Pinhole Camera Calibration in the Presence of Human Noise

Magnus Axholt
Abstract

The research work presented in this thesis is concerned with the analysis of the human body as a calibration platform for estimation of a pinhole camera model used in Augmented Reality environments mediated through Optical See-Through Head-Mounted Display. Since the quality of the calibration ultimately depends on a subject's ability to construct visual alignments, the research effort is initially centered around user studies investigating human-induced noise, such as postural sway and head aiming precision. Knowledge about subject behavior is then applied to a sensitivity analysis in which simulations are used to determine the impact of user noise on camera parameter estimation.

Quantitative evaluation of the calibration procedure is challenging since the current state of the technology does not permit access to the user's view and measurements in the image plane as seen by the user. In an attempt to circumvent this problem, researchers have previously placed a camera in the eye socket of a mannequin, and performed both calibration and evaluation using the auxiliary signal from the camera. However, such a method does not reflect the impact of human noise during the calibration stage, and the calibration is not transferable to a human as the eyepoint of the mannequin and the intended user may not coincide. The experiments performed in this thesis use human subjects for all stages of calibration and evaluation. Moreover, some of the measurable camera parameters are verified with an external reference, addressing not only calibration precision, but also accuracy.
Acknowledgments

As this journey is finally coming to its end, and what a journey it has been - intellectually as well as geographically, I would like to take this moment to express gratitude to the friends and colleagues who have accompanied me along the way.

“Tempus fugit!” he wrote on the whiteboard as we worked on our first experiment - and indeed it does! Seven years has flown by far too quickly, but in the company of Stephen Ellis time has always been well spent. Thank you for the opportunity to work with you in the Advanced Displays and Spatial Perception Laboratory at NASA Ames Research Center in Mountain View, California. I found every day in the lab to be a tremendously rewarding experience. Generously sharing knowledge and experience you are the inspiring mentor I wish every PhD student could have. I am indebted to you in ways I can only reciprocate by paying it forward.

I am also very grateful for the continuous support of my thesis advisors Matthew Cooper and Anders Ynnerman at the division for Media and Information Technology, Linköping University. During the years you have provided advice in all matters, big and small, and helped me pave the way to achieve my goals. Thank you for the flexible ways in which you have allowed me to conduct my research.

My journey into science started at Eurocontrol Experimental Centre in Brétigny-sur-Orge, France. Here, I met Vu Duong, who welcomed me into his lab and offered me the possibility to pursue a doctorate degree. I thank you for this opportunity and want you to know that your lectures on the scientific method have been very useful to me. I also owe gratitude to Marc Bourgois whose pragmatic approach enabled the collaboration with Stephen Ellis.

A travelling companion who deserves special mention is my colleague and dear friend, Dr. Stephen O’Connell. I clearly remember the day we decided to embark on this journey together. The inevitable times of hard work, late hours, and rejection are all forgotten now. What prevails are the memories of happiness, sense of achievement, and the excitement felt when data made sense. It has been a true pleasure to share this adventure with you. What’s next?

In my various workplaces I have collaborated with people whose support I would also like to acknowledge. At Eurocontrol, Anna Wennerberg and Peter Eriksen were most helpful in facilitating research applied to the airport tower. Raymond Dowdall
possesses great knowledge of everything relating to aviation and patiently explained some of it to me. Monica Tavanti taught me the foundations of experimental design. Horst Hering provided lab space. I also appreciate the company and support of my fellow PhD students Ella Pinska-Chauvin, Konrad Hofbauer, Ronish Joyekurun, Antonia Thao (Cokasova), Peter Choroba, Claus Gwiggner, Sonja Straussberger, Simone Rozzi, and Nguyen-Thong Dang. At San José State University I would like to thank Kevin Jordan for enabling parts of the research at NASA Ames and for lending lab equipment to Sweden. I really appreciate the personal interest you took in my research when you occasionally checked in to see how things were going. The exciting time at NASA Ames was a great learning experience due to inspiring colleagues such as Bernard Dov Adelstein, Jeffrey Mulligan, Martine Godfroy-Cooper, and Charles Neveu. I wish lunch discussions were like that all the time! Mark Anderson deserves separate mention for offering tools, help, and great coffee. At Linköping University I would like to thank my colleague Martin Skoglund for introducing a more profound insight into optimization and problem parametrization. I also appreciate your tremendous patience during the Sisyphean task of pilot testing calibration procedures. In moments of intellectual standstill discussions with Per-Erik Forssén, Klas Nordberg, Stefan Gustavson, Joel Kronander, Alexander Fridlund and Miroslav Andel have triggered new steps forward. Thanks to Jonas Unger and Per Larsson for sharing the tools and equipment of the High Dynamic Range Video lab for sharing tools and equipment with me. I also recognize the help of Andreas Lindemark and Eva Skärblom for all practical matters relating to my thesis.

Along the road there have also been people who, in one way or another, have indirectly contributed to the completion of this thesis. I want to thank the extraordinary Forst family who made San Francisco feel like home. In Oakland, the Greek generosity of the Panos-Ellis family knowns no boundaries. I am also still wondering how I shall ever be able to repay the Pettersson-O’Connell family for all the times I ate and slept in their home in Stockholm. Moreover, I want to thank Ana-Gabriela Acosta Cabeda for her countless invaluable advice, Johan Bauhn for sharing life’s ups and downs, and Aurélien Sauty for teaching me how to understand everything French. Diana Muñoz, Philipp Schmidt, and Emmanuelle Bousquet are also people I have reason to thank.

Of course, this thesis would not have been possible without the support of my family. I thank Andreas Axholt and Karin Söderström for their continuous words of encouragement, and my mother and father for their confidence, patience and support.

***

The main part of this thesis was funded through a PhD scholarship from Eurocontrol. Additional funding was provided by Linköping University and the division for Media and Information Technology. The visit and experiments at NASA Ames were also funded in part through the NASA Grant NNA 06 CB28A to the San José State University Research Foundation.
# Contents

I Context of the Work 1

1 Introduction 3
   1.1 Research Challenges 4
   1.2 Thesis Overview 5

2 Augmented Reality 7
   2.1 Historical Review 7
      2.1.1 How Man and Computer Came to Collaborate 8
      2.1.2 The Origin Depends on Definition 9
      2.1.3 Early Development at University of Utah 10
      2.1.4 Early Development at University of North Carolina 10
      2.1.5 Early Development at MIT 11
      2.1.6 Early Development at Governmental Institutions 12
      2.1.7 A Recent and Subsequently Small Research Area 13
   2.2 Definition of Augmented Reality 13
      2.2.1 Early Applications 14
      2.2.2 Taxonomies 14
      2.2.3 Concepts 15
      2.2.4 Consensus 15

3 Subsystems 17
   3.1 The Tracking System 17
      3.1.1 Tracking Techniques 17
      3.1.2 Tracking Metrics 20
      3.1.3 Lack of Recent Surveys 23
   3.2 The Display 23
      3.2.1 The Anatomy of the Head-Mounted Display 24
      3.2.2 The Video See-Through Misnomer 27
      3.2.3 Displaying for the Human Eye 27
   3.3 The User 31
      3.3.1 Postural Sway 31
      3.3.2 Head-Aiming Performance 32
      3.3.3 The Human Eye During Fixation 33
## Contents

### 4 Calibration Theory

4.1 The Pinhole Camera Model .......................... 35  
4.2 Parameter Estimation ................................ 39  
  4.2.1 Isotropic Scaling ................................ 43  
  4.2.2 Non-Linear Optimization ......................... 44  
  4.2.3 Camera Matrix Decomposition .................... 44  
  4.2.4 Degenerate Configurations ....................... 45  
  4.2.5 Filtering of Measurement Noise .................. 46  
  4.2.6 Camera Calibration Techniques .................. 49  
4.3 Asymmetric Camera Models ........................... 52  

### 5 OST HMD Calibration

5.1 Registration Error ................................. 55  
  5.1.1 Motivation for Minimizing the Registration Error 56  
  5.1.2 Statistical Estimation ............................ 57  
5.2 Visual Alignment .................................. 58  
5.3 Direct and Indirect Measurements .................... 59  
5.4 Minimizing Measurement Errors ..................... 59  
5.5 Collecting Individual Measurements .................. 60  
5.6 Camera-Assisted Measurement ....................... 61  
5.7 Model Selection .................................. 62  

## Contributions

### 6 Summary of Studies

6.1 Paper I ........................................ 67  
  6.1.1 Aims ........................................ 67  
  6.1.2 Results ..................................... 68  
  6.1.3 Contributions ................................ 68  
6.2 Paper II ....................................... 69  
  6.2.1 Aims ........................................ 69  
  6.2.2 Results ..................................... 70  
  6.2.3 Contributions ................................ 70  
6.3 Paper III ...................................... 70  
  6.3.1 Aims ........................................ 70  
  6.3.2 Results ..................................... 71  
  6.3.3 Contributions ................................ 74  
6.4 Paper IV ....................................... 75  
  6.4.1 Aims ........................................ 75  
  6.4.2 Results ..................................... 76  
  6.4.3 Contributions ................................ 79  
6.5 Paper V ....................................... 79
List of Publications

I User Boresighting for AR Calibration: A Preliminary Analysis

II User Boresight Calibration for Large-Format Head-Up Displays

III Visual Alignment Precision in Optical See-Through AR Displays: Implications for Potential Accuracy

IV Visual Alignment Accuracy in Head Mounted Optical See-Through AR Displays: Distribution of Head Orientation Noise

V Optical See-Through Head Mounted Display Direct Linear Transformation Calibration Robustness in the Presence of User Alignment Noise

VI Parameter Estimation Variance of the Single Point Active Alignment Method in Optical See-Through Head Mounted Display Calibration

VII Accuracy of Eyepoint Estimation in Optical See-Through Head-Mounted Displays Using the Single Point Active Alignment Method
M. Axholt, M. Skoglund, S. D. O’Connell, M. D. Cooper, S. R. Ellis and A. Ynnerman, Submitted to the IEEE Virtual Reality Conference 2012, Orange County (CA), USA, March 2012
Part I

Context of the Work
Chapter 1

Introduction

With language, humans are gifted with the capacity of describing abstract and complex constructs to one another. With language, experiences and knowledge can be shared between people, in space as well as time. With language, parents can teach offspring, gatherers and hunters can organize groups, and mathematicians can generalize into abstractions. Uniquely to the human species, language has enabled us to build, and pass on, the collective works of human knowledge. Shared and transcribed as songs and legends, cave paintings, scrolls and books, databases and digital media distributed across networks, we now possess enormous amounts of information. With an ever expanding body of knowledge, we need machines to help us make efficient use of the knowledge others have recorded for posterity. In the era of computers, information is digital. Thus we need to program computers such that they understand our intent and assist us in information retrieval in the way a collaborating colleague would, not just execute commands which only reflect the existing knowledge of the computer user. This is the reasoning which motivates research into efficient Human Computer Interaction (HCI).

The concept of Augmented Reality (AR) has the potential of offering the efficient interface between man and machine that researchers are envisioning. As the name implies, AR aims to augment reality with additional layers of information on top of what is accessible solely with the user's existing human senses. To appropriately place the information in the user's surrounding environment, the computer running the AR system must be aware of the user's location, the current viewpoint, and the location and status of the objects the user wants to interact with. Aware of the user's status and intent, the computer has a greater ability to display the information the user might be in search of or find useful for the particular task. The applications are seemingly endless: thermal and "x-ray" vision, navigation, decision support, simulation, and time travel are just a few examples which provoke the imagination of what could be.

A principal challenge in AR systems is to ensure that the information is displayed in its correct location. Most people would agree that a misaligned arrow pointing a taxi driver the wrong way is an inconvenience. The consequences of a blood vessel ren-
dered at the wrong position in a patient’s body could, however, be much more severe. Now, imagine the potential disaster resulting from a discrepancy in AR systems used by airline or fighter pilots. While it could be argued that these examples where chosen for a certain dramatic effect, and that it would perhaps seem more appropriate to only deploy AR systems for applications considered safe, using inaccurate models may still result in inadvertent changes to the user’s behavior, further complicating the interaction between man and machine. The interplay between the human senses is, as we shall see, quite delicate.

1.1 Research Challenges

To correctly overlay information onto the real world, the system model must correctly describe the user and the surroundings. Such a model has parameters which have to be estimated and tuned. The process of determining an appropriate model and populating it with appropriate parameter values is known as calibration. This thesis is concerned with the topic of studying calibration procedures, determining whether it is possible for a human user to gather data appropriate for such a parameter estimation, and investigating whether the pinhole camera is a suitable estimation of the user’s interaction with an Optical See-Through (OST) Head-Mounted Display (HMD) augmenting the user’s vision.

Quantitatively evaluating the success of OST HMD calibration is difficult because with current technology it is not possible to share the user’s view. Since the calibration data is subjectively gathered, and the the result is only visible to the user, some external reference is needed. Previously researchers have used cameras in the place of the human eye, but this changes the calibration conditions by removing the postural noise which is always present in a human subject.

The noise is an important factor to study because it introduces measurement errors in the calibration procedure. This has an impact on the results as most calibration procedures are based on a system of linear equations which has an algebraic solution. The algebraic solution exists in a space separate from the calibration parameters. A small change in the algebraic space can result in drastic changes in parameter space. This is why the calibration procedure must be designed such that the system of linear equations becomes well-conditioned and numerically stable.

Measurement errors are also introduced by the equipment used in the experimental setup. This implies that data from measurement equipment must be filtered to ensure good quality. Moreover, all measurements should be gathered in the same frame of reference to avoid any bias that could occur in the transformation between coordinate systems.

Lastly, the calibration procedure must be appropriate for a human subject. For example, since humans cannot maintain a stable posture and pivot around a single point
1.2 Thesis Overview

The thesis is divided into three parts. The first part, including chapters 2-5, provide the context for this research. The second part, chapters 6-7, summarize the results and provides a discussion of the published papers which are appended in part three.

Chapter 2 introduces AR, presenting its historical background. This helps in understanding the original inspiration and philosophy that fueled the early development of Virtual Environments (VEs) and AR. It also presents alternative definitions of AR through some high-level taxonomies and exemplifies with some early applications.

Chapter 3 is dedicated to the three largest subsystems of an AR system, namely the tracking system, section 3.1, the display system, section 3.2, and the human operator, section 3.3. Properties that have direct impact on design considerations for experiments and applications are presented.

Chapter 4 explains the principals of the most commonly used calibration model, namely the pinhole camera, and some theoretical concepts supporting the major calibration techniques.

Chapter 5 begins by introducing definitions of the misalignment between real and virtual objects in virtual environments often referred to as “the registration error”. It then continues to present previously researched calibration techniques for OST HMDs.

Chapter 6 introduces the aims, results, and contributions of the seven studies included in this thesis.

Chapter 7 concisely summarizes the findings published in the appended papers and discusses the results and contributions.

The end of part two contains a glossary over the abbreviations used in chapters 1-7.
Chapter 2

Augmented Reality

This chapter introduces AR, presenting its historical background. It helps in understanding the original inspiration and philosophy that fueled the early development of VEs and AR. It also presents alternative definitions of AR through some high-level taxonomies and exemplifies with some early applications.

2.1 Historical Review

The term “Augmented Reality” was coined by Thomas Caudell and David Mizell in a paper from 1992 [27]. Thus, up until 1992 there were no keywords or search terms that would uniquely describe the phenomenon which is the main topic of this review. Instead, to trace the origins of AR one has to follow concepts rather than words. Many researchers start a historical review by referencing Sutherland’s early HMD design from 1968 [123], and perhaps also his visionary paper on from 1965 [121], but not much is said about the research in the 27 years between 1965 and 1992 - or research prior and contemporary to Sutherland for that matter. The section below traces how AR sprang from the desire to extend human abilities through technology, and explains the circumstances that made an engineer utter the following words:

“The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. Handcuffs displayed in such a room would be confining, and a bullet displayed in such a room would be fatal. With appropriate programming, such a display could literally be the Wonderland into which Alice walked.”

Ivan E. Sutherland in The Ultimate Display, 1965 [121]
2.1.1 How Man and Computer Came to Collaborate

In his essay “As We May Think” from 1945, Vannevar Bush, US presidential science advisor and in charge of coordinating scientific research for military purposes during World War II, expressed his concern about the rapid increase of information as research branched off into its various specializations. Reasoning that specialization is necessary for progress, Bush discussed how technology can help man to record, compress, and organize information such that it can be efficiently consulted and turned into knowledge [25].

Bush’s inspirational ideas were quoted extensively by Douglas Engelbart who, in 1962, published “A Conceptual Framework for Augmentation of Man’s Intellect” [42]. This paper outlines a research program at Stanford Research Institute (SRI) which aimed to extend humans’ basic information-handling capabilities. While Bush had mentioned photography, typewriters, and mechanical calculation machines as tools for humans to process information, Engelbart instead suggested the use of more contemporary computing machines. Engelbart approached the research problem by describing the human knowledge acquisition process with a very general conceptual framework containing only four classes: “Artifacts”, “Language”, and “Methodology”, all brought together in a human through “Training”. With these four building blocks he described an HCI scheme which was presented in 1967. It consisted of a computer-driven Cathode Ray Tube (CRT) display system allowing a user to interact with a word-processing application through keyboard and mouse [43]. This demonstrated that a computer was not just simply for computation, but could also be a tool to extend an individual’s basic information-handling capabilities.

Bush and Engelbart were not the only visionaries in pursuit of merging the powers of human and machine. In 1960 J. R. C. Licklider of Massachusetts Institute of Technology (MIT) published “Man-Computer Symbiosis” [86] in which he urged the reader “to think in interaction with a computer in the same way that you think with a colleague whose competence supplements your own”. In his review of necessary input and output equipment for humans to communicate with machines, Licklider commented on the lack of simple tools such as pencil and doodle pad to interface with computers. Similar to Bush, Licklider’s contribution was through visionary ideas rather than inventions.

Inspired by Licklider’s ideas, and in parallel with Engelbart’s research, a graduate student in the Lincoln Labs at MIT named Ivan Sutherland develop Sketchpad, an application accepting input via a light pen to allow a user to perform computer-aided drawing on a two-dimensional CRT display system [122]. Sketchpad was presented at the Spring Joint Computer Conference 1963 as the world’s first graphical HCI interface. The idea of a completely natural interface was further expanded in a paper from 1965 where Sutherland described the computer display of the future to be a “looking-glass into the mathematical wonderland constructed in computer memory” serving as many human senses as possible [121]. Since computers at the time could
not convincingly convey taste, smell, or sound, Sutherland instead limited his scope to conceptually describe a kinesthetic display\(^1\). In accordance with human proprioception, the display is thought to provide both visual and physical feedback to user movements. Sutherland also noted that in a simulated world there is no need for the display system to obey ordinary rules of physical reality, thus suggesting the possibility to use computers to create arbitrarily synthesized environments, hence the allusion to “wonderland”. Three years later, now at Harvard University, Sutherland realized the concept of his visionary paper and generalized human gesturing as computer input to also incorporate user head rotation. Using ultrasound and mechanical linkage, his new system follows the user’s head movements and drew corresponding stereoscopic perspective computer graphics in two miniature CRTs such that the user had the sensation of observing a virtual world [123].

Thus it is generally considered that the idea of extending human’s information-handling capabilities through technology originated with Bush, but Engelbart formulated the research plan on how to adapt computers for human interaction. With his series of papers on augmenting man’s intellect, Engelbart is credited with being a pioneer within HCI, but in fact Sutherland was the first to produce hardware and publish results illustrating these ideas – work initially influenced by Licklider.

### 2.1.2 The Origin Depends on Definition

While Sutherland’s work have been inspirational to many researchers within HCI, the actual origin of synthesized environments is a question of definition. Extensive historical reviews made by Scott Fisher [49], Warren Robinett [108], and Stephen Ellis [41] suggest Morton Heilig’s Experience Theater from 1955 [67], later named “Sensorama”, to be the first synthesized environment. Sensorama sprang from cinematography and was intended to expand the movie experience by means of a stereoscopic view projected in a personal half-dome, three-dimensional binaural sound, vibrations, and a fan providing a flow of air passing over compartments with wax pellets impregnated with odors [49].

However, in contrast to the static nature of recorded film, and with emphasis on interactivity, in 1958 Charles Comeau and James Bryan presented a remotely controlled television camera transmitting its live imagery to a head-mounted, biocular, virtual image viewing system which in turn was coupled with the remote camera such that the camera rotated according to the user’s head movements [33].

The two systems are similar to Sutherland’s in that they immerse the user in a synthesized world, but differ on several key points that will be further discussed in section 3.2 on display systems. This review will not go further into telepresence or cinematic experiences but will focus on computer-generated synthesized environments similar

---

\(^1\)Kinesthesia refers to the sense that detects bodily position, or movement of muscles, tendons, and joints. It is sometimes used synonymously with proprioception.
to that of Sutherland’s.

2.1.3 Early Development at University of Utah

After Sutherland moved his research to the University of Utah in 1970, his PhD student Donald Vickers improved the system by including an ultrasonic pointing device which enabled the user to interact with surrounding virtual objects \[135\]. The interaction device as such was not new, as a similar wand had already been used for two-dimensional drawing with Sutherland’s Sketchpad in 1966 \[107\]. Instead Vickers’ contribution consisted of letting the user touch, move, and manipulate the shape of virtual objects in three-dimensional space. James Clark, also a PhD student with Sutherland, expanded on this concept by letting the user interactively manipulate control points describing three-dimensional surfaces \[30\]. Since the interaction device only reported position, the user could not perform any selection task from a distance but had to walk up to the virtual object to touch it. Ray-based selection, or what is more simply known as pointing, was not implemented for interaction with computers until 1980, when Richard Bolt of MIT attached a device which reported orientation of a user’s hand so that pointing gestures could be performed in the MIT Media Room \[17\].

While Vickers and Clark studied basic system functions such as selection and manipulation techniques, Robert Burton, also a PhD student at University of Utah, took on another fundamental function, namely the task of improving the system that tracked the user’s movements. Burton’s system, published in 1974, improved on previous systems in that it did not require mechanical linkage, did not rely on sensitive ultrasonic time-of-flight measurements, and could track several points in space simultaneously, making it possible to track more complicated user movements. Burton’s optical tracking system was based on one-dimensional sensors which reacted to light emitted from time multiplexed lamps which, in turn, were synchronized with slits on a rotating disc \[24\]. The system, called Twinkle Box, was later improved by Henry Fuchs, also PhD student at University of Utah, who replaced the sensors with Charged Coupled Devices (CCDs), used laser points instead of lamps, and did away with the spinning disc such that the system became more precise and also could digitize entire opaque three-dimensional surfaces \[53\].

2.1.4 Early Development at University of North Carolina

University of North Carolina (UNC) also had an early interest in computer-generated synthesized environments, although following a different approach. Lead by Frederick Brooks, the computer science department of UNC initially investigated the topic via force-feedback devices, providing the user with a tactile interface to explore molecular forces (p. 35, fig. 2-16 \[7\]). In a paper presenting the general findings from the
suitably named GROPE projects during 1967 to 1990, Brooks explained how incremental changes to the haptic system evolved it from a state where monetary incentive was not enough to motivate subjects to complete an experiment, to a state where subject performance doubled and Situational Awareness (SA) in expert decisions greatly improved [19]. In a paper from 1988, titled “Grasping Reality Through Illusion – Interactive Graphics Serving Science” [18], Brooks defended three-dimensional interactive graphics as a scientific tool despite its immature technology. In an attempt to push research forward he urged others to report their less rigorous, unproven, results and, at the same time, listed the partially unevaluated display technologies conceived at UNC thus far. Of particular interest for this review, Brooks mentioned a new type of HMD based on a welder’s helmet that, at the time, had not been scientifically evaluated but is described in a technical report from 1986 by Richard Holloway [71]. Fuchs, previously of University of Utah, had moved his research to UNC a few years earlier and now joined in the development of this early HMD [108]. The HMD was revised into a more robust design described by James Chung et al. in a technical report from 1989 [29].

In the years to follow the research on computer-generated synthesized environments at UNC followed two tracks. One track continued to answer fundamental problems such as display calibration and image registration, while the other track dealt with applied research on visualizing large datasets by studying molecular forces with haptics and displaying x-ray images. For example in 1992 Michael Bajura presented how to superimpose ultrasound echography data directly on a patient using an HMD [6]. In 1991 Warren Robinett, formerly of NASA Ames Research Center, and Jannick Rolland, with a genuine background in optics, published a static computational model for the optics in an HMD [109]. This work is highly relevant for this literature study and also provides a great review on prior work. A similar but extended review on prior work can be found in a technical report by Robinett published the year after [108]. The work on image registration continued with PhD students Richard Holloway and Ronald Azuma. Holloway specialized in static registration errors in medical datasets and Azuma investigated dynamic registration errors due to tracker latency. Their work is considered current and is, therefore, presented later in this review.

### 2.1.5 Early Development at MIT

In his paper from 1968 Sutherland noted that motion parallax (kinetic depth effect) and binocular parallax were important factors in conveying a sense of depth in a computer-generated synthesized environment [123]. Scott Fisher of the Architecture Machine Group at MIT made further investigations on these factors in a paper from 1982 [47]. He presented a system where the views through left and right eye were alternated in synchrony with a corresponding view on a TV-set, similar to what later would be known as shutter glasses. The system also tracked the user such that lateral motion would introduce motion parallax. While sufficiently fast image generation...
was the primary challenge in Fisher's paper, it is the subsequent work made by Fisher's
colleague, Christopher Schmandt, that is of interest for this review. Schmandt pivoted
the TV-set and reflected it in a half-silvered mirror such that the computer graphics
was spatially mapped to the real objects on the work area of a desk [113]. In this
paper, on the topic of related work, Schmandt briefly mentions a master's thesis from
1983 written by Mark Callahan, also of the Architecture Machine Group, which also
dealt with the subject of using half-silvered mirrors as optical combiners in HMDs.
While the original thesis has proven hard to find, Chung describes the system as CRTs
mounted facing downward on the forehead of a bicycle helmet and the mirrors worn
slanted in eyeglass frames [29].

2.1.6 Early Development at Governmental Institutions

Outside of academia, the primary application of computer-generated synthesized en-
vironments presented in HMDs was either to construct training and simulation envi-
ronments or to provide tools for teleoperation. Space and aviation was, and still is, a
particularly interesting application area for virtual environments given the multitude
of reference frames and the possibility for a pilot to use his body to give directional
commands to the machine. Furthermore, the advent of glass cockpits, where gauges
and dials to a large extent were replaced by multi-purpose displays, sparked studies
on SA in the context of complex virtual environments [57].

In 1977 a project to develop a so called Visually Coupled Airborne Systems Simu-
lator (VCASS) was initiated at the United States Air Force (USAF) Wright-Patterson Air
Force Base [23]. The VCASS (shown on p. 6 in [96]) was developed to be a ground-
based simulator to test interface concepts considered for implementation in cockpits
of real and remotely controlled aircraft. The VCASS was later incorporated in a Super
Cockpit program [55] directed by Thomas Furness who, in addition to concluding
some observations from the program [57], also had visionary and high-level ideas on
HCI and how to extend human abilities [56], similar to the principles expressed by
Engelbart. It should, however, be noted that not all HMDs developed for military use
aspired to convey a virtual world which corresponds to the definition of computer-
generated synthesized environments covered in this literature review. As seen in the
historical chapter of the extensive book on human factors in HMDs, published by
United States Army Aeromedical Research Laboratorys (USAARLs) [103], and also in
the historical review on Honeywell displays [11], even recent HMDs sometimes suf-
face with non-conformal imagery such as crosshair and icons. These types of displays
can, instead, more generally be categorized as Visually Coupled Systems (VCS).

A good example of a military flight simulator, developed by Canadian Aviation Elec-
tronics (CAE), is described in a report published in 1990 by the Air Force Human Re-
sources Laboratory (AFHRL) at the USAF Williams Air Force Base [9]. The report cov-
vers four years of development of concepts which must be considered quite advanced
2.2. DEFINITION OF AUGMENTED REALITY

for their time. The imagery is relayed from powerful light valves, via fiber optics, to
to optical combiners in the HMD producing a binocular overlap which increases the res-
olution of the user's central vision. The intention was to use eyetracking to generate
imagery with lower Level of Detail (LOD) in the user's peripheral view but, judging
from the report, predicting eye movement was somewhat challenging. Even 20 years
later this report is well worth browsing for the practical problems it addresses.

During the same time, in the mid 1980s, Fisher, formerly of MIT, and Robinett,
later at UNC, teamed with principle investigators Mc Greevy and Humphries at NASA
Ames Research Center and began researching computer-generated synthesized en-
vironments for applications within telerobotics. Built on a motorcycle helmet the
system was flexibly extended to also investigate management of large-scale informa-
tion systems implementing voice command and data glove gesturing [48]. While the
engineers developing simulators for the army had a firm grasp on the optics, using
pupil-forming systems with relaying optics, the solution in the first NASA HMD was
simple but effective. Named Large Expanse, Extra Perspective (LEEP) optics it con-
sisted of two convex lenses that magnified and collimated the rays from a relatively
small display surface. This simpler, less expensive, non-pupil-forming, optical design
is now more common in mid-range HMDs than the pupil-forming systems which are
mainly found in military applications.

2.1.7 A Recent and Subsequently Small Research Area

Researchers have attempted to achieve a natural interface between machine and hu-
man with the intention of augmenting the user’s abilities for several decades, but
when technology was not mature enough to deliver results, visionary ideas were pub-
lished for future researchers to implement. This explains the illustrative wording in
Sutherland’s paper. As technology became available, MIT and University of Utah lead
the development to interface humans to computers via various modalities, but were
soon followed by UNC, the US Army, and NASA. Authors’ different affiliations on
papers seem to suggest that researchers were mobile across institutions. Thus, de-
nspite being named Historical Review, it is worth noting that this research area is fairly
new and subsequently has a relatively small group of researchers associated with it.
The majority of the researchers mentioned in this section are still active and meet at
conferences. In fact, during an internship at NASA Ames Research Center in 2007 I
personally had the privilege to meet with Douglas Engelbart.

2.2 Definition of Augmented Reality

Based on the development presented in the previous section, a definition of AR could
be summarized as “the technology with which computers synthesize signals intended
for interpretation by human senses with the purpose of changing the user’s perception of the surrounding world”. However, the most commonly cited definition of AR is more pragmatic and captured in Ronald Azuma’s three criteria [4]:

- **Combines real and virtual**: Superimposes virtual information on the real world in the same interaction space.
- **Interactive in real time**: The virtual information updates as the user (and objects) moves, in the interaction space.
- **Registered in 3D**: The virtual objects are assigned to, and remain in, a particular place in the interaction space.

### 2.2.1 Early Applications

Papers on early applications generally do not formalize definitions, but are nevertheless interesting because their detailed system descriptions help in understanding what AR is. For example, in the paper from 1992 written by Thomas Caudell and David Mizell of Boeing, in which the term “Augmented Reality” was coined, an application superimposing virtual indicators over a peg board for the purpose of bundling cables for airplane construction is described. A passage describing AR relative to Virtual Reality (VR) is found in section two [27]. The same year Michael Bajura, Henry Fuchs, and Ryutarou Ohbuchi described an application where a user performed an in situ exploration of a 3D medical ultrasound dataset using a head-mounted display [6]. While the term AR is not used, Bajura et al. clearly illustrated the problem of registering the dataset onto the correct part of the patient’s body, a challenge central to all AR systems which will be discussed in section 5.1 on registration errors later in this text. The following year, Steven Feiner, Blair MacIntyre and Dorée Seligmann described a system they named Knowledge-based Augmented Reality for Maintenance Assistance (KARMA), explaining and assisting complex 3D tasks. While it contains detailed explanations and imagery, it also touches on human factors related to application and system requirements [45].

### 2.2.2 Taxonomies

Another method of formulating a definition is to work with established taxonomies. An extensive taxonomy, comprising nine parameters, was suggested by Warren Robinett [108]. It can categorize synthetic information ranging from photos to teleoperation environments and is suitable to classify, for example simulators. A simpler taxonomy which is more often used in the context of AR, is the “Virtuality Continuum” proposed by Paul Milgram and Fumio Kishino [97]. It organizes display types according to their level of immersiveness. An AR display device can also be categorized as head-mounted, hand-held, spatial, or projective according to the taxonomy.
2.2. DEFINITION OF AUGMENTED REALITY

by Oliver Bimber and Ramesh Raskar (p. 72 [13]). Lastly, AR displays are also frequently divided into optical and video see-through.

The taxonomies mentioned so far are listed in ascending order relative to how often they are cited in AR literature, but of course more detailed taxonomies exist. In fact, all AR systems can be categorized based on the specific nomenclature used in the individual research areas that together form AR as a general topic. As an example, displays can also be categorized as “monocular”, “biocular”, or “binocular” depending on whether one or two eyes receive a signal, and whether the two eyes receive the same or individual stimuli. Such nomenclature will be discussed as the various subsystems are later described.

2.2.3 Concepts

A third way to understand the technology which researchers agree to be AR, is to read conceptual texts. In a paper from 1993 Pierre Wellnet et al. suggest expanding the interaction space of the computer outside of the desktop metaphor [141]. The “Paperless Office” touches on a concept known as Ubiquitous Computing which is a more general topic where seemingly mundane objects, such as a piece of paper, serve as computer interfaces allowing the user of an AR system to not only consume information, but also provide feedback. This is a concept which Wendy Mackay revisited in 1998 describing how not only objects, but also the user and their surroundings, can be augmented [93]. Thus, AR can be seen as an interface to access the power of Ubiquitous Computing which is thought to exist as an inextricable and socially invisible part of the surroundings [45].

2.2.4 Consensus

The most widely cited definition of AR is offered by Azuma [4]. Alternatively, by understanding how AR development evolved, AR could also be defined relative its origin through the use of taxonomies. One such example is the relationship between AR and and VR along the Virtuality Continuum [97] which is also a commonly referenced definition. However, since taxonomies may vary depending on the needs of the particular classification, AR has also been defined from different points of reference, for example through a number of parameters describing synthetic experiences [108]. This taxonomy might, however, be too detailed to be useful in practice. So far this work has only been cited three times. The essence of AR can be inferred by reading about early applications [27][6][45] or conceptual texts [141][93]. These works have been cited extensively and seem, in addition to Azuma’s definition, to be the preferred way to describe AR as a research area.
Chapter 3

Subsystems

This chapter is dedicated to the three largest subsystems of an AR system, namely the tracking system, section 3.1, the display system, section 3.2, and the human operator, section 3.3. Properties that have direct impact on design considerations for experiments and applications are presented. This chapter describes an AR system in terms of three subsystems: tracking system, display system, and human operator. This division provides sufficient detail relative to the scope of this review. However, researchers particularly interested in temporal aspects such as pipeline synchronization and processing latency may want to include a rendering unit as a fourth separate system [5][77][3].

3.1 The Tracking System

The tracking system is a very important component in an AR system, because it is responsible for measuring the position and orientation\(^1\) of the user as well as objects in the surrounding space. The data from the tracking system is primarily used to update the user’s view, but also to measure reference points important for calibration. Hence, the quality of the tracker data has a direct impact on the calibration quality. Several tracking techniques exist and they all have intricacies and typical behavior. This subsection introduces the common techniques and also some metrics with which to evaluate them.

3.1.1 Tracking Techniques

In a 1993 survey report on position trackers for HMDs from UNC, Devesh Bhatnagar defines a tracking system as responsible for reporting location of some object, possibly an HMD, to a host computer. He also categorizes tracker types and performance

\(^1\)Position and orientation is collectively referred to as pose or location.
Chapter 3. Subsystems

Bhatnagar’s report is cited and significantly elaborated in a book chapter on motion tracking written by Eric Foxlin of InterSense [51]. Additionally, Foxlin’s book chapter contains an extensive taxonomy, practical advice on implementation, and closes with filtering techniques to mitigate system latency. A condensed version of Foxlin’s book chapter, along with a discussion of how no single optimal tracking method for all applications exists, is available in a journal article from 2002 [15] co-authored with Greg Welch of UNC who, together with Gary Bishop, has a particular interest in filtering techniques [140]. The tracker taxonomy and metrics presented later in this thesis have been compiled from several publications, mainly from UNC. The interested reader might find the extensive review of prerequisites for VEs from 1993 written by Richard Holloway and Anselmo Lastra particularly valuable [73].

- **Mechanical**: Measures angles in the joints in a linkage of rigid units of known lengths.
- **Inertial**: Determines linear and angular acceleration with accelerometers and gyroscopes.
- **Acoustic**: Measures phase difference or time of flight in ultrasonic signals.
- **Magnetic**: Emits a magnetic field inducing voltage in perpendicular coils.
- **Optical**: Uses CCD or Complementary Metal Oxide Semiconductor (CMOS) arrays to detect contrasts in light.
- **Hybrid**: Any combination of the above.
- **Others**: Systems that do not fit the taxonomy above.

**Mechanical**: Common to all mechanical tracking is that the device establishes an object’s location by measuring angles between joints in a structure of rigid units of known lengths [51]. For example, a haptic system typically uses fixed-reference forward kinematics. This could be contrasted with an exo-skeleton which has a moving-reference as the location of its limbs in the surrounding world can only be described relative to the origin of the skeleton, and not relative to some fixed world origin as in the case of a stationary haptic station. Fixed-reference and moving-reference are also known as absolute and relative location reporting respectively. A tracking system designed for relative location reports must use dead reckoning to report its location in absolute world coordinates [73]. Although Sutherland originally made use of a mechanical tracking system [123][139], such systems are practically obsolete for modern HMD applications, mainly because of the limited working volume and obstructing linkage of fixed-reference systems [12].

**Inertial**: Accelerometers and gyroscopes report linear and angular acceleration by measuring displacement of masses reluctant to move because of their inertia. In resting position an accelerometer measures 1 g (9.81 m/s²) in the vertical direction due to earth’s gravity. An erroneous estimate of the direction of the gravity vector will contribute to the accumulation of drift which adversely affects the systems ability to
3.1. THE TRACKING SYSTEM

report its absolute location while dead reckoning. Theoretically, 0.001 g bias error results in 4.5 m drift over 30 s [139]. Furthermore, inertial tracking systems are designed to encompass a particular dynamic range. Acting outside of it, for example by moving extremely slowly, may introduce quantization errors which also cause drift [15].

Acoustic: Similar to how submarines range objects using sonar, acoustic trackers measure the time of flight of ultrasonic signals. By placing the microphones some distance apart on a rigid body the system can also report object orientation in addition to position. If the piezoelectric speakers (emitters) are located in the surrounding and the microphones (sensors) are located on the tracked object the configuration is referred to as inside-out, which means that the system senses in an outward direction from the tracked object [51]. The opposite, outside-in, could be exemplified by wall-mounted microphones listening for signal changes as the object moves. (Some researchers extend this taxonomy by referring to self-contained systems, for example inertial systems, as inside-in [73].) Resolution can be improved by measuring the device’s movement relative the known phase of the signal, a technique which also can be used to discard bouncing (multipath, echoes) signals [139]. Left uncompensated for temperature, measurements can vary by 1.6 mm/m for every degree of deviation from the optimal working temperature [51].

Magnetic: Magnetic trackers consist of a base station that emits an electromagnetic field which, in early versions, had an alternating directionality. Nowadays trackers primarily use directional magnetic fields which are pulsed per hemisphere to avoid interference. The sensor’s orientation and position can be inferred from the voltage generated by the electromagnetic induction as the coils are excited by the electromagnetic pulses [12][89][139]. At close range magnetic tracking is accurate and precise, but Holloway reports that measurement precision deteriorates proportionally to the square of the distance between emitter and sensor (p. 125 [74]). Moreover, Bryson illustrates how accuracy varies throughout the tracked volume, not only spatially but also temporally [21]. Bryson interpolates between discrete measurement points stored in Look-Up Tables (LUTs) to correct for this varying bias and also attempts to fit the measurements to a function modeling the error. Livingston and State argue that such LUTs should be parametrized not only with position but also sensor orientation which increases the number of samples required for correction [89].

Optical: Optical trackers generally refer to one or several cameras, or camera-like sensors, where rays of light are focused with optics and detected on a CCD or CMOS array. As with acoustic systems, cameras can be mounted on the tracked object, “inside-out”, or in the environment surrounding the tracked object, “outside-in” [12]. Inside-out offers the added benefit of improved sensitivity to rotational movement [138]. In specialized systems, the array is normally one-dimensional to enable a faster sensor readout and thus a higher update rate which means that a greater number of time multiplexed blinking light sources can be identified and followed [24]. The basic idea behind optical tracking in early implementations [53] has similarities to the ge-
ometric interpretation of camera planes in modern computer vision [65] which will be discussed in section 4 on calibration theory. In camera-based systems, variance in light rays gives rise to changes in contrast which, in turn, is the basic property in features or feature points (p. 205 [125]). Changes in contrast can be detected in the diffuse light reflection of unprepared objects and is therefore referred to as natural feature tracking or markerless tracking [90]. The tracked object can also be prepared and fitted with fiducial markers which enhances contrast and facilitates image segmentation [104][79]. If the marker simply reflects light it is referred to as a passive marker [105], whereas for example Light Emitting Diodes (LEDs) emit light and are therefore known as active markers or beacons [12][51]. Fiducial markers are often axially non-asymmetric to encode object orientation. Several markers can be used simultaneously if unique marker identification is encoded, either in marker shape, color, or blink frequency.

Hybrid: As pointed out by Welch and Foxlin, it is challenging to find a single tracking technique that satisfies all requirements for a flexible VE [139]. Luckily, some tracking techniques are complementary in the sense that one technique is stable when the other is operating under unfavorable conditions. For example, while inertial tracking is accurate and has a relatively high update rate, it will, as mentioned above, eventually start to drift as quantization errors corrupt the gravity vector. The quantization errors are less significant if the Inertial Measurement Unit (IMU) is in constant motion [15]. Constant motion, however, is not an ideal condition for a camera-based system, in which a stable, or slowly moving, platform is preferable due to the deteriorating effects of motion blur. Thus, camera-based optical tracking and inertial tracking have properties that make them complementary as their sensor data can be fused with, for example, some variety of a Kalman filter [150][149][16][70]. Complementary hybrid configurations also exist for magnetic/optical [119], inertial/acoustic [52][148], and Global Positioning System (GPS)/optical [10] tracking, to mention a few.

Others: Some proximity (for example Radio Frequency Identification (RFID)) and beacon triangularization techniques (for example GPS and wireless Ethernet signal strength) cannot be readily sorted into the UNC taxonomy above. While these techniques in some instances may offer inadequate resolution for exact localization, they are still useful for some AR and Ubiquitous Computing concepts and have been organized in a taxonomy by Jeffrey Hightower and Geatano Borriello [68].

### 3.1.2 Tracking Metrics

The surveys cited in this text [12][73][15][139] specify a number of metrics for quantifying the quality with which a tracker sensor reports the location of a tracked object. Aside from comments on metrics that are characteristic for a particular tracking technology, quantitative data has deliberately been left out as tracker performance is highly dependent on how the system is designed, deployed, and used.
3.1. THE TRACKING SYSTEM

- **Update Rate**: The frequency with which measurements are reported to the host computer.
- **Delay/Lag/Latency**: The amount of time between a change in location of the sensor and the report of the change to the host computer.
- **Precision/Jitter**: The spread (normally standard deviation or root mean square) of position or orientation reports from a stationary sensor over some time period.
- **Accuracy**: The bias (the difference between true and measured value) of position or orientation reports from a stationary sensor over some time period.
- **Resolution**: The smallest change in position or orientation that can be detected by the sensor.
- **Interference/Spatial Distortion**: A change in the measurement bias as a function of the sensor's position and orientation in the tracked volume.
- **Absolute or Relative**: Whether the sensor reports in an absolute coordinate system with a fixed origin, or relative to a coordinate system which is moving or is defined at system start.
- **Working Volume/Range**: The volume within which the sensor can report with a specified quality.
- **Degrees of Freedom**: The number of dimensions of measured variables that the sensor is able to measure.

**Update Rate**: Update Rate refers to how often a tracking system can produce a location report of its sensors. Too low an update rate will prevent the system from faithfully sampling object or user movement. To avoid aliasing, the sampling frequency must be at least twice as high as the highest frequency component of the sampled signal, according to the Nyquist-Shannon sampling theorem (p. 42 [116]). As some trackers produce spurious readings, it is good practice to sample at a rate that allows for additional post-filtering without introducing discernible latency. Additionally, a low tracker update rate will contribute to the AR system's total end-to-end latency [77] and cause a dynamic registration error (further discussed in section 5.1).

**Delay/Lag/Latency**: This metric refers to the tracker's internal latency and reflects how long it takes for the tracker to perform its computation and filtering operations. If Update Rate denotes how often a new location report is available, the Delay/Lag/Latency metric denotes how current that report is. Bernard Adelstein, Eric Johnston and Stephen Ellis investigated amplitude fidelity, latency, and noise in two magnetic trackers [2]. The latency inherent to the magnetic trackers, without filtering, was estimated to be 7-8 ms (which is twice as long as reported by the manufacturer according to Bhatnagar [12]). However, for other types of trackers the delay can be far greater. Marco Jacobs, Mark Livingston, and Andrei State reported the (off-host) tracker latency of a mechanical arm to be an order of magnitude higher.
CHAPTER 3. SUBSYSTEMS

compared to magnetic tracking [77]. At the other end of the spectrum we find inertial tracking in which no post-filtering is necessary, resulting in latencies on the order of a fraction of a millisecond [51].

Precision/Jitter: Foxlin defines jitter as “The portion of the tracker output noise spectrum that causes the perception of image shaking when the tracker is actually still.” [51], but many researchers refer to jitter simply as noise. It can also be interpreted as a measurement of spread and be referred to in terms of standard deviation, root mean square, or sometimes signal to noise ratio. Jitter is present in most trackers, e.g. magnetic (§1.2 and fig. 6-7 [21], p. 125-126 [74]), optical (section IV [63]), and acoustic tracking [139]. The notable exceptions are inertial [51] and mechanical tracking (section 1.2.3. [89]). In addition to the choice of tracker technique, jitter can also arise from poorly tuned prediction algorithms [5]. Jitter can be suppressed with filters. Eric Foxlin, Michael Harrington, and Yury Altshuler of InterSense describe a Perceptual Post-Filter (PPF) which exploits the fact that jitter is only discernible when the user’s head is stationary or moving slowly. When the user’s head is close to stationary the penalty of increased latency is acceptable as it is not likely to result in a dynamic registration error (see section 5.1) [52].

Accuracy: Accuracy is the constant offset between real and measured location and can be interpreted statistically as the average location reported by a stationary sensor over time.

Resolution: Resolution is the minimum amount of change that the sensor can detect. IMUs which are based on displacing a mass may, for example, not be affected by subtle movements that are outside of its dynamic range [15].

Interference/Spatial Distortion: Because of the natural shape of electromagnetic fields, magnetic trackers usually exhibit varying accuracy as a function of distance between tracker sensor and emitter [21][89]. The error vectors illustrating the distortion usually take on the appearance of the curved electromagnetic field, but may also be affected by ferro-magnetic metal objects in the surroundings [151] which, after exposure to the same field, generate eddy currents and develop interfering fields of their own [12]. Another case of degrading performance as a function of distance occurs in optical tracking when the baseline between two (or several) cameras is too small to properly determine the distance to a tracked object [65]. Similarly, insufficient separation between GPS satellites results in poor readings. This uncertainty is called Geometric Dilution of Precision (GDOP) [51].

Absolute or Relative: If the tracker’s location reports are expressed in a coordinate system with a fixed origin, the reports are said to be absolute. If the tracker has a moving frame of reference, or has a frame of reference that is initialized on each system start, the reports are said to be relative. Tracker accuracy cannot easily be determined in a system with relative reports as there is no known point in the surrounding world to use as reference.
3.2 THE DISPLAY

Working Volume/Range: Mechanical trackers are limited to the extent of their linkage [12], and the precision of a magnetic tracker attenuates with distance [21]. These are examples of some of the factors that limit the working volume of a tracker. Principles for scalability to extend the working volume exist for optical tracking [136][138]. Self-contained tracking systems, however, such as internal tracking, do not have a working volume limitation [139].

Degrees of Freedom: Most trackers offer either three or six Degrees of Freedom (DOF). The DOF refer to the three directions around, or along, which the tracker can rotate or translate. For instance, inertial tracker system can incorporate either gyroscopes, or accelerometers, or both, resulting in either three or six DOF [52]. Tracker systems with ability to only report position can be extended to also report orientation, provided that at least three positions on a rigid body are known [62]. Some trackers are designed to report the collective state of articulated parts or limbs. In this case the tracker’s DOF may be higher.

3.1.3 Lack of Recent Surveys

Some of the references cited in this section are rather dated. There are several reasons for this. Firstly, there was a period in the early 1990’s when UNC published a lot of material on tracking. Since then, no particular research institution has distinguished themselves by publishing detailed research results on a wide number of tracking techniques for application within VEs. Moreover, there are very few recent reviews on performance available in general. This may be because there are only a limited number of tracking system manufacturers and improvements to new model revisions may not be substantial enough to warrant a new comparative paper. It may also be because researchers have noted that performance varies a great deal depending on the surrounding environment and how the system is assembled and used. Secondly, early and summarizing works serves as good starting points to understand the various tracking techniques and are, even if old, therefore well suited for a brief survey like this one. Thirdly, in recent years, it seems as if efforts to improve performance have shifted from sensor hardware design to various forms of filtering. Contributions within the field of tracking have become more specialized and are therefore published in conferences and journals outside of the AR and VR community.

3.2 The Display

The second subsystem to be presented is the display system. While a literal interpretation of Sutherland’s vision [121] would suggest a broader use of the word display, as AR can augment any of the human senses (p. 567 [7]), the following section only covers visual signals intended for the human eye. Additionally, although AR can be
CHAPTER 3. SUBSYSTEMS

consumed through hand-held devices, spatial displays (for example Head-Up Displays (HUDs)), and via projected imagery [13], this review will focus only on HMDs, and particularly on OST HMDs.

This subsection is important because it explains how the visual component of a calibration procedure is produced, and how it has been adapted to fit the human eye with the help of lenses. It also explains why systems should be described with spatial resolution rather than screen resolution, and why luminance should be specified in the description of experimental setups.

3.2.1 The Anatomy of the Head-Mounted Display

The viewports of HMD can be intended for monocular (one eye, one viewport), binocular (two eyes, one viewport), or binocular (two eyes, two viewports) viewing (p. 76 fig 3.10 [133]) [95]. With a binocular system, the two viewports can offer stereoscopic vision and binocular overlap to increase Field Of View (FOV) [109]. The overlap can be created by rotating the viewports with a cant angle inwards (convergent) or outwards (divergent) (ch. 3 [96]).

The viewports normally have adjustments that allow them to be positioned in front of the user's eyes, respecting the user's Interpupillary Distance (IPD) and the optical axes of the eyes. The viewports can be divided into pupil-forming and non-pupil-forming systems [26]. The pupil-forming design consists of a system of lenses forming an intermediary image which in turn is magnified by an eyepiece, much like microscopes, gun sights, or periscopes (p. 816 [103]). The rays of the final magnification lens form an exit pupil, a bright disc where the bundle of light rays converge, which must be matched by the location of the user's eye. An exit pupil displaced to either side of the optical axis of the eye causes vignetting and aberrations, and an exit pupil too far from or too close to the eye causes differences in luminance with resulting loss of contrast (pp. 123, 818-819 [103]). The non-pupil-forming design consists only of a magnifying lens [26] (p. 177 [7]). The main difference between the two designs is that pupil-forming systems allow for an image to be relayed over a longer path, which provides the designer with a greater flexibility to, for instance, place the parts of the HMD system such that the weight is more well-balanced on the user's head, while the non-pupil-forming is simpler to manufacture and maintain (p. 817 [103]). (Pupil-forming HMD optics can further be divided into on-axis and off-axis systems as well as refractive and catadioptric optics. For further details, the interested reader is encouraged to read Mordekhai Velger's book on HMD design [133].)

In an AR HMD, each viewport has an optical combiner which fuses the image of some display with the image of objects in the surrounding world. The most intuitive combiner is a half-silvered mirrors placed at an angle against the optical axis (p. 133-135 [133]). However, combiners can also be made in the shape of prisms [133] and “pancake windows” [9][96]). They all serve the purpose of reflecting the imagery on
3.2. THE DISPLAY

a display into the eye. In the field of optics, an image of an object (for instance, the one shown on the display) that has been reflected such that the object appears to be located elsewhere (say, in the real world) is referred to as a virtual image (p. 18 [7]).

In addition to combiners, each viewport is also fitted with optical lenses. The list below is an enumeration of the effects that can be achieved with lenses. The list is adapted from two chapters on HMD design written by James Melzer [95] (p. 816 [103]), and another chapter written by Clarence Rash et al. (p. 109 [103]).

- **Collimate:** Create a virtual image which can be perceived at the same depth of field as the objects in the surrounding world.
- **Magnify:** Make the virtual image appear larger than its actual size on the display surface.
- **Relay:** Move the virtual image away from the display.

Collimate: While collimation can be used reduce the effect of vehicle vibration on symbol legibility (p. 743 [103]), it is more commonly used to provide a relaxed viewing condition in HMDs where the display surface is located so close to the user’s eye that it cannot be brought into focus by the user's eye alone. In these cases, the light rays from a display are instead collimated to produce a virtual image some distance in front of the user which requires less strain to focus on. The effect is normally achieved with diverging lenses. This way, collimation can also be used to present virtual objects at the same depth of field as the real world counterparts. This helps to minimize the accommodation/convergence conflict. Such a conflict occurs when the physiological depth cues of accommodation and vergence, normally operating together, are separated. Ideally the eye is rotated so that its optical axis is pointed towards an object of interest at a certain depth. By previous knowledge, but also reflexively triggered by image blur, the user’s lens is accomodated to bring the object at that depth into focus. In a system using collimated light rays it is possible to decouple accommodation and vergence. [109] [111] (p. 49 [133]).

A collimated bundle of light rays travel approximately in parallel. As such, they appear to emanate from some point in the distance which si commonly referred to as optical infinity [109]. Infinity in this context refers to a distance at which the optical axes of the user’s eyes are already approximately parallel, or rather the object distance at which a user focusing on said object cannot sense a difference between the change in vergence angle between the optical axes of the eyes if the distance to the object were to change. Assuming an IPD of 6.5 cm, the vergence angle of the user’s eyes would depart 0.62°, 0.31°, 0.19° from parallel while observing an object at 3 m, 6 m, and 10 m, respectively. This corresponds to 0.21, 0.05, and 0.02 prismatic diopter².

It should be noted that while parallel rays from a virtual image some distance away result in a more relaxed viewing condition for the user compared to viewing a display

---

²One prismatic diopter refers to the optical strength in a prism (or lens), which can deviate a beam of light 1 cm on a plane placed perpendicularly to the beam at 1 m.
at close distance, there is still some strain in maintaining such a vergence and accom-
modation. A fully relaxed state of vergence and accommodation only occurs when the
eye has no visual stimuli, such as in the case of darkness or clear skies. This occurs
for objects between 0.5-2 m (p. 49 [133]).

The interested reader is encouraged to look at the work of Eli Peli, Professor of Oph-
thalmology at Harvard Medical School, who presents optometric and perceptual is-
ues with HMDs in systems which are designed for distances ranging from 0.5 to 6 m
(p. 246 [99]).

Magnify: A magnifying lens enlarges the image of the display. This is necessary as
some microdisplays have dimensions smaller than 2.5 cm diagonally (p. 167 fig. 4-50
[103]). In a non-pupil-forming system, the size of the display surface and the focal
distance of the magnifying lens governs the FOV, which in turn affects the spatial
resolution, that is the visual angle subtended by each pixel (p. 60 [96]) (p. 243
[99]) (p. 178, 223, fig. 6-25 [7]). The trade-off between FOV and spatial resolution
is also discussed by Cakmakci and Rolland. In their very informative paper on the
optics of head-worn displays, they also explain how the Lagrange invariant governs
the relationship in size between FOV and exit pupil, and how the luminance of a
virtual image remains constant, regardless of magnification [26].

In VR HMDs the FOV can be expanded using LEEP optics [76] which effectively mag-
nify the display surface such that the eye can rotate and perceive a spherical view of a
planar surface. This technique is only suitable for VR systems in completely synthetic
environments and is not suitable for AR applications because the system of lenses
changes the gain of eye rotation in a way which upsets user proprioception (fig. 3
[75]).

Relay: In a pupil-forming system, a system of lenses can be used to relay the path of
light from the display to the user's eye. This means that the display does not have
to be located on the user's forehead which greatly helps in distributing the weight of
the HUD components around the center of gravity of the user's head [95]. Because
of the more advanced optical design and additional lenses, chromatic and geometric
lens aberrations are common [26].

The majority of the off-the-shelf commercially available HMDs used to achieve the
results published in the conferences and journals covered in this review are non-pupil-
forming and thus only use magnifying, and possibly collimating, optics. Pupil-forming
optics are more expensive to produce and are usually only researched for military
applications. These research results are typically published at conferences organized
by such bodies as the Society of Photo-Optical Instrumentation Engineers (SPIE) and
Human Factors and Ergonomics Society (HFES) rather than the ones typically aimed
at VR and AR.
3.2. THE DISPLAY

3.2.2 The Video See-Through Misnomer

Another important difference within HMDs is whether the optical path is direct or indirect. Sutherland’s original HMD used a half-silvered mirror as an optical combiner to directly merge the user’s view of the surrounding world with the magnified image of a CRT display [123]. While some developers of early AR systems followed this principle [29][27], others chose instead to relay a camera feed to indirectly merge computer graphics with the surrounding world [6]. The principal difference between the two systems is the path along which light rays travel from the surrounding world to the user’s eye. In the OST system, the user observes the surrounding world directly via a set of lenses and mirrors which merely bend and reflect the light rays. In the Video See-Through (VST) system, on the other hand, the user observes the world indirectly via a video representation of the surrounding world which has been sampled with the capabilities of a camera. Consequently, referring to VST as being see-through when it is in fact not transparent is not correct, but the term to denote video-based AR systems still remains. The nomenclature non-see-through (Type I) and see-through (Type II) has been suggested (p. 182 [7]), but it is not widely used.

It is important to distinguish between the two types of HMDs as they are quite different in several regards, especially in terms of the importance of eyepoint location and accommodation as mentioned in the following section. There are also important differences in during the calibration procedure as discussed in section 5.2.

3.2.3 Displaying for the Human Eye

In addition to transparency, OST and VST differ in a number of additional key aspects [4]. This has consequences for design variables, both due to human factors and also for simple practical reasons.

- **Eyepoint location**: The need for accurately estimated eyepoint location.
- **Accommodative demand**: The user’s need to adjust the focal power of the lens.
- **Vergence**: The user’s ability to rotate the eyeball.
- **Color perception**: The user’s sensitivity to color.
- **Resolution and Contrast**: The systems’ ability to match the user’s visual acuity.

Eyepoint location: For physical reasons, a camera cannot be placed in the true location of the user’s eye. Subsequently, the user of a VST system will observe the world from an unnatural eyepoint. As a consequence, with vision and proprioception decoupled, the user depends on sensorimotor adaptation to maintain dexterity for intricate tasks [110]. Jannick Rolland, Richard Holloway, and Henry Fuchs have suggested a

---

27

---

5The sensorimotor (or sensory motor) system fuses information from our senses and applies this information to plan and execute appropriate motor responses. One example is hand eye coordination.
way to fuse mismatching viewpoints, but the problem of mismatching perspective and FOVs still remains [112]. In favor of VST, however, Rolland et al. mention that VST systems are easier to calibrate. Generally speaking, in a VST system, the optical axis of the camera and the user's eye do not have to align for overlaid graphics to register properly. A user can look at a display from an oblique angle and still observe the world along the center line of the camera lens. This relaxes the requirement of a correctly positioned viewport. In other words, an VST system has two independent frusta4 - one for the eye, and one for the camera - and they need not align for correct registration. In an OST system, however, the system works differently. The user's eye and the optics of the viewport have two separate frusta. However, the two frusta are interdependent, which requires the apices5, FOVs, and optical axes of the two frusta to be matched in order to achieve good registration. Thus, the need for an accurately estimated eyepoint is greater with OST systems.

As an additional challenge, researchers are not certain of which anatomical point of the eye to use as exact eyepoint location when calibrating an OST system. The eye’s nodal point [109][37][78], pupil entrance [111], and center of rotation [72], have all been suggested as possible candidates for eyepoint location (that is frustum apex, a.k.a. center or projection) in various projection models. Distinguishing between the three locations requires an eyepoint estimation accuracy of 3 mm or less according to a Gullstrand schematic eye [117], an eye model based on a large number of human eye measurements collected by the Swedish ophthalmologist Allvar Gullstrand.

**Accommodative demand:** In a VST system the user accommodates the lens in the eye to focus on a *virtual image* of a display surface which presents an image captured by a camera. In turn, the camera’s system of lenses ensures objects of interest over varying depth are brought into focus in the displayed camera image. With the frusta of the eye and the camera separate, the task of bringing objects into focus primarily resting with the camera’s lens as the distance between the user's eye and the display seldom changes in a HMD system. This is not the case in an OST system in which the frustum of the user's eye and the frustum of the optical system are dependent. In this case, it is, instead, the user who must adjust focus to match the distance to the object of interest. However, since the lens in the user's eye only can be accommodated to one focal length at a time, the object and the overlaid computer graphics must appear within roughly the same depth of field or else object or graphics will appear blurry to the user. Some HMD designs assume that keeping the lens of the eye in a relaxed state is the best approach, but this only works for applications where objects of interest appear at some distance away (p. 516 [103]). For a desktop solution, depth of field of the computer graphics must be more carefully tuned. While the tolerable discrepancy of depth of field between virtual and real objects is individual, and also changes by 0.12 diopter per 1 mm pupil dilation [26], as a general rule, it can be said that the optical power, $f$, required to fuse objects appearing at distances

---

4 A frustum is the pyramid or cone which approximates the field of view. Plural: frusta.
5 An apex is the uppermost point of a frustum. Plural: Apices.
3.2. THE DISPLAY

\( d_{\text{near}} \) and \( d_{\text{far}} \) meters away from the user should in practice not exceed 0.25 diopters \([109]\) according to the Gaussian lens equation (p. 117 \([133]\)), see eq. 3.1. This is the increment with which optometrists prescribe corrective lenses.

\[
f = \left| \frac{1}{d_{\text{far}}} - \frac{1}{d_{\text{near}}} \right| \tag{3.1}
\]

Thus, in a VST system, the object of interest and computer graphics always appear at the same depth of field, while in OST it is possible to experience different depth of field between objects of interest and computer graphics, particularly when objects are closer than 6 m.

Vergence: In the previous section on accommodative demand, Ozan Cakmakci and Jannick Rolland cites sources suggesting that depth of field should be matched to within 0.1 diopters to avoid discomfort \([26]\). The discomfort is mainly caused by the accommodation/convergence conflict which occurs when there is a discrepancy between two physiological depth cues, lens accommodation and eye convergence, which are normally used and interpreted together to estimate object distance (p. 517 \([103]\])\([35]\). The horizontal vergence angle measures the eye’s inward (towards the nose) rotation relative the eye’s optical axes when looking straight ahead. It varies as the user fixates on objects at varying distance and also depends on the user’s IPD \([109]\). The eyes can also rotate vertically, but since the eyes are not separated along this direction, difference in vertical angle is not used as a depth cue. As a consequence, in stereoscopic applications, users are likely to experience diplopia (double vision) when viewports are not vertically level (p. 255 \([99]\]) \([88]\). Herschel Self of Armstrong Aerospace Medical Research Laboratory cites tolerance levels as low as 3.4 arcmin, but notes himself that it may be too conservative a measure \([114]\). The military standard MIL-V-43511C specifies tolerable vertical prismatic deviation in flight helmet visors to be 0.18 diopters (6.2 arcmin). In conclusion, moderate amounts of horizontal vergence can be used as a design feature to create a perception of depth, particularly in OST systems, but vertical vergence should always be avoided.

Color perception: Mark Livingston made a study in which users were asked to identify colors displayed in OST HUDs \([87]\). Although varying combiner transmittance, different Liquid Crystal Display (LCD) backlight, and changing ambient light conditions affect the perception of color, no standardization exists. Therefore lighting condition was not controlled as an experiment variable in the study. However, in constant surrounding light, the study showed that color identification in HMDs is not particularly reliable, especially around the gamut of perceived blue. Livingston mentions in the discussion that color identification may be related to perceived intensity. In the recent book from USAARL, Rash et al. report that the spectral response of the human eye peaks at 507 nm (night) and 555 nm (day) (p. 138, fig. 4-19 \([103]\)). These wavelengths correspond to shades of green. Rash et al. also note that in addition to improved human perception, maximum perceived intensity include additional
benefits such as lower power consumption and less attenuation in relaying optics. Moreover, the superfluous red and blue LEDs can be replaced with two more greens, further increasing luminance. Lastly, monochromatic displays simplify correction of chromatic lens aberration [26].

Resolution and Contrast: A display is a two-dimensional array of pixels with varying brightness in each element. The number of pixels along the horizontal and vertical direction defines what most people would refer to as the display's resolution. However, given the distance from which the screen is viewed, each pixel subtends a visual angle, effectively changing the apparent pixel size. This measurement is referred to as spatial resolution and is usually very small, hence often referred to in arc minutes (arcmin) corresponding to 1/60 of a degree. This is a convenient unit because the limit of human visual acuity is often approximated to 1 arcmin (p. 46 [133]). For instance, a standard Snellen visual acuity chart is designed such that the lines and spaces in letters on the 20/20 line subtend 1 arcmin when read from 20 ft distance (p. 47 [133]) [87]. However, 1 arcmin should only be used as a rough guideline since the performance of human vision is dependent on several additional factors because, even though the cones in the human eye are spaced 2.5 µm apart (corresponding to 0.5 arcmin), it is possible for humans to distinguish patterns with spatial frequencies 1.5 times higher than the nominal Nyquist frequency of the cone spacing [26].

For instance, the perceived difference in brightness between object and object background has a great impact on visual resolution. Therefore, as anyone who has attempted stargazing in central Paris will tell you, visual acuity is governed by contrast. Luminance is an objective measurement of brightness and is defined as luminous intensity per unit area (1 m²) in a given direction, and its effect on HMD optics has been treated in great detail by Cakmakci and Rolland [26]. Relying on the ISO 2720:1974 standard, luminance, \( L \), can easily be approximated using the exposure equations of a digital Single Lens Reflex (SLR) camera, provided that the relative aperture (f-number), \( N \) the exposure time, \( t \) (in seconds), and ISO speed, \( S \), are known. The value for the reflected-light meter calibration constant, \( K \), should range from 10.6-13.4 depending on the brand of camera, see eq. 3.2, [100].

\[
L = \frac{K \cdot N^2}{t \cdot S}
\]

By sampling luminance over the display surface while the HMD is directed towards ambient background light, \( L_a \), representative of the conditions in which the AR system is going to be used, and also sampling the display surface when full white graphics is displayed, \( L_w \), one can calculate the Michelson contrast, see eq. 3.3, (p. 186 [7]). This piece of descriptive information is useful to include together with spatial resolution when reporting results such that other researchers can recreate the experiment under the same conditions (p. 72 [133]).
While the resolution and contrast of the display can be accurately measured, these numbers are not a true representation of the world surrounding the user of an HMD. In the case of VST, both resolution and contrast have been subsampled as the camera is an inferior sensor to the human eye (although how much worse has not been experimentally quantified [87]). The resulting loss of detail affects object detection and also has an impact on depth judgment [82]. While the direct optical path in OST systems does not restrict resolution, contrast is still worsened by half-silvered mirrors which sometimes reduce transmission by 70% [4], or in the case of fighter pilots, sometimes 85% (p. 821 [103]).

These are some of the ways in which the designers of OST HMDs try to match an AR system to the capabilities of the human eye. Most parameters are coarse approximations that achieve acceptable results. For instance, the exact eyepoint location is still an open question [110]. Assuming the center of rotation, Holloway argues that eyetracking is not necessary to follow gaze direction [72], but if research should show the nodal point is a better estimate, as claimed by Robinett and Rolland [109], then the projection changes as the eye rotates and tracking would be necessary for perfect adherence to the projection model. This is, however, only an issue if the user rotates the eye to explore the environment. Consequently, widening the FOV may imply the need for eyetracking. This is an example of how bridging one limitation creates another and how most systems are based of informed compromises.

### 3.3 The User

The optical system of an OST HMD is designed to suit the human eye. Because of the intimate link between display specifications and performance levels of the human eye, some characteristics on human perception have already been discussed in the previous subsection on the display subsystem. Therefore, this subsection will not describe human perception to any greater extent, but rather focus on the user in the role as platform the from which the tracker and display operate.

This subsection also briefly explains the mechanics of head rotation; why some symbology is not suitable for calibration exercises; and how the peripheral view has an impact on postural stability.

#### 3.3.1 Postural Sway

The human balance system consists of the vestibular, visual, and proprioceptive senses, which register body sway away from the vertical and issue compensatory muscu-
lar adjustments [85]. The prevailing notion of the human balance system being a closed-loop system was challenged by James Collins and Carlo De Luca who in 1994 suggested that the balance system instead works as an intermittent closed-loop model and that the seemingly random onset muscular compensations result in an erratic postural sway of about 2 cm amplitude during quiet standing, even in otherwise healthy humans [32]. The postural sway has been matched to a so called inverted pendulum model [58] which is a bit of a misnomer as the movement is not periodic. There have also been contradictory reports as to whether the articulated body parts indeed sway in unison. Keshner and Kenyon suggest that the subjects’ trunk and head can be affected separately through visual stimuli [80]. Furthermore they challenge the notion that inputs from retinal, vestibular, and proprioceptive senses are weighted according to which task that is currently performed and suggest that the reweighting scheme is individual rather than task specific [80].

There are several factors that could affect postural sway. Controlled visual stimuli can bias the perceived vertical and cause a subject to lean in a specific direction away from upright posture [145][80]. In fact, some types of optical flow/vection are compelling enough to induce complete loss of balance [85]. Roll and pitch are particularly effective, regardless of whether stimuli consists of random dots or a realistic environment, and stereo vision seems to amplify the impression [80]. Uncontrolled stimuli, however, may produce conflicting reports between the visual and vestibular senses which can cause balance to deteriorate and result in cybersickness [118][8]. Experiments with varying FOV suggest that it is the peripheral view which is responsible for spatial orientation, sense of locomotion, and posture [39]. Without diminishing the importance of the peripheral view, it has also been shown that motion parallax between foveated objects contribute to postural stability [61]. In addition to perceptual factors, cognitive load have also been shown to affect balance [147].

Postural sway is normally quantified with the Romberg coefficient which compares the ratio between the length of the sway path in a subject with open eyes to the sway path with eyes closed (which normally is about 40% longer) [40].

### 3.3.2 Head-Aiming Performance

The accuracy and precision of visual alignments using head rotations has previously been investigated for Head-Mounted Sights (HMS) which is a form of VCS normally consisting of only one collimated reticle. In 1987 Maxwell Wells and Michael Griffin performed a study in which the average head pointing precision was estimated to 0.14° visual angle for seated subjects with a collimated display aiming at targets at 0.9-1.5 m distance [142]. Where it was possible to determine, due to varying analysis and experimental design, the precision in the 13 works cited by Wells and Griffin varied between 0.13-0.8°.

The human head pitches on a joint formed by protrusions on the underside of the
3.3. THE USER

skull’s occipital bone, called occipital condyles, and the superior facets on the topmost vertebra (a.k.a Atlas or C1) of the spine (p. 827 [103]). Since the joint is located behind the head’s center of mass, the neck extensor muscles must be constantly engaged to prevent the head from pitching forward (p.151 [96]). On the second vertebra (a.k.a. Axis or C2) a vertical protrusion called the odontoid process fits into the first vertebra and provides articulation of the head in the yaw direction [84]. This means that the head does not articulate around a single pivot point, but instead can be categorized as a Fick gimbal, that is a horizontal axis mounted on a vertical axis [84]. Depending on the activity performed by the user, the head articulates differently. An aiming task performed with access to the peripheral view follows Listing’s law, while aiming with restricted FOV results in head motions best described by Donder’s law [28]. This may explain the effect of aiming angle on aiming precision in the Wells and Griffin study where aiming at extreme yaw angles had a pronounced vertical spread, and extreme pitch angles resulted in obvious horizontal spread [142].

3.3.3 The Human Eye During Fixation

When a user fixates a target using some static HMD symbology, for example a crosshair, the constant visual stimulus has a tendency to fall on the same location on the user’s retina. The static image will eventually cause overstimulation of the cones and rods which temporarily desensitizes them and causes the perception of the stimuli to fade. This phenomenon is known as the Ditchburn-Riggs effect and occurs because the eye is not moving sufficiently to maintain sensitivity in the retina [38] [106].

In studies of humans fixating on stationary targets, three types of eye movements have been identified [31]: Small saccades in the range of 2-12 arcmin, lasting for 0.2-10 s, are defined as microsaccades. Between saccades, the eye also exhibits drifts in the range of 1.5-4 arcmin at velocities of 4 arcmin/s. Moreover, the eye also has a tremor causing it to vibrate with 200 Hz at amplitudes of 5-30 arcsec. While slightly larger saccades are used to refoveate visual stimuli during fixations, no evidence has been found to prove that microsaccades play a functional role in maintaining sensitivity in the retina [31].

A similar case of desensitizing the retina, referred to as Troxler fading, has been observed for low contrast eccentric (> 3°) stimuli during prolonged fixation (> 7 s) [127] [31]. While the subject focuses on static symbology (e.g. crosshair), the peripheral view is desensitized which causes the perception of stationary low contrast objects in the outer parts of the subject’s FOV to fade and disappear, even if the low contrast objects flicker and blink.

As seen in the sections above, the human is quite a dynamic platform from which to perform a calibration. There seems to be no general way in which to model postural

\[1\text{arcmin} = 1/60 = 0.01667^\circ \quad 1\text{arcsec} = 1/3600 = 0.00027^\circ\]
sway, aiming performance seems task dependent, and the eye can become desensi-
tized and is highly dependent on surrounding conditions. To establish whether this
platform is stable enough to gather measurements needed to achieve a successful cal-
ibration, the tolerances for noise have to be studied. The following chapter discusses
the effects of measurement noise and ways to improve manual measurements made
by a human operator.
Chapter 4

Calibration Theory

This chapter explains the principals of the most commonly used calibration model, namely the pinhole camera, and some theoretical concepts supporting the major calibration techniques. We start by deriving the camera matrix to understand the eleven parameters that are involved and explain how these parameters can be estimated linearly. The estimation technique depends on a system of linear equations which are sensitive to where and how measurements were made. Therefore, the following subsections are concerned with methods on how to reduce and correct the effects of measurement noise. We also study the expected parameter variance as a function of camera calibration technique and noise, and close by presenting the more exotic asymmetric camera model.

4.1 The Pinhole Camera Model

The introductory chapter in the Computer Vision book written by Boguslaw Cyganek and Paul Siebert contains a retrospect which traces the historical origins of vision research. It mentions how the camera obscura, also known as the pinhole camera, was first used by the arabic scientist Alhazen (a.k.a. Muhammed ibn al-Hasan) as an analogy to the human eye around 1000 AD (p. 9 [36]). While a pinhole camera is a very crude approximation to the human eye, since it lacks a lens and has a fixed aperture, it makes for a simple computational model to estimate perspective projection as all relationships can be made linear.

Represented by linear algebra, the mathematical fundamentals of computer graphics are linear too. To allow objects to be scaled, rotated, but also translated in three dimensions with one single operation, a transformation matrix is augmented with one extra row and column, using a system called homogeneous coordinates (p. 3 [137]). The ability to include translation along with scaling and rotation into an aggregated transformation matrix is however not the only argument to use 4-by-4 matrices. As
shown here below, the homogeneous coordinates can also be used to represent the foreshortening in a perspective view. This has been incorporated in the virtual camera used as an analogy for the human eye in computer graphics (p. 144 [137]).

In a pinhole camera, light rays enter through a small aperture and form an image on the far camera wall called the image plane (or sometimes retinal plane). Similar to the human eye, the image is formed upside down. In a virtual camera used for computations, the image plane has been rotated 180° around the vertical axis, and 180° around the optical axis, which leaves the image plane in front of the aperture, but now with an image oriented the same way as the original image (p. 154 fig. 6.1 [65]). This setup is physically impossible, and this construction is therefore referred to as the virtual image plane. However, it further facilitates the perspective calculations in the pinhole camera model as the height of the imaged object and the image on the virtual image plane can now be described as proportional triangles. This leads to the very foundation of the pinhole camera model, shown in eq. 4.1. $x, y$ and $z$ are coordinates of the image on the image plane, $X, Y$ and $Z$ are world coordinates of the imaged object, $f$ is the focal length and coincides with the distance to the image plane. Point C denotes the camera’s center of projection. See fig. 4.1.

**Figure 4.1:** The proportional relationship between world coordinates, $X Y Z$, and screen coordinates, $x y z$.

$$x = f \frac{X}{Z}, y = f \frac{Y}{Z}, z = f$$  \hspace{1cm} (4.1)

This is the point in the literature from which projective geometry and perspective representation through homogeneous coordinates starts to become useful, but exactly how the homogeneous coordinate works as a perspective divide is seldom very well explained. Olivier Faugeras introduces projective geometry in a quite abstract manner before the reader realizes its use in a pinhole camera model [44], while Emanuele Trucco and Alessandro Verri make use of the homogeneous coordinates directly and leave the projective geometry in the appendix [128]. Richard Hartley and Andrew Zisserman present projective geometry with the help of discussions on the duality principle, showing how the homogeneous coordinates can be interpreted to represent
both a point and a line [65]. Ma et al. also start pragmatically by applying homogeneous coordinates directly and save the elaboration on projective geometry and the duality of homogeneous coordinates for later [91]. This is also the case of Boguslaw Cyganek and Paul Siebert [36]. Thus, in the opinion of the author, the text of Trucco and Verri is the most pragmatic and accessible one, while Hartley and Zisserman give better support for a mental model in that they, chapter by chapter, build on incremental theoretical foundations. Among the more mathematically abstract texts in Faugeras, Cyganek and Sibert, and Ma et al., Faugeras is a bit more extensive and does the better job in introducing the concept of projective geometry. The others are however useful references and alternate explanations are always helpful.

The duality principle means that the dimensionality of an object in Euclidean space $\mathbb{R}^3$ coexists with an object of one dimensionality lower in projective space, $\mathbb{P}^2$, and that the interpretation changes depending on whether $\mathbb{R}^3$ or $\mathbb{P}^2$ is used as reference (p. 30 [65]). A point on the image plane in $\mathbb{P}^2$ is really the image of a line connecting the camera’s center of projection and some point in $\mathbb{R}^3$. A line on the image plane in $\mathbb{P}^2$ is really a plane in $\mathbb{P}^3$ passing through the center of projection. More specifically, take the vector $[XYZ]$ in ordinary (inhomogeneous) Euclidean coordinate space, $\mathbb{R}^3$, to express a point. It can also be expressed in homogeneous coordinates in $\mathbb{R}^3$ by adding the homogeneous coordinate $H$ like this $[XYZH]$. Assuming $H$ is set to 1, dividing the homogeneous vector by its homogeneous coordinate $H$, it can be converted back into an inhomogeneous Euclidean vector again representing a point $[X/HY/HZ/H]$. However, if $H$ is set to 0, the inhomogeneous Euclidean vector will instead represent a point at infinity (since division by zero $\rightarrow \infty$). Thus, choosing a value of $H$ between 0 and 1 in homogeneous coordinates enables the dual notation of letting the $[XYZH]$ express the point $[XYZ]$, the point at infinity, or any point along the line connecting the two. This property can be directly applied to the proportional triangles of eq. 4.1 by first working in homogeneous coordinates in $\mathbb{R}^3$ to establish the perspective lines, as shown in eq. 4.2.

A word on notation: In eq. 4.2, $X Y Z$ still correspond to the world coordinates of the imaged object, and $x y$ still denotes the coordinates on the image plane. However, in the following equations, the geometric interpretation of image plane coordinates will result in an increasingly more complicated expression. Therefore, $u$ and $v$ are introduced to denote image plane coordinates with the geometric expressions implied. $w$ is the homogeneous coordinate of the image plane.

$$
\begin{bmatrix}
  u \\
  v \\
  1
\end{bmatrix}
= Z
\begin{bmatrix}
  x \\
  y \\
  1
\end{bmatrix}
= \begin{bmatrix}
  fX \\
  fY \\
  Z
\end{bmatrix}
= \begin{bmatrix}
  f & 0 & 0 & 0 \\
  0 & f & 0 & 0 \\
  0 & 0 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
  X \\
  Y \\
  Z \\
  1
\end{bmatrix}
$$

So far, the relative perspective relationship between an object in the world and its image on the image plane has been described. However, for practical application,
the absolute relationship between the two coordinate systems must be established. For instance, to express the image plane coordinates in pixel units, the perspective matrix must be multiplied by a scale, $m$, representing the size of a pixel in world units (usually meters). Since the HMD optics produce a virtual image (see section 3.2), the pixel size must be based on the spatial resolution. Assuming an OST HMD, a relatively easy method is to display a line of computer graphics which is 100 px long and match its length to a real world object of known length at known distance. Effectively, this measurement reports the number of pixels per degree of FOV. Without exact knowledge of the optics, or if uncertain of the quality of its mechanical calibration, this measurement specifies pixel size in units of px/m for a screen at any focal length. As a consequence, focal length $f_m$ and $f_m$ in eq. 4.3 will be given in pixels as $f$ was given in meters originally.

Moreover, the image coordinates may not originate from the image center, also known as the principal point. To adjust for this, horizontal $p_x$ and vertical $p_y$ offsets, specified in pixels, can be incorporated into the projection matrix as shown in eq. 4.3.

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} x m_x + p_x \\ y m_y + p_y \\ 1 \end{bmatrix} = \begin{bmatrix} f m_x X + Z p_x \\ f m_y Y + Z p_y \\ 0 \end{bmatrix} = \begin{bmatrix} f m_x & 0 & p_x & 0 \\ 0 & f m_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

(eq. 4.3)

For added generality, but of limited practical use in computer graphics and a frequent source of confusion (as will be shown later), the perpendicularity of axes in the image plane can be modeled as a skew parameter $s$. This completes the matrix of intrinsic parameters which governs the internal workings of the pinhole camera (eq. 4.4) usually denoted $K$ (eq. 4.5). For completeness, it should be noted that Hartley and Zisserman model skew as a contribution to $K_{12}$ (p. 157 [65]) while Faugeras suggests contributions to $K_{12}$ as well as $K_{22}$ (p. 57 [44]).

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} x m_x + s \frac{Z}{f} + p_x \\ y m_y + p_y \\ 1 \end{bmatrix} = \begin{bmatrix} f m_x X + Y s + Z p_x \\ f m_y Y + Z p_y \\ 0 \end{bmatrix} = \begin{bmatrix} f m_x & s & p_x & 0 \\ 0 & f m_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

(eq. 4.4)

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

(eq. 4.5)
4.2. PARAMETER ESTIMATION

The derivation shown here is included in all textbooks mentioned in this section, but in slightly various forms, see (p. 157 [65]) (p. 44 [44]) (p. 54 [91]) for some good examples. It can also be found in a journal paper by Tuceryan et al. (§4.2. [131]). The derivation is included in this thesis to provide context for the following chapters, and because this thesis emphasizes the importance of formulating results in camera parameters as explained in section 5.

The next step is to account for the camera’s rotation and translation relative the surrounding world, referred to as the extrinsic parameters, as they govern the camera’s external workings. For some, it is easier to imagine the opposite, namely the world rotating and translating into place in front of the fixed camera. This is done by applying a 4-by-4 transformation matrix to the point in the world (eq. 4.6). At this stage, for clarity, \( u \) and \( v \) replace the image coordinates which contain more intricate expressions of \( x \) and \( y \). If needed, Maple or Matlab symbol toolbox can be used to expand the expressions.

\[
\begin{pmatrix}
u \\ v \\ 1 
\end{pmatrix} =
\begin{pmatrix}
f m_x & s & p_x & 0 \\ 0 & f m_y & p_y & 0 \\ 0 & 0 & 1 & 0 
\end{pmatrix}
\begin{pmatrix}
r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 
\end{pmatrix}
\begin{pmatrix}
X \\ Y \\ Z \\ 1 
\end{pmatrix}
\]

(eq. 4.6)

In eq. 4.7 the fourth column of zeros in \( K \) is removed along with the last row of the transformation matrix, still expressing the same transformations, which are further simplified in eq. 4.8 by letting \( K[R|t] = P \), henceforth referring to the 3-by-4 matrix \( P \) as the camera matrix.

\[
\begin{pmatrix}
u \\ v \\ 1 
\end{pmatrix} =
\begin{pmatrix}
f m_x & s & p_x \\ 0 & f m_y & p_y \\ 0 & 0 & 1 
\end{pmatrix}
\begin{pmatrix}
r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} 
\end{pmatrix}
\begin{pmatrix}
X \\ Y \\ Z \\ 1 
\end{pmatrix}
\]

(eq. 4.7)

\[
\begin{pmatrix}
u \\ v \\ 1 
\end{pmatrix} = K \begin{bmatrix} R & t \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 
\end{pmatrix}
\]

(eq. 4.8)

4.2 Parameter Estimation

Instead of explicitly setting the parameters of a pinhole camera model to achieve a perspective image, it is also possible to do the inverse, namely take a perspective image and estimate the parameters that went into creating such an image. This is
known as camera resectioning, computation of the camera matrix, parameter estimation, or simply calibration. In essence, it aims to calculate the mapping between a set of points in the world and the corresponding set of points on the image plane. These two sets are known as correspondence points. Some calculations stop at the 3-by-4 camera matrix mapping the projection, while others continue and decompose (or factorize) the camera matrix into parameters that have the physical meanings of the pinhole camera model described in the previous section.

In the exceptionally clear description in the appendix to one of Sutherland’s papers (on a topic seemingly unrelated to AR) [124], a very general outline for calculating any transformation can be found, including the 3D-to-2D projective transform of a camera matrix, see eq. 4.9. (It should however be noted that the following procedure can be adjusted to fit both 2D-to-2D and 3D-to-3D homographies and 3D-to-2D projective transforms (p. 44, 78 [65]), and can be solved with a null space approximation for homogeneous equation systems and pseudoinverse for inhomogeneous equation systems (p. 585, 590-596 [65]).)

\[
\begin{bmatrix}
    u \\
    v \\
    1
\end{bmatrix} =
\begin{bmatrix}
    wu \\
    wv \\
    w
\end{bmatrix} =
\begin{bmatrix}
    p_{11} & p_{12} & p_{13} & p_{14} \\
    p_{21} & p_{22} & p_{23} & p_{24} \\
    p_{31} & p_{32} & p_{33} & p_{34}
\end{bmatrix}
\begin{bmatrix}
    X \\
    Y \\
    Z \\
    1
\end{bmatrix}
\]  

\begin{align}
wu &= Xp_{11} + Yp_{12} + Zp_{13} + p_{14} \\
wv &= Xp_{21} + Yp_{22} + Zp_{23} + p_{24} \\
w &= Xp_{31} + Yp_{32} + Zp_{33} + p_{34}
\end{align}

Substituting eq. 4.12 into eqs. 4.10 and 4.11 and sorting the terms gives the following expression:

\begin{align}
X(p_{11} - p_{31}u) + Y(p_{12} - p_{32}u) + Z(p_{13} - p_{33}u) + (p_{14} - p_{34}u) &= 0 \\
X(p_{21} - p_{31}v) + Y(p_{22} - p_{32}v) + Z(p_{23} - p_{33}v) + (p_{24} - p_{34}v) &= 0
\end{align}

At this point, the eqs. 4.13 and 4.14 can be rearranged and used in three ways, depending on which information is available. If \( P \), and \( X Y Z \), are known, the equations straightforwardly describe a projection of a world point to a point on the image plane. As a second case, if the camera’s center of projection is known, along with \( P \), and \( u \) and \( v \), the equations can be interpreted as planes and be used to back-project an image point into a ray. This form is useful when studying the importance of correspondence point distributions using simulations, as mentioned in the last paragraph of section 5.1.2. However, in the case of two sets of correspondence points, looking
to estimate the camera matrix, $P$, eqs. 4.13 and 4.14 need to be expanded into the system of linear equations in eq. 4.15.

\[
\begin{bmatrix}
X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
- X_1 u_1 & - Y_1 u_1 & - Z_1 u_1 & - u_1
\end{bmatrix}
\begin{bmatrix}
X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
- X_i u_i & - Y_i u_i & - Z_i u_i & - u_i
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & - X_1 v_1 & - Y_1 v_1 & - Z_1 v_1 & - v_1
\end{bmatrix}
\begin{bmatrix}
0 & 0 & 0 & X_i & Y_i & Z_i & 1 & - X_i v_i & - Y_i v_i & - Z_i v_i & - v_i
\end{bmatrix}
\begin{bmatrix}
p_{11} \\
p_{12} \\
p_{13} \\
p_{14} \\
p_{21} \\
p_{22} \\
p_{23} \\
p_{24} \\
p_{31} \\
p_{32} \\
p_{33} \\
p_{34}
\end{bmatrix}
= 0
\]

Denoting the 2i-by-12 matrix as $A$ and the 3-by-4 camera matrix $P$ in vector form as $p$, the eq. 4.15 can be expressed more compactly as shown in eq. 4.16.

\[
Ap = 0
\]

Although the camera matrix $p$ seems to have 12 unknowns, in fact the camera model only has 11 DOF. The 12th unknown is an arbitrary matrix multiplication factor, sometimes referred to as scale (not be confused with the scale of any coordinate system like e.g. $m$ in eq. 4.3). The multiplication factor comes from the fact that the linear equation system is built using homogeneous coordinates. All multiples of a homogeneous coordinate define the same line. \([1/2 3 1]\) is equivalent to \([2/4 6 2]\) because \([1/2 3 1] = [2/4 6 2]\). Sometimes this is written \([1/2 3 1] \sim [2/4 6 2]\), which means “equivalent up to a scale”. With 11 unknowns, five and a half points are needed since each correspondence point result in two equations, see eqs. 4.10 and 4.11. Thus, in practice, six points have to be collected to provide the 12 lines of equations needed to solve for the 11 unknown parameters.

If any additional facts are known about the camera, such as use of weak perspective or orthographic projection models (p. 44, 171 [65]), such simplifications should be incorporated at this stage, moving some elements over to the right hand side resulting in an inhomogeneous equation and shrinking $A$ accordingly. This would also reduce the minimum number of correspondence points required, denoted by subscript $i$ of the elements of $A$ in eq. 4.15.

If six points are provided, they must not be linear combinations of each other. In other words they must not lie along the same line as seen from the camera’s center of projection because, if they do, they will not add any additional information about
CHAPTER 4. CALIBRATION THEORY

the camera rotation, translation, scale or perspective. Nor should they lie in such a configuration that any of them form, or are close to forming, a plane (more on this in section 4.2.4 on Degenerate Configurations). One would therefore expect $A$ in 4.16 to have rank 12 but, remembering that all lines are defined up to a scale leaving scale as a free variable, the rank will effectively be 11, and the exact solution is the vector $p$ corresponding to the one-dimensional right null space of $A$. In practice, however, correspondence points usually contain measurement noise, and therefore no exact solution exists, which results in $A$ having rank 12 and no null space (p. 58 [44]) (p. 179 [65]). A null space can however be approximated using the Singular Value Decomposition (SVD) technique. This method is known as the Direct Linear Transformation (DLT), or simply finding the linear solution.

With SVD, $A$ (i-by-j) is factorized into two orthonormal bases $U$ (i-by-j) and $V$ (i-by-i), representing row space and column space, respectively. The diagonal matrix $\Sigma$ (i-by-i) expresses the scaling between $U$ and $V$ (eq. 4.17). Properties of orthogonal matrices, $V^T = V^{-1}$, and $U^T U^{-1} = I$ (eqs. 4.19 and 4.20) lead to the expression in eq. 4.21 where the $\sigma_i^2$ diagonal elements in $\Sigma$ represent the eigenvalues corresponding to the $i$ columns of $V$ which are the eigenvectors of $AA^T$. $\Sigma$ will have the same rank as $A$, which implies that in the case of an exact solution the eigenvalue $\sigma_i^2$ will be 0.

$$AV = U\Sigma \tag{4.17}$$
$$A = U\Sigma V^{-1} \tag{4.18}$$
$$A = U\Sigma V^T \tag{4.19}$$
$$A^T A = V \Sigma^T U^T U \Sigma V^T \tag{4.20}$$
$$A^T A = V \begin{bmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sigma_i^2 \end{bmatrix} V^T \tag{4.21}$$

The vector in $V$ that changes the least (usually the $i^{th}$ as $\Sigma$ by convention usually is sorted in descending order), is the closest approximation to the null space, and subsequently the best solution for $p$ [64] (p. 90, 179, 585 [65]) and is converted back into $P$ on matrix form (eq. 4.22) according to the original element order in eq. 4.16.

$$P = \begin{bmatrix} V_{1i} & V_{2i} & V_{3i} & V_{4i} \\ V_{5i} & V_{6i} & V_{7i} & V_{8i} \\ V_{9i} & V_{10i} & V_{11i} & V_{12i} \end{bmatrix} \tag{4.22}$$

42
4.2. PARAMETER ESTIMATION

4.2.1 Isotropic Scaling

Regardless of the technique used, there will always be some amount of noise in the measurements of the correspondence points. Let the measurement error be $\epsilon_w$ and $\epsilon_i$ for the world and image plane coordinates, respectively, and assume that they affect all dimensions equally. Using eq. 4.15, the errors will distribute according to eq. 4.23. It is clear to see that camera matrix elements $p_{14}$ and $p_{24}$ are not affected by noise at all, while $p_{31}$, $p_{32}$, and $p_{33}$ will be affected to a much greater extent.

Moreover, the world coordinates are given in meters and image plane coordinates are given in pixels. Specified in these units, the two measurements can differ by two to four orders of magnitude. This means that a relatively small error in pixel image coordinates will have a great impact when multiplied by measurements defined in meters, as is the case for $p_{31}$, $p_{32}$, $p_{33}$. For this reason, Hartley suggested the world and image plane measurements be scaled to an average unit distance from image and world origin, respectively, such that the measurements still preserve their internal relationships \((p. 107 [65])\).

This procedure is called (Hartley) normalization and is effectively a matrix conditioning technique. It can be shown that the condition number, that is the largest over the second to smallest value in $\Sigma$ in 4.21, $(\sigma_1/\sigma_{r-1})$, decreases, indicating that the problem becomes numerically stable after conditioning \([64]\). This effectively means that noisy measurements become less amplified and do not affect the final solution as much. The reason why $\sigma_{r-1}$ is used instead of the customary $\sigma_1$ is to take into account the theoretical case of an exact solution, in which case $\sigma_1$ is 0.

\[
\begin{bmatrix}
\epsilon_w & \epsilon_w & 1 & 0 & 0 & 0 & 0 & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_i \\
0 & 0 & 0 & 0 & \epsilon_w & \epsilon_w & 1 & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_i \\
0 & 0 & 0 & 0 & 0 & \epsilon_w & \epsilon_w & 1 & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_w \epsilon_i & -\epsilon_i \\
\end{bmatrix} = 0
\]

\((4.23)\)
CHAPTER 4. CALIBRATION THEORY

4.2.2 Non-Linear Optimization

Since the DLT minimizes the algebraic distance between correspondence points, this method does not degrade gracefully with increasingly noisy measurements. Instead, the linear solution is often used as an approximation to the camera matrix which is further refined using an iterative approach in which a non-linear optimization algorithm sets out to minimize some target function. Such a target function can be the reprojection error, that is the Euclidean distance between a point on the image plane and its corresponding point in the world as it is projected onto the image plane with a tentative camera matrix.

The optimization algorithm of choice for camera parameter estimation is the Levenberg-Marquardt Algorithm (LMA) because it follows the gradient of the hypersurface in parameter space with a step length adaptable to the curvature of the surface (p. 484 [91]), making it both faster and more robust than, for example the Gauss-Newton algorithm (p. 597 [65]). However, the principle challenge is to parametrize the target function such that the optimization algorithm avoids becoming trapped in local minima. Because a pinhole camera model has 11 DOF, but the reprojection error only has two, it is possible to find a camera matrix $P$ which minimizes the target function but which does not correspond to parameters with true physical meaning. It is therefore useful to decompose the camera matrix into its intrinsic and extrinsic parameters and impose mathematical constraints and restrictions on variable range (p. 62 [44]). Two examples which have the possibility to introduce erroneous minima are focal length and orientation. A proper rotation matrix $R$, as expressed in eq. 4.8, should always be orthogonal ($RR^T = I$) and not change handedness ($\text{det}(R) = 1$), and the focal length, $f$, in $K_{11}$ and $K_{22}$ should have the same sign as the direction of the chosen optical axis. Hartley and Zisserman also describe what they call soft constraints, which effectively punishes the target function if it moves in a direction with unreasonable parameter values (p. 187 [65]). This method can also be used to lessen the impact of gross outlier correspondence points (p. 597 [65]).

At this stage, any additional non-linear lens parameters, like radial pincushion or barrel lens distortions [20] [143] (p. 189 [65]), could be included in the target function for a more complete camera model.

4.2.3 Camera Matrix Decomposition

There are several ways to decompose the camera matrix into its physically representative intrinsic and extrinsic parameters. Faugeras, and also Trucco and Verri, propose a closed form expression in which $P$ can be rewritten as products of the row vectors of $R$, leaving the third row of the rotation matrix, $r_3$, exposed on the last row, as shown in eq. 4.24. With the assumption that $R$ is an orthogonal rotation matrix, the norm of any row or column in $R$, including $r_3$, must be equal to 1. This fact provides the pos-
4.2. PARAMETER ESTIMATION

sibility to define the arbitrary scaling factor introduced by the homogeneous notation and serves as the first step in unraveling the individual parameters of the camera matrix (p. 52 [44]) (p. 134 [128]). The decomposition also uses the geometrical interpretation in which the rows of the camera matrix are thought of as planes (p. 160 [65]). In fact, the planes in $\mathbb{R}^3$ are lines in $\mathbb{P}^2$, and not just any lines, but the image plane axes (p. 40 [44]).

\[
P = K \begin{bmatrix}
  f m_x r_1 + p_x r_3 & f m_y t_x + p_x t_2 \\
  f m_y r_2 + p_y r_3 & f m_y t_y + p_y t_2 \\
  r_3 & t_2
\end{bmatrix}
\]

(4.24)

One drawback with Faugeras’ closed-form approach is that the orthogonality of the rotation matrix is not enforced and can therefore be corrupted by noisy measurements. A method which does guarantee an orthogonal rotation matrix, regardless of noisy data, is RQ decomposition using Givens rotations described in the appendix of Hartley and Zisserman (p. 579 [65]). With RQ decomposition, a succession of orthogonal matrices, $Q_x$, $Q_y$, and $Q_z$, are multiplied by the left 3-by-3 submatrix of camera matrix $P$, here called $S$, such that elements $S_{21}$, $S_{31}$, and lastly $S_{32}$ are canceled and become 0. Like this $SQ_xQ_yQ_z = R$ results in the right upper triangular matrix, $R$, which has the same shape as $K$ in eq. 4.6. The product of the sequence of orthogonal matrices is also an orthogonal matrix $Q = Q_x^TQ_y^TQ_z^T$, which effectively, is the rotation matrix $R$ in eq. 4.6. It should be noted that RQ decomposition with Givens rotations makes the assumption that $K_{21}$, $K_{31}$, and $K_{32}$, are zeros, which they normally are in a standard pinhole camera model, but not necessarily in asymmetric projections.

4.2.4 Degenerate Configurations

In contrast to equation systems with no solution, or systems with exactly one solution, there are correspondence point configurations that lead to a matrix $A$ with rank lower than 11 and an infinite number of solutions. These configurations are known as degenerate or critical (p. 58 [44]) (p. 533 [65]). The problem is not only that the position of the camera center is ambiguously interpreted (p. 534 [65]), but also that the system of linear equations in eq. 4.15 becomes ill-conditioned when the correspondence points are in the vicinity of a degenerate configuration [22]. This can be observed as an increase in the value of the last diagonal element of $\Sigma$ in eq. 4.21, which means that row and column space in eq. 4.17 are becoming increasingly dependent, in turn suggesting that the solution is no longer contained to a point, but rather starts to stretch along a line of infinite solutions. Measurement noise is amplified in ill-conditioned systems because a solution can shift very rapidly along this line.

Degenerate configurations with ambiguous camera center along a curve can be exemplified with the Chasles’ theorem in 2D (p. 536 fig. 22.2. [65]), which in turn
can be generalized to twisted cubics (p. 75 [65]) in 3D. Hartley and Zisserman also explain how a set of coplanar correspondence points are ambiguous up to a projective transformation (p. 534 [65]) and display all possible combinations in an informative illustration (p. 538 fig. 22.3 [65]).

Based on lines and symmetry, it is probably fair to say that degenerate configurations tend to occur more often in man-made structures and scenarios, such as markers placed on walls, or feature points collected from a smoothly curved camera path, compared with natural scenes in which feature points are more randomly scattered. However, Thomas Buchanan's first paper on camera resectioning ambiguities on a twisted cubic closes by specifying a highly structured constellation of points which will enable a camera calibration. The constellation must consist of seven world points which all have to be visible at any given time. Four points should lie on one plane, and the remaining three should be on some another plane. No three points on either plane must ever form a line [22].

In this context it should however be noted that there are camera calibration methods which use the projective transformation ambiguity to their advantage. E.g. Zengyou Zhang's method [152] enables the use of a checker board as a calibration pattern. In this case, the correspondence points are obviously coplanar.

4.2.5 Filtering of Measurement Noise

Measurement noise exists at all levels of magnitude. Theoretically, noise is always present due to the discrete pixel quantization, approximating the projection of world points onto the pixel array of the image plane. Moreover, when cameras are used for tracking, which is commonly the case for VST HMDs, the automated feature point extraction also introduces some level of noise (p. 205 [125]). Also, as previously discussed in section 3.1, trackers are limited in measurement precision. Furthermore, in the case of OST HMDs, the correspondence points are sometimes measured by the user, which also introduces human noise in the shape of head-aiming instability and postural sway as discussed previously in section 3.3. Lastly, if a point on the image plane is aligned with the wrong world point, the mismatched correspondence points will result in a gross error.

Filtering entails keeping good data points while discarding bad ones, and knowing how to distinguish between the two. Determining good and bad requires a reference which is usually calculated relative to other data points in a set. The set can either be a sequence of points (like a series of tracker reports over time) from which to chose one data point as the most representative of the sought measurement, or a collection of points from which to chose the subset that best describes a model.

The median filter is a simple but useful form of filter well suited for one-dimensional data series. It uses a moving window in which the data points are sorted, and the middle value is extracted. Spurious readings are thereby suppressed by their neigh-
boring values inside the window. There is a challenge, however, in how to apply this filter to time series which contain data in multiple dimensions related across tuples like in the case of a location report from a six DOF tracker. While it may be possible to calculate the median over a one-dimensional aggregate, such as the Mahalanobis distance, covariance is more elegantly incorporated in variations of the Kalman filter.

Adding and smoothing positional reports is straightforward as Euclidean space stretches out linearly and infinitely without singularities. The same cannot be said for rotational readings. In a paper from 1985 describing Spherical Linear Interpolation (SLERP), Ken Shoemake argues that quaternions are better suited for interpolation of geodesic distances on the SO(3) unit sphere compared to Euler angles and rotation matrices. Euler angles are complicated to average and interpolate over because of their singularities at \((2\pi, \pi, 2\pi)\). Euler angles are also ambiguous in their representation of orientation unless one of the twelve possible combinations of angle sequences is known. Lastly, a linear interpolation of Euler angles will not result in a smooth rotation, even if converted to a rotation matrix.

However, despite the praise, quaternions have their peculiarities too. For instance, the vector part of a quaternion can ambiguously represent the same rotation although having the opposite direction. Moreover, to counter rounding artifacts in multiplications, it is necessary that the quaternion norm is kept to 1 for it to represent the SO(3) unit sphere.

Assuming that measurement points have been reasonably filtered, it is now time to determine whether they fit to the pinhole camera model or should be discarded as outliers. A very simple approach is to randomly discard one point at a time and re-estimate the camera matrix to see if the reprojection error becomes any smaller. With this technique, for a set of six points, there are six possible combinations, but when excluding any two points there are 15 possible combinations. More generally, eq. 4.25 describes the number of possible combinations of correspondence points, \(c\), that can be generated from a set of correspondence points of size \(n\) when \(r\) points are excluded from the dataset.

\[
c = \frac{n!}{r!(n - r)!}
\]  

(4.25)

As shown in table 4.1, illustrating eq. 4.25 for 6 and 25 correspondence points, it would be feasible to test \(c\) combinations for a dataset comprised of six correspondence points \((n = 6)\), excluding \(r\) points at a time. Unfortunately, even with moderately small datasets of 25 correspondence points, the algorithm would have to test 68,405 camera matrices assuming five (20%) poorly measured correspondence points. It becomes apparent that this exhaustive approach to exclude poorly measured correspondence points does not scale very well.

The opposite philosophy for excluding bad points is to start with a minimal subset of
CHAPTER 4. CALIBRATION THEORY

Table 4.1: Camera matrix combinations for exhaustive outlier detection in datasets of six and 25 points.

<table>
<thead>
<tr>
<th>6 Points</th>
<th>25 Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>..</td>
<td>..</td>
</tr>
<tr>
<td>12</td>
<td>5,200,300</td>
</tr>
</tbody>
</table>

correspondence points and instead look for the largest subset of points that will support a suggested camera matrix. This is the idea behind Random Sample Consensus (RANSAC) presented by Martin Fischler and Robert Bolles [46]. Several variations on RANSAC exist but, in principle, they all start by randomly selecting the minimal number of points, s, needed to estimate a candidate camera matrix. Then RANSAC checks how many correspondence points are reprojected within a tolerance using the candidate camera matrix. If a sufficiently large fraction, w, of the entire dataset of correspondence points fall within the tolerance, consensus is reached, and the RANSAC terminates with success and presents the final camera matrix estimate with the entire consensus subset. If not, then RANSAC draws a new set of random points and repeats until the limit on iterations, L, is reached.

L is defined by the likelihood, p, of finding a set of s good points, see eq. 4.26. Conversely, 1 − p is the likelihood of a set of outliers being found. For each iteration, the fraction of inliers, w, is measured. Thus, w^s is an estimation of the probability that all s points needed for a camera matrix are inliers. Therefore, 1 − w^s is the likelihood of a set of outliers. The likelihood of only outliers in L attempts is (1 − w^s)^L, which should equal the original assumption of 1 − p. Solving for L gives eq. 4.27.

\[
1 - p = (1 - w^s)^L \tag{4.26}
\]

\[
L = \frac{\log(1 - p)}{\log(1 - w^s)} \tag{4.27}
\]

Table 4.2 shows the maximum number of RANSAC iterations needed to estimate a camera matrix based on six correspondence points for datasets that vary in quality between 50-95% inliers with 99% and 95% assumed probability that any six points are good enough to define a camera matrix, as expressed in eq. 4.27.
4.2. PARAMETER ESTIMATION

Table 4.2: Theoretical limits of RANSAC iterations for estimating a six point camera matrix.

<table>
<thead>
<tr>
<th>$w$</th>
<th>$L (p = 0.99)$</th>
<th>$L (p = 0.95)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.95</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>0.90</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>0.85</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>0.80</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>0.75</td>
<td>23</td>
<td>8</td>
</tr>
<tr>
<td>0.70</td>
<td>37</td>
<td>13</td>
</tr>
<tr>
<td>0.65</td>
<td>59</td>
<td>21</td>
</tr>
<tr>
<td>0.60</td>
<td>96</td>
<td>34</td>
</tr>
<tr>
<td>0.55</td>
<td>164</td>
<td>57</td>
</tr>
<tr>
<td>0.50</td>
<td>292</td>
<td>102</td>
</tr>
</tbody>
</table>

Clearly, RANSAC is preferable in larger datasets, but its efficiency is ultimately dependent on the tolerance set for the reprojection error, the measurement that defines consensus. Furthermore, if all correspondence points are equally noisy, then they will all reach consensus at the same tolerance level. Therefore, RANSAC can only be used to single out gross outliers, but not be used to improve a dataset with moderate noise distributed across all points. An exhaustive search, however, could improve a dataset with homogeneously distributed noise, but it would be too inefficient for practical purposes for larger datasets.

In this context it is also worth noting that there is no relationship between the noise of a correspondence point at a specific location on the image plane and the resulting calibration quality over the same area on that image plane. Because the solution is algebraic, and not geometric, the effect of a poorly measured point in the upper left corner of the image plane will affect the projection over the entire image plane, not only in the upper left corner. Moreover, for the same reason, the reprojection error will be smaller closer to the areas on the image plane where correspondence points have been collected compared to areas where none are sampled.

With an acceptable set of correspondence points, it is time to investigate available calibration techniques.

4.2.6 Camera Calibration Techniques

The introduction of Roger Tsai’s paper [129] includes a very thorough review of the calibration techniques available up until 1987. Tsai’s review is particularly focused on the reported accuracy of the techniques and whether they incorporate compensation for lens aberrations. Tsai also divides the techniques he has found into four categories: I. Techniques involving full-scale non-linear optimization; II. Techniques in-
volving computing perspective transformation matrix first using linear equation solving; III. Two-plane methods; and IV Geometric techniques. For example, Sutherland’s approach [124] presented earlier in this section sorts into Tsai’s category II along with the more commonly cited alternate DLT approach by Abdel-Aziz and Karara [1]. A more general form of Fischler and Bolles’ RANSAC [46] fits into category IV. Tsai also proposes a camera calibration technique of his own. One of the main strengths of Tsai’s method is that it offers the ability to calibrate intrinsic and extrinsic parameters separately. This is sometimes referred to as a two-stage calibration, but should not be confused with the two-stage calibration technique for OST HMDs discussed in chapter 5. Moreover, Tsai’s algorithm can also correct for lens aberrations and use coplanar correspondence points, as long as the plane of points is not parallel to the image plane.

In 1997 Janne Heikkilä and Olli Silvén added steps to correct for circular features and distorted image coordinates, effectively resulting in a four-stage camera calibration procedure [66]. Mainly applicable to cameras, and therefore not directly related to OST HMD calibration, the work of Heikkilä and Silvén is included in this review mainly because of their concise reiteration of the DLT technique, parameter decomposition, and parametrization for non-linear optimization.

A more recent and simpler camera calibration technique is that of Zengyou Zhang [152] published in 2000. With the assumption that the planes of world points are located at $Z = 0$, the third column of the rotation matrix can be discarded which enables the use of 2D homographies between two (or more) sets of coplanar world points and their corresponding points in the image plane. It, too, compensates for lens distortions. However, being restricted only to planes, it is of limited use for OST HMD calibration, as will be shown in section 5 on calibration techniques.

Wei Sun and Jeremy Cooperstock have compared the three algorithms from Tsai, Heikkilä, and Zhang, and concluded that the ones from Tsai and Heikkilä perform the best in presence of noise when comparing a normalized reprojection error [120]. In more detail, Sun and Cooperstock list the individual camera parameters for the three techniques when calibrated with hundreds of correspondence points. The estimated principal point and extrinsic parameters are shown in table 4.3 to illustrate the variability between methods.

The purpose of including table 4.3 in this review is to give an indication of the parameter variability as a function of choice of calibration method in an otherwise identical and ideal setup. Other than image resolution (640 by 480 px), the actual camera parameters were not specified by Sun and Cooperstock, thus nothing can be said about accuracy. Nor is it explicitly stated that the camera was static while the three datasets were collected, but the notably small variation in $R_z$ suggests that it was.

In terms of comparing the effect of noise, keeping the calibration method constant, Sunil Kopparapu and Peter Corke have published a sensitivity analysis in which they permute the pixel values of the image plane with noise ranging from 0.05 to 1.0
Table 4.3: Variability in some selected parameters as a function of camera calibration method.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Tsai</th>
<th>Heikkilä</th>
<th>Zhang</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_x ) (px)</td>
<td>339</td>
<td>320</td>
<td>332</td>
<td>19</td>
</tr>
<tr>
<td>( p_y ) (px)</td>
<td>240</td>
<td>263</td>
<td>268</td>
<td>28</td>
</tr>
<tr>
<td>( t_x ) (m)</td>
<td>1.9829</td>
<td>1.9102</td>
<td>1.9635</td>
<td>0.0627</td>
</tr>
<tr>
<td>( t_y ) (m)</td>
<td>1.4555</td>
<td>1.5359</td>
<td>1.5600</td>
<td>0.1045</td>
</tr>
<tr>
<td>( t_z ) (m)</td>
<td>2.3249</td>
<td>2.2564</td>
<td>2.2697</td>
<td>0.0685</td>
</tr>
<tr>
<td>( R_x ) (°)</td>
<td>153.18</td>
<td>154.73</td>
<td>153.07</td>
<td>1.66</td>
</tr>
<tr>
<td>( R_y ) (°)</td>
<td>19.44</td>
<td>17.69</td>
<td>17.33</td>
<td>2.11</td>
</tr>
<tr>
<td>( R_z ) (°)</td>
<td>91.25</td>
<td>91.18</td>
<td>91.45</td>
<td>0.27</td>
</tr>
</tbody>
</table>

The pixel standard deviation [81]. Sunil and Corke use 32 calibration points to estimate a linear solution using a least-square minimization of the algebraic distance. This is comparable to performing a DLT as outlined by Abdel-Aziz or Sutherland, or finding the initial linear solution as described by Tsai or Heikkelä. Table 4.4 shows some results from the sensitivity analysis. Note that, typical to Tsai’s calibration procedure, the extrinsic coordinate system is aligned with the two coplanar calibration structures and not with the camera coordinate system. This means that the world’s z axis is the only axis that coincides with the camera’s coordinate system (yaw). Assuming a screen with 640 by 480 px resolution and a FOV of 37° by 28°, the yaw variance corresponds to 38 px which equals a 6 px standard deviation. Together with the uncertainty of horizontal principle point, the unbiased pooled standard deviation for rendering objects in the horizontal direction is about 6.4 px at noise levels of 1.0 px. While it is only 1 % of the total screen width, it would most certainly still be reported as a noticeable registration error by a subject.

Table 4.4: Variance in some selected parameters as a function of noise (px sd).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>0.05</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p_x ) (px)</td>
<td>2.16</td>
<td>4.42</td>
<td>21.61</td>
<td>43.63</td>
</tr>
<tr>
<td>( p_y ) (px)</td>
<td>0.51</td>
<td>0.96</td>
<td>4.97</td>
<td>10.22</td>
</tr>
<tr>
<td>( t_x ) (m)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
</tr>
<tr>
<td>( t_y ) (m)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>( t_z ) (m)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>( R_x ) (°)</td>
<td>0.00</td>
<td>0.00</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>( R_y ) (°)</td>
<td>0.08</td>
<td>0.15</td>
<td>0.76</td>
<td>1.55</td>
</tr>
<tr>
<td>( R_z ) (°)</td>
<td>0.11</td>
<td>0.22</td>
<td>1.09</td>
<td>2.20</td>
</tr>
</tbody>
</table>

To place this sensitivity analysis in the perspective of human performance, the values reported on head-aiming performance in section 3.3.2 (0.13-0.8°) translate to noise in the range of 2.2-13.8 pixels. Unfortunately, neither Sun and Cooperstock nor Koppa-
rapu and Corke report results using noise of these magnitudes. However, Kopparapu and Corke report that the principal point variance scales proportionally to noise, but that the orientation parameter was more unpredictable. Despite precise measurements, parameters can still vary. Thus noise is not always the reason for poor parameter estimates. Erroneous parameters are also possible if one tries to fit data to a model which does not correspond to the reality in which the data was gathered. The next section introduces some slight variations on the pinhole camera model which will be further expanded in section 5.7 on the practical importance of choosing an appropriate model for eyepoint estimation in OST HMD calibration.

4.3 Asymmetric Camera Models

In a traditional pinhole camera model, the optical axis passes perpendicularly through the center of the image plane, resulting in a symmetric frustum, see fig. 4.2. However, there are instances in VR and AR where such a projection model is not suitable. Ramesh Raskar and Oliver Bimber's book on Spatial AR, dealing with a subset of AR techniques using projectors, contains examples where an originally symmetric frustum of a projector is directed at an oblique angle on a wall, thus resulting in an asymmetric frustum (a.k.a. off-axis), see fig. 4.3. The setup is illustrated in fig. 4.4. If the projection surface is also planar, then the ordinary pinhole camera $P_1$ can be extended with a 2D-2D homography to model $H_1$, the transformation between the camera's image plane and the flat projection surface, denoted $S$. Yet another homography $H_2$, can be added to allow for a changing perspective as the user moves and observes the display surface from changing directions, $P_2$ [13]. In fact, the setup is very similar to that of a tracked VR user moving inside a space enclosed by planar walls, such as in a cube-shaped CAVE Audio Visual Environment (CAVE) [34].

Figure 4.2: Front and top view of a symmetric frustum.

In addition to VR systems with projection surfaces made of stationary walls, asymmetric camera models are also found in stereoscopic HMDs. In an orthostereoscopic setup, the two frusta of the user's left and right eye are pointed straight ahead [47] and are therefore symmetric, see left fig. 4.5. In a binocular display, this presents no problem as each eye has an independent display surface. However, in a biocular display the two image planes have to be separated according to the user's IPD. This
4.3. ASYMMETRIC CAMERA MODELS

Figure 4.3: Front and top view of an asymmetric frustum.

Figure 4.4: Top view of a symmetric frustum projected on a wall, and the frustum of the user’s eye, observing it.

means that the image presented on the display surface observed by the user’s both eyes has to either alternate using some active shutter system, or be separated with a passive filtering technique. Because the outermost part of the left image plane cannot show an image to the right eye, and vice versa, the effective size of an image plane for orthostereoscopic viewing is cropped by two times the user’s IPD. An alternative is to rotate the two frusta with a cant angle. While the middle illustration in 4.5 illustrates an inward rotation such that the optical axes intersect in a “toe-in” configuration, the frusta can also be diverged outwards [109]. In both cases, the two image planes will no longer be coplanar unless the frusta are made asymmetric as shown to the right in fig. 4.5. If the two image planes do not coincide, the presented image will have a lateral displacement which results in erroneous convergence of the user’s eyes and contributes to difficulties in estimating correct distance to virtual objects [113].

It is important to note that the decomposition methods presented in section 4.2.3 assume that the camera matrix only consists of intrinsic and extrinsic parameters. Thus, an asymmetric camera matrix cannot be decomposed into its intrinsic and extrinsic parameters unless the homographies are estimated and canceled out first.
Figure 4.5: Orthostereographic, rotated, and asymmetric frusta.
Chapter 5

OST HMD Calibration

This chapter has essentially two parts. It starts by introducing the misalignment between real and virtual objects in virtual environments, often referred to as “the registration error”, and motivates why we should attempt to minimize it. We also present a method by which it is possible to estimate the magnitude of a registration error even before the system is built. The second part of the chapter presents previously researched calibration techniques for OST HMDs, and explains which parts that we found interesting to use in our calibration procedure. This chapter closes by presenting how an asymmetric camera model could be applied to an OST HMDs in case the experimental setup does not adhere to the basic pinhole camera.

5.1 Registration Error

As shown in the previous sections, an AR system relies on a camera model to render virtual objects in the correct location according to the surrounding world. The model in turn relies on appropriate parameter settings, which should correspond to physical measurements of the relationship between the user and the display. The process of setting model parameters is referred to as calibration. Only with the correct set of parameters can the AR system render a view in which a computer generated object matches the user’s viewpoint.

When a computer generated object in an AR system exactly overlays the real world object it is intended to convey information about, the object is said to be correctly registered. Any perturbation in this alignment is referred to as a registration error. The registration error is usually expressed in pixel units as an Euclidean distance on the image plane. As such it can be seen as a reprojection error between graphics and real world objects. The registration error is sometimes also reported as visual angle independent of display resolution. However, as Richard Holloway thoroughly investi-
CHAPTER 5. OST HMD CALIBRATION

In his thesis [74], the registration error is more than just a lateral perturbation on the image plane, but rather the symptom of discrepancies between system model and the reality it attempts to replicate.

It makes perfect sense to discuss the lateral perturbation in the image plane as a measurement of calibration quality because it is this very image which the user is ultimately presented with. However, if the ambition is to correct for said registration error, more information about cause and origin is needed. As shown in chapter 4, the parameters of the commonly used pinhole camera model have 11 DOF which result in a camera matrix projecting onto an image plane where pixel values can change only in two DOF. This means that there are several combinations of the 11 parameters which could result in an identical registration error. With several possible combinations, it is also a nonintuitive task to find which parameters need to be changed in order to achieve correct registration.

Holloway chose to study the possible error sources and divides the registration error into acquisition errors, tracking errors, display errors, and viewing errors [72]. Acquisition errors refer to the erroneous modeling of the spatial relationships between objects in the world. Tracking errors arise due to poorly calibrated trackers or noisy readings. Display errors are caused by shifts in the virtual image due to poor calibration or lens aberrations. Viewing errors are mainly due to an incorrect eyepoint which may be caused by equipment slippage, but can also be due to poor calibration.

Registration errors can further be divided into those arising from static or dynamic sources. A dynamic registration error has a temporal component and becomes apparent when the user moves. Such artifacts are usually due to system lag [5][77][3], drift or jitter [52]. A static registration error is also visible when the user is stationary. In practice, however, there exists an intermediary class which is not reflected in the literature, namely the registration error which does not become apparent until the user changes head pose from one static location to another. The latter kind is particularly interesting because, as mentioned earlier, many combinations of parameters can achieve a good registration, but only a unique set will ensure good registration from an arbitrary viewpoint.

5.1.1 Motivation for Minimizing the Registration Error

Early applications of AR [27][6][45] were interested in minimizing registration errors mainly for practical reasons as poor alignment decreased the usability of the application. However, since then, it has proven important to study not only what is seen by the user, but also how it is seen. A user may very well temporarily adapt to some viewing condition, but it might be at the expense of discomfort or even injury.

In 1988 Frederick Brooks, who at the time was advising the PhD student Richard Holloway [74], wrote a general paper on interfaces to virtual worlds in which he urged researchers to share not only stringent findings, but also unsupported obser-
5.1. REGISTRATION ERROR

In the paper, Brooks himself shares observations about the 14 projects experimenting with virtual worlds run at UNC at the time. Brooks emphasized the importance of usability and stated that the systems must be “so simple full professors can use them, and so fruitful that they will”. Ten years later, Kay Stanney, Ronald Mourant, and Robert Kennedy published a literature review on human factors issues in virtual worlds in which many of the unsupported observations had been subject to proper experimentation [118]. In addition to the practical perspective, that a system should be simple and efficient, Stanney et al. also emphasize aspects of health. The system must not cause discomfort, or create situations that may harm the user during or after usage. Thus it is no longer sufficient to address calibration requirements pragmatically - the user must also be safe. Stanney et al. were mainly concerned with cybersickness, the undesirable psychophysical side effects caused by contradictory visual, vestibular and proprioceptive stimuli, resulting in nausea or stomach discomfort, disorientation or postural instability, and visual symptoms (p. 172 [9]).

Cybersickness is also the main concern in the more recent review by Judy Barret of the Australian Department of Defence [8]. The sources cited in this work seem to suggest that it is not only the HMD that must be carefully calibrated, but that thought should also be lent to the task at hand.

The literature studies cited here provide excellent references to psychophysical research establishing thresholds and models, but for all practical purposes calibration requirements must be considered individual to both system setup and the intended task. Instead, inspired by the introduction of the paper by Stuart Gilson, Andrew Fitzgibbon, and Andrew Glennerster [60], it is probably more helpful to be aware of the main psychophysical principles and thereafter establish calibration procedures that allow the measurement of parameters important in those principles and attempt to replicate as realistic conditions as possible.

5.1.2 Statistical Estimation

It may be useful to know the level of registration to expect in a specific setup. This knowledge can be used to estimate which tasks are possible and to make design decisions on symbology [92]. For the purposes of estimating the resulting registration error in an HMD, tracker noise and user aiming precision will define the the covariance with which correspondence points are measured. The correspondence point covariance is then propagated through a linearized model of the camera matrix, which results in confidence regions on the image plane (p.138 [65]). To linearize the camera matrix around its mean it is necessary to estimate the partial derivatives, the Jacobian.

Linearizing models is commonly done with Taylor expansion, for example when working with Extended Kalman Filters (EKFs), but Blair MacIntyre, Enylton Machado Coelho, and Simon Julier instead used an unscented transformation which does not
require computation of derivatives. The aim was to compute a convex hull on the image plane inside which the real world object can be expected to be found [92]. Another application of covariance propagation is found in a paper by Daniel Pustka et al. where variance measurements in a camera tracking system is used to estimate the true center of a rigid body of fiducial markers [102]. Pose uncertainty modeled in a similar way is also found in a paper on head pose accuracy by William Hoff and Tyrone Vincent [69].

An alternative to covariance propagation is Monte Carlo simulation (p. 149 [65]). Depending on the size of parameter space, input parameters can be exhaustively or pseudo-randomly sampled to estimate the covariance of the model. In practice this could, for example, imply permuting correspondence points with increasing levels of noise and repeatedly estimate a camera matrix to determine the sensitivity of the camera matrix parameters.

5.2 Visual Alignment

All OST HMD calibration procedures to date are based on the exercise where a user creates a visual alignment between points on the screen with points in the surrounding world. The exercise is called boresighting, although “to boresight” in its strictest sense only includes visual alignments made straight ahead along the barrel of a firearm. Contrary to VST HMD calibration, the OST HMD calibration procedure offers no feedback to aid in the alignment task as it is currently not possible to monitor what the user sees. Thus, it is factors such as the user’s visual acuity, motor skills, postural stability, and subjective judgment that determine the quality of each alignment and subsequently the calibration result.

When a user performs a boresight calibration, the eye is located somewhere along the line connecting the two points over which the alignment is made. To exactly determine the point in space where the eye is located, the user can simultaneously maintain another boresight such that the position of the eye can be inferred to be at the location where the two lines intersect. A more general approach of maintaining multiple boresight alignments is the alignment over two objects of the same relative size. The edge around the objects can be seen as an infinite number of points over which a boresight is held. This technique was used by David Mizell and Thomas Caudell in the first AR application using an OST HMD published in 1992 (fig. 5 [27]). Two circles of different size were placed some distance apart such that when they were observed from a particular point, they appeared to be of equal size. This particular point marks the apex of a cone described by the two circles. Unfortunately, as some back-of-the-envelope calculations will reveal, accounting for limitations in visual acuity and depth of field, the calibration method is too imprecise along the line of sight to be practically useful. Moreover, the calibration over circles only defines five of the required six DOF.
Another example of simultaneous boresight is found in a paper on calibration by Azuma and Bishop from 1994. In their setup three axes of a virtual coordinate system shown on the screen are to be aligned with the perpendicular edges of a wooden crate while the user also boresights over two nails at an oblique angle (fig. 9 [5]). The authors mention that camera calibration techniques were considered, but that they were deemed too difficult to apply in this particular setting. Instead simpler tasks and direct measurement of geometric constraints such as FOV were used to determine the necessary viewing parameters.

5.3 Direct and Indirect Measurements

When Adam Janin, David Mizell, and Thomas Caudell in 1993 revisited the calibration procedure for their application at Boeing [78], they considered the possibility of directly measuring some of the parameters using rulers and protractors. However, plastic covers casing the tracker sensors posed a problem as it was difficult to determine the exact tracker origin. Janin et al. also commented on the inconvenience of their previous calibration procedure involving objects of relative size, and attempted to replace it with measurements using stereo camera metrology. It was eventually concluded that direct measurement, be it with rulers or stereo cameras, was less reliable a technique than optimization methods. Since optimization uses repeated measurements it was more robust to noise compared with a single direct measurement. That said, Janin et al. also noted the difficulties in parameterizing orientation for the target function in the optimization algorithm, similar to the discussion in section 4.2.2.

Even if optimization might be a better approach in noisy setups and where origin and reference points cannot be accurately located, the collected set of indirect measurements related to the viewing parameters say nothing about the resulting accuracy of the calibration, and thus cannot be used to verify the calibration success. Measurement precision may be propagated as explained in section 5.1.2, but an evaluation on calibration accuracy needs absolute reference points, the equivalent of what Janin et al. refer to as direct measurements. This poses a challenge because many of the individual pinhole camera parameters, particularly the intrinsic ones like principle point and focal length, are hard to measure directly.

5.4 Minimizing Measurement Errors

Plastic casings covering the sensor origin are not the only factor preventing accurate direct measurements. Measuring the location of the user’s eye inside concave sockets, or the location of a fiducial marker across the floor and up a wall, are hard to do accurately with tools like a tape measure. In our experience even precise tools using collimated laser beams will diffuse to a spot larger than 5 mm in diameter over
Thus, to minimize measurement errors, it might be preferable to use an additional sensor monitored by the same tracking system as used by the AR system. With measurements contained within the same system, measurement accuracy will only be limited by tracker precision.

Gilson, Fitzgibbon, and Glennerster made use of this fact when they used reflective markers which could be segmented in a camera view, but also tracked by the tracking system, alleviating the need for manual measurement of each marker location. The tracking system need not be camera-based for this principle to be used. It can even be used with only one tracked point, using it like a 3D ruler. For example, Anton Fuhrman, Dieter Schmalstieg, and Werner Purgathofer used a tracked stylus to, one by one, collect the eight corners of a frustum to calculate the camera matrix. A similar approach was used by Artur Tang, Ji Zhou and Charles Owen who collected nine points intended for a Single Point Active Alignment Method (SPAAM) calibration.

Whitaker et al. and McGarrity et al. also used a tracked stylus to collect correspondence points and to measure registration errors, but these pens were tracked by a separate magnetic tracking system. Because the two tracking systems had separate origins, their relationships had to be calibrated as well, thereby potentially introducing additional measurement errors, not benefiting from the virtues of a single system.

The general idea of using the hand to gather measurements is an interesting one. It may be that correspondence points collected using the well-trained motor skills of the hand are more precisely measured than correspondence points collected by aiming with the head. Tang et al. argue this idea, but the source cited is an unpublished thesis from 1973 containing experimental comparisons of Fitt’s Index for various limbs. Unfortunately, Fitt’s law measures completion time to reach some interval along the axis of movement, and is therefore not the ideal measurement to compare precision on point targets. However, the experiments by Tang et al. show that gathering correspondence points with a stylus results in smaller registration error compared to aiming with the head. Therefore it is plausible, but not conclusive, that the correspondence points are indeed measured more precisely by hand.

### 5.5 Collecting Individual Measurements

Be it by head or by hand, the user must collect individual measurements one at a time. When a camera captures an image, it essentially “freezes” a moment in time. With feature point detection algorithms it can then extract hundreds of candidates for a set of correspondence points. Obviously this task is not feasible for humans. Even if one could make do with less correspondence points, it would still require the user to maintain an impossibly rigid pose. As far as calibration techniques go, this
effectively eliminates methods where multiple correspondence points are collected simultaneously, like Zhang's technique and the planar case of Tsai's. Instead, the feature points have to be collected individually and then be merged into the same frame of reference, compensating for user movement between each collected point.

A method for doing this has been presented by Mihran Tuceryan and Nassir Navab [132]. The SPAAM algorithm references the gathered correspondence points in head-relative coordinate space and thus compensates for the user movement between each point. Assuming that the display of the HMD is rigidly attached to the tracker sensor, the user is free to move around a single marker (hence the name), and collect instances of the marker with different head poses such the collected marker points eventually constitutes a calibration pattern in head coordinate space.

SPAAM is well suited for calibration of OST HMDs, but by the same token also difficult to evaluate. The human vision system does not offer any auxiliary signal by which to follow what the user sees. In the original paper, the eye of a mannequin was replaced by a camera, but the results were only reported qualitatively as images [132]. Moreover, the calibration was done on a tripod and did not exhibit any of the human noise typical to postural sway or head rotations. SPAAM has been evaluated using a VST HMD [59] carried by a user, however, it does not model eyepoint location properly since it uses a camera.

5.6 Camera-Assisted Measurement

Some researchers have tried to improve measurement precision by using a camera in lieu of the human eye. As mentioned above, Gilson et al. used the reflective markers of a tracking system to accurately measure the marker locations in a calibration pattern. Then they hollowed out the eye of a mannequin, fitted a camera, mounted the HMD on the mannequin, and performed the calibration with Tsai's technique (see section 4.2.6) using the camera view through the HMD viewport. However, the evaluation which followed was made with the same camera used for calibration. In practice, the HMD would have been taken off the mannequin and placed on a user's head. Thus, at this stage, the calibrated eyepoint of the mannequin would no longer correspond to the eyepoint of the user.

The challenge of mismatching eyepoint after camera-assisted measurements is an issue which Charles Owen, Ji Zhou, Arthur Tang, and Fan Xiao addressed. They offer a two-stage approach where the second stage is a separate eyepoint calibration done after the camera-assisted calibration [98]. An evaluation of this method is found in Zhou's thesis [153].
5.7 Model Selection

Perspective rendering in computer graphics has since long been estimated using the virtual camera analogue. It is an implementation of the pinhole camera model and is found in most textbooks on computer graphics (p. 97 [146])(p. 143 [137]). The pinhole camera model is convenient to work with as it permits the separation of camera movement (extrinsic parameters) from parameters governing the view (intrinsic parameters), see eq. 4.6 and section 4.2.3. Applied to an HMD system, a pinhole camera model might be described with a diagram as presented in Fig. 5.1. The user observes the display with the perspective from eyepoint \( P \), via the reflection in the optical combiner \( S \). The combiner is assumed to be located at 45° relative to the principal axis of the projection \( P \). If it is not, then the homography \( H \) can be used to make the virtual image of the display perpendicular to the image plane of \( P \).

![Figure 5.1: A simple OST HMD model assuming a standard pinhole camera model.](image)

As explained in section 4.3 on asymmetric camera models, the pinhole camera model makes the assumption that the principal axis passes perpendicularly through the image plane. It is only then the matrix decomposition techniques described in section 4.2.3 will function properly. Is this a valid assumption for OST HMDs? If not, what are the effects on the calibration procedure? In the literature reviewed, no study has investigated the tolerances associated with model adherence. No simulations have been done to evaluate how the calibration results are affected as the setup departs from the assumed pinhole camera model.

Although HMD designs vary greatly, an example of departing pinhole camera model
5.7. MODEL SELECTION

in a HMD system with magnifying lenses and optical combiner could perhaps be approximated by the diagram presented in fig. 5.2. It is essentially a variation of Bimber and Kaskar’s projector setup illustrated in Fig. 4.4 [13]. Let $P_1$ be the camera matrix describing the ideal projection for which the system hardware was calibrated. Implicitly the camera matrix also denotes the ideal eyepoint location. The image generated by the display is reflected by the surface $S$ on the optical combiner. As the pixel array in the display and the combiner might not be correctly angled, $H_1$ describes the 2D-to-2D homography needed to turn the image on $S$ into an image that is parallel to the image plane of the pinhole camera described by $P_1$. Now, if the user’s eye is instead located at $P_2$, the compensation of $H_1$ should be undone, and instead $H_2$ should be used to make the virtual image of the display perpendicular to the actual perspective, as seen from $P_2$.

**Figure 5.2: A OST HMD model with displaced eyepoint.**

Assuming that the display, the optical combiner, and the users head are all rigidly attached to the tracker sensor, the correspondence points can be collected individually as explained in section 5.5. The linear solution would, however, express a camera matrix which is the product between $P_2$ and $H_2$, and possibly also an erroneous correction of $H_1$ not needed since the user’s eye is not in $P_1$). Thus, $H_1$ and $H_2$ have to be known to estimate $P_2$, and only then can $P_2$ be decomposed into extrinsic and intrinsic parameters.

Furthermore, while we would prefer $H_2$ to be static, it is unfortunately affected by equipment slippage. Therefore, despite the recommendations against direct measurement by Janin et al. [78], and against the conclusions of Holloway [72], dynamically
maintaining $H_2$ could be a suitable application of eyetracking, provided that the tracking could be performed in the reference frame of the viewport. Naturally, this would of course introduce the challenge of calibrating the eyetracking camera relative to the display surface.

In conclusion, while the pinhole camera is an appropriate model for rendering computer graphics in most other applications, including VR, it may not meet the requirements for calibration of the user’s view in OST HMDs. An erroneously modeled eye-point will lead to a rotated display surface, which introduces measurement noise in the correspondence points. Knowing that the DLT is quite sensitive to noise, it is fair to suspect that eyepoint location has an impact on calibration quality, but we cannot know for certain until it has been tested since no simulations or experiments have been made.

If we chose to study this revised pinhole camera model, we will be faced with the challenge of estimating the relationship between the display surface and the user’s eye if we want to decompose the camera matrix into its extrinsic and intrinsic parameters. If we use a rendering pipeline that does not require the parameters to be specified in two separate sets, we could pass the composite camera matrix to the rendering system, provided that the hardware calibration, $H_1$ is known (or negligible).
Part II

Contributions
Chapter 6

Summary of Studies

This chapter introduces the aims, results, and contributions of the seven studies included in this thesis. The contributions are also summarized in a more concise form in the beginning of chapter 7.

6.1 Paper I

6.1.1 Aims

The purpose of the first study was to investigate the postural stability in subjects performing the boresight aiming exercise commonly used in calibration procedures. The boresight alignment was made between a standing subject, a point on a transparent projection screen 3.40 m in front, and a crosshair on an opaque projection screen an additional 10.40 m behind the transparent screen. Thus, in this setup resembling a large format HUD, the alignment could be constructed along a line and subsequently the primary motion of the subject to achieve alignment would be head translation.

Standing subjects were used because this would more closely resemble the conditions under which a calibration is expected to be made in practical applications. To determine the effect of the boresighting task (Boresighting) we introduced two control conditions, namely quiet standing while not aiming (Eyes Open), and quiet standing with eyes closed (Eyes Closed). The three conditions were measured in terms of the distance accumulated by head translations over 30 s (Sway Path) and the average distance of head translation from its average location (Distance from Center Point). The latter was designed to be a standard deviation in three dimensions, effectively measuring precision. The Sway Path could also be used to calculate the Romberg Coefficient, a common metric to measure postural stability which could be compared to other sources in literature.
6.1.2 Results

We found that the Romberg coefficient for our subjects was, on average, 0.73 m, with a standard deviation of 9 cm. This corresponded well with sources in the literature which indicated that our 12 subjects were representative of the larger population. This was an important step, since the results that followed were at first seemingly counterintuitive. Apparently both Sway Path and Distance from Center Point were significantly greater during the Boresight condition compared to Eyes Open. Intrigued by this finding, we began to investigate Distance from Center Point over time by dividing the 30 s recording into ten bins of three seconds. Since the data points were not normally distributed we used a non-parametric Analysis of Variance (ANOVA) to confirm that the Distance from Center Point in the different time periods were significantly different. We found the data could be fitted to a quadratic trend and hence we could determine that subjects tended to settle and perform best 12-15 s into the calibration exercise. At this point, marked by an arrow in Figure 6.1, the average translational precision is 0.9 mm, which corresponds to 6 mm improved precision compared to an immediate reading in the time period 0-3 s.

![Figure 6.1: Distance from Center Point for the Boresight viewing condition over time and its quadratic approximation.](image)

We then expanded the analysis of the time period between 12-15 s and attempted to fit an omnidirectional probability density function to the subjects’ positional data to determine the likelihood of the precision being 6 mm at 12-15 s. As shown in Figure 6.2 we found a good fit with a Weibull distribution, \( P(x) = A \cdot \alpha \beta^{-\alpha} x^{\alpha-1} e^{-x^\alpha} \).

6.1.3 Contributions

The study found that subjects postural stability worsened during a calibration task compared to quiet standing. It also shows that subjects tend to settle and therefore it is beneficial to wait some 12-15 seconds before collecting a recording. Furthermore,
6.2. Paper II

6.2.1 Aims

The second study treated the same experiment as described in Paper I, but expanded the analysis with data that could not be included in the short paper format of the previous study. For instance, the independent variables Eyes Closed, Eyes Open, and Boresight (collectively referred to as Viewing Conditions) were nested with Background, which implied the use of two different background markers. The first marker was a white Crosshair on a black background. The second marker was a Photo of a view from a control tower. By alternating the two, we wanted to study if there was any effect of peripheral vision during the boresight task.

Moreover, following discussions with Jeffrey Mulligan after a presentation of the preliminary results at a NASA Ames “Brown Bag” meeting, the second study also incorporated some changes to the data preparation which preceded the analysis. In the preliminary data we found a correlation between pitch and anterior-posterior head translation, suggesting that subjects performed head rotations compensating for postural sway, and that these rotations pivoted around the eye. This meant that the position data might have been erroneously amplified by the offset between the tracker sensor and the subject’s eye. Since the subjects were no longer available for measurements to estimate the distance between the tracker sensor and the eye, we used an
anthropological model of an average skull to estimate the offset vector. The vector was then counter-rotated using the orientational readings of the tracker to cancel the effect of head rotation.

### 6.2.2 Results

Comparing the original dataset to the dataset in which the effects of head rotation had been canceled out, it turned out that the effects were often below tracker resolution and therefore negligible. In the instances where corrections indeed had an effect, it was mainly the anterior-posterior direction which had been affected and the corrections were on average 2.9 mm. As a sensitivity study, we also lengthened and shortened the offset vector using the pooled standard deviation of the skull measurements in the anthropological model and found that it had no effect on the results on the subsequent ANOVA.

With the corrected data we asserted our previous finding and confirmed that the boresight task deteriorated postural stability and aiming precision. Moreover, as previously reported, we concluded that it is preferable to wait 12-15 s until recording calibration data. The translational head-aiming precisions in the beginning and the middle of the recording period were 11 mm and 16 mm, respectively. The corresponding measurements reported in the previous study were 9 mm and 15 mm.

Although literature sources had led us to expect otherwise, we did not find any effect suggesting that performing boresight against a background stimulating the peripheral view is any better than a neutral black background. This might be due to that the distant background screen only occupied 10° by 7° of the subject's visual field.

### 6.2.3 Contributions

The study confirmed the previously published results, even when data had been adjusted for compensatory head rotation.

### 6.3 Paper III

#### 6.3.1 Aims

Having found indications that subjects used head rotation to compensate for postural sway during a boresight exercise, the next study was aimed at investigating visual alignment precision with a setup which permitted the subjects to move more freely. To contrast the previous study, in which a subject achieved a visual alignment over two markers in space primarily with translational head movements, this study involved
an HMD with which the subjects would target background markers primarily using head rotations. Moreover, by using an HMD which provided a headslaved foreground marker, we alleviated the need to measure or model the relationship between the tracker sensor and the subject's eye.

We were mainly interested in seeing if there was any difference in visual alignment precision as a function of aiming direction, and decided to distribute white crosshair markers on a black background screen over $\pm 0^\circ, \pm 30^\circ, \pm 60^\circ$ azimuth angle and $0^\circ, \pm 10^\circ$ elevation angle. We then asked subjects to align a headslaved foreground white pixel with the background markers and maintain alignment for 30 s.

Since the HMD would drastically limit the subject's FOV, we decided to closely monitor changes in postural stability as well. Therefore we applied the same three viewing conditions as in the previous study, studying Sway Path as a dependent variable of the viewing conditions Eyes Open, Eyes Closed and Boresight.

### 6.3.2 Results

The use of Sway Path as a metric for postural stability while using an HMD turned out to yield more questions than answers. While there was an effect of viewing condition, there was no significant difference between the Eyes Open and Boresight viewing condition - in fact their distributions looked very similar, see Figure 6.3. Thus we could not conclude that the Boresight task deteriorated the postural stability as we did in our previous study with the HUD. Instead it seemed as if the subject was translating equally far during the boresight exercise as during normal quiet standing.

![Figure 6.3: Sway Path for the three viewing conditions.](image)

It was, however, interesting to note that the Sway Path was about twice as long for all viewing conditions compared to the previous study. When comparing the two datasets...
we noticed that the standard deviation for the Eyes Closed condition was comparable. Since a limited FOV would only be important for conditions involving vision the increased Sway Path in Eyes Closed could not be attributed to the diminished FOV. A longer Sway Path but a similar spread suggested that the subjects behaved quite differently when wearing the HMD compared with the HUD case in the previous study in which the subjects wore nothing but a light tracker sensor. Thus it seems that the longer Sway Path is related to simply wearing the HMD. Our independent variables did not include a condition without HMD, and thus we could not conclusively attribute this finding to the HMD.

Compared to the previous study, the Sway Path was also remarkably longer in the Boresight condition. This might be explained by the fact that the subject was not limited to construct an alignment along a line in space when using a HMD, but instead could translate freely with compensatory head rotations to maintain alignment between the headslaved marker and the background marker. This was something we decided to study in combination with the aiming precision.

Our second metric, the aiming precision, was calculated by finding the average deflection of a unicorn vector from its average position, hence the (Average) Angular Distance from Center Direction is calculated like an angular standard deviation. As seen in Figure 6.4, we found a significant difference between the viewing conditions. During the Boresight condition, the average aiming precision was measured to be $0.25°$.

![Figure 6.4: Angular Distance from Center Direction for the three viewing conditions.](image)

The aiming precision varied depending on the aiming direction. The statistical analysis showed a significant difference in aiming precision between azimuth angles $0°$-$30°$, and $0°$-$60°$, but not between $30°$-$60°$. Subjects' aiming precision straight ahead was $0.21°$, but $0.27°$ when the torso was turned $30°$ and $60°$ relative to the head. As shown in Figure 6.5, the aiming precision was mainly deteriorated in the horizontal
direction with increasing torso rotation.

Figure 6.5: Angular Distance from Center Direction in various aiming directions.

To enable subjects to rotate the head to markers placed at 60° azimuth in our limited lab space, we required our subjects to counter-rotate their torso over a point on the floor in front of the screen on which the background markers were located. After having compensated for said rotation, the stabilogram plots over the subjects’ transversal plane showed a pronounced anterior-posterior postural sway which appeared similar to the one observed in the previous study. Since a correlation between translation and rotation of the head had already been seen, and the current study has an unusually long Sway Path associated with the Boresight viewing condition, investigation of a relationship between Sway Path and Angular Distance from Center Direction seemed appropriate.

Using the tracking data on head rotation and translation, a procedure was created to compensate for the subjects’ head rotation, which we hypothesized occurred mainly due to head translation, be it postural sway or minute voluntary adjustments. One can think of the model as the subject having a laser beam emanating from the eye, leaving a burn pattern around the background marker. Since the background marker is located at a particular distance, the spread of the burn mark can be recomputed to an angular distribution using trigonometry, which in turn is interpreted as the resulting aiming precision. The results indicated that subjects’ aiming precision approached 0.01° which implied that subjects were able to visually make sub-pixel judgments, but
also needed to compensate for the postural sway. As seen in Figure 6.6, the artifacts in the horizontal direction were not very prominent either.

![Figure 6.6: Angular Distance from Center Direction in various aiming directions after head rotation compensation.](image)

With the possibility to correct aiming precision for postural sway, we returned to the aiming precision dataset to make a temporal analysis. Once again dividing the 30 s recording period into 10 three second bins we could no longer see a pronounced quadratic trend, see Figure 6.7. This suggests that it is no longer necessary to wait for the subject to stabilize in order to obtain a reading with optimal precision when using head position and orientation data in conjunction.

### 6.3.3 Contributions

Head-aiming precision using an HMD is 0.21° straight ahead and 0.26° to the sides. Precision can be further improved by canceling the effects of compensatory head rotation. This procedure results in an isotropic head-aiming precision of 0.01° suggesting that human vision and current screen resolution are not limiting factors.
6.4. Paper IV

6.4.1 Aims

To achieve a good calibration the data must be not only precise, but also accurate. The previous studies had only been concerned with investigating the likelihood of consistent reports, that is precision. Therefore we were interested in trying to achieve a study of the bias between a point at which the subject is aiming and the location of the intended target, that is accuracy. To measure accuracy it is necessary to know the actual position of some reference points in the subject's surrounding. While the tracker origin could be determined, it was quite difficult to measure the location of the background markers in the tracker coordinate system, mainly because of their distance from the origin. Plumb lines, laser pointers and a theodolite were used, trying to merge the world and tracker coordinate systems, but it was not possible to define the background markers' coordinates in the reference frame of the tracker using manual methods. Instead we decided to use the boresight data from the previous study, define the middle background marker at \((0^\circ, 0^\circ)\) as reference, and renormalize...
CHAPTER 6. SUMMARY OF STUDIES

all datasets according to this reference point. Then we used head orientation data to measure how accurately the subjects were aiming at the background markers.

Subsequently, the aim of the fourth study was to begin to investigate the accuracy with which a subject can perform a visual alignment and see if there are any changes in bias or distribution with changing viewing direction.

6.4.2 Results

In Figure 6.8 the dataset from the 12 subjects’ three boresighting sessions are plotted over ±0°, ±30°, ±60° azimuth angle and 0°, ±10° elevation angle. The average (green circle) of all subjects’ data coincide with the center (red cross) of panel (0°,0°) since all data was normalized according to this reference.

Looking at Figure 6.8, one is tempted to interpret the vertical scatter and the slanted principal component at the extreme azimuth angles to be indicative of characteristic head mechanics. However, after labeling the individual datasets in Figure 6.9 it becomes visible that the vertical spread is not due to within-subject variability. Instead,
the between-subject is probably due to varying subject height and a mismatch between the subject’s eye and the ideal eyepoint for which the experimental setup was modeled. Theoretically, 6 cm difference in subject height results in approximately 1° offset in the (60°,10°) plot. Another plausible explanation to the scatter in Figure 6.9 is the varying orientation of how the HMD was worn by the different subjects.

![Figure 6.9: The individual contributions (subject ID:repetition) of head-aiming accuracy for direction (60°,10°).](image)

This study brought attention to the fact that even relatively small measurement errors make it difficult to perform a useful accuracy study. Ideally the background markers should have been expressed in tracker coordinates, but the limited range and non-linearity of the magnetic tracker prevented us from obtaining proper measurements. Furthermore, variations in how the HMD was worn affected the measurements. With these experiences noted for future studies, further study of the distribution of the datasets was carried out. However, since the frame of reference was too coarse for a complete study of accuracy, the remaining study of distribution bias fell back on a relative measurement which, once again, only denoted precision.

In terms of studying changes in distribution with changing viewing direction, we looked back to Table 4 in Paper III which already offered an analysis broken down into standard deviation as a function of specific azimuth and elevation angles. How-
ever, hoping to achieve greater statistical power with more data points, a compound analysis across all pitch angles and across all azimuth angles seemed potentially interesting. An ANOVA was performed which found significant effects of azimuth and elevation angle on pitch and yaw standard deviation. (The standard deviations in Table 1 of Paper IV corresponded with the average Angular Distance from Center Point in Table 4 of Paper III which was a reassuring sanity check.) However, despite aggregating across pitch and yaw angles the statistical effects were not convincing enough to be a practically useful result, and so the possibility of biases in the distributions was investigated using other statistical tools.

Since the opposite of bias is a perfect circle, we attempted to prove that the head-aiming distribution could be approximated as a circular distribution, beginning with checking if the dataset had any correlation between head pitch and yaw. The correlation is an indicator of an elliptical, or even linear, distribution. As seen in Figure 6.10, the mode of the correlation coefficient distribution was around zero, which indicated that a majority of the datasets were not elliptical.

![Figure 6.10: The distribution of the correlation between pitch and yaw head rotation.](image)

The distribution of the ratio between pitch and yaw standard deviation also indicated that the data was uniformly scattered around a center. As seen in Figure 6.11, the ratio distribution had a mode around one, which indicated that pitch and yaw had the same standard deviation in most of the datasets.

The third and last test verified that the two marginal distributions of the correlation coefficient and the standard deviation ratio were dependent. $\chi^2$ test was used, comparing observed and expected frequencies in the datasets and determined that the collected data statistically deviates from an independent distribution. The three tests prove that, even if the distributions appear horizontally scattered in Figure 6.5, they
6.4.3 Contributions

While the head-aiming precision is slightly biased, for all practical purposes it can be considered to be a circular distribution. The aiming precision, and subsequently aiming accuracy provided that absolute reference points are known, is 0.2°.

6.5 Paper V

6.5.1 Aims

Armed with some empirical data on the characteristics of noise induced by postural sway and head rotation, the fifth study aimed at studying the effects of such noise on the estimation of parameters in a pinhole camera model. This was particularly interesting since previous sensitivity analyses on camera models have not included noise at levels greater than 1 px standard deviation. It was also desirable to test the effect of noise relative to the number of correspondence points and relative to the depth over which these correspondence points were distributed. To achieve better control over the independent variables, this experiment was set up as a Monte Carlo simulation.
6.5.2 Results

The human noise was modeled as pixel permutations in the image plane in a direction which was randomized using a rectangular distribution. The magnitude of the offset along a direction in the image plane was in turn modeled using a normal distribution, a rectangular distribution, and a fixed magnitude. Thus all three noise models could be parametrized with range. The data generated with the noise model based on a normal distribution turned out to produce the most conservative results and was therefore chosen to be the one presented in the paper. Thus noise of 3 px range approximately corresponds with 1 px standard deviation.

We used 1,000 iterations in which the direction of noise as well as the location of the correspondence points were randomized. The independent variables were the number of correspondence points, the depth over which the correspondence points were located, and the level of noise. As visible in Figure 6.12, illustrating the estimated eyepoint location using 20 correspondence points, it is difficult to estimate the eyepoint along the line of sight (z translation) when the level of noise increases, particularly when the correspondence points are not distributed over any greater depth. The blue boxes denote interquartile ranges (=50% probability) and correspond to the detailed data presented in Table 1 in the paper. For example, one can conclude that if a subject were to aim at 20 correspondence points distributed over 1 m depth with noise equivalent to 10 px range (=3.3 px standard deviation) there is a 50% chance that the estimated eyepoint will vary by 14 mm along the line of sight.

From Table 1 in the paper it is also possible to see that by increasing the depth over which the correspondence points are distributed, we can drastically diminish the number of required correspondence points to achieve a comparable parameter estimation precision. For instance, it takes 81 points distributed over 0.1 m to achieve the same precision as 9 points produce over 1 m.

Lastly, with the exception of 9 or fewer correspondence points, we noted that the variance in all of the estimated camera parameters increases linearly as a function of noise.

6.5.3 Contributions

Distributing correspondence points over greater depth mitigates the effect of noise in the sense that fewer correspondence points are needed to achieve the same level of parameter estimation precision. Noisy correspondence point measurements primarily affect the eyepoint estimation along the line of sight. The sensitivity analysis can be generalized to any level of noise as the parameters deteriorate linearly.
6.6.1 Aims

In the sixth study the time had come to perform some practical experiments on calibration. In the pilot trials it was apparent that the manual acquisition of correspondence points was a cumbersome and tedious task. Therefore it was desirable to keep the number of visual alignment to a minimum. The previous study had shown that by distributing the correspondence points over a greater depth, fewer points would be needed to achieve a stable parameter estimation. Therefore, the aim of the study would be to evaluate three very different patterns in which subjects could acquire correspondence points, namely Static, Sequential, and Magic.

When a subject is focused on the task of performing a calibration exercise with a minimal amount of noise, the instinct to drastically change location from which to create the next alignment probably does not come naturally. Thus, a sequence of correspondence points are usually acquired from nearby locations. This behavior inspired the Static condition, see top Figure 6.13. If the user indeed decides to move, it is probably common to move along some structured pattern, for example along...
rows or columns of some calibration grid. This lead to the desire to test the Sequential pattern, see middle Figure 6.13. Lastly, it seemed advisable to test a pattern in which each correspondence point is as far away as possible from a linear combination of the others, while also optimally spanning the depth available for the calibration exercise. Such a pattern was found in the Magic distribution, see bottom Figure 6.13.

![Figure 6.13: The three correspondence point distributions as seen in a head-relative coordinate system. Top: Static. Middle: Sequential. Bottom: Magic.](image)

We realized that we needed a good metric which was easily observable and intuitive, even for someone not as experienced with the meaning and behavior of the intrinsic calibration parameters. Because the previous studies had illustrated the problem with measuring the offset between tracker sensor and eyepoint, it would be particularly interesting to see whether this parameter could be used as a metric and a dependent variable with which to measure the success of the calibration exercise.

In Paper III the problem of letting a subject pivot around a fixed ideal eyepoint in space was particularly apparent. Therefore it was decided to use SPAAM - a calibration procedure that relaxes the criterion of a consistent and stable eyepoint location. Furthermore, the use of SPAAM was interesting since this particular calibration procedure had not been evaluated for extrinsic parameters in an OST HMD in the past.

Lastly, we also realized that noise in correspondence points acquired close to degenerate surfaces would be amplified in an ill-conditioned calibration problem. Therefore we decided to study the condition number in addition to the variability in the other calibration parameters. However, it was decided to perform this analysis in the form of an offline simulation.

In short, the aim of the study was to let the subjects perform a SPAAM calibration acquiring the correspondence points in three different ways, and to measure calibration success in terms of a correctly estimated eyepoint location.
6.6.2 Results

The results of eyepoint estimation are plotted in Figure 6.14 which illustrates the HMD in a head-relative coordinate system. The actual eyepoints of the subjects were not recorded, but it is still plain to see that the Magic acquisition pattern (blue +) resulted in the most precise eyepoint estimate, followed by Sequential (black o). Eyepoint estimations made with the Static acquisition pattern were too far away to be plotted in the figure. The eyepoint estimations are also shown with an alternate illustration in the box plot of Figure 6.15. The eyepoint estimate along the z axis, a parameter known to be sensitive to noise, is notably less variant in the Magic condition compared with the other conditions. In the statistical analysis a non-parametric ANOVA confirmed a significant difference in parameter variance between acquisition patterns for all parameters except for rotation and horizontal principle point.

In the complementary Monte Carlo simulation it was noted that the condition number decreases with greater correspondence point distribution spread. The almost planar distribution in the Static condition showed a higher condition number than for the Magic condition where the correspondence point distribution is more variable. It was also noted that increased noise can actually lower the condition number.

The Monte Carlo simulation also indicated that the variability in the rotation parameter is not primarily dependent on noise, but rather the number of correspondence points. In Figure 6.16, where variance in the parameters are plotted using parallel coordinates, we noted that 64 correspondence points (colored in red) generally resulted in a lower variance in the rotation parameter (eighth axis from the left) compared to a low number of points (colored in blue) which resulted in higher parameter estimation variance, regardless of whether the Magic or Sequential pattern was used.

Lastly, the simulation also provided data which made it possible to plot the relative improvement in parameter variance as a function of the number of correspondence points used. Figure 6.17 shows the variance in estimated eyepoint for calibrations subjected to 3 and 5 px noise. Using the Magic pattern with 25 instead of 16 points results in an improvement by a factor two and this trend continue up to 81 points. However, with the Sequential pattern, there is a drastic improvement in parameter variability going from 9 to 16 points. This suggests that a previously published evaluation of the SPAAM procedure [126], in which only nine calibration points were used, did not achieve the best results possible.

6.6.3 Contributions

Acquiring correspondence points which are variant in depth improves parameter estimation precision for SPAAM. The exception is the rotation parameter which depends mainly on the number of correspondence points rather than noise level or acquisition pattern. Calibrating with an acquisition pattern which does not have correspondence
points well distributed in depth is possible, but the parameter estimation will be more sensitive to noise. Parameter estimates start converging at 36 correspondence points for representative noise levels, whereas 16 points is sufficient for acquisition patterns which do have correspondence points well distributed in depth. The condition number of a calibration procedure should not be used to evaluate the calibration quality of a particular calibration session as noisy readings may, in fact, lower the condition number. Instead the condition number should be used to compare calibration methods, since methods with correspondence points well distributed in depth will, on average, have a lower condition number than others.

6.7 Paper VII

6.7.1 Aims

While the previous study provided several useful results, the poor precision, and consequent limited practical use, of the eyepoint estimation (illustrated in Figure 6.14) was somewhat puzzling. Results recorded in practice were much more variant than those observed in the ideal environment of simulations. At this point it seemed questionable if the pinhole camera model published in the original paper on SPAAM[132] was indeed the right choice. One last attempt, applying everything learned so far mainly through experiments, but also from reviewing all other OST HMD experiments found in the literature, was made. The subject collected correspondence points with a variant depth distribution; the correspondence points were measured in the same frame of reference as the tracker; seated subjects were used to minimize equipment slippage; the noise introduced by the tracker was controlled; and an eyepoint measurement device which reported in the tracker coordinate system, was constructed. Achieving an experimental setup which could report on accuracy had proven be an elusive task, but in this study the only remaining free parameter would be the measurement error as the experimenter cupped the subject’s eye with the measurement device to verify the calibration (see middle panel Figure 6.18). Thus, this experiment had the potential to deliver accuracy measurement within a few millimeters - provided that the pinhole camera was indeed the right model.

The previous studies had gradually moved focus from the noise in the correspondence point acquisition process to computational challenges in the calibration problem. This particular study concerned a computational step in the calibration process known as matrix decomposition in which the camera matrix is divided into its extrinsic and intrinsic parameters. Of particular interest were compare two matrix decomposition methods which had been developed based on the assumption that the camera matrix was indeed compiled in a pinhole camera model. The results from the two decomposition methods would, hopefully, reveal whether the pinhole camera was a suitable
6.7. PAPER VII

6.7.2 Results

Column 15 of Table 1 in the paper illustrates the difference between the calibrated and measured eyepoints of the nine subjects’ three repetitions. The difference was coded into a Euclidean distance, $t_{diff}$, which on average was found to be 5 cm across all subjects. In terms of practically useful results, this was a disappointment, but in terms of clues for future research the results were important.

A closer look revealed that the aiming precision in seated subjects was 1.7 px (APP, column 7) which was an improvement on the standing subjects in the previous studies which measured 4.3 px standard deviation. Hence, these calibration sessions were of better measurement quality than before. We also compared the depth range over which the correspondence points were collected (R, column 8) to the resulting calibration accuracy, $t_{diff}$, and found no correlation. Thus, none of the parameters previously shown in simulations to have a positive effect improved the calibration, suggesting that the root of the problem may be in the modeling and calculation part rather than in the measurement and manual procedure.

Examination of differences between the three components along the x, y, and z axis, showed that the measured eyepoint was consistently lower (along y) than the calibrated eyepoint. Since this phenomenon was consistent over all 27 datasets, this measurement will be an interesting indicator when testing alternative models in the future.

In the ideal case, without noise and with a perfect fit between calibration model and reality, the two matrix decomposition methods should have produced identical results. Table 2 in the paper shows that there is a considerable difference in results for the two methods. This confirms that the decomposition step in the calibration procedure was a source of error in the pinhole camera model at current noise levels.

6.7.3 Contributions

In an experiment using a human operator for all stages, measuring the eyepoint within the same reference frame as the tracker, the accuracy of the SPAAM algorithm in an OST HMD is 5 cm. The parameter estimates are affected differently depending on the matrix decomposition method used. This confirms the decomposition step to be a source of error, possibly due to a poor fit between reality and the calibration model.
Figure 6.14: The estimated eyepoints for Sequential (black o) and Magic (blue +) condition superimposed on the HMD pictured from above. A plausible eyepoint is in the vicinity of (0.25-0.35, -0.07, -0.20).
Figure 6.15: Variance of the estimated eyepoint for the Static, Sequential and Magic acquisition pattern.

Figure 6.16: Parameter variance plotted in parallel coordinates. Red and blue lines denote 64 and 9 correspondence points respectively. Distribution type has been restricted to Sequential and Magic, and noise has been restricted to 1-5 px.

Figure 6.17: The relative improvement in parameter variance as more correspondence points are used in the three acquisition patterns.
Chapter 7

Discussion

This thesis has involved four user studies and two simulations in seven papers. The following section will discuss the conclusions of the individual papers in a collective perspective, comment on some of the challenges, and describe how the presented work could be continued in the future.

7.1 Main Conclusions

7.1.1 Postural Stability

• A calibration task based on visual alignment deteriorates postural stability in standing subjects.
• Translational head-aiming precision in standing subjects calibrating a HUD...
  ...is 16 mm, but improves to 11 mm after 12-15 seconds.
  ...can be modeled with a Weibull distribution.

Observing the surroundings through one's own eyes during quiet standing, it might be thought that the human body would make a good platform to calibrate from. However, since balance and stable vision are governed by a set of unconscious reflexes, it is generally quite hard to appreciate the dynamic nature of human pose from within one's own body. Only after beginning to experiment with maintaining a boresight over two reference points that it became obvious that the human body is practically always in constant motion. Therefore, it was very useful to start with experiments studying postural sway to gain an understanding of the human body as a calibration platform.

In the state of concentrated aiming most subjects became aware of the relative notion of balance and had a tendency to compensate for subtle movements caused by breathing or heartbeat. Some subjects also exhibited facial spasms as they concentrated on
the aiming task. Covering the non-dominant eye helped the subjects relax. As the room was darkened to limit the effects of peripheral vision and the subjects were presented with only two abstract markers, it was obvious to see, also for a bystander, how postural stability deteriorated noticeably. The postural sway was pronounced in the anterior-posterior direction (see Figure 7 in Paper II), probably due to the bipedal stance, and seemed to originate from the ankles. Autocorrelation of postural sway was investigated by there was insufficient support for periodic head translation to attempt a model.

The postural sway in the anterior-posterior direction resulted in an oblong distribution. Fitting of a bivariate Gaussian distribution to capture this quality was tried, but the kurtosis was too sharp for a good fit. This motivated the use of a Weibull distribution, despite the fact that it reflects an omnidirectional measurement which does not describe the directional properties of the distribution.

### 7.1.2 Head-Aiming Precision

- Orientational head-aiming precision in standing subjects calibrating an OST HMD...
  
  ...is 0.21° straight ahead and 0.26° in directions ≥ 30° azimuth.
  
  ...is not biased and can be approximated with a circular distribution.

- Head-aiming precision in standing subjects calibrating an OST HMD can be improved by considering postural sway and compensatory head rotation together, resulting in a precision of 0.01° which proves that vision and screen resolution are not limiting factors for OST HMD calibration.

- Orientational head-aiming precision in seated subjects calibrating an OST HMD is 0.09° in directions ≤ 15° azimuth.

Whereas the experiment in Papers I and II had been conducted with a HUD, the experiment in Papers III and IV used an HMD. It was interesting to see that both displays induced a postural sway distributed in the anterior-posterior direction (compare Figure 7 in Paper II with Figure 3 in Paper III). It was also noted that the HUD mainly required head translation for alignment, while the HMD was easier to align using just head rotations.

Prior to the second experiment, it was speculated that the typical shape of the distribution was due to the HUD, and the fact that a subject had to translate the eye into alignment along a line fixed in space. With the free movement of an HMD, it was anticipated that the distribution would broaden as a result of translations to the sides and subsequent head rotations. However, data showed that a subject calibrating an HMD remained over the equivalent area compared to a HUD, but that the distributions over time were different. This was illustrated by the fact that the Sway Path for
7.1. MAIN CONCLUSIONS

HMD was almost twice as long compared to HUD. Interestingly, the longer Sway Path was also noticeable for the viewing condition when the subject's eyes were closed. This suggests that deteriorated postural stability was due to the weight of the HMD and not necessarily due to the limited FOV. In previously cited sources, it is rather the importance of a well-balanced HMD than the actual weight which is emphasized. This was an interesting observation which it was not possible to pursue any further without a new experiment since the experiment did not include a condition in which the subject did not wear an HMD.

The second experiment also revealed a correlation between head translation and head orientation. The correlation was illustrated by the isotropic distributions of Figure 7 in Paper III where the effective aiming precision was calculated by removing the compensatory head orientation. The resulting precision corresponded to less than the width of a pixel. This was a surprising result, but still quite possible as it is three times greater than human visual acuity. This observation raised the question of trade-off between display resolution and the precision of the alignment mentioned in Wayne Piekarski's thesis on outdoor AR worlds (p. 83). Contrary to Piekarski's reasoning, our subjects appeared not to completely overlap the foreground and background marker, but instead used a Vernier-like approach in which they aligned marker edges. Thus, screen resolution is not a limiting factor on head-aiming precision using an HMD.

Instead, the limiting factor for head-aiming precision is the postural sway. This is illustrated by the fact that it was possible to cancel out the effect of postural sway, which suggests that the subjects were guided by vision. To investigate whether compensatory head rotation is governed by visual signals or vestibular reflexes, I would propose an experiment where a subject is situated in a panoramic environment to fixate some correspondence point in the distance. In this setup, a head translation requires only a very small head compensatory head rotation. If no compensatory rotation is registered, the experiment would support the theory that compensatory head rotation is guided only by vision.

7.1.3 Parameter Estimation

- Pinhole camera parameter estimation variance...
  ...increases linearly as a function of alignment noise.
  ...increases with diminishing correspondence point distribution depth.
  ...decreases optimally if 25 or more correspondence points are used.
  ...decreases for all parameters with increasing correspondence point distribution depth variance, except for rotation, which primarily depends on the
number of correspondence points.

...can be predicted with a condition number, but only to compare calibration methods.

Although quite a simple method, Monte Carlo simulation was very helpful providing answers that could not be found in existing literature. In particular, a study certain combinations of independent variables in camera algorithm sensitivity analysis as needed, which seemed not to have been published. For instance Faugeras’ sensitivity study (p. 64 [44]) included noise up to 3 px standard deviation, but did not report how the correspondence points were distributed or how many were used in total. Sun and Cooperstock [120] used far too many points compared to our setup and only included noise up to 1 px standard deviation. In this regard, Monte Carlo simulation was a practical tool for pilot studies.

It is also important to realize, however, the limitations of Monte Carlo simulation since the results are based on models which only approximate reality. This is why it was necessary to confirm our simulation findings with live experiments using human subjects. This standpoint is argued in greater detail in section 1.3 of Paper VII.

The first paper using simulation, Paper V, concluded that the distribution of correspondence points in depth has a significant impact on calibration results. Interestingly, this variable has never been controlled in any of sensitivity analyses found in the literature. This variable is particularly important when only a few (<100) correspondence points are available. For future researchers who wish to control this variable, in simulations or experiments, I recommend the use of the method described the appendix to Sutherland’s paper [124] and solving for the depth parameter, Z. Arranging the equation in this form allows the experimenter to parametrize the location of the correspondence point with two screen coordinates and a distance, which is both easier and more accurate compared with most other combinations of trigonometric operations tried in this work.

### 7.1.4 The Pinhole Camera Model

- Pinhole camera parameter estimation accuracy...
  
  ...depends on the camera matrix decomposition method used.
  
  ...is on average 5 cm for eyepoint estimation using SPAAM for seated subjects with an OST HMD, ...
  
  ...which suggests that the pinhole camera model might not be ideal for an OST HMD.

Figures 5 and 6 in the original paper by Tuceryan and Navab illustrate the calibration quality of the SPAAM algorithm [132]. It is clear to see that these results would not have been possible with the magnitude of eyepoint estimation errors reported in
7.1. MAIN CONCLUSIONS

Paper VI and VII. So, wherein lies the difference between the this work and that of Tuceryan and Navab?

The first difference is that Tuceryan and Navab do not decompose the camera matrix into intrinsic and extrinsic parameters. After a linear estimate of the projection using DLT, the entire camera matrix is passed to the Modelview matrix stack of the Open Graphics Library (OpenGL) Application Programming Interface (API). This means that the camera matrix is never subjected to the errors introduced by the matrix decomposition methods discussed in section 4.2.3. Only when correspondence points are free from measurement error, and the relationship between the eye and the image plane corresponds to a perfect pinhole camera, will this decomposition step be error free.

The decomposition step can be motivated with two arguments. Firstly, the OpenGL API has two separate matrix stacks\(^1\). The Projection stack is intended for the intrinsic parameters and the Modelview stack is intended for the extrinsic parameters (p. 138-139 [146]). While the two matrices are eventually multiplied back together again towards the end of the rendering pipeline (p. 98 [146]), any operations that are intended to be made individually either subset of parameters will not have the desired effect. Such operations could be on depth tests and vertex culling operations. The second argument is based on the reasoning of section 5.1, namely that a registration error can only be adjusted if the projection is broken down into its individual camera model parameters.

The second difference is that the i-glasses used by Tuceryan and Navab do not have a documented cant angle. The Kaiser ProView 50ST HMD used in the experiments in this thesis has a measured cant angle of 4\(^\circ\). A cant angle pivots the frustum around the eyepoint and is implemented in systems which are designed to offer binocular overlap to produce a wider FOV or higher resolution, see section 3.2.1. Thus, while both HMDs can depart from a pinhole camera model, the Kaiser ProView 50ST is more likely to have an image plane which is not perpendicular to the principal axis as described in section 4.3.

The third difference is that figures 5 and 6 in the original paper by Tuceryan and Navab are the result of a calibration procedure during which the HMD was worn by a mannequin mounted on a tripod [130]. Thus, the measurement noise in Tuceryan and Navab’s evaluation is not representative for a human operator, whereas the evaluations in Papers VI and VII are made with human subjects.

\(^1\)A matrix stack is a last-in-first-out data structure which is suitable for organizing relative transformations between coordinate systems.
7.2 Main Challenges

7.2.1 Measurements

A principal challenge throughout the work in this thesis has been related to measurements in one way or the other. Tracker sensors introduced noise which is not visible until a sample of the signal was studied separately. Unless a reference movement was known (such as a pendulum), only the static properties of the tracker could be observed. No attempt was made to measure the tracker performance in motion.

In the case of a magnetic tracker, the nonlinear measurements and limited working volume were two large obstacles. No attempt was made to correct the nonlinearities, but instead an estimation of the working volume in which the tracker would provide accurate reports was made. This was a very tedious task which is not recommended. The experimental setups became much easier once an optical outside-in camera-based tracking system with active markers was obtained, since it had a substantially larger working volume while still being very precise. It is essential to keep all entities involved in a calibration procedure within the same frame of reference as manual measurements introduce measurement noise. On that note, camera-based tracking systems still need some tuning in order to achieve precise measurements.

Many of the parameters of the pinhole camera model, particularly the intrinsic parameters, are not easy to measure and therefore difficult to validate. This is unfortunate as a shift in principle point is as detrimental to registration as a frustum rotation (or translation at close range). So far this work only focused on the estimate of eyepoint location, but there are several parameters left for which measurement techniques have to be devised.

7.2.2 Relative Coordinate Systems

Every time the HMD was put on by a subject, the orientation of the tracker sensor would be different. This meant that the offset vector between the head-relative tracker origin and the subject’s eye would act as a lever which drastically changed the eyepoint location estimate as the HMD was donned differently between sessions. For the same reason, HMD slippage during sessions was detrimental to calibration results.

In Paper VI, Figure 6.14, this was solved by recording the position of the rigid body markers for the first subject, and then estimating a compensatory rotation to align all other subjects’ data into this frame of reference. By contrast, no such compensation was implemented in Paper IV, which is why Figure 6.8 has a notable vertical spread.

If one coordinate system needs to be mapped to another, it is my experience that direct measurements are far too noisy to provide an accurate mapping. Instead a 3D-to-3D transformation matrix should be calculated using the statistical power of
many individual measurements. Alternatively, if only a limited number of samples are available, a least-square optimization might be sufficient. Any other approach, particularly when measurements involve orientation, has been unsuccessful. A discrepancy between two coordinate systems can be seen in Figure 6.8 where the average cluster centers (green circle) gradually sink from left to right, indicating that the tracker coordinate system was not level with the world coordinate system in which the background markers were defined.

7.2.3 Collaboration

Studying calibration procedures is a job for two persons. As a calibration is performed, it is only valid for that session as removing the HMD and putting it back on changes the eyepoint location relative to the display surface. Because of the relatively short distance between eye and display surface, the lever effect is quite powerful. Therefore, I have found it very useful to have one person performing the calibration, and have another person by the computer, making the necessary changes.

I also developed a mannequin with camera eyes, see Figure 7.1, but despite the appropriate shape onto which to rest an HMD, the brace always slipped which corrupted the calibration. Furthermore, since the mannequin is static on its tripod, one had to move the background marker instead of the mannequin’s head. Unfortunately, since the acuity and dynamic range of the cameras is much lower than those of the human eye, making visual alignments using the cameras was very imprecise. Moreover, the static mannequin setup could not model human noise as it did not exhibit representative postural sway.

7.3 Future Work

7.3.1 Pinhole Camera Model or Not?

The most recent studies presented here seem to suggest that the pinhole camera is not an appropriate model for the OST HMD. The alternatives are to either achieve a setup in which the pinhole camera model is acceptable, or to adjust the model such that it produces less variant parameter estimates.

These studies have not fully explored the possibility of letting the calibration start with some fitting procedure in which the principal point is matched to the optical axis of the eye, and the image plane is ensured to be perpendicular to the optical axis. Possibly this could be achieved by illuminating the center pixel of the HMD and having the subject look straight into a perpendicular mirror until only one bright spot is visible. The perpendicularity of the image plane could be confirmed by displaying a rectangle, for example, whose corners are to be matched with some real world
counterpart. We have not been successful in matching large geometries to real objects since it probably requires all corners to be foveated simultaneously, but it is possible that there are solutions using the peripheral view.

Another option is to adjust the existing pinhole camera model with a homography which describes the currently unknown transformation between the perpendicular image plane of the pinhole camera and the arbitrarily rotated display surface. While the theoretical framework is already available in projective algebra, the challenge lies rather in designing a calibration procedure which allows us to estimate the camera matrix and the homography independently. It is not until the camera matrix is defined on its own that it can be decomposed into extrinsic and intrinsic parameters.

Perhaps eyetracking could help? It would have the benefit of continuously tracking the location of the eye such that the calibration would be resilient to moderate equipment slippage. It would, however, require that the eyetracking camera was mounted such that it remained rigid compared to the display surface. Furthermore, the camera would have to be calibrated relative to the said display surface, and this would not be an easy task given the close quarters of the HMD viewport.

7.3.2 Objective Registration Error Measurement

While the calibration quality should be discussed in terms of the model parameters which estimate the projection, the registration error is probably the quality measure-
ment most intuitive to the user. However, in an OST HMD the size of the registration error can only be measured subjectively. Provided that it is possible to achieve a good calibration, it would therefore be valuable to perform a study in which the registration error is measured by some external device, such that the error can be objectively quantified, and comparable across subjects. At present, no such study has yet been done.

7.3.3 Optimal Correspondence Point Distribution

Papers V and VI studied the importance of correspondence point distribution in depth. In Paper VI it was concluded that if the points were distributed according to a Magic distribution, parameter estimation variance would be significantly lower. Also fewer points were required to achieve parameters with comparable variance when using a point distribution variant in depth (Magic), compared to when points that were sequentially distributed in depth (Sequential) were used. However, no study was made of the particular statistical properties which made Magic preferable to Sequential to any greater extent.

Figure 7.2 shows the depth distribution of correspondence points as they are acquired in a Sequential and Magic order. Depth Position is the distance between the correspondence point from the subject’s eye and Crosshair Column refers to the columns of foreground crosshair markers over which an alignment is made. In this plot 25 correspondence points were used, hence five columns of five points each.

It is clear to see that the average depth and depth variance per crosshair column is more evenly distributed for the Magic distribution compared to the Sequential distribution. It is possible that there is an optimal correspondence point constellation with which a minimal amount of points can be acquired while still achieving optimal calibration quality. This result would ease the laborious task of calibration, and also be a generally useful result for the Computer Vision domain.
Figure 7.2: A top view of correspondence point (black circles) distributions and the relationship between average depth (green solid line) and depth variance (blue dashed line) per column of foreground crosshair markers.
Bibliography


[53] H. Fuchs, J. Duran, and B. Johnson. A System for Automatic Acquisition of


BIBLIOGRAPHY


Acronyms

AFHRL  Air Force Human Resources Laboratory.  12
ANOVA  Analysis of Variance.  68, 70, 78, 83
API  Application Programming Interface.  95
AR  Augmented Reality.  3–5, 7, 13–15, 17, 20, 21, 23, 24, 26, 27, 30, 31, 40, 52, 55, 56, 58, 60, 93
CAE  Canadian Aviation Electronics.  12
CAVE  CAVE Audio Visual Environment.  52
CCD  Charged Coupled Device.  10, 18, 19
CMOS  Complementary Metal Oxide Semiconductor.  18, 19
CRT  Cathode Ray Tube.  8, 9, 12, 27
DLT  Direct Linear Transformation.  42, 44, 50, 51, 64, 95
DOF  Degrees of Freedom.  23, 41, 44, 47, 56, 58
EKF  Extended Kalman Filter.  57
FOV  Field Of View.  24, 26, 28, 31–33, 38, 51, 59, 71, 72, 93, 95
GDOP  Geometric Dilution of Precision.  22
GPS  Global Positioning System.  20, 22
HCI  Human Computer Interaction.  3, 8, 9, 12
HFES  Human Factors and Ergonomics Society.  26
HMD  Head-Mounted Display.  4, 5, 7, 11–13, 17, 18, 24–31, 33, 38, 46, 50, 52, 55, 57, 58, 61–64, 71, 72, 74, 77, 82–84, 86, 92–98
HMS  Head-Mounted Sights.  32
HUD  Head-Up Display.  24, 26, 29, 67, 71, 72, 91, 92
Acronyms

**IMU**  Inertial Measurement Unit. 20, 22
**IPD**  Interpupillary Distance. 24, 25, 29, 52, 53

**KARMA**  Knowledge-based Augmented Reality for Maintenance Assistance. 14

**LCD**  Liquid Crystal Display. 29
**LED**  Light Emitting Diode. 20, 30
**LEEP**  Large Expanse, Extra Perspective. 13, 26
**LMA**  Levenberg-Marquardt Algorithm. 44
**LOD**  Level of Detail. 13
**LUT**  Look-Up Table. 19

**MIT**  Massachusetts Institute of Technology. 8, 10, 11, 13

**OpenGL**  Open Graphics Library. 95
**OST**  Optical See-Through. 4, 5, 24, 27–29, 31, 38, 46, 50, 52, 55, 58, 61, 62, 64, 82, 84, 86, 92, 94, 98

**PPF**  Perceptual Post-Filter. 22

**RANSAC**  Random Sample Consensus. 48–50
**RFID**  Radio Frequency Identification. 20

**SA**  Situational Awareness. 11, 12
**SLERP**  Spherical Linear Interpolation. 47
**SLR**  Single Lens Reflex. 30

**SPAAM**  Single Point Active Alignment Method. 60, 61, 82–84, 86, 94

**SPIE**  Society of Photo-Optical Instrumentation Engineers. 26
**SRI**  Stanford Research Institute. 8
**SVD**  Singular Value Decomposition. 42

**UNC**  University of North Carolina. 10, 11, 13, 17, 18, 20, 23, 57

**USAARL**  United States Army Aeromedical Research Laboratory. 12, 29
**USAF**  United States Air Force. 12

**VCASS**  Visually Coupled Airborne Systems Simulator. 12
**Acronyms**

**VCS** Visually Coupled Systems. 12, 32
**VE** Virtual Environment. 5, 7, 18, 20, 23
**VR** Virtual Reality. 14, 15, 23, 26, 52, 64
**VST** Video See-Through. 27–29, 31, 46, 58, 61