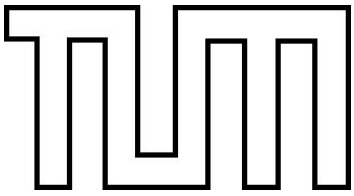


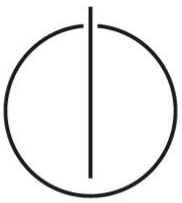
FAKULTÄT FÜR INFORMATIK

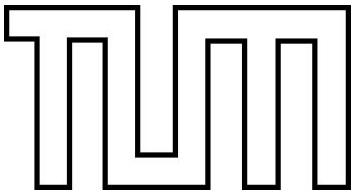
DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Diplomarbeit in Informatik

Active Vision for Interactive Spaces

Henning Herbers





FAKULTÄT FÜR INFORMATIK

DER TECHNISCHEN UNIVERSITÄT MÜNCHEN

Diplomarbeit in Informatik

Active Vision for Interactive Spaces

Aktives Sehen in Interaktiven Räumen

This thesis was conducted in collaboration with the
Insitut Universitari de l'Audiovisual
Universitat Pompeu Fabra, Barcelona

Author:	Henning Herbers
Supervisor:	Prof. Gudrun Klinker, Ph.D.
Advisor:	Dipl.-Inf. Daniel Pustka
	Dr. Sergi Bermudez i Badia (UPF)
Submission Date:	February 15, 2008

I assure the single handed composition of this diploma thesis only supported by declared resources.

Garching, February 15, 2008

Henning Herbers

Abstract

Active Vision goes beyond plain sensing technology and includes strategies for observation. Rather than just processing snapshots, the observer and sensor continually interact to purposefully analyze visual sensory data and answer specific questions posted by the observer. Subject of this thesis is the exploitation of the Active Vision paradigm in interactive spaces. A particular example for such a space is the eXperience Induction Machine (XIM). The human accessible mixed reality space is run at the Institut Universitari de l'Audiovisual (IUA) in Barcelona to enable research applications in the field of mixed reality using biologically inspired models of sensor and effector systems. In this environment, a major challenge lies in the tracking of people by a multi modal tracking system, fusing the input of multiple sensors and maintaining a model of the space in real time.

The specific Active Vision task examined in this thesis is the use of four wall-mounted, movable pan and tilt cameras (gazers) to gain additional information about a person located at a position signalized by the tracking system. This additional acquired information would be helpful, especially if tracking data can not be unambiguously assigned to a specific object contained in the model. Since this deployment of the gazers presupposes an adequate calibration of the individual devices, this thesis emphasizes different calibration techniques and evaluates their strengths and weaknesses. I will introduce an innovative approach that estimates a gazer's pose by finding an optimal fit of the desired parameters to a set of priorly collected correspondences between tracking positions and gazer angles. In order to keep the calibration as independent as possible of any secondary modality, the gazer's pose is computed only from a standardized global tracking input and the gazer's own image scanned by a classifier to detect objects. For verification of this calibration technique and for comparison of the results to those of a state of the art method, a marker based pose estimation was implemented and evaluated. A bundle adjustment thereby serves to globally optimize the poses of the gazers, taking all involved parameters and error sources into consideration. The evaluation of the different calibration techniques allows to draw a conclusion on the accuracy that can be achieved in adjusting the gazers to look at any position in the space. A successfully calibrated system will serve reliable image data for further processing, allowing the assignment of unique attributes such as hue, color or size to the focused object. As an example, I will consider the generation of hue histograms over a specific region of interest. Several distance measures for histogram comparison will be introduced, to see whether a hue histogram carries enough information to distinguish between persons in the space.

Zusammenfassung

Anstatt sich ausschließlich auf die Verarbeitung statischer Bilder zu beschränken, umfasst das Prinzip der Active Vision auch den aktiven Einsatz visueller Sensoren zur Untersuchung bestimmter Ziele. Beobachter und Sensoren tauschen sich dabei stetig untereinander aus, um spezifischere Informationen generieren zu können. Ziel dieser Arbeit ist die Nutzung dieses Ansatzes in interaktiven Räumen. Als Anwendungsfall wird die eXperience Induction Machine (XIM) betrachtet, die am Institut Universitari de l'Audiovisual (IUA) in Barcelona betrieben wird. In diesem begehbaren interaktiven Raum sollen die Möglichkeiten der Mixed Reality auf der Basis biologisch inspirierter Modelle erforscht werden. Eine wesentliche Herausforderung liegt dabei im Tracking, der genauen Positionsbestimmung einzelner Personen im Raum. Ein multi modales Tracking System wird eingesetzt, um die Informationen der einzelnen Sensoren zu verarbeiten und ein Modell des Raums in Echtzeit zu pflegen.

Im Rahmen dieser Arbeit sollen vier schwenkbare Kameras, die Gazer, benutzt werden um zusätzliche Informationen über bestimmte Personen in der XIM zu gewinnen. Diese Informationen könnten insbesondere dann hilfreich sein, wenn zweifelhafte Sensordaten keinem Objekt im Modell eindeutig zugeordnet werden können. Grundvoraussetzung für den Einsatz der Gazer zu diesem Zweck, ist eine genaue Kalibrierung der einzelnen Geräte. Schwerpunkt dieser Arbeit sind deshalb die Umsetzung und Auswertung verschiedener Kalibrierungstechniken. Es wird ein neuer Ansatz vorgestellt, bei dem die extrinsischen Parameter eines Gazers als optimal passende Konfiguration für einen Satz zusammengehörender Paare von Zielpositionen und Gazerausrichtungen gefunden werden. Um die Kalibrierung dabei so weit wie möglich unabhängig von weiteren Modalitäten zu halten, basiert diese nur auf einem globalen Positionssignal und dem eigenen Bild des jeweiligen Gazers. Zur Bestätigung dieses Ansatzes und zum Vergleich der Ergebnisse mit denen einer anerkannten Methode, wurde eine konventionelle Markerkalibrierung implementiert und ausgewertet. In einem Bundle Adjustement werden dabei alle involvierten Parameter und Fehlerquellen berücksichtigt. Auswertung und Vergleich der verschiedenen Ansätze lassen einen Rückschluss auf die mögliche Genauigkeit zu, die bei einer Ausrichtung der Gazer auf bestimmte Stellen im Raum erzielt werden kann. Ein erfolgreich kalibriertes System erlaubt die Gewinnung zuverlässiger Bilder für die weitere Verarbeitung und den Vergleich verschiedener Personen im Raum anhand bestimmter Attribute. So lassen sich zum Beispiel Histogramme bestimmter Bildbereiche erzeugen und vergleichen. Mehrere Abstandsmetriken zum Vergleich zweier Histogramme werden vorgestellt und ausgewertet, um Aufschluss darüber zu geben ob die Informationen ausreichen, um zwei Personen im Raum zu vergleichen.

Acknowledgments

I would like to take the chance and thank everybody who helped me along the way. A special thank goes out to my supervisor Daniel Pustka, who was always there to assist and consult me in any possible way. Lots of thanks also to my professor Gudrun Klinker, for introducing me to the wonderful research field of Augmented Reality and giving me the possibility to write this thesis at her chair. Not to mention the whole SPECS group of the Institut Universitari de'l Audiovisual, especially Zenon Mathews, Sergi Bermudez i Badia and of course Professor Paul Verschure. Thank you for hosting me in Barcelona, for all your support and for giving me the opportunity to spend a great time with you.

Contents

Overview

Before starting with the main content, I would like to give a short overview over the context of this diploma thesis and explain its goals.

The eXperience Induction Machine

The eXperience Induction Machine (XIM) is a human accessible mixed reality space run at the Institut Universitari de l'Audiovisual (IUA) in Barcelona to enable research applications in the field of mixed reality using biologically inspired models of sensor and effector systems. To interact with its visitors, the space is equipped with 72 light emitting floor tiles, that also serve as sensors. Their weight information complements the visual data provided by a infrared camera installed in the ceiling in order to track objects in the room, as well as the data from three triangularly arranged microphones. Inside the space, the visitor is surrounded by projection screens, exposing him to interactive content. Eight movable theater lights can be used for light effects and indication of certain spots, while four wall mounted movable pan-tilt cameras (gazers) may be used to gaze at certain spots and provide an online image.

In this environment, a major challenge lies in the tracking of people in the space. A multi-modal tracking system (MMT) is used to fuse the input of multiple sensors and maintain a model of the space in real time. In its present configuration, with the overhead camera and the pressure sensors in the floor deployed, the MMT encounters difficulties in tracking multiple objects, especially when these move close to each other. Measures to improve the performance include the dynamic filtering of the respective data streams, as well as more sophisticated ways of data fusion. Another approach would be to add a further modality to the setup to provide complementary information that allows the creation of a distinctive model.

Objective of this project is the use of the four gazers to gain additional information about certain entities in the space and provide it to the MMT as additional input. On request the gazers shall be set to look at a desired position and extract additional information, that may be assigned to or compared with an object in question. The acquired data could be helpful, especially if tracking data can not be unambiguously assigned to a specific object contained in the model.

A detailed description of the XIM infrastructure, the tracking system deployed and the specific problem addressed in this project will be given in chapter 1.

Problem Statement

On request the gazers shall look at a defined spot to extract the asked information. In order to do so, the most appropriate gazer needs to be chosen and then set to look in the right

direction. The adequate view direction of the specific gazer can thereby be expressed by two degrees of freedom, the pan and the tilt angle. The pan angles describes the gazers rotation around its main axis, while the tilt angle defines the nod of the camera head. These angles can easily be trigonometrically computed if, apart from the tracking data defining the position in question, the gazer's extrinsic parameters are known. The extrinsic parameters of the gazers are defined by its position and orientation in three dimensional space, when both the pan and the tilt angle are set to zero. While various approaches exist to find these extrinsic parameters, in a prior calibration scenario as well as by a dynamic learning process, this thesis introduces an innovative approach. To keep the calibration as independent as possible of any other means apart from the tracking system, the extrinsic parameters are computed from a set of correspondences between tracking positions and respective gazer angles gained in a calibration scenario. Rather than estimating the position of the gazer in relation to some predefined coordinate system, this setup yields the position in correspondence to the coordinate frame spanned by the tracking system. Keeping in mind that the tracking data is given in exactly this coordinate frame, this is of an enormous advantage. In order to judge on the performance of the calibration and the accuracy of the poses, the results are compared to those of a state of the art marker based calibration method.

Classifier Based Pose Estimation

The extrinsic gazer parameters are computed by finding an optimal fit to an arbitrary set of correspondences between tracking positions and the respective gazer angles. This set of correspondences defines a system of non-linear equations, for which an optimal solution can be found using a Levenberg Marquard Optimizer. To get the required correspondences, an initial calibration scenario is necessary to find the right gazer angles for a number of tracking positions. To get this data the room is scanned with the respective gazer looking for an unmoved person, whereby it follows a search path varying in both rotations. The lightning conditions are kept optimal, to allow the detection of the person in the gazer image by a classifier. The gazer can thereby be adjusted to look straight at the person, whose position is simultaneously tracked by the overhead infrared tracking system of the XIM. Once the gazer is adjusted yielding an image with the person in the center, its angles and the respective tracking position are recorded. The whole process is repeated for different positions in the room.

A sufficient number of correspondences found by the calibration method described above allow the optimization of the desired camera parameters to provide an input for the computational model to determine needed gazer angles online. This optimization is done by a non-linear Levenberg Marquard Optimizer, optimizing position (x, y, z) and orientation of the gazer (yaw, pitch) from a minimum of four correspondences. The optimization was tested for various sets of correspondences to determine a minimal (sufficient) number of pairs required to guarantee satisfying results.

Chapter 2 introduces this innovative approach of extrinsic pose estimation and discusses it's performance and results.

The Control Experiment

For verification of the results from the calibration and for comparison to a state of the art calibration method, a control experiment was run, implementing a marker based calibration. An interactive marker equipped with infrared LEDs was placed in the room, allowing the pose estimation of both the gazer and the infrared overhead tracking camera relative to the marker. In combination these poses allow the estimation of the gazer position and orientation relative to the overhead tracking system, which serves as reference coordinate system for the classifier calibration.

The pose of the gazer in relation to the marker was computed using the Ubitrack framework, a powerful framework for ubiquitous tracking applications develop at the Chair for Computer Aided Medical Procedures and Augmented Reality (CampAR) of TU München. To estimate the pose, the intrinsic camera parameters and distortion coefficients of the specific gazer need to be known. These where determined using the chessboard calibration pattern also implemented in the Ubitrack framework. To allow the estimation of the pose between the marker and the infrared tracking camera installed in the ceiling of the XIM, the marker was equipped with infrared LEDs at five points. After an intrinsic calibration analogous to the one done for the gazers, the pose of the infrared camera relative to the marker could be computed from its images showing the position of the LEDs. To compensate for errors in the marker detection the pose estimation was done for ten different marker positions. The results where then averaged, yielding the estimated position of the gazer in the infrared coordinate system. In a more sophisticated approach, a bundle adjustment implemented in the Ubitrack framework was modified to globally optimize all involved parameters. This optimization method comprises the measurements of both pose estimations, between gazer and marker as well as between overhead camera and marker, and tries to find an optimal fit for all involved parameters.

The results of the control experiments and a comparison to the classifier approach are discussed in detail in chapter 3 of this project.

Attribute Extraction

Using the parameters determined, the gazer pan and tilt angles can be computed online for any given tracking position. A simple protocol for communication with the multi modal tracking system was defined. In a sample application, a saliency map is created to trigger an adequate gazer to look at the point of interest, once a tracked object stands by itself and is not moving. A region of interest is selected and extracted from the gazer image in respect of the distance between gazer and object. This image region will now serve as input for further processing. A first approach of gaining unique information from this region of interest is the creation of a hue histogram with an arbitrary number of bins. In a test scenario, such a histogram was generated for various images taken from all gazers looking at distinguishable persons standing at different positions in the room. The gazer was priorly set to look at the respective position indicated by the overhead tracking system as described above. To determine whether a hue histogram contains enough information to clearly differentiate between the persons, different mathematical approaches of histogram comparison where tested and evaluated.

The description of this experiments and a discussion about its results can be found in

chapter 4.

This thesis finishes in chapter 5 with a recap of the results and the insights gained within the course of the project. Advantages and disadvantages of the different calibration approaches are discussed and an outlook on further work to be done is given.

1. Introduction

In this chapter I will give a description of the context of this thesis and derive the concrete problem statement. After an introduction to the eXperience Induction Machine and the multi modal tracking deployed, my approach of gaining complementary information by use of the gazers will be elucidated in detail. This will lead us to the importance of an accurate camera calibration, an aspect that will be the main subject of this thesis. I will present the different approaches of pose estimation that will be discussed in the following chapters and provide a guideline through the project.

1.1. Environmental context

1.1.1. the eXperience Induction Machine

Introduction

The eXperience Induction Machine (XIM) is a human accessible mixed reality space run by the research group for Synthetic Perceptive, Emotive and Cognitive Systems (SPECS) at the Institut Universitari de l'Audiovisual (IUA) in Barcelona. The room, equipped with a wide range of sensors and effectors, is designed as a general purpose infrastructure to investigate human-artifact interaction and to conduct experiments in mixed reality. XIM is an abstraction and further development of its predecessor, the installation "Ada - the intelligent space", that was build for the Swiss national exhibition Expo02, a fair that hosted over 560.000 visitors over a period of six month. The neuromorphic design and functionality of Ada is described in [27]. Specific research questions within the XIM environment include how a spatial enclosure can affect and interact with its visitors, how humans can act, exist and behave in both physical and virtual spaces, the construction of socially capable and believable synthetic characters and the development of a framework for interactive narratives [12]. While such an installation can also be implemented as a pure input/output device, the conceptual design of the space as an autonomous entity is one of the key features of XIM. It thereby sets itself off against other mixed reality spaces such as the Allosphere at UCSB, the Intelligent House at MIT, the Nanohouse at UTS and the Sentient Lab run by the Faculty of Architecture at the University of Sydney [12].

Infrastructure

XIM in its present configuration is a square room with a 5.5 by 5.5 meter surface and a height of 4 meters. All major instruments are mounted in a rig constructed from a standard truss system. The design of this prototype space is modular in both physical and technological aspects, so that it can easily be expanded to a larger interactive space planned to be deployed as a permanent exhibition at the communication campus of UPF 22@BCN.

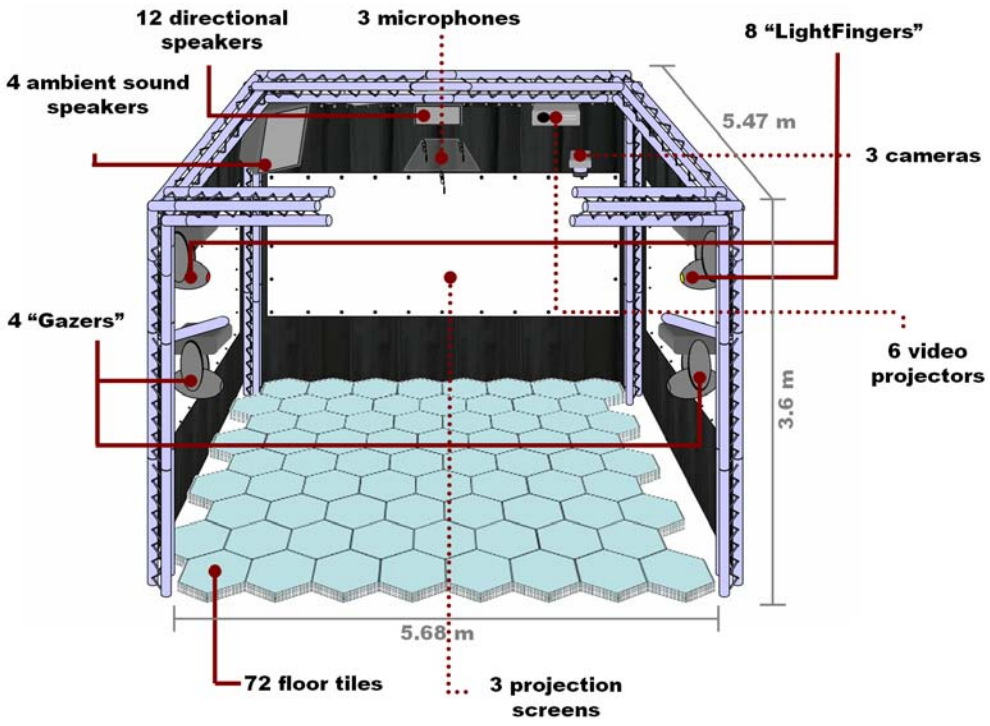


Figure 1.1.: The eXperience Induction Machine

Figure 1.1 shows a scheme of the room in its present state. It is currently equipped with the following devices:

- Two cameras at the top of the rig provide a "bird's eye perspective", that serves as input for the multi modal tracking system. One of the cameras is thereby equipped with a standard infrared filter to deliver infrared images to the system. A major infrared source light is used to enlighten the space with infrared rays.
- Three microphones are attached to the center of the rig and might be used as auditory input source for visitor localization and to recognize specific sounds.
- Eight steerable theater lights (Martin MAC250, Arhus, Denmark), furthermore referred to as "LightFingers", are attached to the top boundary of the rig for light effects from all sides.
- Four steerable color cameras ("gazers"), mechanical constructions adapted from the Martin MAC250 theater lights and equipped with camera blocks from Sony to replace the light bulbs. The gazers are mounted in the corners of the space at head-height to get images of the visitors from all directions and angles.
- A total of 16 speakers with the corresponding sound equipment (MIDI sampler, matrix mixer, amplifiers) provide spatialized sound, while a PA system is used to present soundscapes.

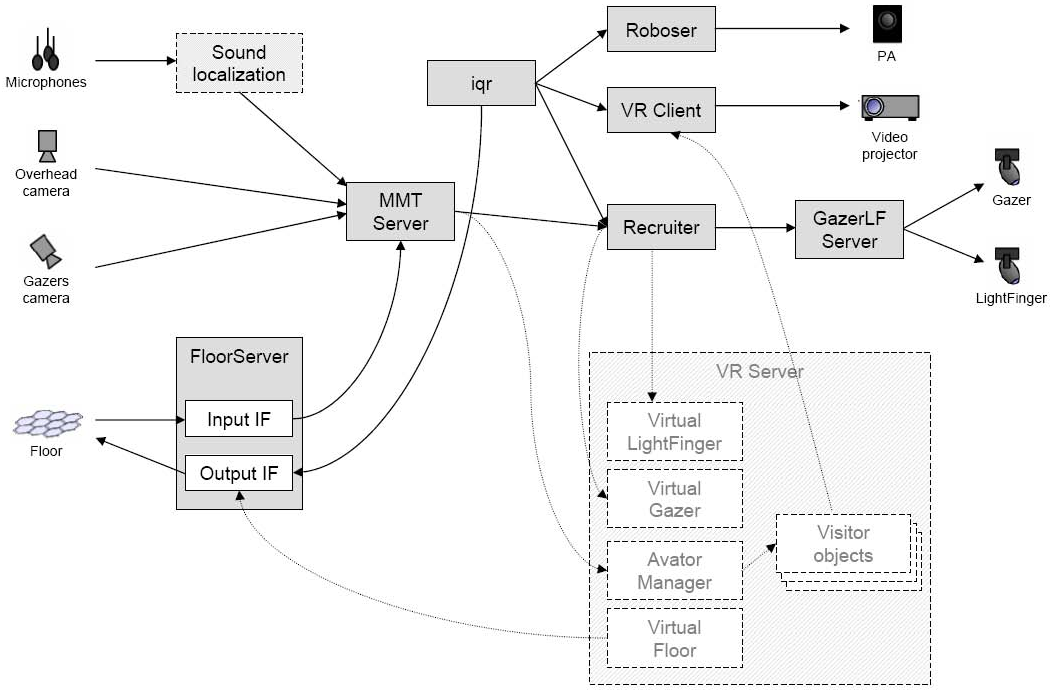


Figure 1.2.: **Simplified XIM connectivity diagram.** *Image adapted from [12]*

- Six video projectors are used to display visual content on the white screens that surround the space.
- A custom construction of 72 hexagonal shaped interactive tiles [25] constitute the floor of the space. The tiles are used as sensors as well as effectors. Equipped with pressure sensors they can provide weight information of the visitors. At the same time, each floor tile incorporates individually controllable RGB neon tubes, which permits the display of patterns and light effects.

Currently XIM is controlled by more than 16 computers, implementing the different subsystems such as sonification, the tracking system or the virtual environment. Figure 1.2 shows a simplified overview over the system architecture. The development of the hardware and software infrastructure was thereby based on the existence of two representations of the XIM, a real and a virtual. This design is emphasized by the aim to create a persistent virtual community (PVC), that will be described in detail later in this chapter. To allow the permanent existence of the persistent virtual community despite its transient and indirect coupling to the physical XIM, the implementation of the system underlies two principles: Real and virtual XIM are considered the same, and the cognitive component of XIM is decoupled from its physical and virtual representations [11]. An emphasis is placed on the first maxim, the functional equivalence. In the virtual counterpart of the XIM effectors have to have the same functional effect on visitors and not necessarily be a faithful representation of the physical device. This functionality is considered crucial for creating a

coherent interaction in the mixed reality environment. The second maxim implies that any event occurring in the real or virtual space is indistinguishable for XIM's cognitive system. In this context, XIM becomes a gateway to a larger virtual environment, the PVC.

The design philosophy of the XIM infrastructure is to reduce complexity in order to cope with these rather complex tasks. Therefore the former system design of Ada was modified, to fulfill requirements such as full scalability or easy communication between the autonomous components. Each of them is assigned with a restricted, well defined task. The communication interfaces are kept thin, i.e. as few commands as possible are transmitted via UDP sockets rather than relying on remote procedure calls (RPC) as it was the case for Ada.

A XIM application

A major application developed in the XIM environment is the Persistent Virtual Community (PVC), which is one of the main goals of the PRESENCCIA project. PRESENCCIA is an integrated project funded under the European Sixth Framework Program, Future and Emerging Technologies (FET), which is tackling the phenomenon of subjective immersion in virtual worlds from a number of different angles [12]. Within the PRESENCCIA project, the PVC serves as a platform to conduct experiments on presence, in particular social presence in mixed reality. The PVC makes use of all aspects of XIM as a mixed reality platform, where entities of different degrees of virtuality can meet and interact. This includes real visitors physically present in the installation, avatars as alter egos of remote visitors and fully synthetic characters controlled by neurobiologically grounded models of perception and behavior.

The mixed reality world of the PVC consists of the Garden, the Cave and the Avatar Heaven (Figure 1.3). The Garden of the PVC is a model ecosystem, which changes its development and state depending on the interaction with and among its visitors [61]. The Clubhouse is a building in this environment and houses the virtual XIM. The virtual version of the XIM is a direct mirror of the physical installation, any events and output from the physical installation are represented in its virtual counterpart and vice versa. This means e.g. that an avatar crossing the virtual XIM will also be represented in the physical installation. Conceptually the physical installation is embedded into the virtual world.

Access to the PVC is given via three portals: Visitors can enter the virtual world physically through the XIM, or virtually either by way of a Cave Automatic Virtual Environment (CAVE) or via the internet from a PC (Figure 1.3). In future the mixed reality installation of the PVC is meant to be open to the general public and will thereby provide a showcase for the key technologies developed in the PRESENCCIA project. In this content, XIM fulfills a double role: On the one hand, the XIM is an interface to the virtual world and hence allows visitors physically present in the room to interact with avatars and synthetic characters. On the other hand, the room has a "ghost in the shell" [12], meaning that it is an autonomous, sentient entity which is engaging in interactions with its visitors and is actively monitoring and modulating their behavior. This design is emphasized by the aim to use XIM to study collective and social presence where groups of visitors share the same frame of reference. While the CAVE can be used by a single user to access the PVC, XIM is an effective approach to allow multiple users to access the space simultaneously. Bernadet [12] describes two major facets of social presence that can be explored in this con-

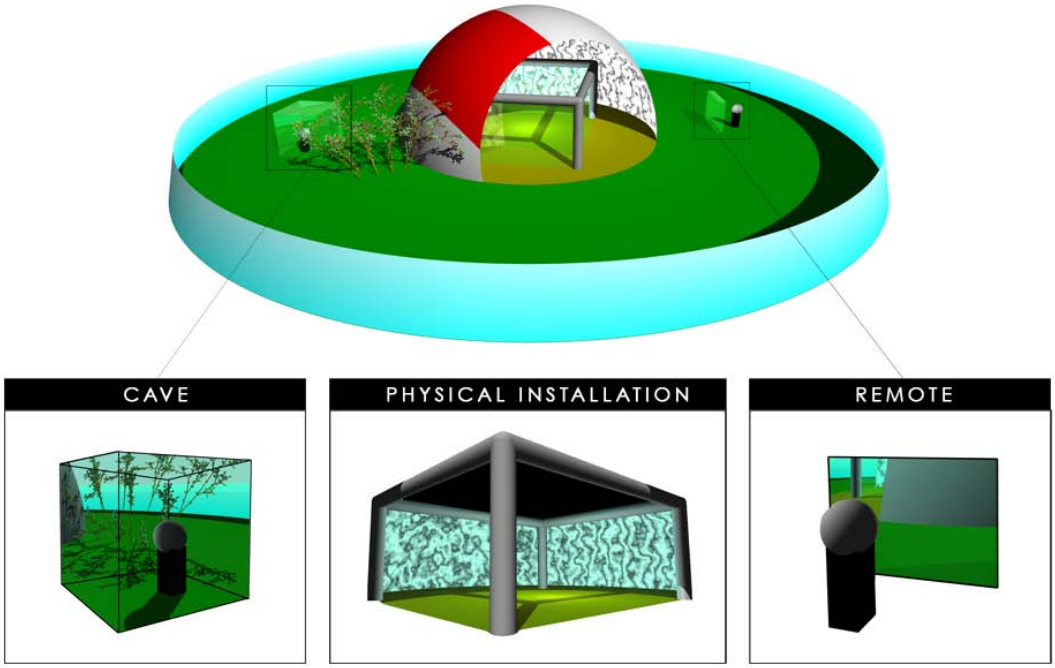


Figure 1.3.: **Layout of the access to the Persistent Virtual Community.** *Image adapted from [12]*

text. Firstly the facet of the perception of the presence of another entity in an immersive context, which depends on the credibility of the entity the visitor is interacting with. In the case of XIM as an entity within the PVC, the credibility of the space is affected by its potential to act and be perceived as a sentient entity and deploy believable characters. The credibility of the synthetic characters representing a user in the CAVE depends on its validity as an authentic antromorphic entity. This could be i.e. the preservation of presence when the synthetic characters change their form of representation. The second facet of social presence covers the collective immersion experienced in a group, opposed to being a single individual in a CAVE. For this purpose the XIM offers a unique platform, since the size of the room permits the hosting of mid-sized groups of visitors.

1.1.2. Multi modal tracking in the XIM

Mixed reality applications like the XIM enable interactions between real, synthetic and virtual characters of diverse nature and quality depending on the technologies available. A major challenge in achieving this interactivity lies in the accurate tracking of the physically present persons or objects in the space. Many interaction scenarios, like interactive games or virtual societies, demand precise information about the position of all entities involved. Also within the context of the research intended to be done in the XIM, the exploration of social presence, accurate tracking is considered a necessity. From the perspective of a real person in a mixed reality space, a sense of presence is generated from the feeling

that he/she exists within the space as a separate entity. The experience of individualism can be enhanced if other existing entities of the virtual world react on ones presence [33, 17]. Subjective personal presence thereby gives a measurement why and to what extend a person feels he/she is in a virtual world [33]. Social presence refers to the extend to which other beings, real or synthetic also exist in the virtual world and in how far they react to a person in the space [33]. It derives from conversing with other humans and the interaction with synthetic entities in the environment. If a person feels he/she is being recognized by someone or something, it is easier for himself/herself to believe that he/she is actually there [17]. Analogously, environmental presence refers to the extend to which the environment itself appears to know that there is an human entity present and can therefore react on its presence.

All of these notions described above are feasible only if information about the position of each individual human visitor of the space is available at any given time, making real time tracking indispensable. Often, as in the case of XIM, single modal person tracking in mixed reality spaces is very difficult as each of the different sensors might have high errors [7]. Therefore tracking in the XIM is realized by a multi modal tracking system (MMT) that can revert on the input of numerous sensors, currently an infrared camera mounted to the ceiling for visual tracking and weight information from each individual floor tile. It seems obvious that, if handled right, additional sensor information can compensate for weaknesses of other modalities and contribute to a more stable and accurate tracking. Therefore one possible approach to enhance the tracking performance of the MMT lies in the use of the individually controllable pan-tilt cameras, the gazers, to provide complementary information. The exploitation of such *selective attentional* mechanisms is the main motivation for this project and will be discussed in the following sections.

The image data from the infrared overhead camera is currently being processed by an adaption of the 3D tracking system AnTS, that was originally developed for behavioral analysis of flying insects and robots [10]. To determine objects in the image, a priorly captured background image is subtracted from the current camera image. The resulting image is then processed by means of simple thresholding, noise reduction and centroid computation to yield the center of all objects that can be silhouetted against the floor. Figure 1.4 shows the captured background (1) and the image at the different stages of the process-

Figure 1.4.: **The image processing pipeline of the 3D tracking system AnTS.** The background image (1) is subtracted from the current camera image (2) to reveal all objects that set themselves off from the floor (3). (4) shows the image after processing, whereby each ellipse indicates the boundaries of an object.

ing pipeline (2,3,4). The coordinates of the objects are then transferred to the MMT, where they serve as major input source, supplemented by the input data from the floor tiles. Each floor tile is equipped with three pressure sensors at opposing corners. The gain of each of theses sensors is constantly written into a shared memory, where from it is extracted by a floor client and forwarded to MMT along with the position of the tile and sensor that caused the output.

As a neuroscientific institute the SPECS group bases their approach for solving data association and fusion tasks on brain mechanisms. The association of different sensory cues with external objects or events, their registration, processing and the subsequent generation of motor commands are critical for the survival of animals. Neurophysiological research suggests that the superior colliculus (SC) is one of the primary areas for sensory data association and appropriate motor action generation for orienting response toward the source of stimulation [51]. It has been shown that the SC possesses sensory maps for individual sensors, from which motor maps for motor action generation are formed. Bayes' rules have been successfully used to model multi-sensory fusion as exhibited by the SC [4]. Further high-level modulation of sensory information could possibly be a key aspect in sensor data processing using limited resources [51]. Based on these considerations, a top-down modulation of bottom-up sensory information has been developed for integrating and deploying the different sensors in the XIM based on Bayesian inference. A first version of this tracking framework is in use, featuring dynamical recruitment of sensors and effectors to enhance tracking and resolve conflicting data. Figure 1.5 shows the concept of top-down modulation of bottom-up sensory information as it is realized for the MMT of the XIM.

1.2. Objective

As described in the previous section, single modal person tracking is in general fragmentary and inefficient, as any sensor has errors and might fail in certain situations. Along with other means, as dynamic filtering of the tracking data, the allocation of additional information might be useful to correct inaccuracies and enhance the tracking performance. This information can be of different nature and either be constantly provided to the tracking system (bottom-up) or requested if needed (top-down). As shown in Figure 1.5 the multi modal tracking system of the XIM implements both bottom-up and top-down sensory information processing. The results of the fusion of all available tracking data, for instance from the overhead visual tracking or the sensory information from the floor, are thereby used to update the world model but also to actively deploy certain sensors to deliver additional and more precise information if needed.

Objective of this project is the use of the four wall mounted pan-tilt cameras (gazers) to gain complementary information about entities in the space and thereby assist the tracking system. This approach is motivated by weaknesses of the existing solution if more than one person is present in the room. Especially if two or more objects move close to each other, difficulties occur in distinguishing between them and assigning ambiguous tracking data to a certain entity. To solve these ambiguities, a detailed model of the world with a list of all entities currently present in the room shall be maintained. The list can constantly

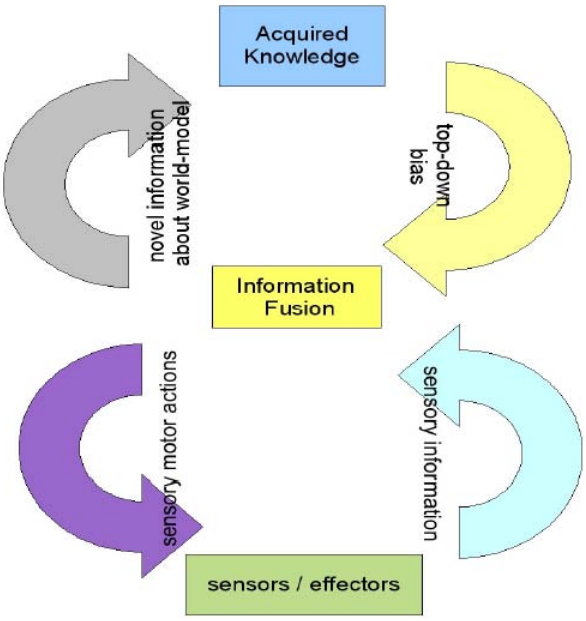


Figure 1.5.: **Top-down modulation of bottom-up sensory information as implemented in the MMT.** The individual sensors deliver data to the data fuser. The data fusion process is modulated by the input from the world-model. The result of the data fusion is then used to update the world-model and also to actively deploy sensors and effectors. *Image adapted from [39]*

be updated with attributes about each single entity. These may be for instance color information or physical appearance, but also more sophisticated measures are thinkable. The gazers might be used to constantly look at persons in the space and thus ”learn” about their features to provide these attributes to the tracking system (bottom-up), but also be triggered by the system if there is uncertain data that can not be unambiguously assigned. In this case an appropriate gazer shall be chosen to look at the spot where the tracking data in question came from and return a variety of attributes to compare the object that emmited the signal to those in the list of existing objects. As a side effect, the active deployment of sensors and effectors will make human participants in the space feel that the space knows where he/she is and that it tries to get to know him/her better. Such display of attention allocation may considerably contribute to an enhanced feeling of social presence as described in chapter 1.1.2.

We will see that the most important premise for a successful deployment of the gazers is an accurate extrinsic calibration. In order to be adjusted to look at a specific spot in the space, the position and orientation of each gazer need to be known. The correctness of all further processing will be strongly dependent of these parameters. Main subject of this thesis will therefore be the pose estimation of the individual gazers and the performance in computing the respective angles that may be achieved with the individual solutions.

1.3. Design issues and definitions

1.3.1. General design issues

In order to adjust the gazers in space, we need to know their poses. The attribute extraction system shall be provided with these poses as an input, without further limitations on how they were estimated. Nevertheless, this estimation is crucial for the correctness of the adjustment. In the course of this project, I will introduce two essentially different approaches of pose estimation and evaluate their performance. Both of them, however, will be implemented as standalone applications and yield the desired parameters as an output according to a predefined syntax. As the accuracy of a later adjustment depends mainly on the choice of the pose parameters, I will but a strong emphasis on the implementation and comparison of the different approaches.

1.3.2. Attribute extraction

It was agreed that the final attribute extraction system shall be designed as a pure pull-application, meaning it should only perform action if triggered by the data fuser. A realization as a push- and pull-application would also be possible, whereby the system would constantly try to gain information and pass it to the data fuser if idle. This would make the whole application much more complex, since it would have to choose points of interest by itself and therefore would need to be aware of the whole world model and all tracking data. Only so it could decide where a person stands motionless and by itself, making it possible to deliver reliable information. It is therefore more reasonable to leave these allocation tasks to the data fuser. The task to be performed by the system in first instance is therefore quite simple. The application will be triggered by the data fuser software with the tracking position specifying the point of interest and the id of the most appropriate gazer being passed according to a specified protocol. On execution, the respective gazer has to orient itself to look at even this spot and extract the requested attribute from its camera image. This attribute will then be returned to data fuser for further handling.

1.3.3. The gazers

The gazers are individually steerable color cameras, mechanical constructions adapted from the Martin MAC250 theater lights that are also in use as LightFingers in the XIM. The light bulbs have been replaced with Sony CCD color camera blocks with a resolution of 768 x 576 pixel (PAL) and a 40x zoom ratio (10x optical and 4x digital). The four gazers are mounted to the rig on head height in the four corners of the XIM (1.6, left) to provide an optimal view on every spot in the space. Connected in series along with the eight LightFingers, they can be individually addressed by an unique ID via the industry standard DMX protocol. In the XIM infrastructure, a server (furthermore referred to as *GazerLFServer*) is constantly running to parse incoming UDP packets and forward the appropriate commands to the respective devices via DMX. For the gazers these commands can be limited to the setting of the absolute orientation, more precisely the *pan* and the *tilt* angle. The *pan* angle defines the gazer's rotation around its main axis while the *tilt* angle specifies the nod of the device's head (1.6, right). The zoom can not be controlled via the

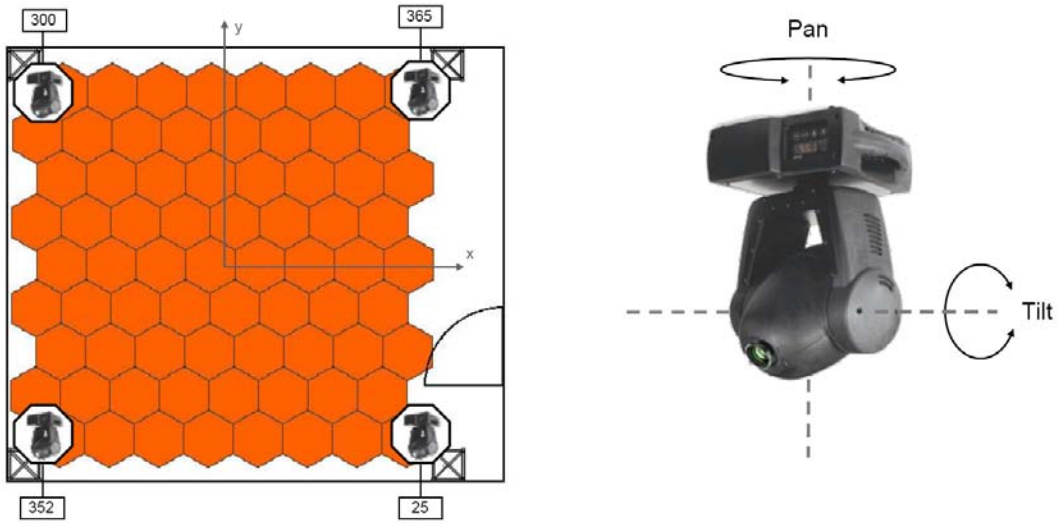


Figure 1.6.: **Gazers in the XIM.** The individually controllable pan and tilt cameras, the "gazers" (right figure), are mounted to the rig on head height in the four corners of the XIM (left figure).

DMX protocol but has to be set separately via the Visca protocol.

In the course of this project I will refer to the individual gazers with their DMX IDs: Gazer 25, gazer 300, gazer 352 and gazer 365. The locations of the individual devices in the XIM are shown in Figure 1.6.

1.3.4. Angle definitions

The pan and tilt angles necessary for a gazer to look at a specific position in the space can be computed trigonometrically. The computational model for these angles is shown in figure 1.7. Necessary for the computation are, along with the tracking data of the spot in question, the gazer's position and orientation in 3D space. The orientation is thereby defined by two angles, furthermore referred to as *yaw* and *pitch*. The *yaw* angle specifies the gazer's offset in rotation around the z-axis of the reference coordinate system as its pan value is set to zero. Analogous the *pitch* angle specifies the offset in rotation around the x-y-plane as the tilt is set to zero. Assuming a correct tracking position T, the accuracy of the computation and therefore the correctness of the resulting image is strongly dependent of these parameters, more precisely the 3D pose of the gazer's base. The estimation of this pose for an arbitrary gazer will be one of the key aspects discussed in this project. While it suggests itself to define the XIM floor as the reference coordinate system and measure the position of the gazers straight forward, this would yield an unprecise pose. In addition the tracking coordinate system is spanned by the overhead infrared camera. Given the floor as reference this would add another coordinate transformation between the floor and the overhead camera to the computation. To avoid this inconvenience, it is more apparent

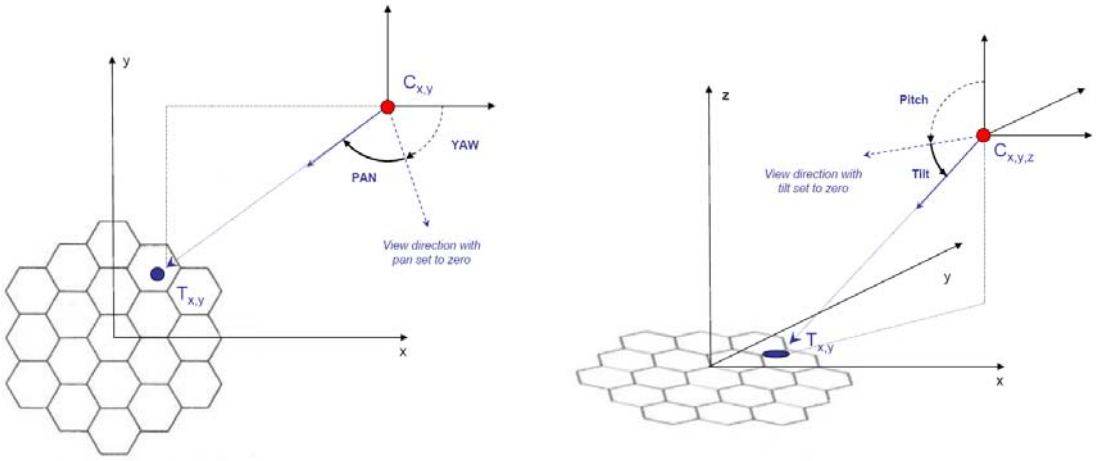


Figure 1.7.: The model for the computation of the pan and tilt angles for a gazer to look at a specific position *T*. Mandatory are, along with the tracking data *T*, the gazer's position *C* and orientation in 3D space, furthermore referred to as *yaw* and *pitch*.

to take the coordinate system of the overhead infrared camera as the reference coordinate system. This project introduces an innovative approach of estimating a camera's extrinsic pose from a set of correspondences between tracking data and gazer angles determined for a gazer to look at exactly the position indicated by the tracking data. This has the advantage of returning the gazer-poses already in the tracking coordinate system, while being independent of any other modalities other than the tracking system.

XIM vs. euclidean rotation

Note that in the following all specific pan, tilt, yaw and pitch angles are listed in terms of the gazers individual rotation scheme. A gazer's complete rotation around its main axis is split into 170 parts. The pan thus ranges from 1 to 170 degrees in XIM rotation. For the tilt, angles from 1 to 318 can be set. To translate from euclidean to XIM rotation and reverse, I thus use the following formulas:

$$pan_{xim} = pan_{deg} \cdot \frac{170}{360} \tag{1.1}$$

$$tilt_{xim} = tilt_{deg} \cdot \frac{318}{360} \tag{1.2}$$

It is possible to set angles in the the ranges stated above with a step size of 0.1 degree.

1.4. Overview over approaches

Based on the considerations from the previous section, this project will be structured as follows:

- Introduction of a novel approach of estimating the extrinsic pose of a gazer from a set of tracking data and gazer angle correspondences. The right angles are determined in a calibration scenario, whereby the spaces is scanned for a range of pan and tilt angles while a classifier is applied to the gazer's image to detect a person standing at a known position in the room. The person's position is tracked online by the overhead infrared tracking system. Once adjusted right, a fitting correspondence of angles and tracking position is found. From a variety of these correspondences the gazer's extrinsic parameters can be estimated by setting up a system of equations and optimizing the desired parameters with a non-linear Levenberg Marquard Optimizer. This approach will be discussed in detail in chapter 2.

- For verification of the results of the classifier approach, a control experiment was run implementing a state of the art marker based pose estimation. Therefore an interactive marker equipped with infrared LEDs at five points was placed in the room. The poses of both the gazer and the infrared camera relative to the marker were estimated. In combination they yield the position of the specific gazer relative to the overhead camera. To compensate for errors in the estimation, a bundle adjustment serves to globally optimize the pose in all constraints. The results of the control experiment and the comparison of the results to those of the classifier approach are discussed in chapter 3.

- Given the results of the pose estimation as an input, the pan and tilt angles for each gazer to look at a specific spot in the room can be computed online. The computation was tested for a number of tracking poses. As a first experiment, hue histograms were generated from the images returned and different algebraic measures of histogram comparison were tested. I further introduce a sample application that creates a saliency map from the incoming tracking data to determine a person standing by itself and does not move. An appropriate gazer is then chosen to look at this person and generate the hue histogram. The results of the experiments on attribute extraction and histogram comparison are discussed in chapter 4.

1.5. Related work

There is a broad band of literature available about Active Vision and Camera Calibration. Also for the particular task of self calibration of active sensors, several approaches can be found. These works however are mostly based on the exploitation of specific landmarks, image sequences or particular camera motions, rather than on a secondary positioning signal as it is the case for this thesis. Nevertheless I would like to present a choice of related work, that deals with similar problems as my thesis.

A nice overview over Active Vision is given by Andrew Blake and Alan Yullie [14]. Their book *Active Vision* explores important themes emerging from the active vision paradigm. The individual contributions look at tracking, the control of vision heads, geometric and task planning and architectures and applications. Covered are traditional works in computer vision, as e.g geometric modeling, but also new areas such as control theory or dynamic modeling.

The early work of Yiannis Aloimonos [3] (who first introduced the term Active Vision) focuses more on the field of active perception, which calls for studying perception coupled with action. It addresses the technical problems related to the design and analysis of intelligent systems possessing perception such as the existing biological organisms and the "seeing" machines of the future.

Song De Ma [50] presents a self calibration technique for active vision systems. The hand eye geometry as well as the intrinsic parameters of a camera are calibrated directly using the images of the environment. The method exploits the flexibility of the active vision system, and bases the camera calibration on a sequence of specially designed motion.

An example of active tracking and pose estimation in an interactive room is given by Darrel et al. [23]. Demonstrated is real-time face tracking and pose estimation in an unconstrained office environment. Previously implemented vision routines are used to determine the spatial location of a user's head and guide an active camera to obtain pitted images of a face. The pose estimation problem is thereby solved in an eigenspace framework, indexed over both pose and world location.

For robust location and pose estimation of an active vision sensor, a further approach is presented by Wang et al. [58]. The calibration technique is valid for a camera system that can acquire omnidirectional panoramic information of the environment and proposes the calculation of panoramic edge histograms. The location and pose of the sensor is then estimated by matching the edge histograms of the images obtained at the present location with those obtained beforehand at reference points.

Two examples of appearance based active object recognition are given by Borotschnig et al. [15] and by Selinger and Nelson [46]. In both works ambiguities in object detection are solved by repositioning the camera to capture additional views. While Borotschnig et al. focus on the appearance-based object representation, Selinger and Nelson discuss the errors and limitations of multi-view performance enhancement.

Zhang et al. [60] present a Joint System for Person Tracking and Face Detection. A noval vision system detects and tracks faces, using the input from multiple calibrated cameras. Calibration is not subject of their work, but many related issues such as classifier based object detection or color comparison are addressed.

A method for determining affine and metric calibration of a camera with unchanging internal parameters is described by Armstrong et al. [5]. It is shown that from planar motion, the affine calibration can be recovered uniquely and the metric calibration up two a

two fold ambiguity. This approach is also taken up by the co-authors of this work, Richard Hartley and Andrew Zisserman, in their book *Multiple View Geometry in Computer Vision* [32], which is an excellent reference for all Computer Vision and Camera Calibration interests.

2. The Classifier Approach

In this chapter I will present a new approach of estimating a gazer's pose as the optimal parameter configuration to a set of correspondences between positions in space and gazer adjustments. Setup and implementation of the individual components will be explained in detail. Furthermore, the performance of this calibration technique will be evaluated in a test scenario.

2.1. Introduction

The four individually steerable color cameras mounted to the rig in the four corners of the XIM, the gazers, shall be used to win complementary information about certain entities in the space. On command they therefore need to be set to look at the desired spot and extract a image, which serves as input for further processing. Each gazer's view direction can be expressed by two degrees of freedom, the pan and the tilt angle. These angles can be computed trigonometrically if, along with the coordinates of the position to be looked at, the extrinsic parameters of the gazer's base are known. A gazer's extrinsic parameters are defined by its position and orientation in 3D space. If we assume the gazer's base to be sufficiently horizontally adjusted, there are five parameters that we need to know for each gazer. The model for the computation of the pan and tilt angles for a gazer with known extrinsic parameters and an arbitrary input position is described in section 2.4.1.

In principle a gazer's extrinsic parameters are not known for an arbitrary setup. Since the correctness of the computation of the pan and tilt angles is strongly dependent of their accuracy, an appropriate finding of these parameters is crucial for all further progress. A gazer's extrinsic parameters are also referred to as its *pose* in 3D space. Since we can neglect the third rotation, this pose consists in our case of five degrees of freedom: the position (x, y, z) and the orientation (yaw, pitch). Figure 2.1 sketches a gazer's pose and the difference between yaw, pitch, pan and tilt. The specific task of determining the pose is called *pose estimation* or also *extrinsic calibration*.

This project introduces an innovative approach of estimating a gazer's pose. Generally the pose estimation problem can be solved in different ways, depending on the information available about the image sensor and a choice of methodology. Two classes of methodologies can be distinguished:

- Analytic or geometric methods. Given that the image sensor (camera) is calibrated, the mapping from 3D points in the scene to 2D points in the image is known. If also the geometry of the object is known, the projected image of the object is a well-known function of the object's pose. Once a set of control points on the object, typically corners or other feature points, has been identified it is possible to solve the pose transformation from a set of equations which relate the 3D coordinates of the points with their 2D image coordinates.

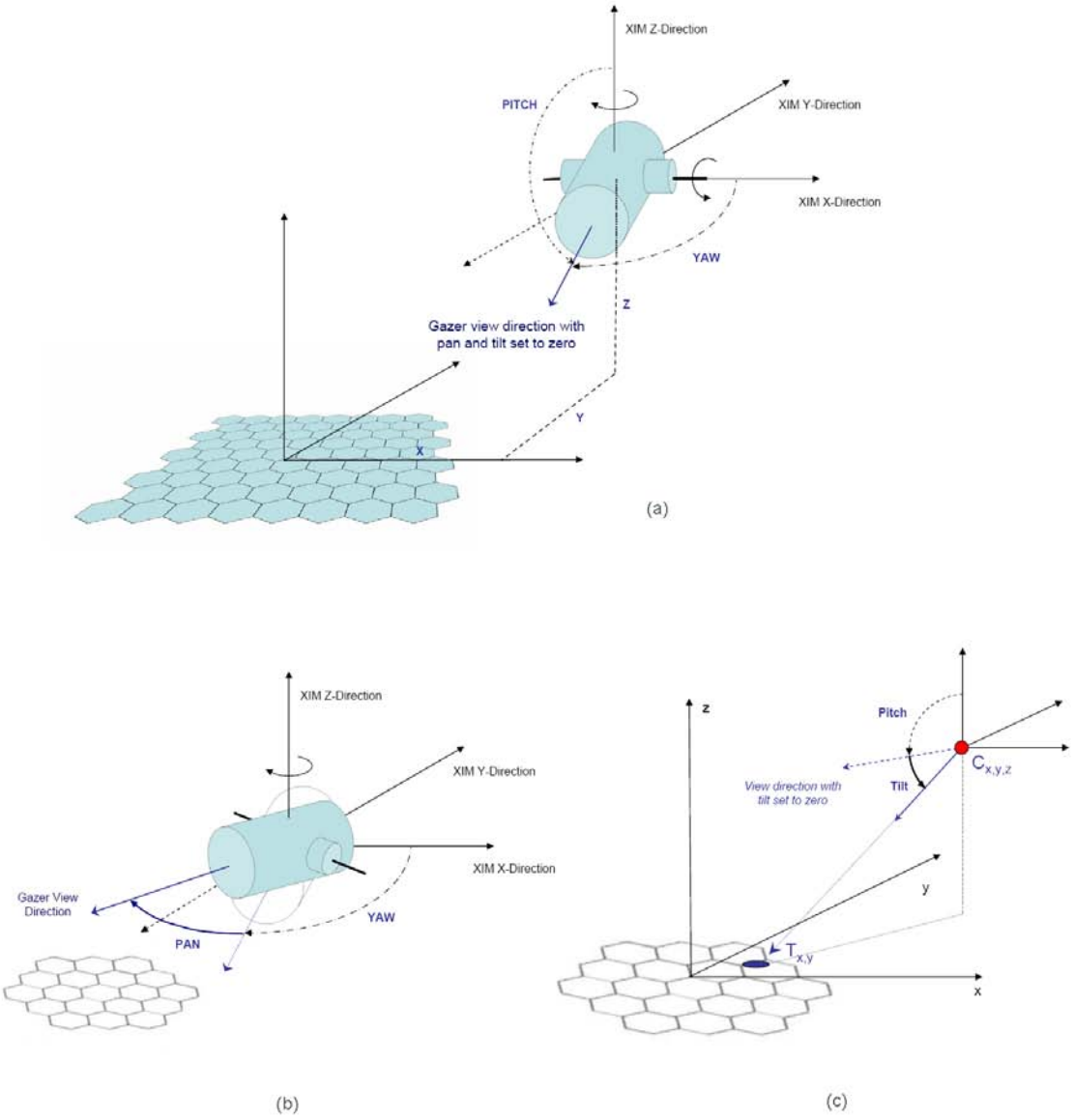


Figure 2.1.: **The pose of a gazer.** Position (x, y, z) and orientation (yaw, pitch) are defined in relation to the reference coordinate system (a). Yaw and pitch thereby refer to the orientation as the pan and tilt angles are set to zero. The pan describes the gazer's internal rotation around its z-axis (b), while the tilt angle describes the internal rotation around the x-axis and thereby the nod of the camera head (c).

- Learning based methods. These methods use an artificial, learning-based system which learns the mapping from 2D image features to pose transformations. Therefore a sufficiently large set of images of the object in different poses hast to be presented to the system during a learning phase. Once the learning phase is completed, the system should be able to present an estimate of the object's pose given an image of the object.

I tackle the pose estimation problem for the gazers by a noval approach that, although geometrically funded, implements a calibration scenario in which continuously valid correspondences between tracking positions and pan and tilt angles are determined. The pose is then computed as the optimal fit to this model. As an input will serve next to the online camera image of the gazer to be calibrated only the live data from the XIM infrared tracking system. This has two major advantages: The usual method of estimating a camera's pose, a marker based analytic approach, returns the pose in relation to the coordinate system, spanning the marker used for calibration. In this case, the transformation between the reference and the marker coordinate systems remains unknown and has to be estimated separately. This leads us to the question, how to define the reference coordinate system. Since all tracking data is always returned relative to the overhead infrared camera, it seems appropriate to take the overhead camera as reference. If we base our pose estimation on the input from even this overhead camera, we profit from the fact of estimating the gazer's pose already in the right coordinate system without having to apply further transformations. As a second advantage, not having to recall on any other modalities but the tracking makes the calibration independent and reproduceable, no matter in what content. We can assume that tracking data will always be available. The only other entity necessary is an object that can be tracked by the overhead camera and also be detected in the gazer image. Since the whole tracking architecture is developed for people tracking, it is reasonable to use a person also for calibration. I propose the use of a *Haar Classifier* to decide if and where a person can be found in the gazer's image.

2.2. Concept

We will introduce a computational model that allows the computation of the pan and tilt angles corresponding to a tracking position for any gazer if its pose is known. The concept of this calibration approach is to estimate the pose by searching an optimal fit of the extrinsic parameters to a set of equations set up by known valid correspondences between tracking positions and pan and tilt angles. Therefore this set of correspondences need to be determined in a prior calibration scenario. Figure 2.2 shows a conceptual overview of the setup and the corresponding spatial relationships. In principle, the calibration process can be structured as follows:

1. A person steps to a position T_i in the space and is thereby tracked by the XIM tracking system.

2. The gazer scans the room rotating around both axes.

    - The gazer's current camera image is grabbed.

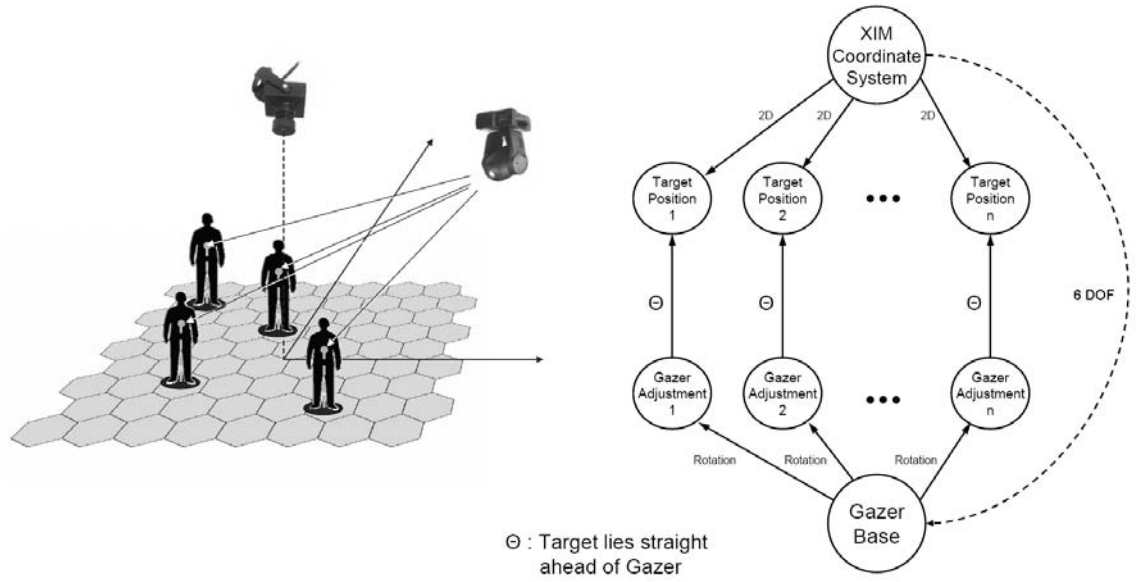


Figure 2.2.: **Conceptual setup for the classifier approach to estimate the extrinsic pose of a gazer.**

- A classifier tries to determine whether the person can be found in the image and if so, where in the image.
- The successive pan and tilt angles are calculated:
  - If a person was detected the gazer is tried to be aligned so that the person is right in the center of the image.
  - Otherwise the gazer continues on its predefined path.
- The gazer's orientation is set to the new angles.
- This step is being repeated until the gazer is adjusted to look straight at the person.

3. The gazer's pan and tilt angles are recorded along with the current tracking position T_i as a correspondence.
4. Steps 1 to 3 are repeated iteratively until an adequate number of correspondences has been found.
5. A system of non-linear equations is set up from the correspondences and the optimal fit of the extrinsic parameters to this system is determined using a Levenberg-Marquard Optimizer.

2.3. Theoretical background

Key hallmark of this approach is the optimization of the pose parameters to serve as an optimal fit to a set of correspondences between target positions and gazer angles, that has

been determined in a prior calibration scenario. In this section, I will provide the theoretical background to the two major aspects of this concept: The gain of the correspondences and the optimization of the parameters. The collection of the correspondences requires the detection of a person in the gazer image, which should be solved by use of a classifier trained on human upper bodies. For parameter optimization, I propose the use of the Levenberg-Marquardt Algorithm, which allows to solve the system of non-linear equations made up by the individual correspondences.

2.3.1. Classifier based object detection

While trying to detect the person in the space through the gazer image, a decision has to be made if the person exists in the current image and if so, where it can be located. Object detection, in particular the detection of human bodies or body parts, is an important element of various computer vision areas, such as image retrieval, shot detection, video surveillance, etc. The goal is to find an object of a pre-defined class in a static image or video frame. Sometimes this task can be accomplished by extracting certain image features, such as edges, color regions, textures or contours and then trying to find configurations or combinations of these features specific to the object in question by some heuristics [16]. For complex objects, such as human body parts, it is hard to find features and heuristics that can handle the huge variety of instances of the object. The shape of a human body may vary strongly and be exposed to different lightning conditions and shadows. For such objects, a statistical model (classifier) may be trained instead and then used to detect the objects [16].

In a statistical, model-based training multiple samples are declared "positive" or "negative", depending on whether an instance of the object in question is contained. Together positive and negative samples make a training set. During training, different features are extracted from the training samples to select distinctive features usable to classify the object. This information is summed up into a row of statistical model parameters. Even when already applied to detect an object, this model can be still be adjusted if the trained classifier does not detect an object or mistakenly detects the object (false alarms) by adding the corresponding positive or negative samples to the training set.

OpenCV's Haar Classifier

The OpenCV computer vision library features such a statistical approach for object detection, an approach originally developed by Viola and Jones [56] and extended by Lienhard [37, 36]. Haar-like features, in their computation similar to the coefficients in Haar wavelet transforms, and a cascade of boosted tree classifiers are used as a statistical model. While in [56] this method is tuned and tested for face detection, a classifier for an arbitrary object class can be trained and used in exactly the same way. Since we want to adjust the gazers to focus on the chest of the person in question, we will use a classifier trained on the upper body of a human. For this specific case, OpenCV provides a precasted cascade that was trained on an MIT pedestrian data set used in experiments on identifying persons in images from surveillance cameras.

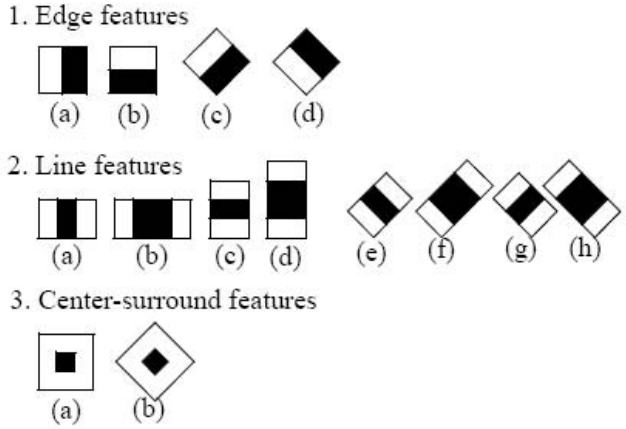


Figure 2.3.: **Extended set of Haar-like features.** *Image adapted from [37]*

The classifier is trained on images of fixed size and detection is done by sliding a search window of that size through the image, checking whether the image region bears the desired features to identify the object. The classifier is scalable to detect objects of different size. Fundamental to the whole approach are Haar-like features and a large set of very simple "weak" classifiers that use a single feature to classify the image region as body or non-body [16]. Each of these features is thereby described by a template specifying the shape of the feature, its coordinate relative to the search window origin and its size (scale factor). Lienhard [37, 36] proposes the use of 14 templates as shown in Figure 2.3, each consisting of two or three joined "black" and "white" rectangles. The Haar feature's value is calculated as a weighted sum of two components: The pixel sum over the black rectangle and the sum over the whole feature area. The weights of these two components are of opposite signs and for normalization, their absolute values are inversely proportional to the areas. As an example, the black feature 3(a) in Figure 2.3 has $weight_{black} = -9 * weight_{whole}$.

Other versions of classifiers use hundreds of features that require the direct computation of pixel sums over multiple small rectangles, making the detection very slow. Viola [56] introduces an elegant method to compute the sums very fast. Therefore first an integral image Summed Area Table (SAT) is computed over the whole image I where

$$SAT(X,Y) = \sum_{x<X, y<Y} I(x,y) \tag{2.1}$$

The pixel sum over a rectangle $r = \{(x,y), x_0 \leq x < x_0 + w, y_0 \leq y < y_0 + h\}$ can then be computed using SAT by using just the corners of the rectangle of size :

$$RecSum(r) = SAT(x_0+w, y_0+h) - SAT(x_0+w, y_0) - SAT(x_0, y_0+h) + SAT(x_0, y_0) \tag{2.2}$$

This equation is valid for up-right rectangles. For rotated rectangles a separate "rotated" integral image must be used.

The computed feature value $x_i = weight_{i,black} * RecSum(r_{i,0}) + weight_{i,whole} * RecSum(r_{i,1})$ is then used as input to a simple decision tree

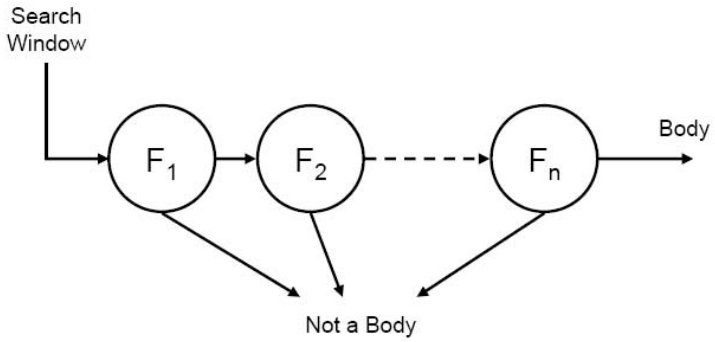


Figure 2.4.: **Object detection cascade of classifiers where rejection can happen at every stage.** *Image adapted from [16]*

$$f_i = \begin{cases} +1, & \text{if } x_i \geq t_i \\ -1, & \text{if } x_i < t_i \end{cases} \tag{2.3}$$

For an arbitrary feature i and a threshold t_i assigned to this feature, such a *weak classifier* thus gives response on whether the feature was detected in the image (1) or not (-1). A weak classifier is not able to detect complex structures such as faces or bodies. It rather reacts to some simple feature in the image that may relate to the object in question. Feature 3(a) in Figure 2.3 for example would, if centered on the eye of a person and scaled, most likely give a large response.

Weak and Boosted Classifiers

For detection of complex structures a *boosted classifier* has to be build iteratively as a weighted sum of weak classifiers as introduced by Freund and Schapire [29]:

$$F = sign\left(c_1 f_1 + c_2 f_2 + \ldots + c_n f_n\right) \tag{2.4}$$

On each iteration, a new weak classifier f_i is trained and added to the sum. If f_i gives a small error on the training set, a large coefficient c_i is assigned to it. The weight of all the training samples is then updated, so that on the next iteration the role of those samples that are misclassified by the already built F are emphasized. It is proven in [7] that if f_i is even slightly more selective than just a random guess, an arbitrarily high hit rate ($<$1) and an arbitrarily small false alarm rate ($>$0) can be achieved for F, if the number of weak classifiers in the sum is large enough. In practice however, that would require a very large training set as well as a very large number of weak classifiers, resulting in a slow processing speed. To avoid this, Viola [56] suggests building several weak classifiers with constantly increasing complexity and chaining them with the simpler classifiers going first. During detection, the current search window is analyzed subsequently by each of them and each of them may reject it or pass it on to the next one. This way, the candidate is only

accepted if all of the classifiers pass it on and each of them may sort it out (Figure 2.4). In experiments, about 70-80% of the canidates where rejected within the first two stages that used the simplest features (about ten weak classifier each), so that this technique speeds up the detection greatly [16]. Also, improvements on performance can be made by choosing the desired hit-rate and false-alarm-rate at every stage.

2.3.2. Levenberg-Marquardt optimization

The estimation of a gazer's pose from a set of correspondences can be reduced to the minimization of some well defined cost function that returns a residual to the model for a specific set of parameters. I propose the use of a Levenberg-Marquard Optimizer to solve this problem. The Levenberg-Marquardt Algorithm (LMA) provides a numerical solution to the problem of minimizing a function, generally non-linear, over a space of parameters of the functions. It works very well in practice and has become the standard of non-linear least square routines that strongly relate to the minimization problem [28].

The Levenberg-Marquardt (LMA) method is a variation on Newton iteration, which is one of the most common iterative parameter minimization methods. The Newton iteration, as a general idea, provides a way of finding the zeros of a function of a single variable. Designed to provide faster convergence and regularization in the case of over-parameterized problems, the LM method may be seen as a hybrid between Newton iteration and a gradient descent method [32]. It is more robust than for example the Gauss-Newton algorithm (GNA), meaning in many cases it finds a solution even if it starts very far off the final minimum. As a trade off, for well-behaved functions and reasonable starting parameters, the LMA tends to be a bit slower.

Main application of the LMA is the least square curve fitting problem: Given a set of empirical data pairs $(\hat{x}_i, \hat{y}_i)$ the parameter vector $\vec{p}$ of the model curve $f\,(\hat{x}|\vec{p})$ shall be optimized so that the sum of the squares of the deviations

$$C\,(\vec{p}) = \sum_{i=1}^{n} [\hat{y}_i \;-\; f\,(\hat{x}_i \mid \vec{p})]^2 \tag{2.5}$$

becomes minimal.

The algorithm starts from an initial guess for the parameter vector $\vec{p}$ that has to be provided by the user. In most cases a uniformed standard guess like $\vec{p}^{\,T} = \{1, 1, \ldots, 1\}$ will work fine. During runtime, the parameter vector $\vec{p}$ is replaced by a new estimate $\vec{p} + \vec{q}$ in each iteration step. To determine $\vec{q}$, the functions $f_i\,(\vec{p} + \vec{q})$ are approximated by their linearizations

$$f\,(\vec{p} + \vec{q}) \approx f\,(\vec{p}) + J\vec{q} \tag{2.6}$$

where J is the Jacobian of f at $\vec{p}$.

At a minimum of the sum of squares C, the gradient of C with respect to $\vec{q}$ is zero. Differentiating the square of the right hand side of the above equation 2.6 and setting it to zero leads to:

$$\left(J^T J\right) \vec{q} = J^T \left[\hat{y} - f\left(\hat{x}|\vec{p}\right)\right] \tag{2.7}$$

from which $\vec{q}$ can be obtained by inverting $J^T J$. The key characteristic of the LMA is the replacement of this equation by a "damped" version

$$\left(J^T J + \lambda I\right) \vec{q} = J^T \left[\hat{y} - f\left(\hat{x}|\vec{p}\right)\right], \tag{2.8}$$

with I being the Identity Matrix. This damped version allows the dynamic adaption of the increment $\vec{q}$ to the estimated parameter vector $\vec{p}$ as

$$\vec{q} = \left(J^T J + \lambda I\right)^{-1} J^T \left[\hat{y} - f\left(\hat{x}|\vec{p}\right)\right]. \tag{2.9}$$

λ is adjusted after each iteration. If C decreases rapidly, a smaller value can be chosen for λ and the iteration is essentially the same as Gauss-Newton iteration. If the iteration gives insufficient reduction in the residual, λ can be increased, whereby the parameter increment approaches that given by gradient descent. Thus, the LM Algorithm moves seamlessly between Gauss-Newton iteration, which will cause rapid convergence in the neighborhood of the solution, and a gradient descent approach, which will guarantee a decrease in the cost function when the going is difficult [32]. The algorithm aborts if either the calculated step length $\vec{q}$ or the reduction in the sum of squares from the previous iteration fall below predefined limits. The last parameter vector $\vec{p}$ is then considered the solution.

2.3.3. The Jacobian matrix

As described above, the LMA replaces the parameter vector $\vec{p}$ in each iterative step by a new estimate $(\vec{p} + \vec{q})$ that is determined by approximating the functions $f_i\left(\vec{p} + \vec{q}\right)$ by their linearizations $f\left(\vec{p} + \vec{q}\right) \approx f\left(\vec{p}\right) + J\vec{q}$. We therefore need to provide the Jacobian matrix J, that contains all first order partial derivatives of the vector valued function f.

Suppose $F : \mathbb{R}^n \to \mathbb{R}^m$ is a function from Euclidean n-space to Euclidean m-space. Such a function is given by m real-valued component functions $f_1\left(x_1, \ldots, x_n\right), \ldots, f_m\left(x_1, \ldots, x_n\right)$. The partial derivatives of all these functions (if they exist) can be organized in a m-by-n matrix, the Jacobian matrix F:

$$J_F\left(x_1, \ldots, x_n\right) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{2.10}$$

2.4. The Computational Model

In the previous section I have provided the theoretical background for the classifier based pose estimation. The computational model derived in this section, gives the mathematical foundation for the specific problem of estimating the poses of the gazers in the XIM.

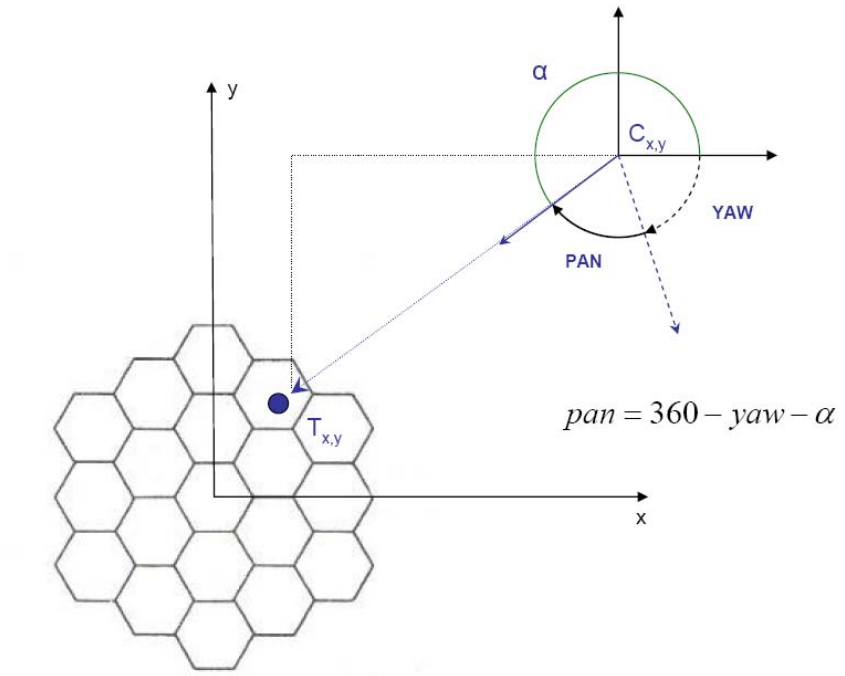


Figure 2.5.: **Computational model for the pan angle.** The angle α can be computed trigonometrically. The pan angle can then be determined by subtracting α and the yaw of the gazer from the full euclidean unit circle.

2.4.1. Computation of the pan and tilt angles

If the position and orientation of the specific gazer's base are known, the necessary rotational angles to turn the gazer to look into a certain direction can easily be computed. We define a computational model, that will later be used as basis for the online computation of these angles for incoming tracking positions. In our calibration scenario we can use the same model for estimating the gazer's pose from a number of known correspondences between tracking positions and the respective gazer angles to look at this position. The computational model for the pan angle is drawn in figure 2.5 while figure 2.6 sketches how to compute the tilt angle.

To determine the pan angle we can first compute an angle α from the gazer's planar position $C_{x,y}$ and the target position $T_{x,y}$ that expresses the target position in polar coordinates. To get α in the interval $[0, 2\pi]$ we have to distinguish five cases depending on how gazer and target relate to each other. Let $x = C_x - T_x$ and $y = C_y - T_y$ be the position of the gazer relative to the target in cartesian coordinates, then

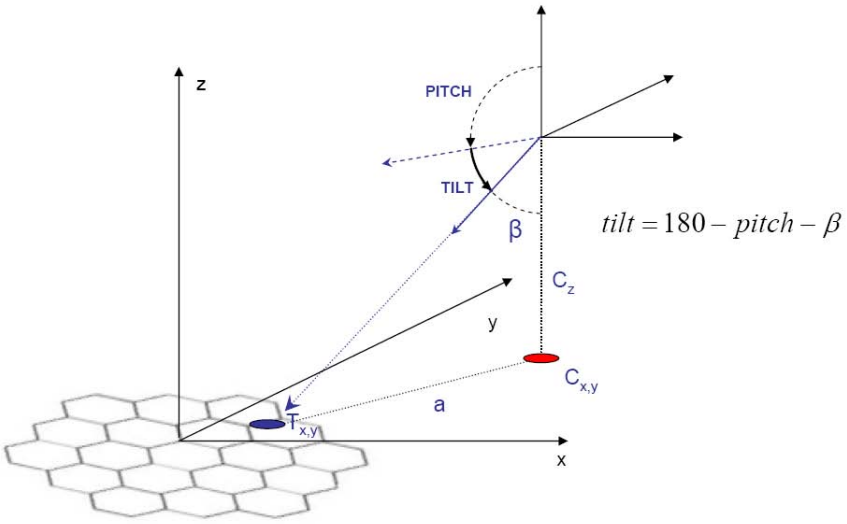


Figure 2.6.: **Computational model for the tilt angle.** The angle β can be computed trigonometrically. The tilt angle can then be determined by subtracting β and the pitch of the gazer from half the unit circle.

$$\alpha = \begin{cases} \arctan\frac{y}{x} & \text{if } x > 0, y \geq 0 \\ \arctan\frac{y}{x} + 2\pi & \text{if } x > 0, y < 0 \\ \arctan\frac{y}{x} + \pi & \text{if } x < 0 \\ \pi/2 & \text{if } x = 0, y > 0 \\ 3\pi/2 & \text{if } x = 0, y < 0 \end{cases} \tag{2.11}$$

In a later computation of α we can recall on the bivariate function of the arc tangent *atan2* featured in C++, which considers all of the above cases internally. It allows us to compute the correct values for α for any given x and y. Knowing α, we can now compute the pan angle as

$$pan = 360 - \alpha * \frac{360}{\pi} - yaw \tag{2.12}$$

or, assuming the availability of the function *atan2*

$$pan = f\left(T_{x,y}\right) = 360 - yaw - \text{atan2}\left(\frac{C_y - T_y}{C_x - T_y}\right) * \frac{360}{\pi} \tag{2.13}$$

The appropriate tilt angle can be computed similarly:

$$a = \overline{T_{x,y}C_{x,y}} = \sqrt{\left(C_x - T_x\right)^2 + \left(C_y - T_y\right)^2} \tag{2.14}$$

$$\beta = \arctan\left(\frac{a}{C_z}\right) * \frac{360}{\pi} \tag{2.15}$$

$$tilt = 180 - \beta - pitch \tag{2.16}$$

$$tilt = g\left(T_{x,y}\right) = 180 - \arctan\left(\frac{\sqrt{\left(C_x - T_x\right)^2 + \left(C_y - T_y\right)^2}}{C_z}\right) * \frac{360}{\pi} - pitch \tag{2.17}$$

2.4.2. Cost functions used for optimization

Equations 2.13 and 2.17 define the basis not only for the later computation of the pan and tilt angles for incoming tracking positions, but also for the estimation of the gazer poses from a set of made measurements. To provide a measure of accuracy for an arbitrary pose to a number of measurement, we need a cost function for the pan and the tilt angle each. For a measurement vector $\vec{m}$ of size n containing corresponding pairs of target positions $\hat{T}_i$ and angles $\hat{pan}_i$ and $\hat{tilt}_i$ these cost functions for a pose vector $\vec{p} = \{C_x, C_y, C_z, yaw, pitch\}$ are defined as

$$C_{pan}\left(\vec{p}\right) = \sum_{i=1}^{n} \left[\hat{pan}_i - f\left(\vec{p}|\hat{T}_i\right)\right]^2 \tag{2.18}$$

$$C_{tilt}\left(\vec{p}\right) = \sum_{i=1}^{n} \left[\hat{tilt}_i - g\left(\vec{p}|\hat{T}_i\right)\right]^2 \tag{2.19}$$

To determine a configuration of pose parameters that gives best results for any target position we therefore first need to collect an adequate number of corresponding pairs between tracking positions ($T_{x,y}$) and internal gazer angles (*yaw, pitch)*. The problem of finding the gazer's pose is then reduced to the problem of finding the optimal fit of these parameters so that the cost functions specified above become minimal. I will tackle this problem using a non-linear Levenberg Marquard Optimizer as described in section 2.3.2.

2.4.3. The Jacobian matrices

During optimization the LMA needs to call up the Jacobian matrices of the functions $f\left(\vec{p}|\hat{T}_i\right)$ and $g\left(\vec{p}|\hat{T}_i\right)$ in each iterative step. These matrices contain the partial derivatives of each f_i and g_i at the individual entries of $\vec{p} = \{C_x, C_y, C_z, yaw, pitch\}$ and are made up according to the matrix form introduced in section 2.3.3 as

$$J_f = \begin{bmatrix} \frac{-1}{(C_y - T_{y,1}) + (C_x - T_{x,1})^2/(C_y - T_{y,1})} & \frac{1}{(C_x - T_{x,1}) + (C_y - T_{y,1})^2/(C_x - T_{x,1})} & 0 & -1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \frac{-1}{(C_y - T_{y,n}) + (C_x - T_{x,n})^2/(C_y - T_{y,n})} & \frac{1}{(C_x - T_{x,n}) + (C_y - T_{y,n})^2/(C_x - T_{x,n})} & 0 & -1 & 0 \end{bmatrix} \tag{2.20}$$

$$J_g = \begin{bmatrix} 0 & 0 & \frac{-\sqrt{(C_x - T_{x,1})^2 + (C_y - T_{y,1})^2}}{(C_x - T_{x,1})^2 + (C_y - T_{y,1})^2 + {C_z}^2} & 0 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \frac{-\sqrt{(C_x - T_{x,n})^2 + (C_y - T_{y,n})^2}}{(C_x - T_{x,n})^2 + (C_y - T_{y,n})^2 + {C_z}^2} & 0 & -1 \end{bmatrix} \tag{2.21}$$

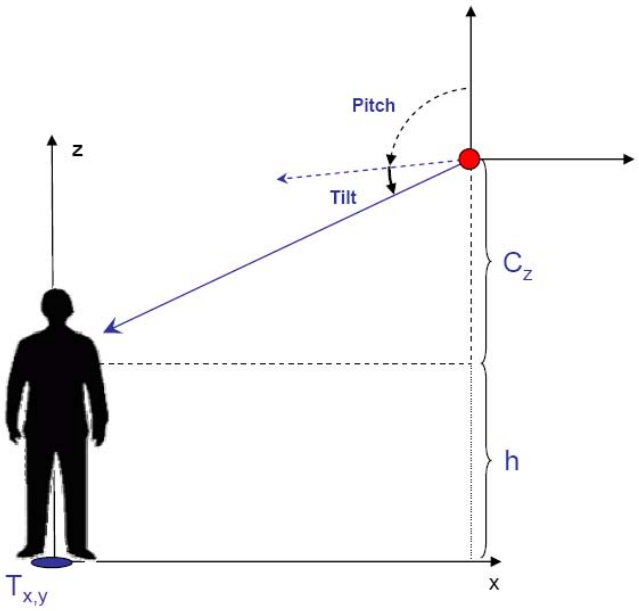


Figure 2.7.: **Significance of the estimated gazer height.** The measured tilt angles, and thus the estimated height of the gazer C_z, relate to the person's upper body height h.

2.4.4. Estimated height vs. real height

I propose the use of a classifier to find the correspondences between tracking positions and gazer angles. In my particular case, the classifier searches for the upper body of a person in the image. Found correspondences thus relate a pair of angles to a 3D position (T_x, T_y, h), where h is the height of the upper body of the person doing the calibration. Consequently, the position of the gazer is estimated relative to his height, wherefore I cannot talk of a "full" calibration. The real height of the gazer in the room would be the sum of the estimated height C_x and the person's upper body height h (Figure 2.7). If I would consider this real height in the computation of the tilt angles, the result would cause the gazer to look at the position $(T_x, T_y, 0)$. Rather than looking on the floor of the XIM, the gazers shall later be adjusted to look at approximately the same height as in calibration. I will therefore maintain the "false" representation of the gazers height nevertheless. Still the reader has to be aware that, when I talk of the gazers pose as estimated by the classifier approach, the z-value does not refer to the gazer's real height.

2.5. Implementation

2.5.1. Environmental constraints

The development of the software that is used for the extrinsic calibration of the gazers faces some environmental constraints. Core of the application is the constant grabbing of images from the camera while the gazer rotates around its z-axis and modifies its tilt to

scan the room for the calibration object. Once the program managed to adjust the gazer so that it looks straight at the object, the object's position has to be cached and recorded along with value of the angles. Therefore interfaces are required to communicate with:

- The `GazerLFServer` that parses the gazer commands and passes them on to the appropriate device in the DMX chain. The communication between the `GazerLFServer` and the gazers is one sided, more specific it can only send commands and does not receive any feedback on their execution. It can not recall any information from the devices. This make it necessary to keep track of the currently set gazer angles on application side. Further we have to make sure that the gazer angles are not modified while the progressing of the image is still in progress, to avoid associating an image - and therefore a possible detection - with the wrong angles. The communication with the `GazerLFServer` is realized via a UDP connection, wherefore we have to integrate a UDP client into our application.

- The framegrabber to constantly update the current camera image from the gazer. All video input from the gazers goes directly to the local machine. We will access it using the BTTV Framegrabber that is based on the standard Video4Linux device.

- The tracking system to be aware of the current tracking data at all times. A UDP server is running on application side, constantly receiving the newest tracking information available.

To decide whether the calibration object, in our case a person, can be found in the image a classifier is used on the camera image. A classifier describes a statistical model that has been priorly trained. Applied to the image, it searches for a cascade - a set of features identifying the seeked object - and gives us an answer on whether an object with the specific features was found in the image. The classifier implements a search window that is shifted over the image, which gives us also feedback on where in the image the object was detected.

Since this classifier works best if the target object sets itself clearly off all background and if there is no light invariation, the calibration scenario should be run under optimal conditions. For best results, no background is projected onto the screen in the XIM. To provide a constant, ambient light we only use the ceiling area light source and turn off all spotlights. To avoid ambiguities and to guarantee that a found gazer configuration can be clearly assigned to a defined tracking position, only one single person may be present in the room during calibration. Further, it has to be assured that the tracking is not disturbed by any other object or modality, for instance light spots on the floor.

2.5.2. System overview

Figure 2.8 gives an overview of the structural design of the system implementing the classifier approach. The core of the application is a class called `Calibrator` that tries to find the correspondences between target positions and the respective gazer adjustments. In an infinite loop the `Calibrator` iteratively modifies the adjustment of the gazer while trying to identify the searched person in the camera image. It thereby switches between three states as shown in figure 2.9. Initially the application resides in a *Seek* state in which the

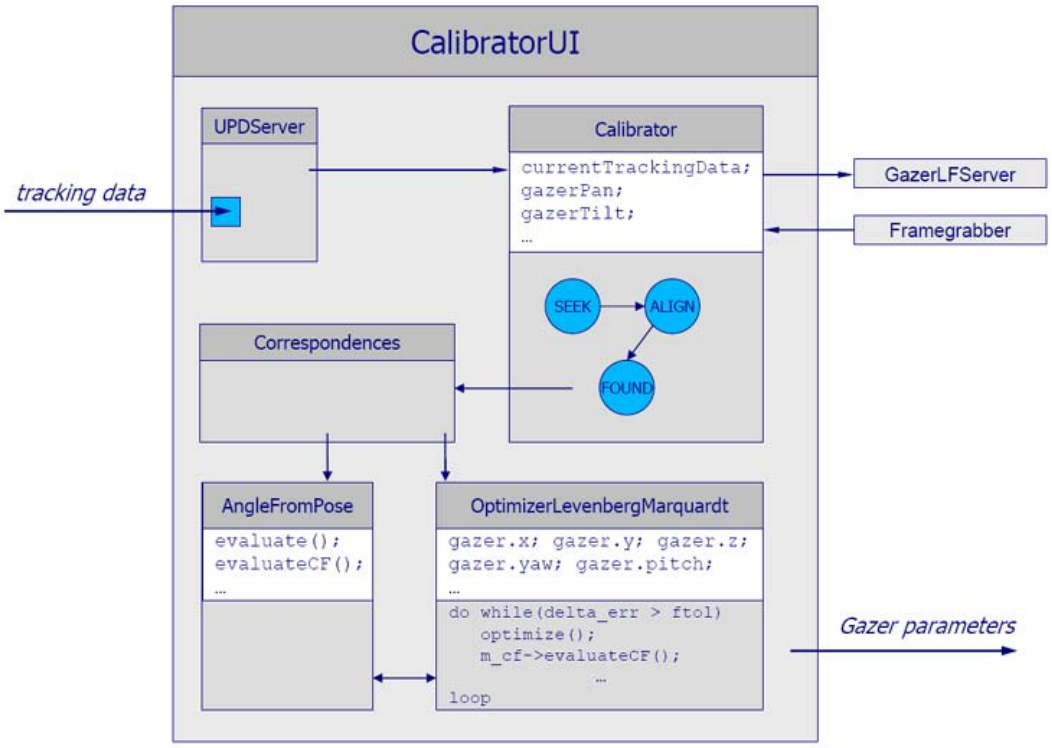


Figure 2.8.: **Overview of the system design for the classifier approach.** The core of the application is a `Calibrator` class that seeks the correspondences. Once enough correspondences have been found, the extrinsic gazer parameters are optimized by a Levenberg-Marquard Optimizer.

gazer is rotated to follow a predefined search path. Pan and tilt angles are modified to scan the room in all dimensions. If a person is detected in the image, the application changes to an *Align* state, in which the successive angles are calculated dynamically trying to adjust the gazer so that the person can be found right in the center of the image. A statement on where in the image the person is located can be made based on the return value from the classifier used for detection. If it was possible to adjust the gazer properly, the application switches to *Found* state and the currently set values for pan and tilt of the gazer are recorded along with the current tracking data. If no proper adjustment is possible within a certain number of steps or if the person can not be identified any more while in *Align* state, the application returns to *Seek* state and continues turning the gazer along the predefined path. After a successful adjustment the system is reset to *Seek* state.

Found correspondences between angles and positions are deposited in a `Correspondence` class. Once an adequate number has been found, a system of non-linear equations can be set up according to the computational model specified in section 2.4.1. The class `AngleFromPose` administrates this system and implements a function `evaluateCF` to evaluate the cost function that returns a measure of accuracy to the found measurements

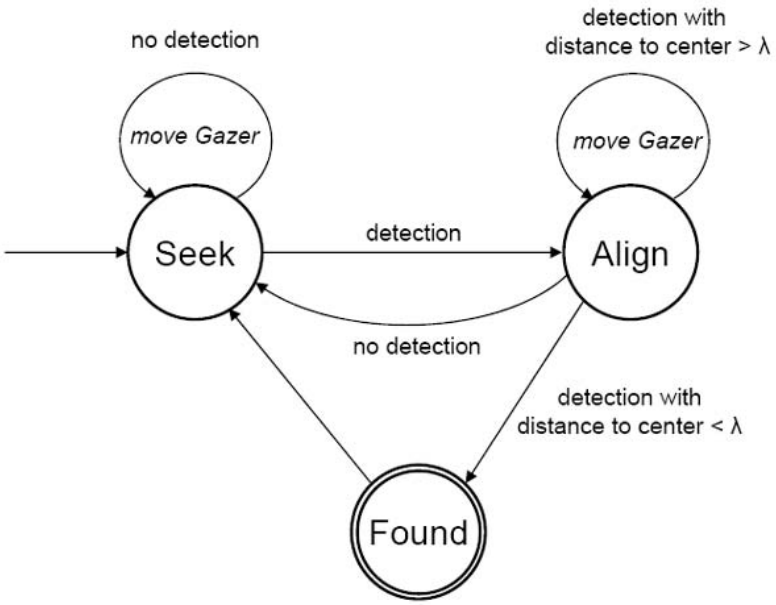


Figure 2.9.: **State chart for the *Calibrator*.** In the *Seek* state the gazer is iteratively rotated around both axes while the classifier scans the camera image to detect the person. If a person was detected the application switches to the *Align* state and tries to adjust the gazer so that the person is located right in the center of the image.

for an arbitrary pose. The cost functions used are stated above in section 2.4.2 as formulas 2.18 and 2.19. To minimize these cost functions and thus find a set of pose parameters that provides an optimal fit to the made measurements, a non-linear Levenberg-Marquard Optimizer is used.

The application returns the estimated pose of an arbitrary gazer. For handling, all components are wrapped up in a Graphical User Interface (GUI). It allows the user to choose which gazer to calibrate and set certain arguments, such as DMX ID, video port or the search range to be covered. The GUI further keeps account of the already collected correspondences and lets the user make modifications on the input passed to the optimizer, in specific the initial parameters or the size of the measurement vector to be used as a subset of the available measurements. Gazers can be added or removed to create a user defined setup that can be saved and loaded at a further point in time. The poses of fully calibrated devices can be exported to a XML file to serve as input for the online computation.

The system has been developed in C++ under Linux. The image processing parts are based on Intel's Open Source Computer Vision Library (OpenCV), a powerful multi-platform and open source image processing library [1]. For linear algebra operations the Template Numerical Toolkit (TNT) created by the U.S. National Institute of Standards and Technol-

ogy was used along with the software library JAMA that is based on TNT [2].

2.5.3. Classifier based object detection in the XIM

In section 2.3.1 I introduced the concept of classifier based object detection and the Haar Classifier included in OpenCV. I use this implementation included in OpenCV for detecting persons in the gazer images. Provided are low-level and high-level APIs for object detection. A low-level API allows the user to check an individual location within the image by a classifier cascade to see whether it contains the object or not. Helper functions calculate integral images and scale the cascade to different sizes by scaling the coordinates of all rectangles of Haar-like features. The higher-level function *cvDetectObjects* wraps up all this functionality. Arguments that have to be passed are a pointer on the image, the cascade, a factor by which the cascade is scaled after each pass, the minimum number of neighboring rectangles that have to be found and the initial search window size. Further a flag can be set to use a Canny edge detector prior to detection to reduce the number of false alarms. We use this function on the current camera image on each iteration of the main loop to decide whether the person carrying out the calibrating can be seen in the gazer image. As described above, the *Upper Body Cascade* available in OpenCV is used and we start with a 60 x 80 pixel search window size that is scale by a factor of 1.2 after each pass.

Figure 2.10 shows examples of the usage of the classifier to detect the person in the calibration scenario. With properly chosen arguments it was possible to reduce the number of false alarms to a minimum. Problems occurred, when dark structures like other gazers, parts of the rig or the door appeared in the background (image (d)). Reducing the number of false detections, by choosing a larger minimum size for the rectangle and a higher number of features that have to be found, results also in a higher rate of false rejects. Thus persons standing at large distance to the gazer or persons not setting themselves clearly off the background (images (b) and (c)) were not detected. But since a false alarm would result in a false correspondence and a messed up pose estimation, the false rejection has to be declared less problematic than the false accept. It is therefore advisable to chose rather defensive parameters. Further, optimal results were achieved when all beamers where turned off and there was no light disturbance from the floor or the LightFingers. Under these constraints, the classifier proved to deliver reasonable and satisfying results in person detection.

2.5.4. Parameter optimization

The `Calibrator` class yields a set of corresponding pairs between measured tracking positions $\hat{T}_i$ and measured gazer angles $\hat{pan}_i$ and $\hat{tilt}_i$. A Levenberg-Marquard Optimizer is used to optimize the parameter vector $\vec{p} = \{C_x, C_y, C_z, yaw, pitch\}$ so that the cost functions 2.18 and 2.19 become minimal for this set of measurements. The cost functions and the Jacobian matrices needed by the optimizer are packed up in the class `AngleFromPose` and provided to the LMA for constant call up.

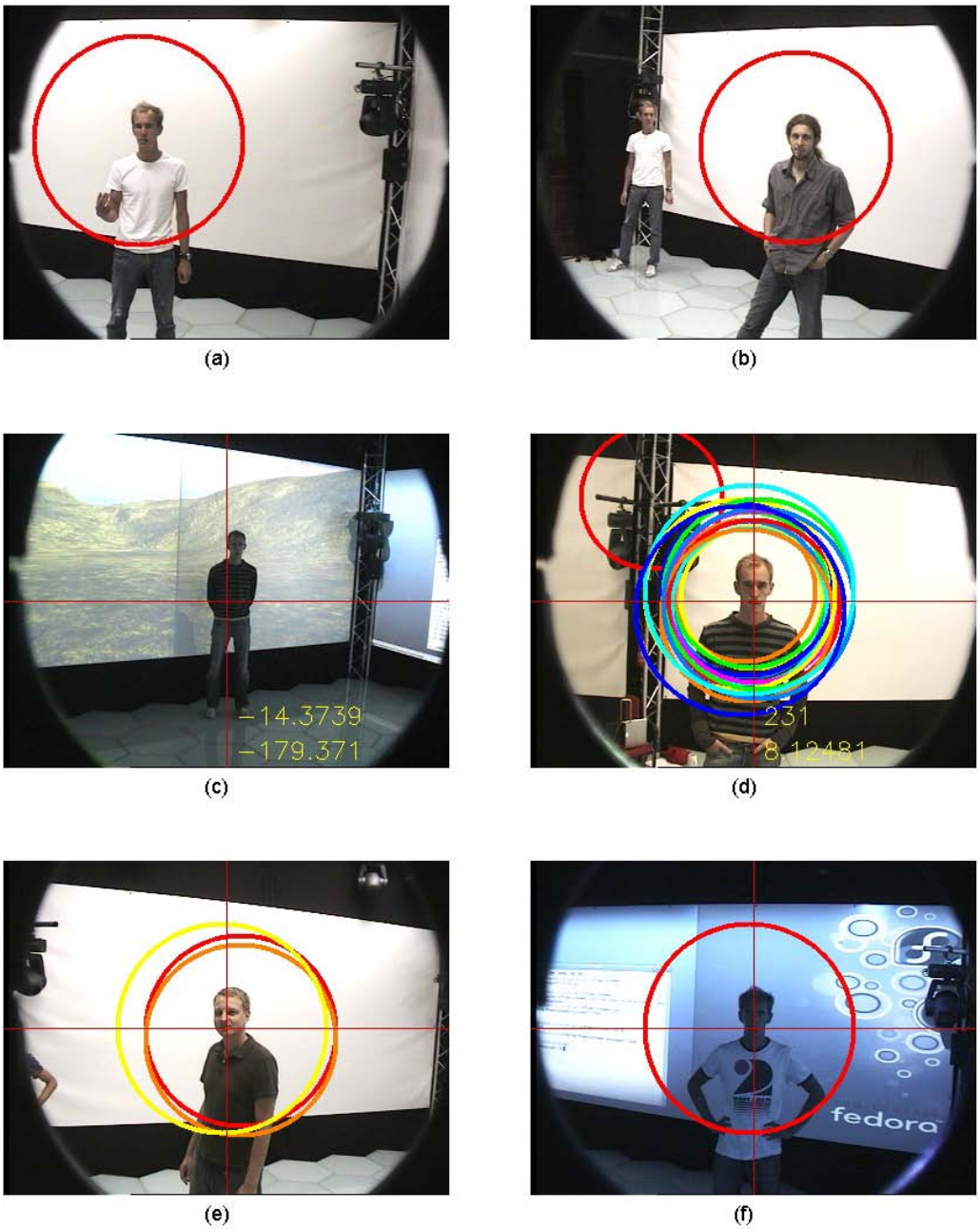


Figure 2.10.: **OpenCV's Haar-Classifier used for object detection in gazer images.** A colored circle indicates a detection and surrounds the persons upper body. The classifier yielded one detection for images (a) and (b), whereby the person in the background of image (b) was not detected. For image (c) the classifier failed to detect the person (false reject), while image (d) shows an example of a false alarm, where a structure in the background was falsely declared to be an upper body. Images (e) and (f) have been taken after the gazer has already been adjusted correctly to feature the person centered in the image.

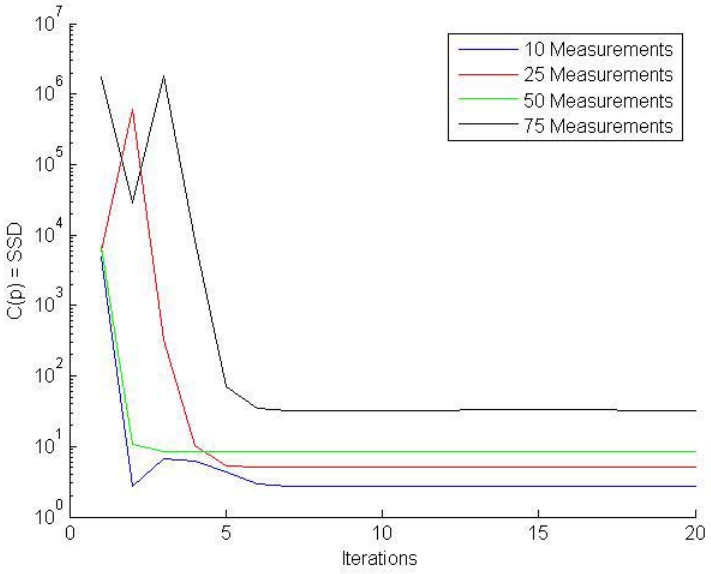


Figure 2.11.: **Convergence of the Levenberg-Marquard Optimizer**. The algorithm was tested for input measurement vectors of different size and proved to converge toward a final parameter vector within few steps. Thereafter the Sum of Squared Differences changed only slightly before the algorithm terminated.

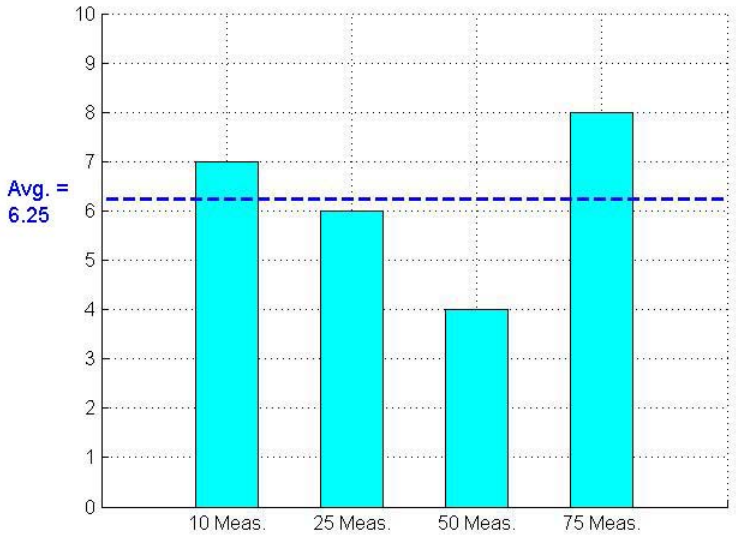


Figure 2.12.: **Steps necessary for the LMA to reach convergence for measurement sets of different size**. After an average of 6.25 steps, the variation in the Sum of Squared Differences was below a limit of 0.1 in each step.

To evaluate the performance of the Levenberg-Marquardt Optimizer and see whether it provides a suitable solution for the pose estimation problem, the algorithm was tested in two respects: How does the size of the measurement vector influence the convergence, and how important is the choice of the initial parameters for the performance. At this point, I evaluate only the optimization of the planar parameters, the gazer's x and y position and its yaw. The cost function recalled by the optimizer therefore is the one in formula 2.18. Optimization of the height and the pitch is done analogously and can be considered easier, since it involves one parameter less.

In order to investigate the first aspect, 75 correspondences where recorded. The algorithm was run on subsets of this measurement vector with sizes of 10, 25, 50 and 75 correspondences, whereby the subset was chosen randomly from the measurements and the initial estimate was set to the same values for all trials. Independent of the input vector's size, the algorithm converged toward a final parameter vector within few steps before only changing the parameters slightly and finally terminating (Figure 2.11). After an average number of 6.25 steps, the change in the Sum of Squared Differences in each step was below a tolerance level of 0.1 (Figure 2.12).

As a second experiment, the optimization was tested for a fixed measurement vector of 50 correspondences. Different starting parameters where provided, some close to the true values and some far off. The resulting parameter vector thereby turned out to be the same, independent of the initial values. Even though the number of iterations before termination varied, the LMA stabilized quickly each time. Figure 2.13 shows how the planar coordinates of the gazer changed within the optimization process for different starting parameters and finally converged toward a well-defined point. The number of iterations needed to do so is shown in Figure 2.14.

Obviously these test don't allow us to judge on the correctness of the resulting parameter vector, which is strongly dependent not only of the size of the measurement vector but also of the accuracy of the individual correspondences. Regarding the performance of the Levenberg-Marquard Optimizer, the algorithm has proven to be a stable and reasonable tool that provides a well-defined solution, even if we don't provide a good initial estimate.

2.6. Performance

In the previous sections we introduced a new approach of estimating the extrinsic pose of movable pan-tilt cameras. In order to judge on the performance of this calibration technique, we have to keep our original aim in mind: For any well-defined position in the space, a gazer's orientation shall be computed as precise as possible. Recalling on our computational model, we know that the gazer's pose is crucial for this computation. This brings up two questions:

- First, is this approach a reasonable way of estimating a gazer's pose regarding usability.
- Second, are the results of the pose estimation accurate enough to guarantee satisfying

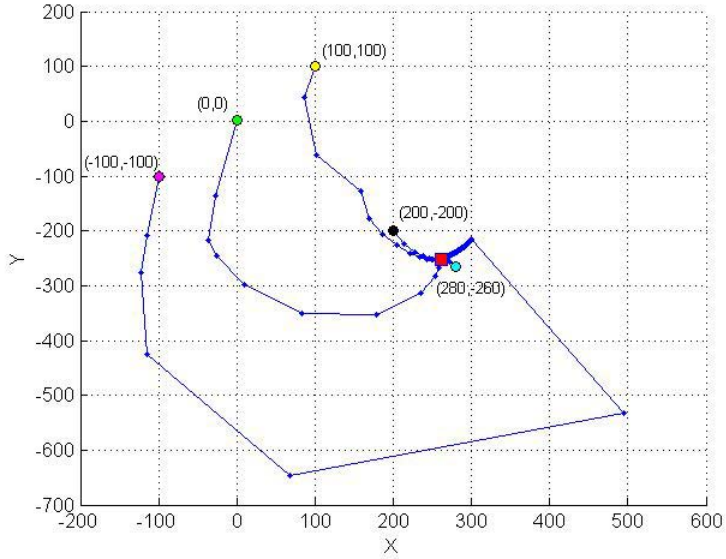


Figure 2.13.: **Optimization path of the gazer position**. Different starting parameters where provided to the LMA as an input. Independent of their choice, the algorithm terminated with a well-defined output (marked as the red square).

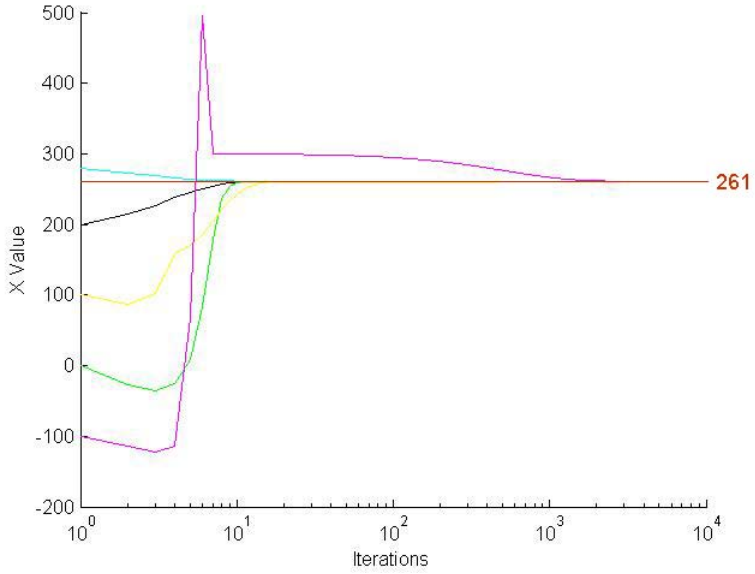


Figure 2.14.: **Convergence of the LMA for different starting parameters**. Regardless of the initial values, the different parameters (shown here is the x value) quickly approached their final value.

results in the computation of the gazer orientation so that we can assume the point of interest to be centered in the image.

Concerning the first question, the individual components of the system, i.e. the classifier to find the correspondences and the optimizer to estimate the pose from even these, have to be evaluated. To see whether the calibration yields satisfactory results, a test scenario was run for each gazer, that will be described later in this section.

2.6.1. Finding the correspondences

In general the Haar-Classifier proved to be a suitable approach for detecting the person in the gazer image. Still problems arise in some situations.

- If the person stands far away from the respective gazer, in specific very close to the projection screens opposing the gazer's position, he/she is often not recognized in the image. This problem can be confronted by making the minimum search window size smaller, so that also smaller objects in the background are recognized. This would on the other hand result in a higher number of false alarms, since smaller areas like shades, irregularities or small objects like the other gazers would falsely be declared to be an object. It is thus not advisable to do so, since it is crucial for the calibration not to have any false alarms.

- Some entities in the room cause confusion and are sometimes falsely declared to be a person. Especially the other gazers are often detected due to their shape. So far this problem could be avoided by declaring the search range so that the gazers are not seen. Also, a single false detection is not critical, as long as the gazer is not aligned to look at the falsely detected object and a correspondence is recorded.

- One wall of the XIM is made up by a semi-transparent mirror to allow the observation of the visitors in the space from outside. During calibration, this mirror causes difficulties. A person in front of the mirror is often not detected, since contours and mirror image blur. Also false detections are possible, if a reflection is declared to be a person. For optimal results it is thus necessary to cover the mirror during calibration with a white sheet.

- Large black areas, such as the transitions between the projection screens and the door, sometimes yield false alarms. This happens especially if the minimum search window size and the minimum number of neighboring features that have to be detected are kept low. Raising them avoids the problem but may cause false rejects of valid objects (persons).

For application in the XIM, it has been possible to choose the settings (search range, minimum search window size, ...) so that no false correspondences were recorded. It was thereby advantageous to do the calibration under optimal conditions, that is with all beamers turned off, no lighting disturbance from the floor or the LightFingers and a steady ambient light. In a fixed, unaltered context the classifier therefore is a reasonable tool to detect a person as a calibration object in the gazer image. In mobile applications though, there could be problems due to inoptimal and unexpected conditions that would have to

be solved individually by adaption of the respective settings or the environmental conditions.

2.6.2. Correctness of the correspondences

The classifier returns a list of corresponding pairs between tracking positions and gazer angles. When estimating the gazer's pose from these correspondences, we hypothesize these correspondences to be correct. The quality of the pose estimation thus is strongly dependent of the correctness of the found correspondences. On the one hand, each tracking position must be uniquely assigned to the right angles. On the other hand, the values contained in each correspondence have to be an accurate measure, for the angles as well as for the tracking positions. We can expect the first premise to be fulfilled, if the classifier works correctly as described above and only one person is present in the room during calibration. Furthermore, a correspondence is only recorded by the classifier, if the gazer was aligned so that the center of the detected person lies within a square of 10 x 10 pixel in the very center of the image. If aligned, the person can therefore always be considered to be in the center of the image with an error $\epsilon_{classifier} < \sqrt{5^2 + 5^2}\, pixel$.

Of more consequence is the inaccuracy of the tracking data. We estimate the pose of a gazer in a cartesian coordinate system and assume the tracking positions to be given in exactly this coordinate system. Subject to perspective and radial distortion in the overhead camera image this assumption is clearly not true. Using the distortion coefficient determined by the intrinsic calibration of the overhead camera's lens the image can be undistorted in sense of radial distortion up to a certain error. The perspective distortion on the other hand remains and turns out to be one of the major flaws in this calibration approach. When determining a person's position from the overhead image, the image processing of the tracking software AnTS segments the visible human body from the image and computes the center of the segmented silhouette. If this person stands close to the center of projection, the camera thereby gets a view from the top. Persons toward the edge of the image and their silhouette are perspectively distorted. The center of the silhouette and thus the returned tracking position is shifted away from the "true" position the further a person moves away from the center. The perspective distortion and the perspective error are illustrated in Figure 2.15.

Considering the perspective distortion, the question comes up in how far the correctness of our computation is affected. On the one hand, the error that arises seems severe. On the other hand, we have to keep in mind that we intend to compute angles corresponding to tracking positions that are afflicted with the same error. The optimizer tries to match the pose parameters, so that the angles are computed as good as possible for any position in the space. Even though the resulting pose may be wrong in terms of geometric interpretation, it may be the set of parameters to get the best result for any given input. At this point we can only draw a conclusion on the accuracy that can be achieved with the pose estimated by the classifier approach. To see whether better results can be achieved with a conventional marker based pose estimation, a control experiment was run. Its setup and the comparison of the results from both calibration techniques are described in chapter 3.

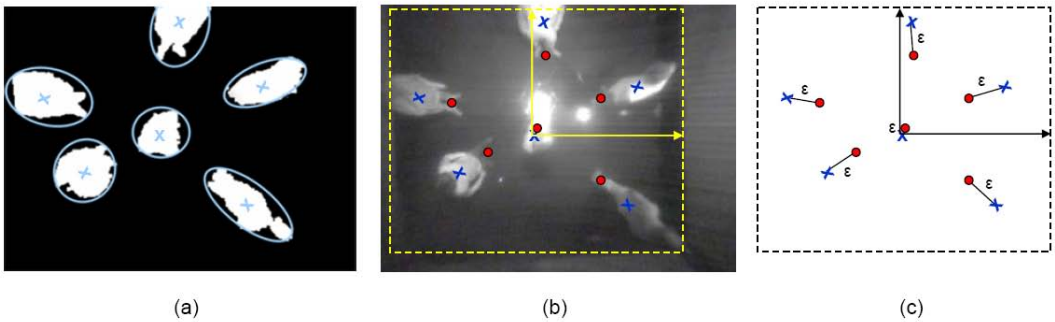


Figure 2.15.: **Perspective distortion in the overhead tracking image**. (a) shows the silhouettes as segmented by the overhead tracking software. As seen in image (b), their centers vary from the positions defined true (the persons positions on the ground floor, marked red in the image). This perspective error ϵ increases as the person moves further away from the center of projection (c).

2.6.3. A test scenario

The estimated poses serve as basis for the computation of the pan and tilt angles that have to be set for each specific gazer to look at a certain position. To make a judgment on the quality of the estimated poses, we thus have to give a measure of accuracy for the angles computed based on the estimations. In a simple test scenario, the computation was therefore tested based on poses estimated by the classifier approach. To evaluate the accuracy of the results, the gazers were set to look at persons standing at different positions in the space. For a perfect computation, the person should then be featured in the very center of the gazer image.

The gazers were calibrated from a set of 50 measurements for each gazer. Table 2.1 shows the results of the estimations. In the course of the experiment, the computation was then tested for 50 different positions for each gazer. A person stepped to an arbitrary spot in the room and his/her position was determined by the tracking system. The pan and tilt angles corresponding to this tracking position were computed and passed to the gazer to

Gazer ID	25	300	352	365
X	261.1	-300.2	-335.1	284.3
Y	-251.9	276.5	-277.9	284.6
Z	11.4	10.3	22.5	20.2
Yaw_{xim}	50.8	39.8	44.3	39.9
Yaw_{deg}	107.6	84.2	93.8	84.5
$Pitch_{xim}$	29.9	29.0	25.7	29.8
$Pitch_{deg}$	33.8	32.8	29.1	33.7

Table 2.1.: **Estimated Poses for the four gazers.**

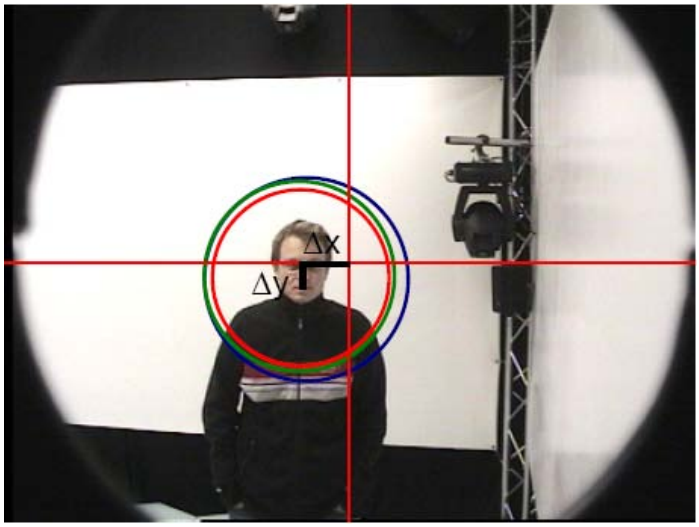


Figure 2.16.: **Deviation in the image after automatic gazer adjustment**. The gazer's orientation was computed from the position of the person seen in the image as signalized by the tracking system. Δx and Δy give a measure of accuracy of the computation. In the perfect case, the person would have to be featured in the very center of the image.

set its orientation. To determine where in the image the person could be found and thus calculate the deviation from the center of the image, the Haar-Classifier that was also used for calibration was applied to the image. Figure 2.16 shows an example of the deviation in x and y direction between the image center and the center of the detected person.

Evaluation of the results

An evaluation of the results of this experiment shows, that the average absolute deviation for the different gazers ranges from approximately 12 to 18 pixels in x-direction, and 7 to 16 pixels in y-direction. The exact averages for each individual device are listed in table 2.2 and plotted in figure 2.17. The fact that all averages lie within the same range shows, that the classifier approach performed consistently for all gazers and that there has been no significant failure. At first sight, a deviation of 12 to 18 pixels seems passable as a foundation for further image processing. Looking at the evaluations of the experiment for the individual gazers on the other hand (figures 2.18, 2.19, 2.20 and 2.21), we see that for some tracking positions the deviation has been enormous. While in some cases a perfect adjustment was achieved, the maximum deviations go up to 64 pixel in x-direction and 68 pixel in y-direction. Even though the maximum deviation was comparably "low" for gazer 365, such *outliers* occurred for all gazers. Interestingly, the deviation in y-direction was always symmetric (except for gazer 300), as the gazer always aimed either to high or to low. The deviations in x-direction on the other hand are widespread in both directions. This shows, that the optimizer truly seeked the pose that fits best for all positions in the space. Still, significant deviations remain for the individual adjustments.

Gazer ID	25	300	352	365
$\oslash\|\Delta x\|$	**15.8**	**18.0**	**15.8**	**11.9**
$min\|\Delta x\|$	1.0	1.0	0.0	0.0
$max\|\Delta x\|$	57.0	61.0	64.0	47.0
$\oslash\|\Delta y\|$	**12.8**	**7.5**	**15.7**	**16.2**
$min\|\Delta y\|$	0.0	0.0	1.0	0.0
$max\|\Delta y\|$	22.0	56.0	68.0	38.0

Table 2.2.: **Average absolute deviations in the images after automatic gazer adjustment.**

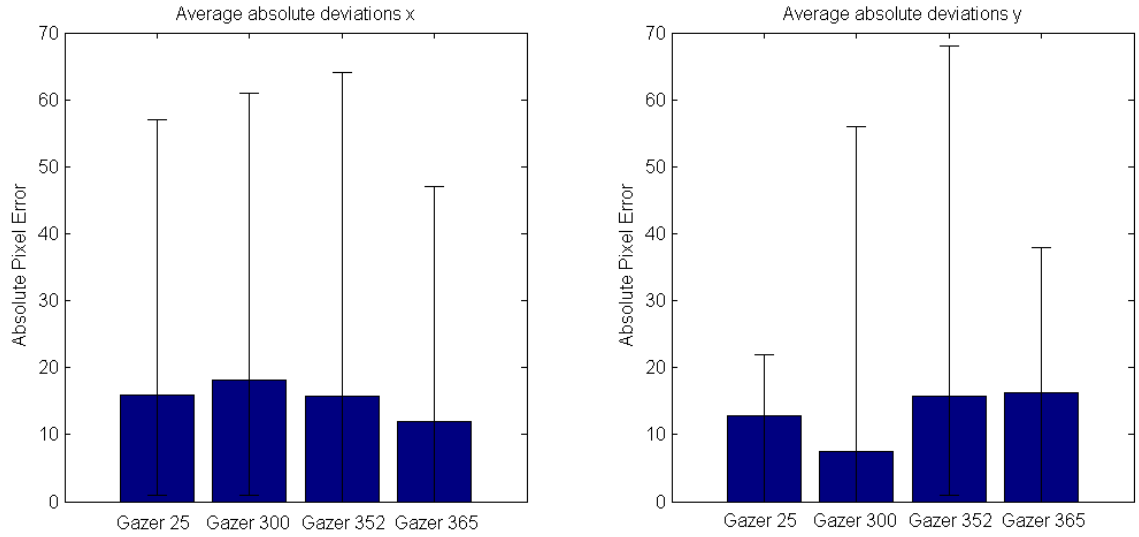


Figure 2.17.: **Average absolute deviations in the images after automatic gazer adjustment.**

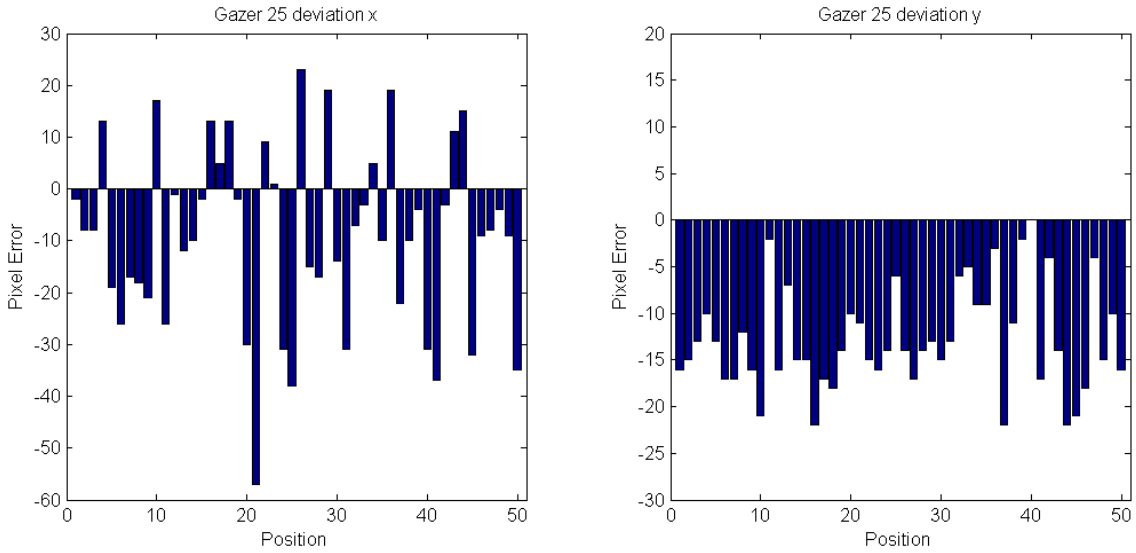


Figure 2.18.: **Accuracy of gazer 25 in the test scenario.**

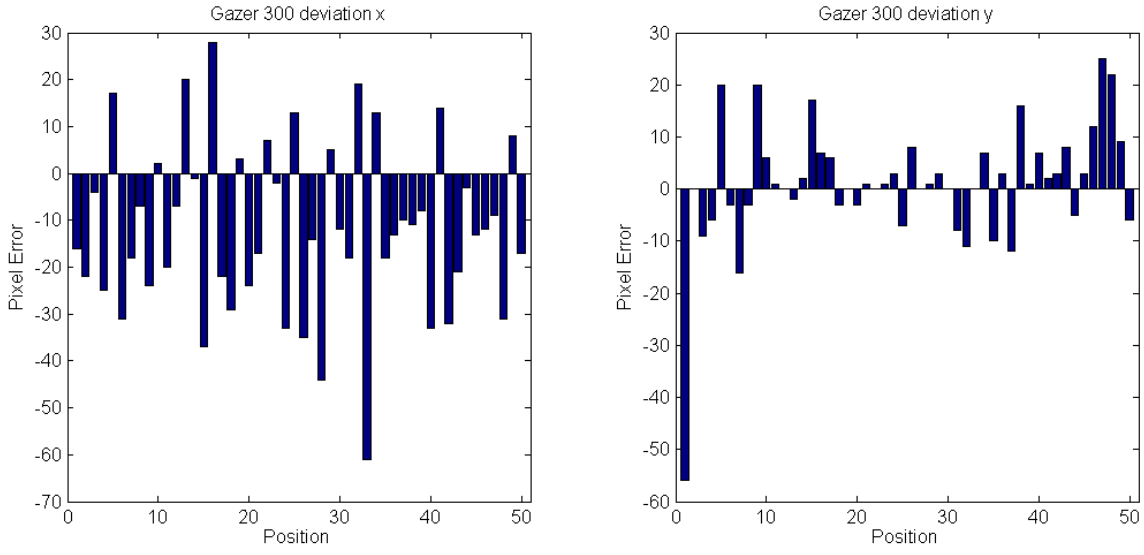


Figure 2.19.: **Accuracy of gazer 300 in the test scenario.**

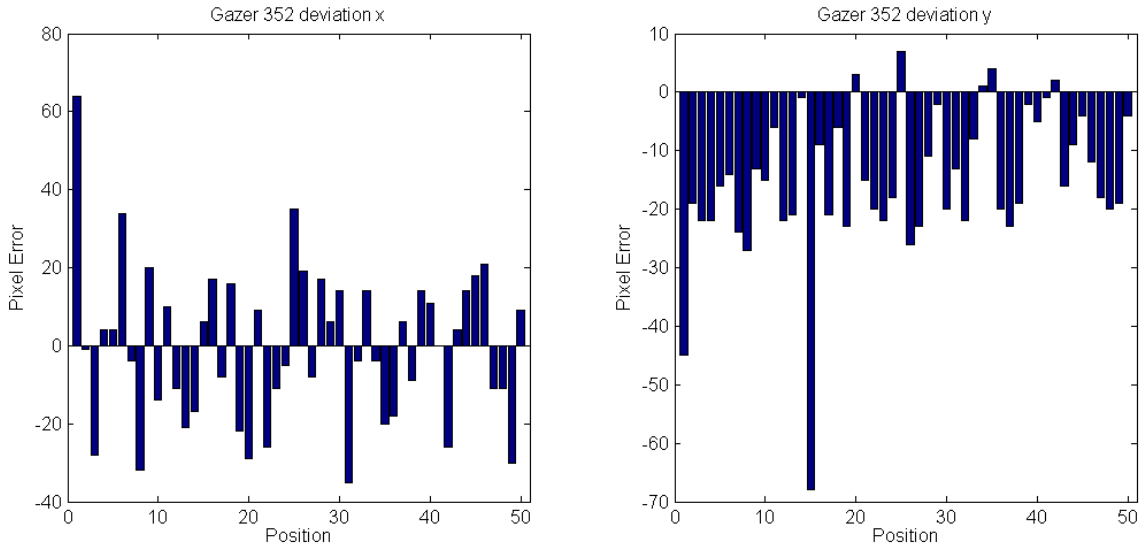


Figure 2.20.: **Accuracy of gazer 352 in the test scenario.**

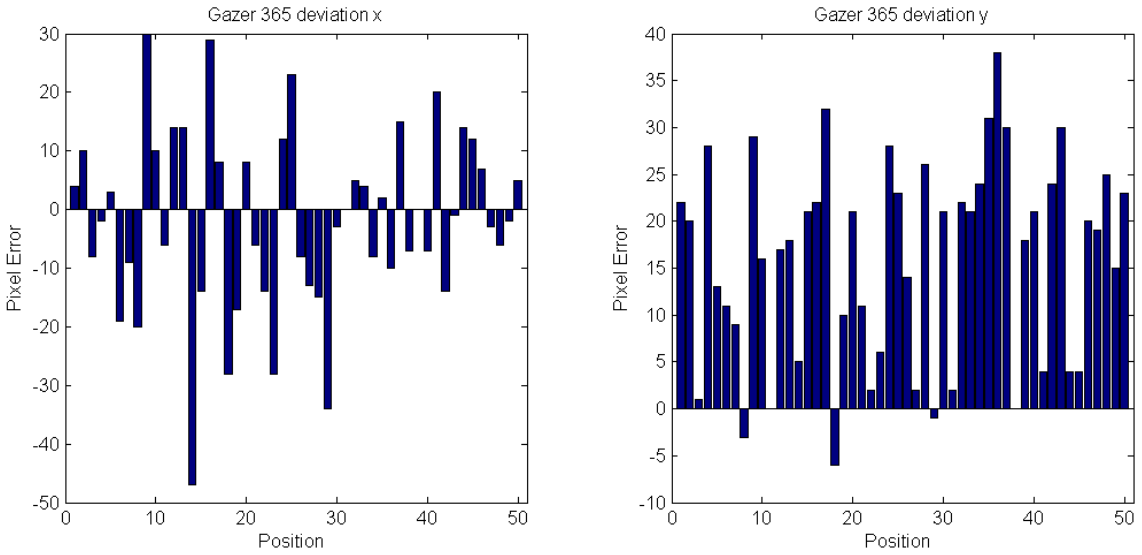


Figure 2.21.: **Accuracy of gazer 365 in the test scenario.**

2.7. Discussion

In this chapter I have introduced a new technique of estimating a gazer's pose. The calibration is based on the collection of valid correspondences between tracking positions and the respective gazer angles. The pose is then determined as the optimal set of parameters to all of these correspondences. Even though the results of this calibration do not represent the true pose of the gazers in a geometric sense, this approach proved to be a suitable way of determining the set of parameters needed to compute the angles corresponding to any position in the XIM. Main advantage of the approach is the independence of any other modality. The gazers learn their position automatically only from their own image and a global tracking signal. In general, the technique could thus be applied to any movable camera in any given context without having to rely on any other entity.

Limitations occur through the classifier that is used to detect a person in the gazer image during calibration. Under certain environmental constraints, the person can non be detected or is falsely classified. Especially these false alarms can significantly affect the performance of the calibration, since also false correspondences are taking into consideration by the optimizer. To avoid these complications, optimal conditions have to be provided throughout the calibration scenario. Possible outliers have to be detected and ruled out by adapting the classifier. Thinkable is also a more robust implementation of the Levenberg-Marquard Optimizer that recognizes false correspondences and disregards them during optimization.

In the particular case of the XIM, the tracking signal is subject to severe perspective distortion. Under the constraint of malicious tracking data, the classifier approach delivered reasonable results. By optimizing the parameters to fit for all correspondences, errors in the tracking data were compensated. Naturally, the more correspondences are provided as an input, the more representable is the resulting pose. In my experiments, 50 pairs proved to be a reasonable basis. The tracking positions should thereby cover all areas of the space equally.

A test scenario showed that decent results in accuracy can be achieved based on the poses estimated in the classifier. After adjusting the gazers to look at specific positions in the space, an average deviation of about 14 pixels could be determined in the gazer images. This seems to be a reasonable basis for further processing. Still there were significant outliers, where the gazer was adjusted way off the target. The question arises, if better results can be achieved if the poses are estimated by a conventional method. I will tackle this question in the following chapter by implementing a state of the art marker based pose estimation.

3. Control experiment

3.1. Introduction

In the previous chapter I introduced a new approach to estimate the position and orientation of a gazer from a set of correspondences between pan and tilt angles and tracking positions. Knowing these parameters, referred to as the *extrinsics* or also the *pose* of a camera, one can compute the orientation that has to be set for a gazer to look at an arbitrary spot in the space. While it has been shown that the approach introduced constitutes a good way of estimating a gazer's pose in sense of usability, the question remains if better results can be achieved by a conventional technique of pose estimation. For verification of the results from the classifier calibration and for comparison to a state of the art calibration technique a control experiment was run, implementing a marker based pose estimation. The implementation and findings of this experiment will be elaborated in this chapter.

3.2. Setup

Aim of the control experiment is the verification of the parameters obtained by the classifier approach. To make a comparison of the results possible, the gazers' poses thus have to be estimated in the same coordinate frame as in the classifier approach. This coordinate frame is spanned by the overhead infrared camera. The pose estimation technique that will be presented in the following is based on the finding of a mapping between a set of 3D world points and the 2D image coordinates of their projection onto the screen. Therefore a marker pattern with known geometry is placed in the space. A number of points that can be unambiguously related to fiducial points of the marker in 3D space must then be identified in the camera image. This makes it possible to compute the mapping from world (marker) to image coordinates from a set of equations that relate the 3D world coordinates of the found fiducial points with their 2D image coordinates. If the internal camera matrix is known, the camera rotation and translation can be decomposed from this mapping.

As described above, this pose estimation technique is based on the mapping from marker to image coordinates and thus estimates the camera's position and orientation relative to the marker coordinate frame. To get the pose of a gazer in the coordinate frame of the overhead tracking camera we therefore have to estimate two coordinate transformations: The one between the marker and the gazer and the one between the marker and the overhead tracking camera. Combined these two transformations give the relationship between the overhead camera and the gazer coordinate system and thereby the desired pose. The setup of the control experiment and the spatial relationships between the entities involved are sketched in figure 3.1. In the Spatial Relationship Graph on the right hand side of the

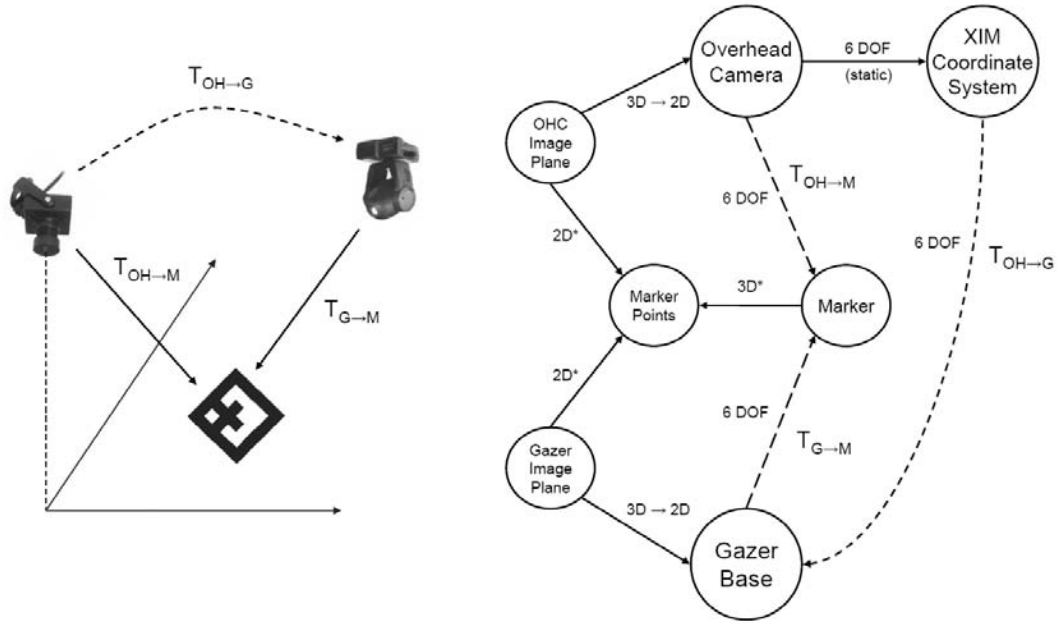


Figure 3.1.: **Setup for the control experiment.** The transformation $T_{OH\rightarrow G}$ between the overhead camera and the gazer is estimated by concatenating the transformations between overhead camera and marker and between gazer and marker.

figure the partial coordinate transformations are denoted by $T_{OH\rightarrow M}$ and $T_{G\rightarrow M}$. Together they yield the transformation

$$T_{OH\rightarrow G} = {T_{G\rightarrow M}}^{-1} T_{OH\rightarrow G} \tag{3.1}$$

that encapsulates the rotation and translation of the gazer's coordinate system relative to the overhead camera's coordinate system and thus the desired pose.

In the following I will briefly describe how to derive a camera's pose from a homography that relates 2D marker to image coordinates and explain how this technique was applied to solve the pose estimation problem for the gazers in the XIM.

3.3. Pose Estimation

The recovery of 3D-geometric information from 2D images is a fundamental problem in computer vision. To automatically compute a rigid-body transformation, the pose, from a single view, it is necessary to match 3D model features with visible 2D image features. The choice of an appropriate model, the identification of the object in the image and the numerical computation of the pose from the correspondences between model and image points have been subject to intensive study. Various approaches exist to solve the correspondence problem by identifying feature points in different configurations: Collinear [31], coplanar

[55] or scattered in 3D [54]. Proposed have also been some pose and displacement algorithms from lines [19, 38, 41, 8], spheres [47] or cylinders [48]. To estimate the pose from the determined correspondences, generally two methods of computation can be distinguished: Algebraic algorithms that provide a numerical solution to the system of linear equations set up from the correspondences and iterative algorithms that start from an initial guess and dynamically optimize the pose by minimizing an appropriate cost function. The numerical solution is easy to implement but often subject to numerical instabilities and noise. Furthermore, numerical solutions like the Direct Linear Transformation (DLT) only minimize the algebraic error but disregard the geometric error. Iterative methods on the other hands minimize the geometric error and are numerically more stable, but their result depends on the initial guess and the algorithm may not converge correctly. It is thus advisable to use a hybrid algorithm that numerically computes an initial guess and then refines it through non-linear optimization.

3.3.1. Pose estimation from the image of a planar marker

The minimum number of point-to-point correspondences needed to estimate the pose depends on the prior knowledge we have about the arrangement of the feature points in the model. In the most general case, if the points are scattered in 3D, the projection matrix P we have to determine is a 3 x 4 matrix with 12 entries, and (ignoring scale) 11 degrees of freedom. Since each point correspondence leads to two linear independent equations, a minimum of 6 correspondences is necessary [32]. If all points lie in a plane, as it is the case for our marker, the projection is equivalent to a homography between two planes that can be expressed by a 3 x 3 matrix. This matrix can be numerically derived from 8 independent equations and therefore a minimum of 4 point correspondences.

Generally, the identification of the internal camera configuration, the intrinsic camera parameters, is needed to achieve the pose. This process is named *camera calibration* and can be done in a preliminary step or it can be achieved in some situations simultaneously with the pose [52].

Homography between two planes

Given a set of points x_i on the marker plane and a set of corresponding points x'_i in the image plane we need to find a projective transformation H that maps each point x_i to x'_i. This interdependency is drawn in figure 3.2 The problem is therefore to compute a 3 x 3 matrix H such that $x'_i = Hx_i$ for each i. A common way to solve this problem is by Direct Linear Transformation (DLT). The algorithm is described in detail in [32] and will only be sketched roughly here.

The equation $x'_i = Hx_i$ may be expressed in terms of the vector cross product as $x'_i \times Hx_i = 0$. This form enables a simple linear solution, where each point correspondence gives rise to two independent equations in the entries of H. Given a set of four such point correspondences, we obtain a set of equations $Ah = 0$, where A is the matrix of equation coefficients from each correspondence and h is the vector of unknown entries of H. We seek a non-zero solution, since the obvious solution $h = 0$ is of no interest. If we use the two independent equations delivered by each of the four correspondences, A has dimension

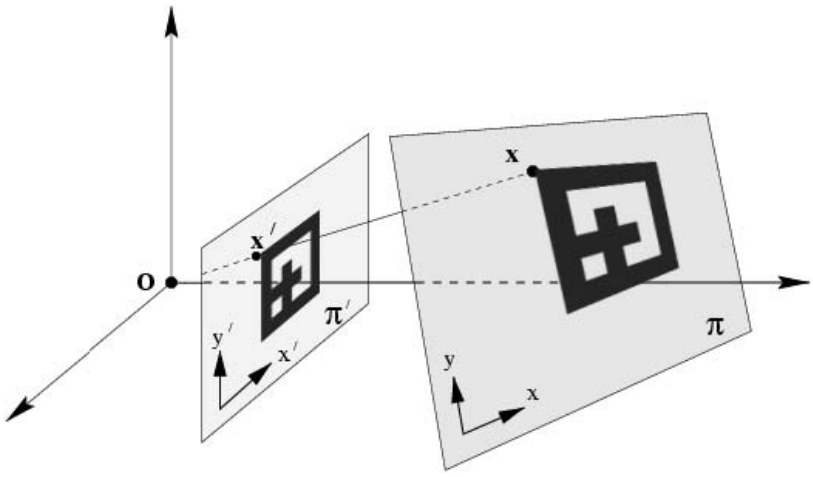


Figure 3.2.: **Homography between marker and image plane.** From a minimum of four point to point correspondences $\{x_i, x_i'\}$ a projective transformation H can be computed that maps each point x_i on the marker plane to its projection x_i' on the image plane. This relationship can be expressed as $x_i' = Hx_i$. *Image adapted from [32]*

8 x 9 and is of rank 8 and thus has a 1-dimensional null-space which provides a solution for h. The equation can be solved by Singular Value Decomposition (SVD), whereby the unit singular vector corresponding to the smallest singular value is the solution h. Such a solution can only be determined up to a non-zero scale factor. However, H is in general only determined up to scale, so the solution h gives the required H. The DLT algorithm minimizes the norm $\|Ah\|$ and thereby the algebraic error vector ϵ_i that is associated with the point correspondences $x_i \leftrightarrow x_i'$ and the homography H.

Retrieving the pose

If the internal camera configuration of the camera, the intrinsic matrix K, is known, the extrinsic parameters can be retrieved from the projection matrix P. Under the constraint that all model points are coplanar, this is also possible for the 2D homography H. Let's assume we computed the homography H mapping points on the planar marker to points on the image plane as described above, so that

$$x' = Hx \tag{3.2}$$

$$\begin{bmatrix} kx' \\ ky' \\ k \end{bmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3.3}$$

We further can define a projective transformation P that projects an arbitrary point in space onto the image plane. For homogeneous coordinates this projection is expressed by a 3 x 4 matrix. If a point X lies on the marker we can define its z-coordinate to be zero. X is then projected onto the image plane as

$$
\begin{aligned}
x' &= PX \\
&= K\,[R\,t] \begin{bmatrix} x \\ y \\ 0 \\ 1 \end{bmatrix} \\
&= K\,[R_1 R_2\,t] \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \\
&= Hx
\end{aligned}
\tag{3.4}
$$

with K being the intrinsic matrix, R a 3 x 3 rotation matrix defining the orientation of the camera coordinate system and t the translation between the camera center and the world coordinate origin. R_i denotes the i-th column of the matrix R. This leads us to

$$H \propto K\,[R_1 R_2\,t] \tag{3.5}$$

where from R_1,R_2 and t can be retrieved by Singular Value Decomposition from the matrix $G = K^{-1}H$. G is defined up to a scale factor and since all axes of the camera coordinate system have to be orthogonal $R_3 = R_1 \times R_3$.

This solution gives a very coarse approximation of the camera's pose and is usually refined with a non-linear optimization method, such as Gauss-Newton or Levenberg-Marquardt.

3.3.2. Pose estimation between infrared tracking system and marker

As described above, when wanting to determine the spatial relationship between overhead camera and gazer, we need to estimate the poses of both the overhead camera and the gazer relative to the marker used as reference. These partial pose estimations are derived from the mapping between the marker in the 3D world and its projection onto the respective 2D image plane. While the marker is obviously visible in the gazer image, we have to face the problem that the overhead camera is equipped with an infrared filter and thus only infrared light can be seen on its image plane. To estimate the pose of the overhead camera in relation to the marker, we therefore have to find a way of making the marker visible to an infrared camera. Furthermore, to guarantee a well-defined reference, a single marker should be used for both pose estimations.

To provide a marker both visible to visual and infrared cameras an interactive marker was used. This interactive marker is the adaption of a regular visual marker augmented with infrared light sources. In specific, five infrared LEDs (Siemens LD271) were installed at well defined spots of the marker (Figure 3.3). In the image of the overhead camera these LEDs and thus the respective vertices of the marker can be clearly extinguished. To get the exact image coordinates, the image is therefore processed with simple thresholding and feature detection as shown in figure 3.4. In my application I use OpenCV library functions for image processing and the OpenCV based blob extraction library *cvBlobsLib* to identify

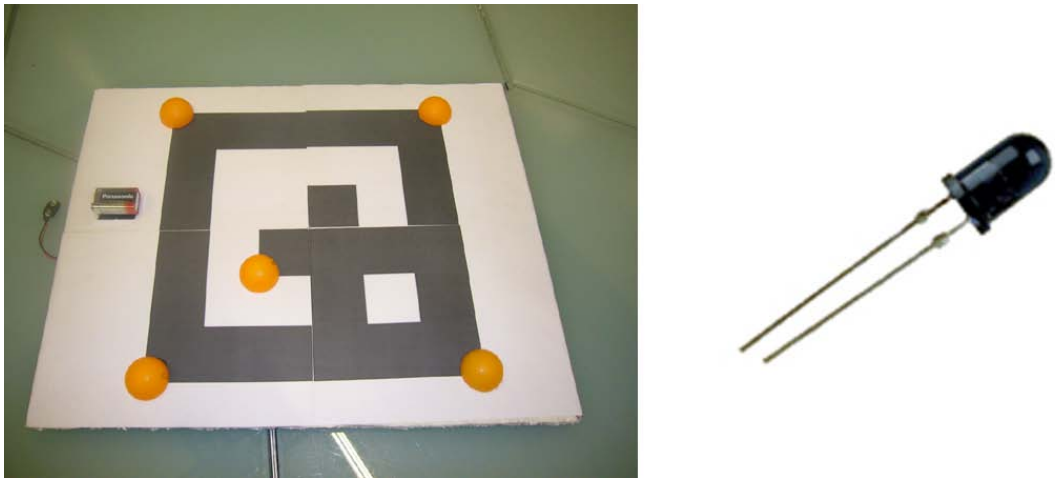


Figure 3.3.: **The interactive marker used for the control experiment.** Five infrared light emitting LEDs were installed at well defined vertices of the marker to make them visible to the infrared overhead camera. To avoid artifacts due to diffusion, the LEDs were covered with semi-transparent hemispheres.

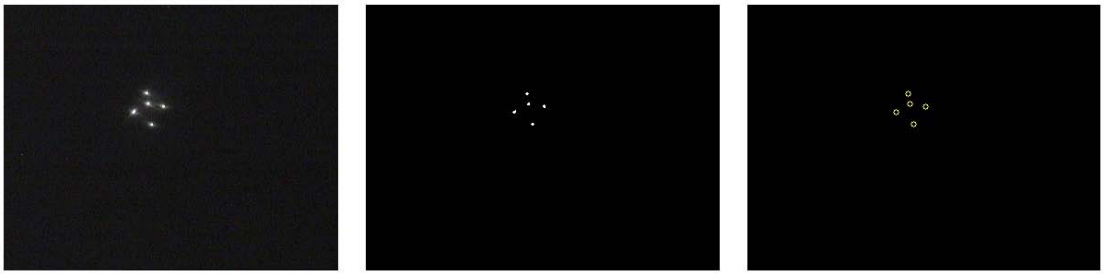


Figure 3.4.: **Processing of the infrared camera image.** The infrared LEDs attached to the marker can be clearly extinguished in the image of the infrared camera (a). By simple thresholding (b) and region detection the marker vertices can be identified (c).

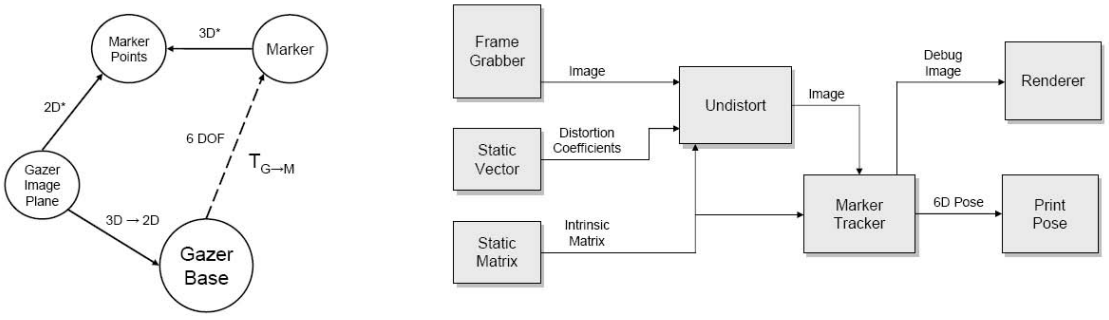


Figure 3.5.: **Spatial Relationship Graph and Dataflow Network for the Marker Tracking.**

the individual regions in the image. Returned is a list of points that indicate the centers of the detected "blobs". We can relate this points on the image plane to the coordinates of the LEDs in the marker coordinate frame and compute the pose of the infrared camera from these relationships as described above in section 3.3.

3.3.3. Pose estimation between gazer and marker

So far we've determined the pose transformation between the marker and the overhead infrared camera. In order to estimate the pose of a gazer in the overhead camera coordinate frame, we still need to find out how this gazer and the marker relate to each other. The identification of a marker in a camera image and the estimation of its pose is a common problem in the context of marker based real time tracking. Challenges thereby lie in the identification and segmentation of a marker with defined pattern in the image, the definition of valid correspondences between points on the marker and the image plane, and the numerical computation of the pose. For estimating the pose between marker and gazer in the XIM I could recall on the marker tracking implemented in the Ubitrack library [26].

The Ubitrack library has been developed at the Chair for Computer Aided Medical Procedures and Augmented Reality (CAMP-AR). It has been designed to provide a framework for the automatic and dynamic fusing of widespread and heterogeneous tracking sensors. Such a hybrid tracking has become a premise in large scale, ubiquitous Augmented Reality applications. The implementation of Ubitrack is based on the formal model of Spatial Relationship Graphs (SRG). In these graphs nodes represent objects or coordinate frames and edges spatial relationships, in specific tracked or known transformations between these objects. All algorithms used for tracking and calibration can be mapped to particular patterns in such a graph, so called Spatial Relationship Patterns. By looking for such patterns in an SRG, the Ubitrack middleware can, given the description of a tracking setup, create data flow networks to fulfill a client's request [44]. One of these patterns, the *Marker Tracker* pattern, fits exactly the problem of estimating the pose transformation between a camera and a marker. From the SRG introduced in figure 3.1 to sketch the overall setup of our control experiment we can segment the relevant part (Figure 3.5, left) and

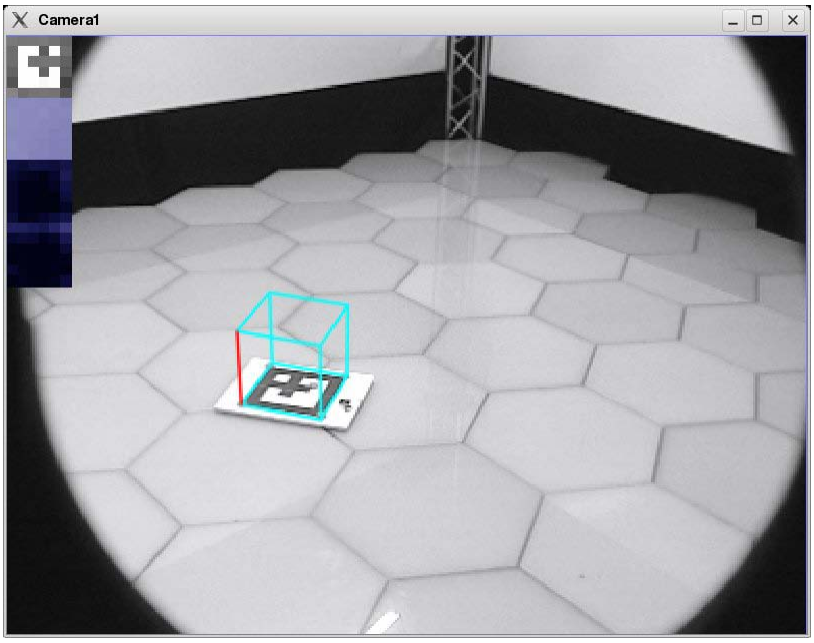


Figure 3.6.: **Debug image of the gazer during pose estimation using the Ubitrack.**

derive the respective Dataflow Network (Figure 3.5, right). The *Marker Tracker* requires as an input, besides the image from the framegrabber, the intrinsic matrix that defines the internal camera configuration of the camera. For the gazers, the intrinsic matrices have been determined in a prior calibration scenario as described in section 3.3.4, as well as a 4-vector specifying the distortion coefficients. Intrinsic matrix and distortion coefficients also allow the undistortion of the gazer image in an *Undistort* pattern before passing it on to the *Marker Tracker*. As an output the *Marker Tracker* delivers the 6D-pose, that gives an estimate to the rotation and translation of the marker in relation to the gazer's coordinate frame. A debug image is rendered and displayed to show if the marker was detected correctly (Figure 3.6).

3.3.4. Intrinsic Calibration

In order to estimate a camera's pose from the image of a marker, the camera's intrinsic parameters need to be known. These parameters are:

- The focal length f, that is the distance between the camera lens and the image plane,
- the location of the image center $p_{x,y}$ in pixel coordinates (the principal point),
- the efective pixel size $m_{x,y}$ and
- the radial distortion coefficient of the lens.

For a general CCD camera these parameters constitute the intrinsic matrix

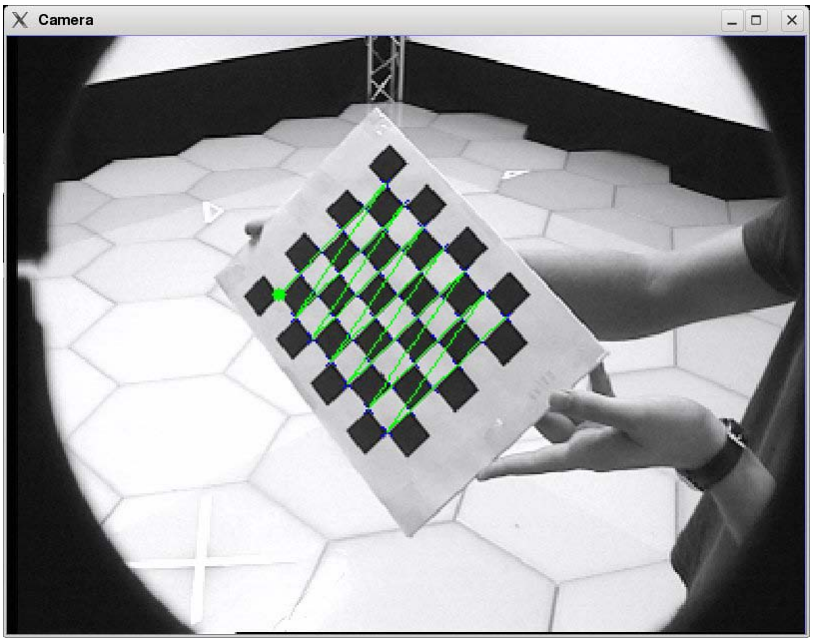


Figure 3.7.: **Intrinsic camera calibration using the Ubitrack chessboard calibration pattern.**

$$K = \begin{bmatrix} \alpha_x & & p_x \\ & \alpha_y & p_y \\ & & 1 \end{bmatrix} \tag{3.6}$$

where $\alpha_x = fm_x$ and $\alpha_y = fm_y$ represent the focal length of the camera in terms of pixel dimensions in the x and y direction respectively. Within the projection matrix $P = K\,[R|t]$, K maps a point in the camera coordinate frame onto the image plane.

The internal camera configuration can be determined from different properties in the image. I will however not cover this subject in this thesis and refer to [32] for further reading. For the specific task of calibrating the gazer cameras in the XIM, the Ubitrack library was used. Ubitrack features a chessboard calibration, where the calibration routine is supplied with a number of images of a planar chessboard pattern with known geometry (Figure 3.7). The corners of the individual squares are identified in the image and related to the world coordinates of the chessboard to compute the intrinsic matrix and the distortion coefficients.

3.4. Implementation and results

A marker based pose estimation is subject to a number of different error sources. Besides numerical errors in the computation, the pose estimation is effected by measurement errors in the image plane as well as on the marker. Furthermore, errors in the pose estimation can be related to the topology of the individual fiducials on the marker. Research on this

issues ranges from the design of the optimal tracking probe [57] to different approaches of predicting the accuracy in pose estimation for marker based tracking [24, 40, 9]. The combination of numerical errors, jitter, tracker bias and probe deformations makes marker tracking inaccurate by nature. Experiments have shown, that the correctness of a pose estimation is thereby effected by the distance to the marker, the size of the marker, but also the slant angle of the camera [43]. As a matter of fact, markers with a slant angle of zero to the camera, that is perpendicular to the camera normal, can have errors up to 15 degrees, which is greater than for any other view angle. As markers do not face the camera, the error decreases, with 45 degree camera tilt seeming to be optimal. In my setup, the accuracy of the pose estimation is severely affected by this error, as the overhead infrared camera looks almost straight onto the marker (at least, if the marker is placed on the floor).

To compensate for errors in the marker based pose estimation, the control experiment was run ten times for each gazer with different marker positions. The pose transformations of both the overhead camera and the respective gazer were computed for each marker positions. To combine the individual transformations and determine an optimal estimate for the gazer's pose, two different approaches have been implemented and will be compared in the following:

- For each marker position i the gazer's pose in the overhead camera's coordinate frame is directly computed as $T_{OH\rightarrow G} = {T_{G\rightarrow M}}^{-1} T_{OH\rightarrow G}$. The final pose is the determined as an average of all ten results.

- In a *Bundle Adjustment* the pose of the gazer is globally optimized by the minimization of a cost function that involves all parameters from both partial pose estimations and the measurements of all ten marker positions.

3.4.1. Direct computation

In homogeneous form a camera's pose is written as

$$T = \begin{bmatrix} R & \vec{t} \\ \vec{0}^T & 1 \end{bmatrix} \tag{3.7}$$

where R is the 3 x 3 rotation matrix and $\vec{t}$ the 3-vector denoting the translation of the camera center. If all transformations are given in such homogeneous form, their composition may be written in terms of matrix multiplication. The matrix $T = T_2 T_1$ for example, first carries out the transformation T_1 and then the transformation T_2. In the case of my control experiment, I have determined the poses of the overhead camera $T_{OH\rightarrow M}$ and $T_{G\rightarrow M}$. In particular, $T_{OH\rightarrow M}$ describes the position and orientation of the marker coordinate frame in overhead camera coordinates and $T_{G\rightarrow M}$ analogously the pose of the marker in relation to the gazer. Consequently, the inverse transformation ${T_{G\rightarrow M}}^{-1}$ specifies the gazer's pose in marker coordinates and

$$T_{OH\rightarrow G} = {T_{G\rightarrow M}}^{-1} T_{OH\rightarrow M} \tag{3.8}$$

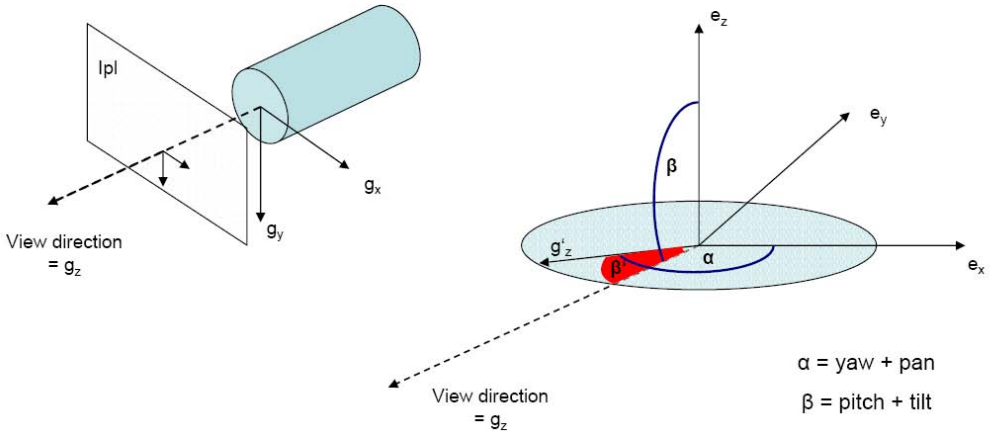


Figure 3.8.: **Model for the retrieval of the yaw and pitch angles from the rotation matrix**.

the gazer's pose in overhead camera coordinates. This relationship is also visualized in the Spatial Relationship Graph in Figure 3.1.

In order to enable a composition of the individual transformations in terms of matrix multiplication, all measured poses $T_{OH\rightarrow M,\,i}$ and $T_{G\rightarrow M,\,i}$ are converted to homogeneous matrix form as in equation 3.7. To receive an average of all measurements, we sum up the individual products $T_{G\rightarrow M,\,1...10}$ and divide the resulting matrix element wise by ten. The average pose estimate for the respective gazer is then

$$T_{avg} = \left(\sum_{i=1}^{10} T_{OH\rightarrow M,\,i}\right)./10 = \begin{bmatrix} R_{avg} & \vec{t}_{avg} \\ \vec{0}^T & 1 \end{bmatrix} \tag{3.9}$$

where from $\vec{t}_{avg}$ can be adopted as the position of the gazer. R_{avg} denotes the average orientation of the gazer in the experimental setup.

Retrieval of yaw and pitch from a rotation matrix

In order to extract the yaw and pitch angles of the gazer from the rotation matrix, we have to consider that the gazer was set to a specific pan and tilt configuration for the duration of the experiments. The found orientation encapsulated in the matrix R_{avg} thus specifies the orientation in space for these specific angles. Since the gazer configuration is known, it is easy to retrieve the desired angles. Note that the z-axis of the gazer coordinate frame is equivalent to the view direction of the gazer (Figure 3.8, left). Within the rotation matrix R, the z-axis is specified by the third column vector, furthermore referred to as $\vec{g}_z$. The angle

$$\beta = \vec{g}_z \times \vec{e}_z \tag{3.10}$$

Gazer ID	25	300	352	365
X	228.7	-230.9	-247.9	252.7
Y	-249.7	240.4	-226.1	244.7
Z	*(340 - 162.5)*	*(340 - 148.5)*	*(340 - 187.1)*	*(340 - 160.9)*
Yaw_{xim}	43.8	38.6	42.3	42.9
Yaw_{deg}	92.6	81.7	89.5	90.8
$Pitch_{xim}$	29.2	28.5	27.4	30.5
$Pitch_{deg}$	33.0	32.2	31.0	34.5

Table 3.1.: **Estimated poses for the four gazers by averaging the results of the marker based pose estimation.**

between $\vec{g}_z$ and the z-axis of the euclidean coordinate frame is the the sum of *pitch* and *tilt*. Rotating $\vec{g}_z$ into the euclidean x-y-plane by an angle $\beta' = \beta - 90°$, we can compute the angle α from the resulting vector $\vec{g}'_z$ as

$$\alpha = \vec{g}'_z \times \vec{e}_x \tag{3.11}$$

where α is the sum of *yaw* and *pan*. With known angles pan and tilt for the done measurements, were can retrieve the *yaw* and *pitch* of the specific gazer from the rotation matrix R.

The results of the pose estimation by averaging the results of the poses measured for ten different markers are shown in table 3.1 and illustrated in figures 3.10 and 3.11. In table 3.1 the poses are thereby listed in terms of XIM coordinates. This includes a further rotation of $180°$ around the x-axis of the overhead camera's coordinate system, since the XIM coordinate system, though spanned by the overhead camera, lies in the floor of the XIM. As a consequence, the z-value of the gazer's pose is approximated by subtracting the z-value of the computed pose from the estimated height of the infrared camera.

3.4.2. Bundle adjustment

So far I have tried to minimize the error in the marker based pose estimation by simply averaging the measurements made for ten different marker positions. This intuitive approach certainly yields a better result than just estimating the pose from one single marker position, but still it remains a very primitive approximation sensitive to the errors that arise from the individual measurements. A more sophisticated approach would be to estimate the gazer's pose globally from measurements made for different marker positions in regard of all involved parameters. One technique to do so is the so called *bundle adjustment*.

> **Bundle adjustment** is the problem of refining a visual reconstruction to produce *jointly optimal* 3D structure and viewing parameter (camera pose and/or calibration) estimates.

In this definition of Triggs et al. [53] *optimal* means that the parameter estimates are found by minimizing some cost function that quantifies the model fitting error, and *jointly* that

the solution is simultaneously optimal with respect to both structure (marker) and camera variations. The name refers to the "bundles" of light rays leaving each 3D feature and converging on each camera center, which are "adjusted" optimally with both feature and camera positions. Thereby all of the structure and camera parameters are adjusted "in one bundle" [53]. In general, bundle adjustment can be seen as a large, sparse geometric parameter estimation problem. The parameters regarded in the optimization go beyond the pure camera pose and include also the 3D feature coordinates (or, in our case, the transformations between the individual coordinates) and eventually the camera's intrinsic calibration. It can be applied to many similar estimation problems in vision, photogrammetry, industrial metrology, surveying and geodesy. The adaption to the particular problem is largely a matter of choosing a numerical optimization scheme that exploits the problem structure and sparsity. There is a wide range of literature on the choice of the optimization scheme and the formulation of an appropriate cost function. Classically bundle adjustments are formulated as non-linear least squares problems [18, 30, 49, 20, 21, 6, 59]. Modern systems on the other hand are often developed for general robust cost functions, rather than restricting attention to traditional non-linear squares [53].

When estimating the pose of the gazers in the XIM, we have to presume noisy measurements in the images and numerical instabilities. A global optimization by use of a bundle adjustment might contribute essentially to better estimation results. For the specific case, I base my derivation on the formal description of the bundle adjustment given by Hartley and Zisserman [32]. We have to consider a situation in which a set of 3D points X_j (the vertices on the marker) is viewed by a set of cameras with matrices P^i. Denote by x^i_j the coordinates of the j-th point as seen by the i-th camera. We then wish to solve the following optimization problem: Given the set of image coordinates x^i_j find the set of camera matrices P^i and the points X_j such that $P^i X_j = x^i_j$. If the image measurements are noisy, this equation of course will not be satisfied exactly. In this case we seek the Maximum Likelihood (ML) solution assuming that the measurement noise is Gaussian. In particular, we wish to estimate projection matrices $\hat{P}^i$ and 3D points $\hat{X}_j$ which project exactly to the image points $\hat{x}^i_j$. Also, we want to minimize the image distance between the reprojected points and the detected (measured) image points x^i_j for every view, i.e.

$$\min_{\hat{P}^i, \hat{X}_j} \sum_{ij} d\left(\hat{P}^i \hat{X}_j, x^i_j\right)^2 \tag{3.12}$$

where $d(x, y)$ is the geometric image distance between the homogeneous points x and y.

To implement the bundle adjustment for the control experiment we have to parameterize the problem and set up a cost function according to equation 3.12 that involves minimizing the reprojection error. The estimation shall be carried out for two cameras $\hat{P}^1$ and $\hat{P}^2$, the overhead camera and the respective gazer, for which the intrinsic parameters are know. The 3D points $\hat{X}_j$ are constituted by the marker vertices at the different marker positions. For ten marker positions we thus get a total of 5 x 10 points, where from five at a time lie in a fixed coordinate frame. The problem is therefore not to optimize the individual points, but rather the transformations between the maker positions. In order to do so we define the coordinate frame of the first marker as reference and a transformation $\hat{M}_k$

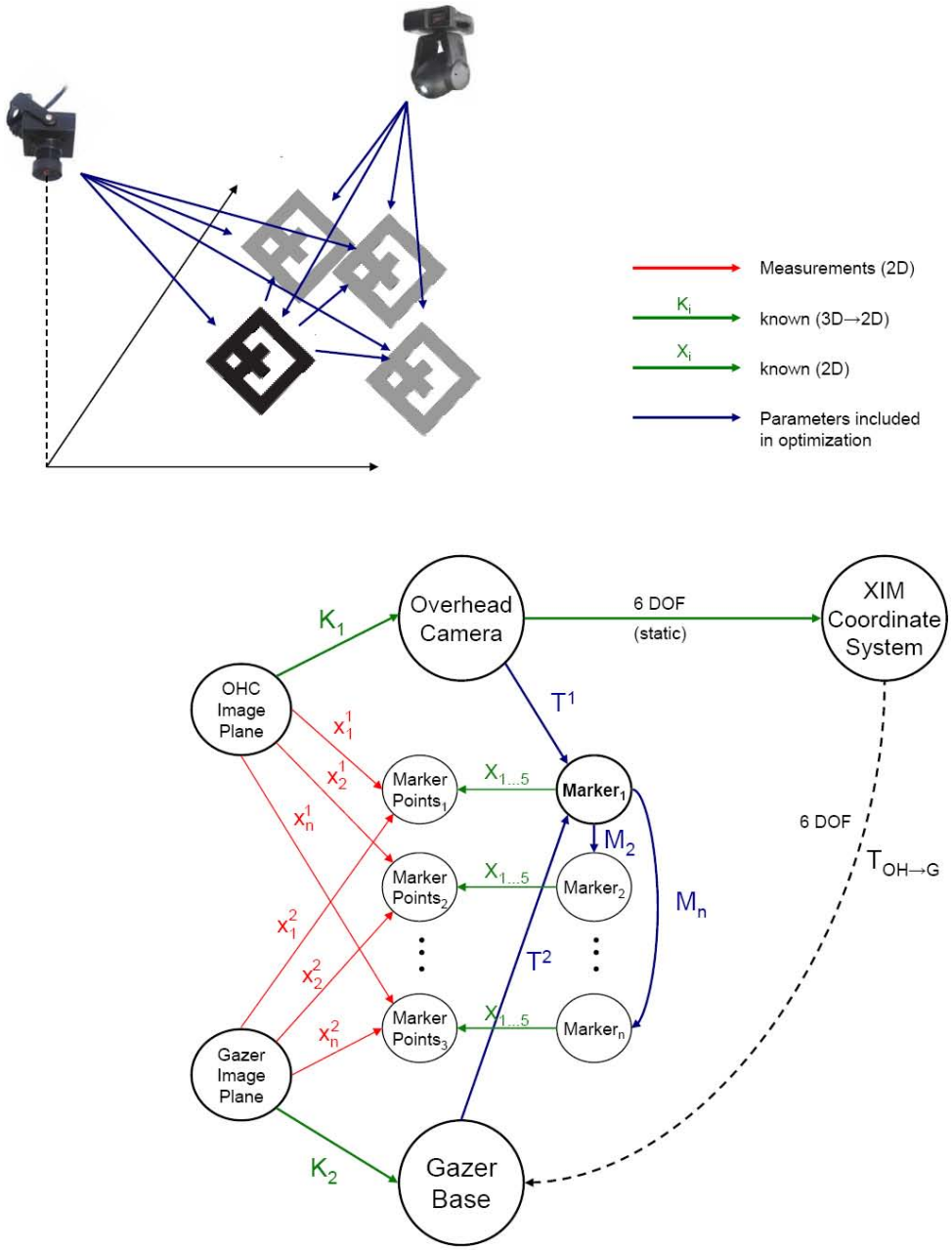


Figure 3.9.: **Spatial Relationship Graph for the bundle adjustment**.

to this reference for each successive marker. These transformation will be included in the optimization. With $\hat{M}_1$ defined as the identity matrix, we get a projection

$$\hat{P}_k^i = K_i \cdot \hat{T}^i \cdot \hat{M}_k \tag{3.13}$$

for each camera i and each marker position k that maps a point from the marker onto the image plane. K_i denotes the intrinsic matrix of the camera i that is considered to be known. The optimization problem can then be formulated as the minimization of the sum of squares

$$\min_{\hat{T}^i, \hat{M}_k} \sum_{ijk} d\left(\hat{P}_k^i \hat{X}_j, x_{jk}^i\right)^2. \tag{3.14}$$

The task is now to find an optimal fit to the parameter vector $\vec{p} = (T_1 \cdots T_i M_1 \cdots M_k)^T$ so that the upper equation becomes minimal. This can be solved by use of any non-linear optimizer. In this particular case, I propose the use of the Levenberg-Marquard Algorithm (LMA) to determine the optimal set of parameters. I therefore set up a system of equations $f_l\left(\vec{p}|X_j\right)$ that relates each 3D point X_j with its 2D images x_l^i for each camera i and each marker transformation M_k. For our setup we thus get a system of $l = 2 \cdot k \cdot j$ equations:

$$\begin{aligned}
x_j^1 &= K_1 T_1 M_1 X_j \\
x_{2j}^1 &= K_1 T_1 M_2 X_j \\
&\vdots \\
x_{kj}^1 &= K_1 T_1 M_k X_j \\
x_j^2 &= K_2 T_2 M_1 X_j \\
x_{2j}^2 &= K_2 T_2 M_2 X_j \\
&\vdots \\
x_{kj}^2 &= K_2 T_2 M_k X_j
\end{aligned} \tag{3.15}$$

As described in chapter 2.3.2, the LMA is an iterative procedure that in each step replaces the parameter vector $\vec{p}$ by a new estimate $\vec{p} + \vec{q}$. To determine $\vec{q}$, each individual function $f\left(\vec{p} + \vec{q}\right)$ is approximated by its linearizations $f\left(\vec{p} + \vec{q}\right) = f\left(\vec{p}\right) + J\vec{q}$, where J is the Jacobian of f at $\vec{p}$. We therefore have to compose the Jacobian matrix J that contains all the partial derivatives and provide it to the LMA. J is thereby sparsely populated and made up by the scheme

$$J = \begin{array}{c|cccccc} & T_1 & T_2 & M_1 & M_2 & \cdots & M_3 \\ x_1^1 & \bullet & & \bullet & & & \\ x_2^1 & \bullet & & & \bullet & & \\ \vdots & \vdots & & & & \ddots & \\ x_{kj}^1 & \bullet & & & & & \bullet \\ x_1^2 & & \bullet & \bullet & & & \\ x_2^2 & & \bullet & & \bullet & & \\ \vdots & & \vdots & & & \ddots & \\ x_{kj}^2 & & \bullet & & & & \bullet \end{array} \tag{3.16}$$

where a bullet denotes the respective partial derivative of x_l^i and all other entries are zero. The sparsity of the Jacobian matrix can be exploited significantly by the implementation of the bundle adjustment to save computational cost.

Bundle Adjustment results

The bundle adjustment was run for the poses measured as described in section 3.3. The overhead camera is defined to be the first camera and the gazer the second, with T^1 consequently being the pose of the overhead camera and T^2 the one of the gazer. Since the implementation of the algorithm in the Ubitrack uses the first marker as the reference coordinate system, the estimates of these poses, $\hat{T}^1$ and $\hat{T}^2$, are returned in relation to even this first marker. The estimated pose of the gazer in relation to the overhead camera is therefore given by

$$T_{bundleadjustment} = \left(\hat{T}^1\right)^{-1} \hat{T}^2. \tag{3.17}$$

The position and orientation of the gazer can be retrieved from this pose as described above in section 3.4.1. The results of the estimation with use of the bundle adjustment are listed in table 3.2 and illustrated in figures 3.10 and 3.11. Interestingly, for three of the gazers the results vary only slightly from the ones determined by averaging the results of the individual measurements. Only for the gazer with the ID 25, a significantly different pose was estimated. The decisive question remains, in how far the different estimation result contribute to a better performance in adjusting the gazers to look at a specific spot in the XIM. To find out, which of them yields the best results, the computation of angles based on the different pose estimations will be compared in the following section.

Robust Bundle Adjustment

When looking at the poses estimated by the bundle adjustment, we see that all of them vary only slightly from the ones estimated directly. Only exception is the gazer with ID 25, for which a significantly different pose has been estimated. Investigating the output of the Levenberg Marquardt Algorithm used by the bundle adjustment showed, that a severe residual error remained, also after optimization. While the LMA converged with a residual error of about four pixels (SSD) for the other gazers, an error of more than 130

Gazer ID	25 w/ Outliers	25	300	352	365
X	185.7	227.0	-226.4	-243.7	255.7
Y	-226.7	-250.2	239.4	-226.9	259.8
Z	*(340 - 100.3)*	*(340 - 161.8)*	*(340 - 141.8)*	*(340 - 184.6)*	*(340 - 183)*
Yaw_{xim}	39.7	43.6	38.4	42.6	42.7
Yaw_{deg}	84.0	92.3	81.7	93.8	84.5
$Pitch_{xim}$	41.2	30.0	29.3	28.2	27.5
$Pitch_{deg}$	46.6	33.9	33.2	31.9	39.0

Table 3.2.: **Estimated poses for the four gazers using the bundle adjustment.**

pixels remained for gazer 25. This high residual is caused by outliers, specific markers or features that carry a high geometric error. As the LMA tries to find an optimal fit to the parameter vector and thereby considers all markers, the final output is severely affected by such outliers. To avoid these complications, a robust implementation of the LMA identifies outliers and disregards them. In the specific case, one such outlier could be identified. Excluding this marker from the optimization resulted in a way lower residual and thus a more significant pose. In the following evaluation, both poses estimated for gazer 25 using the bundle adjustment will be considered, the malicious one with the outliers and the ”enhanced” one with the ouliers removed.

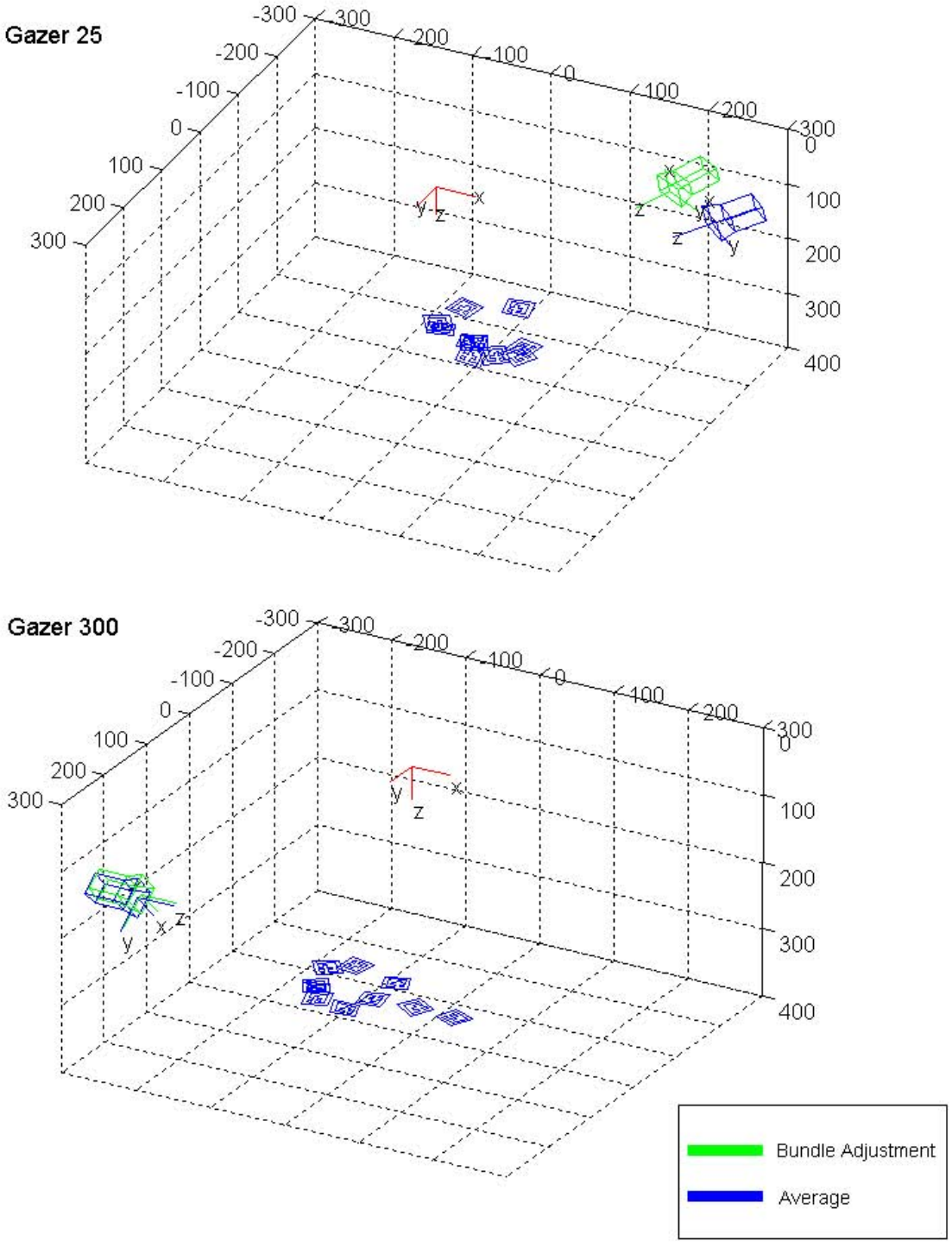


Figure 3.10.: **Estimated poses of the first two gazers (IDs 25 and 300).**

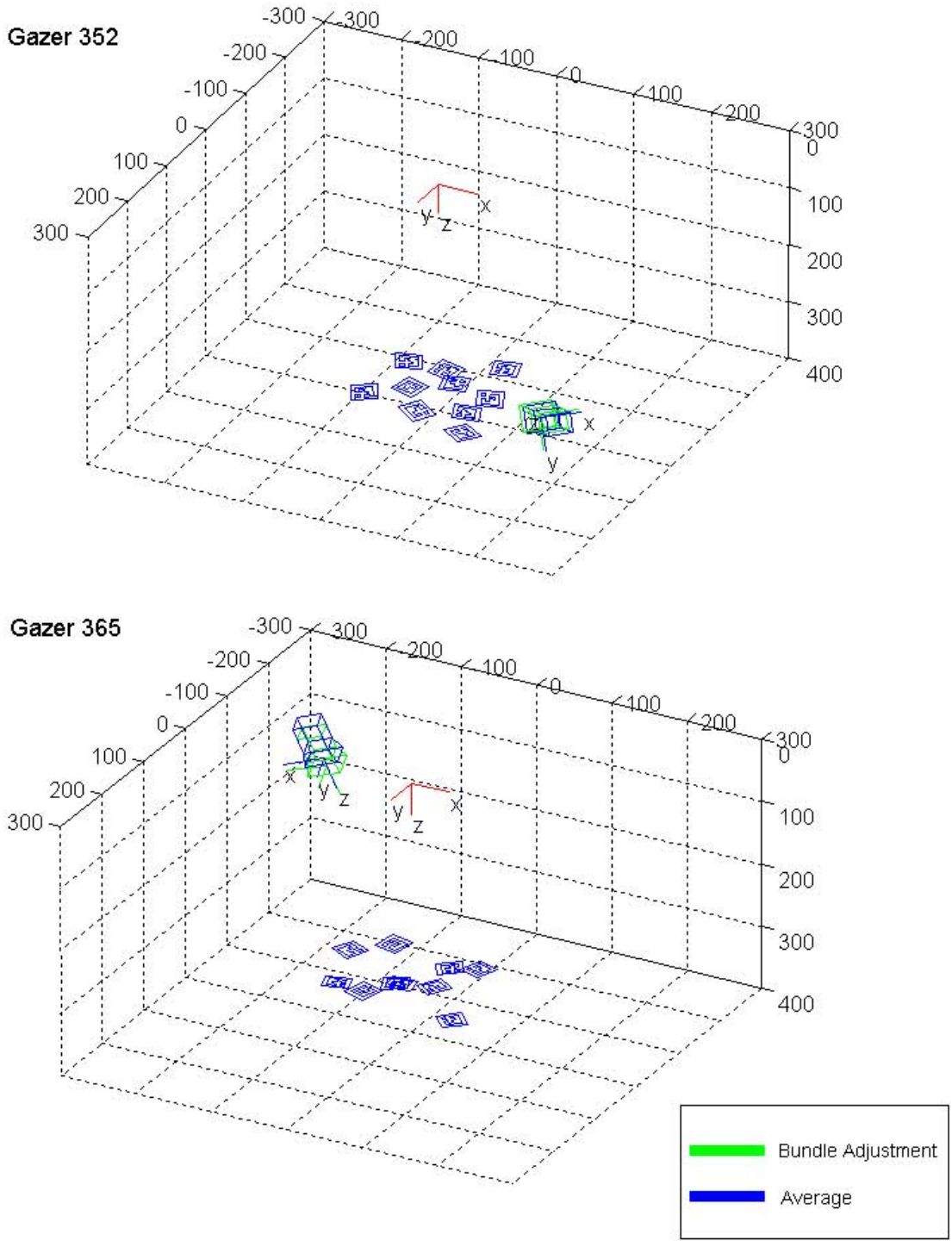


Figure 3.11.: **Estimated poses of the second two gazers (IDs 352 and 365).**

3.5. Comparison of the different approaches

"Only what gets measured gets done!"

Main objective of this project is the use of the four gazers to gain additional information about certain entities in the XIM. On command a gazer therefore needs to look at a certain position in the space. The internal rotation that needs to be set for the gazer to look at this position can be computed if the gazers position and orientation in space, its pose, are known. Consequently, the correctness of the the adjustment is strongly dependent of the correctness of this pose.

In the previous I have introduced three different approaches of estimating a gazers pose. An optimization based on known correspondences gained in a classifier-based calibration scenario, and two implementations of a pose estimation by use of a marker pattern. We have seen approaches that vary strongly and are subject to different sources of error.

- In the classifier approach a set of correspondences between positions in the space, as provided by the tracking system, and gazer adjustments is determined in a calibration scenario. A non-linear optimizer is then used to find the optimal pose parameters. Optimal means that the angles computed for each of the tracking positions based on the pose differ as low as possible from the measured angles. The correctness of the determined pose is affected by a number of errors, particularly severe inaccuracies in the tracking data due to perspective distortion. As a consequence, the result seems to be way of the "real" pose, at least in a geometric sense. Still it may serve as an input to compute the angles corresponding to any specific spot in the space as precise as possible.
- The marker based posed estimation on the other hand determines the position and orientation of the gazer in relation to the tracking camera by seeking an optimal estimate of the spatial relationship between these two entities. As a marker based pose estimation is exposed to various error sources, I have presented two attempts to compensate for these errors: A rather intuitive approach of averaging the results of several individual pose estimations, and a global optimization by use of a bundle adjustment.

In order to judge on the quality of the results from the different approaches, we have to go beyond the correctness of the pose in a geometric sense. The decisive question remains: Which of the determined poses serves best to compute the most precise gazer adjustment for any arbitrary position in the space. I will therefore evaluate the results of the calibrations in two senses:

1. In chapter 2.6.3 I have given a measure of accuracy to the computation based on the pose determined in the classifier approach, by evaluating the deviation in the gazer image after the gazer has been adjusted as computed. We now want to find out, if better results would have been achieved in this test scenario, if the pose estimated by one of the marker based approaches would have been used.
2. The perspective distortion of the tracking camera has been made out as one of the main error sources in the classifier approach. We can try to minimize this error by

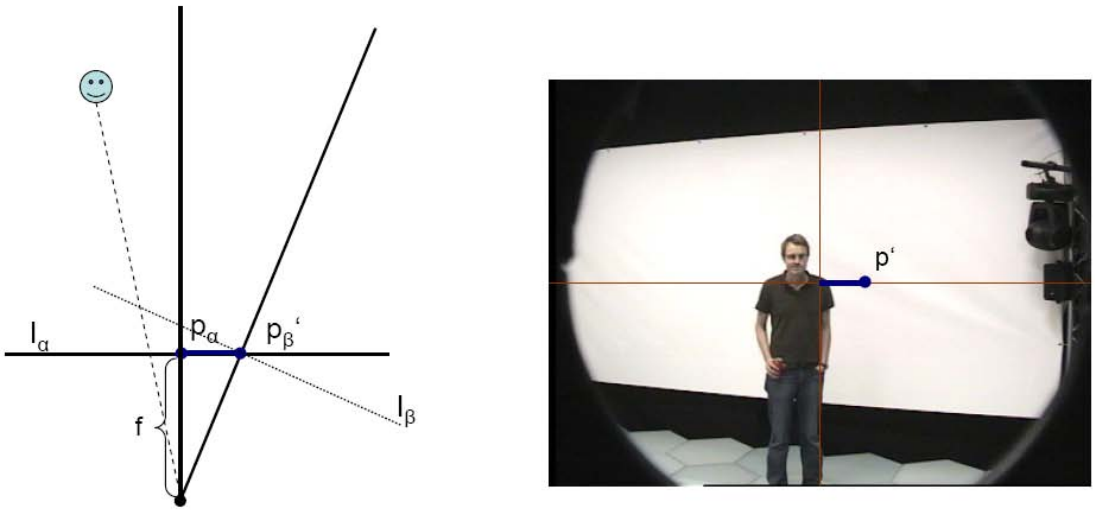


Figure 3.12.: **Computation of the shift between two images I_α and I_β.**

tracking a flat marker for which the position relative to the XIM tracking camera can be measured without perspective distortion. In this case, which of the determined poses will lead us to the best results?

In the following I will describe two experiments that have been conducted to provide an answer to these questions. I thereby confine to comparing the pan angles, since the exact height of the gazers (relative to the upper body of the person from the classifier approach) is not known from the marker based pose estimations. A reliable conclusion about the quality of the tilt computation is thus not possible and will be disregarded at this point.

3.5.1. Comparison based on a tracked person

Based on the experiment described in chapter 2.6.3 to evaluate the performance of the classifier, we want to see if better result could have been achieved if either one of the poses determined in the marker based pose estimations would have been used as a basis of computation. In the classifier experiment, a measure of accuracy was seeked by analyzing the gazer images taken after the gazer was adjusted to look at 50 positions in the space. The positions were provided by the tracking system to indicate the standpoints of a person at various spots throughout the space. After each respective adjustment, the person should be seen in the center of the image. The classifier was applied to the image to detect the true position in image coordinates and see how much this position deviates from the image center.

To provide the same measure for the two other poses, and thus make the different solutions comparable, the experiment would have to be conducted in the very same way. At time of evaluation this was not possible, since there was no further access to the XIM. It is therefore only possible to give a hypothetical solution: If different poses would have

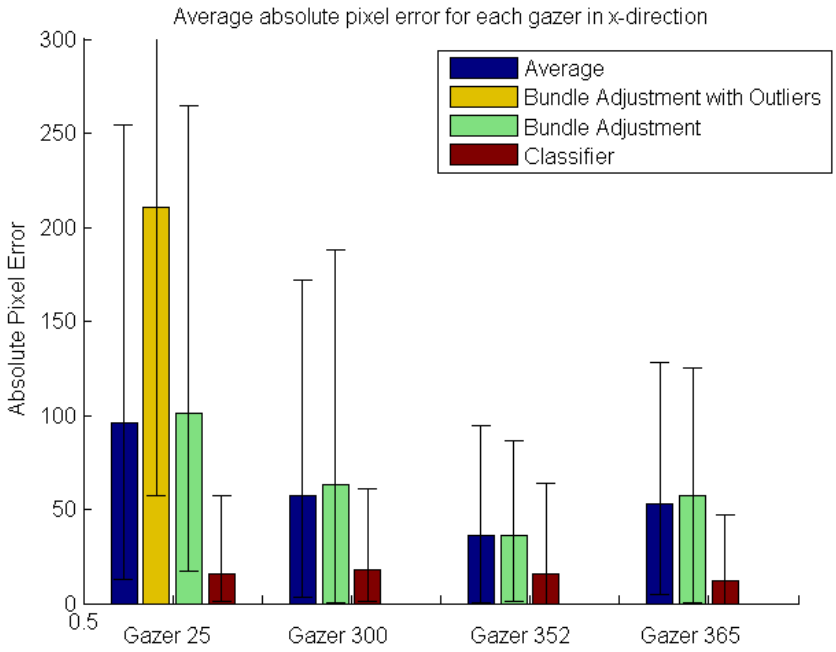


Figure 3.13.: **Average absolute deviations in x-direction for the different calibrations of the gazers.**

been used, how far would the resulting image have been shifted? As a consequence, how would this have changed the deviation?

Let I_α be the image of a specific gazer with a pan of α (considering the tilt to be fixed), and I_β the image of the same gazer with a pan of β. We then need to know, where to find the image center (in particular, the principal point) of I_β in I_α. Denote as p_α the principal point of I_α and as p'_β the imaginary principal point of I_β in I_α (Figure 3.12). Since the intrinsic camera configuration of each gazer is known, we can compute the distance in pixel between p_α and p'_β from the focal length f as

$$\delta x = tan\,(\beta - \alpha) \cdot f. \tag{3.18}$$

If we presume e.g. a focal length of 458.6 (as it is the case for gazer 25), an additional rotation of 1° would result in a δx of approximately 8 pixel. Note that this computation is strictly hypothetical, as we presume a linear shift. This obviously is not true, but for small angles equation 3.18 gives an adequate approximation to the true shift.

In chapter 2.6.3, for each gazer I have examined the images of 50 adjustments $(pan, tilt)_i$ to find the deviations Δx and Δy between the person in the image and the image center. If (x, y) defines the image coordinates of the detected person and c the image center, we can compute the hypothetic deviation $\Delta x'$ for an additional rotation γ relative to the respective adjustment as

$$\Delta x' = x - c_x + tan(\gamma) \cdot f. \tag{3.19}$$

Based on the poses determined in the marker based approaches, the hypothetical deviations Δx_m (for the direct computation) and Δx_{ba} (for the bundle adjustment) have been computed according to formula 3.19. The results for each individual image are drawn in figures 3.14, 3.15, 3.16 and 3.17. Figure 3.13 opposes the average absolute deviations for the different calibrations as listed in table 3.3.

Gazer 25	Direct computation	Bundle Adjustment w/ Outliers	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**95.7**	**210.8**	**101.4**	**15.8**
$min \lvert\Delta x\rvert$	12.9	57.0	17.4	1.0
$max \lvert\Delta x\rvert$	254.3	553.4	264.7	57.0

Gazer 300	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**57.1**	**62.8**	**18.0**
$min \lvert\Delta x\rvert$	3.4	0.1	1.0
$max \lvert\Delta x\rvert$	171.9	188.2	61.0

Gazer 352	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**36.1**	**36.4**	**15.8**
$min \lvert\Delta x\rvert$	0.4	1.2	0.0
$max \lvert\Delta x\rvert$	94.4	86.3	64.0

Gazer 365	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**52.9**	**57.1**	**12.0**
$min \lvert\Delta x\rvert$	4.9	0.1	0.0
$max \lvert\Delta x\rvert$	128.5	125.4	47.0

Table 3.3.: **Average, minimum and maximum deviations in the person tracking experiment under the different calibrations.**

In this experiment, the pose estimated in the classifier approach yields the best results. With average deviations of around 15 pixel, the computation was clearly more accurate than the one based on the marker estimated poses. The average deviations for the direct computation vary from 36 to 95 pixel, likewise for the robust bundle adjustment. Really bad results have been achieved by the unrobust calibration of gazer 25, that has been subject to outliers. In this case, the deviation averaged 210 pixel. At first sight, the fact that

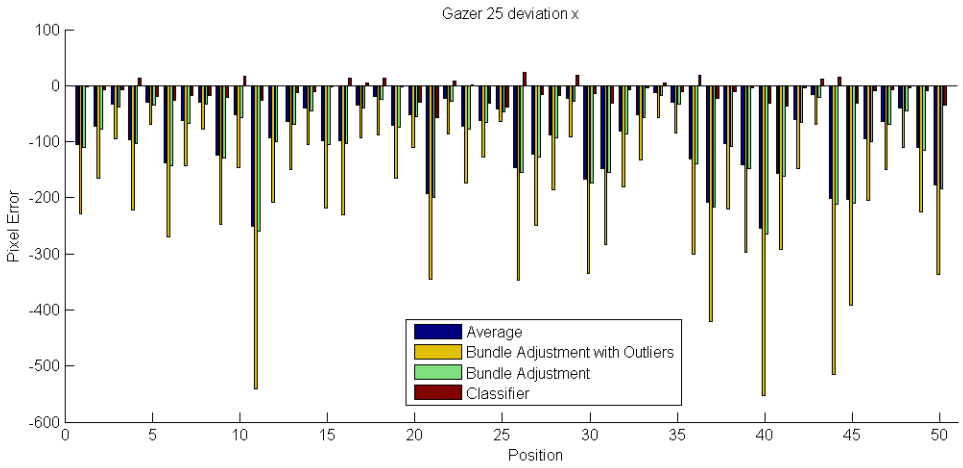


Figure 3.14.: **Accuracy of gazer 25 in the person tracking experiment under the different calibrations.**

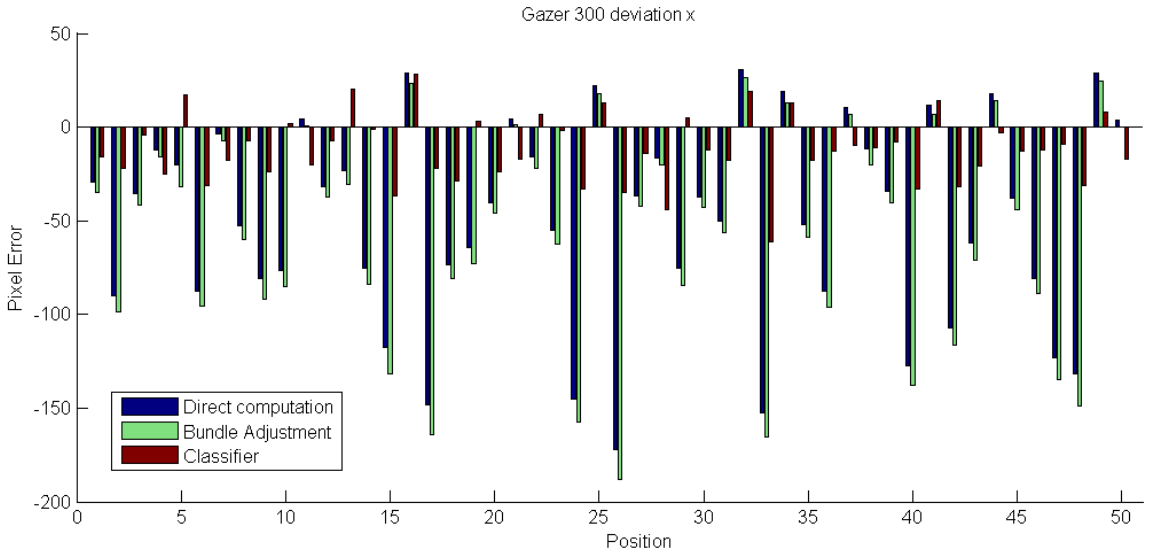


Figure 3.15.: **Accuracy of gazer 300 in the person tracking experiment under the different calibrations.**

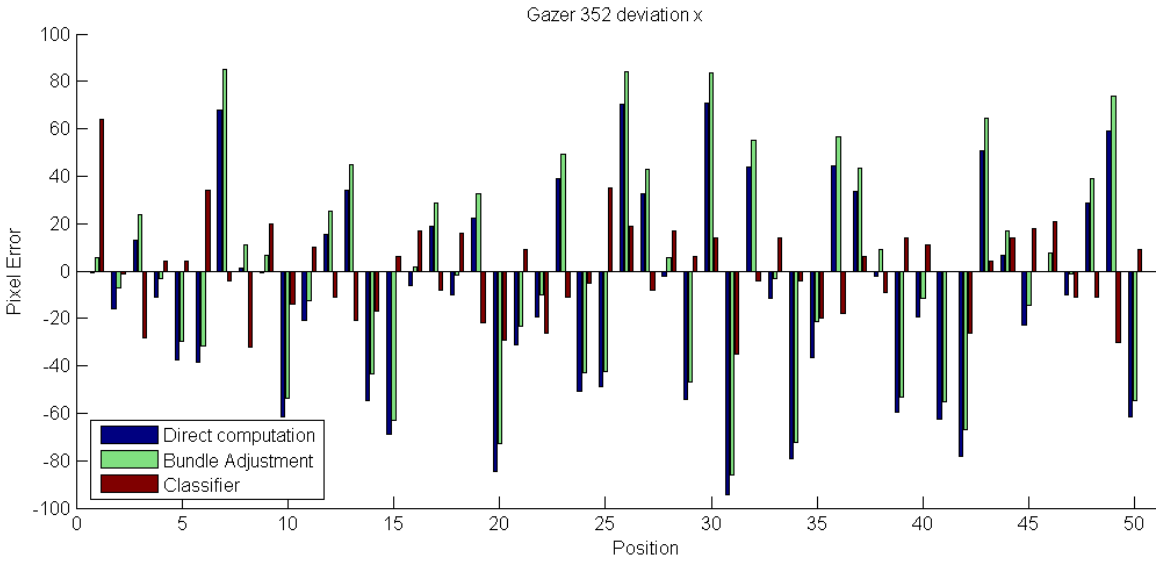


Figure 3.16.: **Accuracy of gazer 352 in the person tracking experiment under the different calibrations.**

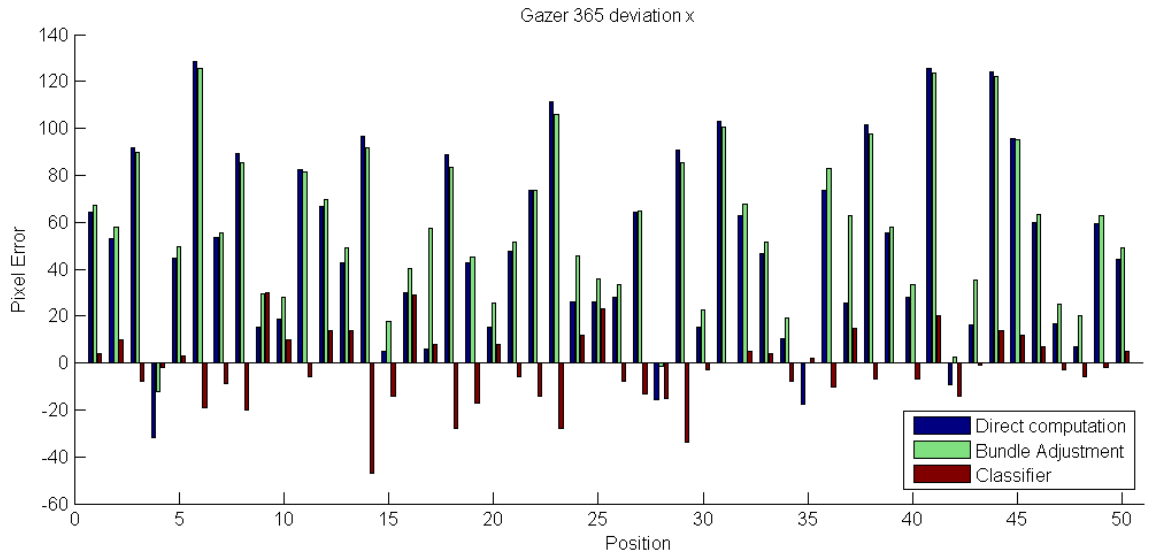


Figure 3.17.: **Accuracy of gazer 365 in the person tracking experiment under the different calibrations.**

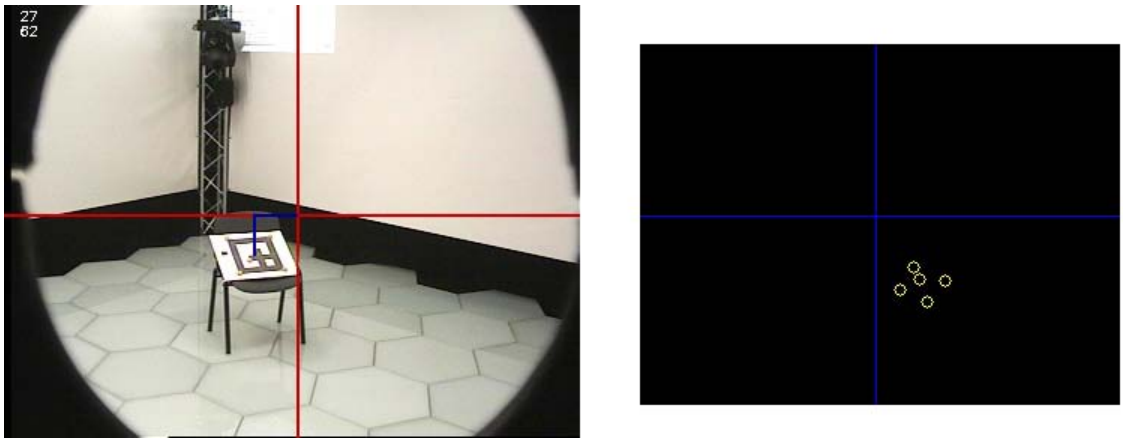


Figure 3.18.: **Setup for the comparison on basis of a reprojected marker.** We want to determine the deviation between marker and image center (left). The pose of the marker as computed from the overhead camera image (right) yields the exact position of the marker in the space.

the classifier approach yields better results seems odd. At least in a geometric sense, the poses estimated in the marker based approaches seem more meaningful. One has to keep in mind though, that the coordinates of the person in the XIM as indicated by the tracking system are afflicted with the same perspective error as the classifier based pose estimation. Under the condition, that the tracking positions are not geometrically correct, the pose estimated by the classifier serves to compute the best possible solution. A geometrically correct pose on the other hand would require geometrically correct tracking positions to generate a valid output. This leads us to the question, how the performance of the individual poses would be if the tracking positions were not subject to perspective distortion. We tackle this question in a second experiment, described in the following section.

3.5.2. Comparison based on a reprojected marker

In this second experiment we assume that there is no perspective distortion in the tracking data. In this case, the deviations resulting from the marker based pose estimations should converge to a minimum. To simulate distortion free tracking data, markers were placed at different positions in the room. The poses of these markes in relation to the overhead camera and therefore the center of the tracking coordinate frame were estimated as described in section 3.3. The translational part of these poses indicates the geometrically correct position of each marker in the room.

As above, the gazer adjustment corresponding to each of these positions was computed on basis of the poses estimated by the classifier approach, the direct marker computation and the bundle adjustment. As a measure of accuracy, we consider the distance between the reprojection of the marker and the image center in the respective gazer image (Figure 3.18). Also in this experiment, we can only evaluate the correctness of the computed pan angles. For each of the gazers, ten different positions were evaluated. Since this evaluation had to be done offline, once again I can only provide a hypothetical solution. I

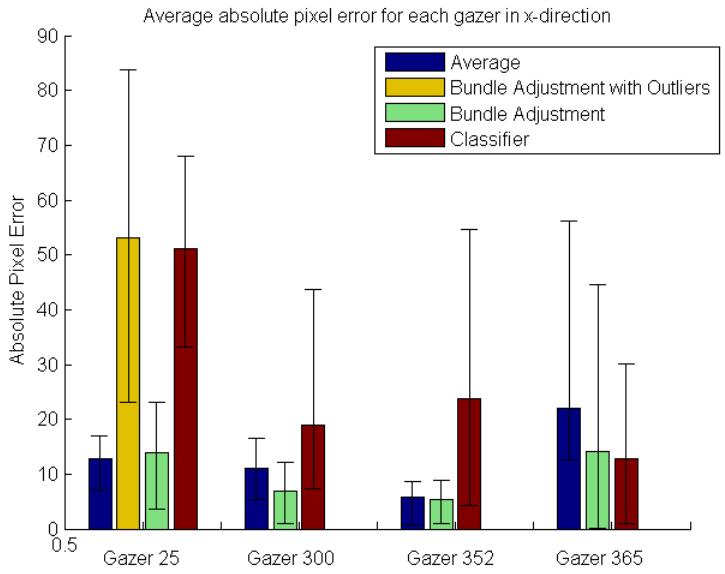


Figure 3.19.: **Average absolute deviations of the marker from the image center for the different calibrations of the gazers.**

thereby recall on images that were recorded for each gazer at different pan angles prior to my departure from Barcelona. For each gazer there are image sets available of ten different marker position. For each set, the space was thereby scanned in pan direction and images were recorded for every second degree. These images now serve as an input for the evaluation of the different approaches. For a computed pan angle pan_i we examine the image corresponding to the closest available pan angle pan'_i. The additional deviation that has to be considered can then be computed as stated in equation 3.18.

Figures 3.20, 3.21, 3.22 and 3.23 illustrate the determined deviations under the different configurations for each gazer. The average absolute deviations are compared in figure 3.19 and listed in table 3.4. Looking at the results we see, that this time the geometrically correct poses yield a much better performance. Except for gazer 365 (and the "malicious" version of gazer 25), clearly better results were achieved with the poses estimated by the marker calibration. The average deviations range from 5.7 to 21.9 pixel for the direct marker computation and 5.4 to 14.0 pixel for the bundle adjustment. In contrast, the average deviations for the classifier approach range up to 51 pixel. High deviations resulted also from the pose estimated by the bundle adjustment for gazer 25, without removing the outlier. Significant is also, that the variations are much lower for the marker based approaches. For gazer 352 for example, the deviations ranged only from 1.0 to 7.7 pixels. One can also see a symmetric deviation to one side for the marker based approaches, while the deviations for the classifier approach spread into both directions. This can be related to the fact, that the classifier estimated the pose based on distorted tracking data and tried to average the error to find an optimal fit for all positions. The deviation in the marker based approaches on the other hand results from a geometric error in the marker tracking and thus affects all

angle computations symmetrically. A more robust marker tracking could help to estimate a gazer's pose more accurately and reduce this error.

Gazer 25	Direct computation	Bundle Adjustment w/ Outliers	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**12.8**	**53.1**	**13.8**	**51.1**
$min \lvert\Delta x\rvert$	7.2	23.1	3.6	33.15
$max \lvert\Delta x\rvert$	17.0	83.8	23.1	68.0

Gazer 300	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**11.1**	**6.8**	**18.8**
$min \lvert\Delta x\rvert$	5.4	1.0	7.2
$max \lvert\Delta x\rvert$	16.6	12.0	43.6

Gazer 352	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**5.7**	**5.4**	**23.7**
$min \lvert\Delta x\rvert$	0.7	1.0	4.2
$max \lvert\Delta x\rvert$	8.7	7.7	54.6

Gazer 365	Direct computation	Bundle Adjustment	Classifier
$\oslash \lvert\Delta x\rvert$	**21.9**	**14.0**	**12.7**
$min \lvert\Delta x\rvert$	12.5	0.1	0.9
$max \lvert\Delta x\rvert$	56.2	44.6	30.1

Table 3.4.: **Average, minimum and maximum deviations in the marker reprojection experiment under the different calibrations.**

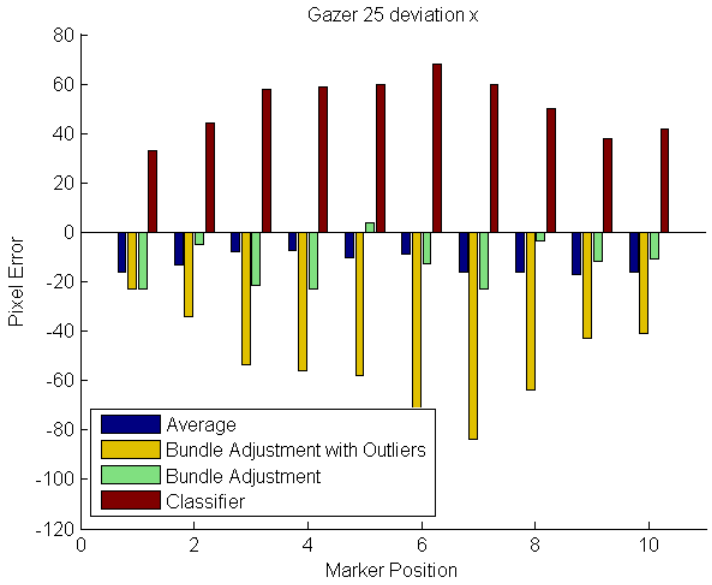


Figure 3.20.: **Accuracy of gazer 25 in the marker reprojection experiment under the different configurations.**

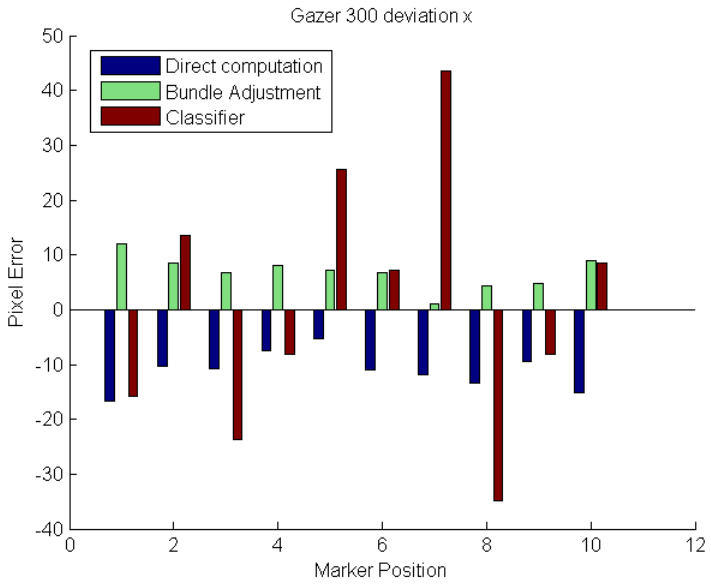


Figure 3.21.: **Accuracy of gazer 300 in the marker reprojection experiment under the different configurations.**

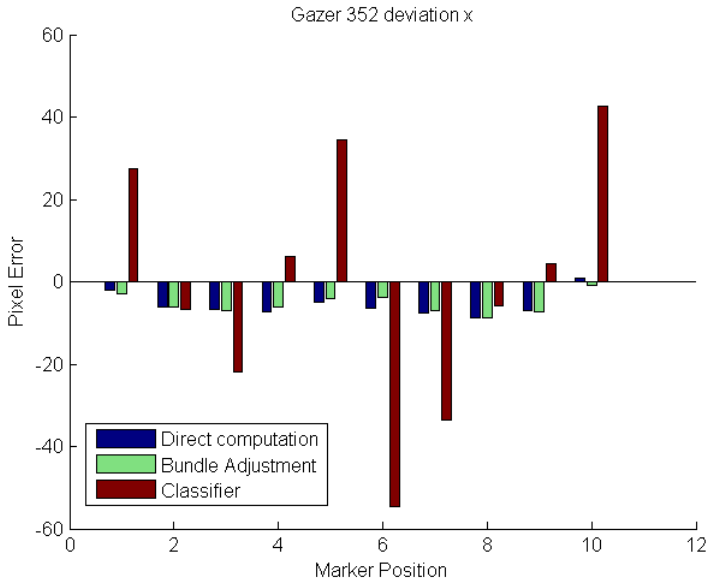


Figure 3.22.: **Accuracy of gazer 352 in the marker reprojection experiment under the different configurations.**

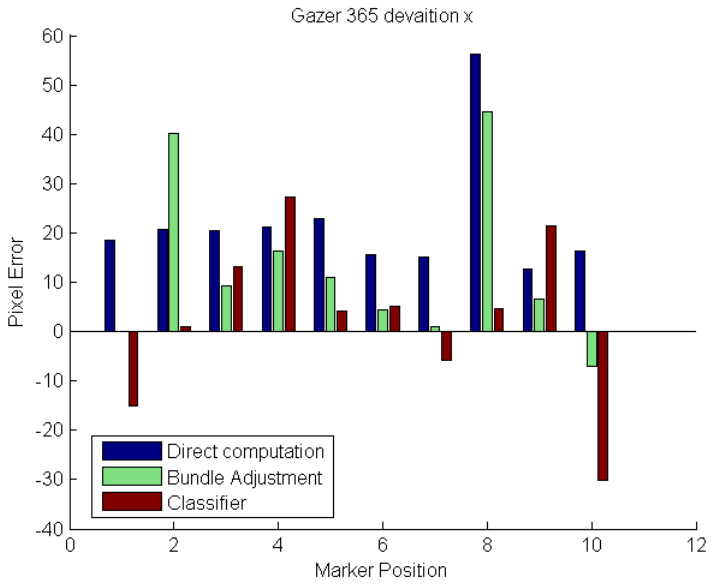


Figure 3.23.: **Accuracy of gazer 365 in the marker reprojection experiment under the different configurations.**

3.6. Discussion

In this chapter I have presented the implementation and results of a marker based pose estimation for the gazers in the XIM. Estimated were both the poses of the overhead infrared camera and the gazers relative to a marker pattern placed in the XIM. Being aware of both these coordinate transformations, we can give a precise estimate of the true position of the gazers in the space. To compensate for numerical instabilities and geometric errors in the marker detection, the estimation was done for several marker positions and a bundle adjustment was used to globally optimize the individual poses.

Motivation for this control experiment was to see, if geometrically correct poses as estimated by a state of the marker calibration, will produce more accurate results in the computation of gazer adjustments then the poses estimated in the classifier approach. To judge on the quality of the individual solutions, the poses were tested under two conditions: Once for tracking data that was afflicted with the same perspective error as the one used for the calibration in the classifier approach, and once for geometrically correct tracking data. The experiments, though carried out on a hypothetical basis, gave a clear indication. The poses estimated by the classifier represent the set of parameters that serves best to compute the angle corresponding to any position in the space, if this position is corrupt in sense of perspective distortion. A geometrically correct pose, in this case leads to significantly wrong results and high deviations of the target in the gazer image. If we would presume a faultless tracking on the other hand, clearly better results can be achieved with a true pose as estimated by a maker calibration.

Consequently, under the presumption that the tracking data is not geometrically representable, the classifier approach is the better way of estimating the gazer poses. On the other hand, if the tracking would not be subject to perspective distortion, it could be of great advantage to estimate the poses precisely, e.g. by a marker based pose estimation as presented in this chapter. In this case, the bundle adjustment proved to be nice way of taking multiple marker positions into consideration and estimating the pose globally in respect of all involved error sources. It showed though, that a robust implementation of the bundle adjustment is a must. Single outliers, that arise from geometric errors in the marker tracking, have to be recognized and discarded. Otherwise the performance may suffer severely.

4. Attribute Extraction

In the course of this thesis I have presented various approaches to estimate a gazer's pose. Knowing this pose, we can set the gazer to look at a specific position in the space. We want to make use of this capability to gain additional information about certain entities present in the XIM. This chapter will describe, how such an attribute extraction could look like and how it could be integrated into the existing software infrastructure. As an example for information that can be gained from an image, I introduce the possibility of generating a hue histogram over a certain image region. Several ways of comparing such histograms will be presented, to decide if the same person can be seen in various images.

4.1. Integration

4.1.1. Objective

Objective of the attribute extraction is to assist the multi modal tracking system (MMT). The MMT fuses the input of multiple sensors to track persons in the XIM and maintain a model of the space in real time. Difficulties occur, when incoming tracking data cannot be unambiguously assigned to a specific object contained in the model. To solve these ambiguities, all entities currently present in the model should be labeled with a set of characteristic attributes. This list can constantly be updated with attributes about each single entity and thus make them distinguishable. The gazers can be used to constantly look at persons in the room and learn about their features. Moreover, in case of doubt, they can be used to look at an object that caused confusing tracking data. In regard of certain attributes, the object in question can then be compared to the ones in the model.

In the previous chapters I have compared different methods of estimating the pose of the gazers, which is necessary to adjust the gazer correctly when ordered to look at a certain spot. In this chapter, I presume that the poses are known and that I can adjust the gazers to look at any spot at any time. The task is now to design a system that interacts with the MMT and delivers reliable information about persons at specific positions. As described above, the gazers shall be used for two purposes. If idle, they shall independently collect information about the individual persons in the room and pass this information on to the MMT. Beyond, if requested, they shall look at specific persons and extract attributes that allow to identify him/her in the model. This leads us to the questions, what concept should be chosen for the implementation and which responsibilities should be transferred to the "Attribute Extractor". One could either implement a pure *pull-application*, that only acts on incoming position information and delivers attributes that relate to even this position. Possible would also be to design a *push-pull-application*, that delivers specified information if requested and tries to collect data independently if idle. When thinking about the proper

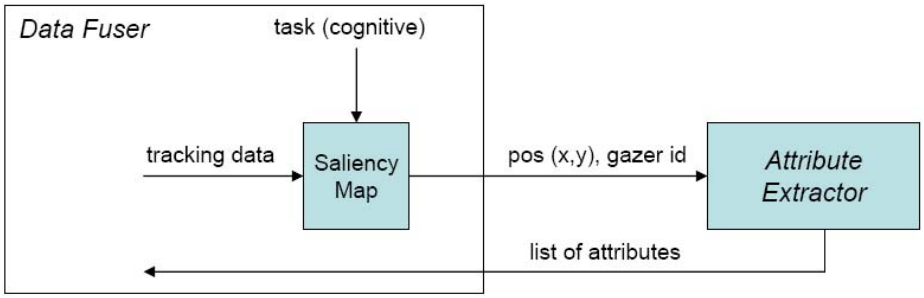


Figure 4.1.: **Interaction between data fuser and attribute extractor.**

solution, we have to consider the information that would have to be provided to the Attribute Extractor in order to fulfill its task. The gazers should only extract information about a person that

a) stands by himself/herself,

b) is not occluded and

c) does not move.

In order to choose points of interest by itself and deliver reliable information, that can be doubtlessly associated with a specific position and consequently a tracked person, the Attribute Extractor would have to be constantly aware of the entire world model. This stands in contrast to the overall design of the multi modal tracking system of the XIM. As described in chapter 1.1.2, the world model is maintained by the data fuser. Even though this data fuser also processes bottom-up sensory information, as e.g from the overhead camera or the floor, these sources require no additional knowledge about the world model. Selective attentional mechanisms on the other hand need to be provided with such knowledge and should be triggered by the data fuser. We therefore agreed to keep the Attribute Extractor as simple as possible and perform action only if requested. The basic interaction between the data fuser and the attribute extractor is drawn in Figure 4.1: The data fuser maintains a *saliency map*, that identifies points of interest from the tracking data and cognitive information about specific entities that require further information. If such a saliency point has been made out, the attribute extractor is triggered with the coordinates of the point of interest and the id of a gazer that has free line of sight. The gained information is then returned to the data fuser for further processing.

4.1.2. Saliency maps

As described in the previous section, the data fuser needs to choose points of interest and command the attribute extractor to gain additional information about the person standing at this point. It thereby has to take several aspects into consideration. If information about a person standing at a specific point is requested, it has to be made sure that this person

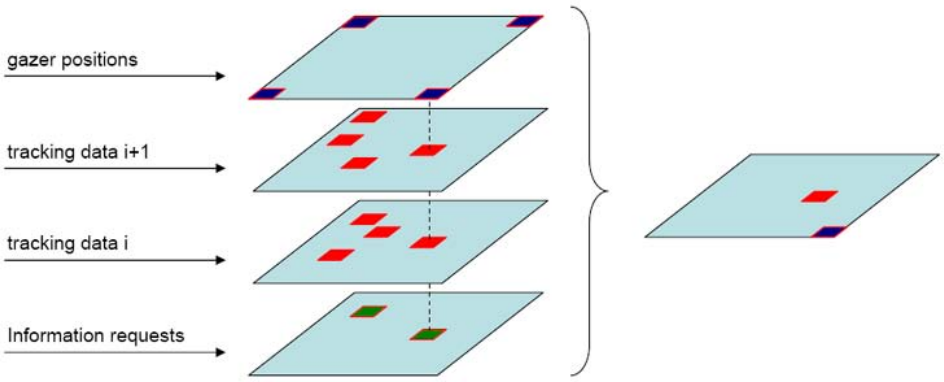


Figure 4.2.: **Creation of a salience map from a set of input sources.**

stands alone and does not move. Also, a gazer has to be chosen that has free line of sight onto the target. As one way to realize this decision tasks, we introduce so called *saliency maps*, topographically arranged maps representing visual saliencies of a corresponding visual scene [34, 22]. Saliency maps are subject of neuroscientific research in perception to handle the problem of information overload. As peripheral sensors generate afferent signals more or less continuously, it would be computationally costly to process all the incoming information all the time. It is therefore important to make decisions which part of the available information should be selected for further processing and which part should be discarded. Furthermore, the selected stimuli need to be prioritized, with the most relevant being processed first and the less important ones later, leading to a sequential treatment of the visual scene. This selection and ordering process is referred to as *selective attention* [35].

We adapt the concept of saliency maps to process all information available and necessary to point out saliency points representing an object that fulfills all the requirements for further examination. Separate maps can be created for all individual sources of information and then be merged to one single map, where a salient point respects all decisive

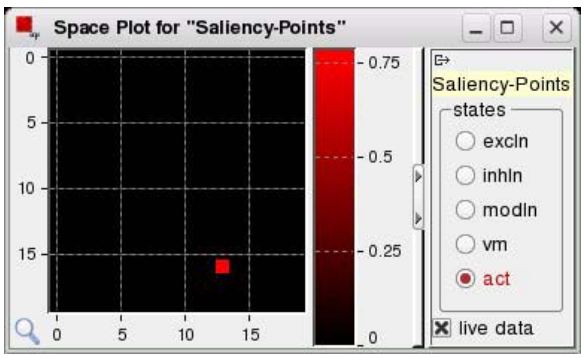


Figure 4.3.: **A saliency map generated by the neural network simulator IQR.**

Figure 4.4.: **Conical representation of the HSV color space (left) and an image along with its h, s and v components (right).**

criterias. This concept is illustrated in Figure 4.2. An object point is only taken over into the saliency map if it occurs at the same spot and separately in all maps. Tracking data (red) can be tacken into consideration over a period of time to decide if the object moves or stands still. Another map can be used to point out objects that need to be examined further (green). While a saliency map could in general be generated in different ways, the merging of the various sensory clues in the XIM is realized by the neural network simulator IQR [13]. Figure 4.3 shows an example for the visual output of the saliency map generated by IQR.

4.2. Hue extraction

So far I have presented, how the Attribute Extractor should interact with the tracking system of the XIM. Given a specific position in the space and the ID of a gazer that has free line of sight onto this position, this gazer should be adjusted correctly and extract certain attributes that allow a unique identification of the person in question. There are various possible attributes that could be used to characterize a person, such as color or height, but also more sophisticates measures that relate certain distinctive features to each other are thinkable. The proper choice of such attributes is not subject of this thesis, but we will consider hue histograms as one possible attribute that could be extracted from the image of a person.

4.2.1. The HSV color space

HSV (Hue, Saturation, Value) is one possible representation of points in an RGB color space, which attempts to describe color relationships more accurately, while remaining computationally simple. Colors in HSV spectrum can be described as points in a conic, whose central axis ranges from black at the bottom to white at the top. The angle around

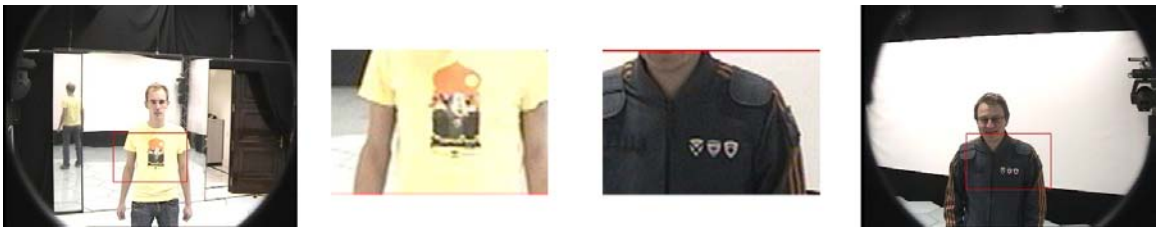


Figure 4.5.: **Particular image regions are compared to see whether two images feature the same person.**

the axis corresponds to the *hue*, the distance from the axis to the *saturation* and the distance along the axis to the *value* of a point. Conceptually, HSV can be seen as an inverted cone of colors, with black at the bottom and fully saturated colors around a circle at the top (Figure 4.4). HSV is thereby a simple transformation of the RGB color spectrum, wherefore each triplet (h, s, v) can be related to a particular color of red, green and blue primaries. In order to compare two color regions, we choose the HSV representation rather then the RGB representation, because a particular hue value represents a color tone in all its saturations and brightnesses, and thus makes it more invariant to lighting conditions.

4.2.2. Histogram comparison

Given two gazer images, we want to check whether they feature the same person. One attribute that would allow a such a comparison, is the color of the person's garment. Since we expect the gazer to be adjusted to look straight at the person's upper body, we can choose a particular region of the image that clips a representative part of his/her torso (Figure 4.5). Naturally, this image region can only be related to the person, if the gazer has been adjusted accurately and the person is really featured in the center of the image. A proper calibration of the gazers is therefore indispensable for the later comparison of objects from the gazer images.

In order to compare the two image regions, we focus on the hue channel of the HSV color space and generate hue histograms for the different samples. Pixels with a similar hue value are therefore grouped into bins, disregarding their saturation or brightness. This makes it possible to compare the images in regard of their hue distribution, while all flexible characteristics of a certain hue value are taken into consideration. Figure 4.6 shows an example of such histograms for two image regions with ten bins each. In order to give a measure of similarity to this image regions, we introduce a few distance metrics that allow the comparison of color histograms.

Distance measures for histogram comparison

The following metrics can be used for the comparison of two histograms at a time. Their choice and systematization follows [45]. In equations 4.1 to 4.6 H and K are the two histograms to be compared. The individual bins are denoted h_i and k_i.

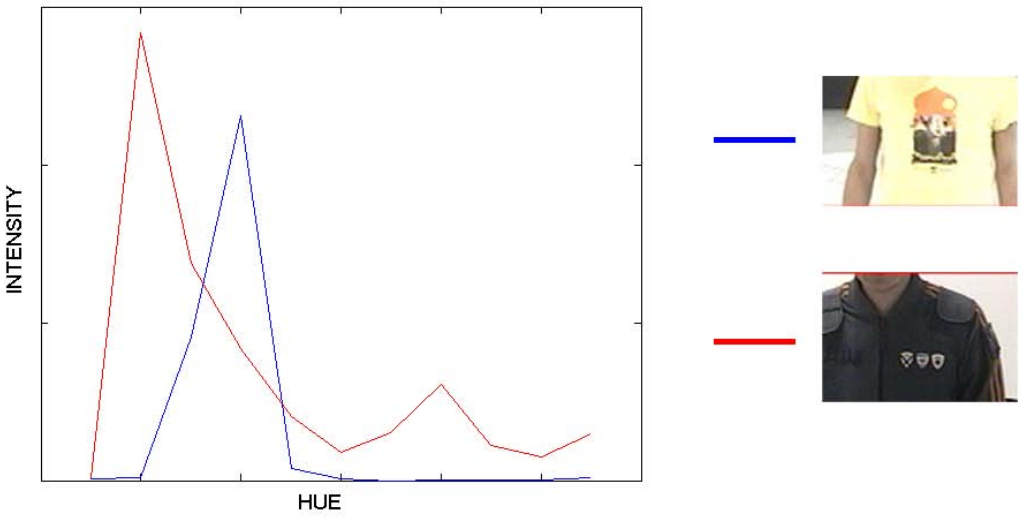


Figure 4.6.: **Hue histograms for two image regions.**

The Minkowski-Distance of Order (MIN_r) is the generalized metric distance and often used for the comparison of color images:

$$d_{L_r}(H,K) = \left(\sum_i |h_i - k_i|^r \right)^{\frac{1}{r}} \tag{4.1}$$

Special cases are the *City Block Distance* (r=1) and the *Euclidean Distance* (r=2).

The Histogram Intersection (HI) compares the common area of two histograms. It is sensitive to partial accordance of two histograms, but vulnerable to light variations.

$$d_{\cap}(H,K) = 1 - \frac{\sum_i min(h_i, k_i)}{\sum_i k_i} \tag{4.2}$$

Founded in statistics is the Chi-Square Distance (χ^2) to give a measure of similarity to two probabiliy distributions:

$$d_{\chi^2}(H,K) = \sum_i \frac{(h_i - m_i)^2}{m_i},\ with\ m_i = \frac{h_i + k_i}{2} \tag{4.3}$$

A further distance measure, that has also been introduced in the context of color based object tracking in multi-camera environments [42], is the Bhattacharya Distance (BD). Similar to the histogram intersection, it gives a measure of similarity for the common area of two histograms:

$$d_B(H,K) = \sum_i \sqrt{h_i \cdot k_i} \tag{4.4}$$

The Kullback-Leibler- (KLD) and Jeffrey-Divergence (JD) derive from information theory. They resemble the computation of an entropy for a distribution. While the Kullback-Leibler-Divergence is no symmetric distance measure, symmetry is given for the Jeffrey-Divergence wich is also more stable to noise:

$$d_{KL}(H,K) = \sum_i h_i \cdot \log \frac{h_i}{k_i} \tag{4.5}$$

$$d_J(H,K) = \sum_i \left(h_i \cdot \log \frac{h_i}{m_i} + k_i \cdot \log \frac{k_i}{m_i} \right), \; with \; m_i = \frac{h_i + k_i}{2} \tag{4.6}$$

All of these metrics compare the histograms on a element wise basis. There are others metrics that work comprehensive, e.g. based on cumulative histograms (Kolmogorov-Smirnov-Distance) or under respect of neighboring elements (Quadratic Distance).

4.3. Experiments and Results

In order to see if a hue histogram carries enough information to distinguish between persons in the XIM, the different distance measures introduced in the previous chapter were applied to hue histograms generated from gazer images. In chapter 2.6.3 the gazers were adjusted to look at different persons in the XIM. I used the images from this experiment as an input. The region of interests to be compared were automatically segmented to show a part of the respective person's shirt. This allows us not only to see which metric serves best to compare the histograms, but also if the gazers were adjusted sufficiently to allow a computation.

We compare a total of 160 images to a model, whereby half of them feature a person with the same t-shirt color and the other half persons wearing different colors. As a result we get a vector with 160 elements for each metric, that is made up by the individual distances for each image to the model. Compared were histograms with ten bins each. Figure 4.7 shows the frequency distributions of the computed distances for each of the introduced metrics: City Block and Euclidean Distance, Histogram Intersection, χ^2-distance, Bhattacharya Distance and Kullback-Leibler- and Jeffrey-Divergence. The results of the comparison of two image regions showing the same t-shirt color are draw as a blue curve (positive samples), the ones of the comparison of different t-shirts as a red curve (negative samples). One can see that the distributions overlap in all cases. It is thus not possible to define a simple threshold that allows a definite decision on whether two images show the same person. To draw a qualitative conclusion about the goodness of the individual metrics, one can examine the respective overlap. It is appropriate to use the intersection of the two frequency distributions as a threshold. The distance measure d_λ associated with this intersection occurs with the same frequency when comparing images of the same shirt color and images of different shirt color. We then get a false acceptance rate (FAR) as the number of negative samples that are falsely declared to be equal, if $d(H,K) \leq d_\lambda$. The false rejection rate (FRR) on the other hand states the amount of positive samples for which a distance greater than d_λ was computed. A different choice of the threshold d_λ would yield a different FAR and FRR. In general, if one of the error rates decreases, the other one increases, as long as the distributions overlap.

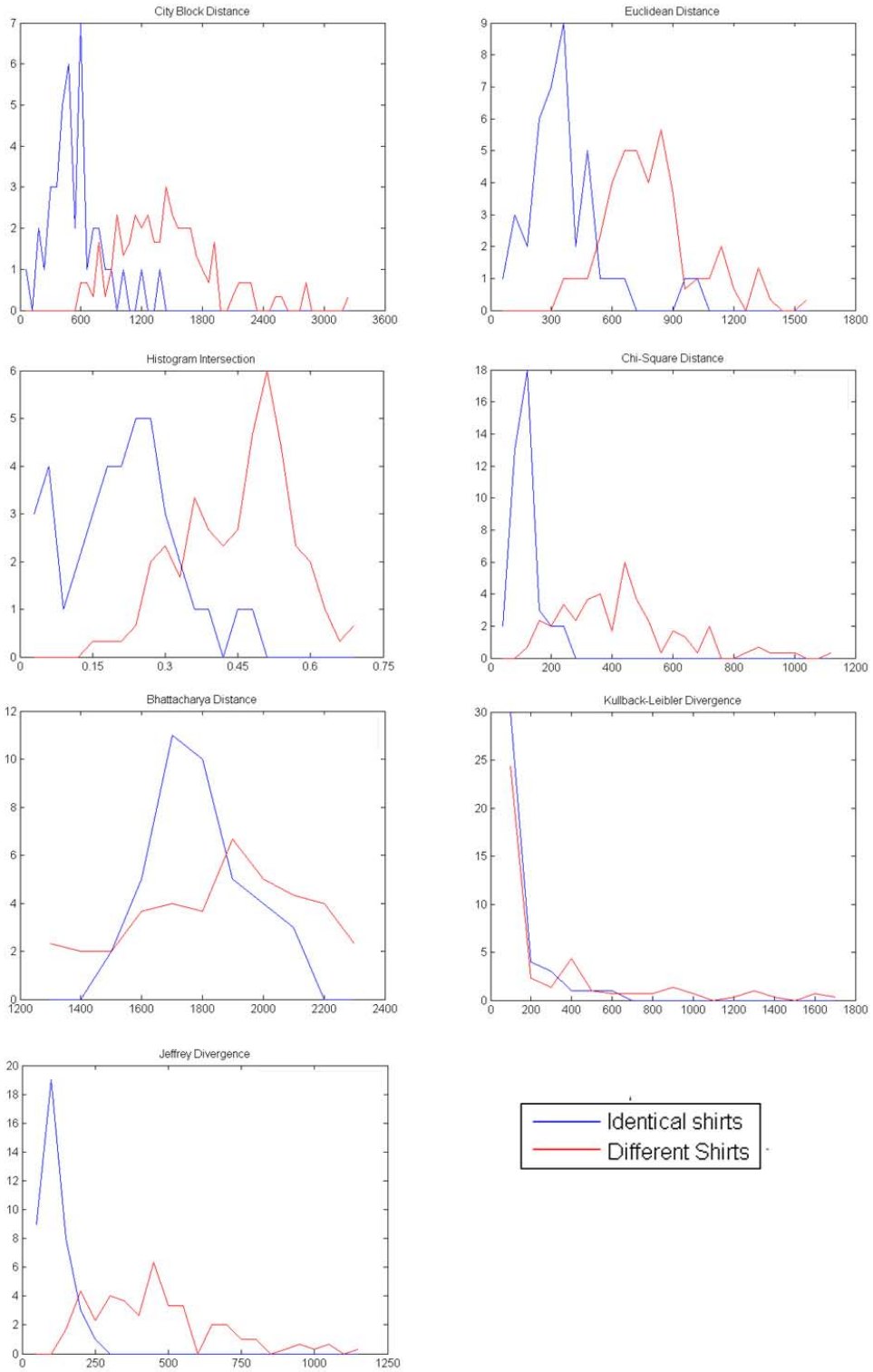


Figure 4.7.: **Frequency distributions of the similarity measures with the different metrics.**

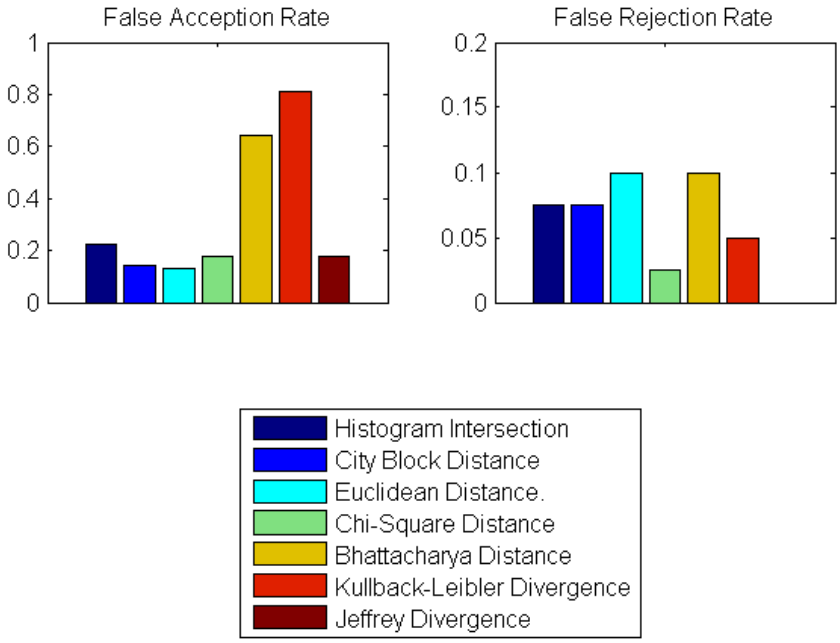


Figure 4.8.: **FAR and FRR for the comparison of 160 samples.**

The FAR and FRR for the comparison experiment are shown in Figure 4.8. The lowest FRR was achieved with the Jeffrey Divergence, for which not a single positive sample was falsely rejected. Also the χ^2-Distance yielded a low FRR, with 2.5 % positive samples declared negative. The smallest FAR resulted from the Euclidean Distance metric with 13 % falsely accepted negative samples. It had however a significantly higher FRR than the other metrics. All FRRs lie below a limit of 10 %, while the FARs vary strongly for the different metrics. Inadequate proved to be the Bhattacharya Distance and the Kullback-Leibler Divergence, with FARs of above 60 %. FAR and FRR are expeced values for the errors in the comparision of two histograms, if the decision is made based on a simple threshold. In the context of the shirt-comparision, this means that from a set of samples those are declared to be the same as a model, if they have a lower distance measure than d_T. The FRR is the expected value of those samples with the same color as the model, but rejected nevertheless. The FAR on the other hand, is the expected value that a different shirt is falsely declared to be the same as the model.

Even though this has only been a first experiment, it has been shown that a comparison of two image regions based on their hue histograms is possible. By choice of the right metric, false accepts could be reduced to 13 % and false rejects always ranged below 10 %. One would have to see, if better results could be achieved if a different number of bins would be used or the threshold would be adapted. Significant is, that these result were achieved with images taken by the gazers after automatic adjustment as described in this thesis. The image regions that were compared were choosen automatically from these images without manual intervention. The gazers, calibrated by the classifier approach, thereby provided suitable images for the comparison of two persons in the XIM.

5. Findings and Conclusions

In the beginning of this thesis we have introduced the mixed reality space XIM and the multi modal tracking system deployed to keep track of its visitors. Main objective of this project has been the use of four movable pan and tilt cameras, the gazers, to assist the tracking system by gaining complementary information about certain entities in case of unambiguities. The gazers therefore need to be set to look at specific positions in the room as indicated by the tracking system. This is only possible, if each gazer's position and orientation in space, its pose, is known. The accurate estimation of this poses must therefore be declared indispensable for a successful deployment of the gazers. Within the course of this thesis we have presented two different ways of estimating a gazer's pose and evaluated which one serves best for application in the XIM. In this context, best not only refers to the most usable approach, but also to the one which yields better results in computing the angles corresponding to any arbitrary position in the space.

5.1. Validity of the different approaches

In order to estimate the poses of the gazers in the XIM, two essentially different approaches have been introduced. On the one hand, the pose was estimated as the optimal fit to a set of valid correspondences between tracking positions and gazer adjustments. On the other hand, a state of the art marker based pose estimation was used to determine the desired parameters. Both approaches showed strengths and weaknesses, depending on different constraints.

First I introduced an innovative approach of optimizing the pose to serve as an optimal fit to a set of tracking data and gazer angle correspondences. The correspondences were thereby determined in a prior calibration scenario, in which the space was scanned with the gazers for a person standing at tracked positions. To detect the person in the gazer image, I proposed the use of a classifier trained on human upper bodies. Main motivation for this approach was the complete independence of any other modality other than a global tracking signal and the gazer's own functionality. The performance of the calibration does not depend on any coordinate transformations and yields the pose in the coordinate frame of the tracking signal received as an input. This is of an enormous advantage, since information requests are given in even this coordinate frame later on. Furthermore, there are no limitations on the tracking source and basically any positioning input can be used. The application thus becomes mobile and reproduceable in any context. Limitations occur through the classifier used for person detection. Under certain environmental constraints the person can not be detected or is falsely classified. These false alarms would however significantly affect the quality of the pose estimation, since all found correspondences are considered by the optimizer. For application in the XIM it was possible to fine tune the

classifier so that there were no false detections, but in different environments problems may arise and would have to be solved accordingly.

In the particular case of the XIM, the tracking signal is subject to severe perspective distortion. The resulting poses are therefore wrong in a geometrical sense, as they do not mirror the gazers real positions and orientation in space. Nevertheless the optimizer finds the set of parameters, that serves best to compute the angle corresponding to any arbitrary position as pointed out by the tracking signal. This proved to be another major advantage, as the tracking data coming in for online angle computation can be expected to be afflicted with the same error.

In a first test scenario, the poses estimated by the classifier approach were used to compute the angles corresponding to 50 target positions. In the perfect case, these angles should adjust the gazer to look straight at the respective target position. For this experiment, average deviations of 12 to 17 pixels were measured in the resulting images. This seems to be a reasonable basis for further processing. Still there were some positions for which the computation yielded tremendous deviations.

In order to see whether better result can be achieved with a state of the art calibration technique that determines the poses geometrically correct, a control experiment was run, implementing a marker based pose estimation. In order to permit a comparison of the results, the poses of the gazers thereby had to be estimated relative to the same coordinate frame as in the classifier approach. This premise made the setup of this calibration rather circuitous, since two individual poses had two be estimated and concatenated to yield the desired coordinate transformation: The pose of the gazer relative to the marker and the pose of the overhead infrared camera relative to the marker. The latter estimation required the application of an interactive marker equipped with infrared LEDs, so that it could be seen by the infrared camera. To compensate for numerical instabilities and geometric errors in the marker detection, the estimation was done for several marker positions and a bundle adjustment was used to globally optimize the poses of the individual gazers, taking all involved parameters and error sources into consideration.

In order to judge on the quality of the individual solutions, the poses were tested and compared under two conditions: Once for tracking data that was afflicted with the same perspective error as the one used for calibration in the classifier approach, and once for geometrically correct tracking data. The experiments were carried out offline and thus only allowed a hypothetical solution. Nevertheless, the evaluation gave a clear indication. The poses estimated by the classifier approach represent the set of parameters that serves best to compute the angles corresponding to any position in the space, if this position is corrupt in sense of perspective distortion. A geometrically correct pose in this case leads to significantly wrong results and high deviations in the gazer image. If we would presume a faultless tracking on the other hand, clearly better results can be achieved with a true pose as estimated by a marker calibration. An experiment with a reprojected marker, for which the position in the room could be tracked free of any distortion, showed, that the deviations in the final gazer image can be reduced to a minimum.

Recapitulating, one can draw a conclusion on the goodness of the individual approaches. Presuming that the tracking data is not geometrically representable, as it is the case for the current tracking in the XIM, the classifier approach is the better way of estimating the gazer poses. If, however, the tracking would not be subject to perspective distortion, it could be a great advantage to estimate the poses precisely. One method to do so, a marker based pose estimation, has been introduced in this thesis. The bundle adjustment thereby proved to be a fancy way of taking multiple marker positions into consideration and estimating the pose globally in respect of all involved error sources.

In regard of the main objective of this thesis, the comparison of different persons in the space by means of specific attributes, I have presented the approach of extracting and comparing hue histograms generated from the gazer images. In an experiment with multiple persons wearing shirts of different color, diverse distance measures for histogram comparision were tested out. The images were thereby taken by the gazers after automatic adjustment, based on poses estimated in the classifier approach. By choice of the right metric, it was possible to reduce the rate of image pairs falsely declared to feature the same person to 13 %. At the same time, less then 10 % of matching pairs were falsely rejected. This rates can surely be reduced, by modification of the histogram comparison as well as by a more accurate calibration. Nevertheless, this experiment was a first proof, that the comparison of two persons in the XIM by use of the gazers is possible.

5.2. Areas for further research

On the way to a successful calibration of the gazers, the most dominant error source proved to be the perspective distortion of the XIM tracking. It has been shown, that the deviation in the gazer images after an adjustment can be reduced to a minimum, if tracking positions and gazer poses are geometrically justly. Best results could thus be achieved, if one would guarantee an undistorted tracking signal. The challenge would then be to estimate the gazer poses as precise as possible. In general, the marker calibration proved to be a valid approach to do so. Still there is room for improvement in various issues. A larger marker pattern and different arrangements of the marker relative to the individual cameras might help to reduce the measurement errors. A particular problem with the setup of the marker estimation in this thesis has been the disadvantageous slant angle of the overhead camera relative to the marker, as the camera looked almost straight onto the pattern. Ideally, both cameras should look at the marker from a 45° angle.

The separate estimation of the individual pose transformation and their concatenation give rise to an increased number of errors. The bundle adjustment is a good way to estimate the final pose in the right coordinate frame in respect of all error sources, numerical as well as geometric ones. Problems occurred however through outliers, single markers that carry a high geometric error. A more robust implementation of the bundle adjustment must therefore recognize these outliers during optimization and discard them. Rather than reading priorly made measurement data, the bundle adjustment could also be implemented to do the optimization online while a marker is moved to different positions in the space.

This thesis has been realized under the presumption, that no qualitative statement about the tracking signal can be made. The gazer poses should be estimated from any arbitrary positioning signal. One can thus not rely on the geometrical correctness of the tracking signal, nor give an approximation of the error that has to be considered. Under this condition, a promising way of estimating the gazer poses was introduced with the classifier approach. A major drawback of this technique were the weaknesses of the classifier in detecting the person in the gazer image. So far, all experiments have been conducted in the same context (the XIM) and under optimal conditions. It was thereby possible to avoid problems by adapting hard- and software setup accordingly. In order to assure applicability in other environments as well, the classifier has to be made more stable and ubiquitous. It could e.g. be advantageous to train the classifier cascade manually over a large set of images taken by the gazers, rather than calling back on the predefined Upper Body Cascade included in OpenCV. Other measures that could be considered include an automatic detection of clearly unfeasible correspondences. This way outliers could be disregarded from the very beginning. In order to draw a representable conclusion on the universal usability of the classifier approach, more experiments have to be conducted under different conditions.

Within this thesis, a strong emphasis was put on the calibration of the gazers. Regarding the original motivation of using the gazers to distinguish between persons in the space, a first approach of winning meaningful attributes from the gazer images has been presented. It has been shown, that hue histograms provide a good foundation for the comparison of two image regions in respect of their color distribution. Still the experiments and results described in this thesis can only be seen as a first hint on the goodness of the technique. All samples were compared to one single model. The drawn conclusions can thus not be considered universally valid. A more meaningful evaluation would have to be done, whereby all samples are compared among each other. Also, it has to be evaluated if better results could be achieved if the individual histograms would be divided into a different number of bins. Additional preprocessing of the images, such as histogram normalization, might also help to improve the comparison results.

With the hue histograms, only one possible attribute information that may be gained by the gazers has been considered so far. There is a number of other possible attributes, that could be assigned to an object in the model and thus make it distinguishable. One could for example use the gazers to gain further attributes that describe the physical appearance of a person, such as height, girth or remarkable features. A variety of such attributes would allow the generation of a detailed model of any person in the XIM, making him or her a unique individual in the world model. In this concern, future research has to be done on choosing and exploiting appropriate attributes and building symbolical models.

Appendix

A. List of Abbreviations

General Context

AnTS	The visual tracking system deployed in the XIM
SRG	Spatial Relationship Graph
PVC	The Persistent Virtual Community
MMT	The multi modal tracking system
SPECS	Laboratory for Synthetic Perceptive, Emotive and Cognitive Systems
XIM	The eXperience Induction Machine

Pose Estimation & Parameter Optimization

DLT	Direct Linear Transformation
GNA	The Gauss-Newton Algorithm
LMA	The Levenberg-Marquardt Algorithm
SAT	Summed Area Table
SSD	Sum of Squared Differences
SVD	Singular Value Decomposition

Histogram Comparison

BA	Bhattacharya Distance
FAR	False Acceptance Rate
FRR	False Rejection Rate
HI	Histogram Intersection
HSV	The Hue Saturation Value Color Space
JD	Jeffrey Divergence
KLD	Kullback-Leibler Divergence
MIN	Minkowski Distance of Order

Implementation & Technology

CCD	Charge-Coupled Device
DMX	Direct Multiplex Protocol
GUI	Graphical User Interface
IR	Infrared
LED	Light Emitting Diode
OpenCV	Intel's Open Source Computer Vision Library
TNT	Template Numerical Toolkit

Bibliography

[1] *OpenCV: Intel Open Source Computer Vision Library. The software library for computer vision. http://www.intel.com/research/mrl/research/opencv*.

[2] *TNT: The Template Numerical Toolkit. An interface for scientific computing in C++. Roldan Pozo, Mathematical and Computational Sciences Division, National Institute of Standards and Technology (NIST). http://math.nist.gov/tnt/*.

[3] J Aloimonos, I Weiss, and A Bandyopadhyay. Active vision. In *Proceedings 1st International Conference on Computer Vision*, pages 35–54, 1987.

[4] T J Anastasio, P E Patton, and K Belkacem-Boussaid. Using bayes' rules to model multisensory enhancement in the superior colliculus. *Neural Computation*, 12:1165–1187, 2000.

[5] M Armstrong, A Zisserman, and R Hartley. Self-calibration from image triplets. In *Lecture Notes in Computer Vision, Vol. 1064; Proceedings of the 4th European Conference on Computer Vision - Volume 1*, pages 3–16. Springer-Verlag, London, UK, 1996.

[6] K B Atkinson. *Close Range Photogrammetry and Machine Vision.* Whittles Publishing, Roseleigh House, Latheronwheel, Caithness, Scotland, 1996.

[7] Y Bar-Shalom. *Tracking and Data Association*. Academic Press Professional Inc., San Diego, CA, USA, 1987.

[8] A Bartoli and P Sturm. The 3-d line motion matrix and alignment of line reconstructions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 287–292, 2001.

[9] Martin Bauer, Michael Schlegel, Daniel Pustka, Nassir Navab, and Gudrun Klinker. Predicting and estimating the accuracy of vision-based optical tracking systems. In *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR'06)*, Santa Barbara (CA), USA, October 2006.

[10] S Bermúdez i Badia, U Bernadet, M Negrello, M Knaden, and P F M J Verschure. *AnTS: A 3-Dimensional Tracking System for Behavioral Analysis of Flying Insects and Robots*. Institute of Neuroinformatics, ETH/University of Zurich, 2005.

[11] U Bernadet, S Bermúdez i Badia, and P F M J Verschure. *Deliverable DA1.1: Upgrade Ada hardware, software and operating system and bring to operational condition*. PRESENCCIA, 2006.

[12] U Bernadet, S Bermúdez i Badia, and P F M J Verschure. *The eXperience Induction Machine and its Role in the Research on Presence*. Presence, 2007.

[13] U Bernadet, M Blanchard, R Wyss, and P F M J Verschure. *IQR: A simulator for large scale neural networks. http://iqr.sourceforge.net/.*

[14] Andrew Blake and Alan Yuille. *Active Vision*. The MIT Press, Cambridge, Massachusetts, 1992.

[15] H Borotschnig, L Paletta, M Prantl, and A Pinz. Appearance-based active object recognition. *Image and Vision Computing*, 18(9):715–727, June 2000.

[16] G Bradski, A Kaehler, and V Pisarevsky. Learning-based computer vision with intel's open source computer vision library. *Intel Technology Journal*, 09, 2005.

[17] M Bricken. *Virtual Worlds: No Interface to Design*. University of Washington, Human Interface Technology Laboratory, 1991.

[18] D C Brown. The bundle adjustment - progress and prospects. *Int. Archives Photogrammetry*, 21(3), 1976.

[19] H Chen. Pose determination from line to plane correspondences: Existence of solutions and closedform solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6):530–541, 1991.

[20] M A R Cooper and P A Cross. Statistical concepts and their application in photogrammetry and surveying. *Photogrammetric Record*, 12(71):637–663, 1988.

[21] M A R Cooper and P A Cross. Statistical concepts and their application in photogrammetry and surveying (continued). *Photogrammetric Record*, 13(77):645–678, 1991.

[22] N Courty and E Marchand. Visual perception based on salient features. In *Proceedings of 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems. Las Vegas, Nevada*, 2003.

[23] Trevor Darrel, Baback Moghaddam, and Alex P Pentland. Active face tracking and pose estimation in an interactive room. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'96)*, pages 67–72, 1996.

[24] L Davis, E Clarkson, and J P Rolland. Predicting accuracy in pose estimation for marker-based tracking. In *Mixed and Augmented Reality, 2003. Proceedings. The Second IEEE and ACM International Symposium on*, pages 28 – 35, October 2003.

[25] T Delbruck, A M Whatley, R J Douglas, K Eng, K Hepp, and P F M J Verschure. A tactile luminous floor for an interactive autonomous space. *Robotics and Autonomous Systems*, 55:433–443, 2007.

[26] F Echtler, M Huber, D Pustka, and G Klinker. Research issue: Ubiquitous tracking - ubitrack. http://campar.in.tum.de/Chair/ResearchIssueUbiTrack.

[27] K Eng, D Klein, A Baebler, U Bernadet, M Blanchard, M Costa, T Delbruck, R J Douglas, K Hepp, J Manzolli, M Mintz, F Roth, U Rutishauser, K Wassermann, A M Whatley, A Wittmann, R Wyss, and P F M J Verschure. Design for a brain revisited: The neuromorphic design and functionality of the interactive space ada. *Reviews in the Neurosciences*, 14:145–180, 2003.

[28] B P Flannery and W H Press. *Numerical Recipies in C - The Art of Scientific Computing*. Cambridge University Press, Cambridge, New York, Oakleigh, 2nd edition, 1992.

[29] Y Freund and R E Schapire. Experiments with a new boosting algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*, pages 148–156, 1996.

[30] S Granshaw. Bundle adjustment methods in engineering photogrammetry. *Photogrammetric Record*, 10(56):181–207, 1980.

[31] R M Haralick and L G Shapiro. *Computer and Robot Vision*, volume 1. Addison Wesley Publishing, 1992.

[32] R Hartley and A Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, New York, Melbourne, 2nd edition, 2003.

[33] C Heeter. Being there: The subjective experience of presence. *Presence: Teleoperators and Virtual Environments. MIT Press, Cambridge, MA, USA*, 1:267–271, 1992.

[34] L Itti, C Koch, and E Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[35] C Koch and S Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[36] A Kuranov, R Lienhard, and V Pisarevsky. An empirical analysis of boosting algorithms for rapid objects with an extended set of haar-like features. Technical Report MRL-TR-July02-01, Intel Technical Report, 2002.

[37] R Lienhard and J Maydt. An extended set of haar-like features for rapid object detection. In *ICIP*, 2002.

[38] Y Liu, T S Huang, and O D Faugeras. Determination of camera location from 2-d to 3-d line and point. *IEEE Transactions on Pattern analysis and Machine Intelligence*, 12(1):28–37, 1990.

[39] Z Mathews, S Bermúdez i Badia, and P F M J Verschure. Active attention for person tracking in mixed reality environments. 2007.

[40] T Morris and M Donath. Using a maximum error statistic to evaluate measurement errors in 3d position and orientation tracking systems. *Presence: Teleoperators and Virtual Environments*, 2(4):314–343, 1993.

[41] N Navab and O D Faugeras. Monocular pose determination from lines: Critical sets and maximum number of solutions. *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 254–260, 1993.

[42] K Nummiaro, E Koller-Meier, T Svoboda, D Roth, and L Van Gool. Color-based object tracking in multi-camera environments. In *DAGM Symposium, Magdeburg, Deutschland*, 2003.

[43] W Piekarski, B Avery, B H Thomas, and P Malbezin. Integrated head and hand tracking for indoor and outdoor augmented reality. In *Proceedings in Virtual Reality*, volume 27, pages 11–276. IEEE, March 2004.

[44] Daniel Pustka, Manuel Huber, Martin Bauer, and Gudrun Klinker. Spatial relationship patterns: Elements of reusable tracking and calibration systems. In *Proc. IEEE International Symposium on Mixed and Augmented Reality (ISMAR'06)*, October 2006.

[45] Y Rubner, C Tomasi, and L J Guibas. A metric for distributions with applications to image databases. In *IEEE International Conference on Computer Vision, Mumbai, India*, 1998.

[46] A Selinger and R C Nelson. Appearance-based object recognition using multiple views. *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01)*, 1, 2001.

[47] Y C Shin and S Ahmad. 3-d location of circular and spherical features by monocular modelbased vision. *Proceedings of the IEEE International Conference on System, Man and Cybernetic*, pages 576–581, 1989.

[48] Y C Shiu and C Huang. Pose determination of circular cylinders using elliptical and side projections. *Proceedings of the International Conference on Systems Engineering*, pages 265–268, 1991.

[49] C C Slama. *Manual of Photogrammetry*. American Society of Photogrammetry and Remote Sensing, Falls Church, Virginia, USA, 1980.

[50] De Ma Song. A self-calibration technique for active vision systems. *IEEE Transactions on Robotics and Automation*, 1:114–120, 1996.

[51] B A Stein and M A Meredith. *The Merging of the Senses*. MIT Press, Cambridge, MA, 1993.

[52] P Sturm. Critical motion sequences for monocular selfcalibration and uncalibrated euclidean reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, Puerto Rico*, 1997.

[53] Bill Triggs, P. McLauchlan, Richard Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *Lecture Notes in Computer Science*, pages 298–372. Springer-Verlag, 2000.

[54] W Triggs. Camera pose and calibration from 4 or 5 known 3-d points. *Proceedings of the IEEE International Conference on Computer Vision*, 1:278–284, 1999.

[55] R Y Tsaï. A versatile camera calibration technique for highaccuracy 3-d machine vision metrology using off the shelf tv cameras and lenses. *IEEE Journal of Robotics and Automation*, 3(4):323–344, 1987.

[56] P Viola and M Jones. Rapid object detection using a boosted cascade of simple features. *IEEE CVPR*, 2001.

[57] S Vogt, A Khamene, F Sauer, and H Niemann. Single camera tracking of marker clusters: Multiparameter cluster optimization and experimental verification. In *Proceedings of the InternationalSymosium on MIxed and Augmented Reality (ISMAR '02)*, pages 127–136. IEEE and ACM, September 2002.

[58] C Wang, H Tanahashi, Y Sato, H Hirayu, Y Niwa, and K Yamamoto. Location and pose estimation for active vision using edge histograms of omni-directional images. In *TENCON '02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, volume 1, pages 578–581, 2002.

[59] P R Wolf and C D Ghilani. *Adjustment Computations: Statistics and Least Squares in Surveying and GIS.* John Wiley & Sons, 1997.

[60] Z Zhang, G Potamianos, A Senior, S Chu, and T S Huang. A joint system for person tracking and face detection. In *Computer Vision in Human-Computer Interaction, ICCV 2005 Workshop on HCI*, 2005.

[61] L Zimmerli, A Duff, A Mura, K Eng, S Bermúdez i Badia, U Bernadet, Z Mathews, and P F M J Verschure. Communication and interaction in the persistent mixed reality environment p-club. In *The enhancement of multilingual communication and learning through technology*, November 2007, Ascona, Switzerland.