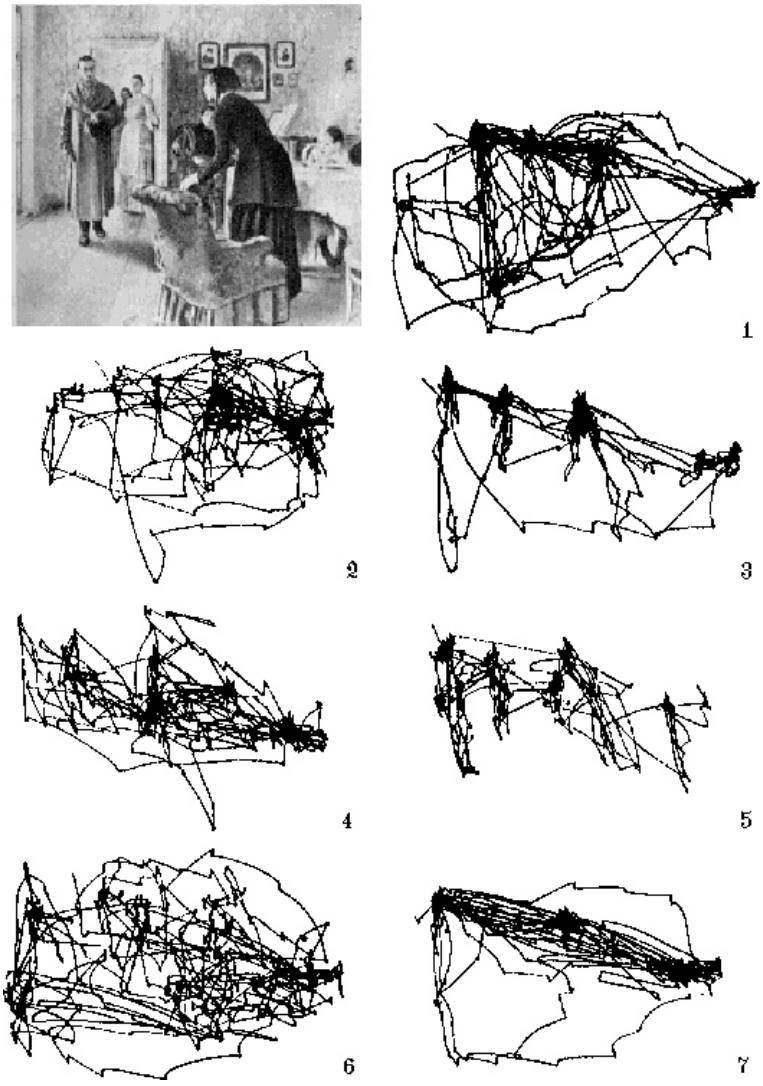


Figure 2.1: Seven records of eye movements by the same subject. Each record lasted 3 minutes. 1) Free examination. Before subsequent recordings, the subject was asked to: 2) estimate the material circumstances of the family; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the "unexpected visitor;" 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor" had been away from the family (from [Yarbus '67], cf. [Glenstrup & Engell-Nielsen '95]).

place between the saccades, the so called fixations, that last for about 200-600ms [Glenstrup & Engell-Nielsen '95].

Researchers have started to use human eye movements to build new human-computer interfaces. Applications include eye controlled interfaces for the disabled [Hutchinson et al. '89], eye gaze word processors [Frey et al. '90] and missile guiding systems. In such interfaces, users can either make use of intentional, manipulatory eye-gaze, or the user's natural eye movements are used, for example when he or she is scanning a screen [Jacob '95] (cf. [Salvucci '99]).

One problem when building interfaces using eye gaze is the difficulty of interpreting eye movement patterns. Raw eye-gaze data does not describe what we think we look at. This is caused by the unconscious eye movements such as saccades and micro-saccades, or due to gaze tracking failure, for example when the user blinked.

The problem of interpreting the raw eye movement patterns has been addressed by a number of researchers. Jacob [Jacob '93] accessed this problem by expecting a series of fixations separated by saccades and trying to fit the raw data to this model. Another approach to interpret eye gaze data has been proposed by Salvucci [Salvucci '99]. In his fixation tracing approach, hidden Markov models are used to map raw eye movements to a cognitive process model. He reports good interpretation results in an eye typing study.

Recent user studies report strong evidence that people naturally look at objects or devices they are interacting with.

Maglio et al., for instance, investigated how people use speech and gaze when interacting with an "office of the future". In their experiment, they used a Wizard-Of-Oz design, where the subjects could interact with speech-understanding office applications, such as a Calendar, a Map and an address book which were represented as different futuristic looking screen displays. They found that subjects nearly always looked at the addressed device before making a request [Maglio et al. 2000]. Furthermore they concluded that in their study gaze information alone was sufficient to disambiguate the addressed devices 98% of the time.

Similar results are reported for example by Brumitt et al. [Brumitt et al. 2000a]. They investigated different interfaces to control lights in a living room. In their study people were able to control the lights in the "Easy Living Lab", a mock up of a small living room, using various non-traditional mechanisms, such as controlling them by speech, speech and gesture, touching or using a wall display. Apart from reporting that people preferred to use their voice to control the lights, they report that subjects typically looked at the lights they wanted to control. Only in 9% of the investigated tasks, people never looked at the light they were controlling for any

commands they issued. In 25% of the tasks, people looked at the light during some of their commands, and in the remaining 66% people always looked at the light they wanted to control [Brumitt et al. 2000a]. They conclude that finding out the place in a room where someone is looking would be the most useful "gesture" to recognize in conjunction with speech recognition.

2.3 Gaze and Attention During Social Interaction

An important ability of humans and other primates is the ability to monitor where other individuals look. This can signal where they are currently attending, it might signal sources of possible interest or of immediate danger.

In the primate literature, there is for instance much evidence suggesting that some primates use gaze to convey information about their intentions. Baboons and vervets for example, use quick glances between an aggressor and a potential helper to gain support from the potential helper. It is also assumed that primates are using gaze to influence the behavior of a human care-giver. Some experiments with monkeys showed evidence that these monkeys link gaze of the human experimenter with his intentional actions [Emery 2000].

The detection of another's gaze is also important to establish joint attention, which is critical for learning and language acquisition. The age at which an infant first follows another's gaze is controversial, ranging from 6 to 18 months of age [Emery 2000]. Before they are 12 months old, human infants can follow their mother's gaze, but cannot direct their attention to the object of her attention and at around 12 months of age, they begin to follow their mother's gaze towards particular objects in their visual field [Emery 2000].

Joint attention may especially be important for language learning in human infants. An early stage in language development is the process of associating a word with the physical presence of an object. This stage of learning is difficult to achieve without the ability to follow gaze. By following a speaker's line of regard, the infant can determine the intended referent of a new word [Baldwin '91].

The close relationship of a person's gaze and his or her direction of attention during social interaction has long been emphasized. In an extensive study, Argyle discriminates between a number of different functions of gaze during conversations [Argyle '69, Argyle & Cook '76]:

Gaze as signal and channel Gaze not only serves as a signal, but also to open and close the visual channel itself; i.e., in order to monitor someone's visual signals,

one first has to point his gaze towards him. During conversations, speakers look up to get feedback from their audience, and listeners look at the speakers to study their facial expressions and their direction of gaze.

Gaze to signal interpersonal attitudes One of the roles of gaze is the signaling of interpersonal attitudes, such as liking, hostility and emotions, such as shame, embarrassment or sorrow. Studies provided evidence that people look more at those they like [Exine & Winter '66], and – with some exceptions – people who look more create a more favorable impression and are liked more.

Dominance and leadership Gaze is related to dominance and leadership during interactions. During communication between two people, people looked more at people of higher status. In addition, persons giving good arguments in group discussion are both looked at more, and are rated higher on leadership qualities [Burroughs et al. '73].

Gaze and speech There is evidence that gaze patterns of speakers and listeners are closely linked to the words spoken, and are also important in handling timing and synchronization of utterances: glances of the speaker are used as grammatical breaks, to emphasize particular words or phrases and gaze sometimes is used to pass the word to the next speaker. On the other side, listeners use glances to signal continued attention, to reinforce particular points and to encourage the speaker or to indicate surprise, disbelief or anger.

Gaze as a signal of attention The most basic meaning of gaze during interaction is as a signal of attention: a person looking at another person signals, that his visual channel is open and that he is paying attention. People who look more are perceived as more attentive and for example looking down during a conversation is interpreted as a sign of inattention [J.W.Tankard '70]. In addition studies prove that it is considered polite to look at people when interacting with them [Kleinke et al. '73].

Similar results are reported by Ruusuvuori [Ruusuvuori 2001]. They studied the coordination of patients' production of their complaint and the doctors' orientation to the patient on the one hand and to medical records on the other. In this study it is suggested that disengaging from interaction by orienting towards the medical records may leave the patient puzzled about whether the doctor is listening or not.

Vertegaal et al. [Vertegaal et al. 2001] investigated the relationship of where people look and whom they attend to during multi-party conversations. They found that subjects looked about 7 times more at the individual they listened to than at others,

and that subjects looked about 3 times more at individuals they spoke to. They conclude that information about who is looking at whom is an ideal candidate to provide addressee information and that it can also be used to predict to whom someone is listening.

2.4 Cues for the Perception of Gaze

The main cue in detecting where other individuals look at are the eyes. In fact, it is assumed that the morphology of the human eye, with its white sclera and the dark pupil, may have evolved to facilitate gaze perception, and thus to facilitate joint attention in our highly social species (cf. [Emery 2000]).

Although the eyes are the primary source for detecting a person's direction of attention, the perception of another person's direction of attention is not limited to information from the eyes alone.

Langton et al. suggest that in addition to gaze, there are also other cues, such as head orientation, body posture and pointing gestures, which make a large contribution to the perception of another's direction of attention [Langton et al. 2000, Langton 2000]. They report several experiments that demonstrate how head orientation influences the perception of gaze even when they eyes are clearly visible. All these experiments indicate that perception of gaze must be based on some combination of information from the eyes and from head orientation.

Perret et al. [Perret & Emery '94] have proposed a model based on neurophysiological research which describes how humans combine information from eye gaze, head orientation and body posture to determine where another individual is attending to. In their model, information from gaze, head orientation and body posture are combined hierarchically: direction-of-attention will be signaled by the eyes if these are visible, but if they are obscured, or if the face is viewed at too great a distance, head orientation will be used to determine direction of attention. If information from the eyes and the head are unavailable, attention direction is signaled by the orientation of the body. All theses cues are likely to be processed automatically by observers and all make contributions to the perceptions of another person's attention [Perret & Emery '94].

The experiments by Langton however showed that even when the eyes are clearly visible, head orientation strongly influences the perception of gaze. This indicates that head orientation and eye gaze may be processed in parallel and play a more equal role for the perception of attention direction. In fact, for children it seems that prior to 14-18 months simply the head orientation is used as an attention-following cue

and that eye-gaze is ignored [Langton et al. 2000]. He concludes that the orientation of the head makes a large contribution to the perception of another's direction of attention [Langton et al. 2000].

Several studies suggest that head orientation is in fact a sufficient indicator of attention direction [Emery 2000, Argyle & Cook '76, Cranach '71]. Cranach [Cranach '71] argued that gaze changes during social interaction are usually accompanied by head orientation changes. Argyle constitutes that this "implies that most lookers in effect cooperate by making head movements, or other special expressive movements accompanying shifts of gaze" [Argyle & Cook '76] (page 49).

2.5 Eye Gaze Tracking Techniques

There are a number of commercially available systems to track a person's eye gaze. The different methods for eye gaze tracking can be classified into the following methods [Glenstrup & Engell-Nielsen '95, Calhoun & McMillan '98]:

1. **Electro-oculography**. Measuring the electric potential of the skin around the eyes. This technique is based on the existence of an electrostatic field that rotates along with the eye.

2. Applying special **contact lenses** that facilitate tracking of the eye-ball. There are two lens techniques: a) engraving plane mirror surfaces on the lens that facilitate tracking and b) implanting a tiny induction coil into the lens. The positioning of the coil can be measured through the use of special magnetic fields placed around the user's head.

3. Measuring the **reflectance of light** – typically infrared light – that is directed onto the eye of the user.

Electrooculography (EOG) is based on the existence of an electrostatic field that rotates with the eye. By detecting differences in the skin potential around the eyes, the position of the eye can be detected [Gips et al. '93]. To measure the potential differences, electrodes have to be placed around the subject's eyes, which makes this method quite intrusive. According to [Calhoun & McMillan '98] there are however some problems associated with this technique, such as varying skin resistance over time and potential changes due to lighting adaption of the eye, which make this method unlikely to work robustly outside the lab.

Since for the methods of type 2 the user has to wear special contact lenses, the practical use of such methods in out-of-the lab scenarios is very limited.

Several commercially available systems use infrared light that is shone into the user's eye. The resulting reflections occurring on the lens and the cornea of the eye, the so-called Purkinje images, can be used to compute the user's eye gaze.

Other vision-based approaches aim at measuring eye gaze by detecting certain features of the eye in the image, such as the boundary and center of the pupil and the corners of the eye. Eye gaze is then computed by estimating the rotation of the eye-ball based on the detected features.

A different eye-gaze tracking technique was proposes by Baluja and Pomerleau [Baluja & Pomerleau '94]. In their system, eye-gaze was estimated with artificial neural networks based on low-resolution images of the user's eyes. They used a stationary light in front of the user and the system started by finding the user's eye by searching the image for the reflection of this light. In [Stiefelhagen et al. '97c] a similar neural network based approach was described, which did not require special lighting.

A main technical problem associated with these vision based approaches is the acquisition of stable frontal images of the user's eyes with good image resolution. Therefore, head-mounted cameras tend to be used, or the user has to be in more or less fixed position with regard to the camera, so that tracking of the eyes is possible. Other disadvantages of these methods are that they are very sensitive to illumination changes and placement of tracking components.

Some available commercial head mounted eye-gaze tracking systems are depicted in Figure 2.2. Using such head mounted eye-gaze trackers eye-gaze can even be measured when a user is moving his head.

Figure 2.3 shows pictures of a few eye-gaze trackers that do not require head mounted cameras. With such systems, the user's head movements are however very restricted since eyes have to stay within the view of the camera. With ASL's eye-gaze tracker Model 504, which is depicted leftmost in Figure 2.3, head movement for instance is restricted to one square foot according to the specifications of that system [ASL].

Figure 2.4 shows typical eye images used in commercial eye tracking systems. It is clear that such high-resolution images are difficult to obtain without precise control of the user's position with respect to the camera.

To summarize: Eye gaze tracking systems of the kind available today are not acceptable to be used in a meeting room. While they are able to monitor a person's eye gaze with high accuracy at high sampling rates, none of them can be used to track


ASL Model 501 [ASL]


Eyelink II by SR Research [SRR


$MI's 3D-VOG system [SMI]

Figure 2.2: Some commercial head mounted eye gaze trackers.


ASL Model 504 [ASL]


LC Technologies' system [LCT]


Arrington's system [ARR]

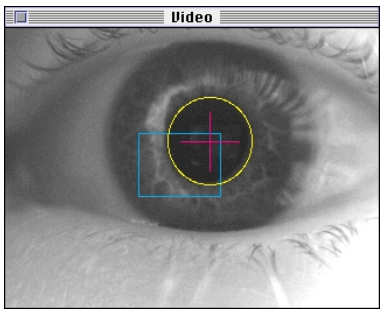
Figure 2.3: Remote eye-gaze tracking systems.

a person's eye gaze without carefully controlling the seating of the user with respect to the tracking system or without having the user wear head mounted cameras.

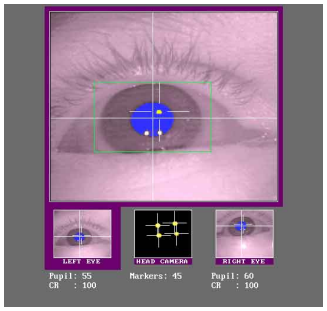
2.6 Head Pose Tracking

If intrusiveness is not an issue, a number of commercial systems can be used track a user's head orientation.

An often used system is the FASTRAK system by Polhemus Inc. [Polhemus]. The tracking system uses electro-magnetic fields to determine the position and orientation (pose) of a sensor that can be attached to a remote object, such as a head. The system consists of a main system electronics unit, one to four pose sensors and a single magnetic transmitter. The pose sensors are connected to the main unit via cables and the system unit can be connected to a computer through a serial interface.



(Picture from [ARR])



Picture from [SRR]

Figure 2.4: Typical image resolutions used in commercial eye gaze tracking systems.

According to the manufacturer's specification, the possible accuracy of the system is 0.15 degrees for the sensor orientation. Such accuracy, however, is only obtained when the pose sensors are located within 30 inches from the magnetic transmitter and when no sources of high magnetic disturbance such as computer monitors or iron shelves are located near the pose sensors. The system delivers the orientation and position parameters at 60 Hz if only one sensor is used.

A magnetic pose tracking system with similar specifications is offered by Ascension Technology Corp [Ascension]. Yet another system is available from iReality.com, Inc. [iReality]. They offer a low-cost 3D orientation tracker intended to be used in virtual reality games, for instance, where high precision is not necessary. This systems utlizes an electronic compass and tilt sensors to compute pose. According to the product specifications, the tracker has an accuracy of +/- 2 degrees for heading and tilt.

Such tracking systems are often used for virtual reality applications, together with head mounted displays or head mounted eye trackers, or to track position and orientation of data gloves. The advantage of such tracking systems is that they can deliver quite accurate head orientation measurements at high frame rates. In order to compute a user's head orientation, these systems, however, would require the user to have a special magnetic sensor attached to his head.

2.6.1 Vision-Based Methods

Vision-based approaches to estimate head orientation from camera images provide a less intrusive alternative to the above tracking systems.

Approaches to vision-based estimation of head orientation can be divided into two categories: model based approaches and appearance based approaches.

In **model-based approaches**, tracking head orientation is usually formulated as a pose estimation problem. Pose Estimation is the task to recover the 3D position and rotation of an object, with respect to a certain coordinate system. Estimating a person's head rotation can be formulated as a pose estimation problem where the task is to recover the 3D rotation and translation of the head. To recover head pose usually a number of facial features, such as eyes, nostrils, lip-corner have to be located. By finding correspondences between the facial landmarks points in the image and their respective locations in a head model, head pose can be computed [Haralick et al. '89], [Gee & Cipolla '94], [Gee & Cipolla '95], [Stiefelhagen et al. '96], [Stiefelhagen et al. '97b], [Jebara & Pentland '97], [Heinzmann & Zelinsky '98].

The main difficulty with these approaches is the reliable tracking of the facial landmark points. It requires rather high resolution facial images, and tracking is likely to fail when quick head movements or occlusions of certain features occurs.

Appearance based approaches on the other hand, estimate head orientation from the whole image of the face. Appearance based approaches either use some function approximation technique, such as a neural network, to estimate head orientation from an image [Beymer et al. '93, Schiele & Waibel '95, Rae & Ritter '98, Kwong & Gong '99], or a face database is used to encode example images [Pentland et al. '94, Ong et al. '98]. Head pose of new images is then estimated using the chosen function approximator, or by matching novel images to the images in the database.

Another appearance based approach is presented by Cascia et al. In their approach 3D head tracking is achieved by registration of a facial image to a 3D surface model of a face. One problem with their approach is the lack of a backup technique when the track is lost. In their model the positioning of the initial model has to be done by hand. And since tracking errors accumulate over time, the performance of the tracker gradually decreases after a few hundred frames [Cascia et al. '98].

The main advantage of using an appearance based approach to estimate head orientation is that no facial landmark points have to be detected and tracked. Only the facial region has to be detected to estimate head pose.

A problem of appearance based approaches is, however, that a sufficient number and variety of example images are necessary for good pose estimation results on unseen images. Furthermore, such approaches tend to be sensitive to different illumination conditions, since these affect the appearance of the facial images. The problem of pose

estimation under different illumination conditions will be discussed in later sections of this thesis.

2.7 Summary

Due to the limited processing capacity of the brain, only a small subset of the available sensory input reaches a level of consciousness and becomes aware to us. This subset is determined by our attention, which is partly an unconscious selection process, and partly can be controlled willfully. Gaze is a good indicator of a subject's focus of attention. We usually look at the objects that are currently of interest to us. This is true for social interaction and for interaction with objects or devices. While gaze is the main cue that humans use to monitor where other individuals look, also head and body orientation influence the perception of another person's gaze. Some studies suggest that head orientation is in effect a sufficient indicator the determine the direction of attention.

Several hardware and software based methods for head pose and eye gaze tracking exist today. The practical usefulness of eye gaze tracking methods is however limited. With state of the art eye tracking technology, the location of the user with respect to the eye tracker has to be carefully controlled or the user has to wear head mounted cameras.

Chapter 3

Detecting and Tracking Faces

Detecting a face in a camera image is a prerequisite for many applications including face recognition, facial expression analysis and audio-visual speech recognition. It is particularly necessary for vision-based approaches to head pose and gaze tracking, such as they are investigated in this work.

In this chapter we will give a brief overview of existing vision based face detection approaches. We will then discuss a color-based approach for face detection and tracking in more detail, since this face detection technique is used in our own system. Finally we describe our use of a panoramic camera to capture a meeting scene and how faces are detected using the panoramic images.

3.1 Appearance Based Face Detection

Several approaches for face locating have been reported: Turk and Pentland described how Eigenfaces, obtained by performing a principal component analysis on a set of faces, can be used to identify a face [Turk & Pentland '91].

Sung and Poggio [Sung & Poggio '94] report a face detection system based on clustering techniques. A similar system with better results has been claimed by Rowley et al. [Rowley et al. '95]. In [Rowley et al. '95] neural nets are used as the basic components to classify whether a sub-image contains a face or not.

Schneiderman and Kanade [Schneiderman & Kanade 2000] recently proposed a statistical method for face detection. In their approach, appearance of both faces and non-faces are represented as the product of histograms, describing joint statistics of wavelet coefficients and their positions on the face. On the test set used in

[Sung & Poggio '94] and in [Rowley et al. '95], their approach appeared to obtain the best results.

A main drawback of these appearance based approaches however is their computational effort. In all the approaches, sub-images at various sizes and at many (if not all) positions of the input image have to be processed to detect faces, which makes them rather slow. Schneiderman for example reports face detection times of 1 minute with a coarse to fine search strategy on 320x240 images and 5 seconds detection time with an optimized version that can only detect frontal faces.

Recently, Viola and Jones [Vioal & Jones 2001] have presented a very fast appearance based face detection approach. In their approach, a "cascade" of increasingly complex classifiers is used to detect faces in images. The key idea of this approach is to start with simple efficient classifiers to reject many sub-windows that are very unlikely to contain a face and then use more and more complex classifiers to investigate whether not rejected sub-windows contain a face or not. In their approach vertical, horizontal and diagonal filters at various sizes are used to compute image features. A learning approach (AdaBoost) is used to find good classification functions and to find an optimal cascade of classifiers. Their upright frontal face detection system achieves results comparable to the systems presented in [Rowley et al. '95] and [Schneiderman & Kanade 2000] while running at 15 frames per second.

3.2 Face Detection Using Color

A different approach for locating and tracking faces is described by Hunke and Waibel [Hunke & Waibel '94]. They locate faces by searching for skin-color in the image and use neural networks to distinguish faces from other skin-colored objects such as hands and arms. The main advantages of a color-based face detection and tracking approach are its simplicity and its speed.

In the research presented here, a skin-color based face detection approach as described in [Yang & Waibel '96] was implemented and used to find and extract faces in images. In the following paragraphs, this face detector is described in more detail.

Most video cameras use an RGB representation of colors. However, RGB is not necessarily the best color representation for characterizing skin-color. In the RGB space, a triple [R,G,B] represents not only color but also brightness.

However, it has been shown that the seemingly strong differences between skin colors of different individuals (including Asian, black, white faces) are mainly based on

brightness of the reflected skin colors, and that skin-colors form a cluster in chromatic color space [Hunke & Waibel '94, Yang & Waibel '96].

Chromatic colors (r,g) [Wyszecki & Styles '82], also known as "pure" colors in the absence of brightness, are defined by a normalization process:

$$r = \frac{R}{R+G+B},$$
$$g = \frac{G}{R+G+B}.$$

In fact, this defines a mapping from a three dimensional color space R^3 to a two dimensional one, R^2. Color blue, which is defined as $b = \frac{B}{R+G+B}$, is redundant after normalization because r+g+b = 1.

3.3 A Stochastic Skin-Color Model

It has been shown that the skin-color distribution in chromatic color space has a regular shape which remains similar under changing lighting conditions [Yang & Waibel '96].

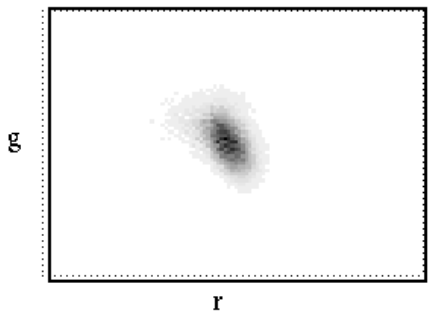


Figure 3.1: Skin-color distribution of forty people

Figure 3.1 shows such a skin color distribution. The histogram shows the skin-color of forty people in the chromatic color space (from [Yang & Waibel '96]). This skin-color distribution was obtained by analyzing faces of different races, including Asian, African American, and Caucasian. The grey-scale in the figure reflects the magnitude of the histogram. It can be seen that skin-colors are clustered in a small area of the

chromatic color space; i.e., only a few of all possible colors actually occur in a human face.

Such a distribution can be represented by a Gaussian model $N(m, \Sigma^2)$, where $m = (\bar{r}, \bar{g})$ with

$$\bar{r} = \frac{1}{N} \sum_{i=1}^{N} r_i, \tag{3.1}$$

$$\bar{g} = \frac{1}{N} \sum_{i=1}^{N} g_i, \tag{3.2}$$

and

$$\sum = \begin{bmatrix} \sigma_{rr} & \sigma_{rg} \\ \sigma_{gr} & \sigma_{gg} \end{bmatrix} \tag{3.3}$$

Following [Yang & Waibel '96], the skin-color model then can be created as follows:

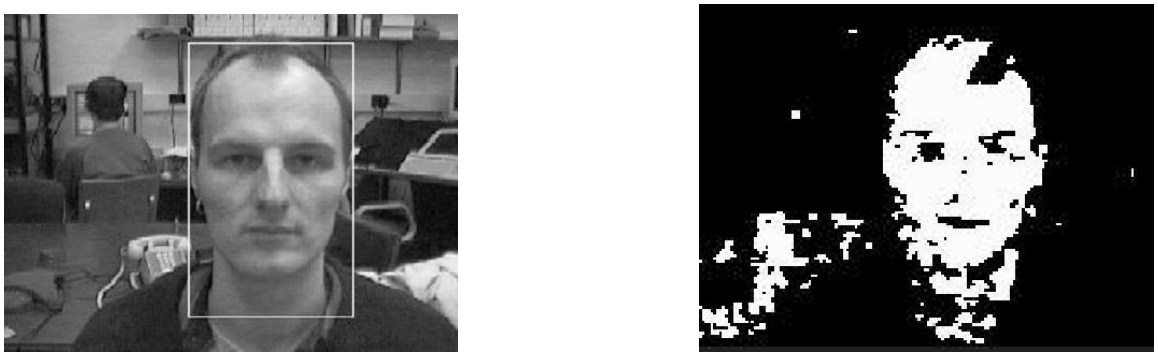
1. Take a face image, or a set of face images if a general model is needed
2. Select the skin-colored region(s) interactively
3. Estimate the mean and the covariance of the color distribution in chromatic color space based on (3.1) - (3.3)
4. Substitute the estimated parameters into the Gaussian distribution model

Since the model has only six parameters, it is easy to estimate and adapt them to different people and lighting conditions.

3.4 Locating Faces Using the Skin-Color Model

A straightforward way to find a face is to match the skin color model with the input image to find the skin color clusters. Each pixel of the original image is converted into the chromatic color space and its probability of being skin colored is computed using the Gaussian skin color model.

Figure 3.2 shows the application of the skin color model to a sample image containing a face. In the image on the right, only pixels with a high probability of being skin-color are depicted white, all other pixels are black. Assuming that there is only one face contained in the image, the face can be located by looking for the largest connected region of skin-colored pixels.



Input image (color!) Skin-colored regions

Figure 3.2: Application of the color model to a sample input image. The face is found in the input image (marked by a white rectangle)

3.5 Tracking Faces With an Omni-Directional Camera

In our system, we use a panoramic camera put on top of the conference table to capture the scene. This has the advantage that only one camera is necessary to capture all participants around the table. Compared to using multiple cameras to capture the scene, no camera calibration is necessary with the omni-directional camera. Furthermore, only one video-stream has to be captured, which eliminates the need for synchronization and reduces hardware needs.

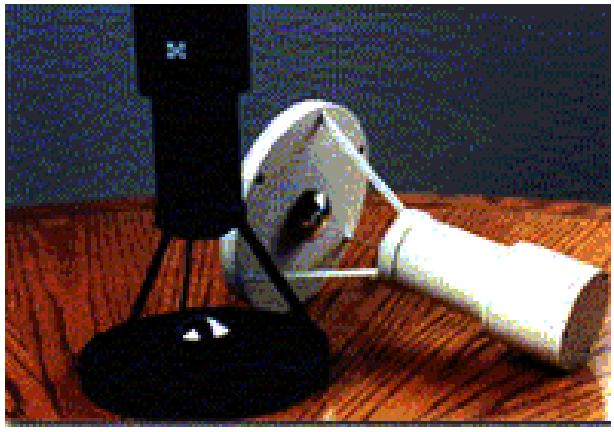


Figure 3.3: The panoramic camera used to capture the scene[1]

Figure 3.4: Meeting scene as captured with the panoramic camera

Figure 3.3 shows a picture of the panoramic camera system we are using. The camera is located in the top cylinder and is focusing on a parabolic mirror on the bottom plate. Through this mirror almost a whole hemisphere of the surrounding scene is visible as shown in Figure 3.4. This figure shows the view of a meeting scene as it is seen in the parabolic mirror and as it is captured with the panoramic camera.

As the topology of the mirror and the optical system are known, it is possible to compute panoramic views of the scene as well as perspective views at different angles of the panoramic view [Baker & Nayar '98]. Figure 3.5 shows the rectified panoramic image (with faces marked) of the camera view depicted in Figure 3.4.

To detect and track faces in the panoramic camera view, we use the skin-color tracker described in the previous section: The input image is searched for pixels with skin colors. Connected regions of skin-colored pixels in the camera image are considered as possible faces. Since humans rarely sit perfectly still for a long time, motion detection is in addition used to reject outliers that might be caused due to noise in the image or skin-like objects in the background of the scene that are not faces or hands. Only regions with a response from the color-classifier and some motion during a period of time are considered as faces.

The drawback of this approach, however, is that faces and hands are not yet dis-

[1]Image courtesy of CycloVision Technologies, Inc.

Figure 3.5: Panoramic view of the scene around the conference table. Faces are automatically detected and tracked (marked with boxes).

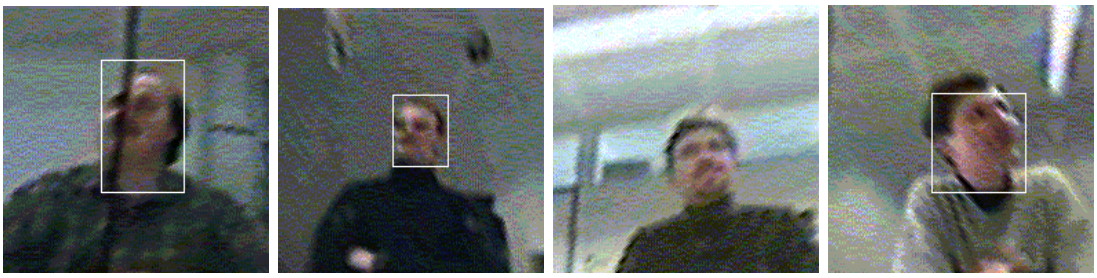


Figure 3.6: Perspective Views of the meeting participants.

tinguished sufficiently. Therefore we consider skin-colored regions as belonging to the same person if the projection of their centers onto the x-axis are close enough together. Among the candidate regions belonging to one person, we consider the uppermost skin-like region to be the face and consider the lower skin-like region to be hands. Figure 3.5 shows the rectified panoramic image (with faces marked) of the camera view depicted in Figure 3.4.

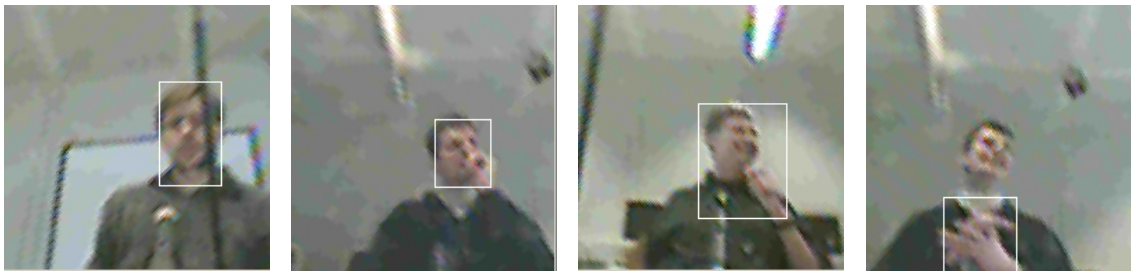
Once the faces in the image are found, perspective views of each person can be computed. Compared to panoramic view as depicted in 3.5, the perspective images are further rectified. Straight lines in the scene now appear as straight lines in the images. Figure 3.6 shows the perspective images of the participants detected in the panoramic image as depicted in 3.5.

Once the perspective images are generated, faces are again searched in these views using the color-based face detector. The faces detected and extracted from these perspective view are later on used to estimate each participant's head pose. This will be described in detail in Chapter 4.

3.5.1 Discussion

On the meetings that we captured with the omni-directional camera (see Chapter 5 for more details), we found that the color-based face tracker correctly detected the participants' faces about 94% of the time.

Errors happened for example because a person's face was occluded by his hand or arms or by a coffee mug. Sometimes a face was also occluded by one of the posts of the camera. In other cases of tracking failure occured because either the hands or arms of the user were falsely classified as the face. Figure 3.7 shows a few examples where faces were not correctly detected.



(a) (b) (c) (d)

Figure 3.7: Some sample images of occluded or not correctly detected faces.

One possibility to detect such outliers could be to use an appearance based face detector (e.g. as described by [Rowley et al. '95] or [Schneiderman & Kanade 2000]) to verify whether the detected candidate region contains a face or not. Such an approach might also be useful to refine face detection, when the face really covers only a part of the candidate region detected by the color-based face tracker (see e.g. Figure 3.7(c)). This has, however, not been investigated in this thesis and remains for further work.

Chapter 4

Head Pose Estimation Using Neural Networks

In this chapter we will describe our approach to estimating head orientation with neural networks.

In this work we aim at estimating head orientation directly from facial images. The main advantage of such an appearance based approach is that no facial landmark points have to be detected in order to compute head pose. Instead, head pose is estimated from the whole facial image and therefore only the face has to be detected and tracked in the camera image. This is especially advantageous if head orientation has to be estimated from small or low resolution facial images, as it is true in our case. In the images captured with the omni-directional camera, which we are using to simultaneously track several participants' faces around a table, detailed facial features such as eyes or lip-corners, mostly cannot be detected nor tracked (see Figure 4.4 for some sample images).

We use neural networks to estimate pan and tilt of a person's head from pre-processed facial images. Neural networks provide a practical learning technique that has been widely used for pattern recognition problems. In the field of computer vision, neural networks have been successfully used for many tasks such as the visual recognition of hand postures [Meyering & Ritter '92], [Drees '95] and facial expressions [Rosenblum et al. '96], the detection of faces [Rowley et al. '98], [Rowley et al. '95], [Hunke & Waibel '94] and pedestrians [Wohler et al. '98], [Zhao & Thorpe '99], for mobile robot guidance [Pomerleau '92] and for the estimation of head poses [Rae & Ritter '98], [Schiele & Waibel '95], to name a few.

Neural networks can be used to "learn" a function that maps the network's input to its output, given input/output vectors as examples during a training phase. In our

approach, the input vectors are preprocessed and vectorized facial images and the outputs are the horizontal (pan) or vertical (tilt) rotation of the input images. When a new image is provided to the trained neural network, it will produce as its output an estimate of the orientation of the face in the input image.

A similar approach is described by Schiele and Waibel [Schiele & Waibel '95]. They describe a system to estimate head pan from facial color images. In their system, faces are classified to belong to a number of head rotation classes, which correspond to 15 quantized rotation angles from -70 to +70 degrees. Their approach, however, differs in several aspects from the approach presented here. Schiele and Waibel's system estimated head pan in ten degree steps, whereas in our approach both the head rotation in pan and tilt direction are estimated without any quantization. In Schiele and Waibel's system, furthermore, only images of faces with no rotation in tilt-direction were used for training and testing, whereas in our system no such restrictions are made, but instead a user's head orientation in any direction is allowed.

Another system that uses neural networks to estimate head pose from images is described in [Rae & Ritter '98]. As compared to the work presented in this thesis, their system, however, is user-dependent and only results on a single user are reported. In their approach, color segmentation, ellipse fitting, and Gabor-filtering on a segmented face are used for preprocessing. They reported an average accuracy of 9 degrees for pan and 7 degrees for tilt for the one user.

We have trained neural networks to estimate a person's head rotation from two kinds of camera images: 1) images from a pan-tilt-zoom camera (Canon VC-C1) and 2) an omni-directional camera.

Pan-tilt-zoom cameras are ideal for tasks where only one person's face and head orientation have to be monitored. An advantage of using an omni-directional camera, however, is that all participants sitting around a table can be simultaneously tracked in one camera view. It is therefore not necessary to synchronize, calibrate and control a number of cameras to track the participants faces and their locations.

4.1 Data Collection

In order to get sufficient generalization to new users, it is necessary to collect training images from different people to train the neural networks. We collected facial images from many members and students of our lab to get a sufficient amount of training data. During data collection, users had to wear a head band with a sensor of a Polhemus pose tracker attached to it [Polhemus]. Using the pose tracker, the head pose with respect to a magnetic transmitter could be collected in real-time.

Figure 4.1: Some sample images from the pan-tilt-zoom camera taken in the computer lab.

Figure 4.2: Some sample images from the pan-tilt-zoom camera taken in a second room with many windows.

4.1.1 Data Collection With a Pan-Tilt-Zoom Camera

In case of the networks for regular CCD-camera images, we collected data in two different rooms with different lighting conditions. One of the rooms was a computer lab which is mainly illuminated by a number of neon lights at the ceiling, the other room had windows all along two of the four walls of the room and was illuminated by daylight only. In the first room, we took images from 14, in the second room images from 16 persons. Altogether more than 14,000 images from 19 different persons were collected. Eighteen of the users were male, five of the nineteen users were wearing glasses. Hair-styles ranged from almost bald to long hair.

To capture the training images, the camera was positioned approximately 1.5 meters in front of the user's head. The user was asked to randomly look around in the room and the images together with the pose sensor readings were recorded. Some sample images taken in the computer lab are shown in Figure 4.1. Figure 4.2 shows some images taken in the room with windows. One can see the users wear the head band with the little magnetic pose sensor attached to it.

The recorded head orientations varied between 90 degrees to the left and to the right and approximately 60 degrees up or down. Figure 4.3 shows the distributions of horizontal and vertical head rotations in the collected data.

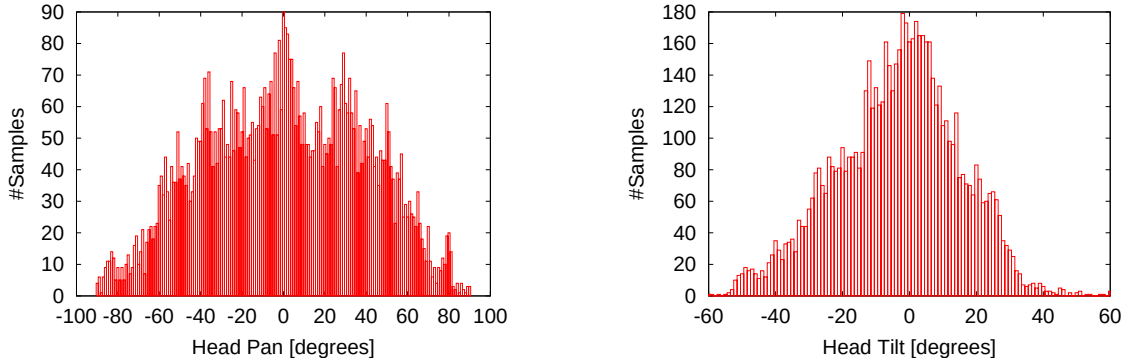


Figure 4.3: Distributions of horizontal (pan) and vertical (tilt) head rotations in the collected data set.

4.1.2 Data Collection With the Omni-Directional Camera

During data collection with the omni-directional camera, we only collected images in the computer lab. Each subject had to sit at the table on which the camera was positioned. We then tracked the subject's face in the panoramic view generated from the omni-directional camera view and computed a perspective view of the subject's face as described in Section 3.5. The subject again had to wear the head band with the pose tracker sensor attached to it, to obtain ground truth of the user's head pose. The subject was asked to look around, and the perspective views of the subject were recorded together with the pose sensor readings.

Each subject was furthermore positioned at four different locations around the table and training data was collected. This was done to obtain training images with different illumination of the faces.

Altogether, we collected data from fourteen users at four positions around a table. All of the users were male, four of them had glasses.

Figure 4.4 shows some sample images taken during data collection with the omni-directional camera. It can be seen that the image resolution is much lower than the resolution of the images obtained with the pan-tilt-zoom camera. In addition, facial images are captured from a different angle than with the pan-tilt-zoom camera. This is due to the fact that the omni-directional camera is positioned at the table, while the pan-tilt-zoom camera was positioned approximately at the height of the user's face.

Table 4.1 summarizes the data that we collected with the different cameras.

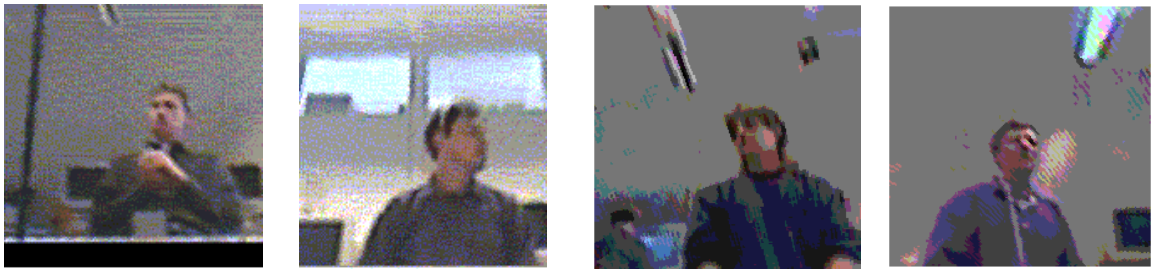


Figure 4.4: Training Samples: The perspective images were generated from a panoramic view. Head pose labels are collected with a magnetic field pose tracker.

Camera type	Room	#subjects	#images
Pan-Tilt-Zoom	Computer Lab	14	6972
Pan-Tilt-Zoom	Seminar Room	16	7468
Omni-directional	Computer Lab	14	10290

Table 4.1: Collected data to train and test networks.

4.2 Image Preprocessing

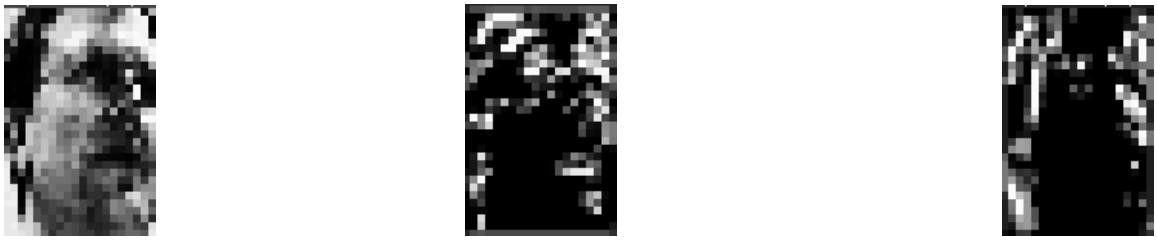
As input to the neural nets, three different approaches were evaluated:

1. Using histogram normalized gray-scale images as input
2. Using horizontal and vertical edge images as input
3. Using both normalized gray-scale plus the edge images as input.

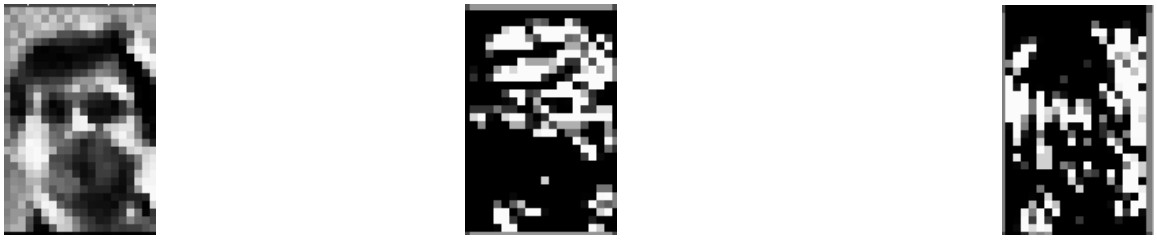
To find and extract faces in the collected images, we use the color-based face detector described in Section 3.2.

4.2.1 Histogram Normalization

In the first preprocessing approach, histogram normalization is applied to the gray-scale face images. No additional feature extraction is performed. The normalized gray-scale images are down-sampled to a fixed size of 20x30 pixels and are then used as input to the nets.



(a) Original image from pan-tilt-zoom camera.



(b) Original image from omni-directional camera.

Figure 4.5: Pre-processed images: normalized gray-scale, horizontal edge and vertical edge image (from left to right).

Histogram normalization defines a mapping of gray levels p into into gray levels q such that the distribution of q matches a certain target distribution (e.g., a uniform distribution). This mapping stretches contrast and usually improves the detectability of many image features [Ballard & Brown '82]. Histogram normalization is also helpful to get some illumination invariance.

4.2.2 Edge Detection

In the second approach, the Sobel operator is used to extract horizontal and vertical edges from the facial gray scale images. The resulting edge images are then binarized (thresholded) and down-sampled to 20x30 pixels and are both used as input to the neural nets.

Figure 4.5(a) shows the pre-processed facial images of a user captured with the pan-tilt-zoom camera. From left to right, the normalized gray-scale image, the horizontal and vertical edge images of a user's face are depicted. Figure 4.5(b) shows the corresponding pre-processed images of a face that was captured with the omni-directional camera.

4.3 Neural Network Architecture

In this work we use multi-layer perceptrons with one hidden layer to estimate head pan and tilt from images. A multi-layer perceptron is simple feed-forward network with differentiable activation functions. Such a network can be efficiently trained using gradient descent (error backpropagation) [Bishop '95, John Hertz '91].

Multilayer perceptrons consist of an input layer, one or more hidden layers and an output layer. Each unit in the hidden and output layer computes a nonlinear function of its input vector $\mathbf{x}$ consisting of a linear activation function followed by a non-linear transfer function. The following activation function is used for all units:

$$a_i(\mathbf{x}) = \sum_{k=1}^{N} w_{ik} x_k + b_i,$$

where the w_{ik} are the weights and b_i are the unit biases. As transfer functions mostly the sigmoid function is used, yielding to the following output function of a unit:

$$y_i(\mathbf{x}) = \frac{1}{1 + exp(-a_i(\mathbf{x}))} = \frac{1}{1 + exp(\sum_{k=1}^{N} w_{ik} x_k + b_i)}.$$

We have trained separate nets to estimate head pan and tilt. For each net, a multi-layer perceptron architecture with one output unit (for pan or tilt) and one hidden layer with 20 to 150 hidden units was used. For both the output units and the hidden units, the sigmoid function was used as transfer function.

The size of the input retina depended on the different number and type of input images that we investigated. When only the histogram normalized gray-scale image was used as input, the size of the input retina was 20x30 units; when the two edge images were used, the input retina had 2x20x30 units; when both the histogram-normalized image and the two edge images were used, 3x20x30 units were used as input units.

Figure 4.6 depicts the architecture of the neural network when both the histogram normalized image and the two edge images are used as input. On the bottom the three pre-processed input images are displayed. The network has the same architecture for estimating pan and tilt.

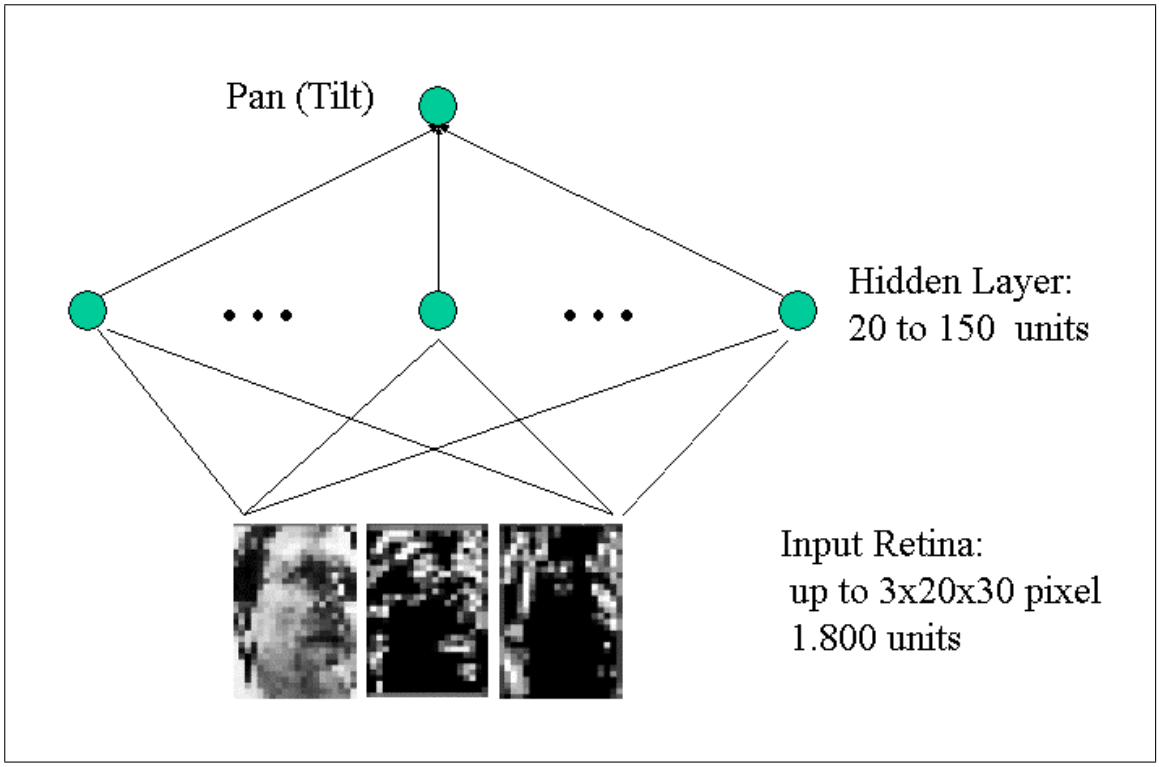


Figure 4.6: Neural network to estimate head pan (or tilt) from pre-processed facial images.

Output activations for pan and tilt were normalized to the range [0,1]. In case of the networks for pan estimation, an output activation of 0 corresponded to a head orientation of 90 degrees to the left and an output activation of 1 corresponded to head orientations of 90 degrees to the right. In case of the networks for tilt estimation, an output activation of 0 corresponded to looking down by 60 degrees and an output activation of 1 corresponded to looking up 60 degrees.

Training of the neural net was done using standard back-propagation.

4.4 Other Network Architectures

We have also experimented with two other neural network architectures to estimate head orientation from the facial images.

First, we tried to estimate head pan and tilt with one network. The network therefore had two output units, one for pan and one for tilt. The output activations for pan

and tilt were normalized to the range [0,1] as in the case with separate networks to estimate pan and tilt. With the networks to jointly estimate head pan and tilt we however achieved poorer estimation results than with the separate networks for pan and tilt estimation.

We also investigated neural networks to *classify* head orientation into different head orientation classes. The output layer of theses networks consisted of 19 units representing 19 different angles (-90, -80, ...,+80, +90 degrees). The output layer of the tilt estimating net consisted of 6 units representing the tilt angles +15, 0, -15, .. -60 degrees. For both nets we used a Gaussian output representation. With such an output representation not only the single correct output unit is activated during training, but also its neighbors receive some training activation decreasing with the distance from the correct label. On a multi-user test set, these networks performed slightly worse than the networks with one output unit. On two new users, the results were slightly better. A drawback of this network architecture is, however, the much higher number of network parameters (nineteen times higher for the pan estimation networks!) which significantly prolonged the time necessary for training the networks, while not leading to significantly better results.

We have therefore decided to use separate networks to estimate pan and tilt and to use networks with one output unit for pan and tilt estimation in this work. In the remainder of this thesis, we therefore only report experiments with such a network architecture.

4.5 Experiments and Results With Pan-Tilt-Zoom Camera Images

To evaluate the different preprocessing methods, we first trained networks on only the images that were taken in one room. The images of 12 users were divided into a training set consisting of 4,750 images, a cross-evaluation set of 600 images and a test set with a size of 600 images. The images of the remaining two users were kept aside to evaluate performance on new users whose images have not been in the training set at all.

Table 4.2 shows the results that we obtained on the twelve-user test set and on the new users using the different preprocessing approaches. Each cell of the table indicates the mean difference between the true pan (tilt) and the estimated pan (tilt) over the whole test set. Results are given in degrees.

It can be seen that the best results were obtained when using both the histogram normalized images and the edge images as input to the neural networks. On the

preprocessing	multi-user	new users
histogram	3.8 / 3.0	9.4 / 10.9
edges	4.6 / 3.6	10.1 / 9.9
histo + edges	**3.5 / 2.8**	**7.5 / 8.9**

Table 4.2: Head pose estimation accuracy from good resolution images on a multi-user test set and on two new users. Results for three different preprocessing methods are indicated: 1) using histogram-normalized images as input, 2) using edge images as input and 3) using both histogram-normalized and edge images as input. The results indicate the mean error in degrees for pan/tilt.

multi-user test set a mean error of 3.5 degrees for pan an 2.8 degrees for tilt was obtained. On new users the mean error was 7.5 degrees for pan and 8.9 degrees for tilt.

4.5.1 Error Analysis

Figure 4.7 shows histograms of the observed errors for pan and tilt estimation on the multi-user test set. For both pan and tilt estimation we see that the error histograms have a Gaussian shape with zero mean. The error histograms for pan and tilt on the new users are given in figure 4.8. On the new users we see the higher variance of the error histogram. We also see that the mean of the error histogram for pan estimation is slightly less than zero, which indicates a tendency of the pan estimation networks to underestimate head rotation on the new users.

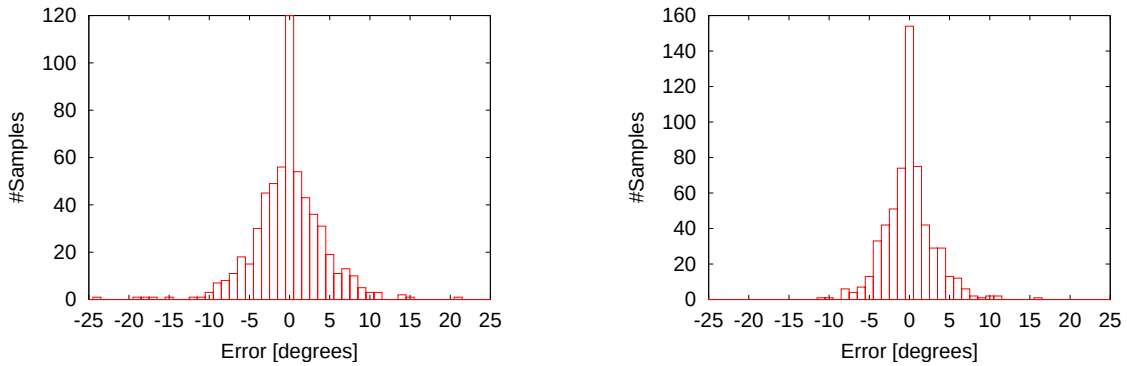


Figure 4.7: Error histograms for pan and tilt on the multi-user test set.

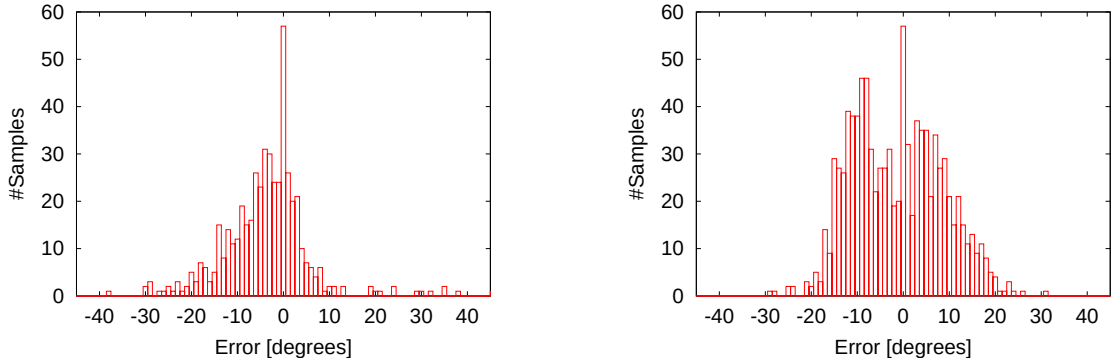


Figure 4.8: Error histograms for pan and tilt on the new users.

Figure 4.9 indicates the mean error for different horizontal head rotations. On the left, the errors for estimating pan on the multi-user test set is depicted; on the right the errors on new users are shown. It can be seen that the average errors are quite similar for head rotations between -60 and +60 degrees. For large head rotations, however, the average errors significantly increase. This is probably due to the distribution of head rotation examples in the collected data set (see Figure 4.3). Since only a small amount of training data was available for large head rotations, the errors are higher for such rotations.

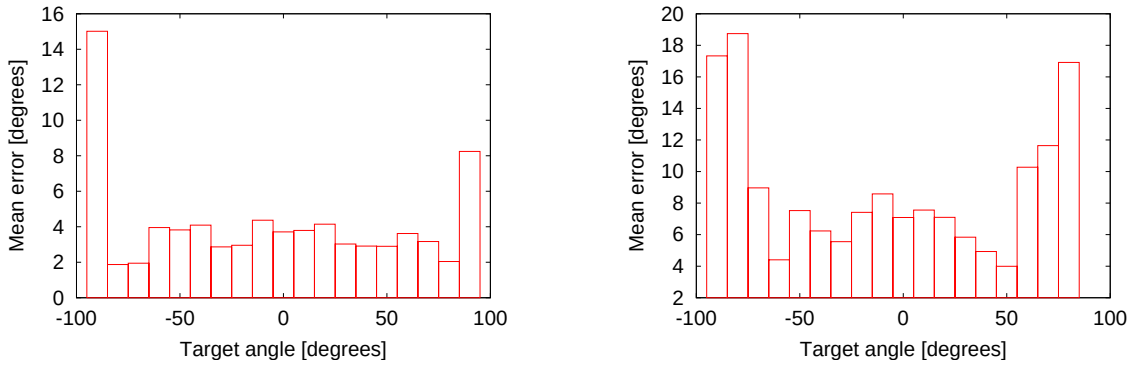


Figure 4.9: Mean errors for different target angles on the multi-user test set and on new users.

4.5.2 Generalization to Different Illumination

We also evaluated the performance of the neural networks on the data collected in the second room and for networks that were trained on images from both rooms. Since the above experiments showed that using the histogram-normalized image and the two edge images together as input worked the best, we only present the results for this preprocessing method here. Table 4.3 summarizes the results.

For training and evaluating the networks on images from both rooms, around 9900 images from 17 different users were used for training. Both the cross-evaluation set and the test-set contained around 1240 images from the same users. User-independent performance was evaluated on 1960 images from two users.

The performance of the network on images from both rooms is approximately the average of the performances obtained on images of each room. This result shows that the network is able to handle some amount of illumination variation, as long as images from both rooms have been in the training set.

Data	multi-user	new users
Room 1	3.5 / 2.6	7.5 / 8.9
Room 2	4.0 / 3.6	6.3 / 12.8
Room 1+2	3.8 / 3.2	7.1 / 9.5

Table 4.3: Average error in estimating head pan and tilt for two "room-dependent" networks and for a network trained on images from two rooms.

To evaluate how different lighting conditions affect the performance of the neural networks, we trained neural networks with images taken only in one room and evaluated the performance on images taken in another room. The first room we collected data in had no windows, but head several light sources on the ceiling. The second room had several windows along two sides of the room, no artificial light was present.

Table 4.5 shows the results that we obtained using these "room-dependent" nets when testing on images from the same room versus testing with images from another room. All the results are user-independent; i.e., no images of the subjects in the test set were present in the training set.

It can be seen that the accuracy of pose estimation decreases when testing the nets on images that were taken under different lighting conditions than during training.

When using images from both rooms for training, however, the pose estimation results remain stable. On the user-independent test set with images from both rooms, for

Training Data	Test Data	E_{pan}	E_{tilt}
Room 1	Room 1	3.5	2.6
Room 2	Room 2	4.0	3.6
Room 1	Room 2	16.8	18.1
Room 2	Room 1	13.9	12.1
Room 1+2	Room 1+2	3.8	3.2

Table 4.4: Results on multi-user test sets, obtained when training and testing on images taken under different lighting conditions. Both histogram-normalized gray-scale image and edge images were used together as input to the nets.

Training Data	Test Data	E_{pan}	E_{tilt}
Room 1	Room 1	7.5	8.9
Room 2	Room 2	6.3	12.8
Room 1	Room 2	13.0	11.4
Room 2	Room 1	15.8	10.9
Room 1+2	Room 1+2	7.1	9.5

Table 4.5: User-independent results obtained when training and testing on images taken under different lighting conditions. Both histogram-normalized gray-scale image and edge images were used together as input to the nets.

instance, an average error of 7.1 degrees for pan estimation and 9.5 degrees for tilt estimation was obtained. This is approximately as good as the average performance obtained on the user-independent test with neural networks that were trained for each of those rooms.

In the results reported here, cross-evaluation sets with images taken in the same room as the images in the training set were used in order to determine when to stop training. Better generalization on images from new rooms can be achieved when a cross-evaluation set with images from the new room is used. We have trained a neural network to estimate pan with the 4754 images from "Room 1" in the training set and used a cross-evaluation set of 648 images from "Room 2" to determine when to stop training. On the test set from "Room 2" the measured average error for estimating head pan then 15.8 degrees, as compared to 16.8 degrees when using images from "Room 1" for cross-evaluation.

Further generalization improvement to a new room could be achieved by also using artificial training images. We artificially mirrored all the images (and the correspond-

ing head orientations) of the training images from "Room1". After training networks with these additional artificial training samples from "Room1", an average error of 15.4 degrees on images from "Room 2" were measured. By furthermore using images from "Room 2" for cross-evaluation, the measured accuracy on images from "Room 2" could be increased to 14.1 degrees. Compared to the initial result of 16.8 degrees average error, this signifies an error reduction of 16%. Table 4.6 summarizes these results.

Training Set	Cross-Evaluation Set	Test Set	E_{pan}
Room 1	Room 1	Room 2	16.8
Room 1	Room 2	Room 2	15.8
Room 1 + mirrored images	Room 1	Room 2	15.4
Room 1 mirrored images	Room 2	Room 2	14.1

Table 4.6: Pan estimation results when training with images from one room and testing on images from another room with different illumination. By using some sample images from the new room for cross-evaluation, generalization is improved. Further improvement could be obtained by also using artificially mirrored training images.

The achieved pan estimation error of 14.1 degrees on images taken in a new room is however still more than twice as high than the average pan estimation error that was achieved on new users, when images from both rooms were available for training (see Table 4.5).

In order to further improve the performance of the neural networks under new illumination conditions we have therefore investigated how the networks can be adapted using training images that are collected in the new location. These experiments will be discussed in Chapter 8.

4.5.3 A Control-Experiment to Show the Usefulness of Edge Features

Our experimental results showed that using histogram-normalized greyscale images leads to better estimation results than using edge images as input. The best results, however, were obtained using both histogram-normalized and edge images as input.

One possibility for the increased accuracy when using both histogram-normalized and edge images as input could be the higher number of parameters of the neural network which is due to three times higher number of input units.

To verify that the increased performance is not only due to the increase of parameters of the network, we have also trained and evaluated networks for estimating head pan with a similar number of input units, but only using histogram-normalized images as input.

To obtain approximately the same number of input units as when using the histogram and edge images as input, we used histogram-normalized images of size 36x54 pixels as input images. Using this image size the aspect ratio of 2:3 of the facial images is preserved and an input retina with 1944 input units is obtained, which is slightly bigger than the input retina of 1800 units, which is used when using histogram-normalized and edge images of size 20x30 pixels.

As it turned out, the best average error for head pan estimation was 3.3 degrees, which is comparable to the results obtained when also using edge images as input. On new users, however, only an average error of 8.6 degrees for pan estimation could be achieved, which is significantly worse than the results obtained when also using edge images as input.

preprocessing	image size	input units	multi-user	new users
histogram	1 x 36 x 54	1944	3.3	8.6
histo + edges	3 x 20 x 30	1800	3.5	7.5

Table 4.7: Results for pan estimation using only histogram normalized images of size 36x54 pixels or using both histogram normalized and edge images of size 20x30 pixels as input.

Similar results were measured when these networks were evaluated on images from another room. When evaluating the networks that were trained with images from "Room1" on a test set with images that were taken in "Room2" the accuracy decreased to 20.1 degrees average error. When edge images were also used as input, the resulting average error for pan estimating on this test set was 16.8 degrees. This is of course much better than the average error of 20.1 degrees obtained without using edge images as additional input.

We therefore conclude that edges are useful features for head pose estimation. The experimental results show that using edge images in addition to the histogram-normalized greyscale images improved the ability of the neural networks to generalize both to new users as well as to images that were taken under different illumination conditions than the images used for training.

Net Input	Training Set	Test Set
Gray-scale	6.6 / 5.0	9.4 / 6.9
Edges	6.0 / 2.6	10.8 / 7.1
Edges + Gray-scale	1.4 / 1.5	**7.8 / 5.4**

Table 4.8: Multi-user results. The mean error in degrees of pan/tilt is shown. Three different types of input images were used. Training was done on twelve users, testing on different images from the same twelve users.

4.6 Experiments and Results With Images From the Omni-Directional Camera

In case of the panoramic image data, we divided the data set of twelve users (of the fourteen users in the whole data set) into a training set consisting of 6080 images, a cross-evaluation set of size 760 images and a test set with a size of 760 images. The images of the remaining two users were kept as a user independent test set.

As input to the neural nets, again three different preprocessing approaches were investigated: using histogram normalized gray-scale images as input, using horizontal and vertical edge images as input and using both normalized gray-scale plus the edge images as input.

Again, we trained the neural networks on the training data set and used the cross-evaluation set to determine when to stop training. The performance of the networks was then evaluated on the test set containing images of the twelve persons that were also in the training set (multi-user case). On the multi-user test set, we obtained the best performance using both normalized gray-scale images and edge images as input. A mean error of 7.8 degrees for pan and 5.4 degrees for tilt was obtained with the best nets. Using only the gray-scale images as input, the results decreased to 9.4 degrees for pan and 6.9 degrees for tilt. With edge images as input, only 10.8 degrees for pan and 7.1 degrees for tilt could be achieved. Table 4.8 summarizes these results together with the accuracies on the corresponding training sets.

To determine how well the neural nets can generalize to new users, we have also evaluated the networks on the two users which have not been in the training set. On the two new users the best result for pan estimation, which was 9.9 degrees mean error, was obtained using normalized gray-scale images plus edge images as input. The best result for tilt-estimation measured was 9.1 degrees mean error and was obtained using only normalized grey-scale images as input. Table 4.9 summarizes the results.

Net Input	Training Set	New Users
Gray-scale	6.6 / 5.0	11.3 / **9.1**
Edges	6.0 / 2.6	13.3 / 10.8
Edges + Gray-scale	1.4 / 1.5	**9.9** / 10.3

Table 4.9: User independent results. The mean error in degrees of pan/tilt is shown. Three different types of input images were used. Training was done on twelve users, testing two new persons.

Net Input	Multi-user Test Set	New Users
Gray-scale	5.5 / 4.1	10.4 / **9.3**
Edges	5.6 / 3.5	12.2 / 10.3
Edges + Gray-scale	**3.1 / 2.5**	**9.5** / 9.8

Table 4.10: Results using additional artificial training data. Results on the multi-user test set and on the two new users are shown for the different preprocessing approaches. The mean error in degrees of pan/tilt is shown.

4.6.1 Adding Artificial Training Data

In order to obtain additional training data, we have artificially mirrored all of the images in the training set, as well as the labels for head pan. As a result, the available amount of data could be doubled without having the effort of additional data collection. Having more training data should especially be helpful in order to get better generalization on images from new, unseen users. Indeed, after training with the additional data, we achieved an average error of only 9.5 degrees for pan and 9.8 degrees for tilt on the two new users. On the multi-user test set the accuracy even was doubled to 3.1 degrees for pan and 2.5 degrees for tilt. Table 4.10 shows the results on the multi-user test set, as well as the new user test set for the different preprocessing approaches.

4.6.2 Comparison

It is interesting to compare the head pose estimation result obtained when using the images from the omni-directional camera as input to the networks with those results obtained with facial images of higher resolution obtained from a pan-tilt-zoom camera. Table 4.11 summarizes the best pose estimation results we obtained with the two kind of input images. It can be seen that on a multi-user test set, both

image types lead to similar results. However, on new users, the mean error for head pan and tilt estimation is higher with images from the omni-directional camera.

Camera Type	Multi-User	New Users
Pan-Tilt-Zoom	3.5 / 2.6	7.1 / 8.3
Omni-directional	3.1 / 2.5	9.5 / 9.8

Table 4.11: Results obtained with good resolution facial images captured with a pan-tilt-zoom camera and results with facial images obtained from the omni-directional camera. The mean difference from true head rotation in degrees is indicated.

Chapter 5

From Head Orientation to Focus of Attention

In this chapter we present a probabilistic approach to detect a person's focus of attention target based on his or her head orientation. We discuss details of the model, how the model parameters can be adapted to different numbers and locations of targets, and we present experimental results on a number of meetings that we recorded in our lab.

In our approach we first estimate a person's head orientation – as described in the previous chapter – and then estimate at whom a person was looking, based on his or her estimated head orientation. Since head tilt is not needed to determine at which of several participants around a table someone is looking, we only use a person's head pan to determine at whom he is looking.

Compared to directly classifying a person's focus of attention target – based on images of the person's face for example – our approach has the advantage that different numbers and positions of participants in the meeting can be handled. If the problem was treated as a multi-class classification problem, and a classifier such as a neural network was trained to directly learn the focus of attention target from the facial images of a user, then the number of possible focus targets would have to be known in advance. Furthermore, with such an approach it would be difficult to handle situations where participants sit at different locations than they were sitting during collection of the training data.

A first solution to find out at whom a person S is looking could be to use the measured head pose of S and look which target person T_i sits nearest the position to which S is looking. Gaze is, however, not only determined by head pose, but also by the

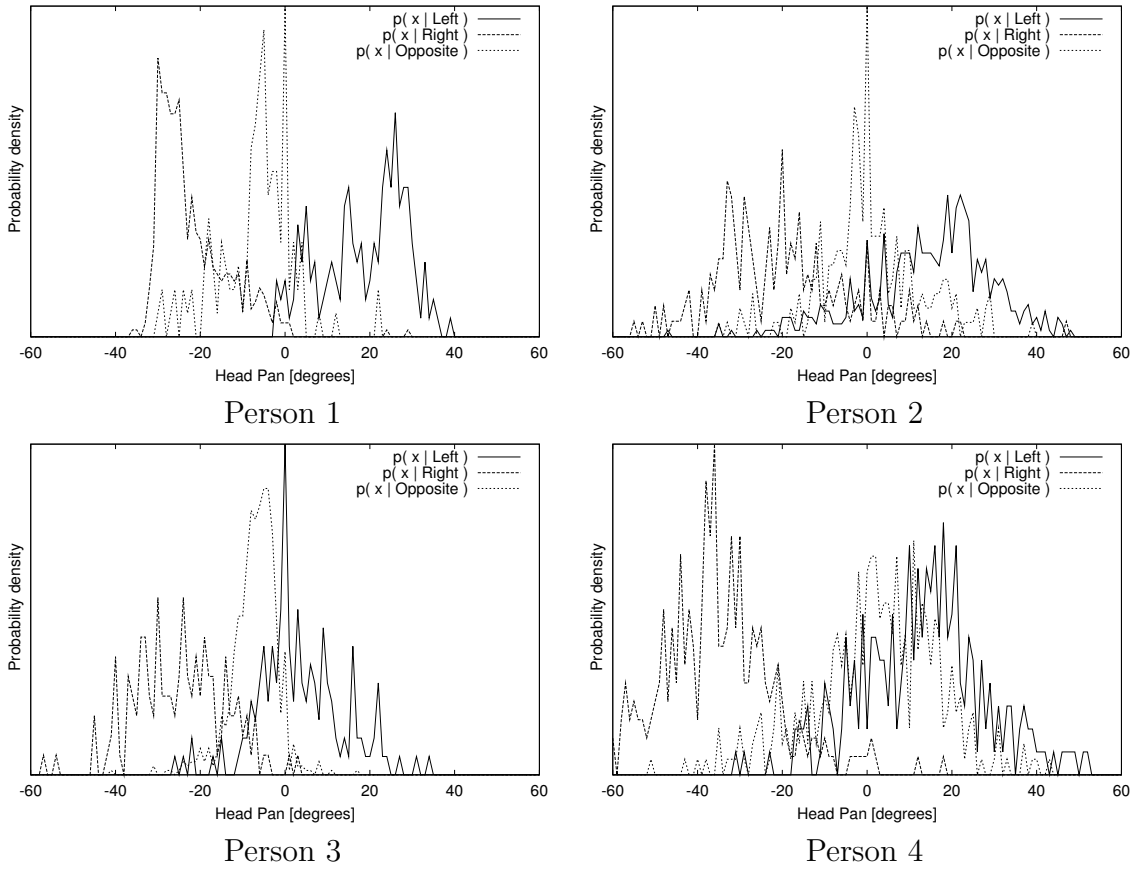


Figure 5.1: Class-conditional head pan distributions of four persons in a meeting when looking to the person to their left, to their right or to the person sitting opposite. Head orientations were estimated using a neural network.

direction of eye gaze. People do not always completely turn their heads toward the person at which they are looking. Instead, they also use their eye gaze direction.

We have therefore developed a Bayesian approach to estimate at which target a person is looking, based on his observed head orientation. More precisely, we wish to find $P(\text{Focus}_S = T_i|x_S)$, the probability that a subject S is looking towards a certain target person T_i, given the subject's observed horizontal head orientation x_S, which is the output of the neural network for head pan estimation. Using Bayes formula, this can of be decomposed into

$$P(\text{Focus}_S = T_i|x_S) = \frac{p(x_S|\text{Focus}_S = T_i)P(\text{Focus}_S = T_i)}{p(x_S)}, \tag{5.1}$$

where x_s denotes the head pan of person S in degrees and T_i is one of the other persons around the table.

Using this framework, given a pan observation x_s for a subject S – as estimated by the neural network for head pan estimation – it is then possible to compute the posterior probabilities $P(\text{Focus}_S = T_i|x_S)$ for all targets T_i and choose the one with highest posterior probability as the subject's focus of attention target in the current frame.

In order to compute $P(\text{Focus}_S = T|x_S)$, it is necessary, however, to estimate the class-conditional probability density function $p(x_S|\text{Focus}_S = T)$, the class prior $P(\text{Focus}_S = T)$ and $p(x_S)$ for each person. Finding $p(x_S)$ is trivial and can be done by just building a histogram of the observed head orientations of a person over time.

One possibility to find the class-conditional probability density function and the prior would be to adjust them on a training set of similar meetings. This, however, would require training data for any possible number of participants at the table and for any possible combination of the participants' locations around the table. Furthermore, adapting on different meetings and different persons would probably not model a certain person's head turning style very well, nor would the priors necessarily be the same in different meetings. In our meeting recordings we observed, for instance, that some people turned their head more than others and some people made stronger use of their eye-gaze and turned their head less when looking at other people. Figure 5.1 shows the head pan distributions of four participants in one of our recorded meetings. The head orientation of the user was estimated with the neural nets. It can be seen, for example, that for Person 1, the three class-conditionals are well separated, whereas for Person 3 or Person 4, the peaks of some distributions are much closer to each other, and a higher overlap of the distributions can be observed.

In order to adapt the parameters of our model to varying target locations and to the different head turning styles of the participants, we have developed an unsupervised learning approach to find the head pan distributions of each participant when looking at the others.

5.1 Unsupervised Adaptation of Model Parameters

In our approach, we assume that the class-conditional head pan distributions, such as depicted in Figure 5.1, can be modeled as Gaussian distributions. Then, the distribution $p(x)$ of all head pan observations from a person will result in a mixture of Gaussians,

$$p(x) \approx \sum_{j=1}^{M} p(x|j)P(j), \tag{5.2}$$

where the individual component densities $p(x|j)$ are given by Gaussian distributions $N_j(\mu_j, \sigma_j^2)$.

In our approach, the number of Gaussians M is set to the number of other participants at the table, because we assume that these are the most likely targets that the person has looked at during the meeting, and because we want to find the individual Gaussian components that correspond to looking at these target persons. This parameter can automatically set to the number of faces detected around the table.

The model parameters of the mixture model can then be adapted so as to maximize the likelihood of the pan observations given the mixture model. This is done using the expectation-maximization algorithm by iteratively updating the parameter values using the following update equations [Bishop '95]:

$$\mu_j^{new} = \frac{\sum_n P^{old}(j|x^n)x^n}{\sum_n P^{old}(j|x^n)} \tag{5.3}$$

$$(\sigma_j^{new})^2 = \frac{1}{d}\frac{\sum_n P^{old}(j|x^n)||x^n - \mu_j^{new}||^2}{\sum_n P^{old}(j|x^n)} \tag{5.4}$$

$$P(j)^{new} = \frac{1}{N}\sum_n P^{old}(j|x^n). \tag{5.5}$$

To initialize the means μ_j of the mixture model, k-means clustering was performed on the pan observations. Adaptation of the parameters was stopped after a maximum of 100 iterations. In a few rare cases, we observed that two means of the mixture model moved very closely together during the adaptation process. To prevent this, we also stopped adaptation when the difference between two means μ_i and μ_j became less than ten degrees.

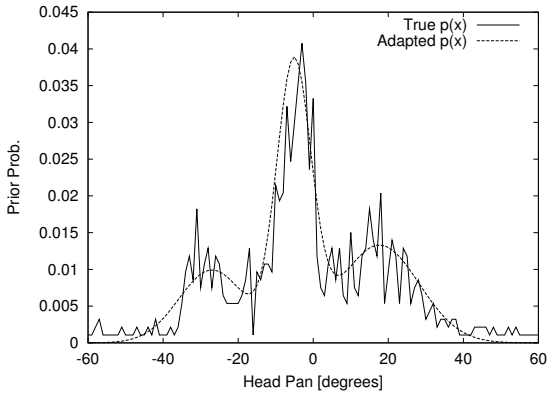
After adaptation of the mixture model to the data, we use the individual Gaussian components of the mixture model as an approximation of the class-conditionals $p(x|\text{Focus} = T)$ of our focus of attention model described in equation (5.1). We furthermore use the priors of the mixture model, $P(j)$, as the focus priors $P(\text{Focus} = T)$. To assign the individual Gaussian components and the priors to their corresponding target persons, the relative position of the participants around the table can be used.

Figure 5.2 shows an example of the adaptation on pan observations from one user. In Figure 5.2(A) the distribution of all head pan observations of the user is depicted together with the Gaussian mixture that was adapted as described above. Figure 5.2(B) depicts the real class-conditional head pan distributions of that person, together with the Gaussian components taken from the Gaussian mixture model depicted in Figure 5.2(A). As can be seen, the Gaussian components provide a good approximation of the real class-conditional distributions of the person. Note that the real class-conditional distributions are just depicted for comparison and are of course not necessary for the adaptation of the Gaussian components. Figure 5.2(C) depicts the posterior probability distribution resulting from the adapted class-conditionals and class priors.

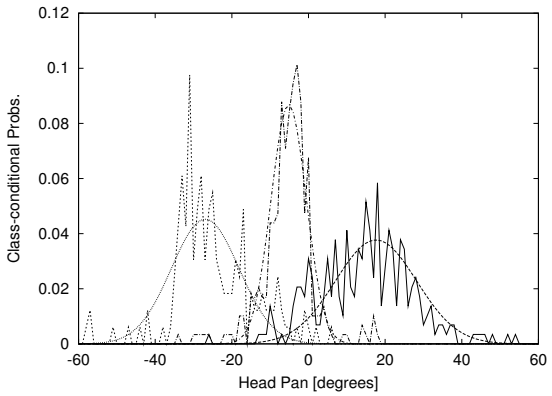
5.2 Experimental Results

To evaluate our approach, several meetings were recorded. In each of the meetings four or five participants were sitting around a table and were discussing a freely chosen topic. Video was captured with the panoramic camera. Each participant had a microphone in front of him so that his speech could be recorded. Using this setup, audio streams for each of the participants plus the panoramic view of the scene could be simultaneously recorded to hard-disk. A typical panoramic view of a recorded meeting is shown in Figure 5.3.

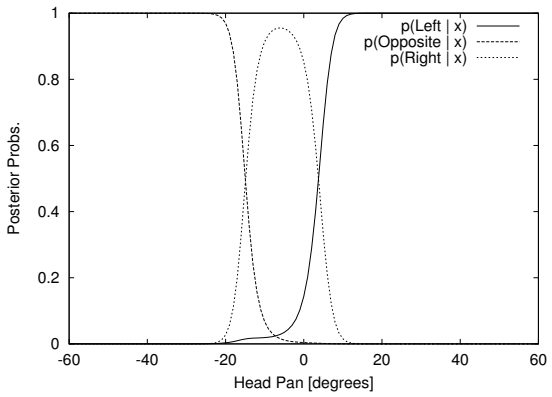
To record the audio- and video streams, a software tool for simultaneous time-aligned capturing of video and up to eight audio-streams was developed. Each audio stream was recorded with a sampling rate of 16kHz. Time-stamps for each video frame were recorded for later synchronization of the audio streams and the video images.



(A)



(B)



(C)

Figure 5.2: a) The distribution $p(x)$ of all head pan observations of one subject in a meeting. Also the adapted mixture of three Gaussians is plotted. b) True and estimated class-conditional distributions of head pan x for the same subject, when he or she is looking to three different targets. The adapted Gaussians, are taken from the adapted Gaussian mixture model depicted in a). c) The posterior probability distributions $P(\text{Focus}|x)$ resulting from the found mixture of Gaussians.

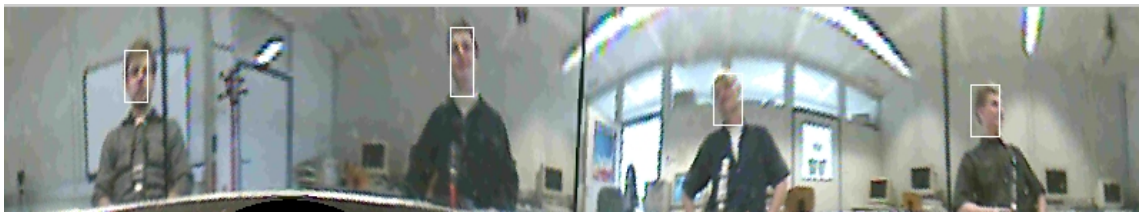


Figure 5.3: A typical meeting scene captured with the panoramic camera.

Meeting	#participants	duration	#frames
A	4	8 min 34 s	1280
B	4	7 min 8 s	1015
C	4	5 min 34 s	874
D	4	4 min 20 s	767
E	5	12 min 52 s	2301
F	5	6 min 16 s	1104
Sum		29 min 2 s	7341

Table 5.1: Overview of the recorded meetings used for evaluation.

.

Since uncompressed video was directly written to hard-disk at a resolution of 640x480 pixels, video could only be captured at a frame-rate of 2-3 images per second.

Altogether, six short meetings were recorded. In four of the meetings, four persons participated and in two of the meetings five participants had joined. The recorded meetings lasted from 5 minutes and 30 seconds to 12 minutes and 50 seconds and contained between 870 to 2300 video frames. Table 5.1 shows the durations and the number of participants in each of the meetings used for evaluation.

Figure 5.4 indicates the locations of the participants around the table in each of the recorded meetings. In meetings A to C the participants were seated symmetrically at each side of the table. In meeting D, the participants were not symmetrically seated. In meeting E and F, five people were seated around the table as depicted.

In each frame of the recorded meetings, we manually labeled at whom each participants was looking. Labeling of the frames was done by looking at the panoramic view of the meeting scene and by looking at the perspective views of each of the tracked persons (see Figures 5.3 and 5.5). In case of the meetings with four participants, these labels could be one of *"Left"*, *"Right"* or *"Straight"*, meaning a person was looking to the person to his left, to his right, or to the person at the opposite. If the

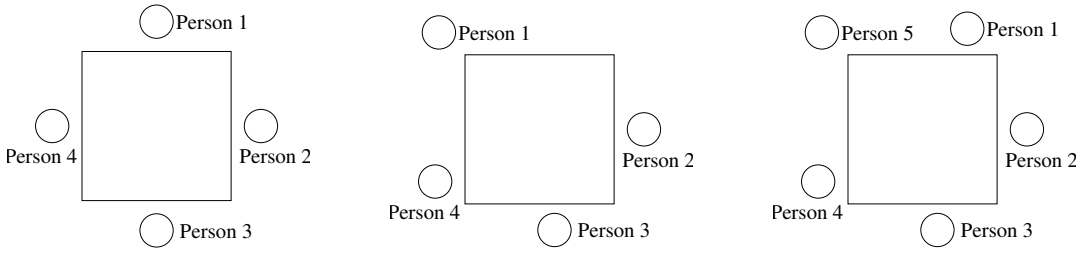


Figure 5.4: Approximate locations of the participants around the table in the recorded meetings (viewed from top).

Figure 5.5: Perspective views of two participants. These views were used together with the panoramic view of the scene to label at whom a participant was looking.

person wasn't looking at one of these targets, e.g., the person was looking down on the table or was staring up to the ceiling, the label *"Other"* was assigned. In case of the meetings with five participants appropriate labels to label the other four target persons were chosen.

In addition, labels indicating whether a person was speaking or not were manually assigned for each participant and each video frame. These labels were assigned by listening to the audio streams of each participant and using the time-stamps that were captured with the video stream.

We have evaluated this approach on the evaluation meetings. In each meeting, the faces of the participants were automatically tracked, and head pan was computed using the neural network-based system to estimate head orientation described in Chapter 4.

Then, for each of the participants in each meeting, the class-conditional head pan distribution $p(x|\text{Focus})$, the class-priors $P(\text{Focus})$ and the observation distributions $p(x)$ were automatically adapted as described in the previous section, and the posterior

probabilities $P(\text{Focus} = T_i|x)$ for each person were computed. During evaluation, in each frame the target with the highest posterior probability was then chosen as the focus of attention target of the person.

For the evaluation, we manually marked frames where a subject's face was occluded or where the face was not correctly tracked. These frames were not used for evaluation. Face occlusion occurred in 1.6% of the captured images. Occlusion sometimes happened when a user covered his face with his arms or with a coffee mug for example; sometimes a face was occluded by one of the posts of the camera. In another 4.2% of the frames the face was not correctly tracked. We also did not use frames where a subject did not look at one of the other persons at the table. This happened in 3.8% of the frames. Overall 8.2% of the frames were not used for evaluation since at least one of the above indications was given.

5.2.1 Meetings With Four Participants

In the meetings with four participants, the correct focus target could be detected on average in 72.9% of the frames. This result was obtained by comparing each of the computed focus targets of each participant with the manually obtained labels. Table 5.2 shows the average results on the three meetings. In the table, the average accuracy on the four participants in each meeting is indicated. The focus detection accuracy for the individual participants ranged from 68.8% to 79.5%.

	P(Focus|Head Pan)
Meeting A (4 participants)	68.8%
Meeting B (4 participants)	73.4%
Meeting C (4 participants)	79.5%
Meeting D (4 participants)	69.8%
Average	72.9%

Table 5.2: Percentage of correctly assigned focus targets based on computing P(Focus|head pan) in meetings with four participants.

While in meetings A to C the four participants were seated symmetrically around the table; i.e., each of the participants was seated exactly in the middle of one side of the table, in meeting D the participants were not evenly distributed around the table. The results show, that our algorithm performs similarly well for different seating arrangements of the participants.

5.2.2 Meetings With Five Participants

Table 5.3 summarizes the evaluation results that could be obtained on the two meetings with five participants. On average the correct focus target of person could be detected in 52.5% of the time.

	P(Focus|Head Pan)
Meeting E (5 participants)	51.9%
Meeting F (5 participants)	53.0%
Average	52.5%

Table 5.3: Percentage of correctly assigned focus targets based on computing P(Focus|head pan) in the meetings with five participants.

As could be expected, this result is worse than the results obtained with four participants. However, the detection accuracy is of course much better than chance, which for the four possible targets would be 25%. An obvious reason for the decreased result is that five people are sitting closer together at the same table as do four people. Therefore, distinguishing at whom someone was looking at based on his head orientation becomes more difficult.

5.2.3 Upper Performance Limits Given Neural Network Outputs

When looking at the class-conditional head pan distribution $p(x|\text{Focus})$ of the participants, such as depicted in Figure 5.1 for example, it is clear that there is some overlap of the distributions.

The overlap in the distribution is due to the subject's head turning behavior, due to the noisy neural network based pan estimation and also dependends on the number of target persons at the table.

By only relying on head pose estimation to detect the focus target of a person, the possible accuracy of the approach is limited by the overlap of these class-conditional distributions.

To evaluate the upper-limit of our approach to detect the focus target based on the estimated head orientations, we have used the true class-conditional distributions of estimated head pan in order to compute each person's posterior probabilities

$p(\text{Focus} = T_i|x)$ and to determine the focus targets. These "true" class-conditionals can be found by looking at the estimated head rotations of a subject, when the subject was known to look at a specific target person.

Upper limit with four participants

	P(Focus|Head Pan)
Meeting A	75.1%
Meeting B	79.5%
Meeting C	81.1%
Meeting D	75.7%
Average	77.9%

Table 5.4: Upper performance limits of focus of attention detection, given estimated head orientations. The percentage of correctly assigned focus targets using true class-conditionals of estimated head pan are indicated. Four subjects participated in each meeting.

Table 5.4 summarizes the results of the two meetings. On average, in 77.9% of the frames the focus target could be correctly determined. This result indicates the accuracy we could obtain, if we knew the true class-conditional distributions of each person's estimated head pan. By estimating the class-conditionals using the unsupervised adaptation approach described above, we obtained an average accuracy of 72.9% on the three meetings, which is 94% of the optimum performance.

Upper limit on Meetings with five people

Table 5.5 summarizes the corresponding baseline results on the meetings with five participants. On average in 67% of the time the correct focus target can be determined based on the estimated head orientations. Since there is a higher overlap of the class-conditional head orientation distributions for four targets than for only three target persons, the possible accuracy is about 11% lower than for three target persons. Figure 5.6 shows an example of the class-conditional distributions of a participant of Meeting E.

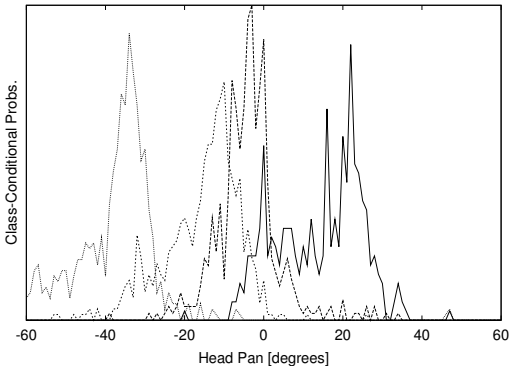


Figure 5.6: Class-conditional distributions of horizontal head rotations of one subject, when he or she is looking at four target persons at the table. A high overlap of the distributions can be observed.

	P(Focus|Head Pan)
Meeting E	68.4%
Meeting F	65.6%
Average	67.0%

Table 5.5: Upper performance limits of focus of attention detection from estimated head orientations with five meeting participants. Percentage of correctly assigned focus targets using true class-conditionals of estimated head pan are indicated.

5.3 Panoramic Images Versus High-Resolution Images

In Section 4.6.2 we demonstrated that slightly better head pose estimation results on new users could be obtained with images captured with a pan-tilt-zoom camera. The best result for head pan estimation with neural networks trained on the omni-directional camera images was 9.5 degrees mean error, whereas with images from the standard camera 7.1 degrees error could be achieved.

We have also evaluated how much the different types of input images affect the accuracy of head pose based focus estimation. In order to do this, we have captured four meetings both with the panoramic camera and with one additional pan-tilt-zoom camera facing one person. For those four persons that were captured with both, the omni-directional and the pan-tilt-zoom camera, focus estimation based on the two

type of facial images could be compared.

Upper performance limit

When the true, manually determined, class-conditional distributions of each person's estimated head pan was used for focus estimation, the correct focus of attention based on images from the omni-directional camera could be estimated for these four persons in 78.9% of the frames. When using the higher resolution facial images, and the neural networks trained for these images, in 79.5% of the frames focus could be detected correctly.

Results with unsupervised adaptation

When the class-conditional head pan distributions were adapted using k-means clustering and the EM-approach as described above, on average in 74.9% of the frames the correct focus target could be detected on the basis of the images from the omni-directional camera. With the facial images of higher resolution the obtained accuracy was 75.5%.

Discussion

This experiments shows that at least when the participants are not too far away from the omni-directional camera, as in our experiment, focus of attention of the participants can be estimated from the omni-directional images with the same accuracy as when using separate cameras to capture each user.

When using the omni-directional camera, however, we expect that head pose estimation results might decrease when a bigger conference table is used and the participants of the meeting sit further away from the camera. Consequently focus of attention detection will be less accurate. Since with a pan-tilt-zoom camera, the size of the face can be kept constant within the image, no matter how far people sit away from the camera, no decrease of focus of attention detection is expected for those cameras.

5.4 Summary

This chapter presented our probabilistic approach to determine focus targets based on observed head orientations. In our approach, a subject's head orientation when

looking at other target is modeled as a mixture of Gaussians. We have demonstrated how the model parameters can automatically be adapted when the number of targets is known. We have demonstrated experiments on meetings with four and five participants. The approach has proven to be able to adapt as well to different seating arrangements as to different numbers of participants. We have furthermore investigated the upper limit of focus of attention detection accuracy, when the focus is determined based on head pan estimations given by the neural nets.

Chapter 6

Head Pose versus Eye-Gaze

In this work, head orientation is used to predict a person's focus of attention in meetings. This is done because head orientation is assumed to be a very reliable indicator of the direction of someone's attention during social interaction, as has been discussed in Chapter 2, and because eye gaze of several meeting participants cannot be easily tracked without the use of intrusive hardware.

Since we estimate where a person is looking based on his head orientation, the following question suggests itself: how well can we predict at whom a person is looking, merely on the basis of his or her head orientation?

To answer this question, we have analyzed the gaze of four people in meetings using special hardware equipment to measure their eye gaze and head orientation [Stiefelhagen & Zhu 2002]. We have then analyzed the gaze and head orientation data of the four people to answer the following questions:

1. How much does head orientation contribute to gaze?
2. How good can we predict at whom the person was looking, based on his head orientation only?

6.1 Data Collection

The setting in this experiment is a round-table meeting. There are four participants in the meeting, and a session of data for about ten minutes with each participant is collected. In every session, one of the participants, the subject, wears a head-mounted

(a)

(b)

Figure 6.1: a) Datacollection with eye and head tracking system during a meeting. b) A participant wearing the head-mounted eye and head tracking system.

gaze tracking system from Iscan Inc. [ISC.]. The system uses a magnetic pose and position tracking subsystem to track the subject's head position and orientation. Another subsystem uses a head-mounted camera to capture images of the subject's eye. Software provided with this system can estimate and record the following data with a frame rate of 60 Hz: the subject's head position, head orientation, eye orientation, pupil diameter, and the overall gaze (line of sight) direction. All these estimations have a precision of better than one degree, which is far beyond the capability of any current non-intrusive tracking methods.

Figure 6.1 (a) shows an image taken during data collection. Note that the second person from the right in the image is wearing the head-mounted gaze-tracker. Figure 6.1 (b) shows a participant wearing the tracking head gear.

A plot of some data captured from one subject is depicted in Figure 6.2. In the figure, the horizontal head orientation, eye orientation, and overall gaze direction over time are shown. Figure 6.3 shows a schematic view of a subject's head orientation, eye orientation and gaze direction. Gaze direction *los* is the sum of head orientation *ho* and eye orientation *eo*.

6.2 Contribution of Head Orientation to Gaze

First, we analyzed the contribution of head orientation and eye orientation to the overall gaze direction along the horizontal axis. On the data from the four participants

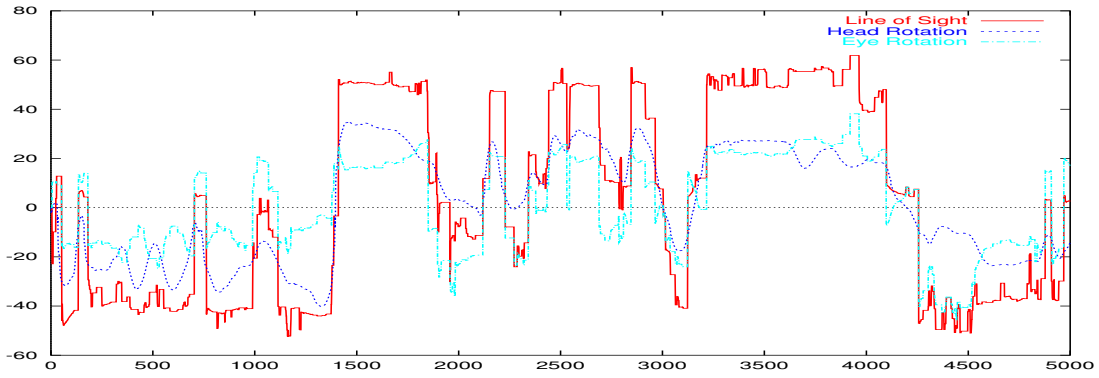


Figure 6.2: Plot of a subjects horizontal head orientation, eye orientation and overall gaze direction in a meeting. Eye orientation is measured relative to head orientation; i.e., the eye orientation within the eye sockets is indicated. The data was captured using an gaze tracking system from Iscan Inc [ISC.].

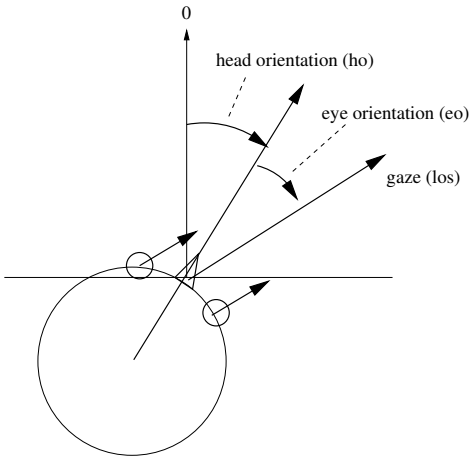


Figure 6.3: Schematic view of head oriention *ho*, eye orientation *eo* and gaze direction *los* of a subject.

Subject	#frames	eye blinks	same direction	head contribution
1	36003	25.4%	83.0%	62.0%
2	35994	22.6%	80.2%	53.0%
3	38071	19.2%	91.9%	63.9%
4	35991	19.5%	92.9%	96.7%
Average		21.7%	87.0%	68.9%

Table 6.1: Eyeblinks and contribution of head orientation to the overall gaze.

.

we found that in 87% of the frames head orientation and eye gaze pointed in the same direction (left or right). In the remaining 13% of the frames, the head orientation is opposite to eye orientation. For the frames in which head orientation and eye gaze point to the same direction, we calculated the contribution of head orientation to the overall line of sight orientation. Since the horizontal component of the line of sight los_x is the sum of horizontal head orientation ho_x and horizontal eye orientation eo_x, the percentage of head orientation to the horizontal direction of gaze is computed as:

$$\text{head contribution} = \frac{ho_x}{los_x}.$$

Table 6.1 summarizes the results of four experiment sessions. From the results, we can see several interesting points:

1. Most of the time, the subjects rotate their heads and eyes in the same direction to look at their focus of attention target (87%).

2. The subjects vary much in their usage of head orientation to change gaze direction: from Subject 2's 53% to Subject 4's 96%, with an average of 68.9%.

3. Even for Subject 2, whose head contribution is the least among the four participants, head orientation still contributes more than half of the overall gaze direction.

4. Eye-blinks (or eye-tracking failures) take about 20% of the frames, which means even for commercial equipments as accurate as the ISCAN system we used, eye orientation, and thus the overall gaze direction, cannot be obtained about a fifth of the time.

From these results it can be concluded that head orientation is the most important and sometimes the only measure to estimate a person's direction of gaze.

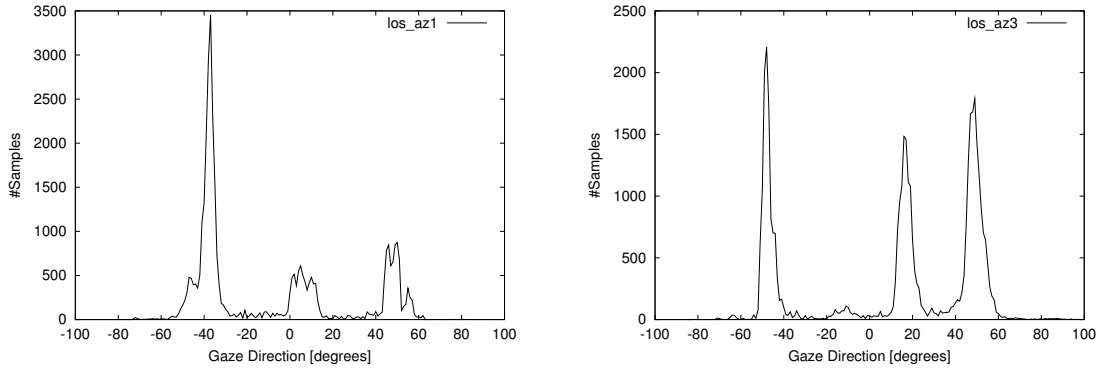


Figure 6.4: Histograms of horizontal gaze directions of two subjects. For both subjects three peaks in the distribution of gaze directions can be seen, which correspond to looking at the three other participants in the meeting.

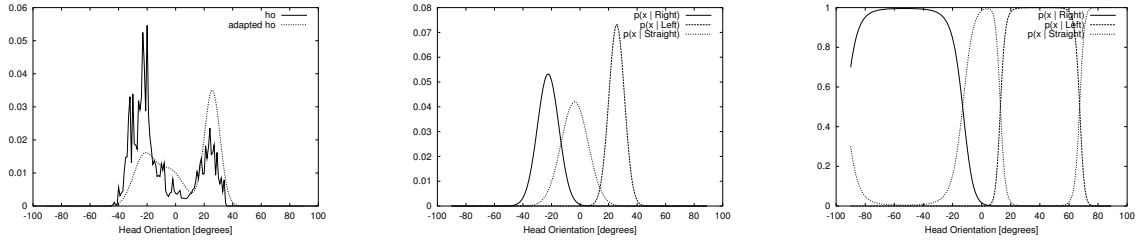
6.3 Predicting the Gaze Target Based on Head Orientation

We approached the second question we proposed before in this particular meeting application: How good can we predict at whom the subject was looking, on the basis of his head orientation? Answering this question gives us an idea of the upper limit of the accuracy that can be obtained when the focus of attention target is estimated based on head orientation alone.

6.3.1 Labeling Based on Gaze Direction

To automatically determine at which target person the subject was looking at (focus of attention), the gaze direction was used. Figure 6.4 shows the histograms of the horizontal gaze direction of two of the participants. In each of the histograms, it can be seen that there are three peaks. These belong to the direction where the other participants at the table were sitting.

We have automatically determined the peaks in the horizontal line-of-sight data-files using the k-means algorithm. The peaks found were then used as the directions where the other persons were sitting, and in each frame, focus of attention labels were assigned based on the least distance of the actual horizontal line-of-sight to the three target directions.



(a) (b) (c)

Figure 6.5: a) The distribution of all head orientation observations $p(x)$ from one subject and the found mixture of Gaussians. b) The three components of the mixture of Gaussians are taken as class-conditional head pan distributions. c) the posterior probability distributions $P(Focus|x)$ resulting from the found mixture of Gaussians.

6.3.2 Prediction Results

To see how accurate the focus target can be estimated based on observing head orientation alone, we used exactly the same method to find the focus targets as described in Chapter 5. The only difference now is, that in the previous chapter, focus was determined based on noisy head pan *estimates* as given by the neural networks, whereas now, focus targets are found based on accurate head pan *measurements* as given from the gaze tracking equipment.

Figure 6.3.2 depicts the unsupervised approach to determine focus posterior probabilites based on head orientations for one subject. As described in Chapter 5, first the EM-approach was used to fit a mixture of three Gaussians to the horizontal head orientation observations of each person (Figure 6.3.2 (a)). The found components of the Gaussian mixuture are then used as class-conditional head pan distributions $p(x|Focus)$ (Figure 6.3.2 (b)) and the mixture weights are used as the focus priors. From these, the posterior probabilities $P(Focus = T_i|\text{head orientation} = x)$ are computed (Figure 6.3.2 (c)). In each frame, the focus target T_i with the highest posterior probability is then chosen for each person.

Finally, focus detection accuracy was determined by comparing the found focus targets with the focus labels. Table 6.2 summarizes the results.

The accuracy result shows that the focus of attention target can be correctly estimated with only head orientation data in 82.6% (Subject 2) to 93.2% (Subject 3 and 4) of the frames, with an average of 88.7%. This can be seen as the upper limit of accuracy that we can get in head orientation based focus of attention estimation in such a scenario.

Subject	accuracy
1	85.7%
2	82.6%
3	93.2%
4	93.2%
Average	88.7%

Table 6.2: Focus detection based on horizontal head orientation measurements.

.

6.4 Discussion

The experiments presented in this chapter show that head orientation is a reliable cue for detecting at whom participants look in meetings. In the recorded meetings, focus of attention of four subjects could be detected 89% of the time based on accurate head orientation measurements. This result also gives an idea of the possible accuracy of focus detection with this approach: Even if we had a near-perfect method to estimate head orientation, focus estimation based on head orientation alone would fail in 11% of the frames in our data.

These experimental results are in accordance with several behavioral studies which suggest that head orientation is in fact a sufficient indicator for attention direction [Emery 2000, Argyle & Cook '76, Cranach '71] (see also Section 2.4).

Chapter 7

Combining Pose Tracking with Likely Targets of Attention

In previous chapters we have discussed that a person's head orientation is coupled with his or her attention, and have presented methods to estimate focus of attention based on head orientation in meetings. As we have seen, however, the proposed approach for focus of attention tracking is of course not perfect. Since eye gaze is not used in our approach, a certain amount of uncertainty is introduced. In addition, the noisy estimation of head orientations from camera images introduces errors.

To improve the performance of focus of attention tracking, we therefore would like to combine various sources of information.

As we have argued before, attention is clearly influenced by external stimuli, such as noises, movements or speech of other persons. Monitoring and using such cues might therefore help us to bias certain targets of interests against others.

Information about who is currently talking in a meeting clearly could be useful for the prediction of where people are attending to. It seems intuitive that participants tend to look at the speaker. Argyle, for instance pointed out, that listeners use glances to signal continued attention, and that gaze patterns of speakers and listeners are closely linked to the words spoken [Argyle & Cook '76].

Support for this idea comes also from a recent study of Vertegaal et al. [Vertegaal et al. 2001]. Their study investigated the relationship of where people look and whom they attend to during multi-party conversations. It was found that subjects looked about 7 times more at the individual they listened to than at others, and that subjects looked about 3 times more at individuals they spoke to. They conclude that information about who is looking at whom is an ideal candidate to provide

addressee information and that it can also be used to predict to whom someone is listening.

7.1 Predicting Focus Based on Sound

We have investigated whether and to what extent it is possible to predict a person's focus of attention based on information about who is speaking.

In our first experiment to predict focus from sound we analyzed at whom the four participants in the recorded meetings were looking during certain "speaking" conditions. Here, "speaking" was treated as a binary vector; i.e., each of the four participants was either labeled as "speaking" or "not speaking" in each video frame. Now, using this binary "speaking" vector and having four participants, there exist 2^4 possible "speaking" conditions in each frame, ranging from none of the participants is speaking to all of the participants are speaking.

Table 7.1 summarizes at whom subjects in our three meetings were looking, based on who was speaking. In the first columns, the different possible speaking conditions are represented. Here, the speakers can be any combination of the participants, which are the subject itself ("Self"), the person sitting left to the subject ("Left"), the person sitting opposite to the subject ("Center") or the person sitting right to the subject ("Right"). The speakers are marked with an "x" in the corresponding columns.

On the right side of the table, the percentages of how often the subject looked at the different other participants during the speaker constellations is indicated. For each person and each case we counted how often the subjects looked to the right, looked straight or looked to the person to their left. For example, when only the person to the subject's left was speaking, in 59% of the cases the subject was looking to the left person (the speaker), in 28% of the cases he was looking straight to the opposite person and in 11% of the cases he was looking to the person to his right.

Overall it can be seen that if there was only one speaker, subjects most often looked to that speaker. This also holds for cases were there was only one *additional* speaker when the subject itself was speaking. The percentages for these cases are indicated in bold font in Table 7.1.

The last row of Table 7.1 indicates in which direction subjects looked on average, regardless of speaking conditions. It can be seen that there is a bias towards looking straight; i.e., regardless who was speaking, in 44% of the cases the person opposite has been looked at, whereas the persons sitting to the side have been looked at in only 26% of the cases.

Speakers				Focus Targets		
Self	Left	Center	Right	Left	Center	Right
				0.26	0.49	0.23
			x	0.11	0.27	**0.60**
		x		0.12	**0.74**	0.11
		x	x	0.07	0.49	0.40
	x			**0.59**	0.28	0.11
	x		x	0.35	0.24	0.37
	x	x		0.33	0.60	0.05
	x	x	x	0.21	0.41	0.38
x				0.24	0.48	0.25
x			**x**	0.09	0.34	**0.53**
x		**x**		0.18	**0.61**	0.18
x		x	x	0.08	0.59	0.30
x	**x**			**0.60**	0.24	0.11
x	x		x	0.29	0.44	0.26
x	x	x		0.35	0.56	0.08
x	x	x	x	0.50	0.50	0.00
all cases				0.26	0.44	0.26

Table 7.1: Table summarizes, how often subjects looked to participants in certain directions, during the different speaking conditions (see text for further explanation).

The entries of Table 7.1 can be directly interpreted as the probability that a subject S was looking to a certain person T, based on a binary audio-observation vector $\vec{A}$ encoding which of the participants are speaking:

$$P(\text{Focus}|\text{Sound}) = P(\text{Focus}_S = T_j|\vec{A}),$$

where T_j with $j \in \{$ "Left", "Straight", "Right" $\}$ denote the possible persons to look at, and where

$$\vec{A} = (a_{\mathbf{S}elf}, a_{\mathbf{L}eft}, a_{\mathbf{C}enter}, a_{\mathbf{R}ight})$$

denotes the audio-observation vector with binary components a_i, indicating whether the subject it*self*, the person to his *right*, *left*, or the person opposite (*center*) to the subject was speaking.

The probability $P(\text{Focus}|\text{Sound})$ can be used directly to predict at whom a participant is looking in a frame, based on who was speaking during that video frame. In each frame, for each subject S the person T_i was chosen as the focus of person S, which maximized $P(\text{Focus}_S = T_i|\vec{A})$.

	P(Focus|Sound)
Meeting A	57.7%
Meeting B	57.6%
Meeting C	46.9%
Meeting D	63.2%
Average	56.3%

(a) Four participants

	P(Focus|Sound)
Meeting E	61.3%
Meeting F	54.2%
Average	57.8%

(b) Five participants

Table 7.2: Focus-prediction using sound only. Percentage of correct assigned focus targets by computing P(Focus|Sound). a) Results with four participants in meetings A to D. b) Results with five participants (Meeting F and G).

7.1.1 Sound-Only Based Prediction Results

By using only the speaker labels to make a sound-based focus prediction, the correct focus of each participants could be predicted with an average accuracy of 56.3% on evaluation meetings with four participants. Table 7.2 (a) summarizes the results on those meetings.

We also investigated how the sound-based prediction performs on the two meetings with five participants. Here, the posterior probabilities used for evaluation on one of the two meetings were adjusted on the other meeting. Table 7.2 (b) shows the corresponding results. Here, the correct focus target could be predicted in 57.8% of the cases on average.

7.2 Combining Head Pose and Sound to Predict Focus

In the previous section it was shown how we can determine the probability P(Focus|Sound); i.e., the probability that a person is looking towards a certain other person, based on the information, about who is currently speaking. By choosing in each frame the target person T_i which maximized $P(\text{Focus}_S = T_i|\vec{A})$ as the focus of person S, a focus prediction accuracy of 56.3% could be achieved on the meetings with four participants.

In section 5 we showed how to compute $P(\text{Focus}_S = T_i|x_S)$, the posterior probability, that a person S is looking towards person T_i, based on his estimated head rotation

x_S. There, by again choosing in each frame the target person T_i which maximized $P(\text{Focus}_S = T_i|x_S)$ as the focus of person S, we achieved correct focus prediction in 72.9% of the frames on the same meetings with four participants.

These two independent predictions of a person's focus – $P(\text{Focus}|\text{Sound})$ and $P(\text{Focus}|HeadPose)$ – can be combined to obtain a multimodal prediction of a person's focus which is based on both the observation, who is speaking, and based on the estimation of the person's head rotation.

A straightforward way to obtain a combined result is to compute the weighted sum of both prediction probabilities:

$$p(\text{Focus}) = (1-\alpha)P(\text{Focus}|\text{Head Pose}) + \alpha P(\text{Focus}|\text{Sound}).$$

We have evaluated the combined prediction results on our meetings for different values of α, ranging from 0.0 to 1.0. On the four meetings, the optimal values of α ranged from 0.3 to 0.6. By setting α to 0.6, good results could be achieved on all meetings. Using this multimodal prediction, an accuracy of 73.6% was achieved on the meetings with four participants. The results are shown in table 7.3 (a). Table 7.3 (b) shows the corresponding results on the meetings with five participants. Here, the combined focus prediction accuracy of 65.1% is 7.4% better than the focus prediction which is based on speaker information alone.

While the presented combination of head pose- and sound-based prediction is done heuristically by choosing a weighting parameter, we expect that by using more advanced and adaptive fusion methods, better combination results will be obtained. Appropriate fusion methods to be investigated could be to train neural networks for fusion of the two modalities, to determine the weighting parameters using error information of the two models or to investigate other feature dependent combinations methods [Miller & Yan '99, Woods et al. '97, Kittler et al. '98, Hansen & Salamon '90].

7.3 Using Temporal Speaker Information to Predict Focus

In the previous section we demonstrated how focus of attention can be estimated based on speaker information alone. The computation of $P(Focus|Sound)$ was based on the observation which participants speak at a given moment.

It is, however, reasonable to assume that temporal information about the speakers will affect the looking behaviour of the participants. When a new speaker starts to

	Head Pose only	Sound only	Combined
Meeting A	68.8%	57.7%	69.7 %
Meeting B	73.4%	57.6%	75.3 %
Meeting C	79.5%	46.9%	79.5 %
Meeting D	69.8%	63.2%	70.0 %
Average	72.9%	56.3%	73.6 %

(a) Meetings with four participants

	Head Pose only	Sound only	Combined
Meeting E	53.3%	61.1%	66.7 %
Meeting F	53.0%	54.2%	63.5 %
Average	53.2%	57.7%	65.1%

(b) Meetings with five participants

Table 7.3: Focus-prediction using only head orientation, using only sound and prediction using both head orientation and sound.

talk, the other participants, for instance, might need some time to shift their focus of attention to the new speaker. On the other hand, if a speaker is addressing several people, he might look at everyone for a while and the probability that the speaker will focus on the same person might decrease over time.

We assume that the prediction of the speaker's focus of attention could benefit from temporal information. Thus we would like to find $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$, the probability of looking at a focus target, based on having observed a history of audio events $A^t, A^{t-1}, ..., A^{t-N}$.

In this work we estimate $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$ using neural networks. It is well known that neural networks can be trained so as to estimate the a posteriori probabilities of target classes, given the input to the neural nets. This can be accomplished by using a neural network in classification mode, using a mean square error criterion and a 1-of-N output representation [Richard & Lippmann '91, Gish '90, Bourlard & Morgan '93].

We trained neural networks to estimate at which target person a subject is looking, given a history of audio-observations as input. Figure 7.1 depicts the resulting architecture of a neural network to predict the focus target for meetings with four participants.

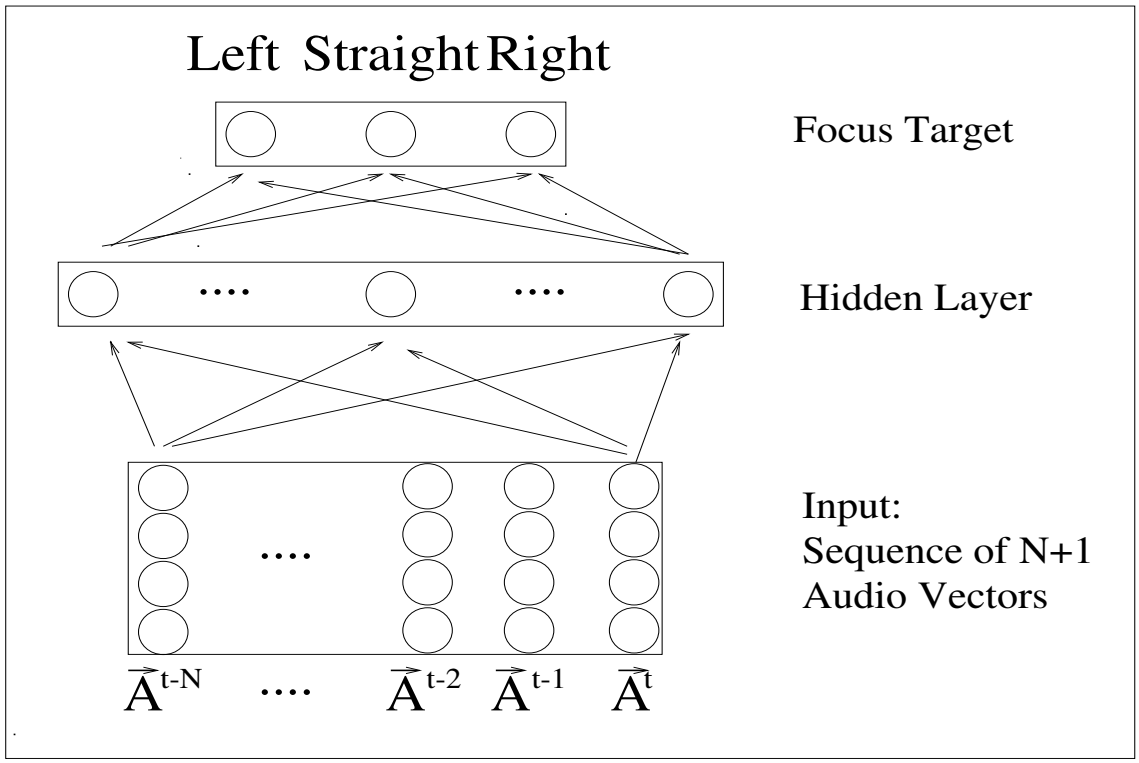


Figure 7.1: Neural net to predict focus target based on who is speaking. A sequence of binary vectors describing who is speaking at a given moment is used as input.

The neural net consists of an input layer of (N+1)*4 input units, corresponding to the (N+1) audio-observation vectors, one hidden layer and three output units, corresponding to the three target persons that a subject can look at. As audio-observations at each time step, again the binary audio-observation vectors $\vec{A} = (a_{\mathbf{S}}, a_{\mathbf{L}}, a_{\mathbf{C}}, a_{\mathbf{R}})$, described in the previous section, were chosen.

As output representation a 1-of-N representation was used; i.e., during training the output corresponding to the correct target class was set to 1 and the other output units were set to zero. As error criterion, the commonly used mean square error criterion was used.

After training, such a network will approximate the a posteriori probabilities of the focus targets F_i given the observered audio-information $A^t, A^{t-1}, ..., A^{t-N}$: $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$.

7.3.1 Experimental Results

To evaluate the performance of the prediction from temporal speaker information, the networks were trained round-robin; i.e., the neural nets were trained on data from two out of four meetings, cross-evaluation was done one a third meeting, and the networks were evaluated on the remaining meeting.

Neural networks were trained with different number of audio-events as input to find an appropriate length of the history that should be used. The investigated range of audio history ranged from only using the current audio-vector as input to using 40 audio-vectors as input. Since the audio-vectors were computed approximately 2.5 times per second, this corresponds to using up to 16 seconds of audio-history to predict the current focus of person.

During testing, the output unit which obtained the highest output activation was choosen as the winning unit and the corresponding target person was considered as the subject's current focus of attention.

Figure 7.2 shows the average sound-based focus prediction results on the four meetings for the different numbers of audio-vectors that were used as input and for different numbers of hidden units of the neural network. The best accuracy is 66.1%. This was achieved using three hidden units and a history of 20 audio-vectors, corresponding to approximately eight seconds of audio-information. In the figure, we see that the accuracy for all investigated numbers of hidden units strongly increases until ten audio-frames are used as input. Then the curve somehow flattens out and seems to asymptotically reach the 66% accuracy boundary.

Using the audio-history based prediction of focus, an average prediction accuracy of 66.1% on the four meetings could be achieved. Compared to the 56.3% achieved with the prediction based on a single audio-frame, this is a relative error reduction of 22%. The audio-based prediction results are summarized in Table 7.4.

	$P(\text{Focus}\|A^t)$	$P(\text{Focus}\|A^t, ..., A^{t-N})$
Meeting A	57.7%	59.2%
Meeting B	57.6%	69.6%
Meeting C	46.9%	61.3%
Meeting D	63.2%	74.3%
Average	56.3%	66.1%

Table 7.4: Focus-prediction using twenty frames of speaker information. Neural networks were trained to predict $P(\text{Focus}|A^t, A^{t-1}, ..., A^{t-N})$.

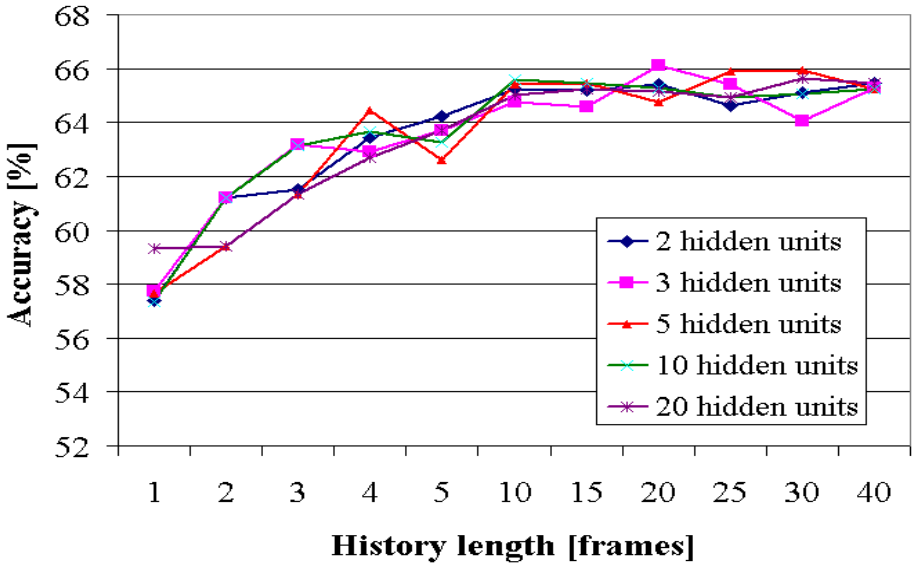


Figure 7.2: Sound-based focus prediction results with different audio-history lengths and different number of hidden units.

7.3.2 Combined Prediction Results

Again we can compute a combined, head orientation- and sound-based prediction by computing the weighted sum of $P(\text{Focus}|\text{Head Pose})$ and $P(\text{Focus}|\text{Sound})$:

$$P(\text{Focus}) = (1-\alpha)P(\text{Focus}|\text{Head Pose}) + \alpha P(\text{Focus}|A^t, ..., A^{t-N}).$$

By setting α to 0.6, we achieved an average accuracy of 75.6% on the meetings with four participants. Table 7.5(a) summarizes the results we obtained by using sound-only based focus prediction, head orientation-only based focus estimation and combined estimation.

Meetings with five participants

We also trained neural networks to predict a subject's focus of attention for the meetings with five participants. In this case a subject could look at four other participants. The output of the neural networks therefore consisted of four output units corresponding to the four possible targets. For the five participants case, we didn't evaluate different network architectures, but chose five hidden units and a history of 10 audio frames.

	Head Pose only	Sound only	Combined
Meeting A	68.8	59.2%	69.1%
Meeting B	73.4	69.6%	77.8%
Meeting C	79.5	61.3%	80.6%
Meeting D	69.8	74.3%	74.7%
Average	72.9	66.1%	75.6%

(a) Meetings with four participants

	Head Pose only	Sound only	Combined
Meeting E	51.9	65.4%	69.7%
Meeting F	53.0	59.7%	68.0%
Average	52.5	62.6%	68.9%

(b) Meetings with five participants

Table 7.5: Focus-prediction using only head orientation, only sound and prediction using both. Sound-based focus prediction is done with a neural network, using twenty frames of speaker information as input. Four persons participated in the meetings.

As in the four-participants case, the best accuracy was obtained by setting alpha to 0.6. Using the focus prediction based on head pose and sound, 68.9% accuracy was achieved on the two meetings. Compared to the focus prediction accuracy of 52.5% when using head orientation alone on these meetings, this is a huge increase in performance. Table 7.5(b) summarizes the results on the two meetings with five participants.

7.4 Summary

In this chapter we introduced the idea of estimating focus of attention from various cues in addition to a subject's head orientation.

We demonstrated that information about who is speaking is a good cue to predict the participants focus of attention in meetings and discussed two methods how focus of attention could be probabilistically predicted based on information who is speaking. We showed how neural networks can be used to predict focus from temporal speaker information. Experimental results indicate that temporal information about the speakers improves sound-only based focus prediction accuracy.

The combination of head-orientation based and speaker-based focus of attention prediction lead to significantly improved accuracy of focus prediction as compared to using one modality alone.

We think that the accuracy of focus of attention prediction can be furthermore improved by investigating more sophisticated fusion methods.

In addition, other cues such as movements, gesture tracking, or detecting certain keywords in the spoken content, such as the names of the participants, could be used in a similiar way to bias certain focus targets against others.

Chapter 8

Portability

In this chapter we discuss how the presented system for focus of attention tracking can be installed in a new location.

The focus of attention tracking experiments reported so far in this thesis were all performed using training data and recored meetings collected in our lab at the Universität Karlsruhe. To investigate which steps are necessary to successfully move the focus of attention tracking system to a new location, we have also installed the system in our lab at Carnegie Mellon University in Pittsburgh, USA.

The main problem when installing the system in a new location is that the illumination conditions in the new location might be completely different from the conditions in which the training data for the neural networks for head orientation estimation was collected. As we have already discussed in Section 4.5.2, the accuracy of head orientation estimation then seriously degrades under the new conditions.

One possibility to make the focus of attention tracking system perform well in a new location, is to collect training data in the new location and train a neural network for head orientation estimation by either using only the new images for training or by using them together with the training images that were collected in the other room. Such experiments were described in Chapter 4.5.2.

In this chapter we discuss how a neural network for head pan estimation can be *adapted* to work under new conditions by using some adaptation data collected in the new location. We examine how much adaptation data is necessary to obtain reasonable focus of attention tracking performance and compare the adaptation results to the results obtained with neural networks that are trained from scratch with the new data.

Figure 8.1: The data collection setup at CMU (see text).

8.1 Data Collection at CMU

In order to train neural networks for head pan estimation, we have collected training images from twelve users in our lab at CMU (the new location).

As during the data collection in Karlsruhe, subjects had to wear a head band with a Polhemus pose tracker sensor attached to it so that true head pose could be determined. Images of the person's head were captured with an omni-directional camera as described in Chapter 4 and were recorded together with the person's head pose. From each person, we collected training images at several locations around the meeting table. The data collection took about fifteen minutes for each participant. Altogether we collected around 27,000 training images from twelve persons.

Figure 8.1 shows an image of the data collection setup at CMU. The subject wears a head band to which the Polhemus pose tracking sensor is attached. To the right of the subject, the magnetic emitter is placed on a tripod. On the table, the omni-directional camera system can be seen.

8.2 Head Pan Estimation Experiments

Having collected the data at CMU we performed the following experiments:

1. First, the head pan estimation accuracy of a neural network which was trained on images that were taken in our lab in Karlsruhe (the "UKA-net") was determined on a test set of the new data from CMU. Using the UKA-net the average error for head pan estimation was 19 degrees on the data from CMU.

2. New neural networks to estimate head pan were trained using increasing amounts of training data from CMU.

3. The UKA-network was adapted using increasing amounts of data from collected at CMU.

For all experiments, we used three pre-processed facial images together as input to the neural networks: the histogram-normalized image, the horizontal edge image and the vertical edge image. Details on these pre-processing methods were given in Chapter 4.

8.2.1 Training New Networks from Scratch

We first trained neural networks for head pan estimation using only the data that was collected at CMU. To see how much training data is necessary for reasonable generalization, we trained different networks using increasing subsets of the data. To evaluate the performance of the networks, data from four subjects was kept aside as a user-independent test set.

We trained networks on images from one up to all eight subjects in the training set. The neural network architecture and training was identical to those used with the networks trained with the data from Karlsruhe as described in Chapter 4. The networks were trained on the training data set and a cross-evaluation set was used to determine the number of training iterations.

Figure 8.2 shows the results obtained on the user independent test set from CMU (top curve). It can be seen that the average pan estimation error on the test set is as high as twenty degrees when only images from one subject were used for training. The pan estimation error then gradually decreases, when training images from more subjects are added. When all eight subjects were used for training, an average pan estimation error of 13 degrees was obtained.

We also trained one neural network on images on all the available twelve subjects. For training we used 80% of all the images. 10% of the images were used for cross-evaluation and the remaining 10% of the images were used as a test set. With this multi-user network for pan estimation, we achieved an average error of 7.6 degrees on the test-set.

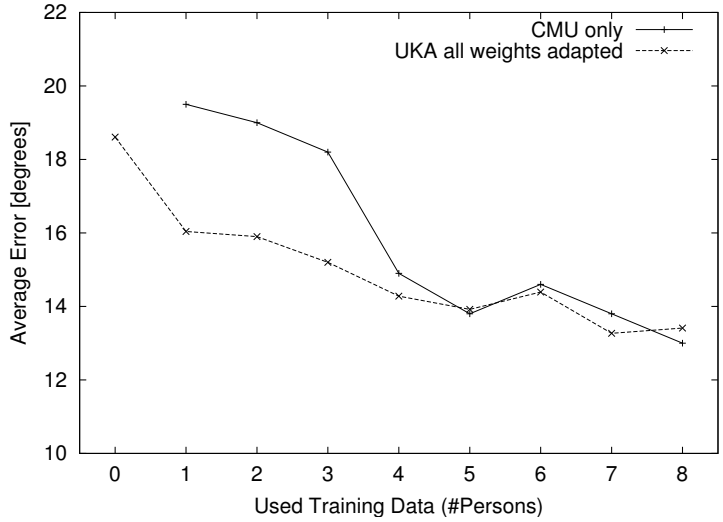


Figure 8.2: Pan estimation results on a user-independent test set from CMU. Shown are the results for networks trained from scratch with data from CMU and the results of the UKA-network when all weights were adapted using the data from CMU. For both approaches, results using images from an increasing number of persons for training/adaptation are shown.

8.2.2 Adapting a Trained Network

We then investigated whether and how well the network which was previously trained on data collected in Karlsruhe – the "UKA-network"' – could be adapted to the new CMU images, by using the different training data sets from CMU for adaptation.

We adapted the UKA-net using the images from one to all eight subjects of the CMU training set for adaptation. The performance of the adapted networks was then also evaluated on the four other persons in the user-independent test-set collected at CMU.

Adapting All Weights

We first adapted the UKA-network by retraining all its weights on the different adaptation data sets from CMU. Training was done using standard back-propagation with a learning parameter of 0.1. To determine when the adaptation process should stop, a cross-evaluation set containing images from an additional subject was used. The

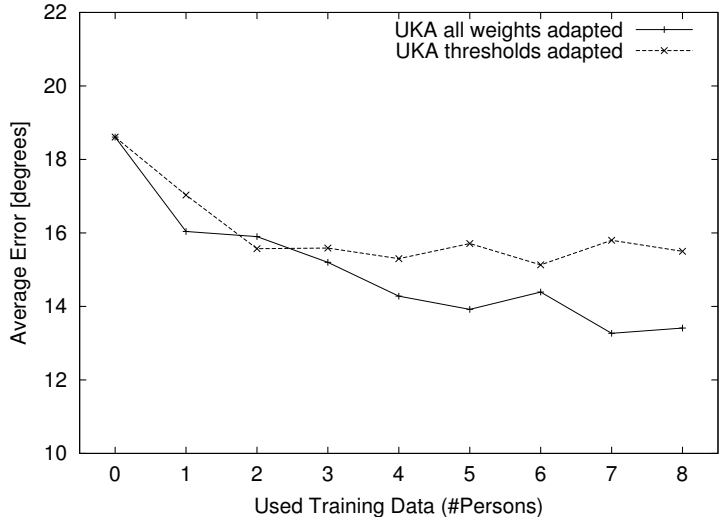


Figure 8.3: Pan estimation results on a user-independent test set from CMU. Shown are the results with the adapted UKA-network. The lower curve indicates the mean pan estimation errors when all weights were adapted; the upper curve indicates the results when only the unit biases of the network were adapted.

images in the cross-evaluation set were also collected at CMU. Typically, adaptation stopped after two to six iterations.

With the unadapted UKA-network an average error of 19 degrees was obtained on the test set. By using images from one subject from CMU for adaptation, the average error decreases to 15.6 degrees. When all training data from eight subjects is used for adaptation, the average pan estimation error decreases to 13 degrees. The results are also shown in Figure 8.2 (lower curve).

It can be seen that pan estimation works significantly better with the adapted networks when only little data is available for training or adaptation. In our experiments, the newly trained network only reached the performance of the adapted UKA-network, when training images from at least five subjects were available for training.

Adapting the Unit Biases Only

We also investigated how well the UKA-network would perform on the CMU images, when only the biases (thresholds) of the network's units are adapted. Since many fewer parameters have to be retrained when only the unit thresholds are adapted,

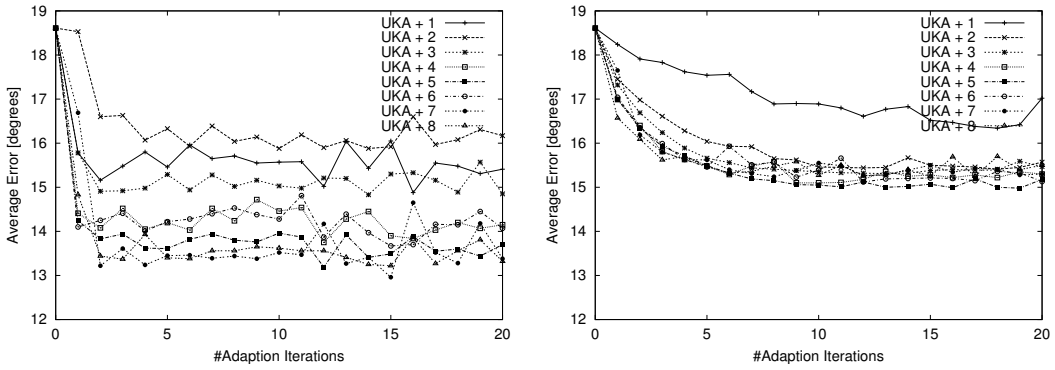


Figure 8.4: Adaption results when adapting all weights (a) or unit biases only (b). Shown are the average pan estimation errors for increasing numbers of training iterations and using images from an increasing number of subjects for adaptation.

this approach might lead to better generalization when only little training data is available for adaptation.

Figure 8.3 compares the results obtained when only the unit biases of the UKA-network were adapted with the results obtained when all weights were adapted. For both approaches, the pan estimation results when images from more and more subjects were used for adaptation are shown.

It can be seen that some gain in pan estimation performance can be obtained by adapting only the unit biases. By using images from only two subjects for adaptation, the average error for pan estimation already decreased from 18.6 degrees to 15.6 degrees. This result is in fact slightly better than the the pan estimation result that was obtained when all network parameters were adapted using the same training set. When more data was used for adaptation, however, no further significant improvement could be achieve when only the unit thresholds were adapted. Thus, adapting all network parameters led to significantly better pan estimation results when more training data was available.

Figure 8.4 shows the pan estimation results on the CMU test set for increasing number of adaptation iterations and for increasing numbers of images used for adaptation. In Figure 8.4 (a) the results when all weights are adapted are shown. It can be seen that for all adaptation sets, the pan estimation error is close to its minimum already after two training iterations. We furthermore see, that the pan estimation error is decreasing when images from more subjects are used for adaptation. Figure 8.4(b) shows the results when only the unit biases are adapted. Here, the learning curve

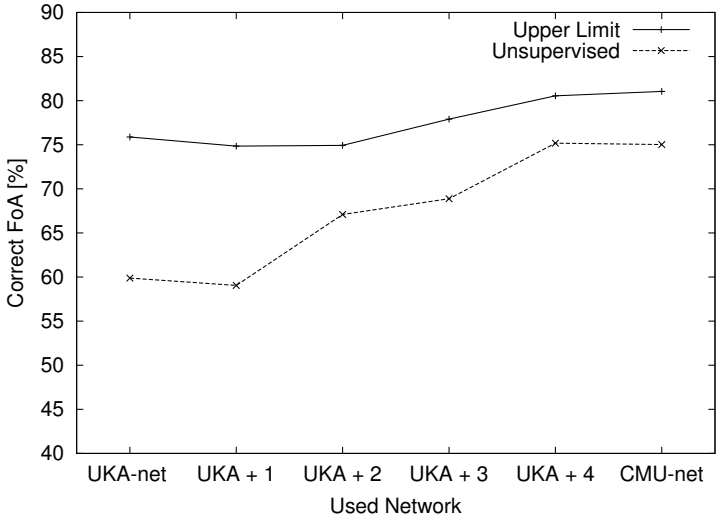


Figure 8.5: Accuracy of focus of attention detection on a meeting recorded at CMU. Both the upper limit and the result using unsupervised adaptation of the model parameters is indicated for the different neural networks (see text).

flattens out after nine to ten iterations for most adaptation sets containing more than one person. We again see that the performance does not improve when more than two subjects are used in the adaptation set.

8.3 Focus of Attention Detection Results

To measure how well focus of attention can be estimated using the different neural networks, we have collected two meetings with four participants in our lab at CMU.

The focus of attention tracking system was run on the recorded meetings with different networks for pan estimation. For the evaluation we used the unadapted UKA-network, the UKA-networks where all weights were adapted with images from one to four subjects and the neural network that was trained on images from all twelve subjects in our data set from CMU.

For each network we evaluated the focus of attention detection accuracy using the mixture of Gaussian approach presented in Chapter 5. All parameters of the Gaussian mixture model were adapted completely unsupervised. We also measured the upper limit of focus of attention detection accuracy that is possible given the estimated head pan observations from the different networks (see section 5.2.3 for more details).

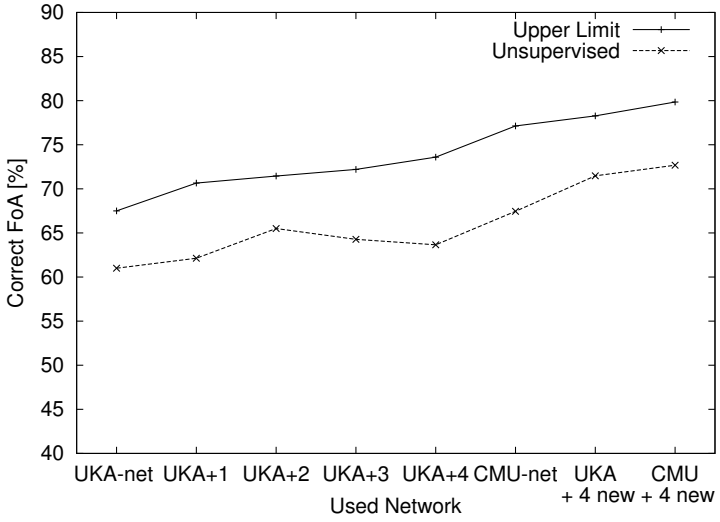


Figure 8.6: Accuracy of focus of attention detection on a second meeting recorded at CMU. Both the upper limit and the result using unsupervised adaptation of the model parameters is indicated for the different neural networks (see text).

Figure 8.5 shows focus of attention detection accuracy on the first meeting for the different networks used for head pan estimation. In the figure as well the unsupervised focus of attention detection results as the supervised upper focus of attention detection limits are indicated.

Using the UKA-network for head pan estimation, focus of attention could be detected in only 60% of the time on the meeting, with a possible upper limit of 76%. By adapting the UKA-network with data collected at CMU the performance increases to 75% focus of attention detection accuracy when images from four subjects were in the adaptation set ("UKA + 4"). This performance is already as good as the performance obtained with the CMU-network, which was trained on images from twelve subjects collected at CMU.

Figure 8.6 shows the results on the second meeting. On this meeting both the results with the CMU-network and the adapted UKA-networks are worse than the results on the first meeting. Using the best adapted network – using two subjects for adaptation – we obtained 66% focus of attention detection accuracy. Using the CMU-network 67% accuracy could be achieved.

This second meeting was recorded approximately two weeks after the training images at CMU were collected. After recording (and evaluating) the meeting, we discovered that the camera gain and the focus of the omni-directional camera system had been

changed since the training data for the neural networks was collected at CMU. This most likely caused the poorer results.

In order to improve focus of attention tracking we therefore again collected training images from four subjects with the current camera settings to adapt the neural networks with the new data. Only one of the subjects from which we collected further training data was a participant of this evaluation meeting. Figure 8.6 also indicates the obtained focus of attention tracking results obtained with the UKA-network that was adapted using only this new adaptation data ("UKA + new") and the results with adapted CMU-network ("CMU + new"). With the newly adapted UKA-network, 71% accuracy was obtained (78% upper limit), with the adapted CMU-network we achieved 73% accuracy, with a possible upper limit of 80%.

8.4 Discussion

In this chapter we discussed how the system for focus of attention tracking can be ported to a new location. Our experiments suggest that a network which has already been trained to estimate head pan from images taken in one location can be adapted to work in a new location and under different illumination conditions by collecting a limited number of images in the new location and adapting the networks' weights with the new images. In our experiments we achieved good focus of attention tracking results in the new location by using adaptation images from only four subjects. These images could be collected in approximately one hour. Our experiments also showed that adapting an existing network for pan estimation, which has been trained on images taken in different lighting and camera conditions, leads to better pan estimation results than training networks from scratch with images from the new location when only a small amount of training images are available.

Chapter 9

Focus of Attention in Context-Aware Multimodal Interaction

The components to estimate a user's focus of attention can also be applied to other situations than meetings. To demonstrate the generality of our approach, we have applied the developed components to a human-robot interaction scenario [Stiefelhagen et al. 2001b].

Advancing human-robot interaction has been an active research field in recent years [Perzanowski et al. 2001, Agah 2001, Koku et al. 2000, Adams et al. 2000, Matsusaka et al. '99]. A major challenge is to develop robots that can behave like and interact with humans. In an intelligent working space, social robots should be capable of detecting and understanding human communicative cues. Tracking a user's focus of attention can be useful to improve human-robot interaction.

In an intelligent working space, where humans and robots may interact with each other, information about a user's focus of attention could for instance be useful to interpret what object or place a person is referring to when talking with a robot, or to determine whether a person is talking to the robot or not.

We have built a prototype system to demonstrate focus of attention aware interaction with a household robot and other smart appliances in a room using the components for focus of attention tracking presented in this thesis.

Figure 9.1 outlines the idea of the demonstration system.

In the demonstration environment, a subject could interact with a simulated household robot, a speech-enabled VCR or with other people in the room, and the recipient

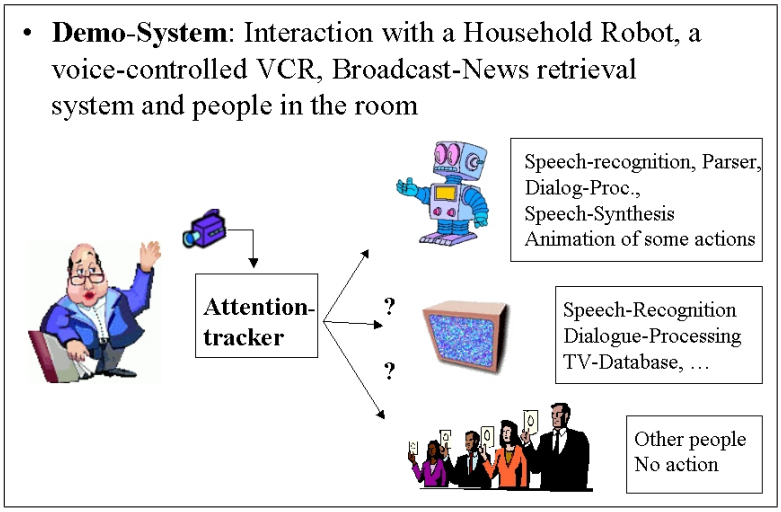


Figure 9.1: A demonstration prototype system to show focus of attention aware interaction with several appliances in a smart room. See text for details.

of the subject's speech was disambiguated using focus of attention tracking.

The system consisted of the following main components:

Robot Visualization For the demonstration we have simulated a robot using a 3D visualization toolkit and projected the robot onto one of the walls of our lab.

Speech Recognition A speaker independent large-vocabulary continuous speech recognizer was used for understanding the users' commands [Soltau et al. 2001].

Parser A parser based on the system described in [Gavalda 2000] was used to analyze the hypothesis received from the speech recognition module and to generate action commands that were sent to the robot visualization module.

Dialog Manager This module enabled the virtual robot to lead simple clarification dialogues, if necessary information is missing.

Speech Synthesis A Speech synthesis system [Black & Taylor '97] is used to provide spoken feedback to the user.

Focus of Attention Tracker To observe the user's focus of attention, a pan-tilt-zoom camera was placed next to the simulated robot. The face of the user was tracked in the camera image and the user's head pose was estimated with a neural net as described in Chapter 4.

Communication of all the components - recording, speech recognizer, parser, dialogue manager, visualization and focus-of-attention-tracker - was done using a client-server architecture that we adapted from [Fügen et al. 2001].

For the demonstration three focus-targets were chosen: a) the robot, which was displayed in the front of the user, next to the camera; b) the VCR located at the left and c) an area right to the user, where other people in the room were located.

During an initialization phase, the user had to look at each of the three targets for while. During this phase, his head orientation was continuously estimated and stored to a file. From the observed head orientations during the initialization phase, the class-conditional probability density functions of his head orientations when looking at the three targets could then automatically be determined using the approach described in Chapter 5.

After initialization, the most likely target could continously be determined based on the user's head orientation and the class-conditionals found during initialization. Since we assumed equal priors for each of the targets, the computation of the most likely target could be simplified to choosing the target which maximized class-conditional probability for the users' observed horizontal head orientation.

Now, whenever the user was looking towards where the VCR was placed, the focus of attention module identified the VCR as target and the output of the speech recognizer was sent to the the VCR.

Whenever the user was looking towards the simulated robot, the robot was chosen as the focus target, and therefore recognized speech was directed to the robot; i.e., the robot's parser, dialogue and visualization module, to generate appropriate actions of the robot simulation. Whenever the user was neither looking at the VCR nor to the robot the user's speech was not recorded at all and neither the robot nor the VCR were responding.

This demonstration shows how the developed components for focus of attention tracking can be used to enhance interaction with smart appliances such as household robot or a speech enabled VCR.

In the presented demonstration, a user's focus of attention is only used to determine the current addressee of the user's speech. Focus of attention, however, could also be used during multimodal communication to determine to what object or place a person is referring to ("Put that there!").

The demonstration also shows one limitation of the developed approach for attention tracking. Within the current approach, it is assumed that the subject (the user) and the focus targets do not change their position after an initialization phase. While

this can be assumed for meetings, where people sit around a table, this assumption is too restrictive when a person is interacting with appliances in a smart room. In a smart room, the user will most likely not stand at the same position in the room when he is interacting with appliances. A less restrictive system should allow the user to move freely in a room. A possibility to overcome this problem could for instance be to develop some sort of online-adaptation of the class-conditional head pose distributions and to use a 3D model of the scene and the interesting targets in it in order to facilitate the detection of the correct focus target based on the user's head orientation at a give position in the room.

Chapter 10

Conclusions

This thesis has addressed the problem of tracking focus of attention of participants in meetings. In this we work we studied why and how focus of attention tracking in meetings could be beneficial. To our knowledge, this is the first work presenting a system capable of tracking focus of attention in meetings.

In our system, focus of attention of each meeting participant is estimated based on his or her head orientation. We have discussed relevant literature suggesting that head orientation is a reliable cue for detecting to whom someone is attending. In addition we have experimentally demonstrated that head orientation can be used to predict where a person is attending in meetings. A user study has has been conducted investigating how precisely focus of attention can be predicted in a meeting with four participants by using only the participants' head orientation. The user study clearly demonstrated that head orientation is a very reliable cue to detect to whom someone is attending. In the meetings which we recorded for this study, we were able to correctly determine at whom the subject was looking 89% of the time based solely on the subject's head orientation.

In order to build a working system to track the participants' focus of attention based on their head orientation, the following problems were addressed in this thesis:

1. Detecting and tracking the locations and the faces of all the participants in a meeting.
2. Estimating each participants' head orientation from facial images.
3. Building a general probabilistic framework to determine at whom each participant is looking based on his or her observed head orientation.

In the system developed an omni-directional camera is used to capture the scene around a meeting table. Participants and their faces are detected and tracked using a skin-color based face tracker.

To estimate participants' head orientations, a neural network-based system for estimating head orientations has been developed. We have demonstrated that this system is capable of predicting head orientation from both high- and low-resolution images that were obtained from an omni-directional camera. With a multi-user system that was trained to estimate head orientations of 12 different users from images taken with an omni-directional camera, an average error in estimating horizontal head rotation of only 3.1 degrees was obtained. On new users we achieved an average error of 9.5 degrees for estimating horizontal head orientation. Using higher-resolution facial images, an average error of 7.1 degrees could be obtained for new users. This compares favorably with the results obtained with other vision-based head orientation estimation systems reported in the literature.

A major contribution of this thesis is the design of a probabilistic framework to determine at which target a person is attending, based on his or her head orientation. With the proposed model, we estimate the *a-posteriori* probability that a person is looking at a certain target, given the subject's observed head pose. Furthermore, we have presented an approach as to how the underlying class-conditional probability density functions and priors can be adapted in an unsupervised, data-driven way, given that the number of possible targets at which the subject might have been looking is known. The proposed approach automatically adapts to different numbers of participants and to different locations of the participants relative to each other. We have experimentally evaluated this focus of attention detection approach on several recorded meetings, each containing four or five participants. On the meetings with four participants 72% accuracy in detecting the correct focus of attention of each participant could be achieved. On meetings with five participants, 53% accuracy could be achieved.

Another important contribution of this thesis is the investigation of whether a person's focus of attention in meetings can be predicted based on information about who is speaking. Our work demonstrates that information about who is speaking is a good cue to predict where participants are looking. On our recorded meetings with four participants we have demonstrated that the participants' focus of attention can be predicted based only on information about which participants are speaking at a given moment. Based on this information, we were able to correctly predict focus of attention 56% of the time. In addition we have demonstrated that this sound- or speaker-based prediction can be significantly improved by also taking into account who *was* talking before a given moment. By training neural networks to predict focus of attention based on a time window of information about who was speaking,

sound-based prediction of focus could be increased to 66% accuracy on the recorded meetings.

Finally, we have shown that head pose-based and sound-based prediction of focus of attention can be combined in order to get an improved accuracy of focus of attention detection. By combining the head pose-based and sound-based posterior probabilities of the different targets, we have achieved 77% accuracy in detecting focus of attention on the recorded meetings with four participants. This amounts to a relative error reduction of 18% compared to using only head orientation for prediction (72% accuracy). On the meetings with five participants the accuracy achieved by combining both methods resulted in 69% accuracy versus 63% accuracy using only sound and 53% using only head orientation for focus estimation.

The presented system for focus of attention tracking has been successfully installed in both our labs at the Universität Karlsruhe, Germany and at Carnegie Mellon University in Pittsburgh, USA. A problem when porting the system to a new location is the need for appropriate training images for the neural network based approach for head orientation estimation. We have therefore investigated how a neural network for head pan estimation can be *adapted* to work under new conditions by using some adaptation data collected in the new location. We have examined how much adaptation data is necessary to obtain reasonable performance and have compared the adaptation results to the results obtained with neural networks that are trained from scratch with the new data. Our experiments showed that adaptation images from only four subjects were sufficient to achieve good focus of attention detection accuracy in a new location with completely different illumination conditions.

Tracking a user's focus of attention could also be useful for several other application areas such as gaze-aware human-computer interaction, shared collaborative workspaces or psychological experiments where monitoring a subject's focus of attention might be of interest.

To demonstrate how our approach for focus of attention tracking can be used to enhance multimodal human-computer interaction, we have integrated the developed components into a prototype system for multimodal focus-aware human computer interaction in a smart room. In the demonstration scenario, we showed how focus of attention tracking can be used to determine which of several possible speech-enabled appliances the user addressed.

10.1 Future Work

The work presented in this thesis could lay the foundation for many future studies.

The investigation of how focus of attention can be used for the analysis and the better understanding of meetings opens new research directions. A number of suggestions how focus of attention can be used here have been given in this study: to deliver deictic information, which is often missing in the spoken content; to monitor activity and attentiveness of participants; perhaps focus of attention could also be a good cue for the classification and analysis of meeting types and discourse segments.

Focus-of-attention-tracking could be used in many ways to enhance human-computer interaction. A straightforward example is given in this thesis. There, focus of attention is used to disambiguate between a number of possible target applications, which can be controlled via speech. However, many other uses could be thought of (dimming room lights when a user is looking at the TV screen for a while, switching on and off computer displays based on the user's focus, ...).

This work has presented a first system for focus of attention tracking in meetings. The developed system of course has several limitations which should be addressed in future work:

In order to reliably estimate head orientation from facial images, unoccluded and more or less correctly extracted facial images are necessary. As we have discussed in Chapter 3, our – and probably any other – automatic face detection method will sometimes wrongly detect outliers as a face, such as a subject's hands for example; sometimes only parts of a face are detected, for example due to extreme shadow in the face; in other cases, the face might be occluded by the subject's hand or arms for example. In order to reliably estimate the subject's head orientation, such outliers should be detected automatically in the future.

The current approach to estimate a person's focus of attention target assumes that the locations of the subject and the target persons remains relatively stable, which is true for the meetings we evaluated and which is likely to be a reasonable assumption for many meetings. For other applications of focus of attention tracking, such as for focus-aware human-computer interaction in smart rooms, we, however, cannot assume the user to remain standing at the same location when interacting with smart appliances. For such scenarios, the current approach will have to be extended.

Bibliography

Abowd Gregory D., Atkeson Chris, Feinstein Ami, Hmelo Cindy, Kooper Rob, Long Sue, Sawhney Nitin, Tani Mikiya (1996). Teaching and Learning as Multimedia Authoring: The Classroom 2000 Project. In *Proceedings of the ACM Multimedia'96 Conference*, pp. 187–198, November 1996.

Adams B., Breazeal C., Brooks B.A., Scassellati B. (2000). Humanoid robots: a new kind of tool. *IEEE Intelligent Systems*, 15(4):25–31.

Agah A. (2001). Human interactions with intelligent systems: research taxonomy. *Computers and Electrical Engineering*, 27(1):71–107.

Argyle Michael, Cook Mark (1976). *Gaze and Mutual Gaze*. Cambridge University Press.

Argyle Michael (1969). *Social Interaction*. Methuen, London.

ARR (). Arrington Research. http://www.arringtonresearch.com/.

Ascension (). Ascension Technology Corporation. http://www.ascension-tech.com/.

ASL (). Applied Science Laboratories. http://www.a-s-l.com/.

Baker S., Nayar S. K. (1998). A Theory of Catadioptric Image Formation. In *Proceedings of the 6th International Conference on Computer Vision , ICCV 98*, pp. 35–42, Bombay, India, January 1998.

Baldwin D. (1991). Infants' contribution to the achievement of joint reference. *Child Development*, 62:875–890.

Ballard Dana H., Brown Christopher M. (1982). *Computer Vision*. Prentice Hall, New Jersey.

Baluja Shumet, Pomerleau Dean (1994). Non-Intrusive Gaze Tracking Using Artificial Neural Networks. Technical Report CMU-CS-94-102, Carnegie Mellon University, 1994.

Barber P.J., Legge D. (1976). *Perception and Information*, chapter 4: Information Acquisition. Methuen, London, 1976.

Bett Michael, Gross Ralph, Yu Hua, Zhu Xiaojin, Pan Yue, Yang Jie, Waibel Alex (2000). Multimodal Meeting Tracker. In *RIAO 2000 : Content-Based Multimedia Information Access*, pp. 32–45, Paris, France, April 2000.

Beymer D., Shashua A., Poggio T. (1993). Example-based Image Analysis and Synthesis. Technical Report TR-1431, MIT AI Lab, 1993.

Bishop Christopher M. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press.

Black A.W., Taylor P. (1997). The Festival Speech Synthesis System: system documentation. Technical Report, Human Communication Research Center, University of Edinburgh, UK, 1997.

Black M., Brard F., Jepson A., Newman W., Saund E., Socher G., Taylor M. (1998). The Digital Office: Overview. In *Proceedings of the 1998 AAAI Spring Symposium on Intelligent Environments*, volume AAAI Technical Report SS-98-02. AAAI, AAAI Press, March 1998.

Bourlard H., Morgan N. (1993). *Connectionist Speech Recognition*. Kluwer Academic Publishers, Boston.

Broadbent D. E. (1958). *Perception and communication*. Pergamon Press, London.

Brumitt B., Krumm J., Meyers B., Shafer S. (2000a). Let There Be Light: Comparing Interfaces for Homes of the Future. *IEEE Personal Communications*.

Brumitt B., Meyers B., Krumm J., Kern A., Shafer S. (2000b). EasyLiving: Technologies for Intelligent Environments. In *Handheld and Ubiquitous Computing*, September 2000.

Burroughs W., Schultz W., Aubrey S. (1973). Quality of argument, leadership roles and eye contact in three person leaderless groups. *Journal of Social Psychology*, 90:89–93.

Calhoun Gloria L., McMillan Grant R. (1998). Hands-free input devices for wearable computers. In *Proceedings of Fourth Annual Symposium on Human Interaction with Complex Systems*, pp. 118–123. IEEE, 1998.

Cascia M. La, Isidoro J., Sclaroff S. (1998). Head tracking via robust registration in texture map images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, Santa Barbara, CA, 1998.

Cherry E.C. (1957). *On human communication: A review, a summary and a criticism*. MIT Press, Cambridge, MA.

Cranach von M. (1971). The role of orienting behaviour in human interaction. In Esser A. H. (Eds.), *Environmental Space and Behaviour*. Plenum Press, New York, 1971.

Drees A. (1995). *Visuelle Erkennung von Handstellungen mit neuronalen Netzen*. PhD Thesis, Univ. Bielefeld, Germany, 1995.

Emery N.J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews*, 24:581–604.

Encyclopaedia Britannica, 2002. http://www.britannica.com.

Eriksen C. W., Yeh Y. Y. (1985). Allocation of attention in the visual field. *Journal of Experimental Psychology: Human Perception and Performance*, 11:583–597.

Exine R.V., Winter L.C. (1966). *Affect, Cognition and Personality*, chapter Affective relations and mutual glances in dyads. Tavistock, London, 1966.

Frey L. A., K. P. White jr., Hutchinson T. E. (1990). Eye-gaze word processing. *IEEE Transactions on Systems, Man and Cybernetics*, 20(4):944–950.

Fügen Christian, Westphal Martin, Schneider Mike, Schultz Tanja, Waibel Alex (2001). LingWear: A Mobil Tourist Information System. In *Proceedings of the of the First International Conference on Human Language Technology (HLT 2001)*, San Diego, March 2001.

Gavalda M. (2000). SOUP: A parser for Real-World Spontaneous Speech. In *Proceedings of the 6th International Workshop on Parsing Technologies, IWPT-2000*, Trento, Italy, February 2000.

Gee Andrew H., Cipolla Roberto (1994). Non-Intrusive Gaze Tracking for Human-Computer Interaction. In *Proc. Mechatronics and Machine Vision in Practise*, pp. 112–117, 1994.

Gee Andrew H., Cipolla Roberto (1995). Fast Visual Tracking by Temporal Consensus. Technical Report CUED/F-INFENG/TR-207, University of Cambridge, February 1995.

Gips J., Olivieri P., Tecce J. (1993). Direct control of the computer through electrodes placed around the eyes. In *Porceedings of the Fifth International Conference on Human-Computer Interaction*, 1993.

Gish Herbert (1990). A probabilistic approach to the understanding and training of neural network classifiers. In *Proceedings of IEEE Conference on Acoustics Speech and Signal Processing*, pp. 1361–1364, 1990.

Glenstrup Arne John, Engell-Nielsen Theo (1995). Eye Controlled Media: Present and Future State. Technical Report, University of Copenhagen, http://www.diku.dk/users/panic/eyegaze/, 1995.

Goodwin Charles (1981). *Conversational Organization: Interaction between Speakers and Hearers*. New York: Academic Press.

Gopher Daniel (1990). *The Blackwell dictionary of Cognitive Psychology*, chapter Attention, pp. 23–28. Basil Blackwell Inc., 1990.

Gross Ralph, Bett Michael, Yu Hua, Zhu Xiaojin, Pan Yue, Yang Jie, Waibel Alex (2000). Towards a Multimodal Meeting Record. In *IEEE International Conference on Multimedia and Expo*, pp. 1593–1596, 2000.

Hansen L.K., Salamon P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:993–1001.

Haralick R. B., Joo H., Lee C-N., Zhuang X., Vaidya V.G., Kim M. B. (1989). Pose estimation from corresponding point data. *IEEE Transactions on Systems, Man, and Cybernetics*, 19(6):1426–1445.

Heinzmann Jochen, Zelinsky Alexander (1998). 3-D Facial Pose and Gaze Point Estimation using a robust real-time tracking paradigm. In *International Conference on Face and Gesture Recognition*, pp. 142–147, Nara, Japan, April, 14-16 1998.

Hunke Martin, Waibel Alex (1994). Face Locating and Tracking for Human-Computer Interaction. In *Twenty-Eight Asilomar Conference on Signals, Systems and Computers*, Monterey, California, November 1994.

Hutchinson T. E., K. P. White jr., Martin W. N., Reichert K. C., Frey L. A. (1989). Human-computer interaction using eye-gaze input. In *IEEE Transaction on Systems, Man, and Cybernetics*, volume 19, pp. 1527–1534, 1989.

iReality (). iReality.com, Inc. http://www.genreality.com/.

ISC. (). Iscan Inc. http://www.iscaninc.com/.

Itti L., Koch C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.*, 40(10-12):1489–1506.

Itti L., Koch C., Niebur E. (1998). A model of saliency-based visual-attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jacob R. J. K. (1993). Eye-movement-based human-computer interaction techniques. In Hartson H. R., Hix D. (Eds.), *Advances in Human-Computer Interaction*, volume 4, pp. 151–190. Ablex Publishing Corporation, Norwood, NJ, 1993.

Jacob R. J. K. (1995). Eye tracking in advanced interface design. In Barfield W., Furness T. A. (Eds.), *Virtual Environments and Advanced Interfaces*, pp. 258–288. Oxford University Press, New York, 1995.

James William (1890/1981). *The Principles of Psychology*. Harvard UP, Cambridge,MA.

Jebara T.S., Pentland A. (1997). Parametrized structure from motion for 3D adaptive feedback tracking of faces. In *Proceedings of Computer Vision and Pattern Recognition*, 1997.

John Hertz Richard G. Palmer, Anders Krogh (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley.

J.W.Tankard (1970). Effects of eye position on person perception. *Perc. Mot. Skills*, (31):883–93.

Kittler J., Hatef M., Duin R.P.W., Matas J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20:226–239.

Kleinke C. L., Bustos A. A., Meeker F. B., Staneski R. A. (1973). Effects of self-attributed and other-attributed gaze in interpersonal evaluations between males and females. *Journal of experimental social Psychology*, (9):154–63.

Koch C., Ullman S. (1985). Shifts in selective visual attention: towards the underlying circuitry. *Human Neurobiology*, 4:219–227.

Koku A.B., Sekmen A., Alford A. (2000). Towards socially acceptable robots. In *Proceedings of 2000 IEEE International Conference on Systems, Man and Cybernetics*, pp. 894–899, 2000.

Kwong J., Gong S. (1999). Learning support vector machines for a multi-view face model. In Pridmore T., Elliman D. (Eds.), *10 th British Machine Vison Conference*, volume 2, pp. 503–512, Notthingham, UK, Sept. 1999.

LaBerge D. (1983). Spatial extent of attention to letters and words. *Journal of Experimental Psychology: Human Perception and Performance*, 9:371–379.

Langton Stephen R.H., Watt Roger J., Bruce Vicki (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Neuroscience*, 4(2):50–58.

Langton Stephen R. H. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quaterly Journal of Experimental Psychology*, 53A(3):825–845.

LCT (). LC Technologies. http://www.eyegaze.com/.

Maglio Paul P., Matlock Teenie, Campbell Christopher S., Zhai Shumin, Smith Barton A. (2000). Gaze and speech in attentive user interfaces. In *Proceedings of the International Conference on Multimodal Interfaces*, volume 1948 from *LNCS*. Springer, 2000.

Matsusaka Y., Tojo T., Kubota S., Furukawa K., Tamiya D., Hayata K., Nakano Y., Kobayashi T. (1999). Multi-person conversation via multimodal interface -A robot who communicate with multi-user-. In *Proc. Eurospeech 99*, volume 4, pp. 1723–1726, Sep. 1999.

Meier Uwe, Stiefelhagen Rainer, Yang Jie (1997). Preprocessing of visual speech under real world conditions. In *Proceedings of European Tutorial & Research Workshop on Audio-Visual Speech Processing*, pp. 113–116, 1997.

Meier Uwe, Stiefelhagen Rainer, Yang Jie, Waibel Alex (1999). Towards unrestricted Lipreading. In Tang Y. (Eds.), *Proceedings of the Second International Conference on Multimodal Interfaces*, Hong Kong, China, January 1999.

Meier Uwe, Stiefelhagen Rainer, Yang Jie, Waibel Alex (2000). Towards Unrestricted Lipreading. *International Journal of Pattern Recognition and Artificial Intelligence*, 14(5):571–785.

Meyering A., Ritter H. (1992). Learning 3-D-shape perception with local linear maps. In *Proceedings of the International Joint Conference on Neural Networks*, Baltimore, MD, 1992.

Miller David J., Yan Lian (1999). Critic-driven ensemble classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10):2833–2844.

Mozer Michael (1998). The neural network house: An Environment that adapts to its Inhabitants. In *Intelligent Environments, Papers from the 1998 AAAI Spring Symposium*, number Technical Report SS-98-92, pp. 110–114. AAAI, AAAI Press, 1998.

Ong Eng-Jon, McKenna Stephen, Gong Shaogang (1998). Tracking Head Pose for Inferring Intention. In *Procceedings ECCV 98 Workshop on Perception of Human Action*, University of Freiburg, Germany, June 1998.

Pentland A., Moghaddam B., Starner T. (1994). View-based and Modular Eigenspaces for Face Recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994.

Perret D.I., Emery N.J. (1994). Understanding the intentions of others from visual signals: neurophysiological evidence. *Cahiers de Psychologie Cognitive*, 13:683–694.

Perzanowski D., Schultz A.C., Adams W., Marsh E., Bugajska M. (2001). Building a multimodal human-robot interface. *IEEE Intelligent Systems*, 16(1):16–21.

Polhemus (). http://www.polhemus.com.

Pomerleau Dean (1992). *Neural Network Perception for Mobile Robot Guidance*. PhD Thesis, Carnegie Mellon University, Pittsburgh, PA, February 1992.

Posner M. I., Snyder C.R.R., Davidson B.J. (1980). Attention and the detection of signals. *Journal of Experimental Psychology: General*, 109:160–174.

Rae Robert, Ritter Helge J. (1998). Recognition of Human Head Orientation based on Artificial Neural Networks. *IEEE Transactions on neural networks*, 9(2):257–265.

Rao R.P.N., Zelinsky G.J., Hayhoe M.M., Ballard D.H. (1995). Modeling Saccadic Targeting in Visual Searchd. In *Advances in Neural Information Processing Systems 8 (NIPS 95)*. MIT Press, 1995.

Richard M. D., Lippmann R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, 3:461–483.

Ries Klaus, Waibel Alex (2001). Activity Detection for Information Access to Oral Communication. In *Proceedings of the of the First International Conference on Human Language Technology Conference (HLT 2001*, San Diego, March 2001.

Rosenblum M., Yacoob Y., Davis L.S. (1996). Human Expression Recognition from Motion Using a Radial Basis Function Network Architecture. *IEEE Trans. on Neural Networks*, 7(5):1121–1138.

Rowley H. A., Baluja S., Kanade T. (1995). Human face detection in visual scenes. Technical Report CMU-CS-95-158R, CMU, 1995.

Rowley Henry A., Baluja Shumeet, Kanade Takeo (1998). Neural Network-Based Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38.

Ruusuvuori Johanna (2001). Looking means listening: coordinating displays of engagement in doctor-patient interaction. *Social Science & Medicine*, 52:1093–1108.

Salvucci Dario D. (1999). Inferring Intent in Eye-Based Interfaces: Tracing Eye Movements with Process Models. In *Human Factors in Computing Systems: CHI 99*, 1999.

Schiele Bernt, Waibel Alex (1995). Gaze Tracking Based on Face-Color. In *International Workshop on Automatic Face- and Gesture-Recognition*, pp. 344–348, 1995.

Schneiderman Henry, Kanade Takeo (2000). A Statistical Method for 3D Object Detection Applied to Faces and Cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 746–751, 2000.

Shiffrin R., Schneider W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84:127–190.

SMI (). SensoMotoric Instruments. http://www.smi.de/.

Soltau Hagen, Schaaf Thomas, Metze Florian, Waibel Alex (2001). The ISL Evaluation System for Verbmobil - II. In *ICASSP 2001*, Salt Lake City, May 2001.

SRR (). SR Research. http://www.eyelinkinfo.com/.

Stiefelhagen Rainer, Yang Jie (1997). Gaze Tracking for Multimodal Human-Computer Interaction. In *Proceedings of International Conf. on Acoustics, Speech, and Signal Processing*, April 1997.

Stiefelhagen Rainer, Zhu Jie (2002). Head Orientation and Gaze Direction in Meetings. In *Conference on Human Factors in Computing Systems (CHI2002)*, Minneapolis, April 2002.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (1996). A Model-Based Gaze Tracking System. In *Proceedings of IEEE International Joint Symposia on Intelligence and Systems*, pp. 304 – 310, 1996.

Stiefelhagen Rainer, Meier Uwe, Yang Jie (1997a). Real-Time Lip-Tracking for Lipreading. In *Proc. of Eurospeech*, Rhodes, Greece, 1997. lipreading, lip tracking.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (1997b). A Model-Based Gaze Tracking System. *International Journal of Artificial Intelligence Tools*, 6(2):193–209.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (1997c). Tracking Eyes and Monitoring Eye Gaze. In *Workshop on Perceptual User Interfaces*, Banff, Canada, October 1997.

Stiefelhagen Rainer, Finke Michael, Yang Jie, Waibel Alex (1998a). From Gaze to Focus of Attention. In Turk Matthew (Eds.), *Proceedings of Workshop on Perceptual User Interfaces: PUI 98*, pp. 25–30, San Francisco, CA, November, 4-6th 1998.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (1998b). Towards Tracking Interaction Between People. In *Intelligent Environments. Papers from the 1998 AAAI Spring Symposium*, Technical Report SS-98-02, pp. 123–127, Menlo Park, California 94025, March 1998. AAAI, AAAI Press.

Stiefelhagen Rainer, Finke Michael, Yang Jie, Waibel Alex (1999a). From Gaze to Focus of Attention. *Lecture Notes in Computer Science*, 1614:761–768.

Stiefelhagen Rainer, Finke Michael, Yang Jie, Waibel Alex (1999b). From Gaze to Focus of Attention. In Huijsmans D. P., Smeulders A. W. M. (Eds.), *Visual Information and Information Systems, Third International Conference, VISUAL '99*, number 1614 in Lecture Notes in Computer Science, pp. 761–768. Springer, June 1999.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (1999c). Modeling Focus of Attention for Meeting Indexing. In *Proceedings of ACM Multimedia '99*, pp. 3–10. ACM, 1999.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (2000). Simultaneous Tracking of Head Poses in a Panoramic View. In *International Conference on Pattern Recognition*, volume 3, pp. 726–729, September 2000.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (2001a). Estimating Focus of Attention based on Gaze and Sound. In *Workshop on Perceptive User Interfaces (PUI'01)*, Orlando, Florida, November 2001.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (2001b). Tracking Focus of Attention for Human-Robot Communication. In *IEEE-RAS International Conference on Humanoid Robots - Humanoids 2001*, 2001.

Stiefelhagen Rainer, Yang Jie, Waibel Alex (2002). Modeling Focus of Attention for Meeting Indexing based on Multiple Cues. *IEEE Transactions on Neural Networks. Special Issue on Intelligent Multimedia Processing.*

Stiefelhagen Rainer (1996). Gaze Tracking for Multimodal Human Computer Interaction. Master's thesis, Universität Karlsruhe (Technische Hochschule), Germany, http://werner.ira.uka.de/ISL.multimodal.publications.html, July 1996. Diplomarbeit.

Sung K., Poggio T. (1994). Example-based learning for view-bawsed human face detection. Technical Report 1521, MIT AI Lab, 1994. face detection.

Theeuwes J. (1993). Visual selective attention: A theoretical analysis. *Acta Psychologica*, 83:93–154.

Turk Matthew A., Pentland Alex (1991). Face recognition using eigenfaces. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–591, Maui, HI, USA, 1991.

Vertegaal Roel, Slagter Robert, Veer van der Gerrit, Nijholt Anton (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *SIGCHI'01*, Seattle, March 2001. ACM.

Vioal Paul, Jones Michael (2001). Rapid Object Detection using a Boosted Cascade of Simple Features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - CVPR 2001*, volume 1, pp. 511–518, 2001.

Waibel Alex, Bett Michael, Finke Michael, Stiefelhagen Rainer (1998). Meeting Browser: Tracking and Summarizing Meetings. In Penrose Denise E. M. (Eds.), *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pp. 281–286, Lansdowne, Virginia, February. 8-11 1998. DARPA, Morgan Kaufmann.

Waibel Alex, Bett Michael, Metze Florian, Ries Klaus, Schaaf Thomas, Schultz Tanja, Soltau Hagen, Yu Hua, Zechner Klaus (2001a). Advances in Automatic Meeting Record Creation and Access. In *ICASSP 2001*, Salt Lake City, May 2001.

Waibel Alex, Yue Hua, Soltau Hagen, Schultz Tanja, Schaaf Thomas, Pan Yue, Metze Florian, , Bett Michael (2001b). Advances in Meeting Recognition. In *Proceedings of the Human Technology Conference*, San Diego, 2001.

Wohler C., Aulanf J. K., Portner T., Franke U. (1998). A time delay neural network algorithm for real-time pedestrian recognition. In *Int. Conf. Intelligent Vehicle*, Germany, 1998.

Wolfe J. (1994). Guided Search 2.0: A Revised Model of Visual Search. *Psychonomic Bulltetin and Review*, 1(2):202–238.

Woods K., W.P. Kegelmeyer Jr., Bowyer K. (1997). Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410.

Wyszecki G., Styles W. S. (1982). *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley and Sons, New York, second edition.

Yang Jie, Waibel Alex (1996). A real-time face tracker. In *Proceedings of WACV*, pp. 142–147, 1996.

Yang Jie, Stiefelhagen Rainer, Meier Uwe, Waibel Alex (1998a). Real-Time Face and Facial Feature Tracking and Applications. In *Proceedings of Auditory-Visual Speech Processing Conference*, pp. 79–84, Terrigal, South Wales, Australia, December 1998.

Yang Jie, Stiefelhagen Rainer, Meier Uwe, Waibel Alex (1998b). Visual Tracking for Multimodal Human Computer Interaction. In Karat Clare-Marie, Karat John, Horrocks Ian (Eds.), *Human Factors in Computing Systems: CHI 98*, pp. 140–147, Los Angeles, CA, USA, April 1998. ACM SIGCHI, Addison-Wesley Pub. Co.

Yang Jie, Zhu Xiojin, Gross Ralph, Kominek John, Pan Yue, Waibel Alex (1999). Multimodal People ID for a Multimedia Meeting Browser. In *Proceedings of ACM Multimedia '99*. ACM, 1999.

Yarbus A. L. (1967). Eye movements during perception of complex objects. In Riggs L.A. (Eds.), *Eye Movements and Vision*, pp. 171–196. Plenum Press, New York, 1967.

Zhao L., Thorpe C. (1999). Stereo and Neural Network-Based Pedestrian Detection. In *Proc. Int'l Conf. on Intelligent Transportation Systems*, Tokyo, Japan, October 1999.