A mediated reality suite for spatial interaction

Symbiosis of physical and virtual environments for forensic analysis

Poelman, Ronald

DOI
10.4233/uuid:5fc08214-3d74-434d-ba02-097d221e26ea

Publication date
2017

Document Version
Final published version

Citation (APA)
Poelman, R. (2017). *A mediated reality suite for spatial interaction: Symbiosis of physical and virtual environments for forensic analysis*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:5fc08214-3d74-434d-ba02-097d221e26ea

A Mediated Reality Suite for Spatial Interaction

Symbiosis of Physical and Virtual Environments for Forensic Analysis

Ronald Poelman

A mediated reality suite for spatial interaction

Ronald Poelman

A mediated reality suite for spatial interaction

Symbiosis of physical and virtual environments for forensic analysis

Proefschrift

ter verkrijging van de graad van doctor

aan de Technische Universiteit Delft,

op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,

voorzitter van het College voor Promoties,

in het openbaar te verdedigen op

maandag 4 december 2017 om 10:00 uur

door

Ronald POELMAN

Master of Science in Engineering Product Design,

Open Universiteit Nederland,

geboren te Zeist, Nederland.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr .ir. A. Verbraeck

Copromotor:

Dr. S.G. Lukosch

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. A. Verbraeck	Technische Universiteit Delft
Dr. S.G. Lukosch	Technische Universiteit Delft

Onafhankelijke leden:

Prof. K. Kiyokawa	Nara Institute of Science and Technology
Prof. dr. D.K.J. Heylen	Universiteit Twente
Prof. dr. M. A. Neerincx	Technische Universiteit Delft
Prof. dr. ir. P.P. Jonker	Technische Universiteit Delft
Prof. dr. B. A. Van de Walle	Technische Universiteit Delft

Cover design: Ronald Poelman

ISBN: 978-94-028-0880-3

Author email: ronaldpoelman@gmail.com

Acknowledgment

My life altered significantly during the writing of this thesis, I suddenly found myself working in San Francisco without properly finishing this dissertation. Fortunately, a handful of remarkable people are responsible for its existence and I would like to thank them for not giving up on me.

Alexander, I wouldn't even have been excited about academia without your enthusiasm and exquisite academic mind. Even after moving to the United States you took the effort to look me up and gently nudge me towards finishing this dissertation. I have learned so much and you can take most of the credit. I hope we can keep our conversations going after this milestone. I`m very appreciative toward my co-promotor Stephan, you kept my focus on the task, it was a pleasure to work with you, your constant reminders kept me from swaying.

Sometimes we are lucky enough to meet people that change the course of events in the best possible way, in my case that's meeting my fellow Ph.D. student Oytun Akman. My big ideas would never have been realized without your deep knowledge and academic mind. I thoroughly enjoyed our discussions and learned so much that it still propels me today. Because of you I`m able to work in the domain I enjoy so much.

I was lucky enough to have a great roommate. Martijn, you always gave me good advice and made me understand what an academic is all about. Furthermore, without Jonatan's software development skills, this thesis would not exist, what I lacked he made up for. I cannot thank him enough for all the hard work and great conversations.

The use case of this thesis is developed with two people; Hans and Jurrien. I enjoyed writing the proposal and encouragements during the execution. The experience helped me to write many more proposals and would like to thank you both.

I am especially grateful to the members of my promotion committee for investing their precious time to comment the green light version of this thesis. Thank you for the good critiques and interesting discussions.

The ones who probably suffered the most are my wife Caroline and my kids, Thijs and Mark. I stole holidays, evenings and weekends to be able to finish this thesis. Caroline, thank you for believing in me and allowing me to indulge in this academic research. I love you and will try to make up for lost time!

CONTENTS

ABBREVIATIONS

Abbreviation	Description	First appears in section
BIM	Building Information Model	1.3
BMI	Brain–Machine Interface	4.5.1
CAD	Computer Aided Design	1.1
CAM	Crime Analysis Meeting	2.2.2
CCD	Charge-Coupled Device	1.1
CMOS	Complementary Metal-Oxide Semiconductor	4.2.3
GIS	Geographical Information Systems	1.3
GPS	Global Positioning System	1.1
HMD	Head-Mounted Display	4.1.4
ICP	Iterative Closest Point	4.2.7
LTE	Long Term Evolution	1.3
NFI	Nederlands Forensisch instituut	1.2
MEMS	Micro Electro Mechanical Systems	1.3
MRI	Magnetic Resonance Imaging	1.3
OLED	Organic Light Emitting Diode	1.3
SFM	Structure from Motion	4.2.5
SLAM	Simultaneous Localization and Mapping	4.2.5
SWIR	Short Wave Infrared	4.2.3
USB	Universal Serial Bus	1.3
UVC	USB Video Class	5.5.2
VRD	Virtual Retinal Display	4.4.3
WIMP	Windows, Icons, Menus, Pointer	1.3

A Mediated Reality Suite for Spatial Interaction

Symbiosis of Physical and Virtual Environments for Forensic Analysis

Ronald Poelman

1. INTRODUCTION

Our digital world provides seemingly limitless opportunities, which are not easily replicated in the physical world. Would it not be convenient if software and hardware solutions could leverage the best of both worlds? Fortunately, there are technologies that allow us to bring the physical and virtual world closer together. Connected smart devices with a multitude of sensors enable our digital world to better understand the physical world. The digital map of the physical world is continuously improving and can be accessed from anywhere; that information can be used to analyze the physical world or to digitally transport people to any mapped environment.

This thesis concerns an interface to physical environments that benefits from the competence that exists in the digital world. Our main interest is to digitally support professionals in pristine environments with spatial related tasks.

This chapter will preface the motivation for this research and outline the challenges. Relevant concepts and domains will be introduced before the research question is presented, after which the research approach and philosophy will be discussed.

1.1. CHALLENGE

People capable of handling information intensive tasks are in high demand. While for some tasks, it is sufficient to make use of information sitting at a desk, increasingly, people need information 'on premise' that is relative to the context of the environment. If the context of the environment is considered next to the incoming task, the complexity of the information grows. A well-known example of this complexity comes from the military, i.e., "friendly fire" in which information data and geographical data needs to be combined on the fly to prevent casualties (Blair & Johns, 1993). Another example, this time from the medical domain, are laparoscopic surgeons, who rely on the merging of body scans, the position of surgery instruments and video imagery (Botden & Jakimowicz, 2009). Systems that are similar but less demanding are electronic location aware museum guides, car navigation systems and housekeeping robots. The knowledge of a scene, its context, is therefore of increasing importance. Hence information is more frequently coupled with a geospatial reference. Nowadays, even when taking a photograph with a digital camera, a geographical tag is automatically added.

Our present-day society is witnessing an explosion of information and knowledge, available to many, and an increasing complexity of subject matter in many domains. Information has never flowed so rapidly and in such large quantities (Castells, 1996; Flew, 2009). The changes in the amount and complexity of knowledge and information, as well as changes in requirements for coping effectively with increasingly complex tasks, challenge us to find solutions. Today, we are still bound to a desktop environment for our work, while the data we are working with tends not to be workplace bound, but virtually available. Smartphones have powerful capabilities that enable us to do tasks that recently were only possible in a desktop environment. But smartphones are also handicapped - they make use of ~2% of our visual field and currently only work with pre-created context information, while, by contrast, 50 % of the cerebral cortex is used to interpret visual stimuli (Milner, 1998). Spatial related tasks are generally not yet associated with mobile computing, but the sensors and capabilities are under development (Klein & Murray, 2007). Spatial related analyses are conducted by professionals on a daily basis. The information is used for various goals; maintenance, simulation, design, leisure or for practical things such as property and tax measures.

There is today an abundance of use cases, but very few solutions. The army, for example, would like to be able to detect changes in the road to find possible locations for road mines and be able to compare the data from previous patrols to the real-time collected data, preferably visualized in a 3D overlay. Just as important is the potential for medical applications: hybrid digital and physical 3D data sets are regularly used in brain and laparoscopic surgery. Common to both examples is the urgency for correct 3D data and the acquisition at interactive speeds. It is just as essential for forensics, which deals with crime scenes in different states. Airplane maintenance people have piles of manuals that provide information about the hyper complexity of an airplane, but going through piles of physical information is very inefficient. Architects currently render images of their design and paste the results in photographs. Providing a view of a newly designed building in its physical world context, while being able to freely walk around on the site can furnish valuable information. And by extension, for heritage purposes, too: reliving the former glory of a historic site that is currently an old ruin appeals to experts and novices alike. The visualization in a real-world context of the design or historic reconstruction can help to detect design flaws, provide new insights, create a shared understanding, situational awareness and serve various communication means, among other things. What all scenarios have in common: physical world context, an ability to communicate, visualization, data capture and analysis.

Currently, many in-between steps must be accomplished to get from the physical world through the digital world and back to the physical world, which introduces noise in communication and data quality loss due to abstraction. Asynchronous data streams are inevitable and influence decision making. Because many disciplines are involved, there is a high risk of confusion and misunderstanding. The current workflows can therefore be said to be downright cumbersome: specific equipment must be brought in to capture the environment (i.e.; room, building, manufacturing plant), processing must be done with very specific computer aided design software (CAD) which requires elaborate training.

"Personal computers have evolved in an office environment in which you sit on your butt, moving only your fingers, entering and receiving information censored by your conscious mind" (O'Sullivan & Igoe, 2004). Mankind is forced to understand the ways of a computer, to interact with it, while a computer can be fashioned to understand our vocabulary of communication to a much higher degree. Much like Alan Cooper's "The inmates are running the asylum", in which, describing the terrors of bad software, a dancing bear is compared to modern day software: people are so excited by the dancing bear that they don't notice how well the bear is actually dancing. It is the author's opinion that a computer should be able to support us more naturally, even on location, with spatial tasks. To accomplish this, the computer should be able to sense the environment, not merely use the digital context.

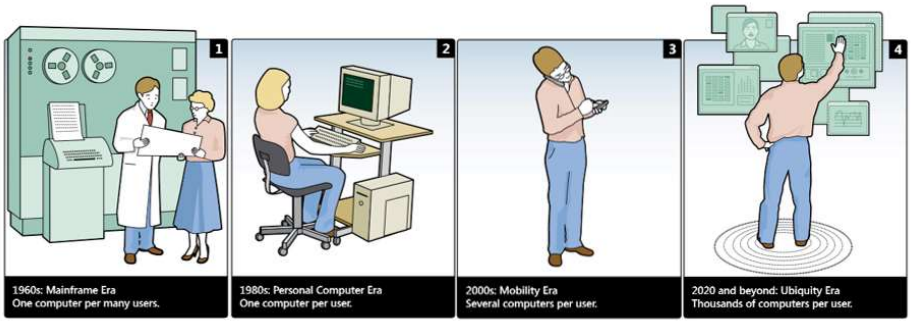


FIGURE 1 FOUR COMPUTING ERAS (Harper, Rodden, Rogers, & Sellen, 2008)

Nowadays, graphic hardware can render near- to photorealistic results at interactive speeds. This effectively means that the border between physical reality and virtual reality from a vision perspective is blurring. The display market is experimenting with stereo vision, similar to our own human vision.

Most cars and mobile phones are equipped with a global positioning system (GPS) and navigation software. This has triggered an explosion of services that make use of spatial information. Google Lens[1] and Pokémon GO[2] are just two of the best-known applications that have been created to serve this market. Increasing chip process power, decreasing the energy consumption and miniaturization are all ongoing developments: today's mobile phones have the capabilities of a personal computer of 10 years ago. Sensors are getting more powerful and are integrated in many professional and consumer products, such as; range sensors and charge-coupled devices (CCD) in cars, mobile phones, and consoles. Furthermore, even small mobile devices are able to capture high definition content at more than interactive speed. More smartphones are sold than personal computers (Canalys, 2011); they are equipped with multi-touch, voice control and accelerometers for optimal control. Furthermore, many capabilities are comparable to desktop computing, with some functionalities even surpassing that.

To conduct spatial tasks on location we need to look at a post-desktop model of human-computer interaction (Weiser, 1991). This is known as ubiquitous or pervasive computing; terms that are used when information processing has been thoroughly integrated into everyday objects and activities - obviously, a rather broad description. A more fitting term, according to this author, is physical computing (O'Sullivan & Igoe, 2004). In its broadest sense, physical computing means building interactive physical systems using software and hardware that can sense and respond to the analog world. In the visual domain, a system that can augment, diminish, or otherwise alter the visual perception of reality is called a "Reality Mediator" (Mann, 2003).

Logically, the paradigms for interaction with computers have changed with the mobility trend. It is hard to use a mouse when walking. Gesture, voice and sensor rich attachments are much better suited to support on-the-move tasks. Vision technologies can precisely track hands for freehand control, and voice recognition is the default in most car navigation systems.

If humans are to be supported on location, understanding the environment in question is essential. A Geographical Information Systems (GIS) database can provide information that facilitates navigation, but the database might be outdated, not detailed enough, initially wrong or incomplete. Spatial analyses of

[1]https://www.wired.com/2017/05/google-lens-turns-camera-search-box/, last visited June 2017

[2] http://www.pokemongo.com/ last visited June 2017

the environments for an up-to-date and detailed model of the surrounding constitute a rudimentary need for support on location. Real-time sensing provides the most recent information, which guarantees a higher level of autonomous freedom. Our environment can be stored as pre-knowledge (i.e. context aware information) or context free knowledge (i.e. sensory information).

1.2. GENESIS

At some point in the year 2000, the author of the present study watched a documentary[3] about Ivan Sutherland, which showed the first (1968) Augmented Reality head-mounted device, which opened his eyes to the significance of blending the virtual world and the real world. Since then, he has kept abreast of the domain and followed related research. After earning his M.Sc., he started working for an engineering company that used laser scanners to automatically map environments, instead of classical survey-based mapping methods. The demand for digitization of physical environments was growing significantly, both for high-tech engineering and for less complex use cases, such as serious gaming, virtual reality, architecture, etc. The insights gathered from mapping environments proved to be foundational to understanding how Sutherland's original ideas might be extended.

Humans are magnificent at imagination, but imagination is a hard thing to share. The virtual world is the closest to sharable imagination that we can currently come. That sharable characteristic and the anchoring of virtual content in the physical world provided the thrust that fueled this research. It is not hard to imagine a digital overlay of the physical world that provides capabilities - browsing spatial historical data on the spot, looking through walls and projecting loved ones from the other side of the world - that were inconceivable 50 years ago.

The author's network provided a viable opportunity for advancing Sutherland`s ideas in a specific domain. The Netherlands Forensic Institute (NFI) announced a project call for innovation in crime scene investigation. A co-authored proposal nicknamed "CSI The Hague" was written, which had the desired prerequisites to be used as a dominant case. The most important elements will be briefly discussed in this chapter and detailed in chapter 2.

In CSI The Hague, various technology companies and research institutes were granted the opportunity to experiment, adapt and validate their technology in

[3] https://www.youtube.com/watch?v=NtwZXGprxag, last visited June 2017

the Forensic Field Lab. The condition for participation was that their technology should have the potential to improve crime scene investigation. Furthermore, the companies and institutes could discuss and validate their technology with crime scene experts in close-to-real use cases. The proposed topic for this project was mediated reality in crime scenes: "A digital layer on a crime scene as a collaborative environment".

Immediately, a list of interesting challenges emerged. As every crime scene is unique, what tools allow us to work with digital overlays in continuously changing conditions? What kind of information sharing is necessary to gain shared situational awareness?

This case provided a great opportunity and had clear relevance:

- A crime scene is a unique *pristine* environment. Although 3D models of the environments may exist, they don't reflect reality.
- Many *spatial* related tasks take place at crime scenes, i.e., line of sight verifications, bullet trajectory analysis and blood pattern analysis.
- Preferably, research on crime scenes should be *contactless*; contamination of a crime scene needs to be avoided at all cost.
- Shared *understanding* of a crime scene is important due to the number of people and different types of expertise involved in crime scene investigation.
- Too *many* people at a crime scene will increase the chance of contamination.
- A crime scene *degrades* quickly over time; a body or artifact is removed, chemical degradation, disappearance or other changes.
- Experts in many associated domains are sparse; the chances of getting expert knowledge within a reasonable amount of *time* are slim.
- A vast body of people needs to obtain situational awareness quickly.

An important motivator for taking on crime scene investigation as a case for this research was an influential report written by Bernard Welten[4] (2004), which posited that the role and significance of forensic investigations will greatly increase in the near future. In his view, forensic investigations will no longer only rely on tactical processes but will be increasingly controlling and direction-giving in investigations. He envisioned a not-so-far-away future where smart technology will aid investigators. *"Technical evidence is worth more than the*

[4] Chief of Police, Amsterdam 2004-2011

statement of people. People make mistakes, suspects invoke their right to remain silent, but the technical evidence says a lot, if not all" (Welten, 2004). Following this report is a more specific report on imaging by Flight and Hulshof (2010), who identified the following three milestones to be achieved in the use of imaging in the security domain: the ability to follow objects and subjects, to reconstruct incidents and to add metadata. Of the greatest interest to this thesis in relation to augmented reality is the reconstruction of incidents: *"In 2015, parties in the security domain can reconstruct, based on images, events and incidents, so that useful information becomes available to the safety chain"* (Flight & Hulshof, 2010). Furthermore, working with external image data will become more relevant: *"Observers in the security field will have to meet more stringent requirements in terms of education, skills and competencies. Observers will increasingly have to work with image data of places that they themselves do not know. They also need to have the skills and powers of police officers, social workers, guards and security guards to control from behind the screen"* (Flight & Hulshof, 2010). These quotes illustrate the growing importance of trust in data and improved or diminished situational awareness.

Fortunately, the use case applies to many domains that face similar constraints, as summarized in chapter 1.1. This will aid the generalization of the technological and social impact to other domains, such as the medical, military and engineering worlds.

1.3. MEDIATED REALITY

As will be elaborated on in following chapters, there are many names for the technologies researched in this thesis. A less used term is "mediated reality", where mediation refers to the process in which reality is brought into alignment with what humans perceive as being real. The goal of this thesis is not to only augment a scene but to act as the intermediary for associated participants. So, what is mediated reality and what is necessary to allow for mediated reality? *"By way of explanation, 'virtual reality' creates a completely computer-generated environment, 'augmented reality' uses an existing, real-life environment, and adds computer-generated information (virtual objects) thereto, 'diminished reality' filters the environment (i.e., it alters real objects, replaces them with virtual ones, or renders them imperceptible), and mediated reality combines augmented and diminished reality…..allowing individuals to communicate with one another by altering each other's perception of reality"* (Mann, 2003). In this thesis, collaborative augmented reality would be a correct terminology too. However, mediated reality also covers other scenarios, such as replacing regular cameras with infrared or X-ray modalities, thus rendering the real world imperceptible.

This broader term covers more but does not encompass all possible aspects of mediated reality. Mann's quote highlights a few prerequisites that need to be in place to allow for mediated reality. A 'map' of the environment is needed to create augmentations, a 'virtual reality' layer is necessary to overlay the intent and the principles governing 'augmented reality' must be in place. The father of augmented reality, Ivan Sutherland, explains augmented reality as follows: *"A display connected to a digital computer gives us a chance to gain familiarity with concepts not realizable in the physical world. It is a looking glass into a mathematical wonderland"* (1965). In (1968) he developed the first head mounted display that was capable of merging the virtual and the real world. Krevelen & Poelman (2010) have compiled a survey of the history of augmented reality. Many research domains contributed to making augmentation possible, which allows this thesis to be multi-disciplinary. In order of appearance, the related domains are briefly discussed to provide relevant background on the domains.

Mediated reality needs cornerstone technologies to exist:

- A virtual environment where sensed data, user input and library data comes together and can be shared.
- An outlet for the composited information where effective digital information overlays the physical.
- A "computer" understanding of the environment, for overlay, interaction and analysis. a.k.a. a 3D map.
- An interaction paradigm for interacting with the presented information and which allows for collaboration.

Overlay digital content onto the real world

Mixed reality refers to the merging of real and virtual worlds to produce new environments and visualizations where physical and digital objects co-exist and interact in real-time (P. Milgram & Colquhoun, 1999). Milgram and Kishino (1994) defined a mixed reality as: *"...anywhere between the extrema of the virtuality continuum."* To be able to allow for mediated reality, we need to know what this spectrum looks like. Figure 2 shows that the Virtuality Continuum extends from the completely real to the completely virtual environment, with augmented reality and augmented virtuality in between the two. Spatial information and virtual reality enable and improve mixed reality forms.

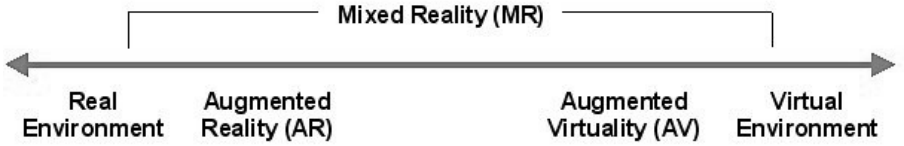


FIGURE 2 REALITY-VIRTUALITY CONTINUUM, ADAPTED FROM (P. MILGRAM & COLQUHOUN, 1999)

The continuum ranges from the purely virtual, without any restrictions on transportation, to reality with its concomitant restrictions. Virtual reality is useful for many applications, but there are still limits to what we can achieve with it, such as physical fatigue, personal contact and full sensory usage. While much research used to be directed at virtual reality, lately it is the mixed forms that have been receiving more attention because of the advancements in sensors, including the micro electro mechanical systems (MEMS) being used in smartphones and game controllers.

Mediated reality was defined by Mann (2003) and as he explained, the constraints and challenges of augmented reality apply. The following augmented reality laws were composed by Azuma (Azuma 1997):

- Combines real and virtual objects in a real environment;
- Registers (aligns) real and virtual objects with each other; and
- Runs interactively, in three dimensions, and in real time.

A virtual object is a computer generated real or imaginary object (P. Milgram & Kishino, 1994). The ability of augmented reality to present information superimposed on our view of the world opens many interesting opportunities for graphical interaction with our direct environment. Up until now research had mainly been focused on the technology that enabled mixed reality (Bimber & Raskar, 2005a), but as explained in earlier in this chapter, advances have been made that allow technological barriers to be breached. The rules formulated by Azuma (1997) demand much from both software and hardware: a representation of reality is needed, as is an alignment with reality so that the artifacts exist in the same space, and everything needs to run at interactive speeds.

Some industries are already making use of advanced mixed reality forms. The best and previously mentioned example is laparoscopic surgery (Fuchs et al., 1998). The operating devices used by the surgeon are spatially tracked while he operates without physically seeing the operating space directly; instead, he is

provided with updates from micro cameras. The rest of the information he relies on derives from a magnetic resonance imaging (MRI) scan that has been previously recorded and overlaid on the camera information, which in turn is overlaid with the tracked operational devices. A much simpler example of mixed realty is that of the navigation system in cars: because of the GPS information, the car knows where it is, plans its route and displays that to its driver.

Augmented reality is mostly used on mobile phones and tablets (Daley, 2015). These devices effectively use only ~2% of our visible area. The use of head worn displays that provide a considerably larger viewing area is still a niche market (Daley, 2015). However, big commercial entities, like; Sony, Facebook, Microsoft and Google are (again) starting to experiment with head worn displays. The strong economic incentive of the smartphone market is pushing for new ways of consumption and the number of augmented reality related applications are increasing (Daley, 2015).

3D Mapping

Creating a map of a pristine environment is a critical aspect to the use case; a crime scene is by default something not encountered before. To be able to augment a scene, the map must be high fidelity and spatial. Many disciplines require three dimensional maps; the technologies for generating these are still improving.

Spatial Information describes the physical location and dimension of objects, and the relationship between objects. The spatial information domain is a sub-set of the broader information technology domain and is closely related to metrology, geographical information systems and geometrics. Many tools are being developed to measure space, such as measurement tape, theodolites, photogrammetry and laser scanning, to name but a few (Kavanagh, 2008). With these tools, virtual representations of real-world objects are created and software tools can be used to manipulate data in design or analysis processes.

For creating real-world 3D maps, two foundational technologies are used: active sensing and passive sensing (Beraldin, Blais, Cournoyer, Godin, & Rioux, 2000). An active sensor has its own energy source to reach the scene with; a laser, pattern or other projection. Passive sensors wait for the environment to emit data that can be captured. Examples of active sensors are laser scanners, Microsoft's Kinect and white light scanners. Examples of passive sensors are video cameras, DSLR's and infrared cameras. Both technologies need software processes that merge data into a coherent model (Hartley & Mundy, 1993).

With today's hardware and software, it is possible to sense our environment at interactive speeds and to enable the user to interact with digitized versions of reality. There is a known pipeline for 3D modeling available that shows potential for the automation of the processes as described by the author (Fumarola & Poelman, 2011).

Virtual reality

Apart from having a digital spatial 3D description of a pristine environment, the information needs to be visualized. It is important to know what is mapped and what is virtual. Virtual reality is a term that applies to computer-simulated environments that can simulate places in the real world as well as in imaginary worlds (P. Milgram & Colquhoun, 1999). Currently, virtual reality environments are primarily visual experiences, displayed either on a computer screen, a projective display or on wearable displays, i.e., a mobile phone or head mounted display. Some simulations include additional sensory information, such as sound through speakers or headphones and even haptic feedback (Poelman & Fumarola, 2009). Some industries, such as the movie industry can create photorealistic renderings that are characterized by labor intensive non-real-time procedures. However, over the past few years, physically based rendering and 3D data structures have become highly optimized for the real-time interaction needed to create high-fidelity believable games (Stricker, Vigueras-Gomez, Gibson, & Ledda, 2004). This move from offline to online high fidelity rendering in response to the demands of the game industry is leapfrogging the virtual reality domain. As discussed in the previous paragraph, the tools for capturing real world environments are commoditizing rapidly, which is putting a strain on visualization (Meager, 1982). Fortunately, software and hardware surfacing has been developed that counterbalances this and allows real-time interaction. For raw 3D detail rendering, deferred rendering, hardware tessellation of polygon models and sparse voxel-octrees offer suitable solutions (Laine & Karras, 2010).

3D Game engines are responsible for quite a few of the advancements in virtual reality; they consist of multiple modules, including a rendering, sound, physics and artificial intelligence module (Poelman & Fumarola, 2009). Most high-end game engines are capable of rendering to different types of displays, have an authoring environment and are capable of handling vast amounts of 3D data. There are other virtual reality related domains, such as CAD, building information modeling (BIM) and GIS. To visualize reality in high fidelity, the scene detail must be high and three dimensional. This means a high bandwidth for data and a lot of processing power. Looking at the state of the art of graphics

chips and taking Moor's law into account, it is obvious that this is becoming less of an issue. For example, NVidia's Tegra's, Qualcomm's Snapdragon and Texas Instrument's PowerVR chips are built for mobile devices, with low power consumption, multiple graphic processing units; they can display multiple high resolutions and are therefore able to run high-end games. Virtual reality is becoming mobile, widespread and affordable (Daley, 2015).

Human Computer Interaction

The widespread adoption of electronic devices in all shapes and forms has encouraged the development of alternatives to the keyboard and mouse, the classic Windows, Icons, Menus, and Pointer (WIMP) (Daley, 2015). These include one-handed keyboards, digitizing tablets, movement tracking devices, voice recognition and glove-based devices. The domain that researches this is called human computer interaction. It proceeds on the assumption that, as the attention of a user has to be directed at the task at hand, a user interface should support this in the best possible way (Weiser, 1991). Mark Weiser's (1991) main concern was that computer interfaces were too demanding of human attention; *"Unlike good tools that become an extension of ourselves, computers often do not allow us to focus on the task at hand but rather divert us into figuring out how to get the tool to work properly".*

Smartphones and gaming consoles have introduced a new breed of interface that does not require a steep learning curve. Multi-touch is easing mobile interface usage and the game consoles are integrating body movement as an input device. There are two main directions; sensing movement with Microsoft Kinect, Intel Realsense and the WII control with its accurate device movement detection. Furthermore, Connexion's space mouse and Leap Motion's vision based tracker is improving desktop interfacing.

Natural 3D interaction is still a challenge for the HCI community; mixed reality interfacing is still under development and mostly resides in research departments. Previous research conducted by the author to validate the effectiveness and ease of use of 2D, 2,5D and 3D displays in spatial tasks showed that, while 3D was ranked highest in potential, it was still considered to be immature (Poelman, Rusak, Verbraeck, & Alcubilla, 2010).

Wearable Computing

Wearable computing did not disappear with the Walkman; the device was soon followed by portable DVD players, video players and mobile phones. There are many examples of mobile computing devices, such as GPS watches, Universal Serial Bus (USB) necklaces, Google glass and more serious applications like health monitoring devices and guiding devices for the impaired, to name but a few.

Mobile technology is characterized by a low power consumption and small components, high bandwidth wireless communication and durable batteries. At the heart of power consumption is the phenomenon known as "Die shrink", referring to the ongoing shrinking of silicon geometries. Die shrink is beneficial, as shrinking a die reduces the current leakage in semiconductor devices while maintaining the same clock frequency of a chip, which produces a product with less power consumption, increased clock rate headroom, and lower prices (Kosonocky & Collins, 2013). The displays are also getting smaller, as the resolution rises, and they increasingly require less power, with organic light emitting diodes (OLED) emerging as the preferred technology: OLED screens require only a fraction of the power conventional screens do, due to the absence of a lighting source (Kamtekar, Monkman, & Bryce, 2010). Most users have access to the current 4th and 5th generations (4G LTE, 5G) of the mobile phone network. Other technologies, e.g. WiMax, are waiting at our doorsteps. Long Term Evolution (LTE) promises peak download rates of 1 Gbit/s and up 500 Mbit/s, which should make it possible to stream high definition content with ease (Woyke, 2011). Looking at battery life, new developments are also on their way. EEStor, claims their solution can get 280 watt hours per kilogram compared with 120 for lithium-ion battery (Dean, 2004). Stanford researchers (Liu et al., 2014) use nanowires to remake lithium-ion batteries. This new technology has the capability of lasting eight times longer than current batteries (Venman, 2015).

Leveraging these innovations, the smart phone is the accelerator for a great many advances in mobile technology: a new smartphone is introduced with new features every few months. However, smartphones and tablet have limitations too. They are not capable of powerful text editing or advanced 3d modeling and are less creation platforms, than they are platforms for information consumption (Harrison, Flood, & Duce, 2013).

Wearable head mounted displays, considered to be classical virtual reality equipment, have improved considerably as well. Two examples are Facebook's Oculus rift and Sony HMZ. An important reason to focus on large field-of-view

devices in wearable computing is provided by Tor Norretranders (1999) in his book 'The user illusion'. According to Norretranders, sight is the dominant human perceptual channel; the majority of the information that we process is vision related.

Collaboration

Some types of work require people to be onsite, because the real world is a basis for their analysis or design. Spatial challenges are mostly solved by teams of people, especially because they involve multiple domains of expertise (Dong, Behzadan, Chen, & Kamat, 2013).

Experts are rarer than non-experts and allowing them to collaborate effectively in co-located situations increases their reach. According to Dong et al. (2013), for effective collaboration, especially co-located, shared situational awareness is essential. The focus of this dissertation will be predominantly on one-to-one collaboration.

Over the last few decades, many augmented reality systems have been developed that focused on collaboration and shared situational awareness (Arayici & Aouad, 2004; Broll et al., 2004; Kiyokawa, Billinghurst, Campbell, & Woods, 2003; Szalavari, Schmalstieg, Fuhrmann, & Gervautz, 1998). What the majority of these systems had in common were the favorable effects of having virtual objects grounded in reality, which lessens the cognitive load and provides a synchronized workspace that avoids misunderstanding in communication caused by the distortion of time or viewpoint (Bujak et al., 2013).

One of the distinct advantages of augmented reality is that it can enable the communication and discussion of a validation analysis using a collaborative environment, where the field experts can quickly appreciate the visual analysis displayed and are able to interactively participate in a discussion that helps to understand, validate and improve the analysis processes (Dong et al., 2013).

Endsley (1995) has an encompassing theory on situation awareness. He distinguishes three levels; (1) the perception of elements in the environment, (2) comprehension of the current situation, and (3) projection of future status. Augmented reality systems adhere to those phases by having the information jointly available.

When designing a system (Figure 3) that facilitates situational awareness, Endsley stresses that interface knowledge (e3), which leans on the systems

knowledge (e2) should not misalign with perceived (e4) human sensed real world conditions. In that light, augmentation in mediated reality offers a satisfactory method.

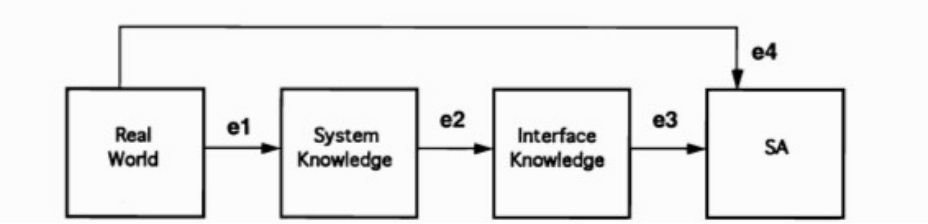


FIGURE 3 SITUATION AWARENESS INPUTS (ENDSLEY, 1995)

It is intriguing to investigate whether Endsley's situational awareness theory and augmented reality can improve collaboration, especially between co-located people.

1.4. RESEARCH QUESTION

In chapter 1.1, the challenges that arise in everyday life are discussed. These challenges require new ways of dealing with complex shared tasks. The relatively unexplored domain of mediated reality was identified as an opportunity to solve complex spatial related tasks. A use case was described in chapter 1.2. which added detail to the challenges, without being too specific. In chapter 1.3 the domains related to mediated reality were explored and the opportunities explained. The challenges concur with the findings of various celebrated researchers (Azuma, 1997; P. Milgram & Colquhoun, 1999; I. Sutherland, 1968), who lacked the technological means that are currently available to fulfill the promise of mediated reality.

RESEARCH QUESTION

> ***How can we support collaborative spatial interaction in a pristine environment applying mediated reality?***

Apart from the "how", the research question also implies utility; will the product fulfil the needs? From this, various sub research questions also ensued. First, we must build an artifact that allows us to support collaborative spatial interaction in mediated reality for a pristine environment, which gave rise to the sub questions:

A) What architecture allows for collaborative spatial interaction in mediated reality?

The goal of this sub-question is to validate the functionality requirements, i.e. how the problem is technically solved. Next, questions related to the remote and on-premise interaction must be answered. To collaborate, individuals should be able to work with the system, leading to the following sub-questions.

B) Does the architecture support an on-premise user with meaningful interaction of the digital overlay?
C) Does the architecture support the interaction of a remote user with the digital pristine environment?
D) Does the architecture support spatial collaboration between an on-premise and a remote user?

The sub-questions facilitate answering the main research question in a stepped approach.

1.5. Research Approach

Research philosophies, approaches and strategies are used to create knowledge and to construct this in a rigorous and meaningful way to answer a research question. There are different research beliefs and philosophies, among which a student must find the way. The decisions regarding these philosophies are fundamental and will determine the way knowledge is constructed.

The research questions already allude to the need to look at both the human and empirical aspects. As stated in chapter 1.2., experts in the field will contribute their expertise to validate the desired system. Discrete answers are highly unlikely once humans with all their complexity are in the loop. First, the philosophy of this research will be discussed before continuing with the strategy and instruments.

1.5.1. Research Philosophy

An artefact for a socio-technical system that resides in a multi-actor environment cannot rely on just one research approach. An artifact can be empirically validated, but an interface for an artifact is subjective and therefore biased, requiring a different approach. According to Dobsen (2002) a socio-technical artifact *"cannot be understood independently of the social actors involved in the knowledge derivation process".* Our methodology follows the approach taken by Hevner (2004), which is otherwise known as design science in information systems. We elaborate on this in the next chapter.

We have adopted critical realism as our ontological stance. Critical realism holds that for scientific investigation to take place, the object of that investigation must have real, adaptable, internal mechanisms which can be actualized to produce outcomes. The science should be understood as an ongoing process in which scientists improve the concepts they use to understand the mechanisms that they investigate. At the same time the resulting production of knowledge is viewed as a human, socially and historically conditioned activity. To understand the intervention of the artifact on the environment and actors, an interpretivist approach is the epistemological choice that fits this thesis best. To that end, we have worked with a small number of experts to interpret and justify the knowledge produced.

1.5.2. Research Strategy

Fundamentally, the research philosophy shapes the research strategy. March and Smith (1995) clarified how design science can be applied in the field of information technology. This clarification and the personal belief of the author

overlap. Essentially, design science considers research as "devising artifacts to attain goals". While justification and discovery are part of design science, unlike in traditional science, these are not the fundament: in design science, building and evaluating artifacts are the essence.

According to Hevner et al. (2004), design science and natural science are complementary. While design science relies on the use of existing theories and is pro-active with technology, natural science builds theory and takes the use of technology for granted. While design science uses relevant theory from natural science, it advocates a research cycle in which artifacts targeted at solving information problems are built. As explained by Hevner, the methodologies of design science and natural science can be used in conjunction, whereby the research cycle from design science is augmented with natural science to engage and anticipate the created artifact that fits the research questions.

Design science distinguishes three types of research contributions; artifact design, foundations and methodologies (Hevner et al., 2004). This research focuses on the design of an artifact that needs to be validated as a socio-technical system. Although currently regarded as unorthodox, the activity is iterative and incremental. Design science provides us with requirements upon which the evaluation of an artifact is based. According to Hevner et al. (2004), the use of a prototype is necessary. The evaluation will be based on the integration of the artifact within the current workflow processes of the domain as described in chapter 1.2. The foundational information and methodologies can be collected and should inform the artifact design.

1.5.3. INSTRUMENTS

The design evaluation methods described by Hevner et al. (2004) are used in this research. In this thesis case, using only a single method would not be sufficient, as the participants are observed, interviewed, logged; experiments are run to validate sub- systems and the entire artifact. Some of the experiments are analytical, some experimental, and in accordance with Hevner, testing and descriptive methods are used.

The information systems research framework provides instruments that must be adapted to the designed artefact and selected evaluation. The case study provided in chapter 1.2 is used to research the artefact in the domain environment and will be used for evaluation. The research and sub- research questions indicate that the systems need individual validation: in some cases, metric and in others, humanistic. Multiple tools were used to ensure the rigor of

the research and to cast a wide net; these comprised questionnaires, after action reviews, system performance validation and expert interviews. Because of the expected iterations, a spiral modal of the design cycle could be adopted for the information systems research framework.

Of course, the literature research that will be used to build the knowledge base is complementary to the evaluation. By using Hevner's design science tools, both a positivist and interpretivist perspective can be used to challenge the artifact.

1.6. RELEVANCE

Nowadays, companies can make use of co-location, or work in geographically dispersed or in virtual teams. The tools that are currently used reflect the traditional ways of working, but with a digital finish. Most of the world is covered by communication networks; there is an increasing body of knowledge accessible by anyone having access to information networks. Yet still, people are stuck in traffic jams, have longer and longer commutes and attempt to communicate increasing complex information through means that were not designed for this. The tool discussed in this thesis can provide solace in multiple areas that can potentially impact positively on society.

Travel: it limits the number of visits to physical locations. By having access to digital replicas of virtual environments and having means of communications hat allow collaboration in the virtual space that augments environments, fewer people will be required to be on-location, leading to less travel.

Hazardous environments: fewer people are required to work in hazardous or contaminated environments, while co-located experts can digitally aid those on-site. Analyses can run in real-time based on the acquired 3D map data, which allows experts to respond to early warning signs.

Complex operations: some human computer interfaces or mechanical environments are so complex that digital aid is required. Operating a factory or a plane can be daunting; the ability to ask for help sharing real-time visual information can save the day.

Collaborate: more effectively- not only sharing visuals but real-time sensed data allows for richer collaboration; the context adds to the collaboration.

1.7. RESEARCH OUTLINE

This thesis is organized according to the design science information systems approach discussed in chapter 1.5 and illustrated in Figure 2. The following chapter elaborates on the dominant use case and the requirements that were abstracted from the case. An architectural chapter then follows, to highlight the multidisciplinary domain characteristics of this thesis. The architectural guidance is used to evolve the literature research into relevant background knowledge. Next, the iterative design methodology is described and the individual subsystems validated. In the penultimate chapter, the entire system is validated with the dominant use case. The final chapter generalizes conclusions from the use case, reflects on the research approach and puts forward suggestions for future research.

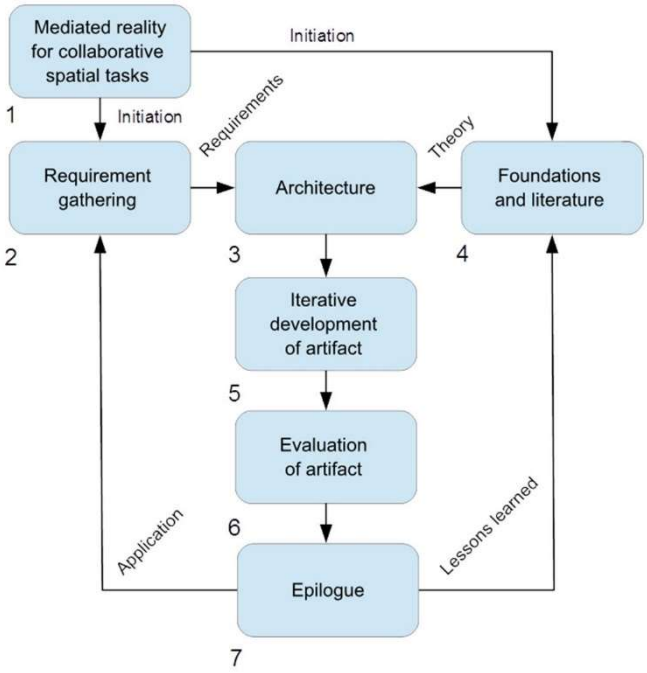


FIGURE 4 RESEARCH OUTLINE

2. DOMAIN RELATED REQUIREMENTS

In section 1.2, the case study of collaborative mediated reality in crime scene investigation was briefly introduced. To establish the relevance of this for the domain of crime scene investigation, three methods are used; (1) literature research on 3D crime scene reconstruction for crime scene investigation, (2) workflow analysis and (3) interviews with experts. The result of this chapter is a list of requirements based on the obtained domain knowledge. First, the field of 3D scene reconstruction in crime scene investigation is explored.

2.1. 3D RECONSTRUCTION

Three dimensional (3D) acquisition and virtual crime scene construction are a sub category of Computational Forensics, and an emerging interdisciplinary research domain (Franke & Srihari, 2008). This type of research is understood as the hypothesis-driven investigation of a specific forensic problem in which virtual construction is one of the available tools. The primary goal is the discovery and the advancement of forensic knowledge (Franke & Srihari, 2007). Acquisition and virtual crime scene construction (in short 3D reconstruction) involves active and passive sensors, modelling, simulation, analysis and recognition in studying and solving forensic problems. By overcoming the limitations of human cognitive and physical abilities, the crime scene investigator can better detect and analyze evidence. The real-time sensors and the 3D reconstruction of the crime location itself can reveal and improve traces of evidence in a reproducible and objective way.

Classically, the most commonly used technologies for spatial 3D in crime scene investigation are photogrammetry and tachymetry (Flight & Hulshof, 2010). In photogrammetry, a different perspective between two or more images is used to acquire 3D coordinates, a technique generally called structure-from-motion (Dellaert, Seitz, Thorpe, & Thrun, 2000). Tachymetry involves the use of either a GPS or the positions of known objects in space to retrace the location of the equipment, which is then used to obtain new 3D coordinate measurements. More recent (Flight & Hulshof, 2010) is the use of laser scanning to obtain 3D measurements, in which the device rotates around and acquires a panoramic image with depth information. Creating a digital "copy" of the crime scene serves multiple goals: it freezes a crime scene in time, it allows for 3D interaction and analysis and it provides a communication means for the investigators (Buck et al., 2011).

3D reconstruction and analysis are regularly used by hundreds of crime scene investigators worldwide (Fries, 2006; Jenkins, 2005). The use of sensors and computational support is not meant to replace the investigator. They are tools that are intended to assist in basic and applied investigation and to support investigators in their quest for truth. The equipment used by forensic investigators for 3D imaging was initially designed for survey related domains. It proved its use in forensics some years ago, as described at the International Association of Forensic and Security Metrology (IAFSM[5]) conference. Over the past few years, the equipment has become more sophisticated and better suited for forensics[5] . By being able to capture the crime scene in 3D, new opportunities for the investigator have emerged. Many successful 3D reconstructions have been conducted by forensic labs. When boats, cars, trains or airplanes crash, the 3D data obtained by the capturing process is used to understand and simulate the incident. Case studies have been described by a number of researchers (Fries, 2006; Jenkins, 2005). Another example is its use in witness verifications, where a 3D virtual model of the environment is created to validate scenarios: a suspect's height is gathered from imagery created by security cameras, scenarios are checked in different orders in virtual reality, blood patterns are analyzed to reconstruct impact locations (Figure 5), a ballistic trajectory is evaluated for scenario testing, form reconstruction is used to show what something looked like before it broke, etc. All these examples have in common the fact that they make use of metric, spatial oriented 3D data to support or contradict scenarios in the investigation (Buck et al., 2011).

[5] http://www.iafsm.org/, last visited July 2017

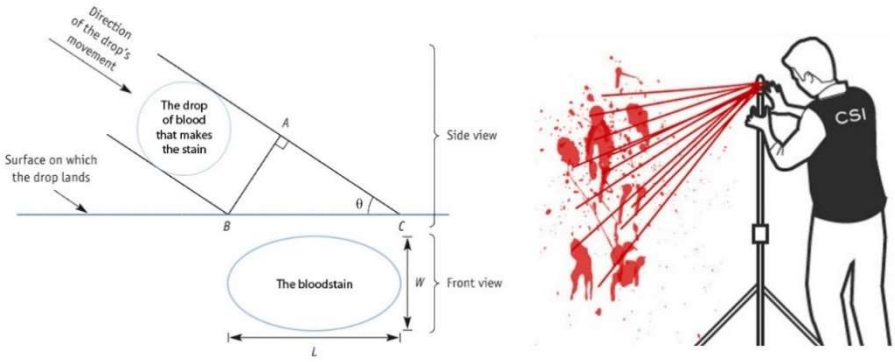


FIGURE 5 BLOOD PATTERN ANALYSIS, THEORY (LEFT), PHYSICAL RECONSTRUCTION (RIGHT), COURTESY OF JACKSON & JACKSON (JACKSON & JACKSON, 2004)

A model of the real environment with all its nuances is required to accurately reconstruct a crime scene; hence the use of 3D measurement technologies. To illustrate 3D reconstruction in crime scene investigation, we use the example of blood pattern analysis, as illustrated in Figure 5. The goal of 3D blood pattern analysis is to estimate the origin of impact, which can be derived from the blood stains. The result of 3D blood pattern analysis is used to find the impact location; whether a victim was standing, sitting or lying down can be crucial information in court to verify witness testimonies. When blood impacts a surface, the stain explains something about the direction before impact. With respect to the left image in Figure 5, sinα = Width/Length, which results in a collimation as shown in the right image (Jackson & Jackson, 2004). 3D blood pattern analyses are a sub-discipline of blood pattern analysis, which is much broader and includes interpretation of other than oval stains, age of stains and selection of the best samples. *Requirement [01:] the system must be able to acquire and store, spatial oriented metric 3D data from a pristine environment.*

2.2. WORKFLOWS

A crime scene investigation process follows strict phases (Figure 6). In this section, the phases are summarized and reflected on in relation to crime scene analysis and 3D reconstruction. Depending on the crime scene size and scope there can be inner loops that iterate between planning and execution. The European Network of Forensic Science Institutes (ENFSI) uses different wording but the essence of the process is the same; Discovery, Collection, Enhancement, Comparison, Interpretation (Jackson & Jackson, 2004).

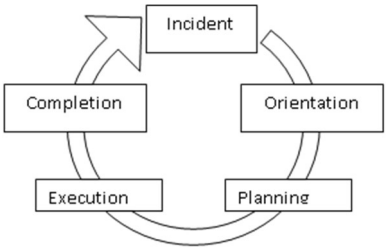


FIGURE 6 CRIME SCENE INVESTIGATION PHASES ACCORDING TO THE NFI (2011)

An elaborate overview of the best practices for crime scene investigators can be found in the guide compiled by Kevin Lothridge and Frank Fitzpatrick (Lothridge & Fitzpatrick, 2013). A summary is provided in the following paragraphs. The notification of the incident can have a wide variety of origins, from direct contact with citizens up to an emergency call. The orientation phase is meant to scope the extent of the incident and will generally involve two officials. This phase does not involve sophisticated means; only a mobile phone, notebook and, in most cases, a digital camera. Orientation is crucial in providing directions for further research and is also the first phase with contamination danger of the crime scene. The main goal of this phase is to scope the extent of the incident: should the investigation be scaled up, are more people required, is it safe, etc. In many cases, the first investigator on the scene must testify in court about what he witnessed. That first look at the crime scene is therefore critical and preferably no self-imposed obstructions should hinder the investigator. The technology cannot get in the way in this crucial phase, where scoping and safety are the primary concerns. *Requirement [02]: the system should allow an investigator unhindered view of the crime scene.*

There are three investigation means at the disposal of the crime scene investigators at a pristine scene: 1) registration of the crime scene, 2) indicative resources and 3) securing traces of evidence. Just using registration tools will provide approximately the same results as our eyes can perceive; to enhance our senses we can use indicative tools, such as alternative light sources. Furthermore, there are physical traces that can be gathered and secured, such as biological traces, fibers, entomological proof, deformation, and micro traces. However, these are usually destructive to the incident scene and therefore fit better in the execution phase.

The official(s) who arrived first on the crime scene and who were responsible for gathering the orientation material must brief the team that is installed by their superior. Based on the briefing with the gathered experts, a plan is created that

will involve a more detailed investigation of the incident scene. The necessary specialist expertise is contacted and planned. Then the execution phase starts, during which detailed research takes place. In case of a severe crime scene for which a 3D reconstruction is needed, the scene is often digitized, either by detailed photography, filming, panoramic scanning, or laser scanning.

The type of acquisition is directed by the type of analysis necessary, including bullet trajectories, blood pattern analysis or line of sight analysis (Franke & Srihari, 2007). The team that facilitates in the 3D acquisition of the crime scene is not automatically the same team that performs the virtual construction and analyses. Such analysis needs specialist training. However, the expert's analysis generally guides the surveyors on-site by specifying detail, areas and suitable technology. The results are used to build virtual representations and provide the input for simulations or analysis. During the process, new information might become available that requires the team to go back to the planning phase, re-evaluate the current research and conduct new data acquisition. The results of the work are documented in the last phase of the investigation and are used in court. In court, the officials who virtually built and captured the crime scene must also be able to provide additional information and insight on the quality of the investigation. They will be questioned as witnesses on what they saw and did.

Figure 6 and the associated paragraphs show that the process is sequential in nature and that the 3D reconstruction takes place during the execution phase. There are multiple people and domains involved in crime scene investigation; examples are a team leader, a prosecutor, coordinators, technical police, tactic police, etc. (Lothridge & Fitzpatrick, 2013). In the execution phase, crime scene investigators divide tasks and have regular meetings about investigation results. During the investigation, information is shared orally using mobile phones or radiotelephones, whereas the pictures, film and 3D models are shared mostly during the scheduled meetings in the planning and execution phases. Currently, collaboration is mostly synchronous (one-to-one) with regards to the 3D reconstruction, and the surveyors and experts are at the same location.

3D reconstruction and spatial analysis in crime scene investigation is currently considered to be predominantly available in the execution phase of research (Fries, 2006; Jenkins, 2005). It is used as an offline tool that takes a considerable amount of time with few experts available to conduct the analysis. The obtained information is shared during meetings and the expertise that is needed to use the data is scarcely available (Franke & Srihari, 2007). The stakeholders involved in the legal chain have a different background, which makes explaining

investigation results difficult without a common language that is shared between them. Not being able to directly see on-going investigations in which reconstructions are required slow down the process, among others, because of the considerable time needed for 3D reconstructions. The fidelity of 3D acquired data provides a lot of context to discussions, and having this data available can help shared situational understanding (Vosinakis, Koutsabasis, Stavrakis, Viorres, & Darzentas, 2008). *Requirement [03]: allow the system to share information with the investigation team during or shortly after acquisition.*

A contradiction surfaces when considering the current use of spatial analysis. Few tools are available during the orientation phase which provides the direction for the upcoming phases and is therefore critical for crime scene investigation. A crime scene changes over time and the means to freeze the scene are used in later phases of the process (Lothridge & Fitzpatrick, 2013). This information should preferably be obtained as early as possible, before contamination of the scene occurs. It would therefore seem logical to place scene capture in an early phase of the process, as stated in requirement [3].

Furthermore, there is a mismatch between the intended users and the complexity of current equipment and associated software. The hard- and software is not specifically designed for crime scene investigation and involves switching between a variety of software suites, and sometimes hardware platforms, e.g., Leica Geosystems Cyclone[6], Geomagic Studio[7], Autodesk Maya[8]. The software suites require extensive training and maintenance of skills, and this type of training is not part of the default training curriculum (Boel et al. (2009). The 3D reconstruction phases are explained in section 2.2.1. The work of a CSI is practical, which is reflected in education, hands-on training and real world practice (Buck et al., 2011). The majority of investigators are target users for 3D reconstruction methods, from new recruits through seasoned investigators, most will already be trained in the physical (analogue) investigation methods (Buck et al., 2011). The current comparable means used are photo and video solutions; using analogue photography methods, the investigators are up and running with all the equipment within less than 30 minutes (Weiss, 2008). *Requirement [04]: the time between setting up the system and the start of using the system for geometry capture should be less than 30 minutes.*

[6] http://hds.leica-geosystems.com/en/, last revisited 12/03/2017
[7] http://www.geomagic.com/en/, last revisited 12/03/2017
[8] http://usa.autodesk.com/maya/, last revisited 12/03/2017

2.2.1. 3D RECONSTRUCTION

The 3D acquisition and modelling pipeline has clear distinguishable steps in both image-based and range-based approaches, but there is no significant difference for use with crime scene investigation. A typical reconstruction example of a reconstruction is illustrated in Figure 7. Triangle-based interpolation was used to fill the data gaps.

The literature indicates that the preferred method for 3D reconstruction is laser scanning (Jenkins, 2005). Laser scanners require less processing and are less reliant on the operator's knowledge level. As can be seen in Table 1, a comparison of steps to be followed in each case shows that less processing is required for laser scanning, because the 3D data does not have to be computed.

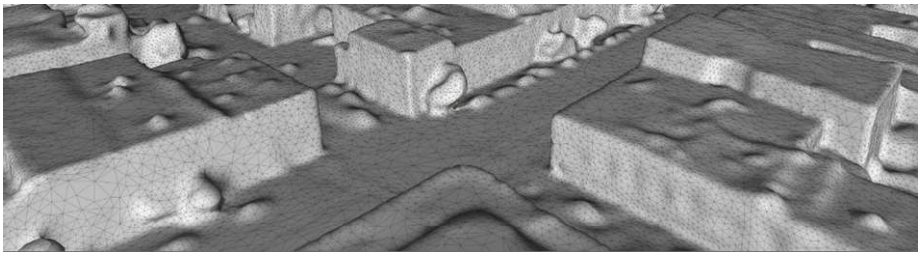


FIGURE 7 PHOTOGRAMMETRIC ACQUISITION WITH INTERPOLATION.

Step one is the acquisition of the data. The acquisition device must be placed at strategic positions in the scene and create 360-degree range image data, normal panoramas or regular images. This is similar to the tripod setup for a professional camera man. Specific to CSI is the way the equipment is treated: it has to be sterilized and the footprint has to be minimal (Weiss, 2008). The exposed parts, such as the tripod feet, have socks and the user must wear protective clothing. The skill of the investigator is very important for deciding on the positions with the best coverage (Jenkins, 2005).

	Photogrammetric	3D Laser
1	Image acquisition	Scan acquisition
2	Feature extraction (automated)	
3	Image calibration & orientation	Registration/Alignment
3a	Dense point reconstruction	(Directly available)
4	Surface generation	Surface generation
5	Texturing	Texturing
6	Final 3D model	Final 3D model

TABLE 1 SIMPLIFIED 3D MODELING PROCESS, BASED ON (REMONDINO & EL-HAKIM, 2006)

The second step is the organization of the acquired data. The data has no knowledge of its contents. It must be organized to facilitate the subsequent step, i.e., alignment. The questions to answer in this phase are: what pre-knowledge can we use (survey network, targets), how is the data oriented in relation to each other (clusters) and what data is correct? (Kavanagh, 2008). Sometimes data must be added to make it recognizable. In CSI, a sketch is made representing the scanner positions that can be used to search for overlap (Lothridge & Fitzpatrick, 2013).

The third step is the alignment of the data. Without alignment, the data is a bag of puzzle pieces: they need to be connected in the right way to create a coherent 3D model. A surveyed network, targets, inertia data and overlap can all be leveraged to calculate the complete 3D model (Fumarola & Poelman, 2011). Crime scenes are usually confined enough to allow overlap between the scans, which is generally used to build a model (Jenkins, 2005). This is the preferred approach, as the crime scene itself cannot be touched or disturbed. Using physical targets for surveying contaminates the scene, and a crime scene should be left in a pristine state during surveying. *Requirement [05:] the system must be able to align acquisition data without disturbing the pristine characteristics of the scene.*

Surface generation is the fourth step. Line of sight and a limited number of acquisitions do not cover the full scene; the data rarely has 100% coverage. Re-modelling is used to make the data available to other software (Fumarola & Poelman, 2011). If a surface is approximately flat, it is preferable to represent this as a plane instead of as thousands of triangles. For CSI, the scans are modelled with primitives that approximate the scans to make the model as lightweight as possible while maintaining quality (Buck et al., 2011). The advantage is that this improves the distribution of the data, facilitating its use in other software. The disadvantage is that some of the data is interpreted, which might change the results.

During the fifth step, the model is textured, so that non-experts can understand it. Figure 8 demonstrates what laser data looks like when not textured: most users do not understand what they are looking at.

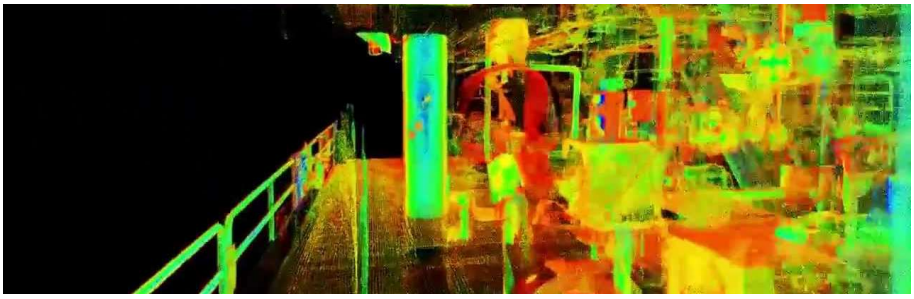


FIGURE 8 ALIGNED REFLECTANCE BASED LASER DATA.

Texturing allows photographic data to be projected on top of the laser data or surface data. Figure 8 clearly shows that intensity point clouds are hard to read because of non-real world coloring and because of the see-through effect of clouds of x,y,z coordinates. Surfaces and mesh representations are easier to understand than a cloud of measurements, and projected photo data makes it even more recognizable. *Requirement [06]: the data which represents the 3d structure of the scene needs to be presented to the users as surface data.*

Although, the preferred reconstruction method is laser scanning, it is not the only method of spatial/digital acquisition in crime scenes. Investigators take images for additional details or to cover areas that are occluded for the scanner. Furthermore, it is common to record audio material, which adds additional insight. Currently, that information is not spatially connected and the complete "picture" is difficult to assemble. *Requirement [07]: the acquired data from the*

different sensors (sound, imagery, measurements) need to be spatially indexed and fused into a global 3D model.

2.2.2. COLLABORATION

By law, multiple investigators are required to solve a crime. The investigators need various expertise and they need to work together (chapter 2.2). The investigators need situational awareness, i.e. awareness of the position with respect to conditions and circumstances, relative position or combination of circumstances at a certain moment in time, including the awareness of having realization and knowledge. The goal of the collaborative effort is to acquire scientific evidence that can be used in court.

There are two types of research that can be distinguished, tactical and technical, both of which must be performed by the team appointed to solve a crime (Jackson & Jackson, 2004). The tactical part of the team is responsible for interrogation, background checks, phone communication, etc. The technical part of the team is responsible for crime scene analysis and therefore they are the ones who visit the crime scene most often. As explained in section 2.2, the team is headed by a team leader and a prosecutor, and they work according to the phases listed in Figure 6.

Manning (2008) uses the term Crime Analysis Meeting (CAM) for the regularly planned collaboration meetings, during which an agenda of topics is used and investigators inform their colleagues about any new findings. Graphical material is used to present and discuss the findings, such as interrogation videos, images of victims, building plans and areal imagery. All this imagery material is used to create the shared situational awareness needed to advance the investigation (Boel et al., 2009). In addition, there is other material that is used in a-synchronous communication with colleagues (Lothridge & Fitzpatrick, 2013):

- On a crime scene, there is a lot of material, placed by investigators, that informs new arrivals about an ongoing investigation. E.g. special tape that seals locations, bullet location signs, measurement marker strips.
- A map or floor plan of the crime scene with important locations and additional remarks.
- Documents with hypotheses, background information and relevant audio and video material.

There are various kinds of markings for specific use cases, including special bullet markings for a victim or suspect, different types of tape for various conditions, etc. (Lothridge & Fitzpatrick, 2013).

It is common practice to build timelines of events: what material originates from where, when did a witness claim to have seen an event, what can we see at a certain point in time on video footage, etc. (Manning, 2008). On a smaller scale, the timeline of a crime scene is critical: who visited the crime scene when, what was moved, when did a certain part collapse, etc. This information is acquired over time and a-synchronously, and it is very relevant for obtaining situational awareness. Being able to retrace the steps of an investigator is important for co-investigators who become part of the research team. What parts and assets of the crime scene did their colleague cover, what was not covered? *Requirement [08]: the system has to be able to differentiate and visualize the regions that are mapped by multiple investigators.*

While, as explained, there is a great deal of asynchronous data sharing, there is also a need for synchronous data sharing. Few people are allowed on a crime scene simultaneously, which especially limits the 3D analysis investigators in starting with their work. The 3D reconstruction experts are consulted late in the investigation phase. Yet, especially for line of sight analysis, which is used in witness interrogation, there is a need for 3D information early on. Just as in video conferencing, the aim is iterative feedback with minimal latency (Kraut, Miller, & Siegel, 1996). *Requirement [09]: the system must have low latency when team members interact.*

2.3. INTERVIEWS

Spatial analysis in crime scene investigation is a young discipline; therefore, we cannot rely on just studying current workflows and the literature.

2.3.1. INTERVIEW SETUP

As an instrument to acquire additional knowledge, we conducted structured interviews with 5 leading experts in the field of 3D crime scene reconstruction. The interviews took place during the "International Association of Forensic and Security Metrology" (IAFSM) workshop in 2010. The interviewed experts were selected based on references from the initially contacted experts at the NFI. They were all heads of national crime scene investigation departments, with multiple years of experience. In this thesis, a person was designated as an expert if they had worked with 3D imaging soft- and hardware for at least 5 years. In general, only large police agencies in developed countries have the means and manpower to conduct 3D reconstructions. This type of expertise is situated in the larger cities, and typically only a few police officers are trained with the required equipment.

The interviewed experts were from the US, UK and the Netherlands. The interviews were conducted in a face-to-face setting and by telephone. The research was introduced in non-specific terms to make the interviewees aware of the reasoning in the interview. In total, the interviews took approximately 90 minutes per expert. The following section is based on the results of the interviews, and may be considered a shared opinion among the experts. An open-ended questionnaire was used to guide the interviews; see Appendix I - Questionnaire expert form for details. Quotes of the experts are used to reinforce statements. The interviews were not recorded but notes were written on the interview form and transcoded as US1, US2, UK1, NL1 and NL2.

2.3.2. SUMMARY OF INTERVIEWS

As explained in section 2.2, 3D reconstructions are conducted in the execution phase of the investigation, during which the responsible person, usually a judge, orders the reconstruction and analysis. In rare cases, a lawyer will order 3D reconstructions.

When asked whether judges are all aware of the recent advancements in 3D reconstructions the reactions were unanimously negative. Quote: *'It is not part of a judge`s education or curriculum to know about 3D reconstruction'.* Only those who keep track of advancements in crime scene investigation are aware of the capabilities, which is an indicator that diffusion in the application area is incomplete.

The main reason for 3D reconstructions is the severity of the crime, which, in most situations, means murder (Jenkins, 2005). Some agencies order 3D imaging for less severe crimes, as well. It is difficult to get exact numbers on 3D reconstructions that are conducted by crime scene investigators. However, in 2011, in the Netherlands, a crime scene was captured approximately 100 times (NL1, NL2). Few captured crime scenes are fully reconstructed; a full reconstruction is expensive and in many cases not required. Approximately one third of the 2D/3D captured scenes were reconstructed to some extent. *Quote (NL1): 'In most of the cases our scans go directly to long term storage (vault) because they are created as a precaution'.* In big cases, the data might be assessed years later, in which case the documentation needs to be precise, e.g. time stamped (UK1, NL2, US1). A new team might access the data without the original investigators being present; in that case, the data should speak for itself. *Requirement [10]: all steps in the process to acquire spatial 3D data need to be logged and time-stamped.*

The time needed to create a 3D reconstruction that can be used in court is considered to be extensive by all the experts we interviewed. Only a few people have the requisite skills to work with the software and the hardware and it is a laborious process. The reconstruction pipeline is discussed in section 2.2.1. The experts mentioned various issues that clutter the reconstruction pipelines, such as: *Quote (US1, UK1): 'Based on either inoperability between software packages and/or the differences between experts'*. More stringent than the time for reconstruction is the number of times the scene is captured. More often than not, the scene is captured once (one time slice), while an actual crime scene changes over time in many ways. For example, when a body is removed, biological traces transform and heat traces disappear. The moment the investigator notices something, the technology should be able to keep up and capture the moment of importance (NL1, NL2). Time is critical, both in respect of the degradation of the crime scene and the time needed to perform the 3D reconstruction (NL1, NL2).

The interviewed experts with experience in conducting physical blood pattern analyses or ballistic trajectories analyses had reservations about virtual reconstructions (US2, UK1). Fortunately, after working with the tools and comparing the results from both physical and real reconstructions, the level of confidence increased. There is evidence that this confidence is justified (Maloney et al., 2009), although it must be noted that older generations of professionals were reluctant to use the new tools. *Quote (UK1): 'I'm a happy man if I can watch and edit the documentation of a case that is physically reconstructed; doing this all virtual seems like a lot to learn'.* Tools that are daunting are less used; therefore, use of the system should be highly intuitive. There are many different coordination systems, hotkeys and workflow types that need to be learned to create a 3D reconstruction. Experts like the simplicity of a digital camera and the corresponding software.

The experts that execute the 3D capturing and virtual crime construction are generally the younger members of the police force (NL1, US2). The police force scouts informally for potentials with the right affinity, and sometimes hobby users surface within its own ranks (NL1, US2). Often, those recruited receive no specialized training for blood pattern analysis or other 3D reconstruction courses. The experts in the domains are generally the mature investigators who are trained in physical 3D reconstruction. Domain experts would like to be involved in the virtual reconstruction process and generally sit together with the virtual reality expert after the acquisition (US1, US2). Guidance works best when the investigators can have a dialog. By focusing on the 3D reconstruction together, it is much easier for experienced investigators to focus on the

reconstructions that matter, (NL1, NL2). *Requirement [11]: enable spatial collaboration by creating common ground in the form of a 3D model between domain experts and on-location investigator. Requirement [12]: Enable spatial collaboration by enabling conversation between domain experts and on-location investigator.* Expertise in 3D reconstruction is important; problem-solving knowledge is divided among different experts. Specific experts are able to reduce the time needed for reconstruction, because they rely on previous knowledge and can therefore separate the weak leads from the strong (US2, NL2, UK1). Reconstruction of a crime scene is an interpretation of recorded data, which is a domain in itself; accuracy, for example, is something few 3D reconstruction experts can prove.

Collaboration on 3D interpretation is difficult; it is easier to switch control than to explain what kind of interaction is desired (US1, UK1, NL2). Collaborating in a 3D space on a flat screen is difficult; having a 3D pointing/selecting tool is much easier than verbally explaining navigation in a complex 3D space. Angles of blood patterns or the line of sight exercise require complex navigation. Collaboration becomes easier when both can navigate and use tools in 3D space (Buck et al., 2011). *Requirement [13]: enable spatial collaboration by enabling 3D interaction between domain experts and on-location investigator.* Although there may always be reasons for revisiting the physical crime scene, 3D acquisition of the crime scene for analysis has advantages, such as freezing the time, forming a powerful communication means and the sheer infinite possibilities of what virtual reality can do e.g. simulation, timelines, problem segmentation, etc. (NL1, US1, US2). However, there is a downside, as noted by one of the interviewees: (NL1) '*The initial cases that used 3D visualization did not end well because the visualization experts interpreted too much of the data to their own liking'.* Because of this, visualizing the 3D reconstructed data in an as un-interpreted form as possible is most convincing to lawyers and judges. Effectively, this means that laser scans should be unedited and displayed as original un-interpreted data, with the different image types being spatially aligned. *Quote (NL2):* '*It is even becoming common to use the raw data for virtual visits'.* It is one of the reasons laser scanning data is replacing the inspection of the real crime scene for lawyers and judges. This shows that the virtual version of the crime scene is considered to be valuable and convincing enough to replace visiting the actual scene. *Requirement [14]: be able to capture and store raw data.*

Apart from the level of expertise needed to operate the hardware and the software, many complaints were heard about the portability of the equipment (US1, UK1, NL2). Having to move a tripod around is undesirable; the feet of the

tripod must be tied with special material and all equipment touching the scene must be sterilized every time it is used (NL1). *Requirement [15]: the system is not allowed to induce contamination of the scene.* Quote (*US1: 'The weight*[9] *of the scanners is too high, most colleagues complain about shoulder pain after a day's scanning'. Requirement [16]: the equipment's weight should not exceed ergonomic guidelines.* Next to the weight, there is another side to portability. *Quote (UK1): 'Most crime scene investigators are reluctant to take more than a digital camera to the crime scene, especially for the initial look around'.* Clearly, the investigators brook no interference in or interruption of what they are set to do. The equipment must be non-intrusive; sometimes an investigator needs his hands to investigate, but cannot lay the equipment down because of scene contamination. *Requirement [17]: the system is not allowed to interfere with the investigation.*

All the 3D reconstruction investigators stressed the importance of regularly sharing visual information within the task force to provide all team members with the most recent information. The meetings during the investigation make heavy use of images, film and sometimes laser scans (NL1, NL2). All the interviewed experts felt that 3D imaging technology was beneficial for collaboration; they claimed it was very good for communication. A crime scene investigator might build on the work of colleagues, but is currently hamstrung by the lack of collaboration tools.

Too few (US1, US2, UK1) investigation forces are equipped with 3D imaging technology. The good news is that, according to the HDS surveying group from Leica Geosystems, many are being sold to police forces all around the globe. While it cannot yet be said to be a standard tool in the box, there are indications that it is going to be (Jenkins, 2005).

A fair warning from the interviewed specialists. *"Do not try and replace all current ways of working. Crime scene investigators favor face-to-face communication, so the tool should not try to replace this phenomenon and revisiting the actual crime scene might be still be needed for physical testing of a scene hypothesis, such as bullet impacts in certain material or new insights that needs a wider search area or further specifics".*

[9] According to the latest lifting ergonomic guidelines, objects that are to be lifted for just a few minutes up to shoulder height may not weigh more than ~7kg (Mutual, 2004).

2.4. SUMMARY OF REQUIREMENTS

The requirements compiled from the previous sections are summarized in Table **2**. They are categorized in order of appearance in the text and according to their relation to either; Hardware (HW), Software (SW), Interaction (INT) and collaboration (COL).

	System requirements				
Nr.	*Description*	HW	SW	INT	**COL**
01	*The system must be able to acquire and store spatial oriented metric 3D data from a pristine environment.*		*x*		
02	*The system should allow an investigator unhindered view of the crime scene.*	*x*			
03	*Allow the system to share information to the investigation team during or shortly after acquisition.*		*x*		
04	*The time between setting up the system and the start of using the system for geometry capture should be less than 30 minutes.*		*x*		***x***
05	*The system must be able to align acquisition data without disturbing the pristine characteristics of the scene.*	*x*	*x*		
06	*The data which represents the 3d structure of the scene needs to be presented to the users as surface data.*		*x*	*x*	***x***
07	*The acquired data from the different sensors (sound, imagery, measurements) need to be spatially indexed and fused into a global 3D model.*		*x*		
08	*The system must be able to differentiate and visualize the regions that are mapped by multiple investigators.*		*x*		
09	*The system must have low latency when team members interact.*	*x*	*x*		
10	*All steps in the process to acquire spatial 3d data need to be logged and time-stamped.*		*x*		
11	*Enable spatial collaboration by creating common ground in the form of a 3D model between domain experts and on-location investigator.*		*x*		***x***
12	*Enable spatial collaboration by enabling conversation between domain experts and on-location investigator*		*x*		

13	*Enable spatial collaboration by enabling 3D interaction between domain experts and on-location investigator.*			*x*	
14	*Be able to capture and store raw data.*		*x*		
15	*The system is not allowed to induce contamination of the scene*	*x*	*x*		
16	*The equipment's weight should not exceed ergonomic guidelines.*	*x*			
17	*The system is not allowed to interfere with the investigation.*	*x*			

Table 2 Summary of requirements

3. ARCHITECTURE

As introduced in chapter 1, many domains are touched by this research. Most multidisciplinary research is characterized by a wide solution space (Andreasen & Brown, 2004). Exploration of deep domain crossovers can easily consume too much time without being effective. To mitigate the risk, the approach in this thesis is to compartmentalize the research in the areas that matter. By examining the research questions and the requirements, a preliminary supporting architecture is exposed. However, it is acknowledged that multiple architectures might fulfil the requirements and that the proposed architecture matches the requirements enough to satisfy the need.

This architecture chapter outlines the dominant building blocks that require exploration for answering the research question. The system design approach from INCOSE (Walden & Roedler, 2015) is used throughout the chapter. The guidelines that are explained in INCOSE provide guidance in designing the system, and the steps are adopted. The architectural building blocks will be used to focus the background literature study and function as a basis for the design chapter. In line with Figure 9 from INCOSE, the objectives and mission of the system have been formulated in Chapter 1, followed by the functional requirements in Chapter 2. Accordingly, the time has arrived for the next key task, the development of concepts and architectures.

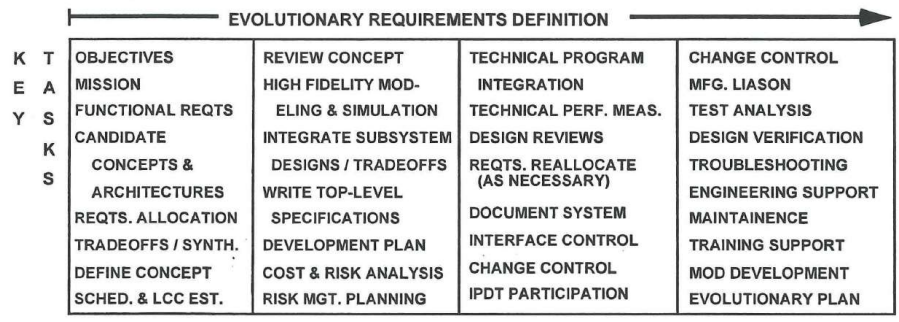


FIGURE 9 KEY PHASES AND TASKS IN PROGRAM LIFE CYCLE, INCOSE (WALDEN & ROEDLER, 2015)

According to INCOSE, a system is: An interacting combination of elements to accomplish a defined objective. These include hardware, software, firmware, people, information, techniques, facilities, services, and other support elements. To define the high-level architecture in a system, it must be broken down into

elements or segments. The elements of this system are identified in the next chapter.

A rudimentary set of necessary capabilities for the architecture can be derived from the requirements in Table **2**. The high-level architecture description is a formal description and a representation of the system, organized in a way that supports reasoning about the structure of the system.

3.1. HIGH LEVEL ARCHITECTURE

The high-level architecture needs to reflect all the high-level requirements listed in Chapter 2. The approach is to layer requirement specifics on a base system, and thus gradually increase the complexity. The highest level is a system, consequently broken down by elements, subsystems, assembly, subassembly, components and lastly, parts. The approach we take is to start with elements that are formed by combining subsystems in which the elements are roughly on par with requirements.

For mediated reality, a virtual reality element is a necessity (Mann, 2003). A virtual reality element makes the virtualized physical world digitally accessible and is therefore, supported by requirements, a foundational first step. (06, 07). A virtual reality element has at least three basic subsystems: (1) virtual reality content, (2) a scene handler and a (3) renderer that is able to display the results (Zwern, 1995). Data flows from digital content, to handler to renderer (Figure 10 A). Software needs to run on a hardware system, which is not defined here, as this can currently be anywhere. The scene manager is the central subsystem that handles incoming and outgoing data streams. The off-line world data represents the digital data needs to be visualized, such as 3D meshes, solids, textures, material definitions, etc. The scene manager is the orchestrator of digital assets; it is able to load, assemble and manipulate models, as well as having state knowledge. The renderer is the virtual camera in the scene: it allows a rasterized view (render) of any given position in the scene.

The requirements (01, 06) dictate that the virtual reality content needs to overlay a physical scene, the physical scene needs to be captured and a position with respect to the physical world must be established. Both the virtual reality content and the real-world data needs to be combined to augment a scene (Figure 10 B). The scene manager is therefore the most central component of the architecture; the manager needs to be able to cope with incoming sensor data, library objects and other inputs and to feed the compositor with the relevant

input. Furthermore, the hardware needs to be portable, as to create augmentations it needs to be present on location.

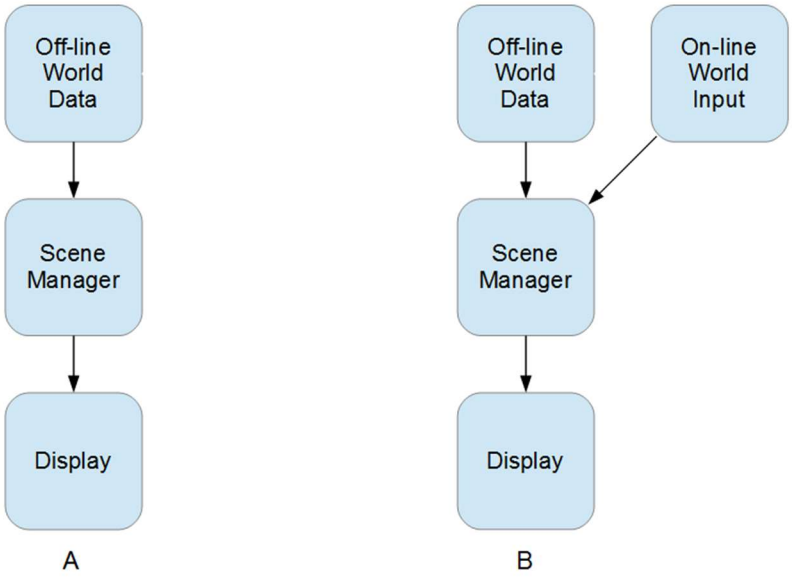


FIGURE 10 A MINIMAL VIRTUAL REALITY SYSTEM (A) AND A MINIMAL AUGMENTED REALITY SYSTEM (B)

With these subsystems, the system is only capable of overlaying content. Requirement (13) states that users need to be able to influence the digital representation of the scene. This adds the next level of complexity to the system; a user should be able to provide input to the system, and hence a software tool is necessary to replicate current physical actions (Figure 11). Current physical actions that require 3D analysis include blood pattern analysis, discussed in section 2.1. As the interviewed experts discussed in chapter 2 noted, blood pattern analysis is a good candidate for virtualization. The tool needs to allow for bi-directional traffic in order to exchange information with the scene manager. An example might be that the locations of digital assets need to be updated. The scene manager now needs to be extended with user input and tool capabilities, regulating what a user can do, its state and the tools' state.

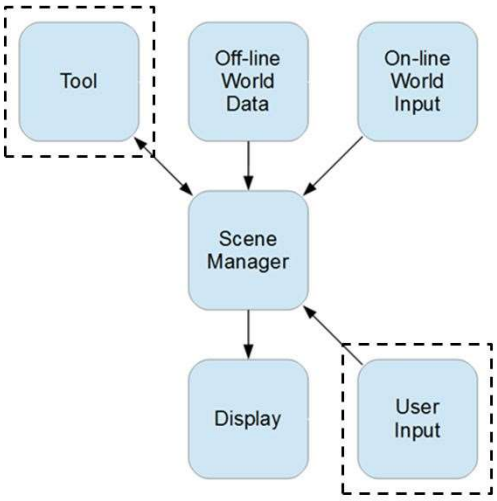


FIGURE 11 A MINIMAL AUGMENTED REALITY ARCHITECTURE THAT ALLOWS USER INPUT

Next, as requirements 10 and 14 state, the system needs to be able to record the sessions for replay. To be able to replay, we need to know the state of the scene in the scene manager, the input the user provided and the raw acquisition data. All communication is bi-directional, as for replay, the system needs the original data and states (Figure 12). The states provide the result of the system and the original data provides the real-world content.

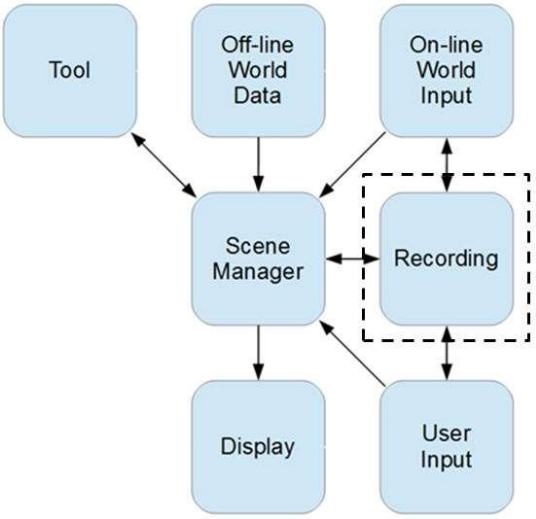


FIGURE 12 NEAR COMPLETE ARCHITECTURE OF AUGMENTED REALITY SYSTEM WITH RECORDING FUNCTIONALITY

Furthermore, to complement the system, logic needs to be represented and the system needs to talk to a similar system though a communication network. The scene logic guards the rules of the system, i.e. which user is allowed to do what, and restricts a scene to follow rules. Although the logic is not a requirement formulated in chapter 2, the functionality is required for tooling. Multiple people will contribute to the scene; the scene logic must track who can do what, when and mitigate collision. The communication demanded pursuant to requirements 03, 08, 11, and 12 is reflected in the instanced subsystem from Figure 13. The architecture can communicate with an instance of itself that runs somewhere else (Figure 13). A user in another instance might change the scene, which should be reflected throughout all the connected systems, in accordance with requirement 11. By default, this is bi-directional traffic.

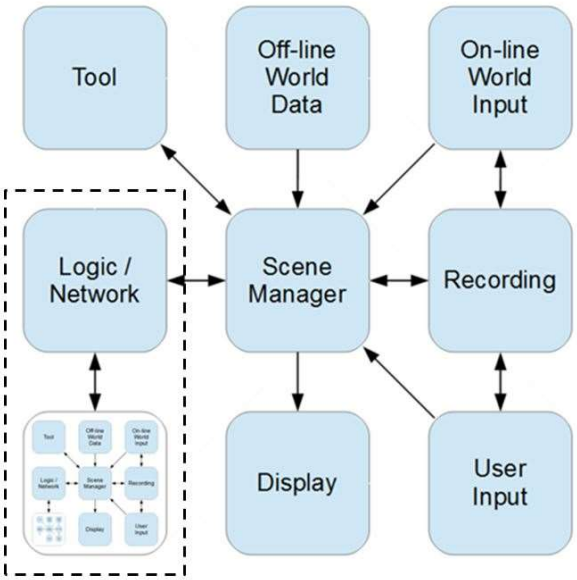


FIGURE 13 COMPLETE HIGH LEVEL ARCHITECTURE

3.2. EXPOSING THE HIGH-LEVEL ARCHITECTURE

The high-level architecture is too coarse to base a background research strategy on. This chapter is dedicated to refining the required depth to gain a better understanding of the relations and interactions between the different components of the high-level architecture.

Figure 14 portrays the unfolding of the complete high level architecture. Apart from more logic blocks, two symbols have been added to the figure, multiple versions of components in the architecture and data bases (DB). Multiple versions are depicted as empty blocks behind blocks; DBs are the cylindrical containers. Multiple versions of components facilitate in understanding the flexibility of data flows and allow for swapping out subsystem with variations. The user input might, for example, be keyboard, gesture or voice. Especially where research is required, flexibility is important.

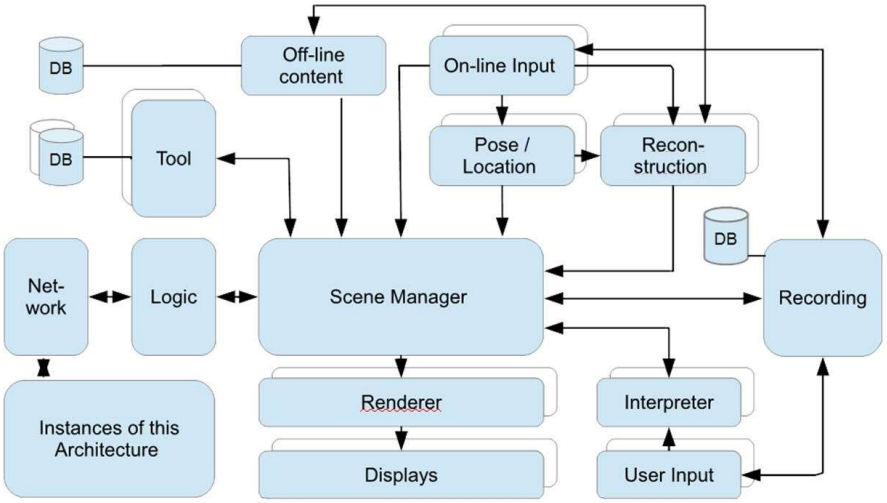


FIGURE 14 UNFOLDING OF THE HIGH-LEVEL ARCHITECTURE

3.2.1. SCENE MANAGER

The scene manager is the central piece of the system; off-line and on-line data needs to be aggregated into a renderable asset, including the tool and user actions. This module always has the latest stage of the virtual scene and the position of the scene. Data will accumulate and updates to the scene need to be handled. The scene manager handles the orientation and position of the scene in a Cartesian coordinate system.

The scene manager controls, among other things, user menus, occlusion of objects/assets, which user is in the scene and the state. Correct replay of the scene requires the usage needs to be recordable. The assembled state of the scene is used as the input for rendering.

3.2.2. OFF-LINE CONTENT

As discussed in chapter 2, investigators need tools to conduct their work and the tools need to be virtually shared for collaboration according the requirements. A 3D/2D library of objects and symbols for the tools needs to be available to the users (Figure 14). Furthermore, a reconstruction by previous users of the system needs to be stored and available, as depicted by the arrow from reconstruction.

3.2.3. On-line input, pose/location and reconstruction

Sensors gather information from the crime scene. That data needs to be stored for reviewing purposes and needs to be interpreted correctly for use in the pose/location subsystem (Figure 14). Sensors will have deficiencies that will require filtering (Durrant-Whyte & Bailey, 2006) and it is likely that multiple sensors will run next to each other. Commonly used sensors are infrared, RGB and depth that have complementary capabilities. Replay of the events will require processing of recorded data as if it is a live feed for an exact reconstruction.

Next to the sensors that gather the scene information, a module needs to compute the location of the sensors with respect to the scene at all times. This information is necessary to correctly overlay digital content onto the scene. The sensor pose location will be used for multiple purposes, including localizing possibly multiple sensor outputs with respect to each other, feeding a reconstruction engine for aggregating the data and allowing virtual objects or markers to be placed in the scene at their correct scale and location, as stated in requirements (05, 07).

The reconstruction engine will aggregate the 3D information into a coherent map. It will represent the pristine representation within the sensor constraints. Prior information from previous system users will have data that complements the reconstruction, hence the arrow from the off-line content database.

3.2.4. Logic

The logic can be best described as a helper to the scene manager. While the scene manager has the latest state of the scene, the logic informs the scene manager of impossible combinations of objects or interactions. The logic controls the order in which scene manipulations can take place, who has the right to what, etc. The scene is also governed by virtual environment laws, such as that certain objects should stick to the ground and not float, or if there is no data, analyses are impossible.

Furthermore, the logic is also responsible for syncing the multiple instances of the system that might exist, co-located colleagues and observers.

3.2.5. Network

Because of the requirements 03, 08, 11, 12, we already know that on-line connections between co-located colleagues are essential. Off-line and on-line users must be able to exchange information and collaboratively work on a scene.

Off-line in this case also entails that after the 3D acquisition is done, the data needs to be accessible.

The exchanged information needs to be optimized for network protocols and must only contain the relevant information and the state of multiple instances that need to be synced. State, data and direct means of communication are likely to need to be streamlined in the network module. The tools that act on the scene need to sync their status, the scene map needs to be constantly updated and the off-line and on-line users need to be able to communicate.

3.2.6. DISPLAY AND RENDERER

Both the investigators on location as well as the co-located colleagues need 'renderings' of the most recent state of the scene. The scene manager has all relevant information available in a scene graph; the scene graph arranges the logical and spatial representation of a graphical scene. The scene graph information is the input for the renderer, but the renderer needs to know the specifics of the display it needs to render and to compensate for the device specifics. The virtual camera in the scene needs to know its required position, what deformation needs to be applied and other rendering specifics such as resolution, stereoscopic and bit depth.

3.2.7. USER INPUT AND INTERPRETER

In the investigators' use case, the collaborative efforts are related to 3D interactions with the scene. Generally, when user interaction with a scene or tool is required, the input of the user needs to be interpreted in the light of the scene data and feedback he or she has received, to be able to understand whether the interpretation is correct. If a user attempts to select a virtual object, a selection command is given in a 2D or 3D space with an intersection ray. The ray collides with the object if the object is selectable and the bounding volume is hit. The accuracy of the selection, the hit target and other collision objects might all affect the success.

Because reality captured data is not as 'clean' as CAD modeled data, the interpretation of intent and actual response requires special attention. Both the environment data as well as the input data will be noisier than a CAD or 3D game environment.

For action replay to be possible, the raw input needs to be recorded and the interpreter needs to be able to read raw results from the recorder database in order to apply the interpretation.

3.2.8. TOOLS

Recording a scene and displaying the results to a co-located colleague is not the same as in-depth collaboration. Tools that conduct scene analysis need to be created to support joint interaction and collaboration.

The tools will replace current physical equivalents, must be 3D compatible and allow for collaborative workflows.

3.2.9. RECORDING

As discussed in chapter 2, a judge or colleague needs to be able to rewind the time to understand prior work on the crime scene. The recording module that takes in raw data from the used sensors and records the state of the scene is critical to the success of the system.

Apart from being able to replay, the recording subsystems also allow researchers to playback experiments and leverage the data to tune algorithms.

3.3. CONCLUSIONS

This chapter framed the architectural elements that reflect the majority of the requirements derived in chapter 2 and made a start with addressing the research questions. Creating the conceptual architecture simplified the framing of the challenges: (1) to develop a software system as discussed in this chapter, (2) hardware that supports the software capabilities, (3) interaction with the system and (4) bridging collaboration.

According to INCOSE (Walden & Roedler, 2015), the functional requirements evolve into concept architectures, which are the stepping stones to the tradeoffs and synthesis. To be able to weigh tradeoffs, the challenges must be explored in more detail. The architectural components discussed in this chapter will therefore facilitate background and related work research.

4. Background and Related Work

To understand the state of the art in augmented and mediated reality, first we introduce systems that are closest to the architecture outlined in chapter 3. Studying these state-of-the-art systems will elicit the components that require deeper research. This is explored in the remaining chapters.

4.1. State-of-the-art "Augmented" Reality Systems

The literature shows that few systems are referred to as mediated reality systems, but that various terms are used for similar capabilities. To gain a better understanding of what is available, search terms such as augmented reality, real world interface, mixed reality and location aware virtual reality were used. The number of systems in this domain is vast. As discussing them is neither feasible or desirable, the author of the present study selected those that best represent the state of the art. For the sake of simplicity, the requirements that require regular software design practices have been left out.

4.1.1. SixthSense

This is a wearable gestural interface that augments the physical world around us with digital information and lets us use natural hand gestures to interact with that information (Mistry, Maes, & Chang, 2009). The augmentation takes place with projected imagery. The ideas behind this system overlap with our requirements. The user is free to walk around, does not need to prepare the environment and has digital augmentation. However, there are discrepancies too: the workflow has no mapping of the environment and there is no alignment of the physical and virtual world (Figure 15).

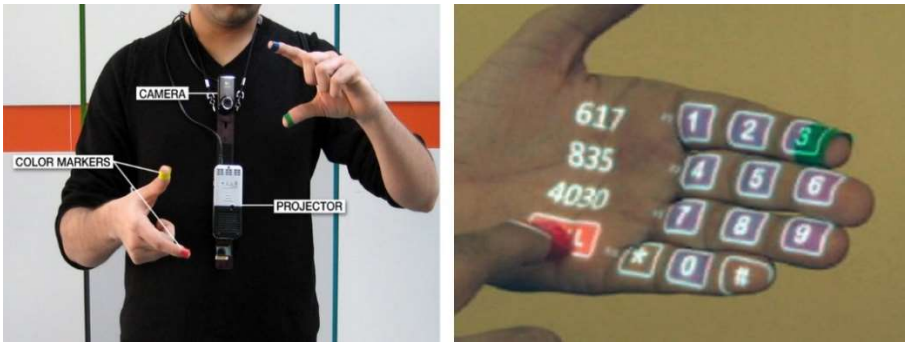


FIGURE 15 SIXTH SENSE, COURTESY OF PRANAV MISTRY (MISTRY ET AL., 2009)

There are certainly elements in the system that are favorable, including interaction with gestures; gestures supported by multi-touch systems, freehand gestures and iconic gestures (in-the-air drawings). The intuitiveness of the integration with easily recognizable and memorable gestures deserves further investigation.

The dominant requirements that are missing are: (1) the mapping of environments and (12) collaboration. Important requirements that are fulfilled are (13) spatial interaction and (6) overlay of digital data.

4.1.2. ARTHUR

The name is an abbreviation of: A Collaborative Augmented Environment for Architectural Design and Urban Planning (Broll et al., 2004). It is an augmented reality enhanced round table to support complex design and planning decisions for architects. With this system, the user is not free to walk around but confined to a round table. There is gesture interaction that is detected by a camera placed above the round table (Figure 16). The system corresponds to a large degree with the architecture described in chapter 0, as here, augmented reality is not only used to add information for one user, but also serves as a collaborative platform. The authors claim that "The system enables designers to truly enter into a collaborative form of design, which is beyond the mode of taking turns or creating individually, thus far not provided by any other design tool".

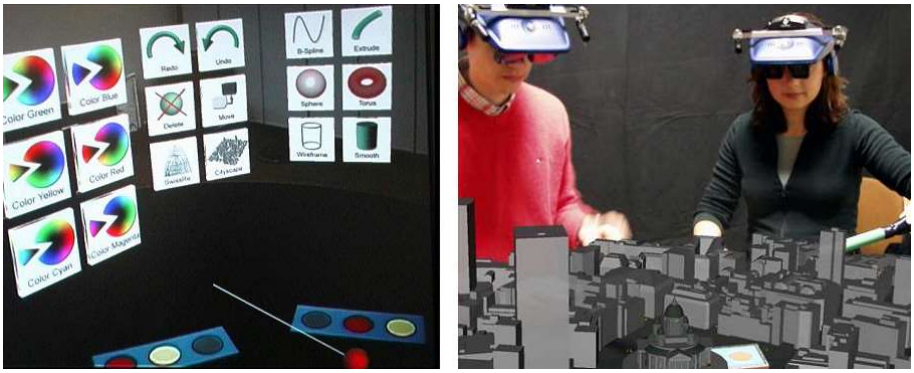


FIGURE 16 ARTHUR, COURTESY OF (BROLL ET AL., 2004)

There are a number of interesting architectural components in ARTHUR. First, the morgan architecture that enables the communication, rendering and states of all the information, shares similarities with the description of the scene

manager in section 3.2.1. Secondly, the hardware, AddVisor 150[10], is modified with two cameras to detect universal interaction handlers, which is similar to our interaction module and display. Thirdly, the collaboration aspect that is built into the system: colleagues interact with each other in the virtual space, but must be physically in the same space, as well.

Apart from mapping the environment, working in a pre-set environment and having colleagues in the same room, this system is a valuable reference to fulfil the requirements from chapter 2. The authors claim positive collaborative side effects of working in an augmented environment.

The requirements that are missing are: (1) the capability to map environments. Requirements that are fulfilled are (13) spatial interaction and (6) overlay of digital data.

4.1.3. FARPDA

The title is an abbreviation of First Augmented Reality Personal Digital Assistant. This system is described by the authors as the first stand-alone augmented reality system with self-tracking running on an unmodified personal digital assistant (PDA) with a commercial camera (Wagner & Schmalstieg, 2003). The device knows where it is in relation to the environment by scene prepared targets, is mobile and runs on a 3D engine for augmenting the imagery (Figure 17).

FIGURE 17 FARPDA, COURTESY OF (WAGNER & SCHMALSTIEG, 2003)

What is interesting to this research is the optional dynamic workload-sharing with a backend server, which allows the computationally expensive computer

[10] http://products.saab.se/ or http://en.souvr.com/product/200712/185.html, last visited May 2017

vision calculations to be outsourced to the server via a wireless network, as well as the integration of the handheld platform's software into the Studierstube (Szalavari et al., 1998), a research framework for mutual re-use of resulting software components between workstation/notebook and PDA-based augmented reality. This experiment had already proved as early as in 2003 that 3D augmented reality applications could run on mobile devices.

The interaction is based on pen-based touch with the personal digital assistant and is expressed as 2D overlays to the user. The system can be best described as GPS based car navigation, in which the pose does not come from the GPS but from target recognition. The goal of the system is augmented reality for everybody with custom-off-the-shelf hardware; although the author acknowledges that Head-Mounted Displays (HMD) give the highest immersion among the device classes for augmented reality (Wagner, 2007), these were deemed not to be sufficiently commoditized.

The requirements that are missing are: (1) the mapping of environments and (13) spatial interaction. However, the systems allow for (12) collaboration through server side communication and an overlay of digital data (6).

4.1.4. MARS

The title is an abbreviation of Mobile Augmented Reality Systems. As described by Hollerer, Feiner et al. (1999), a mobile user, tracked by a centimeter level real-time-kinematic global positioning system (GPS) and an inertial/magnetometer orientation sensor, and equipped with their prototype backpack computer system, experiences the world augmented by multimedia material displayed on a see-through and hear-through head-worn display. The interaction relies on see-through head-worn displays, in conjunction with 6DOF head and hand trackers, and 3DOF object trackers, to overlay and manipulate virtual information.

The pose estimation works in unprepared environments as does the corresponding software architecture. Furthermore, it is one of the first mobile augmented reality systems. There are several comparable systems, such as TinMith (Piekarski & Thomas, 2001), or BARS (Julier, Baillot, Lanzagorta, Brown, & Rosenblum, 2000). Comparing these systems to our requirements, yielded a few similar systems such as: TinMith uses 3D interaction with objects in the scene with data gloves for modelling and analysis, while BARS has smart information filters that, depending on the relevant mode, decides on the necessary information. The key learning from MARS is related to describing the

components necessary to create the system with the hardware available to the researcher, which is significantly improved today.

The major requirements that are missing are (1) the mapping of environments; and (12) collaboration capabilities. Special interaction (13) and the overlay of digital data (6) are provided for.

4.1.5. DWARF

The title is an abbreviation of Distributed Wearable Augmented Reality Framework (Bauer et al., 2001). The authors propose a new approach to building an augmented reality system using a component-based software framework.

The proposed framework consists of reusable distributed services for key sub problems of augmented reality such as: the middleware to combine them, and an extensible software architecture. The authors have implemented services for tracking, modelling real and virtual objects, modelling structured navigation or maintenance instructions, and multimodal user interfaces.

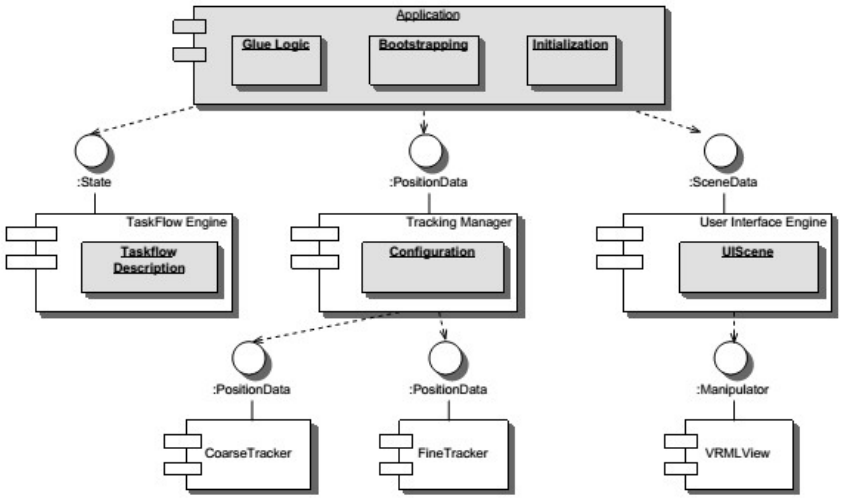


FIGURE 18 THE DWARF ARCHITECTURE, COURTESY OF (BAUER ET AL., 2001)

As a proof of the DWARF concept, an indoor and outdoor campus navigation system using different modes of tracking and user interaction has been developed. The architecture blueprint (Figure 18) appears to be similar to that of the architecture discussed in chapter 3, and includes clear modules, such as a tracking manager, online and off-line data separation and a state engine.

Important requirements that are missing are the capabilities to map environments (1) and collaboration (12). Spatial interaction (13) and overlay of digital data (6) are well represented.

4.1.6. SHAREDVIEW

SharedView (Kuzuoka, 1992) is spatial workspace collaboration system. An approach supporting its use via a video mediated communication system is described. Based on experiment results, the movability of a focal point, the sharing of focal points, the movability of a shared workspace and the ability to confirm viewing intentions and movements were determined (Figure 19).

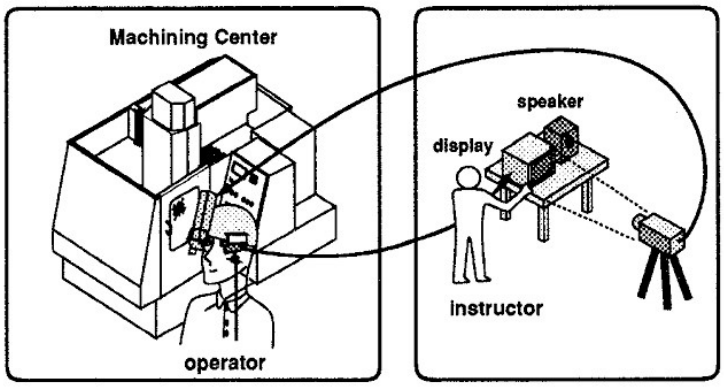


FIGURE 19 SHAREDVIEW'S CAPABILITIES EXPRESSED (KUZUOKA, 1992)

The systems discussed up to this point did not separate the physical location of the users. SharedView's capabilities allowing communication between co-located users and with digital means in an instructor-to-operator configuration makes it relevant in answering the research question.

There are advantages to using the SharedVlew system in a scenario where an instructor can easily show, with the help of gestures, what they want an operator to see. The need for the instructor to confirm where the operator is looking is also important.

Once again, however, no capabilities to map environments (1) is provided. However, the collaboration requirement (12) is especially well done. Important requirements that are partly fulfilled are (13) spatial interaction and (6) the overlay of digital data.

4.1.7. EXISTING SYSTEMS DISCUSSION

The literature study of existing augmented reality systems showed that none of the systems fulfil the requirements from chapter 2, nor could they offer the full architecture stated in chapter 3. Table 3 summarizes the results, from which it is evident that the requirement lacking most frequently concerns the 3D mapping capabilities. The first column shows whether the system can create, at the very least, a sparse map; the second column indicates whether interaction mechanisms for operating the system are available for a user. In the third column, it is shown whether or not augmentation occurs according to a tracked pose perspective, while the fourth column clarifies whether the system also offers visual collaboration means. Purely from an architectural standpoint, DWARF is the most interesting, while from a collaborative standpoint both ARTHUR and SharedView provide interesting insights. From an interaction perspective, the Sixth Sense system has gesture qualities worth exploring.

	3D Mapping (1)	**Interaction (13)**	**Digital overlay (6)**	**Collaboration (12)**
SixthSense	X	V	X	X
ARTHUR	X	V	V	V
FARPDA	X	V	V	V
MARS	X	V	V	X
DWARF	V	V	V	V
Sharedview	X	V	V	V

TABLE 3 AUGMENTED AND MEDIATED REALITY SYSTEMS SCORES, REQUIREMENTS BETWEEN BRACKETS

The systems discussed favor head mounted displays over hand held displays for spatial related tasks. This preference is thanks to advantages such as a the more complete user view offered by head mounted displays, the stereoscopic characteristics and hands-free operation. Interpretation of 3D on a mobile 2D display and the limited user interface paradigms narrow this research to head mounted displays. This is a new requirement evolving from the background research. *Requirement [18]: the system needs to use head mounted displays for digital overlay.*

None of the systems use real-time map making as a requirement. This appears to be a somewhat neglected topic in augmented reality research, which is not surprising: a pristine environment is not a common requirement. The most used approaches for aligning the digital and virtual environments are either marker or global positioning system based.

Interaction with spatial data (a 3D view) is most often achieved with either data gloves or a handheld marker. Neither seems to be a valid solution with respect to the hands-free requirement. Further research is necessary to determine how the investigators interact with the overlay on the scene.

The collaboration paradigm from SharedView is too limiting, both because it restricts the control of the remote experts and because the hardware setup prohibits its applicability at a crime scene. Few augmented reality systems focus on co-located collaboration. On the continuum from Miligram (2006), this is mainly addressed by tele-presence. Although tele-presence does not address virtual reality as a shared environment, generally the domain addresses digitally transporting a person's presence. This research resides on the boundary of both domains, which is not addressed by the above systems that function in a 3D integrated environment.

While not mentioned in the individual chapters, most of the systems use Linux as the development environment and rely heavily on existing libraries[11].

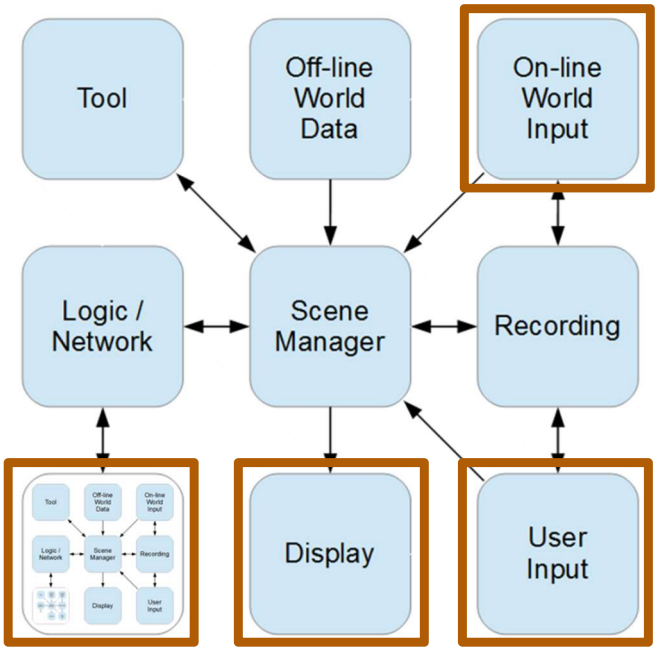


FIGURE 20 SIMPLIFIED ARCHITECTURE WITH RESEARCH AREAS HIGHLIGHTED

11 https://en.wikipedia.org/wiki/List_of_augmented_reality_software, last visited June 2017.

The following chapters highlight the less mature elements in the elaborated systems (Figure 20), namely, mapping pristine environments, co-located collaboration, displays and utilitarian spatial interaction.

4.2. MAPPING PRISTINE ENVIRONMENTS

As mentioned in chapter 2, crime scene investigators use mapping technology that is mobile and contactless. The created spatial maps are input to the analysis. Focus is on line-of-sight[12] mapping technologies. The following chapter will provide insight into the technologies that are used to geometrically map environments. A single measurement is not a map. To create a geometric map, the measurement device must be positioned in space. In this case, the fact that the environments to be captured must be pristine narrows the field of candidate systems quite considerably.

4.2.1. MEASURING ENVIRONMENTS

Pristine environments entail that no pre-knowledge of the environment is available. A spatial metric understanding of the environment is required and the method used to acquire this understanding cannot be destructive to the environment. This means that sensors must be used to map the environment. Ideally, the investigators will carry sensors on their bodies to circumvent contamination. First, however, it is essential to understand the basics of remote sensing.

The electromagnetic spectrum is used to measure our environment. The reason we see artifacts is because they emit, reflect or transmit a part of the visible spectrum of the electromagnetic spectrum which we call visible radiation (light). Most measuring techniques make use of this portion of the electromagnetic spectrum. Electromagnetic radiation with a wavelength between 380 nm and 760 nm (790–400 terahertz) is detected by the human eye and perceived as visible light, Figure 21

[12] Although mobile CT and MRI technologies exist for acquiring volumetric data, the author does not discuss volumetric acquisition, because it is currently not used on location and the type of cases outlined do not require it.

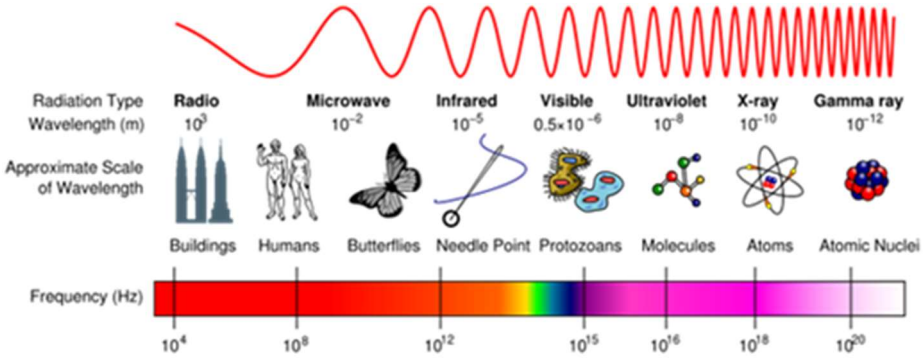


FIGURE 21 ELECTROMAGNETIC SPECTRUM, COURTESY OF (BALWER, 2013)

For our environment, we are interested in the sensors that can obtain metric data that can be used for spatial reconstruction, interaction and analysis. Technologies that change the scene are out of the question. Ultraviolet light kills DNA and beyond ultraviolet is even more destructive. On the other side of the spectrum are the longer wavelengths. These are also unsuitable for measurements because the wavelengths exceed the size of the objects to be measured. *Requirement [19]: the system senses with technologies that function in the visible light and infrared light.*

Light has been used in several ways to measure artifacts. Fundamental research in sensor technology and computer vision has been conducted by Beraldin et al (Beraldin et al., 2000), who often use a separator for the technologies depicted in Figure 22. Passive scanners do not emit any kind of radiation themselves but detect reflecting ambient radiation. Most of these scanner types use visible light, which is detected by common image sensors.

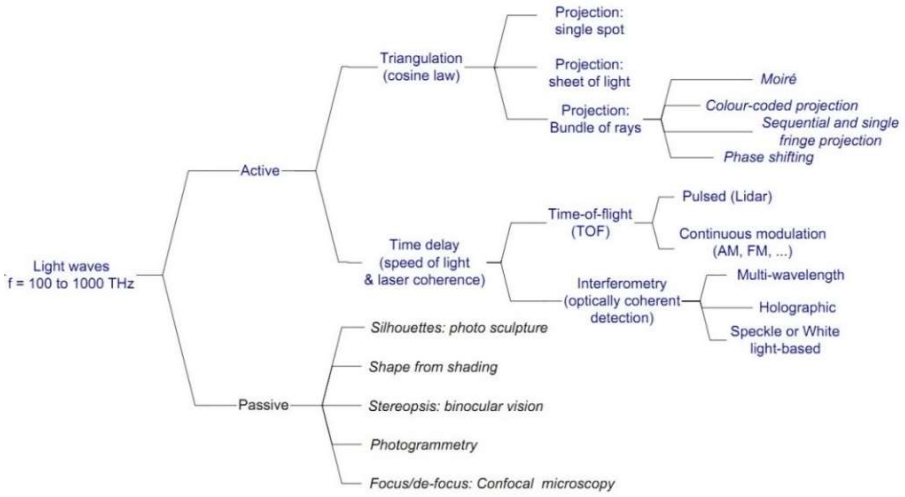


FIGURE 22 MEASURING 3D SHAPE: LIGHT WAVES, COURTESY OF (BERALDIN ET AL., 2000)

The sensing conditions narrow down the likely candidates too; the intended users work both indoors and outdoors, and have relatively short windows to capture. With active measurement technology, the device sends out a signal and measures on the returned response. This is different from passive technology, in which the device monitors and measures external signals.

Principles underlying range measurements are triangulation and time-of-flight techniques (pulse or phase-shift). Triangulation, the basis for many measurement techniques, was used by the ancient Greeks to make geodetic measurements and can still be found in the laser-based 3D cameras. The basics of triangulation are depicted in Figure 23. In simple terms, it involves finding the value of 'z' based on the known distance 'd' and angle 'α'. Using time-of-flight principles, the distance is calculated based on the knowledge of the speed of light and the Δ time between sent and received. There are many hybrid solutions and variations, including phase shift, in which the short wavelengths are integrated with a long wave to differentiate the returned pulses, resulting in much higher sample rates.

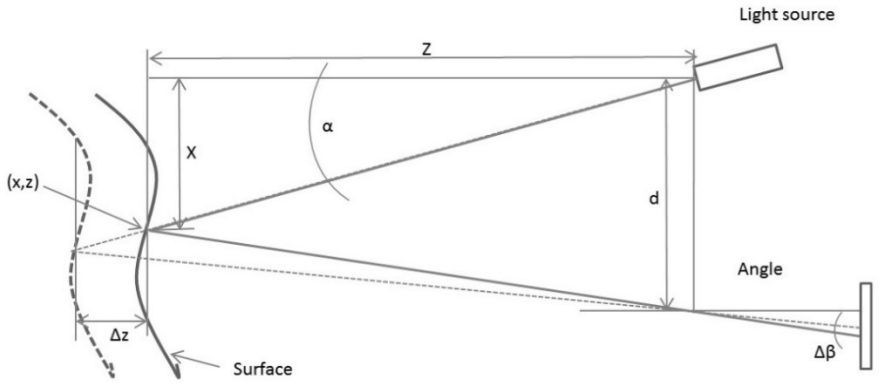


FIGURE 23 TRIANGULATION-BASED LASER PROBE WITH ACTIVE LASER (BERALDIN ET AL., 2000)

Frequency, resolution and accuracy are other important technology decision criteria. Requirement 9 stated that low latency speeds are necessary, which means the graphics are updated frequently, so the user experiences no noticeable delay. The frame rate for film that is generally accepted as fluent is ~25 frames t (Watson, 1986). The resolution of the map depends on the distance and footprint of the technology, which, in laser scanning, is defined by the collimated laser beam size and in photogrammetry is limited by the resolution of the sensor. The accuracy is not only dependent on the resolution, but also on the material properties, pose estimation and environmental conditions (Wehr & Lohr, 1999).

Acquiring a three-dimensional model of complex environments is thoroughly discussed by El-Hakim (2001), who highlights a number of important aspects. Acquiring a 3D dimensional model requires more than a single point measurement. There are roughly two methods: know the angle at which a pulse is sent and received with respect to the previous signal or maintain some distance between the send and receive (or second send). Of course there are all kinds of variations, such as projected lines/patterns, but effectively, these are the two ways to achieve this (Beraldin et al., 2000).

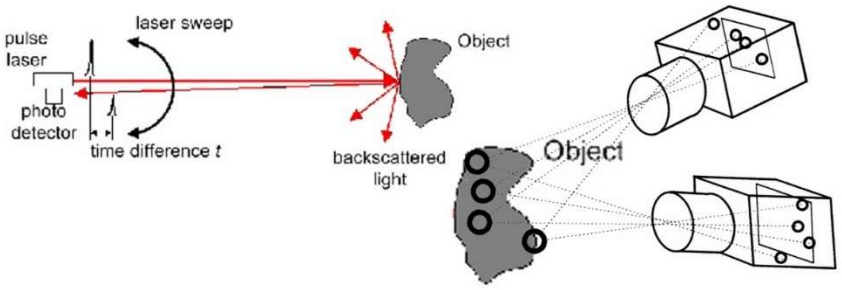


FIGURE 24 TWO METHODS TO EXTRACT 2,5D SPATIAL KNOWLEDGE

Depending on the solution, the third dimension is extractable by knowing the angle of the emitted signal or the offset between the send-receive (send-send), as shown in Figure 24. The method on the left-hand side of the figure is mostly used by active technologies; that on the right by passive technologies.

4.2.2. ACTIVE RANGE SENSING

Both active and passive range sensing methods have advantages and disadvantages. In this section, the dominant technologies in active sensing are highlighted.

A device that can generate a wave of light using only a very narrow band of the spectrum is called a laser; it emits light in a narrow, low-divergence beam with a well-defined wavelength. This, in contrast to a light bulb, which emits into a large solid angle and over a wide spectrum of wavelengths. Because of these properties, a laser can be used to measure reflectance, and thus metric distance. They can be found in all kinds of appliances, such as DVD players, laser pointers, mousse, etc. In general, the more energy a laser possesses, the longer the distances we can measure. The laser must be "aimed" to obtain more measurements, which are usually done with a directable mirror. Modern developments are moving in the direction of an increasingly favorable form factor for this technology (Lincoln, 2010). A recent example is Microvision's Pico projector SHOWWX,[13] which is the size of a smart phone. However, every coordinate has to be acquired separately.

The laser can also be used per pixel. These are usually called time-of-flight cameras. The idealized form factor is currently not available, but, contrary to classical 3D imaging methods such as those based on triangulation, TOF cameras

[13] http://www.microvision.com/category/showwx/ Last visited February 2017

can be miniaturized up to a point without compromising their performance. An example is the camera from CSEM (2009). Compared to digital cameras, current models have very low resolutions.

LADAR – Laser Detection and Ranging is a common choice when it comes to the use of flash-like technology. With Flash LADAR, the scene is flooded with a diffuse laser light and a focal plane array (FPA) is used as a detector to acquire a frame of 3D data each time the laser is fired (Anderson, Herman et al. 2005). The detector concept resembles the FPA in a 2D digital camera, and the flash is similar to the flash of a camera. The potential significant advantage of flash is the speed of data collection. Instead of a single measurement, as is common in time of flight (ToF) scanners, a full frame is recorded. Currently, both power consumption and form factor prohibit the use of this technology in highly mobile devices.

The Kinect from Microsoft introduced a large audience to affordable range sensing. This device was followed by the Capri[14] from Primesense, which has a more favorable form factor. The camera interprets the 3D scene information from a continuously-projected infrared structured light source (Peng & Gupta, 2007). This 3D scanner system, referred to as 'light coding', employs a variant of image-based 3D reconstruction. Its main disadvantages are the form factor, fixed range and the sensibility in outdoor environments: sunlight contains infrared light.

Major advantages of active sensors are the relative insensitivity to various lighting conditions and the constant (predictable) quality of the data. The pitfall is the form factor of the equipment, power consumption and the locked range. The mobility of a current generation laser scanner still requires a tripod. In conclusion, although the active range sensing devices are promising and provide considerable advantages over passive devices, the form factor is currently not favorable (req. [16]). In the near future, they may become the dominant method for acquisition; the Kinect One[15] from Microsoft and the DS311 from Soft Kinect[16] are on a promising track.

4.2.3. INFERRED RANGE SENSING

In 1993, researchers from Carnegie Mellon University and University of Pennsylvania compiled a report for DARPA on computational sensors (Kanade &

[14] http://www.primesense.com/news/primesense-unveils-capri/. Last visited June 2015
[15] http://www.xbox.com/en-US/xbox-one/innovation , Last visited July 2017
[16] http://www.softkinetic.com/, Last visited, November 2017

Bajcsy, 1993), which is in line with what today is known as computational photography. As discussed by Marc Levoy from Stanford University (Levoy, 2010), *"the principles have been there for a long time but due to corporate secrecy, hardware-vs.-software, conservatism and still some research gaps the field did not take of yet!"*

As explained in Figure 24, at least 2 locations are necessary to obtain 3D images. CCD and complementary metal-oxide semiconductor (CMOS) image sensors are two different technologies for capturing images. Both types of imagers convert light into an electric charge and process it into electronic signals; this effectively provides a 2D raster of RGB values. They are extremely cheap, power friendly and small. Over the past 10 years, the resolution has increased considerably. It is relatively normal for smartphones to have high definition movie capture capabilities. This type of sensor can be used to capture more than just the visible light i.e. imaging in the Short Wave Infrared (SWIR) provides unique capabilities such as insensibility to dust or fog (Battaglia, Brubaker, Ettenberg, & Malchow, 2007).

Angle sensitive pixels in CMOS for Lensless 3D Imaging (Wang, Gill et al. 2009). This angle-sensitive pixel uses local, stacked diffraction gratings over a photodiode to discriminate the incident angle of incoming light. This type of imaging sensor provides a fundamentally richer description of the light it detects compared to a standard CMOS imager, while maintaining small size and robustness. However, they are not yet commercially available.

The light field, first described in a paper authored by Arun Gershun in 1936, is defined as radiance as a function of position and direction in regions of space free of occluders. In free space, the light field is a 4D function - scalar or vector, depending on the exact definition employed. The technology has high potential because it does not change the form factor and provides much more information to work with, as is extensively discussed in Ng's dissertation, entitled "Digital Light Field Photography" (Ng, 2006). Marc Levoy discusses 3D reconstruction aided by light field in a specimen from a single photographic exposure (Levoy, 2010). The technology is not yet commercially available, but is expected to be within a few years. However, MIT has a solution that enables current cameras to be converted into light field cameras (Marwah, Wetzstein, Bando, & Raskar, 2013)

Stevens and Harvey (2002) researched an optical system for reproducing three-dimensional images using the principle of integral photography. The key components are multiple arrays of lenses which relay, invert and encode a range

of views of the objects. Although, to our knowledge, the system is not commercially available, the potential is evident. Nano technology can be used for mass production when needed.

The system features a lens configuration resembling an insect's compound eye, that transmits several smaller images to the camera. The result is a photograph with multiple sub-views, each taken from a slightly different vantage point at exactly the same time (Shankland, 2007). Computational power can then be used to derive a 3D model of the scene. This therefore also falls in the computational photography domain.

The big advantage of computational range sensing is the favorable form factor (requirement 16). This can be micro scale, which is important for a lightweight acquisition. Furthermore, as this field emerges, resilience against bad environment lighting conditions and the quality/accuracy of 3D reconstruction is also steadily improving. The disadvantages are the need for multiple setups to obtain accurate results and the reliance on environmental light conditions. Because of the numerous opportunities in passive acquisition and the potential this has to match the capability to create maps of pristine environments, but with less weight, this thesis will focus on inferred sensing technologies. The bandwidth within this thesis is insufficient to explore both technologies thoroughly. However, when relying on inferred sensing, the acquisition quality will be largely dependent on knowing the position of the devices in space.

4.2.4. POSITIONING

When reconstructing more than the point-of-view from a measurement device, additional measurement positions must be known in space. There are many methods to obtain these positions, sometimes referred to as odometry. Especially when the weakness of one technology has to be patched by another technology is it important to know how and when that is possible.

If there is no relation to an overall coordinate system but the positioning is only known to the first known position at the start of the measurements, this is called relative positioning. Orientation is obtained from many cues, such as the earth's magnetic field, measuring the behavior of mass during movement or a spinning gyroscope (Caarls, Jonker, & Persa, 2003).

A magnetic compass contains a magnet that interacts with the earth's magnetic field and aligns itself to point to the magnetic poles. A compass provides a direction, but is not suitable for accurate measurements due to the many

disturbances in the magnetic field, both natural and man-made (Caarls et al., 2003).

An accelerometer is a device that measures acceleration. Accelerometers are associated with the phenomenon of weight experienced by a known mass that resides in the frame of reference of the accelerometer. Many devices such as inertia trackers use accelerometers, but they have limited refresh rates (Hz) and are prone to drift (Caarls et al., 2003).

A gyroscope is a device for maintaining or measuring orientation based on the principles of conservation of angular momentum. Currently, the most commonly used gyroscopes are based on microchip-packaged MEMS. Gyroscopes are used when magnetic compasses do not work and high precision is required. Most inertial navigation systems rely on gyroscopes (Caarls et al., 2003).

Ultrasonic tracking can also be used to position a sensor in space. An omni-directional, ultrasonic transmitter must be attached to the sensor to be tracked. This transmitter produces brief, periodic bursts of sound at frequencies above the range of human hearing. However, receivers have to be placed around the environment to triangulate the transmitter, which does not comply with requirement 15 regarding pristine environments (Caarls et al., 2003).

Vision-based pose estimation is a relative technology that will be thoroughly discussed in chapter 4.2.5. The recent advances made in this technology, the simple hardware and potential for 3D mapping make this a potential candidate.

As discussed in the requirements chapter, the pristine crime scene might be anywhere and, in most cases, a relative positioning technology is good enough. However, there are cases where the relative reconstructions must be tied to an absolute reference, such as large scenes and areas that have no line of sight overlap.

To have a common frame of reference between countries, researchers have agreed on a world coordinate system. If a position is known in relation to an agreed coordinate system it is generally known as absolute positioning. Earth observation technologies in particular rely heavily on this agreed frame of reference. Many objects in our everyday environment are known in relation to a local coordinate system. There are many ways to find a position in relation to a local coordinate system (Evans, 1998).

A Global Positioning System (GPS) is a space-based global navigation satellite system that provides location and time information on any location where there is an unobstructed line of sight to four or more GPS satellites. Many devices today have a GPS receiver inbuilt. There are two big disadvantages to GPS positioning; inaccuracy and line-of-sight restrictions. Current accuracy for everyday users is in the order of meters and areas without enough satellites visible to the receiver cannot find a pose (Loomis, Golledge, & Klatzky, 1993).

Next to a GPS system there are earthbound networks that facilitate in absolute positioning. There are geodetic reference grids (known xyz coordinates) generally with specific nails on the ground or passive landmarks with known positions, such as church tower tops and antennas. These networks are regularly maintained to guarantee high accuracy. Having such a recognizable object or marker in view provides absolute positioning (Evans, 1998).

Another much-used method to get absolute positioning is the use of wireless location systems such as ground-based radio frequency systems (loran), Wifi networks, cell phone antenna triangulation, etc. This is a relevant technology especially in cities with a large coverage of these networks (Evennou & Marx, 2006).

4.2.5. VISION BASED POSE ESTIMATION

The requirements (1,5,16) narrow down the type of range sensing systems that are acceptable; the pose of our equipment in relation to our environment must be known in metrics to overlay the digital imagery correctly and augment in the correct scale. Preparing the environment to support the pose estimation is not an option (pristine environment (16)), and there are no constraints on indoor or outdoor conditions.

These constraints force us to look at solutions that provide absolute positioning and rule out flucidal solutions. The system from Caarls et al.(2009) relies primarily on markers and 6 degrees of freedom inertial navigation, which we unfortunately cannot use in our design. The requirement of a pristine environment compels us to find a solution using the information present at the scene.

The system must rely on what a pristine environment can emit or reflect for finding its pose. Visual odometry or SLAM (Durrant-Whyte & Bailey, 2006) first extracts unique features from a scene, then matches the features and derives 3D information from corresponding features. This is reliant on what the environment emits or reflects.

There is a large body of literature on natural features, effectively disguisable contrast rich patches in an image that are re-detectable between frames (Lowe, 2004). To better understand features, we first distinguish between *feature point detection,* in which the chief task is to select suitable salient points in an image, and *description* in which the task of robustly transforming a small image patch around a feature point into a vector representation suitable for further processing is key.

The goal of the detection process is to find key points in an image that can be robustly detected against light changes, distributed over the image but still distinctive and efficiently identifiable. Early detectors were Harris corners (Harris & Stephens, 1988) and Features from Accelerated Segment Test (FAST), which was recently improved and renamed (FASTER) (Rosten, Porter, & Drummond, 2010). The big advantages of these detectors are the speed of detection and the quantity in which they can be found; however, they are not combined with a descriptor nor scale invariant. Currently, the more commonly used detection methods are in SIFT (Lowe, 2004), SURF (Bay, Ess, Tuytelaars, & Gool, 2008) and, more recently, DAISY (Tola, Lepetit, & Fua, 2010). The most important aspects of the modern detection methods are their reliability and, especially of the newer incarnations, speed.

Features must be matched against other features from different images and therefore need a description that can be used in a comparison process. Especially in SIFT, much effort was put into defining a description that would make it very robust against false matching. However, with the arrival of large scale photogrammetric reconstructions (Strecha, Bronstein, Bronstein, & Fua, 2010) based on photo archives, the 64- and 128bit descriptors were simply too large in memory footprint to be efficiently used. A solution was found (LDAHash) in which the descriptors were trained on a database, making them much smaller and representing them as short binary strings, as proposed by Strecha (Strecha et al., 2010).

After clear descriptions have been defined, the matching process can take place. There are many methods for figuring out which features correspond, but most will use a variation of random sample consensus RANSAC (Fischler & Bolles, 1981). To speed up the matching process, the search space can be reduced based on pixel distance, regions, descriptors or the epipolar line. Once the corresponding matches between two or more images are known, the features can be used to compute the third dimension. For robustly obtaining the third dimension, both the intrinsic and the extrinsic camera parameters must be

known. The most common technology for creating 3D from 2D is called structure from motion (SFM) or simultaneous localization and mapping (SLAM). Feature points need to be tracked from one image to the next and the feature trajectories over time are then used to reconstruct their 3D positions and the camera's motion (Dellaert et al., 2000).

This literature survey shows that pristine environment mapping with SLAM is feasible within the constraints of the requirements. The ergonomic guidelines are not crossed (16), the pristine characteristics (5) are not a burden and fusion of multiple sources can be done (7).

4.2.6. RECONSTRUCTION REFINEMENT

A “sparse” map from the structure from motion pipeline is not descriptive enough to conduct 3D analyses like blood pattern analysis. Only the features that are relevant for the pose estimation represent the scene. They represent less than 1/1000 of the available pixels in the image. The maximum density of the map equals the number of pixels in the images. Structure from motion on an image collection does not provide scale by default; secondary sensors or inputs will need to provide scale. There are different approaches to obtain a per pixel spatial estimation. The dominant approaches are stereo and multi-view reconstruction. Single view reconstruction has not been considered, as currently no metric data can be obtained (Saxena, Sun, & Ng, 2007).

In stereo reconstruction, the disparity between two images, on a per pixel basis, is used to derive the 3D information. There are two input criteria; the pose of the two cameras in space and the camera model; and the intrinsic and extrinsic parameters (Durrant-Whyte & Bailey, 2006). A lot of spatial movement is necessary to obtain a robust map with a monocular setup. As it is difficult to know which pixel corresponds to which pixel in the other image, most methods rely on epipolar line searches to detect correspondence. Because there is a slight view angle difference between the images, there will be areas that are occluded or only exist in one of the two images. Extensive overviews are provided by various researchers in the field (Scharstein & Hirschmüller, 2007; Scharstein & Szeliski, 2002).

In multi-view reconstruction, there are N numbers of images that represent the scene. There are roughly two approaches to construct dense 3D information: stereo-based multi-view and true multi-view. When using stereo-based multi-view, the first phase is to find the best stereo pairs within the multi-view set and to reconstruct those, followed by fusion of the stereo images. In true multi-view,

per pixel comparison is done in all candidate images. There are many variations in multi-view photogrammetry that try to circumvent some of the challenges, such as patch-based multi-view (Furukawa & Ponce, 2009), inverse ray-tracing (Lui & Cooper, 2011) and the use of shading (Wu, Wilburn, Matsushita, & Theobalt, 2011).

Some supporting technologies, such as gyroscopes and accelerometers, improve the reconstruction, but are generally not metric (Caarls et al., 2003). Apart from the technologies mentioned in chapter 4.2., there are three areas that provide additional information: light, lens system and ccd/cmos. Arguably, flash light is a fourth, but in Figure 22 it would be positioned in the active branch.

In most photogrammetric methods, the lighting information is used to differentiate the pixel intensity while searching for correspondence. The shape of an artifact can be derived from the shading of the artifact (R. Zhang, Tsai, Cryer, & Sha, 1999). An integration with multi-view provides additional detail that is ignored by generic pixel matching. This has been researched by Wu, Wilburn et al. (2011).

A lens is used to focus the desired artifact; however, controlled defocus is an option too. Watanabe, Nayar et al.(1996) explain how focus/defocus can be harvested to reconstruct depth. Using this relatively old method, real-time reconstruction can be obtained.

Most camera sensors capture 8bits per pixel in the Red, Green and Blue channels. However, camera sensors that reach beyond these restrictions are becoming more mainstream, most notably the high dynamic range (HDR) cameras, which can commonly capture 16bits per color channel. By increasing the light sensitivity of the sensor, the capabilities in light and dark areas are boosted as well as the distinguishing properties of pixels, as demonstrated by Cui, Pagani et al. (2011).

High resolution 3D models can be derived from cameras alone, once the camera positions are known in space (Newcombe, Lovegrove, & Davison, 2011). The reconstructions in the papers in this chapter are promising. In the next chapter, the systems that use the described theory are discussed to understand the limitations of image based reconstructions.

4.2.7. Recent visual localization and mapping systems

The best known methods for estimating the pose of a vision system are simultaneous localization and mapping. Visual odometry or SLAM can be

compared with the SFM problem. In SFM, the goal is to determine, from a collection of images and up to an unrecoverable scale factor, the 3D structure of the environment and all the 6-D camera poses from where the images were captured. The differences between SFM and SLAM are not only in the methods but also in the objectives. That is, similar aspects of similar problems are given different priorities. On the other hand, in visual odometry, the robot's ego motion must be obtained from a sequence of images. This can be seen as a similar problem to stereovision SLAM, where features must be matched across two or more stereo pair of images. Simultaneous localization and mapping, as pioneered by Davison (Davison, Mayol, & Murray, 2003) has seen significant improvement over the years and can be considered mature (Durrant-Whyte & Bailey, 2006) The following systems do not require any pre-knowledge of the scene and hence are suitable for pristine environments.

A System for Large-Scale Mapping in Constant-Time using Stereo

Mei et al. (2010) describe a relative simultaneous localization and mapping system for the constant-time estimation of structure and motion using a binocular stereo camera system as the sole sensor. Achieving robustness in the presence of difficult and changing lighting conditions and rapid motion requires careful engineering of the visual processing, and Mei et al. (2010) describe a number of innovations which they show lead to high accuracy and robustness. To achieve real-time performance without placing severe limits on the size of the map that can be built, they use a topometric representation in terms of a sequence of relative locations. When combined with fast and reliable loop closing, they mitigate the drift to produce highly accurate global position estimates without any global minimization. They evaluate their system on long sequences processed at a constant 30-45 Hz in which they obtain precisions down to a few meters over distances of a few kilometers.

Parallel Tracking and Mapping

Parallel tracking and mapping is a camera tracking system for augmented reality. In their paper, Klein & Murray (Klein & Murray, 2007) present a novel method for estimating camera pose in an unknown scene. While this has previously been attempted by adapting SLAM algorithms developed for robotic exploration, Klein & Murray propose a system specifically designed to track a hand-held camera in a small augmented reality workspace. They propose to split tracking and mapping into two separate tasks, processed in parallel threads on a dual-core, with one thread dealing with the task of robustly tracking erratic hand-held motion, while the other produces a 3D map of point features from previously

observed video frames. This allows the use of computationally expensive batch optimization techniques not usually associated with real-time operation. The result is a system that produces detailed maps with thousands of landmarks which can be tracked at frame-rate, with an accuracy and robustness rivalling that of state-of-the-art model-based systems.

Dense Tracking and Mapping in Real-Time

Newcombe et. al. (2011) developed a real-time dense camera tracking and mapping system, which relies not on feature extraction but on dense, every pixel methods. As a single hand-held RGB camera flies over a static scene, Newcombe et. al. estimate detailed textured depth maps at selected key frames to produce a surface patchwork with millions of vertices. They use the hundreds of images available in a video stream to improve the quality of a simple photometric data term, and minimize a global spatially regularized energy functional in a novel non-convex optimization framework. Interleaved, they track the camera's 6DOF motion precisely by frame-rate whole image alignment against the entire dense model. The algorithms are highly parallelizable throughout, achieving real-time performance using current commodity GPU hardware. They demonstrate that a dense model permits superior tracking performance under rapid motion compared to a state-of-the-art method using only features; and show the additional usefulness of the dense model for real-time scene interaction in a physics-enhanced augmented reality application.

Simultaneous Range and Color Tracking

Whelan et al. (Whelan et al., 2012) developed an algorithm that permits dense mesh-based mapping of extended scale environments in real-time. This is achieved by (a) altering algorithms such that the region of space being mapped by the pure range algorithm can vary dynamically, (b) extracting a dense point cloud from the regions and (c), by incrementally adding the resulting points to a triangular mesh representation of the environment. The system is implemented as a set of hierarchical multi-threaded components which can operate in real-time. The architecture facilitates the creation and integration of new modules with minimal impact on the performance on the dense volume tracking and surface reconstruction modules. Whelan et al. show trade-offs between the reduced drift of the visual odometry approach and the higher local mesh quality of the iterative closest point (ICP) -based approach.

4.2.8. SUMMARY

In chapter 4.2.1, range sensing was introduced and relevant technologies were discussed. In chapter 4.2.2, active versus passive sensing was explained followed by, in chapter 4.2.5, the technologies that are necessary to acquire a full 3D model. Chapter 4.2.7 closed with an overview of recent research on visual-based 3D mapping technologies.

It is evident that mapping is an active research domain and that current state-of-the-art technologies promise to be a viable solution for our pristine environment requirement. Research in SLAM and SFM have reached a maturity level that allow these to be viable alternatives in certain conditions to time-of-flight based acquisition methods.

It is important to keep in mind that monocular SLAM has some characteristics that need to be patched:

- Structure from motion on an image collection does not provide scale; secondary sensors or inputs will need to provide scale.
- A lot of spatial movement is necessary to get a robust map with a monocular setup.
- The technique relies on light being detected by the sensor.

Workflows heavily rely on tripod-based measurements, as described in chapter 2.2.1. These contaminate the crime scene and require considerable post processing. This background chapter shows that it should be possible to create a 3D map with mobile sensors without the need for the acquisition steps discussed in chapter 2.2.1.

4.3. COLLABORATIVE VIRTUAL REALITY

Once a 3D map of our environment exists, it is foundational for interaction between participants. The nature of information in pristine environments is three dimensional and virtual reality is the domain associated with this type of information. Virtual reality is an all-encompassing term; the following chapters will describe the sub-domains in virtual reality that are of interest to the architecture discussed in chapter 3.

4.3.1. VIRTUAL REALITY ENGINES

Virtual reality is a term that applies to computer-simulated environments that can simulate physical presence in places in the real world or/and in imaginary worlds. Most virtual reality environments are primarily visual experiences,

displayed either on a computer screen or using a special display (P. Milgram, 2006). A virtual reality setup might include additional sensory information, such as sound through speakers or headphones. Furthermore, virtual reality covers remote communication environments which provide for the virtual presence of users through the concepts of tele-presence and tele-existence, using standard input devices such as a keyboard and mouse, or through multimodal devices such as trackers and haptic devices. The simulated environment can resemble the real world to create a lifelike experience—for example, in simulations for pilots or combat training—or it can differ significantly from reality, such as in virtual reality games. In practice, it is currently relatively easy to create a high-fidelity virtual reality experience, largely due to technical advances in processing power, image resolution, and communication bandwidth. Game engines and virtual realities are similar (Fumarola & Poelman, 2011) and can be used interchangeably in our case; they share many components, such as a render engine, scene graph, etc. (cf. Figure 25).

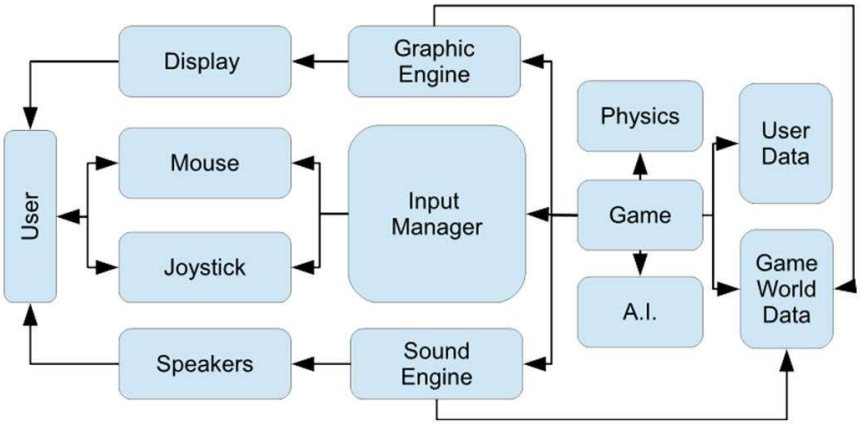


FIGURE 25 GENERIC GAME ENGINE ARCHITECTURE, ADAPTED FROM (GRINBLAT & PETERSON, 2012)

The leading open source game engines (Poelman & Fumarola, 2009) provide a software framework that developers use to create games for video game consoles, mobile devices and personal computers. The core functionality typically provided by a game engine includes a rendering engine for 2D or 3D graphics, a physics engine, collision detection and response, sound, scripting, animation, artificial intelligence, networking, streaming, memory management, threading, localization support, and a scene graph. The process of game

development is often economized, in large part, by reusing/adapting the same game engine to create different games, or to make it easier to "port" games to multiple platforms. Fortunately, there are many free to use and open source game engines available, as described in a selection method paper by the author (Poelman & Fumarola, 2009). As explained in the paper, there are a few important selection criteria:

- Compositing is an important quality in a game engine. To display reality capture data accurately, a correct camera must be definable. Correct overlay of a real-world scene with digital content requires a flexible virtual camera model. The human eye is complex and not easily fooled with misaligned data.
- Multiple people need to have access to the mediated reality space. The network capabilities need to be able to pass through scene interaction for multiple people, updates to the scene form the reality capture device and library objects shared between the participants.
- Overlaying a scene with digital content, while the system must also create a map and handle interaction requiring high frame rates, pushes approaches that can do different processes in a multi-threaded approach - much like what the parallel tracking and mapping algorithms are doing.

Nowadays, game engines effectively provide the componentized libraries that are needed to create the required artifact or at least most of the software infrastructure. The architectural components sketched in chapter 3 nicely map onto the basic game engine architecture. The design chapter must select the most appropriate engine and adapt it. More background knowledge on game engines is provided in Poelman & Fumarola (Poelman & Fumarola, 2009).

4.3.2. VISUALIZATION OF PRISTINE MAPS

3D reconstruction software can output various data structures. The raw format is generally described as a point cloud, and the interpretation of a point cloud is generally a mesh. Reality captured data is inarticulate; it is a raw stream of data that takes considerable space to store or to render. For an extensive overview of out-of-core-visualization, see (Silva, Chiang, Corrêa, El-sana, & Lindstrom, 2002). A brief overview of related technologies is provided, as game engines are not built for this data type.

State-of-the-art in the online rendering of point clouds is based on octrees. An arbitrary 3-D object can be represented to any specified resolution in a

hierarchical 8-ary tree structure or "octree" (Meager, 1982). A point cloud is encapsulated in a (mostly) squared volume and divided by eight equal volumes called leaves; this subdivision by eight (child leaves) continues until the point cloud is divided into manageable point sets or up to ??? voxels.

A description of an octree with manageable leaf nodes is provided by Wand, Berner et al. (2008). The octree structure allows them to display a subset of the data and to work with the full data because of the octree heritage. A percentage of the points in a leaf node are loaded, which represents the full set: a leaf node consists of 500k randomly stored points, evenly distributed; if 10% are loaded. this approximates the set.

Another approach to point cloud visualization is based on level of detail. The data is stored in different detail layers, comparable to the mip-mapping of textures. The camera position and image resolution determine the detail, which prevents loading too much data (Rusinkiewicz & Levoy, 2001).

The data structure is designed to get access to the right points in the dataset quickly; however, the points need to be visualized too. Common representations are: raw points (i.e. GL_POINTS), sprites and splats. In the case of voxels, the node size can be used to draw the correct screen size. The data structures are not only relevant for visualization, but also for the interpretation of the data, clash detection and surfacing.

It is also possible to surface the points at interactive speeds, as proposed by Erik Hubo (2007). A reconstruction pipeline can interpret the point cloud into a mesh representation. Meshes differ from point clouds because they contain additional topological information and are not based on a predetermined resolution. Like octrees for voxels, out of core meshes are possible, too, as Isenbrug & Gumhold show (Isenbrug & Gumhold, 2003).

Whatever the representation of the acquired map, there are many solutions to make these approachable out of core. The cost of out of core is generally time to prebake the necessary information for efficient retrieval (Kot, Chernikov, & Chrisochoides, 2006).

4.3.3. AUGMENTATION OF IMAGES

Overlaying digital assets are an important aspect of augmented reality. As defined by Ronald Azuma (Azuma, 1997), the first rule is "Combines real and virtual", which implies that at least some of the scene is virtual. In an ideal world, the digital assets are indistinguishable from reality. Discussing methods for

rendering generic geometric shapes is beyond the scope of this thesis. Game engines have the necessary components. However, overlaying digital assets onto imagery is not standard and therefore requires further literature research.

Geometric shapes and live camera feeds should preferably co-exist without noticeable difference to the human eye (Wann, Rushton, & Mon-Williams, 1994). However, this means that material properties and lighting conditions need to be known. Lighting conditions are generally captured by placing an object with known properties in a scene and extracting the reflections (Drora, Adelsonb, & Willskya, 2001). The material properties are more difficult to extract but promising research is ongoing (Lamond, Peers, Ghosh, & Debevec, 2009). Rendering photorealistic content in augmented space has been researched at Fraunhofer, where researchers were able to create seamless augmented pictures (Stricker et al., 2004), (sf. Figure 26).

FIGURE 26 AUGMENTED TABLE (LEFT), REAL TABLE (RIGHT), COURTESY OF (STRICKER ET AL., 2004)

The full pipeline for rendering augmented reality has also been described by Santos, Gierlinger et al.(2007), and a ray trace implementation was created by Scheer, Abert et al. (2007). Because of the pristine requirement of our scene, we cannot rely on pre-knowledge, which means our capturing device needs to have light condition capturing capabilities in cases where the picture needs to be perfectly matched.

Although some papers offer solutions for indistinguishable overlays that show promise, the computational cost and amount of additional research required to implement these seems problematic, but might not be necessary.

A relevant question that emerges when considering augmentation is how to overlay accurately. From the reconstruction pipeline, Section 4.2.5, we know that

camera models must be known, both intrinsically and extrinsically. This challenge has been addressed (Kolb, Mitchell, & Hanrahan, 1995), by using the actual lens setup, instead of the regular pinhole model for rendering. The accurate virtual placement of contents in a real scene either needs to be compensated on the render side or on the optics side.

4.3.4. COLLABORATION IN AUGMENTED REALITY ENVIRONMENTS

A considerable amount of research has been conducted on 3D augmented and virtual environments to support spatial decision making, including the interaction in virtual environments and virtual environments as support for collaboration (Bouras, Giannaka, Panagopoulos, & Tsiatsos, 2006) (Pekkola, 2002). These 3D virtual environments can be connected in networked virtual environments, making collaboration with remote users possible. It makes this kind of environment more interesting for projects where people must collaborate from different locations. These virtual environments are used to visualize information, such as providing a representation of the physical world regarding which users must make decisions. Users can have an integrated set of tools wherein collaboration, visualization and communication is fully supported. The collaboration we are focused on has a dominant spatial component. Rohrer suggested that *"Seeing is believing":* a 3D visualization is important for understanding a complex system which supports collaboration (Rohrer, 2000).

Furthermore, Kraut et. al. (1996) report on an empirical study of people using mobile collaborative systems to support maintenance tasks on a bicycle. Kraut`s results showed that field workers make repairs more quickly and accurately when they have a remote expert helping them. The pairs were connected by a shared video system, where the video camera focused on the active workspace, and they communicated with full duplex audio. Help was more proactive and coordination was less explicit when colleagues had video connections.

Follow-up studies on shared visual spaces by the same team (Gergle, Kraut, & Fussell, 2013) focused on another aspect of collocated collaboration: the use of shared visual information to provide communicative cues. Visual information impacts situation awareness by providing feedback about the state of a joint task and facilitates conversational grounding by providing a resource that pairs can use to communicate efficiently. Technologies to support remote collaboration can selectively disrupt people's ability to use visual information for situation awareness and grounding, and the extent of this disruption depends in part on task characteristics, such as the linguistic complexity of objects.

Many tasks require people to work together and there is great interest in using technology to improve the effectiveness of group activities, especially when they are co-located. Groups, unlike individuals working alone, communicate and exchange information (Agrawala et al., 1997): *"A whiteboard provides a single shared drawing surface that facilitates such collaborative interaction. Users can communicate by voice, gestures and by writing on the shared surface"*. A virtual space can be used to communicate in similar ways.

An example of a collaborative augmented and virtual reality system is shown in Figure 27, (Kuzuoka, 1992), also highlighted in chapter 4.1.

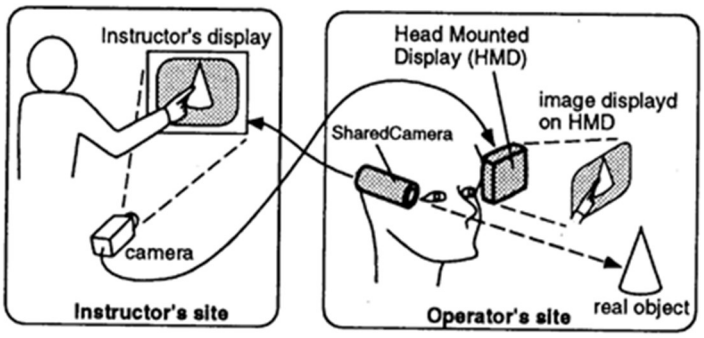


FIGURE 27 THE SHAREDVIEW SYSTEM LATER UPGRADED TO GESTURECAM (KUZUOKA, 1992)

The experiments with sharedview show that the system requirements necessary to support spatial workspace collaboration were the movability of a focal point, sharing focal points, movability of a shared workspace, and the ability to confirm viewing intentions and movements. The gestures of the instructor were visible on their head mounted device and the researchers claim that "having gestures significantly decreased the required number of verbal expressions, especially declarative expressions such as modifiers" (Kuzuoka, 1992).

Interesting for this research are the qualities from the GestureCam system (Kuzuoka, Kosuge, & Tanaka, 1994). As a basic function, GestureCam supports collaboration and offers the ability to acquire information from the real 3D environment and process the information using a computer. Also, the system can merge information provided by a computer with the real world. Moreover, not only a computer, but a human can also show information to be merged with the real world.

Furthermore, there are some soft constraints such as the fact that it is desirable for an instructor to have a free view, and that the virtual camera should be visible to the operator. At the time GestureCam was designed, the researchers used state-of-the-art technology, i.e. a laser pointer on an actuator, to visualize where and at what the instructor was pointing. Most of the technical limitations confronting the designers at that time have since been resolved, as shown in the previous sections.

More recently, another development in spatial workplace collaboration has been researched, namely the wearable active camera/laser (Kurata, Sakata, Kourogi, Kuzuoka, & Billinghurst, 2004). The main difference with GestureCam is that the HMD is replaced by a device worn on the shoulder. Interesting, the difference between both systems was marginal, except that the shoulder system was more comfortable to wear. The authors also claim this is largely due to the limitations of the used HMD. In the case of the HMD, less time was needed to talk during selections.

From a technology perspective, the work done by Adachi, Ogawa et al. (2005) is interesting, because they use live video information to overlay a pre-recorded scene model to provide the user with a photorealistic model. According to these authors, the scene presence for a remote user is improved by providing video information that is mapped live on a 3D model. The system enables an observer at a remote site to independently investigate a colleague's activity through the virtual environment.

In the visual co-presence research of Kraut, Miller et al. (1996), the emphasis is on remote collaboration aided with technological means. They conclude that a remote expert significantly enhances task performance (Kraut et al., 1996), i.e., "Workers were less explicit in describing the state of the physical world and what they had accomplished when they shared a view of the work environment with their collaborators" and "When they shared this view, experts were more likely to offer proactive instruction, basing the instruction they delivered and when they delivered it on a combination of the worker's explicit descriptions and their visual inspection of the worker's behavior" (Kraut et al., 1996). In follow up research, they are even more explicit about why visual grounding works: "Shared visual space is essential for collaborative helpers to determine: (1) worker's readiness to receive help, (2) the nature of the help the worker needs, and (3) worker's comprehension of new information" (Kraut, Siegel, Hanson, & Lerch, 2000).

The researchers Fussell, Setlock et al. (2003) considered the ways that participants use visual information to help coordinate their activities when performing collaborative physical tasks - tasks in which two or more individuals work together to perform actions on concrete objects in the three-dimensional world. A comparison between audio, side-by-side and an HMD with eye-trackers revealed that side-by-side clearly had the advantage, with audio coming in last. Interestingly, this research also showed that the use of eye-trackers did not contribute to a better understanding of the user's attention, although it theoretically should have. The authors claim that this might well be because of their implementation.

Collaboration can also take place in the same room, with an overlay of virtual reality (face-to-face). As stated by Kiyokawa, Billinghurst et al. (2003) an HMD that is used for collaboration in the real world should be naturally and clearly visible, which is important for face-to-face conversation. This is confirmed by Nilsson, Johansson et al. (2009), who researched co-located collaborative augmented reality suite for military strategy planning.

However, linguists and psychologists have observed that in reality, meaning is often negotiated or jointly constructed (Clark, 1996). Although providing the same view of a situation to two or more people is a good starting point for a shared understanding, things like professional and cultural background, as well as expectations formed by beliefs about the current situation, clearly shape the individual interpretation of a situation (Clark, 1996).

The maintaining of common ground is an ongoing process, which demands both attention and coordination between the participants (Nilsson et al., 2009). The system not only allows augmentation of the individual users view, but it allows each user to affect and change their team members' view of the ongoing situation, which is fundamental to the definition of a collaborative augmented reality system (Nilsson et al., 2009).

4.3.5. SUMMARY

Collaborative augmented reality is certainly not new; there are many predecessor systems that are in line with what we are trying to achieve. (Billinghurst & Kato, 1999; Butz, Hollerer, Feiner, MacIntyre, & Beshers, 1999; Kritzenberger, Winkler, & Herczeg, 2002; Nilsson et al., 2009; Pan, Zhigeng, Yang, Zhu, & Shi, 2006; Rekimoto, 1996; Rolland, Biocca, Hamza-Lub, Ha, & Martins, 2005; Vosinakis et al., 2008; Winkler, Kritzenberger, & Herczeg, 2003).

This chapter elaborated on the aspects of collaborative augmented reality that are unique to what needs to be researched. Co-located collaboration in virtual spaces has been successfully conducted by multiple research teams and the findings share many similarities. Because colleagues are not able to see each other the technological means should facilitate in maximizing the bandwidth of communication. Knowing what the other person is looking at, pointing at the topic of discussion and the ability to talk freely are successful practices.

4.4. DISPLAY HARDWARE

There are many devices that are capable of blending virtual and real environments (Benford, Greenhalgh, Reynard, Brown, & Koleva, 1998). However, as discussed in chapter 4.1, the systems that could successfully support 3D interaction and collaboration used an HMD as display hardware. Furthermore, as Daniel Wagner (2007), who researched the use of smart phones over HMD's, concluded: *"Backpack setups with HMDs have the advantage of providing high processing power and immersion, while HMDs have clear advantages in application areas that require stereoscopic augmentations or hands-free interaction."* An HMD covers a maximal area of human sight (Kern & Riedel, 1996) and allows for free-hand movement, while depth perspective/perception can be obtained to support spatial related tasks (Bimber & Raskar, 2005b), as formulated in requirement 18.

The tasks to be performed in mediated augmented space are physical, i.e. blood pattern analysis, line of sight, etc. In recent research by the author (Poelman et al., 2010), different display types were tested during 3D manipulation tasks. Most of the participants preferred the head mounted device for human computer interaction to holographic and desktop displays.

There are two flavors of HMD available: image blending in a display or in optics. Both have advantages and disadvantages. A see-through HMD is a device used to combine optical real and virtual-world views. The next chapters compare the methods and explain the trade-offs.

4.4.1. OPTICAL SEE-THROUGH

The notion of optical see-through displays started with that of Ivan Sutherland (I. Sutherland, 1968), and even today, most optical designs combine computer generated imagery with the real world using a beam splitter. In contrast to closed-view HMDs, which do not allow any direct view of the real world, a see-through HMD lets the user see the real world, with virtual objects superimposed by optical or video technologies (Azuma, 1997).

Optical see-through HMDs work by placing “combiners” in front of the user's eyes. These combiners are partially transmissive, so that the user can look directly through them to see the real world. Sometimes they are also partially reflective, so that the user can see virtual images bounced off the combiners from head mounted displays (Azuma, 1997). Most existing optical see-through HMDs act like a pair of sunglasses (cf. Figure 28) when the power is cut-off, because the combiners reduce the amount of light from the real world.

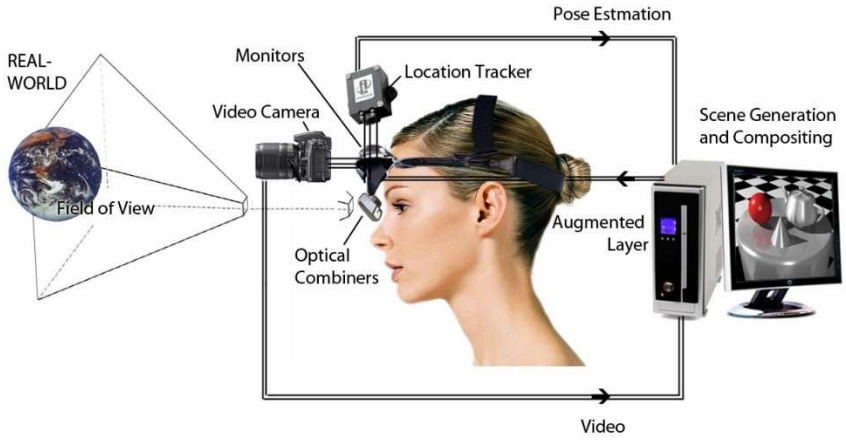


FIGURE 28 A CONCEPTUAL DIAGRAM OF AN OPTICAL SEE-THROUGH HMD

FIGURE 29 EXAMPLES OF SEE-THROUGH HMDS (LEFT) VUZIX RAPTYR, (RIGHT) SAABTECH ADDVISOR 150

An optical see-through HMD offers advantages over a video see-through. The real world is viewed “as is”, only slightly darkened; the virtual object is constrained by the video output and the eye position does not suffer from an offset. Furthermore, the video images must be distorted to fit the eyes, not the other way around.

As explained in the research by Caarls and Jonker (2003), the calibration that is necessary per user and session is daunting. A slight shift of the HMD misaligns all the data, initiating a new calibration procedure. Other disadvantages are that the virtual objects are transparent, there is no occlusion and the imagery is not grounded. Although most see-through devices are still grotesque, new advances have made them much lighter; see Figure 29.

4.4.2. VIDEO SEE-THROUGH

Video see-through HMDs work by combining a closed-view HMD with one or two head-mounted video cameras. The video cameras provide the user's view of the real world. Video streams from the cameras are combined with the graphic images created by the scene generator, blending the real and virtual.

The result is sent to the monitors in front of the user's eyes in the closed-view HMD (Azuma, 1997). A video see-through HMD (cf Figure 30) offers advantages over an optical see-through HMD. The temporal mismatches between the real world and the virtual images can be reduced. The image quality of the real-world images and the virtual images can be matched to one another. Occlusion between real objects and virtual objects can be expressed correctly. The images captured from the real world can be used in additional registration methods. The lag with the real world is not noticeable because the video steam is the only reference.

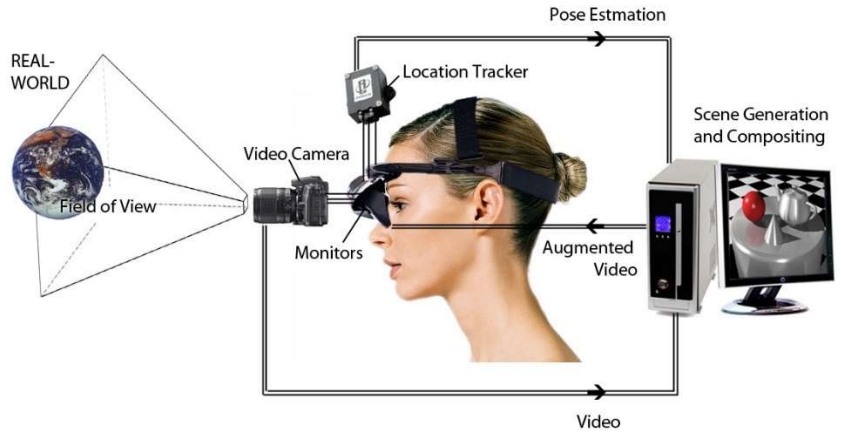


FIGURE 30 A VIDEO SEE-THROUGH HMD CONCEPTUAL DIAGRAM

4.4.3. VIRTUAL RETINAL DISPLAYS

Rather than looking at a screen through a magnifier or optical relay system as in previous HMD method descriptions, the viewer of the virtual retinal display (VRD) has a scanned beam of light enter the pupil of the eye and focused to a spot on the retina, Figure 31.

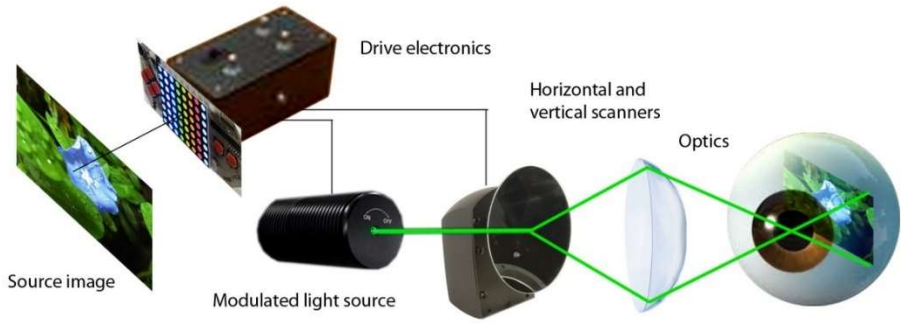


FIGURE 31 VIRTUAL RETINAL DISPLAY SCHEME

The VRD has several advantages over CRTs, LCD, and other addressable-screen displays, (Kollin & Tidwell, 1995):

- Resolution is limited by beam diffraction and optical aberrations, not by the size of an addressable pixel in a matrix. Very high resolution images are therefore possible without extensive advances in micro-fabrication technology. Also, the VRD does not suffer from pixel defects.

- The display can be made as bright as desired simply by controlling the intensity of the scanned beam. This makes it much easier to use the display in "see-though" configuration on a bright day. Because the light is projected into the eye and the scanner is electro-mechanically efficient, the display uses very little power.
- In theory, the VRD allows for accommodation to be modulated pixel by pixel as the image is being scanned.

FIGURE 32 MICROVISION PICO PROJECTOR SHOWWX[17], COURTESY OF MICROVISION

Typical examples of the required components can be found in, for example, Pico projectors; see Figure 32. Unfortunately, this technology for augmentation was not commercially available for research in the thesis time frame. The current units are relatively heavy and have an extremely limited field of view, especially when the user moves his or her eye around. The projector does not move to compensate for this behavior.

4.4.4. ANTHROPOMETRY FOR A HEAD MOUNTED DISPLAY

The optics in front of our eyes should take the ergonomic aspects of the human head into account. Multiple aspects must be considered in deciding on an HMD: center of gravity, maximal weight, overall dimensions and the effects of long time use.

The center of gravity of the HMD should best be placed on the same line as the center of gravity of the head to circumvent unnecessary momentum (Yoganandan, Pintar, Zhang, & Baisden, 2009). The center of gravity of the head is the point that crosses the midsagittal plane (+- 0,3 cm), a plane that floats 2,2-

[17] http://www.microvision.com/product-support/showwx/, last visited July 2017

4,3 cm above the Frankfort plane (Figure 33) and a line that lies 0,2-1,3 cm in front of an axis connecting the external auditory meati (ear openings).

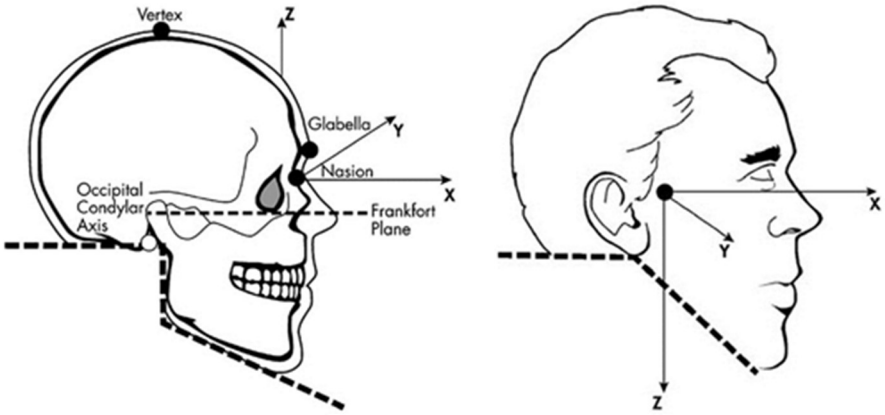


FIGURE 33 ANATOMICAL KEY AREAS AND SCHEMATIC HEAD CENTER (YOGANANDAN ET AL., 2009)

When wearing an HMD, the load will probably be on the front of the head. Knight and Baber (2004) found that a frontal head load of > 500 gram resulted in a significant increase in neck muscle activity, culminating in an increased level of perceived pain and discomfort in the head and neck. Counterbalancing that with weight on the back of the head failed to provide more comfort, which meant that increasing the total amount of weight for balance is not the answer.

When designing an HMD, among other things, the interpupillary distance needs to be known, as varying results have been reported, depending on age, race and measurement method. An overview study conducted by Dodgson (2004) showed that the vast majority of adults fit within a range of 50–75 mm. Further important dimensions are the nose bridge width and the arm length for ear support. Glasses manufacturers catalogues report nose bridge width dimensions from 10-30 mm and arm lengths of 115 to155 mm (cf Figure 34).

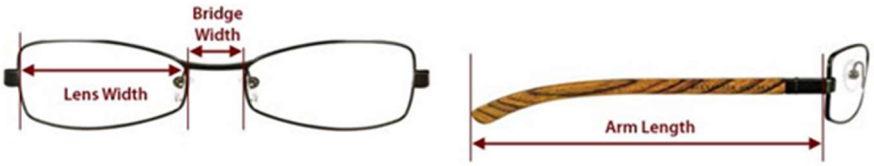


FIGURE 34 EXPRESSION OF COMMON GLASSES DIMENSIONS

Researchers James F. Knight et al. (2005) assessed the wearability of head mounted displays and subsequently defined a framework that detailed the different aspects of wearability. Important to note: several participants adopted a posture with the left arm raised to hold the HMD (650 gram) after 8-16 minutes, indicating that this was still too heavy.

The human vision system consists of more than the eyes alone; the brain and the vestibular apparatus (inner ear) are just as important. The human eye is an organ which reacts to light, as a conscious sense organ, and which allows vision (cf Figure 35). Rod and cone cells in the retina allow for conscious light perception and vision, including color differentiation. The rods are equipped for high sensitivity (low light, scoptic vision) and the cones have low sensitivity (day light vision, photopic vision). The human eye's non-image-forming photosensitive ganglion cells in the retina receive the light signals, which affect adjustment of the size of the pupil. The receptors are not evenly distributed along the retina and have a blind spot (Figure 35), which is patched by our brain. The diameter is about 24 mm among adults and the retina has a static contrast ratio of around 100:1, approximately 6½ f-stops, and a dynamic contrast ratio of about 1,000,000:1, approximately 20 f-stops (Barton & Byrne, 2007).

The pupil of the human eye can be likened to a camera lens aperture, where the iris is the diaphragm that serves as the aperture stop. Refraction in the cornea causes the effective aperture to differ slightly from the physical pupil diameter. The entrance pupil is typically about 4 mm in diameter, although it can range from 2 mm (f/8.3) in a bright light up to 8 mm (f/2.1) in the dark (Barton & Byrne, 2007). The angular resolution of the human eye is generally about 0.02°-0.03°, which corresponds to 30–60 cm at a 1 km distance and simultaneous visual perception in an area of about 160° × 175° (Wandell, 1995). For the brain to obtain a clear view of the world, the eyes must turn so that the image of the object of interest falls on the fovea. Eye movements are therefore very important for visual perception.

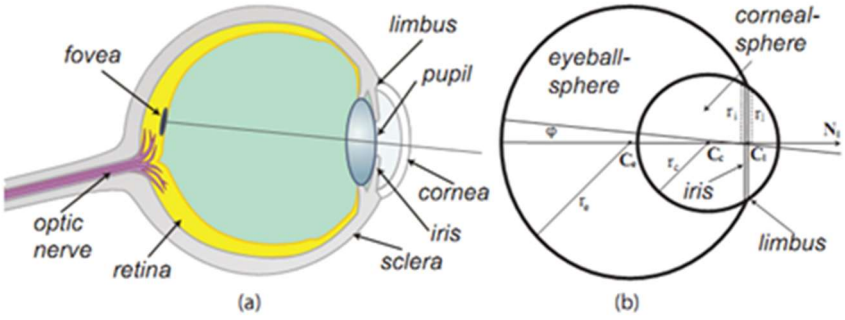


FIGURE 35 SCHEMATIC OF THE EYE (A) AND TWO QUASI SPHERES (B), COURTESY OF (SCHNIEDERS, FU, & WONG, 2010)

Although, the eye is of an approximate spherical shape, it cannot just be modelled by a single sphere but can be approximated by segments of two quasi spheres (Schnieders et al., 2010). This effectively means that an HMD with a wide field of view cannot just use the center of the eye to line up with the eye in its display curvature. This model fits well with Rolland, Ha et al. (2004), in which the eye is described in relation to an HMD (cf Figure 36)

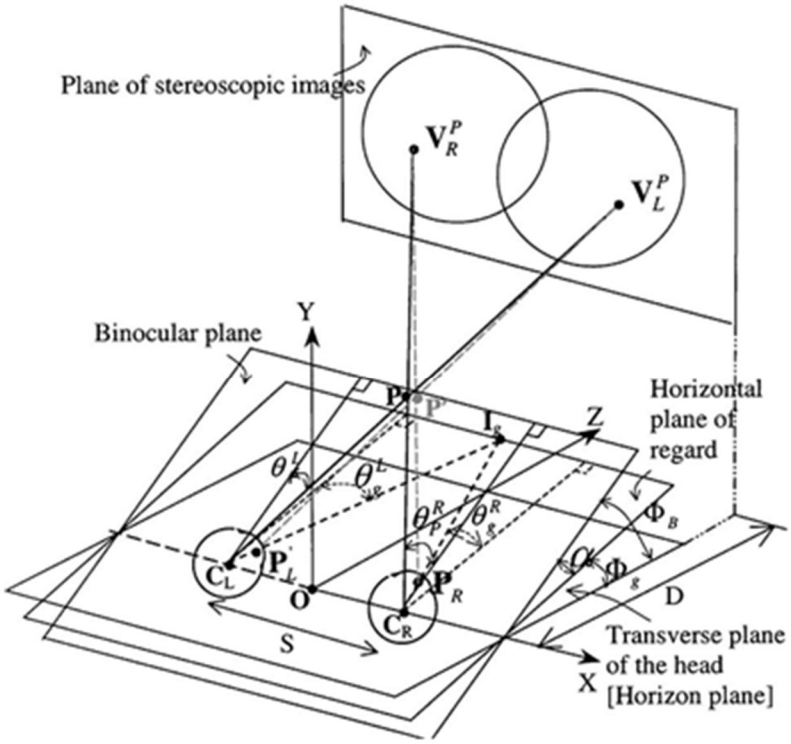


FIGURE 36 DEFINITION OF FRAME OF COORDINATES (ROLLAND ET AL., 2004)

The best visual acuity, of the human eye at its optical center (the fovea) is less than 1 arc minute per line pair[18], reducing rapidly away from the fovea. The human brain requires more than just a line pair to understand what the eye is imaging (Kopeika, 1998). Johnson's criteria define the number of line pairs of ocular resolution needed to recognize or identify an item (Johnson, 1958). Visual acuity is acuteness or clearness of vision, which is dependent on the sharpness of the retinal focus within the eye and the sensitivity of the interpretative faculty of the brain. With respect to HMDs, 60 pixels/° (1 arcmin/pixel) is usually referred to as eye limiting resolution, where typical HMDs offer 10-20 pixels/°. As explained in the above, the retina's surface has light-sensitive receptors, namely rods and cones. There are three kinds of cones, each "tuned" to absorb light from a portion of the spectrum of visible light. There are cones that are receptive to long-wavelength light (red), middle-wavelength light (green) and short-wavelength light (blue). The vestibular apparatus in the inner ear (gravity

[18] In degrees this means ° = MOA ÷60 thus 1 MOA = 0.0166666666667 °

organ) also plays an important role, as does the brain, that interprets the signals from the eyes. Together, this apparatus and the brain are responsible for the oculovestibular reflex: a reflex eye movement that stabilizes images on the retina during head movement by producing an eye movement in the direction opposite to head movement, and by doing so, preserving the image on the center of the visual field.

A saccade is a fast movement of an eye, or other part of a body or device. Our interest is mainly on the functions of the saccadic movements of the eyes. Saccades are quick, simultaneous movements of both eyes in the same direction. Our eyes locate interesting parts of a scene, building up a mental, three-dimensional 'map', corresponding to the scene. By moving around, by moving the eye, especially the fovea, smaller parts of a scene can be sensed with greater resolution. With an unexpected stimulus, it normally takes about 200 milliseconds to initiate and then lasts about 20–200 milliseconds. The main reasons for saccadic movement are that it provides additional resolution, makes a quick reaction to visual stimuli possible, avoids the blurring of an image and the anticipation of predictive happenings (Cassin & Solomon, 1990).

Depth is not only perceived by having binocular vision; our visual system has additional cues to perceive depth. Strong *pictorial depth* cues are occlusion and shadows, *kinetic depth* cues can provide depth information obtained by changing the viewpoint, *physiological depth* cues are convergence (fixate at a certain depth) and accommodation (changing the shape of the eye's lens) and *binocular disparity* provides depth cues by combining the two horizontally offset views of the scene (E. Kruijff, Swan, & Feiner, 2010).

A default rule is that the more the left and right image overlap, the better the depth perception. When the two eyes receive different stimulation on corresponding retinal areas, which precludes binocular fusion, a condition exists for creating a phenomenon known as binocular rivalry (Patterson, Winterbottom, Pierce, & Fox, 2007). Two images must be fused to provide a coherent image and binocular rivalry refers to a state of competition between the eyes. One eye inhibits the visual processing of the other eye. Binocular fusion can minimize the occurrence of binocular rivalry. At least 40% of the images must overlap to decrease the effect of binocular rivalry; furthermore, false contour lines between the monocular and binocular decrease the effect, too. Convergent design seems to have fewer side effects, but the periphery area is always more sensitive to some sort of suppression.

With respect to the human visual perception system, quite some unfavorable side effects are introduced by wearing an HMD. The origin is generally found in a mismatch of technology with regard to the human vision system, which can induce symptoms of motion sickness, including nausea, drowsiness, general discomfort, apathy, headache, disorientation and fatigue (Kennedy, Berbaum, & Lilienthal, 1992; Kern & Riedel, 1996). Most effects are the result of a-synchronous information delivery, a delay or incomplete data provided to our senses, because:

- Our vestibular system works closely together with our vision system and we only manipulate our vision, thereby ignoring our vestibular system;
- The refresh rate of displays is slow in comparison to human vision;
- The field of view is too narrow to understand the position of the rest of the body;
- Resolution and contrast provide too few details to understand our environment;
- Default reactions of the eyes (saccade) are not anticipated or reflected; the technology in which the focus is locked and convergence is ignored;
- In case of video see through, the cameras create an offset between our world perception and the recorded (spatial perception error).

Research conducted by Boger (2007) provided meaningful tips, such as that the acceptable weight for HMD is 225 grams and 100 degrees field of view is preferable (Boger, 2007).

4.4.5. SUMMARY

Both optical see-through and video see-through can be viable solutions. VRD's are not feasible because of availability/maturity issues. Recently, the number of currently available HMDs has increased considerably. The Oculus rift, Google glasses, Sony's HMZ and Silicon Micro Display`s ST1080 are just a few of the available options.

Considering the complexity of the human visual system, trade-offs will be necessary. Implementation becomes particularly difficult when pixel perfect alignment is necessary, as pointed out in the work of Rolland et al. (Rolland et al., 2004). There are many reasons that can cause unfavorable side effects to arise, hence minimizing the chance of negative side effects is important.

It was explained in the previous chapters what kind of ergonomic and anthropometric aspects are important for the design of an HMD. There is preferably 100% overlap between the images with a 1 arcmin/pixel resolution with 100 degrees' field of view. Furthermore, the device should not weigh more than 225 grams. With a biocular display, where both eyes are presented with identical images that are tilted or displaced, best practice is to allow the images to be fused at a quasi-planar surface at a specific image plane depth (Wann et al., 1994).

To overcome motion sickness, rendering latencies lower than 10 milliseconds are necessary (Aw et al., 1996) Conventional systems have a refresh rate of 50Hz which means 20 milliseconds to display a single frame. The time to render that frame will add to the latency. It is not possible to reach the required latency for augmented reality (<10 milliseconds) by sequentially rendering and displaying a frame.

4.5. HUMAN INTERACTION

In this chapter, we will highlight the user interaction that is suitable for wearable augmented reality systems. Req. [17] states that the tool should not interfere with standard procedures. From the interviews leading to the requirements, it clearly emerged that the participants preferred not to have to use physical interaction devices with their hands.

To perform a task within a 3D augmented environment, an interaction technique is required that translates the user's intentions captured by an input device into system actions, whereby the action should result in an output from the computer. Main categories can be defined that overlap with most tasks. These tasks are selection and manipulation, navigation, system control and symbolic input (Bowman, Kruijff, LaViola, & Poupyrev, 2005).

- Manipulation is probably the most fundamental task; it allows a user to adapt the content of an environment, thereby getting away from just being a passive observer.
- Navigation is less important, because in this case physical navigation is used; this generally refers to the tasks of moving through a virtual environment.
- System control methods are techniques to send commands to an application, to modify a parameter or to change a mode and they are inherently equivalent to using a widget in desktop environment.

- Symbolic input is generally the task of communicating symbolic information such as text or numbers to the system. However, in our system, we likely do not need this type of interaction. It is the "traditional" desktop task, normally bound to keyboard input.

The interaction is considered as a human input/output system and its relation to human computer interfaces is best described by the information processing diagram depicted in Figure 37.

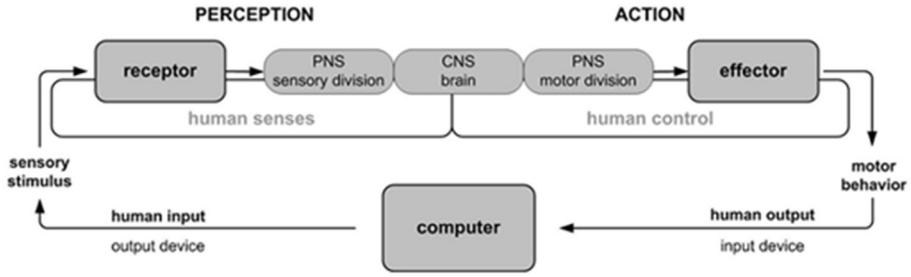


FIGURE 37 INFORMATION PROCESSING IN THE HUMAN I/O SYSTEM, OBTAINED FROM (E. KRUIJFF, 2006)

Since Sutherland (1968), there have been many attempts to create a wearable augmented reality system with 3D interfaces. The more recent systems will be briefly discussed, compared with the requirements and the effectiveness of the system considered.

4.5.1. INTERACTION MODALITIES

Users can interact with a system in different ways, from touch to speech and from muscle tension to eye blinking. A thorough overview is provided by Hahn (2010). The selection of relevant research is largely based on the applicability to the following scenarios:

- Audio communication between on location participant and remote expert;
- Interaction with tools to accomplish 3D tasks;
- The research scenes are native 3D; the manipulation tasks need an interaction technique that is suitable for full scale 3D;

An in-depth discussion of audio communication is unnecessary, as this is a very mature field and many tools - so-called Voice over internet protocol (VOIP) tools

- are commercially and freely available. This topic has been discussed many times at computer supported collaborative work conferences (CSCW) and in related books (Rutter, 1987) Relatively new are the recently developed bone-conduction headsets, which allow the operator to hear radio communications through bone-conduction technology without curtailing ambient sound. Standard audio headphones are useful in many applications, but they cover the ears of the listener and thus may impair the perception of ambient sounds. Bone-conduction headphones offer a possible alternative (Walker, Stanley, Iyer, Simpson, & Brungart, 2005). Giving speech-based commands works best when only the user's commands are accepted; 3D control is not mentioned as a solution for 3D manipulation (Quek et al., 2002). When multiple people are speaking, it is difficult to determine whether the captured audio signal is from the speaker or from other people. The recognition error is much larger when the speech is overlapped with other people's speech (Z. Zhang et al., 2004). Ambient sound is sometimes disturbing, but there are solutions. Active noise control is a technology that uses a noise-cancellation speaker, which emits a sound wave with the same amplitude but with inverted phase to the original sound. The waves combine to form a new wave, in a process called interference, and effectively cancel each other out (Elliott & Nelson, 1993). This technology cannot be used for 3D manipulation, but can be very effective for enabling/disabling the system.

The classic user interface is the WIMP interface, an acronym for windows, icons, menus and pointer. It was coined by Merzouga Wilbert in 1980 and expanded and referenced by Hinckley (Hinckley, 1996). Other expansions are sometimes used, substituting "mouse" and "mice" or "pull-down menu" and "pointing", for menus and pointer. This system is still the most used for desktop computers. This principle has largely been discarded for stereo HMDs, because selection in depth of a scene is not considered in this paradigm (Haan, 2006). However, the paradigm is still useable for remote clients that participate in the scene and for selecting tools.

Although voice and hand gestures are dominant in human-computer interaction research (Dix, Finlay, Abowd, & Beale, 2004), there are other technologies that are promising that can also be used for control. There are many things we can control on our body, and they can be leveraged.

- Electromyography or myoelectric signals can be used to control virtual objects (Takeuchi, Wada, Mukobaru, & Doi, 2007) . They are the signals that make the muscles contract. The technology is mostly used to control

prosthetics. However, this type of controlling requires sensors on the body and there is a difficulty in determining strength. Non-invasive receiving methods do not create accurate enough signals (Takeuchi et al., 2007).

- A brain–machine interface (BMI) uses neurophysiological signals from the brain to control external devices. (Kansaku, Hata, & Takano, 2010) The researchers show that the user's brain signals successfully control virtual agents. Although not yet used for precision tasks, the method might be used to browse a graphical user interface. In Kansaku, Hata et al. (2010), the user's thoughts became reality through the robot's eyes, enabling the augmentation of real environments outside the anatomy of the human body.
- According to Hua, Krishnaswamy et al. (2006) it is highly desirable to integrate eye-tracking capability into HMDs in various applications. Researchers (Rolland et al., 2004) quantified that an accurate representation of eyepoint can minimize angular and depth errors in high-precision displays.

Eye-movement can be used to operate a system but much like BMI's the accuracy and the repeatability is lacking.

4.5.2. GESTURES AS INTERACTION

"There is strong evidence that future human-computer interfaces will enable more natural, intuitive communication between people and all kinds of sensor-based devices, thus more closely resembling human-human communication" (Wachs, Kölsch, Stern, & Edan, 2011). They advocate vision-based hand gestures as one of the dominant technologies, and lay emphasis on responsiveness, user adaptability and feedback, learnability and accuracy. They also claim that no current system has been perfected yet, thus much is still to be gained in this field. One of the main problems is the lack of haptic feedback.

Wang and MacKenzie (2000) found that performance degraded significantly when there was no physical surface to touch when manipulating virtual objects with the hand. Thus, with free space hand gestures, other sensory replacements need to be researched to mitigate the effects of lost kinesthetic feedback.

Many researchers have classified gesture interactions (Quek et al., 2002; Wexelblat, 1998) and a clear categorization was created by Wachs (Wachs et al., 2011). The typical types that are needed in our project are deictic gestures. Most important to this research are the manipulative gestures, i.e., selecting tooling in

the interface and interacting with 3D objects in the scene. Quek et al. (2002) formulate manipulative gestures as follows: *"Gestures whose intended purpose is to control some entity by applying a tight relationship between the actual movement of the gesturing hand with the entity being manipulated".*

Different technologies can be distinguished and used to detect hand gestures, i.e., acoustic, mechanical, magnetic or visual tracking. The perceptual types are relevant since they enable gestures to be recognized without requiring any physical contact with the input device or with any physical object. The user can therefore communicate gestures without having to hold or make physical contact with an intermediate device. In the past, a custom marker on the hand plus a circular Hough transform to calculate the positions of the finger tips using a single camera have been used (Persa & Jonker, 2000). A set of custom fiducial markers on plain gloves (Piekarski & Thomas, 2002) tracked by a camera has also been used to control the AR system. The position of the markers relative to the camera can be calculated because the size of the marker is known. Although they report that the system is stable in outdoor conditions and that positioning is quite accurate, the downside is that additional hardware is needed.

Another AR interface is HandVu (Koelsch, Bane, Hoellerer, & Turk, 2006), in which the researchers explore "flock-of-features" tracking for hand recognition. While they claim the system is reasonably stable for outdoor tracking, when both the background and the foreground move, the center of the flock jumps to various positions (Poelman et al., 2010). A solid requirement match is the system developed by (T. Lee & Hollerer, 2007) named "Handy AR", in which a bare hand is used to interact as a 6 DOF device. The hand color is used to segment the region of interest and detail steps find the fingers, after which a hand coordinate system is established. Lee's (2007) algorithm for hand segmentation is among the best based on color. However, they report that their system is not stable in outdoor use. Furthermore, the system is not used for interaction but for view purposes.

Also related to this research is the system developed by Lee (M. Lee, Green, & Billinghurst, 2008) in which the user can move around virtual artifacts on a tabletop with bare hands. Collision of the object with the hand is a selection procedure, deselection occurs when the user's hand moves backward until the finger ray is far enough away not to touch the object. However, their implementation does not use the depth or marker position to obtain absolute positions. Another option is the work conducted by Heidemann et al. (Heidemann, Bax, & Bekel, 2004), in which interaction with both virtual data as well as physical data is combined in AR space. They use hand gestures to activate

menu items and to select artifacts on a tabletop. Their tasks do not require precision or absolute metrics, but the system works well in controlled indoor environments.

Computer vision hands pose estimation has shown improvements since the arrival of the Microsoft Kinect, because of the added depth. The following type of artifacts can be detected in a video stream: lines, scale (in) variant features and blobs. There are usually three types of information that can be used to extract the artifacts: optical flow, color and depth (Akman, Poelman, Caarls, & Jonker, 2013).

A technique that is used to extract hands based on features and that relies on color space is presented by Lee (T. Lee & Hollerer, 2008). This technique involves detecting whether the finger tips fit a plane by using the fingertips as a coordinate system. Because the system is pre-trained, the approximate distance is known (cf. Figure 38).

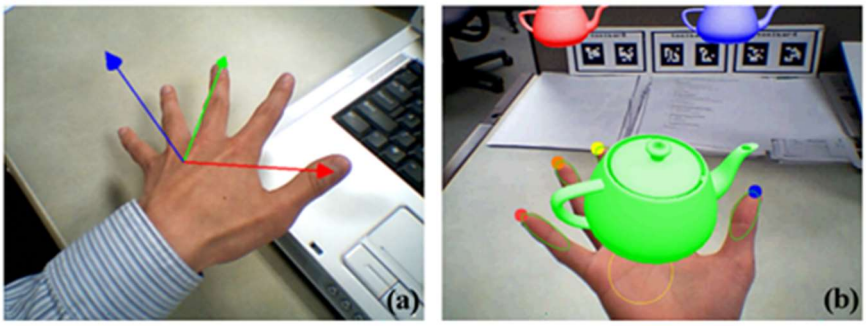


FIGURE 38 HAND COORDINATE SYSTEM, (B) SELECTING AND INSPECTING OBJECTS, OBTAINED FROM (T. LEE & HOLLERER, 2008)

Optical flow (Chik, 2006) can handle more complex hand gestures, which results in different segments of the hand moving at different speeds. By using optical flow, significant improvements in tracking accuracy have been observed, especially for gripping motions of the hand (Chik, 2006). Occlusion of two hands has in most cases been left out of consideration, with the exception of e.g. the research of Argyros and Lourakis (2004). They use a training database both online and offline and the trajectories of the hands to handle occlusion.

If more cameras are used, the accuracy of the hand detection increases. In the work of Schlattmann and Klein (2007) multiple cameras were used to construct a volumetric model of the hand, in which they were able to detect multiple poses

in 6 DOF with a high probability of success. Further to this work, the same authors (M. Schlattmann, Nakorn, & Klein, 2009) then showed that the proposed technique for grabbing and releasing enabled efficient manipulations and precise releasing.

Another system from Vogel et al. (2005) called AirTap is based on a click technique similar to how we move our index finger when clicking a mouse button or tapping a touch screen. Vogel's AirTap evaluation demonstrates the usability of relative hand-based pointing techniques with error rates in the same low range typically seen with status-quo devices like mice.

Of course, pre-knowledge can also be used to estimate the hand pose in 3D space such as been proposed by Stenger et al. (Stenger, Mendonc, & Cipolla, 2001). They built an anatomically accurate hand model from truncated quadrics from which they generated 2D profiles. The profiles were matched to the image profiles and filtered. It is fast and robust against self-occlusion because a full match is not necessary. An improvement on the use of pre-knowledge is from (Gorce, Paragios, & Fleet, 2008) in which shading and self-occlusion were taken into account to create stable outcomes.

Another hand tracking solution is the use of color (R. Wang & Popovic, 2009). If gloves are permitted, this is a viable solution. A single camera is used to track a hand wearing an ordinary cloth glove that is imprinted with a custom pattern that is designed to simplify the pose estimation problem. Performance, precision and robustness are unmatched by bare hand solutions, and the cost is very low.

4.5.3. Summary of Human Interaction

In brief, collaborative augmented reality systems facilitate an open audio channel; participants can talk to each other, which is a pre-requisite. As noted in the above, in co-located interaction, anything that allows for maximizing the bandwidth between users is generally beneficial. However, no evidence was found that audio could be successfully applied to 3D interaction, which is a requirement for the type of task investigators must conduct.

There are many papers on gesture-based interaction. Vogel's AirTap (Vogel & Balakrishnan, 2005), depicted in Figure 39, shows a natural interaction with 3D virtual content. The evaluation demonstrates the usability of relative hand-based pointing techniques with error rates in the same low range one typically sees with status-quo devices like mice.

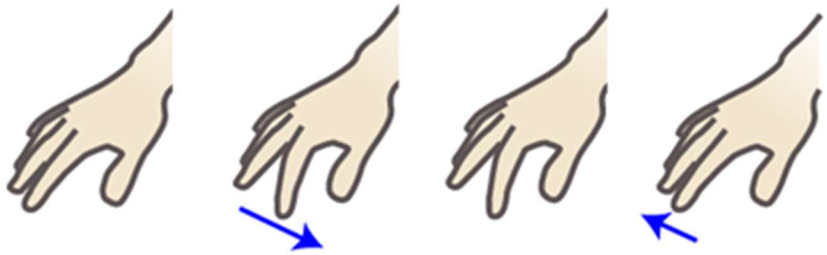


FIGURE 39 AIRTAB, IMAGE OBTAINED FROM (VOGEL & BALAKRISHNAN, 2005)

Furthermore the implementation of Schlattmann, Nakorn et al. (2009) is interesting. These researchers used bare tracked hand gestures to manipulate virtual objects in 6 DOF. The researchers showed that the proposed technique for grabbing and releasing enables efficient manipulations and precise releasing, even compared to a 3D mouse and traditional mouse. The more successful systems either use a stereo setup or a range camera to detect bare hands. It is especially important when using these systems to remove background noise and ruggedize against changing lighting conditions.

4.6. LITERATURE RESEARCH SUMMARY

As stated earlier, this research is multidisciplinary in nature, touching domains, such as computer-vision, human computer interaction, optical hardware and computer supported collaboration. The introduction to chapter 4 distinguished four areas that required additional background, provided in the present chapters, namely; mapping pristine environments, human Interaction, collaborative virtual reality and display hardware.

A rich set of comparable systems is discussed in chapter 4.1. While none of the systems share the same requirements, there are part-solutions that can be leveraged in the design chapter. Especially the real-time mapping for pristine environments is a new concept in augmentation systems.

Recent advancements in computer vision allow for high fidelity map creation suitable for mapping pristine environments. The frameworks necessary to create real-time maps are well documented and available.

Interactions with 3D environments that are real-world scale to match augmentation are sparsely researched. Occasional references can be found in the

literature as mentioned in chapter 4.5. However, additional research will be required.

From a collaboration perspective, various systems are available that allow an expert to help a novice while collocated. It is clear that most related research was conducted at a time when the technology was less sophisticated than today. Furthermore, the co-located collaborative environment was either a video stream or a CAD model, not a real-time map.

HMDs have significantly improved over the past few years: their weight, resolution, refresh rate and field of view all got better. Specifications show that the ergonomic constraints that used to create nausea are becoming less of an issue.

Furthermore, two additional requirements emerged from this background research, which needed to be added to the list in chapter 2.

Nr.	*Description*	HW	SW	INT	COL
18	*The system needs to use head mounted displays for digital overlay.*	*x*			
19	*The system senses with technologies that function in the visible light and infrared light.*	x			

TABLE 4 EXTENDED REQUIREMENTS

Lastly, the software frameworks used to create 3D games have the flexibility to be adapted for the purposes of this thesis. Unfortunately, not all the systems discussed are open source or they deviate too much from our purpose to be effective. It is beyond the scope of this thesis to come up with a new system from scratch.

5. Design of a Mediated Reality System

Previous chapters were devoted to discussion of the research questions, requirements and background. This chapter will be dedicated to the creation of the artifact. The goal of the artifact has now been defined; the people that are going to work with the system have been described and high-level workflows documented. The high-level architecture presented in chapter 3, and repeated in Figure 40 for convenience, shows systems that can be designed as subsystems until they fold back into the main system.

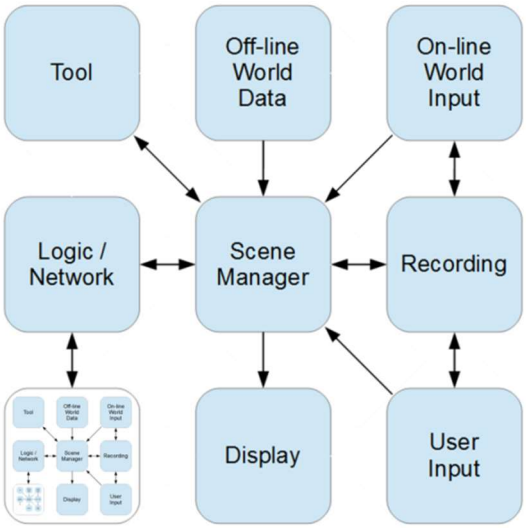


FIGURE 40 HIGH LEVEL ARCHITECTURE

We selected the INCOSE design approach (Walden & Roedler, 2015) to construct our artifact. INCOSE's methodology was developed with complex systems or systems of systems in mind. The methodology has been successfully used to develop complex military and medical systems (Walden & Roedler, 2015). The intended system has many systems and subsystems and consists of both hard- and software elements and therefore shares similarities with, among others, medical systems. In chapter 3, the requirements were translated into system elements to focus on background research. INCOSE's V-model will be used as the main approach for development; see Figure 41.

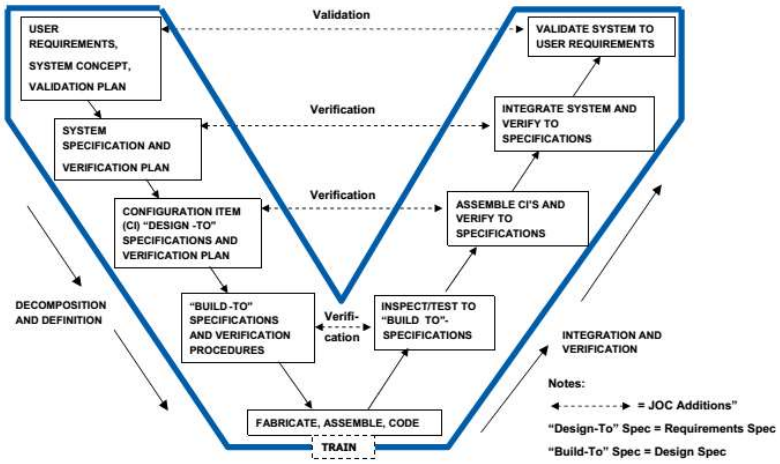


FIGURE 41 CLASSICAL V-MODEL, (FORSBERG, MOOZ, & COTTERMAN, 2005)

The V-model's first iterative step is to validate the requirements and concept with users. This verification is discussed before detailed approaches of methodologies are elaborated on.

5.1. A PROTOTYPE FOR VALIDATION

Based on the literature research from chapter 4 and the expert interviews from chapter 2.3, mock-ups were created that showed the individual subsystems functioning. The mock-ups and concepts were elaborated on at a target users conference[19] to acquire input from practicing crime scene investigators. The goal of the sessions was to acquire early feedback from potential users on the artifact. To that end, the practitioners were asked to evaluate the design of the mock-up, in the light of the fulfillment of basic requirements, the capabilities, usability, and work practice.

5.1.1. PROTOTYPE

Some of the proposed capabilities, such as augmented reality and hand tracking are not experienced by many people. The mock-ups therefore consisted of the four main subsystems working in isolation: mapping, interaction, collaboration and augmentation.

[19] 'Forensische Visualisatie - Visualisatie en Reconstructie van de Plaats Delict', 23-24 June 2010, Nijkerk, the Netherlands.

From our background chapters, we know that mapping a scene is possible with computer vision and that the requirements implicitly need a 3D Map. The mapping mock-up consisted of a webcam with a laptop that was running the parallel tracking and mapping module from Klein & Murray (Klein & Murray, 2007). Using PTAM, the capabilities of a module that could create a map of an arbitrary pristine environment could be demonstrated; see Figure 42 (A)

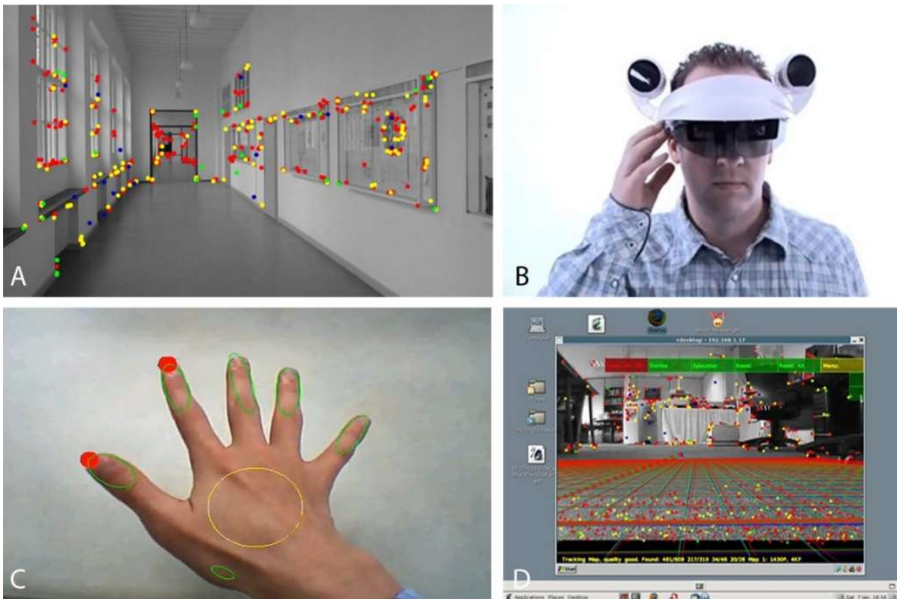


FIGURE 42 3 MOCK-UPS OF MAPPING (A), HEAD MOUNTED DEVICE (B), HAND TRACKING (C) AND COLLABORATION (D)

The literature shows us that interaction with an augmented reality system without additional equipment is possible. To demonstrate interaction in a virtual space, the Handy AR system from (T. Lee & Hollerer, 2007) is used. In the mock-up, the participants can see primitive 3D model pieces sticking to their hand; see Figure 42(C).

To demonstrate what collaboration looks like, Klein's tracking system (2007) was used. Two laptops were running; one for PTAM and the other for connecting a remote desktop into the running PTAM session. This is a cheap mock-up to show that it is possible to see a colleague making a map from a remote location connected to the internet; see Figure 42 (D).

Augmentation is part of the research question; it is therefore critical to show wearable augmented reality to the participants. To demonstrate augmented

reality, the system developed by Caarls and Pieter (Caarls & Pieter, 2009) was used; see Figure 42 (B). Caarls' system allowed participants to see what an augmented reality experience looks like. The optical see through system showed 3D floating objects; the tracking was not working.

Although the elements of the prototypes did not allow the participants to undergo a full experience, the imaginary step is not massive. Both pictures A and D from Figure 42 were taken in-situ.

5.1.2. PARTICIPANTS

The author hosted two 90 minute sessions with approximately 30 practicing crime scene investigators per session. The audience signed up voluntarily for the event, which focussed on forensic visualisation, with as main topics advances in (1) crime scene photography, (3) 3D laserscanning, (3) video analysis, (4) beyond visual traces (spectral, infrared), (5) crime scene reconstruction and (6) forensic visualisation.

As the events were mainly for educational purposes, the audience was composed of investigators interested in the latest advances in their domain. The events featured multiple tracks simultaneously, allowing participants to attend the topics of their interest.

It is understood that the audience represented the more progressive group in crime scene investigation and digital technologies. Although this biases the opinions, this group will likely be first adopters.

5.1.3. COLLECTION METHOD

The sessions were stuctured into three parts: (1) a presentation provided by the author on the innovations depicted in Figure 42, including an overview of the 'CSI The Hague' project, (2) hands on with the provided prototypes and (3) discussion, questions and aswers.

The author took notes during the sessions that were structured around the four elements of (1) mapping, (2) the head mounted device, (3) 3D interaction and (4) collaboration. Where relevant, the notes from the session were then used to buttress the general consensus. Hands-on participation was possible in front of the presentation room, but due to the fragility of the prototypes this was only feasible in a one-on-one setting with the author.

5.1.4. FEEDBACK

With respect to intended capabilities, fast 3D scene mapping was considered by both groups to be the most relevant application out of all the technologies presented, especially in the early phase of crime scene investigation when the crime scene is still untouched. During the orientation phase of the investigation, when as few investigators as possible enter the crime scene, it is helpful for investigators to have visual support when providing information to their colleagues. The temporal aspect of a degrading crime scene can therefore also be captured. *Quote: "Many investigators want a floorplan and photos as soon as possible. They use that to piece the scene together in their mind; many times they have to wait until the next day to get a floorplan which is used as a important communication means."*

Although both blood pattern analysis and bullet trajectory analysis were used as cases for 3D reconstruction, the fast mapping capabilities were considered great opportunities. When asked whether speed or quality were more important, fast communication of the data was clearly given priority over quality. A delay of minutes was acceptable with regard to quality. Although when asked about the quality of the data, the author admitted it would in all probability not be laserscan quality, the reponse was not negative. Using Klein's (2007) technology, a good reconstruction is unlikely, so care must be taken. This, however, did not scare off the audience.

When discussing the head mounted device, a few notable topics emerged that are worth mentioning, such as whether it is a problem to wear a head mounted device, how often and when would it be worn and whether information provided as an overlay would be useful. The discussion on whether it would be a problem to wear was quickly dismissed, as, when investigators enter crime scene, they already wear special suits and, in some cases, a helmet. *Quote:"This example looks goofy, but if it's either integrated in a regular helment or somewhat beefy glasses I`m fine.* And another: *For me it's just another tool I use on a crime scene, like any other tool I use it when required or when it's useful. Ruggedization, easy mount-dismount and comfort are important. This "helmet" does not scream comfort.*

When asked about weight, the general consensus was that it should be as light as possible, and if it would fit over regular glasses, that would be a plus. Most said they would not wear the HMD all the time. *Quote: "It's just like a mobile phone or other equipment, if needed we will use it."* In the prototype, the audience could see objects superimposed on the footage, which could also represent work from previous collegues. This stirred the crowd a little; yes, it would be great to have

an good overview of what has been done and of who did what, but it also meant an invasion, almost, of their privacy. Officers will not accept a tool that can be used to judge their behavior. *Quote: "A head camera is collecting dust although it proved successful in investigation (project NFI/Amsterdam) and just like a digital camera, officers will not be constantly walking around with the tool."* Too much information or information that is incorrect too often will also constitute a reason not to use a tool. *Quote: "GIS databases that were used to quickly find owners were out of date and therefore not used"* – in other words, what is in the scene needs to be relevant and cannot be incomplete.

The collaborative capabilities stirred a similar bifuricated discussion; on the one hand, discussion arose about the experts who would be able to remotely help and on the other, about the privacy of the participants. For this group, it was clearly a no-brainer that allowing experts to remotely access the crime scene would be incredibly beneficial. There are always too few experts and they need to travel a lot. It was clearly hard for the audience to imagine what such collaboration would look like. *Quote: "So I can imagine calling my collegue to get his expertise but he will be able to see what I see and even do measurements and such?"* On the privacy side, there was clearly some hesitation. Especially the fact that all their actions were traceable tended to be regarded wih some disfavour.

The least discussion was on 3D interaction. It was understood that a new paradigm was necessary for operating such a system. It was repeatedly stressed that the interface needed to be simple and intuitive. A modern phone was mentioned multiple times as an example and not a personal computer program with many windows.

Disussing interaction in the scene - such as scene tagging and placing signs for collegues - was considered very interesting, especially in areas that are hard to reach or cannot be touched yet. *Quote: "This is already done with physical means today but without the flexibility that's proposed."* Other scene interaction that proved interesting to the audience was a timeline concept. The fact that a previously removed body or hazardous scene element could be visualized in the scene was considered a very attractive feature; and if the interface had that time line capability, they would use it.

5.1.5. CONCLUSION

With minimal means, the concepts extracted from the requirements, and background research were demonstrated to a relevant audience. No significant

changes to the original intent had to be made. Especially the quick mapping of environments and expert collaboration received a large amount of feedback.

The comments from the participants will be incorporated into the design where feasible.

5.2. Design Approach

"It is important to associate the problem that is being solved and concisely describe what the intended artifact will be expected to do" - INCOSE (Walden & Roedler, 2015). This includes defining available inputs, necessary functions, expected outputs, desired runtimes, requirements for user friendliness, and the level of fidelity expected. The objectives will form a baseline against verification and validation to be established.

This design exercise has two types of design problems: (1) as described in the background research, there are a few possible candidates for elements that show promise; further research is required, as the exact solution is not yet known. (2) Many of the more standard elements are well defined and require a rigorous design approach. The Classical V-model (Figure 42) is well known to work for a complex system and a spiral development model is suited to iterate prototypes quickly; see Figure 43 (Boehm, 1986).

The design approach is split into the following chapters. First, the basic requirements of the shelf components are identified. Secondly, the subsystems are identified and described in more detail and with interface descriptions according the V-model methodology. Thirdly, the subsystems are developed according the spiral model methodology. The individual component iterative cycles are described and the impact in full architecture tested. Fourthly, the chapter concludes with looking at the full system.

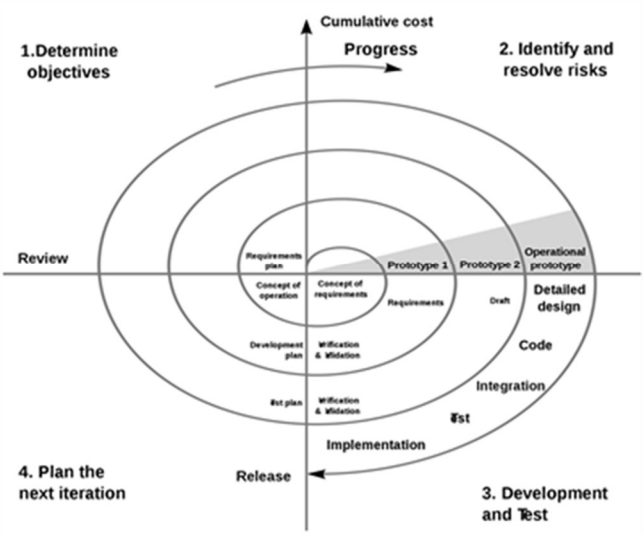


FIGURE 43 A SPIRAL MODEL OF SOFTWARE DEVELOPMENT AND ENHANCEMENT, ADAPTED FROM BOEHM (BOEHM, 1986)

5.3. CUSTOMIZING THE SHELF COMPONENTS

A software framework is needed that is flexible enough to be adapted to the specifics of the requirements. As discussed in chapter 4.3.1, game engines share many similarities with augmented reality systems. The modular characteristics and the availability of many variations help to narrow down the search to that domain (Poelman & Fumarola, 2009; Rocha & Araújo, 2010). The selected engine needs to be adaptable enough to handle the special requirements.

5.3.1. DEVELOPMENT ENVIRONMENT

An important aspect of the development is the selection of a suitable operating system. From a development perspective, there are no real differences between the big three: Microsoft Windows, Apple's OSX or Linux. However, the availability of useful libraries accelerates development. In chapter 4 systems and individual components were discussed that shared commonalities with the intended system. It was found that a significant number of the researched augmented reality systems use Linux as a development environment and rely heavily on existing libraries.

The requirements from chapter 2 do not state any specifics regarding the operating system, and the background research from chapter 4 revealed no

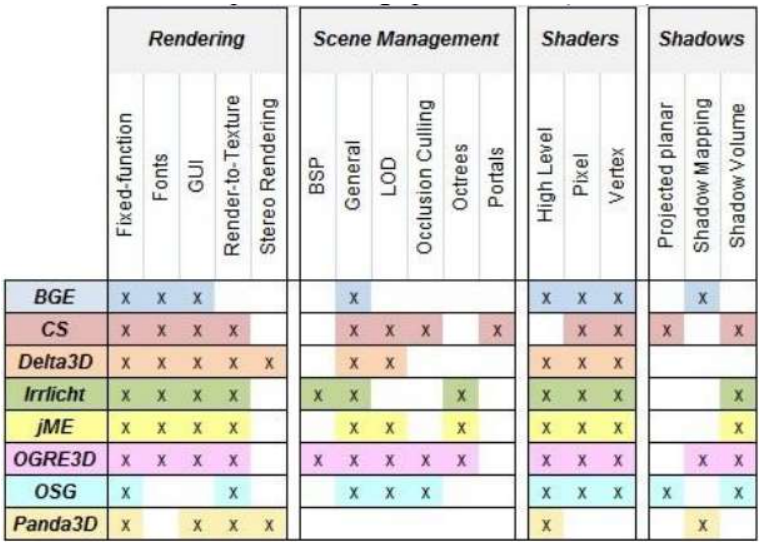
constraints prohibiting the use of Linux. This made the decision to use Linux as an operating system a straightforward one.

Next to an operating system, a suitable hardware platform is required. From the background research on simultaneously localization, mapping and head mounted displays, it is evident that process power and multicore capabilities are important. Next to processing power, a multitude of devices must be able to be connected to the hardware. Cameras, displays and other communication means must be simultaneously connected. It is not feasible to walk around with a personal computer, but making the system run on a mobile device creates a lot of complications for development. A powerful laptop is therefore the indicated choice for development. The decision was underscored by comparison to the systems discussed in chapter 4: the majority used a powerful laptop.

Remote collaboration is part of the main research question. A modality of co-location is an open audio line. However, it is beyond the scope of this thesis to develop such a system, and therefore an off- the-shelf system is preferred. Skype runs on Linux, can be recorded and therefore can be used as an off-the-shelf audio system that does not have to be specifically created to fulfill requirement 12.

5.3.2. SELECTION OF GAME ENGINE

A multitude of game engines is purchasable or available as open source. Only a few of the engines run on Linux, our choice of development environment. A brief list of serious game engine candidates was compiled by Rocha and Araujo (Rocha & Araújo, 2010); see Figure 44.



	Rendering					Scene Management						Shaders			Shadows		
	Fixed-function	Fonts	GUI	Render-to-Texture	Stereo Rendering	BSP	General	LOD	Occlusion Culling	Octrees	Portals	High Level	Pixel	Vertex	Projected planar	Shadow Mapping	Shadow Volume
BGE	x	x	x				x					x	x	x		x	
CS	x	x	x	x			x	x	x		x		x	x	x		x
Delta3D	x	x	x	x	x		x	x				x	x	x			
Irrlicht	x	x	x	x		x	x			x		x	x	x			x
jME	x	x	x	x			x	x		x		x	x	x			x
OGRE3D	x	x	x	x		x	x	x	x	x		x	x	x		x	x
OSG	x			x			x	x	x			x	x	x	x		x
Panda3D	x		x	x	x							x				x	

FIGURE 44 COMPARISON OF LINUX OPEN SOURCE GAME ENGINES, COURTESY OF ROCHA & ARAÚJO (ROCHA & ARAÚJO, 2010)

The game engine frameworks Delta3D, Irrlicht and OGRE3D are the most elaborate according Rocha et al. (2010). When mapping the high-level architecture to the capabilities of the engine, the major criteria were support to client-server network, mature scene management and a solid 3D rendering pipeline. Open scene graph (OSG), although used by many research institutes, lacked some fundamental components that make network capability as straightforward to implement as the other contenders.

Delta3D, at the time of this selection, was missing Linux functionality and was more mature on Windows. Comparing both Irrlicht as well as OGRE3D was more difficult; on paper both shared many similarities. The final selection was based on the better network support in OGRE3D and the more elaborate active community of OGRE3D. OGRE3D matches our high-level requirements; it has network capabilities, elaborate scene management, a solid rendering pipeline and an input devices framework.

It may be argued that a lot of the game engines and scene graph renderers can ultimately be used for the systems, which is true. However, answering the research question is the goal of this thesis, so building this system is a necessity. A well-documented framework featuring all the major components is preferred over custom design.

5.4. INTRODUCTION TO SUBSYSTEMS

After discussing comparable systems in chapter 4, we now direct our focus at subsystems that require to be newly developed or integration which has not been previously validated. The subsystems are categorized into 4 logical modules; see Figure 45. Together with the selected engine framework, better interface definitions and requirements can now be derived.

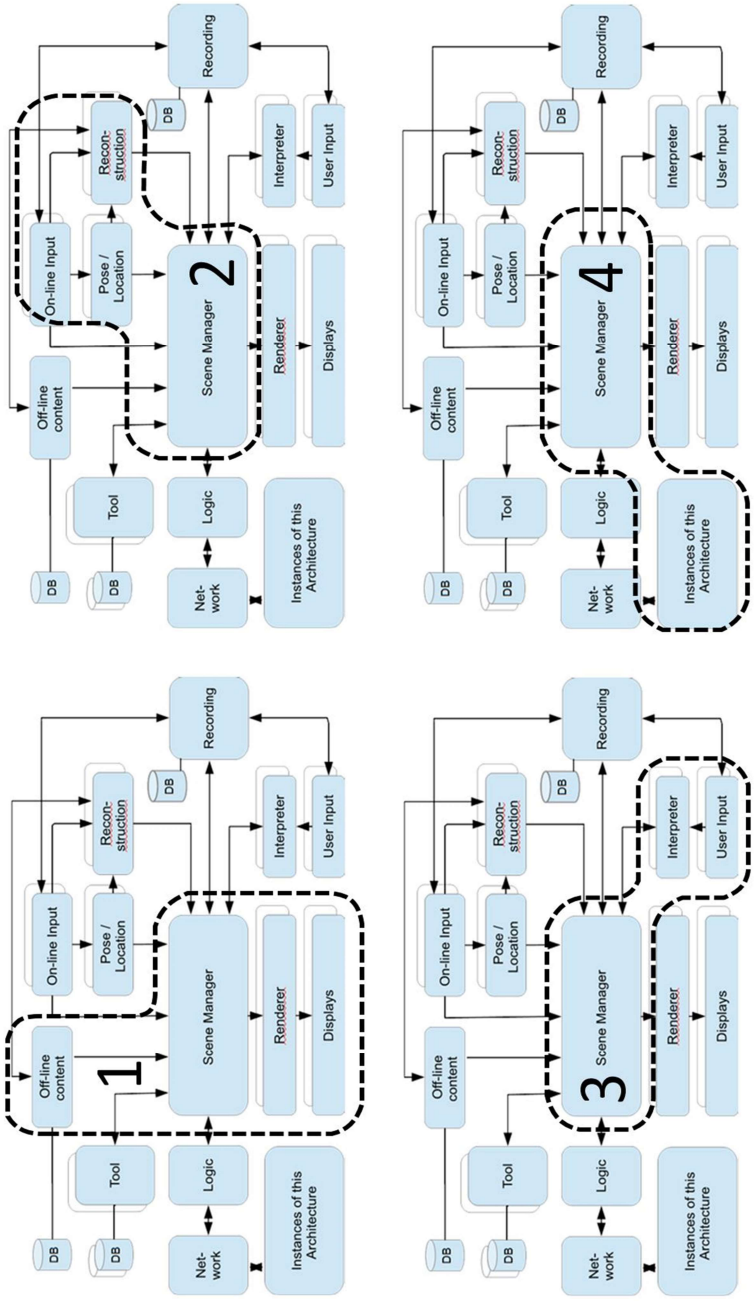


FIGURE 45 THE FOUR SUBSYSTEMS TO BE NEWLY DEVELOPED WITHIN THE ARCHITECTURE

The subsystems all connect to the central module, the scene manager. Subsystem (1) is basically the see-through rendering system. Subsystem (2) represents the simultaneous localization and mapping functionality. Subsystem (3) allows user intervention in the scene, and subsystem (4) allows users to collaborate. Although these are the systems that require to be newly developed, they are not the only subsystems. There are two other subsystems that are not depicted in Figure 45, namely the recording module and tooling.

5.5. ITERATIONS AND DESIGN OF THE SEE-THROUGH SUBSYSTEM

Mediated reality requires an overlay of reality, in the case of this thesis a see-through head mounted display. While this type of system is very straightforward to develop, matching a scene with an optical see-through system has many additional challenges (Caarls & Pieter, 2009). The cost of developing and maintaining an optical see-through system was deemed too high. As Caarls (2009) noted, every time the device is used, the calibration needs to be redone. The decision was therefore made to use a video see-through system; this can always be upgraded to an optical see-through if required. In this chapter, the iterative steps to arrive at a subsystem that suited the needs of this research questions for digital overlay will be explained. The top-level requirement is that it must be able to render a stereo video stream to the displays of the head mounted device. The challenge is depicted in Figure 46.

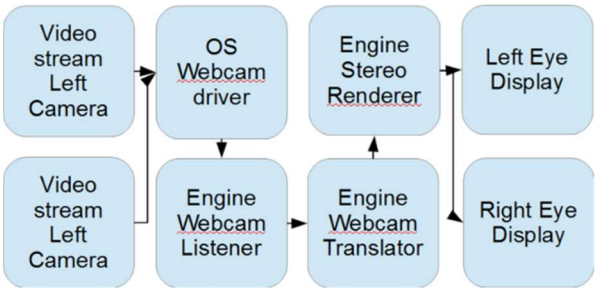


FIGURE 46 STEREO PIPELINE FOR HEAD MOUNTED DEVICES

The operating system has drivers to control cameras, in this case, webcams. The webcams need to be controlled, toggle on and off, resolution, frame rate, etc. For spatial interaction, we need stereo, and thus two webcams. A 3D object needs to stick to the environment at the correct location in the correct size. OGRE does not use webcam data into its renderer by default. A module needs to be designed that controls the webcams from OGRE. In that module, the data needs to be converted

to imagery that is OGRE compatible. The renderer needs to use the imagery according the physical distance of the webcams to output a render for each eye in the correct scale and distortion. The hardware setup reflects the same subsystem, i.e. two webcams wired to a laptop running the software and the head mounted device.

To validate whether the subsystem was working correctly, a virtual box was created at an estimable size and projected into the user's view. The size of the object needed to resemble an existing object in the real environment. If the two object feel comparable, the overlay with the head mounted device succeeded. The second test tested for performance: when the user moved his or her head, the rendering should not trail more than 40 milliseconds behind; Caarls (2009)

5.5.1. HARDWARE ITERATIONS

The relevant lessons learned are discussed. Although military grade off-the-shelf video-see through systems exist, the availability and custom modification capabilities of such systems prevented us from taking that route.

Iteration (1)

The main goal of this first iteration was to get a full functional pipeline working and to learn what is relevant. The hardware of the first system consisted of the off-the-shelf AV 920 Vuzix HMD, two stripped Logitech Quickcam pro 5000 cameras connected with the wires to a laptop. (cf. Figure 47)

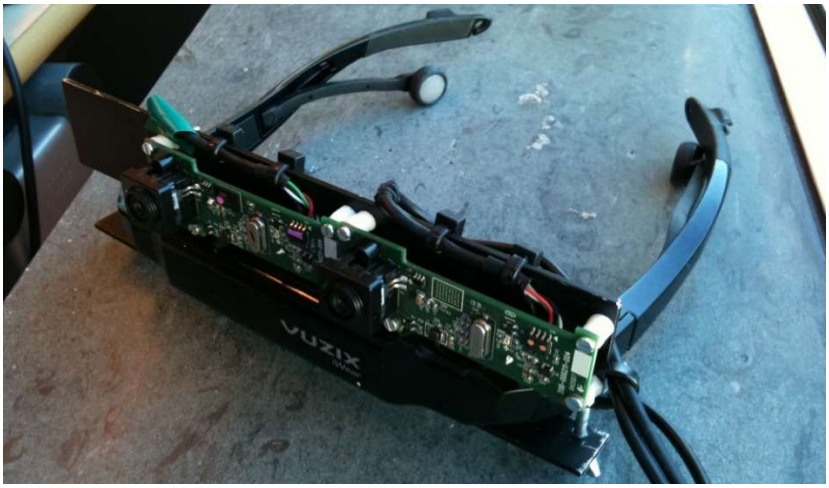


FIGURE 47 FIRST ITERATION OF THE HARDWARE FOR VIDEO-SEE THROUGH

Two webcams were mounted on top of the AV 920 at the same distance as the displays. The USB cables of both the webcams connected to USB ports in the laptop (not on the picture) and the AV 920 received its picture through a VGA port, with the syncing being done through USB connection. The software implications will be discussed in the following chapter, the webcams produced 30 fps @ 800x600 max resolution and the display could either use 60hz side-by-side imagery or 30hz per eye field sequential, both at VGA resolution. The webcams had a diagonal field of view of 75° , displays of 32°. To circumvent the field of view differences, the images from the webcams were cropped, leaving less resolution but a correct match. Had we not done that, the scale of the scene would have been off, with the digital object looking as if it were too far away.

The system worked correctly and a virtual object could be drawn at the right size. Two issues haunted this system. First, the resolution was considered too low fidelity when shown to a CSI core team to effectively do tasks in 3D. Secondly, the offset between the cameras and the displays above each other was not favorable. Warping the images to the correct viewpoint did not produce better results.

Iteration (2)

For the second iteration, the resolution needed to be higher; this meant both the recordings and displays needed an upgrade. A prototype from Carl Zeiss, the Cinemizer, was used to improve the image resolution. The Cinemizer was complemented with Logitech HD C500 webcams with 1280 X 720-pixel video, 30 frames per second (cf. Figure 48).

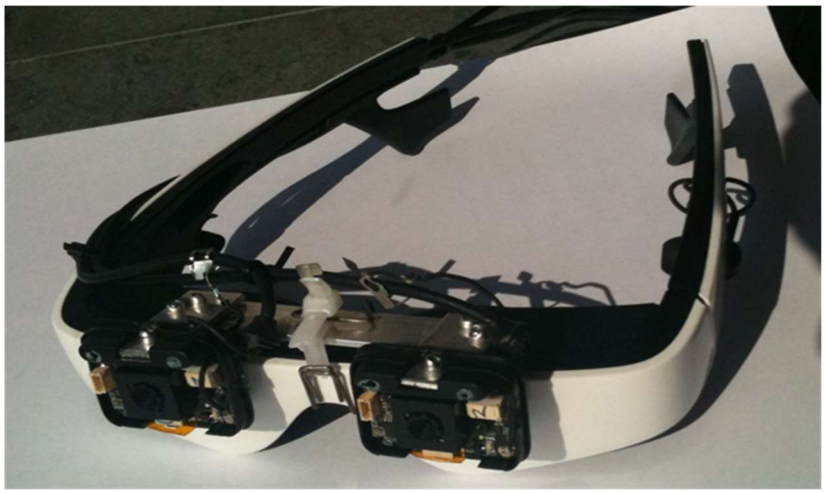


FIGURE 48 SECOND ITERATION OF THE HARDWARE FOR VIDEO-SEE THROUGH

This second iteration patched the flaws of the previous system, the resolution was better and the webcams' cameras were in front of the displays. Most of the principles from the first version could be re-used in this iteration. Two new challenges emerged from this iteration. First, the webcams' mounts were found to be too flexible. Drawing a virtual cube on the screen resulted in a slight offset every time a webcam received a slight bump. And secondly, the laptop we were using was not able to cope with the bandwidth of 2x a 720p video stream on a single USB controller; frames were dropped and syncing became a problem.

Iteration (3)

Fewer modifications were necessary for iteration 3; basically, slightly better webcams were used with a little smaller board size, namely the Microsoft's LifeCam HD-5000 webcams. The laptop was switched with an USB 3 capable laptop with multiple USB controllers (cf. Figure 49)

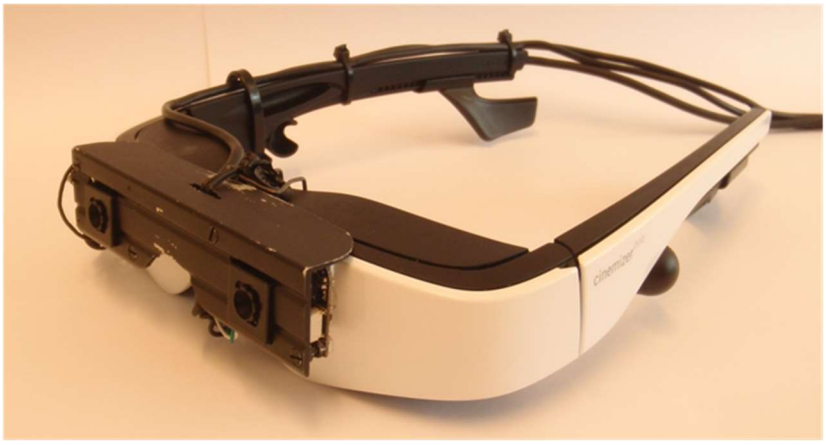


FIGURE 49 THIRD ITERATION OF THE HARDWARE FOR VIDEO SEE-THROUGH

This third iteration did not have any of the drawbacks noted in the previous systems. The pipeline allowed for a resolution and frame rate throughput that had room to spare. The result was a state-of-the-art video see-through head mounted device capable of digital overlays with correct scaling. A 30 Hz refresh was maintained thought-out the pipeline and the resolution maintained the display max resolution at all times. The weight of the Cinemizer OLED was ~115

grams and the attached stereo rig ~65 grams, well below the 225 grams mark advised by Wann (Wann et al., 1994).

A laptop must still be attached to the device, but it can be run without the power cord, and carried in a backpack. In the next chapter, the software iterations will be discussed that were required to obtain an acceptable system throughput and frame rate.

5.5.2. SOFTWARE ITERATIONS

The software reflects the iteration steps of the hardware. Development effort was directed at the optimizations necessary to achieve high frame rate/throughput. Figure 46 depicts the building blocks that were used to iterate to an acceptable subsystem.

Iteration (1)

The first software subsystem relied upon the hardware from Iteration 1: webcams are the Logitech Quickcam pro 5000's and the AV 920 Vuzix. As discussed in the introduction (cf. Figure 46), different architectural components are necessary to create the software pipeline associated with the previously discussed hardware. In this section, the components will be examined and in following iterations, the major changes will be discussed.

OS Webcam driver

The first module has access to the stream of captures from the webcams. The pro 5000 was fortunately compliant with USB Video Class (UVC) drivers. This software component fully supports UVC and a wide range of compliant devices in Linux, including a V4L2 kernel device driver and patches for user-space tools. UVC has direct programmatic control of the settings of the webcams, the id can be extracted, and the resolution and modes can be accessed. A frame grabber module was created to change the settings of the webcams and to gain access to the frame buffers for frame extraction.

Engine webcam listener

The main application loop of OGRE needs to know that the webcams are connected and it needs to be able to set the resolution, frame rate, etc. Feedback must be provided to the user if there are any problems and the frame memory buffers must be known. An OGRE Engine webcam listener was created to take care of the webcam within the main application loop.

Scene manager

Before the images from the webcams can be used by the real-time renderer, a few steps need to be taken. First, the webcams need to get the correct offset, aka the inter pupil distance from the physical webcams, which, depending on the person, is approximately 65mm. Secondly, as only 32° of the image is of interest, a crop of the image needs to be rendered. Thirdly, the webcam streams need to be viewable in the scene. To that end, image planes are created at the same locations as with the physical cameras, which use the webcam data as live textures. An OGRE scene manager is created that facilitates in setting up de scene for rendering.

OGRE Stereoscopic renderer

With sufficient light, the webcams provide 30 frames per second of images. The renderer needs to either output one single image with a left and right frame squashed into one 640*640 pixel at 60 Hz or sequential left/right with 30 Hz per eye. The higher update rate of side-by-side was preferred to sequential. A workaround[20] was necessary to control the stereoscopic rendering on the AV 920 in Linux: the device needs to know, through USB, what mode and frame is sent through the VGA signal. A custom view is calculated and projection matrices were set for each eye by using a set frustum, which mimics the eye position. To obtain the correct view frustum of the webcams, the camera calibration toolkit from Jean-Yves Bouguet[21] was used. The distortion of the webcams and displays was calculated to match the distortions. OGRE's compositor was used to composite the left and right eye frames side-by-side; the frames were packed into a single buffer, which the GPU sent to the display. A hardware specific OGRE stereoscopic renderer profile was created that could be re-used and edited if required.

The pipeline discussed zooms in on the flow starting from recording up to display; other aspects had to be set up, too. In augmented reality, frame rate is of essence, as, if it is too low (+200ms), people get sick. Timers had to be built in to know precisely how much time was spent from the moment the video frame buffer was available up to rendering the frame again. To maintain frame rate, the capture module was placed on a different thread than the renderer, which helped to maintain frame rate below throughput times. Furthermore, a visual feedback

[20] http://www.pabr.org/wxhmd/doc/wxhmd.en.html#biblio, last visited July 2017.
[21] http://www.vision.caltech.edu/bouguetj/calib_doc/, last visited July 2017.

environment for debugging was necessary. A secondary monitor was used to be able to look at the renders in left, right or stereo mode. The render environment was set up with specifics for overlay; to keep them neutral, the webcam video textures could not receive shadow or bounce light and the render planes were set beyond possible scene elements. If the planes are near the eye they occlude everything else in the scene; this way, they serve as a backdrop.

Iteration (2)

The first system did not have sufficient resolution; hence the hardware was upgraded. The hardware for this iteration was the Cinemizer OLED with two HD C500 webcams. Fortunately, many of the software components could be re-used.

Engine webcam listener

The OGRE frame manager was revamped in this iteration, to be able to cope with a bandwidth of 2x 720p video streams. An abstraction class was created that served frames. Some webcam configuration frames needed frame flipping, color correction or cropping, and knowledge of skipped frames. Serving the exact frame to be used in the render pass, helped reduce the delay caused by working with high res imagery. Switching from the default MJPEG compression to uncompressed YUV had a dramatic impact on quality. The amount of image distortion introduced by MJPEG was significant. The OGRE frame manager was improved to serve personalized frames.

Scene manager

Constant calibration issues plagued this iteration of the device, as the bare webcams were very sensitive. To cope with this, an in-application calibration module was created. Basically, by pointing the webcams at a checkerboard multiple times, the matrices could be created that are used by the renderer. The OGRE scene manager could dynamically create the new matrices, correcting the webcam adjustments, and the results were directly viewable on the displays. A benefit of this procedure was that it allowed correct stereo and scaling to be verified on the fly. The cost of using YUV in the pipeline was significant but worth it and ffmpeg can be used to translate YUV -> RGBA. In the first iteration, the calibration image distortion parameters were not visible. However, because of the higher pixel density and better verification methods, it became obvious that these parameters influenced the scene display. Basically, there are two methods to address this: either the 3D scene elements can be modified to match the cameras or the images can be modified. Modifying the images proved to be costly,

and to cause a 10-15 milliseconds delay. A more effective method was to apply the warping on a tessellated version of the image planes. This only needed to be done after the calibration, and the rest was for free. The verification method for correct display was refined in this iteration, as the checkerboard proved to be very effective. Warping the rendering results to match the distorted images proved to be less effective and harder to control.

Stereoscopic renderer

The rendering was simplified in this iteration, the syncing of the device per USB was not necessary with the Cinemizer OLED, directly rendering a side-by-side image proved to be all that was needed. The main change to the rendering was serving a 2x higher resolution image to the device, as the hardware downscaling of the image proved to deliver higher image quality then serving the exact resolution. The goal was to keep the frame rate faster than the acquisition frame rate.

Iteration (3)

Only the webcams were changed in this configuration, making the hardware more stable and requiring less calibration. Most of the improvements at this stage were directed at compatibility and documentation for other subsystems. The frame rate, the resolution and the overlay capabilities met the requirements. The only small enhancement was the slight backward movement of the virtual cameras to be closer to the position of the physical eyes. This helped to lessen the feeling of offset when putting on the head mounted device.

The OGRE frame manager was not only useful for providing frames to the overlay, other subsystems needed frames too. Cameras can be added; different resolutions extracted and efficiency needs to be maintained. Furthermore, based on the base architecture, the raw data was known to require recording.

The OGRE scene manager is currently without a user interface; other subsystems need user interfaces to influence the scene. This is not part of this subsystem, but a requirement (13), none the less. In the next chapters, the interface on top of the OGRE Scene manager will be detailed, as well as the tooling. The current renderer, the OGRE stereoscopic renderer, needs to be set up to accept other rendering subsystems from the base architecture. Objects, tracking results and user interface components all need to exist in the same scene graph and Cartesian position.

5.5.3. CONCLUSION

The result of the iterative steps led to a subsystem that fulfills the requirements (6, 16) stated in chapter 2 and that forms the basis for augmentation of any scene. Audio, the last part of the requirement is not considered in this subsystem, it is customized off the shelf.

From a performance and visual fidelity level, our system is capable. And while this may not be the most capable hardware solution money can buy, our system is fully configurable to what we need.

FIGURE 50 VALIDATION OF THE VIDEO SEE-THROUGH HEAD MOUNTED DEVICE AUGMENTATION CAPABILITIES

Figure 50 depicts the head mounted device (bottom middle), the fixed checkerboard (right) setup and the calibration toolbox matches (upper middle and left). The performance is better than 30 fps.

The first steps to interface with other subsystems have been created. They will be discussed in the corresponding chapters.

5.6. ITERATIONS AND DESIGN OF A 3D SIMULTANEOUS LOCALIZATION AND MAPPING SUBSYSTEM

As a result of the collaboration with the TU Delft Robotics Institute, the bulk of the work on implementing the computer vision technology was accomplished by Oytun Akman (Akman, 2012). A more in-depth analysis of the algorithms can be found in Akman's work (2012). In chapter 5.5, a video see-through system was developed that allowed virtual 3D content to be rendered on top of a stereoscopic image stream. The present chapter will be dedicated to the creation of the subsystem allowing the user of the head mounted device to know, at all times, where he is in relation to a mapped scene and facilitating the creation of a 3D map of the scene.

State-of-the-art systems have shown that visual tracking can provide us with a map of the scene. The see-through subsystem supplies a stereoscopic image stream that can be leveraged in pursuit of the goal of this chapter, because it provides all the information used in metric localization and mapping. Furthermore, based on the discussion of the state of the art in chapter 4.2, it is evident that PTAM (Klein & Murray, 2007) offers a good methodology, implementation and open source code base.

The challenge of this chapter comprises three distinct, albeit related problems: map making, pose estimation and re- localization

5.6.1. LOCALIZATION AND MAP MAKING

The localization and mapping processes are intertwined, as a map is needed to compute a pose, and without a new pose, the map will not expand (Durrant-Whyte & Bailey, 2006). The frame manager and the scene manager have already been discussed in chapter 5.5. The frame manager's purpose needs to be extended with serving frames to the pose estimation pipeline and the scene manager needs to accept the translation and rotation of the origin for every frame; see Figure 51.

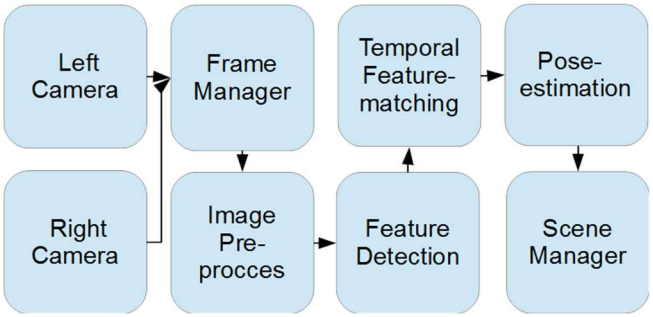


FIGURE 51 INTEGRATING POSE ESTIMATION INTO THE VIRTUAL SCENE.

Iteration (1)

By natively using PTAM, a few requirements were not met:

- PTAM was created for a monocular camera, hence scale is not derived; our requirements dictate correct metrics (01);
- The initialization of the tracker is based on a sideward movement of the camera. During the demos in 5.1.1, this proved to be unreliable, as most of the demos needed multiple attempts to start;
- Bigger than tabletop scales introduced drift quickly.

The hardware discussed in the previous chapter is stereoscopic, high resolution and with a wide field of view. From the literature we know that tracking becomes more stable with a wider field of view (Klein & Murray, 2007). The first step is to leverage the hardware to improve ease of use, stability and precision.

For tracking, the frames are in grayscale: the frame manager has been modified to serve the same frames in grayscale to the image preprocessing module. The contrast in the images has been improved to acquire more feature candidates (Klein & Murray, 2007). In the current PTAM pipeline, the matching starts with comparing the first two frames acquired with the side sweep. Because of the stereo rig, the first two frames are from the left and right camera. If features are detected in both frames, the initialization is based on the stereo rig; this eliminated having to side sweep for initialization.

Because the stereo rig is calibrated, the exact position of the cameras with respect to each other is known. Any feature that can be matched between the two cameras has a metric depth based on triangulation, providing metric scale to the scene.

Temporal feature matching offers a track table of matches both between sequential frames as well as between the stereo pair. The pose is derived from the track table and provided to the scene manager in the same package as the color image used for augmentation. Because the field of view of the stereo rig is more than a monocular setup, the reliability of the system increases; double the image data and slightly more field of view.

When moving the stereo rig, the map of features grows, basically providing a 3D map of the scene in metric 3D. The scene is sparsely represented but up to scale and already is recognizable. Figure 52, shows a sparse feature map of a corridor reconstructed with the stereo rig. The two blue lines represent the left and right cameras

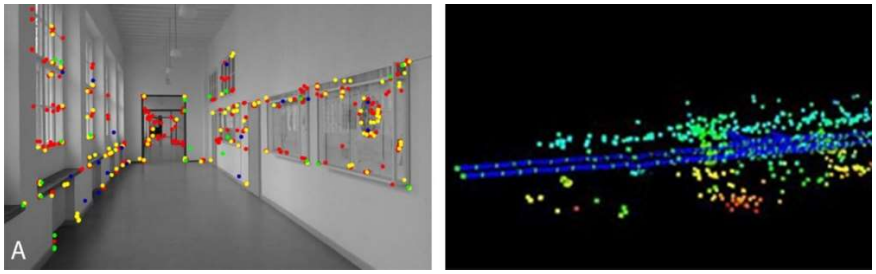


FIGURE 52 A SPARSE FEATURE MAP OF A CORRIDOR; RIGHT DETECTED FEATURES, LEFT STEREO TRACK

With this first iteration of the tracking and map making processes, the earlier mentioned disadvantages of PTAM are circumvented.

- A 3D sparse map is created with metric scale;
- The initialization of the system is based on the hardware, no manual steps necessary;
- The drift is significantly reduced because of wide angle stereo setup cameras.

Iteration (2)

With the introduction of different webcams, a new problem occurred. The webcams differentiated significantly in color representation. The automatic settings to color balance yielded very different results, even with very controlled tests. Modifying the driver software to cope with the color differentiation did not yield results. This negatively affected the matching of features as well as the visualization, and to cope with this new problem, a color equalization function

was added to the image preprocessor module. Effectively, the color space of one image was used to provide the other camera with a color template.

Although the precision of PTAM was improved, the frame rate went from 60+ Hz to ~10 Hz. The full system needed to function with ~25-30 frames per second, so improvements were necessary. Tracking constantly with two cameras proved to be unnecessary, as once an initial map was created, the tracker would function just as well with one camera. It was only when key frames or bundle adjustment was required that the imagery from the second camera was leveraged. The frame rate after this modification went up to ~20 frames per second. The PTAM feature matching relies on normal direction corrected patches; the more patches that warped, the slower the matching process. The new cameras recorded in 720p and therefore had more resolution to be processed. Downscaling the images improved the performance, pushing the frame rate to the desired ~30 Hz.

With the tracker working, the 3D maps required more fidelity. A dense 3D map of the crime scene is relevant information for collaboration; it provides a more detailed copy of the crime scene, allowing the analysis to be more detailed. The pose information coming from the pose estimation module was used to construct the 3D map of the scene. A continuous stream of disparity maps was generated while the user moved around the scene. Each new disparity map was registered (combined) using the pose information from the module to construct the 3D map of the scene. The dense maps are created at full 720p resolution, as opposed to the pose estimation, which uses limited resolution. Effectively, a dense map was generated every 0.5-2 seconds, and stored as a point cloud. The time and processor friendly block matching algorithm was used to generate the depth maps (Hirschmuller, 2008).

Iteration (3)

The tracking and map making processes proved to be less reliable when used on bigger scenes (multiple rooms) and when the system was used for longer periods of time (10 minutes+), two deficiencies that are acknowledged by Klein and Murray (Klein & Murray, 2007). Slightly out of sync cameras and frame dropping introduced errors that needed addressing. The frame manager was extended with syncing functionality that allowed for less syncing error; the initialization of the cameras was controlled. The webcams decide whether or not to drop a frame: this is uncontrollable from the software, as this is dictated by the lighting conditions. The system marks dropped frames, preventing the use of wrong frames. With a monocular setup, this is less of a problem, but because the system has a render pipeline that relies on a certain frame rate in stereo it is relevant.

The dense structure of the motion point cloud, although colored, is relatively noisy compared to a mesh model; meshing the data and re-projecting the key frames provides cleaner data (Wand et al., 2008). A simple scheme was used per key-frame depth image, running Poisson surface reconstruction (Kazhdan, Bolitho, & Hoppe, 2006). No attempt was made to merge into one mesh.

A side effect of using a vision-based system is that one camera might decide to drop a few frames or is occluded by an object or person. The software was adapted to switch main tracker camera instantly if occlusion or failure occurs: without features a vision base system cannot function.

A not yet addressed capability of the subsystem is the ability to re-localize itself if the tracker is lost. In the previous two iterations, the default PTAM method was used. Searching through key frames is the basic method for re-localization. The look-up database of key frames was extended to take the additional stereo frames into account.

The used tracker attempts to fit a plain on the sparse feature set with Random sample consensus; the xyz origin is placed on the detected plane. This proved to be unreliable, as in practice, the plane was often found at odd angles, connecting closets and chairs. In the video see-through chapter, it was seen that a calibration procedure was implemented. This procedure was now also used in the initialization; if a checkerboard is detected, a plane is fit and the origin constructed. For our test environment, a checker board could be placed on the floor and the system would always have a usable origin.

5.6.2. CONCLUSIONS

A tracking subsystem was created that automatically creates sparse maps of environments. Any object placed in this map will have the correct scale and can augment the scene. Every frame captured by the stereo rig is associated with a pose of the rig in relation to the environment. Augmentation is a necessary step for mediated reality and one of the key requirements (01, 05).

The iterations resolved into a subsystem that functioned at ~30 Hz. The subsystem was tested indoors and outdoors. The corridor test (cf. Figure 52), that failed with PTAM native because of the relatively featureless long distances worked well with the enhanced system.

Although other sensors and devices, such as GPS and inertia sensors potentially improved the results, they were deemed unnecessary to answer the research question. Furthermore, no additional hardware was necessary to achieve this

functionality. If improved tracking methods for wearable hardware should emerge, based on depth sensors or otherwise, the subsystem can be easily replaced or enhanced.

5.7. ITERATIONS AND DESIGN OF A 3D USER INTERACTION SUBSYSTEM

This chapter relies heavily on the collaboration with the TU Delft Robotics Institute. The technical details can be found in a study conducted by Akman et al. (Akman et al., 2013). In the previous chapters, the subsystems were designed and prototyped that allowed for a virtual overlay in a re-time tracked environment. Nothing happens in the 3D scene. An investigator walks around with the device that merely "scans" the environment. In this chapter, the interaction with the virtual space will be discussed.

In chapter 5.1, the use cases are discussed with the investigators. Based on the discussions, two user interactions can be derived: tool (de)activation and interaction with the 3D scene. In both scenarios, a user needs to make an intent apparent to the system.

Actions are taken in the field of view of the user; it makes no sense to allow actions that have no visual feedback. The hardware discussed in chapter 5.5.1 already monitors the field of view with a stereo rig, and this will be further exploited. While this chapter is not dedicated to investigation tools, in some cases they will be used to illustrate the goal. The focus is on 3D interaction with the scene, as tool (de)activation is considered a subset of 3D interaction.

The analysis discussed in 2.2, where, among others, requirements (11, 13) were formulated, means that objects must be able to be placed in the 3D scene. The subsystem must replace the use of a mouse and keyboard to move a 3D object around. However, the investigators must have free use of their hands without having to hold any objects. Background research showed that hand gestures can be effectively used to control 3D objects in 3D space; audio fails for 3D control.

Currently, controlling a 3D object in a 3D scene with the WIMP paradigm is possible because the viewport has scaling and a free camera. An investigator can walk around, but does not have that kind of infinite camera freedom. Furthermore, the hands have a limited reach, making placing something beyond arms' reach difficult. What is the equivalent of a mouse click? How does the system know that a selection or placement has been acknowledged?

5.7.1. INTERACTION PARADIGM

The first iteration was focused on scene interaction, given that hand tracking is possible (chapter 4.5.2.). A minimal finger tracking pipeline is depicted in Figure 53.

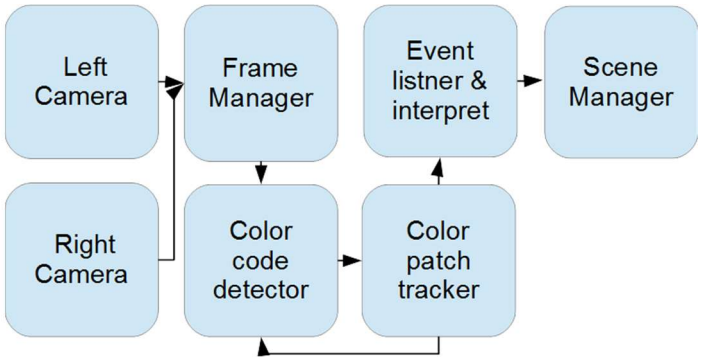


FIGURE 53 COLOR CODED FINGER TRACKING

Iteration (1)

Using color coded sleeves on three fingers per hand, a rudimentary hand tracking system was developed. The goal of this system was to study interaction paradigms. The description of the individual components will be postponed to the next iteration, due to the large number of changes that were made. The subsystem is depicted in Figure 54.

FIGURE 54 COLOR CODED HAND TRACKING SUBSYSTEM

An algorithm was designed that distinguished three types of bare-hand gestures: left hand thumb-up, left hand thumb-down, and right hand thumb-down. This system is similar to AirTap (Vogel & Balakrishnan, 2005), as for the sake of reliability, the gestures needed to be easily distinguishable and simple to learn.

Figure 55 (left) shows the gestures distinguished with the defining hand postures. The algorithm was designed as follows: a click is performed by moving the left or right recognized segmented hand forward quickly, and moving it backward again. The direction of movement of the segmented hand is continuously monitored to recognize this gesture. When the pointer moves in a forward direction, the path over which it is moving is tracked. As soon as it has moved forward and then backward more than halfway along the same path, this is registered as a click at the furthest point of the path. If anywhere in this sequence the segmented hand deviates more than a pre-defined angle from the path, the event is not recognized as a click. In this way, both small and big gestures are recognized, as long as the direction of the movement is right. A more elaborate description is given by Lukosch, Poelman Akman & Jonker in their article that appeared in 2012 (Lukosch, Poelman, Akman, & Jonker, 2012).

FIGURE 55 SUPPORTED HAND GESTURES (LEFT), EVALUATION SETUP (RIGHT) ADAPTED FROM LUKOSCH ET AL. (LUKOSCH ET AL., 2012)

5.7.2. EXPERIMENT

To validate 3D user interaction and to assess whether it was suitable for mediated reality, a simple experiment was designed. The goal of the experiment was to validate if 3D gestures would allow participants to control what they are doing in a 3D scene.

Participants

The setup was created in the forensic field lab in Delft. 10 people from the NFI and the CSI The Hague project volunteered to participate, 6 professionals from the NFI and 4 CSI The Hague non- professionals.

Measures

The performance of the participants was measured in three ways. First, by logging all the hand movements and secondly, by having the participants complete a small questionnaire (Appendix II – Questionnaire for 3D interaction) to establish their experience level, the perceived ease-of-use and usability. The questionnaire was inspired by Grinblat & Peterson (Grinblat & Peterson, 2012). Lastly, an after-action group discussion took place.

Setup

The participant looked at a large wall projection of a stereo pre-recorded dummy 3D crime scene. A stereo camera rig was mounted on a baseball cap to simulate wearable interaction (Figure 55). We instructed the investigators to look at the projection, which depicted a prerecorded crime scene. Their hands could operate in the pre-recorded crime scene like a video see-through subsystem; the user and screen were fixed in space. The experts had to conduct two tasks: 1) browsing through the options menu, in which only basic 2D GUI tasks had been loaded, and 2) tagging the crime scene by manually selecting 3D points and placing virtual objects. The participants received a five-minute introduction in which the test set-up and interface was explained.

Technical results

The evaluation of our tests involved the analysis of the log files that recorded the gestures of the participants, the TAM-based questionnaires, and the after-action group discussion. Selecting the appropriate tool from the menu took most experts just one trial. The system failed if participants needed three trials and had to be restarted. The conclusion was that 2D tool selection worked well from a technical perspective.

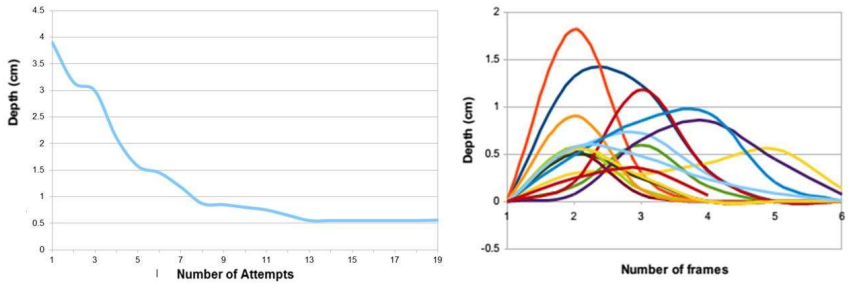


FIGURE 56 GESTURE MOTION FOR 3D POINT SELECTION OF AN EXPERIENCED PARTICIPANT

For the selection of points in real 3D, scene metrics needed to be derived. The depth from motion recognition was observed for accepting the select command and the number of frames it took to recognize the command.

Figure 56 (right) shows the results from one experienced participant; with the horizontal axis plotting the number of image frames (and hence with 30 Hz implicitly the time) and the vertical axis the observed path length of the index finger. The higher curves show the first attempts of the participant. The right side of the graph is a visualization of the exact numbers, which are interpolated to show the median more clearly; the lines do not exist in the measurements. The curve on the left shows that the selection became stable at 0.5 mm and that after 11 attempts, the user had optimized the motions such that he could perform the required 3D point selection in a minimal amount of time with minimal motion. The depth of the finger motion was less than a centimeter. The operations took on average just 3-8 frames. This shows that the algorithm can deal with very slight movements in a short time. For more details, see Lukosch et al. (Lukosch et al., 2012)

In the questionnaire, each participant was asked to provide information about their age, previous experience and field experience. By reviewing the log files and plotting them in Excel, the only clearly noticeable difference between the participants was their experience in working with software-based 3D models. Hence the group was divided into three classes of users; experienced users who use 3D models daily in their work, normal users who do not use it on a daily basis but who are familiar with 3D models, and inexperienced users. All participants accomplished at least 15 3D placement actions to learn the 3D point selection action. The evaluation shows that the experienced users have one main bump in their motions, indicating there is not much difference between their attempts and that they learned the minimal quick motions to trigger the 3D point selection. The overall results of the normal users showed that they were able to master the minimal depth to trigger the action, although their motions were less crisp than those of the experienced users. The inexperienced users were also able to learn the trigger motion, but they performed this slowly and with abundant motion. The questionnaire showed similarities for both the experienced and novice users. They evaluated the gesture system as easy to use and easy to learn. They liked and felt confident in using the system, which was confirmed in the control questions. Interestingly, the feedback from the less experienced participants was the most positive.

Results from the after-action review

The after-action group discussion provided us with additional insights. The experienced users compared the interaction with the performance of their everyday work, as opposed to the inexperienced users, who were impressed by the system. In our log files, we could also see that the experienced users invoked 3 times as much action as the inexperienced users, which indicates that they were testing the system more thoroughly. Furthermore, the participants were asked why they were making small or large gestures. We had expected gestures of 3-4 centimeters and in the test, most motions had been less than a centimeter. The experienced users asserted that small, quick gestures provided them with more control and precision.

Conclusion

Gesture-based interaction provides sufficient control to place objects in 3D environments, even when the objects to be placed are out of tangible reach. Both for advanced 3D users and novices, the system satisfied the requirements.

Iteration (2)

In the interaction just discussed, a minimal hand tracking solution was implemented. As color was used, a very small variation in light could throw the system off. In our experiment, the light was controlled. The validation was done under strict lighting conditions. This second iteration was used to get rid of color coded sleeves, see Figure 54.

Some important lessons were learned during the development of the iteration (1).

- Limiting the search space in depth eliminated many outliers;
- Using the full 720p resolution is needed to obtain depth accuracy.
- Dynamic backgrounds need to be filtered.

Based on the lessons learned, a new system was designed to which we added stereo depth reconstruction; see Figure 57. The.

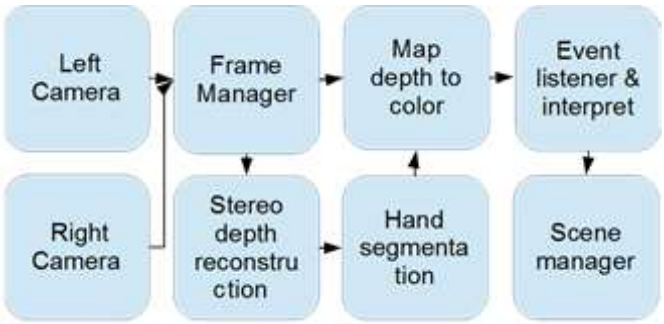


FIGURE 57 HAND GESTURE INTERPRETATION PIPELINE

Frame Manager

The frame manager was extended to serve gray scale images to the stereo depth reconstruction module for every frame acquired. Hand stereo and scene stereo are separate threads.

Stereo depth reconstruction module

To eliminate the background and select likely candidates for hands in the scene, a stereo module was created. This significantly reduced the search space; background is automatically ignored. With block matching (Hirschmuller, 2008), a depth per frame was extracted and clipped to the maximum and minimum range of hand reach. Candidates were supplied for the next module.

Combine color and depth candidates

Both color and depth are candidates with the potential to be hands. This module is responsible for deriving the best candidate hands in the scene. Color space facilitates hand candidates and depth provides a logical candidate depth space. A prediction for the tracker was applied to ruggedize the tracking.

Event listener and interpretation

Left hand, right hand or both remained to be determined, as well as the intent of the user. Clicks must be detected and the position in space needs to be derived. This module interprets the detected hands to be used by the scene manager. The pose of the hand and relative position in space provides the correct labeling; see Figure 55.

In this second iteration for hand gesture recognition, the color-coded finger sleeves were replaced by recognizing the hands in natural images without any physical cues. The only impact on the system was that the full hand was now used to detect a click, instead of a single finger. A more detailed description on the implementation can be found in Akman et al. (Akman et al., 2013).

Iteration (3)

The subsystem works at ~30fps and can detect hands with acceptable reliability, see previous iterations. However, users quickly became frustrated with the system, as there was no way a user could tell whether an intended action had been registered or not. As a result, the activation of unintended gestures as selections was a frequent occurrence.

A feedback system was designed to aid in solving this problem. Small boxes at the bottom of the screen were introduced that allowed the user to see whether a left, right or both hands had been detected. A small picture of the detected hand is displayed. The color of the box changes when a click is detected and the gesture

recognition could be turned off and on with a broad gesture, preventing undesired action.

5.7.3. CONCLUSION

A gesture-based subsystem was designed that allowed users to place 3D objects in a virtual scene with bare hand gestures. A more detailed description of the technical implementations and testing can be found in Akman et al. (Akman et al., 2013) and the experiment is described in Lukosch et al. (Lukosch et al., 2012)

The performance of the system was comparable to the previously discussed subsystems, approximately 30 Hz. The system surpasses the current state of the art stereo systems based on mini rigs. In the future, when mini depth cameras have been miniaturized, the components that are responsible for recognizing the hand can be replaced.

5.8. ITERATIONS AND DESIGN OF A REMOTE COLLABORATOR SUBSYSTEM

In this thesis, collaboration is understood to mean that the pristine environment is not only available to the investigator wearing the video see-through system, but also available to colleagues who are providing aid in the pristine environment. The pristine environment is the subject, and the colleagues need a view of the data and the aid provided, as reflected in requirements (11, 13).

Tele-presence manifests itself through technical means, a display, cameras and maybe robotic components instead of another human being (Adachi et al., 2005). In the case of this thesis, the environment and the position of the person is shared and linked in the intended system, which is different from tele-presence

There are four types of data being shared; an open communication line with audio (1), the view of the user of the scene (2), a 3D reconstruction of the scene (3) and virtual overlay actions on the scene (4). Audio (1) will be ignored in the design of this subsystem because an off-the-shelf solution will be used that runs separately. In this chapter, the development of the subsystem that allows for collaboration will be explained.

During the sessions with the investigators, discussed in chapter 5.1, the following scenario was proposed based on the technology discussed.

"The investigator that's first on the crime scene has to mark an area that cannot be contaminated because of tracks and blood. A virtual obstruction is necessary to

block others and point experts in the right direction. The areas must be correctly detected and marked. An expert aids the investigator on location with correctly marking the scene, if the expert needs to make changes he should be able to do so."

This scenario was used to design the subsystem. The validation of the subsystem is based on two network connected laptops that sync; objects and object changes in the scene, which displays a stereo recording and frequently updating scene reconstruction.

There are two types of cameras in the scene: the stereo camera of the investigator on location and the mouse-controlled camera of the remote collaborator. The stereo camera only moves when the investigator moves, the remote collaborator's "free" camera can move to any location at any point in time.

5.8.1. REMOTE COLLABORATOR SUBSYSTEM

Rather than designing different subsystems for the various instances, we chose to design one subsystem with various modes, depending on the view and capabilities needed. There are only a few technical differences between the two types of users:

	On location	Remote	Synchronization
View on the data	Stereo perspective	Stereo perspective, free camera	Stereo
Scene objects (locally stored)	Add, move, remove	Add, move, remove	Bidirectional
Scene reconstruction	Yes, local	No, needs to be extracted	Unidirectional
3D spatial interaction	Hand tracking	Hand tracking, mouse	Bidirectional
Audio communication	Yes, Skype	Yes, Skype	Bidirectional

TABLE 5 DIFFERENCES BETWEEN REMOTE AND LOCAL SUBSYSTEM FOR CO-LOCATED REMOTE COLLABORATION.

The major differences are the additional free camera with a mouse interface and that the reconstruction either results from a local source or a remote source, (Table 5). One system was designed that, depending on the configuration file, toggles capabilities on and off.

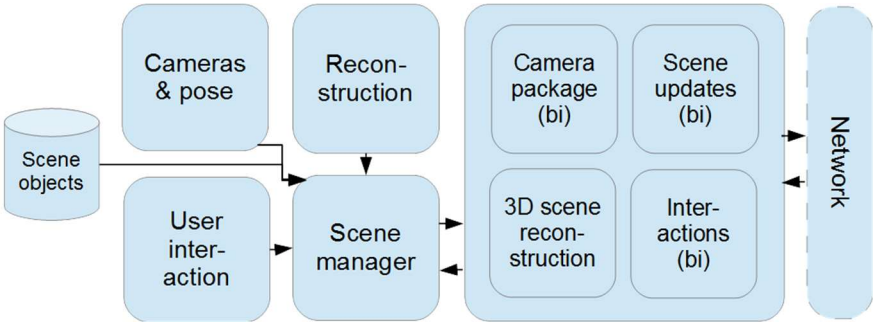


FIGURE 58 SCHEMATIC COLLABORATION SUBSYSTEM

The subsystems were sketched in the previous chapters. Figure 58 shows the breakdown of the syncing and network module.

Iteration (1)

While OGRE has multiple networking modules that would work, Raknet[22] proved to be the most straightforward to implement. This was used as the basis for our syncing module.

Cameras

The scene manager must constantly update the syncing module. The stereo rig images are piped through the scene manager and are compressed with JPG-compression to make them as small as possible without losing fidelity in the syncing module. The package with the left and right image was enhanced with the camera matrices. By keeping the package below 1500 bytes, with the UDP[23] header accounting for 8 bytes, the transfer is fast. Package loss is handled by tagging the packages. This is one-directional data and the unpacking works with the same module.

Scene updates

Objects are placed by the users in the virtual space. The objects that are visible, what the transforms are and whether they are active or not, need to be synced. A scene update package is created that is sent when something in the scene changes. Both users can update the scene.

---

[22] http://www.jenkinssoftware.com/, last visited June 2017
[23] http://www.sop.inria.fr/members/Vincenzo.Mancuso/ReteInternet/05_udp.pdf, last visited June 2017

3D reconstruction

3D scene elements do not update every frame; when a new key frame is added, the dataset expands and a package is created with updated coordinates. This data is only sent in one direction.

Interactions

When either of the users wants to perform an action, or wants to point out a detail, that user's intention must also be displayed on the other side. This is the tele-presence aspect of this system. Coordinates of mouse or gesture interactions require syncing. This module monitors the actions and sends packages when required.

Iteration (2)

The camera module was updated to work in two directions. From initial tests, it became evident that knowing where the remote investigator was with respect to the scene was important. The remote investigator has a free camera, and the position of that camera needs to be communicated to the on-location investigator.

The reconstruction data package was extended to include dense reconstruction and meshes, too. This prepared the pipeline to accept scanner-like data types.

Iteration (3)

Apart from knowing where the remote investigator was, the status of the remote investigator was added. Three modes were distinguished: viewing with stereo, a free camera, or not active.

5.8.2. EXPERIMENT

A test setup was conducted to technically validate the collaborative capabilities of the subsystem. A pre-created scene, as described in the interaction validation shown in Figure 55, formed the basis for the validation. Participants in a different room than the mock-up crime scene were able to create, move and select objects from the shared scene. The actions of the first participant were reflected on the screen of the second participant. Figure 59 shows the view of the participant wearing the head mounted device on the left; on the right, the screen of the participant using the free camera is shown.

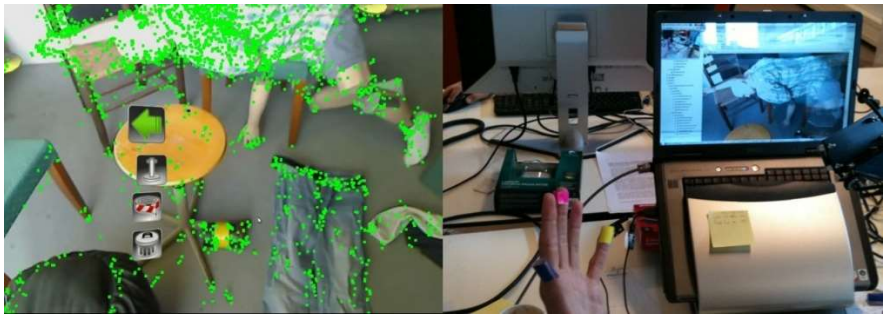


FIGURE 59 EVALUATION OF CO-LOCATED COLLABORATION

The goal of the experiment was to test the feasibility of the technical collaborative capabilities. A demo of the system was provided and the 'remote' laptop and head mounted device were placed in different rooms. Because this experiment concerned a technical validation, experts from the NFI were not required.

Participants

Three Ph.D. candidates from the TU Delft and one colleagues from the CSI The Hague project took part in the experiment. They were all familiar with the setup, and part of the inner circle of people familiar with the details of this research. To be able to cross validate, the participants switched roles after successfully placing and moving objects.

Measures

The goal of the experiment was to measure the technical capabilities of the system crucial to the collaboration described in requirements 11, 12 and 13. Three minor experiments were conducted;

- Audio experiment (req. (12)); are the `on-premise` and `remote` colleagues able to verbally indicate what they are doing.
- Placement experiment (req. (11)); are both `on-premise` and `remote` colleagues able to place a virtual pole in the scene.
- Cross selecting experiment (req. (13)); are the `on-premise` and `remote` colleagues able to select the pole placed by the other.

The author of this thesis communicated between the participants by walking between the two locations (one door apart). The first question was whether they could talk to each other and whether the audio was working, the second was

whether they were able to place the pole in the scene; first, the on-premise participant, quickly followed by the remote participant. The participants were then asked whether they could select the pole which had been placed by the other. If successful they traded places: first A, then B.

Because it was the technical feasibility that was being tested, only a 'pass' or 'no pass' was needed as a measurement. However, of course, the session was also conducted to gain additional insights that would help improve the setup, and those were noted.

Setup

As shown in Figure 59, a dummy doll was placed on the floor to represent a crime scene. In order to create a feature-rich scene, the doll was surrounded by a variety of objects. The processing laptop with the cables for the HMD was placed on a table next to the dummy. The hand-tracking in this experiment was done using the color finger sleeves (Figure 54) and the lighting in the room was controlled. The 'remote' participant was sitting behind the laptop, controlling the scene action with a mouse. No menu was necessary, as the only action possible was placing a pole. When the pole was approached, it highlighted and could be selected.

Results

The experiment was straightforward without surprises. The placement of the pole, a.k.a. the xyz location in the scene, was correctly communicated through the networks software. Basically, this is nothing more than what happens in a virtual reality game, when, if someone places an object in a 3D world, this can be manipulated by others as well.

What is different is the real-world map; the map is created on the fly while the other can see the virtual world being extended. The other difference is that the remote participant can see exactly what the HMD wearer is seeing.

Task	**1A**	**2A**	**3A**	**4A**	**1B**	**2B**	**3B**	**4B**
Audio (req. 12)	y	y	y	y	y	y	y	y
Placement (req. 11)	y	y	y	y	y	y	y	y
Cross selecting (req. 13)	y	y	x/y	y	y	y	y	y

TABLE 6 COLLABORATION EXPERIMENT RESULTS TABLE

There was one x/y in the table, which means the experiment needed to be redone. Otherwise, the experiment scored only y's (yes, succeeded). The rerun of two of the experiments was because the location of the virtual camera was replicated for the 'remote' collaborator: the HMD wearers were too active, so the remote collaborator could not select because of the rapid movement.

Apart from the sessions being successful, it was remarkable how, after just a few seconds, the remote participants started to adapt to the on-premise participant. They were obviously not in control, but seemed to be able to anticipate the movement on-premise quite easily. In the discussions afterwards, the participants talked about a "being there" experience. When asked "why" the answer was that, because they could reach into the world, they felt more connected than just watching video. Quick movements seemed to confuse this a little, either due to the refresh rate, blur, or too big of a disconnect

The experiment worked out well, and collaboration through this system was validated.

5.9. INTERFACES AND ENGINEERING COMPONENTS

In the previous chapters, the iterative development of the subsystems that were required to be newly developed were discussed. However, some requirements are not part of these subsystems. Subsystems that are important too, such as: requirement (10), logging and timestamping, and requirement (14), store raw data, still needed to be addressed.

5.9.1. RECORDING

An important requirement of the system is the recording capability. There are multiple reasons for recording, including, for example, the evolution of the scene, i.e. who did what and when. This information is managed by the scene manager. The stereoscopic video data is used for pose estimation and reconstruction. The source video is required to improve the technical performance and to be able to review what investigators could see. The image manager is responsible for handling the frames. The investigators influence the scene with gestures, hence, both from a technical performance improvement perspective as well as replay ability, recording is required (cf. Figure 60).

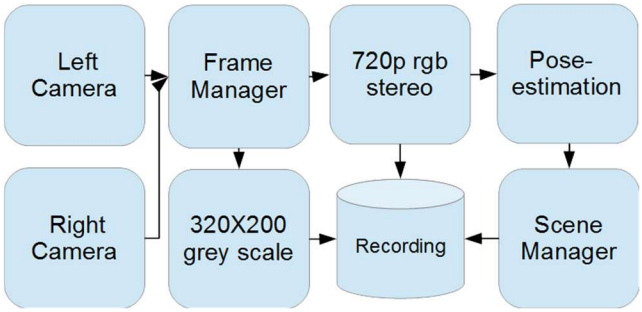


FIGURE 60 RECORDING SUBSYSTEM

The frame manager can directly tap into the webcams or play from the repository. Both pose estimation and hand tracking can be re-run with the image streams recorded, reproducing the same results, which can be validated by the scene manager recording. The data is stored with time stamps and version control.

5.9.2. TOOLS

To be able to have a minimal viable system, a few tools need to be present as well as barebone system functionality.

The system needs to be initialized within 30 minutes (requirement 4), and a user needs to be able to start and stop the system and to start and stop recording, as is discussed in chapter 5.9.1.

Collaboration also requires a few tools. The system therefore needed basic operational tools. The following tools were implemented: system settings, connect to network, feature overlay, marker in scene and rendering occlusion, (requirement 6).

In order to have functionality in the scene, a few tools specific to crime scene investigation were implemented.

- Marking zones with poles and ribbons
- Attach notes to the scene
- Bullet rods for trajectories

The poles snap to the reconstruction and ribbons appear between the poles. The 'up' direction of de poles was established by the scene up direction. The bullet rods also snap to the reconstruction and can be elongated.

5.9.3. 3D USER INTERFACE

The research discussed in chapter 4.5.2. clarified that the number of user interface widgets in the scene should be minimal, as they increased the number of mistakes. The interface is operated by gestures and the hands need to be able to select/invoke the tool desired.

No official experiments were done to design and iterate on the best possible interface. When developing the system, a few iterations on the interface were done. The most relevant findings are listed below.

- Initial tests with a fixed location interface (aka windows menu) in stereoscopic view proved to be cumbersome. Participants found grasping a GUI component with their hand, while their heads were not perfectly still, difficult. This is logical: the head, environment and hands all have freedom of movement with respect to each other. It is easier to place an interface component near the location that requires the tool and lock it to the scene instead of on the peripheral of vision.
- Our hands are used to working together, and looking at the hands gives the same depth/location cues that looking through whatever HMD does.
- A minimum and maximum size was determined that would neither block a large part of the scene nor be too small to recognize.

To achieve these, users can place the interface widget anywhere in space based on a left-hand thumb up recognition; see Figure 61.

FIGURE 61 GRAPHICAL USER INTERFACE OPTIONS MENU

A menu surrounding the hand will appear when the left-hand thumb-up is detected: the menu will stick to the hand and be locked in space until it no longer detects that posture. Hence, when the thumb points downwards, options can be selected. The right hand, as a pointing device, is used to select objects in the virtual scene. Effectively, recasting is used to determine which scene point should be interacted with.

5.9.4. RESOURCE & INTERFACES

Although the subsystems have now already been designed with scene manager communication criteria, the components do not yet work as a single system. Hence some adjustments had to be made

The dedicated machine has 4 CPU kernels, and 4 demanding processes must run simultaneously. The following processes were distinguished, each with a dedicated core:

1. Tracking, the image analysis to match images for pose estimation;
2. Mapping, the creation of a map of the environment;
3. Hand tracking, being able to detect and analyze the position of the hand for gesture recognition;
4. Stereoscopic rendering and scene management.

Another integration step is related to the hand gestures. A hand close to the camera might block ~25% of the screen, while the tracker needs to track the environment correctly relying on that imagery. A feedback loop was created from the hand tracker to the pose estimator to mask the pixels that were related to detected hands. Features detected on those pixels are ignored for pose estimation.

5.10. REQUIREMENT VALIDATION

Chapter 5 is focused on the design of a system that satisfies the requirements derived in chapters 2 and 4. The previous sections of this chapter iteratively completed the system that started out as a coarse outline in chapter 3.

The following table briefly explains how the requirements were addressed.

	System requirements			
Nr.	*Description*	*Explanation*	*Full-filled*	***Section***
01	*The system must be able to acquire and store, spatial oriented metric 3D data from a pristine environment.*	*A stereo vision based localization and mapping subsystem is created.*	*Yes*	***5.6. & 5.9.***
02	*The system should allow an investigator unhindered view of the crime scene.*	*Because of the difficulties of see-through head mounted devices, a fall back with a video see-though device was decided upon. When technology improves, this requirement can be fulfilled*	*No*	***5.5.***
03	*Allow the system to share information to the investigation team during or shortly after acquisition.*	*The networking subsystem shares the data; data is shared instantaneously.*	*Yes*	***5.8.***
04	*The time between setting up the system and the start of using the system for geometry capture should be less than 30 minutes.*	*Initialization of the system amounts to little more than booting Linux and launching the application requires less than 5 minutes.*	*Yes*	***5.5.***
05	*The system must be able to align acquisition data without disturbing the pristine characteristics of the scene.*	*The tracking and mapping module is vision based and does not require any physical preparation. Both audio and video can be recorded and are time stamped for alignment.*	*Yes*	***5.6.***

06	*The data which represents the 3d structure of the scene needs be presented to the users as surface data.*	*A mesh reconstruction module creates a textured mesh.*	*Yes*	***5.5.***
07	*The acquired data from the different sensors (sound, imagery, measurements) need to be spatially indexed and fused into a global 3D model.*	*The coordinate system of the acquisition is the same coordinate system that aligns everything else.*	*Yes*	***5.6.***
08	*The system must be able to differentiate and visualize the regions that are mapped by multiple investigators.*	*Every mapping exercise has identification information*	*Yes*	***5.9.***
09	*The system must have low latency when team members interact.*	*The system runs ~25 Hz.*	*Yes*	***5.5.***
10	*All steps in the process to acquire spatial 3d data need to be logged and time-stamped.*	*The data can be called upon based on its time-stamped storage.*	*Yes*	***5.9.***
11	*Enable spatial collaboration by creating common ground in the form of a 3D model between domain experts and on-location investigator.*	*Both collaborators have access to the same environment at the same time.*	*Yes*	***5.7. & 5.8.***
12	*Enable spatial collaboration by enabling conversation between domain experts and on-location investigator*	*Skype is used to establish this connection.*	*Yes*	***5.9.***
13	*Enable spatial collaboration by enabling 3D interaction between domain experts and on-location investigator.*	*The 3D reconstructed scene and the interactions are reflected in the same 3D environment for the collaborators. With this capability, they can jointly interact with the*	*Yes*	***5.7.***

		scene, and thus collaborate.		
14	*Be able to capture and store raw data.*	*The unprocessed image data is stored, YUV images.*	*Yes*	***5.9.***
15	*The system is not allowed to induce contamination of the scene*	*The system is mobile, no physical interaction with the scene is required.*	*Yes*	***5.5.***
16	*The equipment's weight should not exceed ergonomic guidelines.*	*The recommend max weight of 225 grams is not exceeded.*	*Yes*	***5.5.***
17	*The system is not allowed to interfere with the investigation.*	*It can be turned off; the hands and movement of the investigator are not restricted. However, human computer interaction is necessary for the use of the system, possibly distracting from current work.*	*Parti al*	***5.5.***
18	*The system needs to use head mounted displays for digital overlay*	*A stereo head mounted display was created.*	*Yes*	***5.5.***
19	*The system senses with technologies that function in the visible light and infrared light.*	*Regular RGB cameras are used.*	*Yes*	***5.5.***

TABLE 7 REQUIREMENTS VALIDATION

All except one requirement has been validated in this assessment; see Table 7. The hardware for optical see-through head mounted devices with enough resolution, field of view, and refresh rate was unavailable at the time the experiments were performed. Furthermore, the optical challenges of aligning the virtual world with the digital world per each user's specifics is time consuming (Caarls & Pieter, 2009).

Most of the requirements are not out of the ordinary for embedded hard- and software systems; logging, discerning user information and communication are a part of many systems. However, real-time mapping and aggregation of information required a rethinking of the traditional use case pipelines. Controlling a user interface for head mounted display tasks as reflected in requirement 13 appeared to be simple at first, but ultimately became one of the more difficult requirements to tackle, requiring many iterations.

Connecting peers with a common ground 3D map required a rethinking of information sharing. Classical systems, i.e., 3D games, update status but do not share new 3D information, and classical telepresence information systems sync system updates.

5.11. TECHNICAL EVALUATION OF SUB-RESEARCH QUESTIONS

With the creation of the mediated reality system in chapter 5 and the subsequent requirements validation, the focus can now turn to the sub-research questions. The rest of this chapter will be devoted to the sequential discussion of the sub-research questions.

5.11.1. ARCHITECTURE

A major contribution of this thesis is the generic architecture developed to answer the research questions. Although the backbone of the architecture is based on an established open source game engine, it was necessary to develop a great many extensions to fulfill the requirements. The most imported aspects of the architecture will be discussed in this chapter and linked to the sub-research questions.

For the sake of convenience, the sub-research question is stated below, together with the generic architecture; see Figure 62.

A) What architecture allows for collaborative spatial interaction in mediated reality?

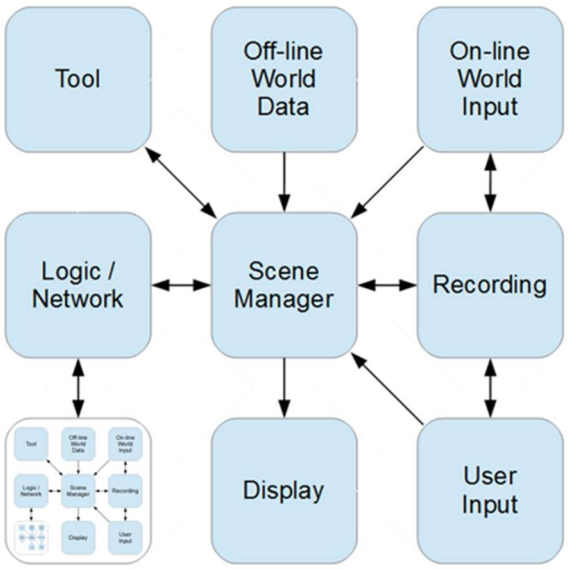


FIGURE 62 HIGH-LEVEL ARCHITECTURE

Chapter 3 introduced the generic architecture depicted in Figure 62. As stated in chapter 4.1, many systems, in some way, offer a solution to this question. However, none of the systems could answer the question and at the same time satisfy the domain specific requirements, which meant that a new system had to be developed.

The architecture was required to support collaboration, spatial interaction and mediated reality:

- Spatial interaction: The bottom right node of Figure 62 shows that the user can control spatial interaction with the scene manager. A gesture based subsystem was designed that allows the user to place 3D objects in a virtual scene with bare hand gestures. A more detailed description of the technical implementations and testing can be found in the literature (Akman et al., 2013; Lukosch et al., 2012). The system is generic enough to allow the input to change from hand gesture recognition to mouse and keyboard input, which is used for remote interaction.

- Collaboration: The bottom left node in Figure 62 shows that collaboration is supported. The full architecture can be cloned as many times as required and modalities to the interface changed. One of the

systems will be a “master”, the rest of the systems will be slaves. The master system collects the mapping information and distributes updates to the slave systems. Two-way communication was established for scene updates like tools and people acting on the digital overlay of the scene, including the state of people. All concurrent users are visualized as active (or not) and their interactions with the scene visualized and distributed. If any of the users point or move anything, this is visible to all. Furthermore, an open audio communication line allows for additional communication.

- Mediated reality is created by two nodes from Figure 62. The top right corner creates a real-time digital copy of the scene and a head mounted display takes care of digital overlays with the lower middle node. A real-time map of the environment is created with a stereo camera setup and used to merge the virtual objects with the real environment. The eyes of the user are represented as virtual cameras in the scene and provide stereo depth view of the scene with a real-world perspective. This architecture allows multiple collaborators to jointly see the same environment, influence the digital overlay of the environment and communicate with multiple means.

The keystone component of this architecture is the scene manager, visualized in the middle of Figure 62. The scene manager integrates all other components. The scene manager communicates the data from real-time mapping inputs, to user interaction to scene layering with updating 3D model data. Multiple video input streams, sensor based meshes/point clouds and multiple user interactions were required to significantly extend the default capabilities of OGRE.

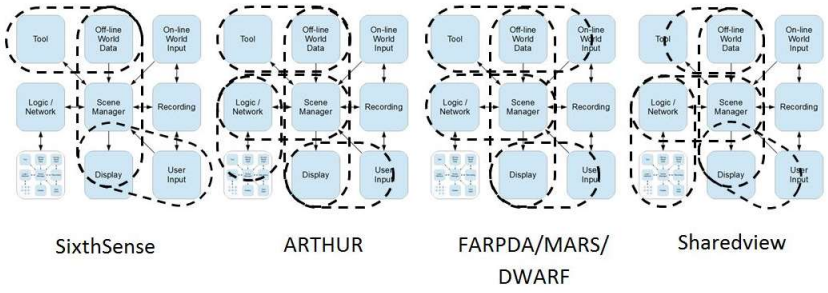


FIGURE 63 MAPPING COMPARABLE ARCHITECTURES

Figure 63 depicts comparable architectures, selected in chapter 4, and mapped against the generic high level architecture created in this thesis. The flexibility of the high-level architecture is illustrated by the fact that all the researched architectures map onto it. The mapping is natural, although the half touching circles require explanation, which can be found in section 4.1. In summary, the sixthSense has a hand-held display, ARTHUR lacks 3D tracking, Sharedview has a fixed camera and FARPDA/MARS/DWARF have less sophisticated 3D interaction or no 3D tracking.

With the creation of this architecture, sub- research question (A) is answered and satisfied. An architecture has been created that is flexible and easily allows for extensions in collaborative mediated space.

5.11.2. ON-PREMISE INTERACTION OF THE DIGITAL OVERLAY

Sub-research question (B) invokes nearly the full capability of the system; from gesture interaction to tracking and mapping. To recapitulate, sub-research question (B) asked:

B) Does the architecture support an on-premise user with meaningful interaction of the digital overlay?

Figure 64 depicts a scenario, which is a subset of the scenario used in chapter 5.8.2. The test subjects needed to place a restricted area ribbon in the scene. This action invoked the use of almost the full system architecture; only collaboration and remote interaction were not required.

- User Input; gesture interfacing with the scene
- Tool; tool interacting with the scene
- Off-line World data; models of manipulatable objects.
- On-line world input; tracking and mapping of the scene
- Scene manager; assemble all scene elements, act on the scene and send to HMD.
- Display; the HMD

The poles are virtual 3D objects, 3D modeled and textured with code created ribbons between the poles to form a barrier. The scene itself is tracked and a map of the scene generated, a plane is fitted to the scene, thereby establishing a coordinate system to work on. Because the tracking is from a stereo setup, the scale of the scene is known, thus the correctly scaled 3D object fits right in. The placement of the poles is established by gesture movements, as described in chapter 5.7 and elaborated on in Lukosch et al (Lukosch et al., 2012). The

gestures trigger intersection rays with the mapped scene, making it possible to snap objects to the scene. The composition of the scene is handled by the scene manager and rendered as stereo imagery to the HMD. The HMD's cameras are compensated by offsetting the image data to correspond to the displays in the HMD and by matching the field of view. Figure 64 shows the composited scene for the right eye of the HMD wearer.

FIGURE 64 AUGMENTED ON-PREMISE SCENE

The setup was created at the FFL in their new test facility. A more thorough explanation is provided in chapter 5.7 and chapter 5.8 The utilitarian aspects are folded into the task; correctly fencing a scene is a scenario discussed in chapter 2. It is normally a physical task that potentially interrupts the pristine aspects of a scene. By completing the task of putting virtual poles in the scene (Figure 64), and selecting them for removal, all the aspects requested in sub-research question (B) have been answered, and hence validated. In other words, the architecture supports the on-premise user with utilitarian interaction of the digital overlay.

5.11.3. REMOTE INTERACTION OF THE DIGITAL OVERLAY

Sub-research question (C), repeated below, is addressed and answered in the following section.

C) Does the architecture support a remote user in interacting with the digital pristine environment?

The previous section did not address the remote aspects of the architecture. In this scenario, the scene needs to be streamed to a remote participant to validate the requested interaction. Figure 65 shows the same view as in Figure 64, but from the remote expert's perspective.

FIGURE 65 AUGMENTED SCENE FOR REMOTE PARTICIPANT

The imagery from the staged crime scene from chapter 5.8.2 is streamed through the network, and the coordinates of the poles are used to place and extract the local models. The network packages are partly bi-directional; the updated pole positions are communicated both ways. When the remote expert positions a pole, the coordinates and the ID of the object are communicated to the on-premise participant and placement can be observed in both instances. The remote participant is able to use the mouse and laptop to place poles in the scene, thereby updating the scene. Both selection and placement can be witnessed by both participants.

Various data types are communicated to the remote user:

- Stereo video streams including pose estimation and frustums packed together. The remote expert can potentially switch to an HMD too, although a regular 3D view or a view such as in the picture, via the left camera, is also possible.
- 3D scene and scene updates. Following the initialization of the map, the scene is regularly updated after the map maker creates a new key frame.

The points or polygons are added to the scene information with every key frame.

- Active tool, virtual cameras, object locations and live objects. This communicates the activities in the scene for both to see.

This setup is validated in chapter 5.8.2. In the above section, additional background is provided together with imagery based on experiment 2. In this chapter, the remote user interaction with a digital pristine environment is validated and supported by the designed architecture.

5.11.4. COLLABORATION BETWEEN THE ON-PREMISE AND REMOTE USERS

This section addresses and answers sub-research question (D), repeated below.

Does the architecture support spatial collaboration between a on-premise and a remote user?

Chapter 5.8.2, 5.11.2 and 5.11.3 provide the background for this section. Here, specifically the collaboration between on-premise and remote users is addressed.

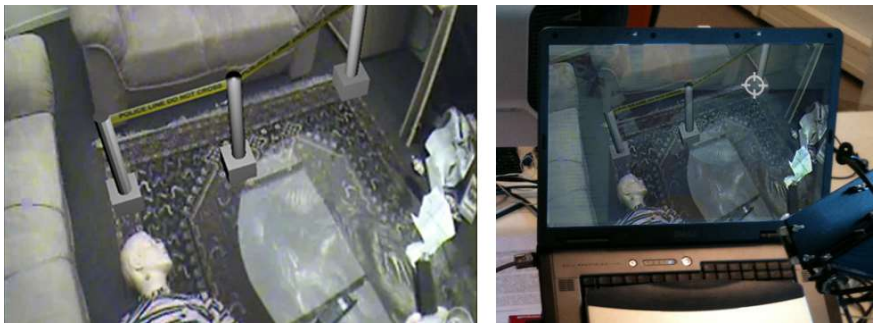


FIGURE 66 VIEW OF REMOTE (RIGHT) AND ON-PREMISE PARTICIPANTS (LEFT)

Looking at the same scene from two different rooms, Figure 66 depicts the perspective of the on-premise user and the remote user. In Figure 66 (right) the user of the laptop is active and is busy placing the third pole. As elaborated on in chapter 5.8.2, the audio communication worked well and the participants took turns placing poles in the scene.

The architecture supports collaboration between the on-premise and remote users by allowing the users to connect through audio, offline and online 3D data and 3D tools.

5.12. CONCLUSIONS

The V-model structures the development of the system into 3 levels: the systems level (1), the subsystems level (2), and the functions level (3). Research was required for four subsystems. The V-model in these cases was augmented with Boehm's iterative development cycle.

The four subsystems that were required to be newly developed were: (1) the see-through rendering system, (2) simultaneous localization and mapping system, (3) 3D scene interaction, and (4) collaboration between users. The subsystems were built on the same software kernel, namely the OGRE platform. The systems have been designed to supply the data to a central hub and are loosely coupled, working on different CPU threads.

In chapter 5.5 and 5.6, a head mounted display was created that allows the digital overlay of physical environments. It is capable of autonomously mapping pristine environments and supports the gesture control needed for the requirements formulated in chapter 5.7. In chapter 5.7, the interaction paradigms subsystem for controlling a head mounted display was designed. In chapter 5.8, the subsystems were bound together, to enable co-located scene intervention. The development of the system was finalized in chapter 5.9, which wrapped up with a discussion of the subsystems needed to complete the requirements.

With the engineering done in chapter 5.5 to 5.9, the requirements were validated in chapter 5.10. All except one requirement was fulfilled; a video see-through HMD was selected instead of a non-intrusive optical see-though HMD. The complexities of an optical see-though HMD proved to be too daunting and hardware in that space is evolving rapidly.

The sub-research questions were validated in chapter 5.11. Individual users on-premise and off-site were able to interact with the virtual overlay in the pristine environment. Communication between both types of participants was validated, allowing them to collaborate. This chapter concludes the creation of the artifact required to answer the main research question.

6. EVALUATING MEDIATED REALITY SUITE

Up to the present chapter, the emphasis has been on building the mediated reality artifact, with a predominate focus on the technical aspects. The purpose of this chapter is to be able to answer the main research question. Now that we have the mediated reality system created in chapter 5, the foundation for validation is present. In the next sections the experiment, the setup and evaluation methods will be described, followed by the evaluation and conclusions. The results of this chapter are also discussed in the conference proceedings of the Computer Supported Cooperative Work (Poelman, Akman, Lukosch, & Jonker, 2012).

Although the individual components of the system have been validated and discussed in previous chapters, the summation of the parts still require intense handholding. With such fragile and new technology, the qualitative aspects of this research outweigh the quantitative results.

6.1. INTRODUCTION TO THE EXPERIMENT

As introduced in chapter 1, the emphasis of the main research question is collaboration, which, as yet, has gone unaddressed. In chapters 5 and 6, the pristine environments, 3D interaction and mediated aspects were validated. As stated in the main research question, repeated below for the sake of convenience:

1) ***How can we support collaborative spatial interaction in a pristine environment applying mediated reality?***

The premise of this chapter is to validate whether a remote expert can aid a novice on location using this mediated reality system. To be able to measure this type of support, a scenario and experimental setup must be created and it must be measured to validate success. As crime scene investigation is the main use case, the scenario should fit a specific crime scene scenario. To measure remote collaboration, the expert and novice cannot be in the same location, the expert must have more information than the novice and the system should mitigate their not being in the same location within the constraints given. For the "CSI The Hague" project and in agreement with the project participants, a multi-purpose scenario environment was created (Figure 67).

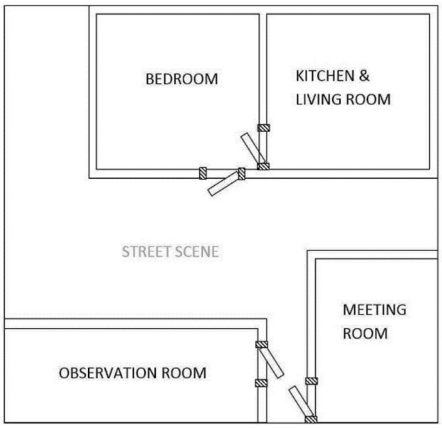


FIGURE 67 FORENSIC FIELD LAB (FFL) SCENARIO SPACES

The kitchen and meeting room are a natural fit for a remote scenario, because the rooms lack sound or visual communication capabilities. Furthermore, a scenario is required that showcases an information exchange that benefits this type of collaboration, in which the delta in information is effectively communicated through the mediated reality system. The information exchange through the system is depicted in Figure 68.

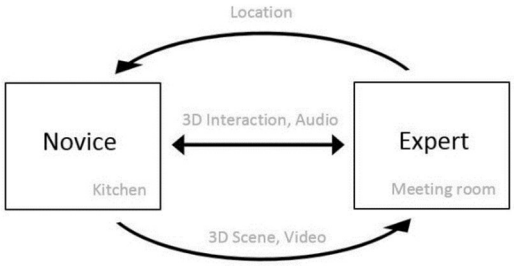


FIGURE 68 SCHEME FOR INFORMATION EXCHANGE

Through the system, the expert can communicate by means of audio, 3D interaction and his location in the virtual space. He must obtain information through the 3D scene, the 3D interaction, audio and a stereo video feed. The novice must provide as complete an overview of the scene as possible through 3D interaction, audio, the 3D scene and stereo video.

The effectiveness of the information exchange must be measured to answer the research question. This requires a methodology that provides insight on how good collaboration is supported. Several distinctive processes are important

when collaboratively solving tasks, such as the exchange of knowledge that is relevant for the task at hand, argumentation processes, coordination processes and motivational processes. Through these dimensions, symmetric individual contributions are measured to provide complementary aspects of collaboration. Considering the system from a communication bandwidth perspective, the communication can be said to have been widened with a spatial component.

Communication processes are important to ensure a common referential within a group of collaborators. Establishment of common ground in collaborative processes in which co-workers mutually establish what they know is critical to the artifact, and is necessary for proceeding with the task at hand. Important characteristics of collocated synchronous interactions are assumed to support grounding (Burkhardt et al., 2009). Multimodal channels for communication allow for multiple ways to convey complex messages and provide redundancy. Furthermore, a shared local context allows for mutual understanding. A balance in the roles of the participants in communication, group management and task management is considered a good indicator for collaboration (Spada, Meier, Rummel, & Hauser, 2005). In their view, the quality of collaborative learning is linked to the symmetry of the interaction. This standpoint has been adopted for the purpose of this research.

Technology supported collaboration is rarely tested in a multidimensional and generic manner. And particularly as the systems offer additional communication possibilities which were previously not available, an evaluation method is needed that measures a broad range of collaboration aspects. Therefore the research conducted by Spada (2005) in the closely related domain of computer supported collaborative learning (CSCL) has been partly adopted, as this was developed to compare and assess collaboration in collaborative learning tasks with a wide range of collaboration dimensions. Spada (2005) distinguishes nine qualitatively defined dimensions that cover five broad aspects of the collaboration process, namely: communication, joint information processing, coordination, interpersonal relationship and motivation. However, according to Burkhardt et al (2009), indicators exploited by observers in order to assess collaboration are underspecified. Their method relies on the subjective evaluation of 7 dimensions (Table 8) on a 5-grade Likert-like scale supported with a training session (Burkhardt et al., 2009). The drawback to their approach is that it fails to capture the observables from the collaborative situation (a bird's eye view). As a result, it is not possible to track back the assessment value to the original data. We modified the assessment procedure to make observable indicators underlying the evaluation explicit.

Dimensions	Definition	Indicators
1. Fluidity of collaboration	It assesses the management of verbal communication (verbal turns), of actions (tool use) and of attention orientation.	Fluidity of verbal turns, Fluidity of tools use (stylet, menu), Coherency of attention orientation
2. Sustaining mutual understanding	It assesses the grounding processes concerning the artefact (problem, solutions), the investigators actions and the state of the mediated reality disposal (e.g., activated functions).	Mutual understanding of the state of the analysis, Mutual understanding of the actions in progress and next actions, Mutual understanding of the state of the system (active functions, open documents)
3. Information exchanges for problem solving	It assesses analysis ideas pooling, refinement of ideas and coherency of ideas.	Generation of analysis ideas (problem, solutions, past cases, constraints), Refinement of analysis ideas, Coherency and follow up of ideas
4. Argumentation and reaching consensus	It assesses whether there is argumentation and decision taken on common consensus.	Criticisms and argumentation, Checking solutions adequacy with analysis constraints, Common decision taking
5. Task and time management	It assesses the planning (e.g. task allocation) and time management.	Work planning, Task division, Distribution and management of tasks interdependencies, Time management
6. Cooperative orientation	It assesses the balance of contribution of the actors in analysis, planning, and in verbal and graphical actions.	Symmetry of verbal contributions, Symmetry of use of graphical tools, Symmetry in task management, Symmetry in analysis choices
7. Individual task orientation	It assesses, for each contributor, its motivation (marks of interest in the collaboration), implication (actions) and involvement (attention orientation).	Showing up motivation and encouraging others motivation, Constancy of effort put in the task, Attention orientation in relation with the analysis task

TABLE 8 DIMENSIONS AND INDICATORS OF OUR METHOD, ADOPTED FROM (BURKHARDT ET AL., 2009)

The selected method is based on Burkhardt et al. (2009), which again is based on Spada's ideas (2005). It extends the assessment procedure by having experts observe and evaluate the collaboration target (Burkhardt et al., 2009); a design task, but one that is general enough to be adopted to the spatial analysis task, which uses the dimensions explained. The questions on the seven dimensions of this scale reflect the essential humanist aspects of collaboration. With the extended sharable information that is provided by the system, this should show.

For each indicator, the questions are balanced with positive valence and negative valence; see Appendix III. To obtain a good spatial analysis, the questions distinguish between "good" collaboration and "low quality" collaboration. A sample positive question relating to dimension 3 "Information exchanges for problem solving" is: Do the investigators improve the spatial analysis by having their shared ideas available, pooled and coherent? Or the negative version: Is there miscommunication because refinement during the analysis takes place and is pooled? In total, 28 questions were formulated for the seven dimensions. The questions were formulated to fit the spatial tasks. The same set of questions was given to the on-premise participant and the expert; furthermore, observers were also used to acquire additional data. The interest, of course, was in seeing whether there is coherence on the questions for all participants and what the qualitative interviews with the observers would bring.

To be able to acquire an as complete as possible understanding of the collaborative aspects, interviews with all the participants were conducted. The questions from Burkhardt et al. were used a guide.

6.2. EXPERIMENT

The goal of the experiment was to validate whether the participants felt supported by the created system. As suggested in the previous chapter, the social aspect plays a major role in judging the system's success.

Experiment details

At the FFL a staged crime scene was created to facilitate the experiment. Two separate rooms were used to conduct the experiment: (1) a room (meeting room, Figure 67) containing the remote user, who was monitored by the observers, left in Figure 69, and (2) a physical crime scene for the wearer of the HMD (kitchen, Figure 67), right in Figure 69. The design of the setup was similar to the setup created in section 5.8.2.

FIGURE 69 EXPERIMENT SETUP

Participants

Three types of volunteers participated in the experiment. (1) Observers, whose goal was to observe the collaboration and to provide feedback in the after-action reviews, just like the other participants. (2) On-premise participants, who wore the HMD and collaborated with the remote expert, and (3) Remote experts, who interacted with the HMD wearer and provided remote guidance. The group of participants consisted of volunteers from the NFI, experts from other CSI The Hague project, the author and a colleague from the TU Delft. The group consisted of nine males, aged 25-55.

Measures

Three types of measures were used: (1) all the system data was logged during the experiment, except the audio through skype, (2) the questionnaire with the 27 questions, and (3) an after- action review, with notes. Four participants, in turn, wore the HMD for the full experiment. To enable the participants to be able to validate the scenario from both a novice perspective and from an expert perspective, the tasks were swapped after the first 2 rounds in the following order of participants: A, B, C and D; A-B, C-D, D-A and C-B. No external recording was required, as the video steam and actions were recorded, providing a complete overview.

Setup

The on-premise participant received the mediated reality HMD and was positioned in room 2. The remote participant was placed behind a large TV screen in room 1, where he could see the video stream from the head mounted device, the 3D model being created by the map maker and the interactive

elements put there either by himself using a generic keyboard and mouse or by the on-premise participant using our gesture-based interface. The remote expert and the on-premise participant could also communicate via Skype. From behind the expert, the observers could see both the actions of the experts and of the on-premise participant through the TV (left side, Figure 69).

The artifact is a complex prototype so the tasks were kept simple, to ensure the success and repeatability of the experiment. The following three tasks needed to be accomplished:

1. Mapping the pristine environment;
2. Tagging a specific part of the scene with information tag;
3. Using barrier tape on poles to spatially secure the body in the crime scene.

FIGURE 70 TASK FOR THE EXPERIMENT OF THE MEDIATED REALITY SYSTEM AT THE FFL

The experiment started with the on-premise participant stationed in the door opening. By walking a few steps and looking to the left, a 3D map was easily created (right side, Figure 58). When the interactive map visualized for the remote participant, he was required to guide the on-premise participant to stick a note on the faucet to secure for fingerprints and to digitally secure the body by placing barrier tape on poles.

No information was provided to the on-premise participant, who had to wait in the door opening for instructions. The remote expert was given two goals: to secure the faucet for fingerprints and to place barrier tape in a professional way around the body. Prior to the start of the experiments, the participants were given a mini training to familiarize themselves with the system.

After the sessions ended, the participants were given the time to fill in the questionnaire. The same 7 topics (Burkhardt et al., 2009) were used to discuss

the questionnaire results as an after action review. The following chapter discusses the results.

6.2.1. EVALUATION QUESTIONNAIRE & FEEDBACK

Because statistical significance relevance is negligible and qualitative results are more meaningful for this research, quotes from participants during the experiment are used to emphasize relevant aspects. The seven collaboration dimensions shown in Table 8 are used to structure this evaluation. The quotes have been translated from Dutch, interpreted by the author of this thesis and represent thoughts from the novices, experts and observers in their mixed roles.

Fluidity of collaboration

The first quote from an expert was: *'What's the protocol for interaction?'* From his perspective, there was no clear protocol in place to support the fluidity of verbal turns, except by interacting on the audio channel. Although the 3D data was visualized and the actions of the on-premise participant visible, the trigger for starting to talk was obviously missing.

Nevertheless, the attention of both the remote participant as well as the on-premise participant was rated as high; the gaze of the on-premise participant neatly guided the shared focus. One observer noted: *'Having the expert present in the scene as a virtual camera with tracked hands would give him an identity in the scene'.* The actions themselves were coordinated well and the verbal communication dominated the actions.

One on-premise participant in the experiment commented: "*Just like in a regular phone conversation, who speaks when is well orchestrated and when this is coupled with seeing what the other is doing "physically", it enforces ease of communication.* The remote participant was allowed to toggle view modes; when it came to collaboration, the same mode was picked by all remote participants. The choice was between 3D views, on-premise participant view or 3D view with on-premise participant camera view in the top right corner and vice versa. The participants all picked the 3D view with the camera in the right corner. As a remote participant pointed out*: 'Not seeing what the layman is seeing severed the intuitiveness because orientation was disconnected'.*

Sustaining mutual understanding

The on-premise, observer and remote participants all gave this this dimension a high rating (strong agree): the actions were visible to both participants and the

visual feedback provided a clear mutual understanding of the state of the interaction. The system provided visual feedback on active and non-active items that were visible to the users of the system. One observer remarked: *'The expert had a better overview due to the free 3rd person view he has, i.e. he was virtually able to walk around in the obtained 3D environment'.* Another remark from a remote participant was*: "What's not to mutually understand? He can see what I'm doing and I can follow his every move.*

Information exchange for problem solving

The limited tasks did not provide a clear understanding of this dimension (neutral or not used). The participants jointly followed up on their ideas but no clear problem solving was involved in the tasks. An observer mentioned: *'I can imagine that with some tasks like completeness of mapping or guidance through complex scenes joint problem solving can be achieved through this system'.*

An on-premise participant mentioned that *'Like in the CSI The Hague video24, a remote participant could look at a multitude of sensor channels and therefore has different information than me'*, and *'I don't feel comfortable looking at the overview of the map, it distracts me so the bird's eye view that the remote participant has can guide me'.*

Argumentation and reaching consensus

The common consensus was unanimously rated with a "strong agree" by the participants and the observers. And as remarked by the participant, *'The strong visual feedback quickly leads to consensus, as the system supports instant feedback.* The observers agreed that*: 'The instant visual feedback on hypothesis testing with spatial interaction is very convincing'.*

The collaborative discussions tended to be task related, as a remote participant remarked: *'Don't put the poles so close to the couch, we need more space'.* The follow-up remote participant remarked that: *'Because you are viewing directly what he is doing, as if you're next to him, you can easily predict ahead, which is super easy to for creating consensus'.*

[24] Official CSI The Hague video http://www.csithehague.com/, last visited February 2016

Task and time management

All participants agreed (mid/high scores) that the task division and planning followed naturally from the different tasks in the experiment. During the third task the remote participant, e.g., simultaneously worked on the barrier tapes and poles. According to the observers, this was really interesting: *'As these clearly showed how one can work on a scene jointly as spatial interaction'.* Basically, the on-premise participant put the couch-kitchen barrier in place at the same time as the remote participant put the barrier on the other side in place. However, this was the only time it was jointly done; in the other sessions, the task was performed sequentially.

An observer remarked, *'When it was my turn I did not realize that I could also start doing things, this is so new'.* He asked the person who carried out the task simultaneously how he knew how to do that, and this person replied: 'It's a virtual world and my tools are active.

From a task perspective, an observer remarked: *The remote participant needs a good map first and because of how the experiment was laid out the task could be done with minimal mapping, the dependency is mostly on a good base map'.*

Cooperative orientation

Except from the initial discussion on the interaction protocol, there was clearly symmetry in the experiment. The remote participant remarked that*: 'The same visual focus aligned their cooperative effort',* although the on-premise participant subsequently noted that: *'Knowing that somebody can see exactly what you see and not vice versa is slightly stressing '.* However, all voted for a "strong agree" on this dimension.

A curious thing occurred when an on-premise participant wanted to block a pole that the remote participant was attempting to move – which, of course, was not possible. Standing in front of the pole did not prevent the movement at all. This is interesting: the digital world is shared, but not the physical one; all the rules of the digital world apply. For the on-premise participant, the digital and the physical are one.

An observer pointed out there was symmetry distortion because of missing technology*: 'A remote expert was pointing at the faucet to put a marker on and saying' there'',* while, of course, the finger direction was not noticeable to the on-premise participant.

Individual task orientation

It was obvious that both participants motivated each other during the analysis. All participants gave this a high rating. Again, the on-premise participant remarked that: *'Being observed all the time results in a slight discomfort'.*

The tasks were clearly separated: the remote participant had the information on what needed to be tagged and the on-premise participant needed to provide the information for the task to happen. An on-premise participant remarked*: 'I feel pretty relaxed, my task is to be the eyes and ears, I just have to wait for what's next'.* And, during the placement of the poles, he remarked: '*It feels like a hotline for support'.*

In conclusion, the questionnaire provided insight into the experiment on the 7 dimensions. The scores were all positive, with the majority of specific information coming from the remarks during the sessions.

6.3. REFLECTION

Because of group of participants was so small, the qualitative aspects of the experiment are extremely important. In the discussions with the participants and observers afterwards, the following discussion topics arose.

6.3.1. PRIVACY

According to the on-premise head mounted display wearing investigators, there was something intimate about sharing what they saw with the remote expert. In a pure virtual reality environment, everybody is equal and has the same limitations. In the case of a mediated reality environment, you share what you see at all times. It apparently does not feel like "being a 3D scanner"; there is a feeling of being observed.

A lot of the crime scene investigators that are the first to arrive on a crime scene handle the case from experience. When they are asked to write a report about their ongoing investigation they are generally incomplete in their reporting (per the experts). Quote: *'When they are asked to write about researching a certain trace they will tell you but hardly write that down'*. They prefer not to be bothered with desk work and the team must place its trust in their experience. With the recordings of this system: rewards, mistakes, hours spent, etc. - all is transparent. The feedback was that this feels dis-humanizing.

Recording the investigators' every activity can be perceived either positively or negatively. It is no longer necessary to write everything down, as every move is

monitored, but the fact that everything is being recorded is perceived as a tool to monitor the way investigators perform their job, making resistance inevitable. If the experts are to be believed, only new recruits are likely to accept the change in procedures without problems. In the Netherlands, a pilot was carried out in 2009 in which the police wore video monitoring equipment all day, which met with quite some resistance.

6.3.2. PRESENCE

During our experiment, it became clear that the presence of the expert was not made sufficiently explicit. This was particularly an issue when the expert started creating or moving objects in the scene, without ensuring that the layman was fully aware that he was doing this. It is evident that visual awareness indicators to aid are missing.

When two or more people are in the same space, there generally is a feeling of where they are or what they are doing based on verbal communication, visibility or sound. In the case of the present system, most sense cues are missing.

According to the participants in the experiment, indicators and some representation of the spectators are necessary to at least know who is there and whether they are active. The minimalist camera representation of the current system did not suffice.

6.3.3. GROUNDING VIRTUAL DATA

Something surprising happened to one user of the system, as briefly described in the previous chapter: he attempted to protect a virtual pole from being moved by standing in front of it. For this user, the object had become physical.

Although, the objects used for the experiment were very clearly artificial, as seen in Figure 64, the object in question was nevertheless treated as if it were physical. Whether this is a temporal effect of a new system or this type of augmentation simply works well for the human brain is up for discussion. When digital fences were placed, the on-premise investigators also tended to avoid walking through them, which is a similar effect to that discussed above.

This behavior towards the assigned tasks aggregates real and virtual into a coherent situational awareness model which loops through Endsley's (1995) situational awareness steps of (1) the perception of elements in the environment, (2) comprehension of the current situation, and (3) projection of future status.

6.3.4. COLLABORATION

It might be stating the obvious that collaboration goes well when synchronously looking at the same environment and performing relatively simple tasks. However, some nuance is required: in this case, the expert is not on-premise.

The questionnaires and the interviews made it clear that remotely supporting spatial interaction is feasible with this mediated reality system. An expert from a remote location can guide and intervene in scenes that are reconstructed on the fly. A shared digital model that can be extended on demand provides a sufficient degree of shared situational awareness, thus allowing for collaboration. In the experiment, the tasks were jointly conducted.

The telepresence jump to the scene experts not only provides them with eyes at the scene, but also with scene assessment tools, such as measurement and analysis. The on-premise investigator can see the results of being helped directly and has a direct audio connection for clean communication.

6.3.5. MEDIATED REALITY SYSTEM PERFORMANCE

The system did not perform optimally throughout the full experiment. The experiment was shortened, as calibration of the lenses, starting all the subsystems in the correct order and orienting the scene correctly proved to be tedious. The experiments therefore needed to pause now and then. The downtimes were used to talk about the implications of the system and provided valuable insights sprinkled throughout the previous chapters.

The 4th experiment, not mentioned earlier, proved to be too difficult. The goal was to create virtual rods indicating blood pattern directions, like the physical equivalent. The quality of the scene reconstruction was not high enough to be able to accurately measure the direction. Due to difficulties experienced by the first participant, the experiment was not repeated.

6.4. CONCLUSION

Crime scene investigation often requires a spatial analysis to obtain evidence of a delict at the location of a crime. Spatial analysis is time consuming and requires specific knowledge that is sparsely available. This thesis presents a mediated reality system which is based on a novel gesture-based user interface and real-time 3D map making. This mediated reality system allows crime scene investigators to collaboratively conduct a first analysis at the crime scene while being remotely supported by expert colleagues. The design of our system is based on extensive requirements analysis with experts in crime scene investigation.

We evaluated our approach with an experiment in a staged crime scene in which an investigator at the location of the crime scene solved a spatial challenge jointly with a remote expert investigator. The collaboration between the on-premise investigator at the crime scene and the expert was observed by a domain expert panel and evaluated along the 7 different dimensions of the fluidity of collaboration; sustaining mutual information exchanges for problem solving; argumentation and reaching consensus; task and time management; cooperative orientation; and individual task orientation.

As shown in the previous chapters, the author was able to stack multiple capabilities together in a soft- and hardware system that was tested in a simulated setup at the FFL. The qualitative results were promising and provide solid hooks for further research.

With this experiment, the main research question can be answered, which is repeated below for convenience;

1) ***How can we support collaborative spatial interaction in a pristine environment applying mediated reality?***

Because the real-world is in constant flux, yesterday's digital representation of the real world will be out of date today. A crime scene is an excellent example of this, hence the need for a method that does not have to rely on previously gathered data. To answer this part of the research question, a real-time mapping system has been developed that relies on parallel tracking and mapping technology. A miniature stereo rig was created that captures HD imagery data for use in 3D scene reconstruction.

Spatial interaction is fulfilled by gesture recognition. A user interface is controlled using specific gestures through which direct scene interaction was

established. Gestures can be used to place objects in the scene and multiple users can jointly interact. No additional hardware is required to accomplish this, making the system appropriately mobile.

A head mounted display is used to augment the real environment; the generated 3D map represents that digital overlay of the environment. With this augmentation, a mediated reality is created that allows on-premise and remote experts to visually share an environment.

The system allows for the sharing of stereo video, real-time reconstructed models, 3D interaction tools and audio. The shared environment allows on-premise and remote experts to establish situational understanding. The use case experiment proved that collaborative spatial interaction is possible by two remote individuals.

The participants of the experiment could collaboratively engage with a simulated crime scene. Expert information was successfully used to aid and support a novice in 3D interactive tasks on a crime scene.

This research contributes to computer supported collaborative work. It proves that remote experts can engage with on-premise investigators in 3D collaboration in pristine environments in real-time.

7. EPILOGUE

Sutherland's first head mounted device in 1968 was astonishingly insightful at the time. His research showed that the virtual worlds can spatially co-exist with the physical world, and blend in, which still feels fantastical. In line with Sutherland's original ideas, this research was conducted to leverage mediated reality for remote collaboration purposes. Recent advances in technology allow for elaborate sharing of communication modalities. The basic idea, in which a novice not only shares what he can see from a pristine environment but also allows others to 'look into' and influence his constantly updated environment, allows for new collaborative interactions, effectively creating a new tool that boosts human capabilities. One of the first things that set early humans apart is the use of tools. Tools are extensions that enhance the physical capabilities to overcome limitations. Preceding this type of research are the mental extensions, such as the written text and photography. The conscious mind can hold only so much information; now, because of a lasting medium, information can suddenly cross generations.

We are currently in the digital age, a powerful new extension to humans. Digitization allows us to virtualize physical aspects and transport them across space. Most of the human senses have a digitization equivalent that, in a lot of cases, outperforms the human version. Basically, sensors can be made that look further, smell better and are more sensitive than their human counterparts. However, access to improved information is the same as the ability to process this. In other words, scalability is not always possible, as the same brain must cope with all the information. Scalability to solve challenges can be achieved by distributing the information amongst many. When people work together, much bigger problems can be solved than what would be achievable by a singular exemplar. The key to working together effectively is information sharing,

Our artifact attempts to leverage digital extensions that allow for collaboration in novel ways. It replicates the eyes by having computer vision, and offers the option to add sensors that are currently outside of the human capabilities. It presents this information to the user in a familiar comprehensive way and can store all the information for review from the moment of acquisition to whenever required. The artifact does not just present the information in real-time to the user but also to others; a shared 3D space that is constantly updated with incoming data, and affording a means to collaborate in the shared space that overlays the physical space. The physical space of the user becomes digitally enhanced and allows others to be immersed in that same space while not being

physically there. A novel aspect of the system is the autonomous way it functions. No prior knowledge needs to be introduced to the system, allowing it to function in pristine environments.

Developments in this domain have moved rapidly since this research was conducted. On the technology front, many technologies used in this thesis have already significantly improved. This is extremely encouraging; it means that it is desirable and that it has value. With better technology, the underdeveloped parts of the system can be improved.

7.1. REFLECTION

To be able to answer the research questions formulated in chapter 1, a sequence of research tasks was conducted. The multi domain characteristics proved to be daunting, making it harder to excel in one domain.

We started by picking a use case that provided the ingredients required to facilitate this research. Fortunately, the use case lived up to expectation, as already stated by Welten (2004). Required expertise per domain will be deeper and deeper, and therefore will be harder to obtain. The use case in which an expert must provide aid while not being on location is one that is highly feasible. Furthermore, digitization of crime scenes is becoming the norm. However, this does not mean that other domains lack use cases that can answer the viability of the proposed system.

To design a system able to meet the needs of the domain, a series of interviews, literature research and domain sessions were conducted. This yielded the necessary requirements to guide this research. I would consider this part of the thesis "a chicken and egg" or the "faster horse" problem. Although the requirements are real, they were used directly as constraints in the design process, while some of the requirements could change based on actual use and others might not have been thought of yet.

The research questions and the requirements were used to put some boundaries around the research direction before the literature research began. Without this architecture setup chapter, the search space would have been too large and unfocused. This helped significantly in the following chapters and in hindsight, did not block opportunities, if cross correlated with use cases from other domains.

In-depth literature research was needed in areas hitherto unexplored for augmented and mediated reality. Tracking, mapping and 3D interaction required

significant literature research. Those topics are easily theses in themselves and were only lightly touched on in this thesis to be able to answer the research question. Furthermore, while creating this thesis, the commercial sector started to create compelling hardware solutions.

To be able to experiment with the collaborative aspect of mediated reality, an artifact had to be created together with a software platform that would be able to provide the tooling required to do the experiments. This also proved to be daunting, as milling, soldering and modeling was required to create an acceptable head mounted device. Especially the software architecture setup worked out well. The architecture proved to be flexible and able to accommodate a wide variety of scenarios. Furthermore, the system was open source based n and could be extended for further research.

Experiments that were conducted to validate the subsystems have been individually published. Although they might not represent the state of the art in the individual domains, they can hold their own because of their integration into a broader story. The 3D interaction with a real scene overlay is particularly a very good starting point for further research. Because the system as a whole is very new, the final experiment strained it near the breaking point. In further research, a lot of work hardening would be required to do bigger scale experiments.

7.2. GENERALIZABILITY OF RESULTS

Although this thesis relies on one dominant use case, the same patterns of successfully applying similar systems can be observed across domains.

Early research on shared view (Kuzuoka, 1992) shows that the system requirements necessary to support spatial workspace collaboration are the movability of a focal point, the sharing of focal points, movability of a shared workspace, and the ability to confirm viewing intentions and movements, which are confirmed in this thesis, as claimed by Kuzuoka and confirmed by the participants in the experiment described in chapter 7.

Linguists and psychologists have observed that in reality, meaning is often negotiated or constructed jointly (Clark, 1996). Although providing the same view of a situation to two or more people is a good starting point for a shared understanding, things like professional and cultural background, as well as expectations formed by beliefs about the current situation, clearly shape the individual interpretation of a situation (Clark, 1996). This was emphasized by the participants of the research in chapter 7, too. Because they could see what the other participant was seeing and shared the same base scene, shared

situational understanding was accomplished. Furthermore, the maintenance of common ground is an ongoing process, which demands both attention and coordination between the participants (Nilsson et al., 2009). The system not only allows augmentation of the individual user's view, but also allows each user to affect and change their team members' view of the ongoing situation, which is fundamental to the definition of a collaborative augmented reality systems (Nilsson et al., 2009).

Dong et al (2013) showed that users made significantly fewer mistakes in inspection and assembly tasks, gained stronger spatial cognition and memory, and thus experienced less mental workload within a collaborative AR environment compared with VR. Although our research does not directly compare to VR, the grounding in real data was considered very favorable.

A similar setup was created in more recent research (Tait & Billinghurst, 2015) that included a toggleable view in which it was found that, for the best performance, systems for remote assistance in an AR interface should grant the remote expert an independent view of the local user's workspace, which was the preference indicated in our research as well. A simpler version for providing remote instruction was created by Adcock and Gunn (Adcock & Gunn, 2015), in which the light feature was rated highly by users in terms of efficiency, instructiveness and overall preference. Remote controlled AR projection is also described in a study from Gurevich et al (Gurevich, Lanir, & Cohen, 2015), who conclude that remote influence on a scene is a successful way of transferring information.

Kraut et al. phrased this as follows: "*Shared visual space is essential for collaborative helpers to determine: (1) worker's readiness to receive help, (2) the nature of the help the worker needs, and (3) worker's comprehension of new information"* (Kraut et al., 2000). All three requirements for determining help are represented in the artifact. Adding to Kraut`s statement is the favorable effect of having virtual objects grounded in reality, which lessens the cognitive load, and the synchronized workspace that avoids misunderstanding in communication caused by the distortion of time or viewpoint (Bujak et al., 2013)

Our lab in the TU Delft continues to push AR research, too. Researchers there (Lukosch, Lukosch, Datcu, & Cidota, 2015) found that the biggest advantage of AR technology in the security domain was the fact that a remote user could be introduced, who was virtually co-located with the users at the crime scene. Such virtual co-location not only allows the remote user to see what the local users see, but also provides additional information on the spot by augmenting the real

environment with virtual objects. Furthermore, they set an agenda for further research in this domain (Lukosch, Billinghurst, Alem, & Kiyokawa, 2015).

The principles used to remotely aid a user by directly influencing the scene and thereby establishing situational awareness and collaboration is a proven principle in multiple domains and does not only apply to the one used here.

7.3. CHALLENGES

In some ways, this thesis creates more questions than it answers. Fortunately, many of the newly manifested questions could be sidestepped to answer the main research question. In this chapter, the main nagging unresolved challenges are discussed.

7.3.1. AIR TAPPING

Although some success can be claimed using the proposed solution as addressed by Lukosch (Lukosch et al., 2012), many challenges still remain. Microsoft Kinect, the most adopted gesture device in the market, is no longer sold with the console as this was not used anyway. Selecting with 3D gesture interaction is very difficult without haptic feedback. The challenges with 3D gesture interactions can be divided into a few problem spaces: physical, feedback and intuitiveness.

With consoles, relatively coarse interaction is all that is required: hand waving, swipes and keeping still in one location. With HMD gesture interaction, precision is required, as this requires concentration and motor system effort. During the experiments, it was only fair to ask the participants to work on a scene for a few minutes. In longer experiments, participants started to complain.

Because no natural haptic feedback is present, the feedback must come from somewhere else. Experiments were conducted with different feedback cues and the results deviated significantly. By providing a clear cue of what would be selected, together with a symbol change in the case of a successful selection, the success rate went up. However, this still barely scratches the surface in feedback to the user if haptic feedback is lacking.

There is as yet no equivalent in 3D gesture interaction to the ‘swipe’ movement on a smartphone. The chosen left hand thumbs-up selection menu seems to work well, but this only replaces the selection menu from a pulldown menu to space-based one. Furthermore, the system only used 3D selection to place and replace objects. Even 3D specialists had to be told how to interact with the system. The complexity of interaction was kept simple on purpose; as soon as complex 3D interaction was required, we struggled with the paradigm.

7.3.2. VIRTUAL REPRESENTATIONS

Especially in the final experiment, it became clear that it was important to provide a representation of a co-worker in virtual space. Although a representation was visualized as a virtual camera, this was clearly not enough. There are multiple aspects to this representation, including knowing who is watching, knowing what they are doing and exchanging information with the virtual presence.

Knowing who is watching was found to be important. One of the investigators explained this well*: 'It's as if you are working on a painting and you don't want people to see the results until you are finished, it's still dirty laundry'.* What an investigator sees and does is personal and open to interpretation, especially when superiors or unfamiliar people are involved. This appears to be less of an issue for the camera feed then for scene interaction. Another aspect is whether they are engaged or not: whether they are intently following or merely occasionally watching what is going on.

When two co-located colleagues are working together virtually at a scene, it is important that each knows what the other is doing, what are they looking at, which tool each has activated, what are they clicking on, etc. In one of the experiments not documented in this thesis, two people used gestures in the same scene, both between different systems, which helped significantly in feeling connected.

How to interact with the other virtual presence in the environment? A document that is relevant to the scene, such as a similar pattern to be used for comparison can be handed over as something new in the inbox or as a virtual hand providing the document. Is the physically mimicked equivalent better than the digital method?

In multiplayer games, avatars constantly interact. However a pure virtual space is different than an augmented environment. In a 3D game environment, events are pre-scripted; the controllers are mouse/keyboard or gamepads and the virtual camera is free. There is a much more personal aspect to a system that requires additional presence cues.

7.3.3. AUGMENTATION

At the beginning of this thesis, we thought it would be easy to hook up an optical see-through HMD. After running several experiments, it became clear that the eye/brain visual system is an extremely fine-tuned sense that is not easily fooled, hence the fallback to video see-through. There are different challenges in

augmented overlay: surpassing human performance, compensate for anthropometric aspects, blending and rendering.

The inner ear bone, the motor system, the hearing and the eyes work together seamlessly. As described in much HMD research, any disruption to this system causes nausea in many people. This is also the reason that the newer HMDs have an external tracker to follow the HMD's position in space, providing not only jaw, role and pitch but also translation. For an HMD to work optimally, it would need to outperform the reaction time of humans. This effectively means that the positional location tracking, the rendering and the display all need to preferably react under 40ms. If the reaction time is higher, the human sense system notices.

An HMD is strapped to the head, which means the only moving parts are the eyes. The eyes are used to perceive depth because of disparity and convergence; furthermore, there is rapid eye movement for additional resolution and the principle of human vision is based on contrast patches. Under low light, the eyes switch to contrast enhancing gray scale, and with enough light, colors can be very vivid. Also, the eyes move in their sockets to quickly change the focus of attention and only the attention area is processed at full resolution; the rest is coarser. In contrast, an HMD is a fixed position lens in front of the eye with a display and a smaller field of view. Basically, most of the capabilities of the eyes are ignored in HMDs. In this thesis, a video see-through HMD was used to invoke augmented overlay, and although successful overlay was accomplished, removing the HMD after a session showed how different seeing for real and displays are. The tests with optical see-through were not as different.

Because the blending of optical see-through was hard to accomplish, video see-through was used, eliminating alignment issues. Our eyes only have accurate depth perception at a few meters' range and rely heavily on other depth-related cues to determine distance. The first tests with the HMD did not use any occlusion information to render the 3D objects in place which, although correctly placed in the scene, felt off. As soon as depth was used to occlude correctly, the placement felt more accurate. Still, the data was obviously alien to the scene. Of course, that can be a good thing for tools and photorealism might not be necessary, but grounding in the scene still seems important, to generate confidence about the fact that the virtual data is accurate.

The initial few seconds of putting on the HMD seem magical; the depth feels so real and virtual objects are extremely stable; however, as soon as the HMD is removed, the eyes adjust and it not only feels very disorienting, but also tiring.

7.3.4. MONITORING

Apparently, there is boundary that separates what people accept that others may observe and what is private. We ran up against this problem in the present work, which, it should be noted, we had not anticipated. While wearing cameras causes resistance, monitoring arouses that much more. This topic was often addressed in after action reviews. Although not a specialist in this area, it appeared that there were several reasons for the angst displayed: de-humanizing judgement, too much monitoring and trust.

If there is a feeling that managers are going to judge the performance of employees based on numbers that can be extracted from the constant monitoring, this would face resistance. So many aspects are important in judging employees that it feels like a threat.

When the augmented sessions are open to too many people, the feeling of being watched seems to increase. During the final experiment, an on-premise participant would not perform an action until it had been triple checked, because he did not want to make a mistake in front of such an audience. This might be because everything is so new, however, being watched all the time does not seem to be a desirable situation.

Regular collaboration uses voice intonations, body language and micro face expressions to create trust and symmetry between participants. In this mediated system, this symmetry is partly established by the actions taken through the video feed. The artifact users must place a great deal of trust in the other person.

7.3.5. RECONSTRUCTIONS

Although the reconstructions based on imagery are still improving, there are still many challenges to be overcome in this space. Fortunately, simple reconstructions were sufficient to prove the principle for this thesis. So, many details remain to be addressed in this challenge, in which whole domains are devoted to accurately copying physical spaces. Relevant challenges in this thesis were: tracking and mapping, dense 3D reconstruction and multimodality fusion.

Tracking and mapping worked well for room-sized spaces with sufficient features. The experiments were controlled to function well with enough geometry and texture in the scene and enough light. They system was tested in many conditions, but curating was required for stable performance. Speed was another challenge: 30hz was obtained, but for fluid tracking much higher frame rates are preferable. A quick prototype test showed that with much higher fps the fluidity felt much better. The scene orientation is based on fitting the

dominant plane on the first feature set, which proved to be troublesome, as, if the cameras are pointed to a table or wall, the origin ends up in an undesired place which makes it hard to collaborate and for the tools to work correctly.

With the pre-calibrated stereo camera and key frames, a dense map of the scene could be reconstructed. However, the quality of the map could not compete with currently used scanner data; the results were especially poor on monochromatic surfaces. Meshing the point clouds provided a cleaner looking model but it smoothed out many of the details. Much better results could be achieved with better cameras, but the form factor was undesirable.

The odometry was only based on vision tracking; the inertia sensors of the phone/HMD proved to be unreliable. Ideally, drift, fast movement and orientation can be extracted from other sensors instead of relying purely on vision. Pre-calibrated cameras allowed for easy scene texturing, however, the pre-calibration was easily disturbed. Automated in scene calibration would have been desirable.

7.3.6. PRESENCE

The research questions did not address presence as a topic to be researched in this thesis. However, the topic is undeniably part of the research space. Especially the remote participant projects himself in the scene and has no physical manifestation on the work scene.

Although the point cloud or mesh environments are not pretty, it feels authentic, much different than the second life-like 3D worlds. Moreover, there is another person walking around and mapping the environment. It feels like being there for most people, especially when they can manipulate scene elements.

The opposite might be true for the on-premise participant, for whom the virtual overlay starts to disconnect his feeling of reality. In theory, that person can only see a digital overlay and never see the real world directly. What if they always look at a virtual overlay? The sessions with the HMD were short, so no real understanding of this aspect surfaced.

7.3.7. COLLABORATION

The premise of this thesis is to understand how we can allow people to collaborate in pristine environments. The answer to this research question is detailed in the previous chapters; however, many side topics on collaboration have been ignored that are still important. Dong et al. (2013) showed that users made significantly fewer mistakes in inspection and assembly tasks, gained

stronger spatial cognition and memory, and thus experienced less mental workload within a collaborative AR environment compared with VR.

The research question did not require a comparison between physical and mediated reality supported equivalents. Participants frequently asked for examples in which the system would outperform what they were already doing. Because the tools are relatively crude, the effects of spatial cognition are unclear, does having real footage support the spatial tasks?

The system aggregated audio, timings, interactions, video, etc. into one scene. Although specified according to the requirements, does this help the collaboration or is there too much of an information threshold? When does the aggregation of information hamper what a team is trying to accomplish?

There was no difficult problem solving involved in the scenarios, hence it is unknown whether this setup will work when more complex assessments must be made.

7.4. COMMERCIAL SOLUTIONS

The industry is catching up on the idea of augmented overlay in pristine environments. One of the best examples is Google's Project Tango. No HMD is used to overlay, but a smart phone with depth sensor is used to create a course map of the environment, mainly for indoor localization but the principle is the same. A user looks though his mobile screen and can overlay any 3D object and it will be correctly located in space. Associated with project Tango are a lot of companies that deliver pieces of technology, such as Bsquare, paracosm and Mantis vision. 13th lab has a similar approach, although without a depth sensor, as demonstrated in the rescape project. The advances and commoditization of tracking and mapping technology are very favorable to the research conducted in this thesis.

A mobile phone does not have the immersion of a head mounted display, but in the head mounted display domain, too, much progress has been made by the industry. Microsoft has the Hololens project, in which an autonomous HMD accomplishes similar capabilities to the system described in this thesis. GPS-based positioning information is used to deliver context-based messages. The oculus rift is another example of industry progress; a wider field of view, low persistence displays and increased resolution are just a few of the list of improvements.

3D user interfaces to augmented reality environments are also being improved by the industry. Project Meta uses a depth sensor integrated with a see-through head mounted display to interact with the virtual space. Leap motion has an extremely fast depth hand tracking camera for 3D interaction and the DAQRI Smart Helmet integrates the interaction by overlaying menu options on top of a phone.

Although the hardware in the commercial sector has improved, the software still has to catch up. No solutions were found that facilitated the collaboration capabilities.

7.5. FURTHER RESEARCH

Many of the challenges proposed in the previous chapter are natural candidates for further research and are addressed in this chapter. I would like to start with a slight modification of Figure 1 in chapter 1, here shown as Figure 71. As described in chapter 4.2, sensors are becoming more sophisticated, cheaper and ubiquitous, which means that sensing technology will increasingly be monitoring our environment, giving machines a sense of our environment. If this trend continues, real-time sensed environments would allow for virtual arch angles that sense with us. Combined with better artificial intelligence and 'always on' connectivity, human - computer interaction might look very different in a few years' time.

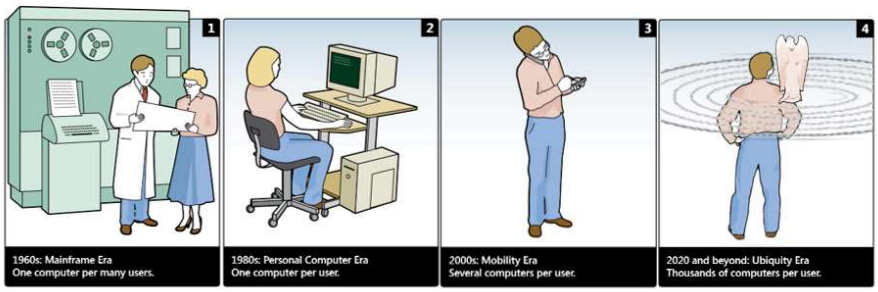


FIGURE 71 ADAPTION OF (HARPER ET AL., 2008), MODIFIED FIGURE 4.

7.5.1. DIGITIZATION

Scene reconstruction based on sensor data has come a long way, but also has a long way to go. If the right constraints are leveraged, a sensor pose and a reconstruction can be achieved in real-time, as proven in this thesis. However, many aspects of scene reconstruction are immature or non-existent.

When there is less geometric information to latch on to, most systems must rely on other sensors, such as the odometer, compass, etc. This introduces drift, which inevitability leads to bad data. If a large percentage of the environment is dynamic, moving people, cars, waving trees, etc., the reconstruction generally fails. The precision of the reconstruction relies the stability of an environment, i.e. it is constantly the same. Glass, reflective material, translucence and changing light impact the reconstruction significantly. The material properties of the environment are generally ignored and a lambertian generic description is used to predict properties.

Apart from the reconstruction, the sensors themselves are very capable. There are passive and active sensors for the full electrometric spectrum. The challenges are not what a sensor is capable of outputting, it is the backend that needs to extract the useable information. The human body is bombarded by 2 billion senses every second and filters these to what is required for the task at hand. Vision systems do not have filters that aggregate the information like the human brain. Although the sensors are very capable, many have a form factor that is not portable or energy sensitive.

Leveraging the knowledge from multiple types of sensors presents opportunities for the future. The Kinect v2 already senses depth, RGB and infrared. By having all three channels available, the detection and precision rate has improved significantly.

7.5.2. SPATIAL INTERACTION

Although we proved that spatial interaction in an overlay of a physical scene is feasible, this research is far from complete. The intuitiveness and discoverability that is associated with 2D touchscreen devices is not yet present in 3D. Serious attempts have been made to interact with virtual spaces, such as the Microsoft depth sensor and game console, but it still feels clunky. Fatigue still plays a major role in interaction paradigms with bare hand gestures. The motor system of the shoulders and arms is not ideal for precision, and, while fingers are extremely accurate, they are harder to detect accurately because of occlusion. Furthermore, the “enter” button is missing: what is a click or an acknowledgement? Apart from haptics, there are other feedback channels that must be leveraged, such as visual, auditive and HMD vibrations.

For further research, there are a few clear research topics. What interface paradigm comes intuitively to humans for 3D interaction in digital overlays in physical environments? What are the best places to put sensors for precise and

non-intrusive tracking of gestures? Which modalities work best for what kind of operation? And so forth. An overlay of a handheld device requires a different approach than a head mounted display based approach, 3D interfaces in stereo are still novel.

7.5.3. HEAD MOUNTED DISPLAYS

Head mounted devices are becoming more and more popular and their capabilities are improving rapidly. However, even with the larger field of view and the high refresh rates, nausea remains a problem. The eye is required to focus on infinity and accommodation of the eye has no effect on the 3D rendering.

More novel methods of using a contact lens together with a display or direct retinal display projection are still immature. If a digital overlay is going to be mass adopted, many of the side effects must be substantially reduced.

Considering just an overlay of sensors, like infrared or ultraviolet is another challenge. Where are the sensors, how are they overlaid, how is occlusion handled or the movement of the eye? Next, what needs to be done to accommodate the movement of the eye, are eye trackers good enough, is there a constant updating of the eye model?

Is it possible to miniaturize the technology enough so it becomes unobtrusive, will the brightness and color of the projection ever be on par with the physical world, and do we want a difference? Are the current rendering paradigms good enough? Would we need light fields and bidirectional reflectance distribution functions kind of quality or are there other better approaches?

7.5.4. MEDIATED COLLABORATION

Most further studies based on the findings in this thesis are in this research bucket. The simple fact that an off-site human can look into the replicated environment of an on-site human does something with the mind. It is not like the tele- presence face-to-face communication in which both persons look into the same scene or even through the eyes of each other. On top of that, the off-site person has a digital replica that is able to virtually be present and influence the digital overlay.

There is a growing body of knowledge on presence in the psychology domain. However, this line of research currently has no access to this mediated reality system. It is certainly worth teaming researchers from this domain with this type of mediation research. What can we do to help people feel more comfortable with someone constantly looking over their shoulder? Will people eventually accept

such a system if it is mandatory? Is the human mind capable of constant transitions to that of others, is the feeling of presence too overwhelming or is more needed to invoke presence?

There are big ethical questions to be answered, too. Can this information be used against the user if they were to do something wrong? Should some of the information be shielded or censored? A constantly recording camera in devices such as Google's glass has already set off heated debate. What if that device can record much more then imagery - high quality sound, chemical sensors and x-ray?

This research can be dramatically extended into the situational awareness space. Currently, the tests we performed focused on an expert helping a novice. The team play aspect was not addressed. How does this system scale in the event of a major hazard involving multiple people with mediated systems and experts? How is situational awareness improved?

What physical aspects are still missing for an expert not on location, what is he or she missing by not being there? How important is a physical aspect, if all the data is digitized? Does the human brain need to adapt so that eventually his senses - in tandem with technology - will provide better results?

7.6. CONCLUSIONS

This thesis is a contribution to physical computing; the building of interactive physical systems with the use of software and hardware that can sense and respond to the analog world. As Mark Weiser (1991) observed, a main concern is that computer interfaces are too demanding of human attention; *'Unlike good tools that become an extension of ourselves, computers often do not allow us to focus on the task at hand but rather divert us into figuring out how to get the tool to work properly'.* Much earlier Sutherland (1965) was already building an augmentation device that would make human computer interaction more intuitive.

As humans, we are magnificent at imagination, but imagination is hard to share. The virtual world is the closest to sharable imagination we can currently get. This thesis attempts to push that barrier a little closer in collocation.

REFERENCES

Adachi, T., Ogawa, T., Kiyokawa, K., & Takemura, H. (2005). *A Telepresence System by Using Live Video Projection of a Wearable Camera onto a 3D Scene Model.* Paper presented at the International Conference on Human-Computer Interaction 3-12.

Adcock, M., & Gunn, C. (2015). *Using Projected Light for Mobile Remote Guidance.* Paper presented at the Computer Supported Cooperative Work.

Agrawala, M., Beers, A., McDowall, I., Fröhlich, B., Bolas, M., & Hanrahan, P. (1997). *The Two-User Responsive Workbench: Support for Collaboration Through Individual Views of a Shared Space.* Paper presented at the SIGGRAPH.

Akman, O. (2012). *Robust Augmented Reality.* PHD, TU Delft, Delft.

Akman, O., Poelman, R., Caarls, J., & Jonker, P. (2013). Multi-cue hand detection and tracking for a head-mounted augmented reality system. *Machine vision and applications, 24*(5), 931-946.

Andreasen, N., & Brown, T. (2004). *Facilitating Interdisciplinary Research*. Washington DC: National Academies Press.

Arayici, Y., & Aouad, G. (2004). DIVERCITY: distributed virtual workspace for enhancing communication and collaboration within the construction industry. In A. Dikbas & R. Scherer (Eds.), *eWork and eBusiness in Architecture, Engineering and Construction*. London: Taylor & Francis Group.

Argyros, A., & Lourakis, M. (2004). *Real-time Tracking of Multiple Skin-Colored Objects with a Possibly Moving Camera.* Paper presented at the European Conference on Computer Vision.

Aw, S., Halmagyi, G., Haslwanter, T., Curthoys, I., Yavor, R., & Todd, M. (1996). Three-Dimensional Vector Analysis of the Human Vestibuloocular Reflex in Response to High-Acceleration Head Rotations II. Responses in Subjects With Unilateral Vestibular Loss and Selective Semicircular Canal Occlusion. *Journal of Neurophysiology, 76*(6), 4021-4030.

Azuma, T. (1997). A Survey of Augmented Reality. *Presence, 6*(4), 355--385.

Balwer, A. (2013). Electromagnetic WavesPhysics Course: Wikibooks.

Barton, H., & Byrne, K. (2007). *Introduction to Human Vision, Visual Defects & Eye Tests.* Paper presented at the Information and Communication Technologies.

Battaglia, J., Brubaker, R., Ettenberg, M., & Malchow, D. (2007). High speed Short Wave Infrared (SWIR) imaging and range gating cameras *Photo-Optical Instrument Engineers*.

Bauer, M., Bruegge, B., Klinker, G., MacWilliams, A., Reicher, T., Riß, S., . . . Wagner, M. (2001). *Design of a Component–Based Augmented Reality Framework.* Paper presented at the International Symposium on Augmented Reality.

Bay, B., Ess, A., Tuytelaars, T., & Gool, L. (2008). SURF: Speeded Up Robust Features. *Computer Vision and Image Understanding, 110*(3), 346-359.

Benford, S., Greenhalgh, C., Reynard, G., Brown, C., & Koleva, B. (1998). Understanding and Constructing Shared Spaces with Mixed-Reality Boundaries. *Computer-Human Interaction, 5*(3), 185--223.

Beraldin, J., Blais, F., Cournoyer, L., Godin, G., & Rioux, M. (2000). *Active 3D sensing.* Paper presented at the Modelli E Metodi per lo studio e la conservazione dell'architettura storica, Pisa, IT.

Billinghurst, M., & Kato, H. (1999, Mar.). *Collaborative Mixed Reality.* Paper presented at the International Symposium on Mixed and Augmented Reality, Yokohama, Japan.

Bimber, O., & Raskar, R. (2005a). *Modern approaches to augmented reality*. Paper presented at the Computer Graphics and Interactive Techniques.

Bimber, O., & Raskar, R. (2005b). *Spatial Augmented Reality: Merging Real and Virtual Worlds*. Wellesley, MA, USA: A. K. Peters, Ltd.

Blair, P., & Johns, L. (1993). Who Goes There: Friend or Foe? (pp. 84): Office of Technology Assessment, Congress of the U.S.

Boehm, B. (1986). A Spiral Model of Software Development and Enhancement. *Software Engineering, 11*(4), 22-42.

Boel, P., DeCloet, V., DeKinder, J., Mahieu, J., & VanVarenbergh, D. (2009). *Handboek forensisch onderzoek*. Politeia.

Boger, Y. (2007). Are Existing Head-Mounted Displays 'Good Enough'? , 11. Retrieved from http://sensics.com/are-existing-head-mounted-displays-good-enough/

Botden, S., & Jakimowicz, J. (2009). What is going on in augmented reality simulation in laparoscopic surgery? *Surgical Endoscopy, 23*(1), 1693–1700.

Bouras, C., Giannaka, E., Panagopoulos, A., & Tsiatsos, T. (2006). *A platform for virtual collaboration spaces and educational communities: the case of eve.* Paper presented at the Multimedia Systems.

Bowman, D., Kruijff, E., LaViola, J., & Poupyrev, I. (2005). *3D User Interfaces: Theory and Practice*: Addison-Wesley

Broll, W., Lindt, I., Ohlenburg, J., Wittkamper, M., Yuan, C., Novotny, T., . . . Strothmann, A. (2004). ARTHUR: A Collaborative Augmented Environment for Architectural Design and Urban Planning. *Journal of Virtual Reality and Broadcasting, 1*(1).

Buck, U., Kneubuehl, B., Näther, S., Albertini, N., Schmidt, L., & Thali, M. (2011). 3D bloodstain pattern analysis: Ballistic reconstruction of the trajectories of blood drops and determination of the centres of origin of the bloodstains. *Forensic Science International, 206*(1-3), 22-28. doi: http://dx.doi.org/10.1016/j.forsciint.2010.06.010

Bujak, K., Radu, I., Catrambone, R., MacIntyre, B., Zheng, R., & Golubski, G. (2013). A psychological perspective on augmented reality in the mathematics classroom. *Computers & Education, 68*, 536-544. doi: 10.1016/j.compedu.2013.02.017

Burkhardt, J., Détienne, J., Hébert, H., Perron, L., Safin, S., & Leclercq, P. (2009). *An approach to assess the quality of collaboration in technology-mediated design situation.* Paper presented at the European Conference on Cognitive Ergonomics.

Butz, A., Hollerer, T., Feiner, S., MacIntyre, B., & Beshers, C. (1999). *Enveloping Users and Computers in a Collaborative 3D Augmented Reality.* Paper presented at the Workshop on Augmented Reality, Washington, DC, USA.

Caarls, J., Jonker, P., Kolstee, Y., Rotteveel, J., & Eck, W. (2009). Augmented Reality for Art, Design and Cultural Heritage - System Design and Evaluation. *Journal on Image and Video Processing, 2009*, 16.

Caarls, J., Jonker, P., & Persa, S. (2003). Sensor fusion for augmented reality *Ambient Intelligence* (pp. 160--176): Springer Verlag.

Caarls, J., & Pieter, J. (2009). Wearable Augmented Reality System. *Journal on Image and Video Processing* doi: 10.1155/2009/716160

Canalys, D. (2011). Smart phones overtake client PCs. Retrieved from https://www.canalys.com/newsroom/smart-phones-overtake-client-pcs-2011

Cassin, B., & Solomon, S. (1990). *Dictionary of Eye Terminology*: Triad Pub. Co.

Castells, M. (1996). *The Rise of the Network Society: The Information Age: Economy, Society, and Culture Volume I*.

Chik, D. (2006). *Using Optical Flow for Step Size Initialisation in Hand Tracking by Stochastic Optimisation.* Paper presented at the Vision in Human-Computer Interaction, Canberra, Australia.

Clark, H. (1996). Using language. *Journal of Linguistics*(35), 167-222.

Cui, Y., Pagani, A., & Stricker, D. (2011). Robust point matching in HDRI through estimation of illumination distribution. *German Association for Pattern Recognition*.

Daley, B. (2015). Smart Augmented Reality Glasses. https://www.tractica.com/research/augmented-reality-for-mobile-devices/: Tractica.

Davison, A. J., Mayol, W. W., & Murray, D. W. (2003). *Real-Time Localisation and Mapping with Wearable Active Vision.* Paper presented at the International Symposium on Mixed and Augmented Reality, Tokyo, Japan.

Dean, W. (2004). Electrical-energy-storage unit (EESU) utilizing ceramic and integrated-circuit technologies for replacement of electrochemical batteries In EESTOR (Ed.), *US2004071944 (A1) — 2004-04-15*. USA.

Dellaert, F., Seitz, S., Thorpe, C., & Thrun, S. (2000). Structure from Motion without Correspondence. *Computer Society Conference on Computer Vision and Pattern Recognition*.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (2004). *Human-Computer Interaction* (3rd ed.). Harlow, England: Pearson Education Ltd.

Dobson, P. (2002). Critical realism and information systems research: why bother with philosophy? *Information Research, 7*(2).

Dodgson, N. (2004). *Variation and extrema of human interpupillary distance.* Paper presented at the Stereoscopic Displays and Virtual Reality Systems, San Jose, California, USA.

Dong, S., Behzadan, A., Chen, F., & Kamat, V. (2013). Collaborative visualization of engineering processes using tabletop augmented reality. *Advances in Engineering Software, 55*, 45-55.

Drora, R., Adelsonb, E., & Willskya, A. (2001). *Estimating Surface Reflectance Properties from Images under Unknown Illumination.* Paper presented at the Human Vision and Electronic Imaging, San Jose, CA, USA.

Durrant-Whyte, H., & Bailey, T. (2006). Simultaneous Localisation and Mapping (SLAM): Part I The Essential Algorithms. *Robotics and Automation Magazine, 13*(2), 99-110.

El-Hakim, S. (2001). *Three-dimensional modeling of complex environments*. Paper presented at the Videometrics and Optical Methods for three-dimensional Shape Measurement, San Jose.

Elliott, S., & Nelson, P. (1993). *Active Noise Control, Low-frequency techniques for suppressing acoustic noise leap forward with signal processing* Paper presented at the Signal processing magazine.

Endsley, M. (1995). Towards a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64.

Evans, J. (1998). *The History and Practice of Ancient Astronomy*. Oxford Oxford University Press.

Evennou, F., & Marx, F. (2006). Advanced integration of WIFI and inertial navigation systems for indoor mobile positioning. *Journal on Applied Signal Processing, 2006*(1), 164-168. doi: 10.1155/ASP/2006/86706

Fischler, M., & Bolles, R. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM, 24*, 381-395.

Flew, T. (2009). *New Media: An Introduction 3rd edition Edition*: Oxford.

Flight, S., & Hulshof, P. (2010). Roadmap beeldtechnologie veiligheidsdomein Amsterdam: DSP - groep BV.

Forsberg, K., Mooz, H., & Cotterman, H. (2005). *Visualizing Project Management*. New York: John Wiley and Sons.

Franke, K., & Srihari, S. (2007). *Computational Forensics: Towards Hybrid-Intelligent Crime Investigation.* Paper presented at the Information Assurance and Security, Manchester, UK.

Franke, K., & Srihari, S. (2008). Computational Forensics: An Overview *Computational Forensics* (Vol. 5158/2008, pp. 1-10): Springer Berlin / Heidelberg.

Fries, C. (2006). Cutting Edge 3-D Reconstruction. *Forensic Magazine*.

Fuchs, H., Livingston, M., Raskar, R., Colucci, D., Keller, K., State, A., . . . Meyer, A. (1998). *Augmented Reality Visualization for Laparoscopic Surgery.* Paper presented at the Medical Image Computing and Computer-Assisted Intervention, Heidelberg, Germany.

Fumarola, M., & Poelman, R. (2011). Generating virtual environments of real world facilities: Discussing four different approaches. *Automation in construction 20*(3), 263-269.

Furukawa, Y., & Ponce, J. (2009). Accurate, Dense, and Robust Multi-View Stereopsis. *Pattern Analysis and Machine Intelligence*.

Fussell, S., Setlock, L., & Kraut, R. (2003). *Effects of Head-Mounted and Scene-Oriented Video Systems on Remote Collaboration on Physical Tasks.* Paper presented at the Human Factors in Computing Systems, Ft. Lauderdale, Florida, USA.

Gergle, D., Kraut, R., & Fussell, S. (2013). Using Visual Information for Grounding and Awareness in Collaborative Tasks. *Human Computer Interaction*(28:1), 1-39.

Gorce, M., Paragios, N., & Fleet, D. J. (2008). *Model-Based Hand Tracking with Texture, Shading and Self-occlusions.* Paper presented at the Computer Vision and Pattern Recognition, Anchorage, Alaska.

Grinblat, I., & Peterson, A. (2012). OGRE 3D Application Development Cookbook (pp. 306).

Gurevich, P., Lanir, J., & Cohen, B. (2015). *Design and Implementation of TeleAdvisor: a Projection-Based Augmented Reality System for Remote Collaboration.* Paper presented at the Computer Supported Cooperative Work.

Haan, G. (2006). *Hybrid Interfaces in VEs: Intent and Interaction.* Paper presented at the Eurographics Symposium on Virtual Environments

Hahn, T. (2010). Future Human Computer Interaction with special focus on input and output techniques.

Harper, R., Rodden, T., Rogers, Y., & Sellen, A. (2008). Being human: human-computer interaction in the year 2020: Microsoft Research Ltd.

Harris, C., & Stephens, M. (1988). A combined corner and edge detector (pp. 147-152). United Kingdom: Chris Harris & Mike Stephens.

Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science*. doi: 10.1186/2194-0827-1-1

Hartley, R., & Mundy, J. (1993). *Relationship between photogrammmetry and computer vision*. Paper presented at the Integrating photogrammetric techniques with scene analysis and machine vision, Orlando, Fl, USA.

Heidemann, G., Bax, I., & Bekel, H. (2004). *Multimodal interaction in an augmented reality scenario.* Paper presented at the International Conference on Multimodal Interfaces, New York, USA.

Hevner, A., March, S., Park, J., & Ram, S. (2004). *Design Science in Information System Reseach*.

Hinckley, K. (1996). Haptic Issues for Virtual Manipulation. Retrieved from http://research.microsoft.com/apps/pubs/default.aspx?id=68165 website:

Hirschmuller, H. (2008). Stereo processing by semiglobal matching and mutual information. *Pattern Analysis and Machine Intelligence, 30*(2), 328–341.

Hollerer, T., Feiner, S., Terauchi, T., Rashid, G., & Hallaway, D. (1999). Exploring MARS: Developing Indoor and Outdoor User Interfaces to a Mobile Augmented Reality System. *Computers and Graphics, 23*(6), 779--785.

Hua, H., Krishnaswamy, K., & Rolland, J. (2006). Video-based eyetracking methods and algorithms in head-mounted displays *Optics Express, 14*(10).

Hubo, E. (2007). *Compression Techniques for Massive Point Set Surfaces, with Application to Ray Tracing.* PHD, University Hasselt, Hasselt, Begium.

Isenbrug, M., & Gumhold, S. (2003, 3 July). *Out-of-Core Compression for Gigantic Polygon Meshes.* Paper presented at the Graphics.

Jackson, R., & Jackson, J. (2004). *Forensic Science*: Prentice Hall.

Jenkins, B. (2005). Laser Scanning for Forensic Investigation. *SparView™, 3*.

Johnson, J. (1958). *Analysis of image forming systems.* Paper presented at the Image Intensifier Symposium, Warfare Electrical Engineering Department, U.S. Army Research and Development Laboratories, Ft. Belvoir, Va.

Julier, S., Baillot, Y., Lanzagorta, M., Brown, D., & Rosenblum, L. (2000). *BARS: Battlefield Augmented Reality System.* Paper presented at the NATO Symposium on Information Processing Techniques for Military Systems, Istanbul, Turkey.

Kamtekar, K., Monkman, A., & Bryce, M. (2010). Recent Advances in White Organic Light-Emitting Materials and Devices *Advanced Materials, 22*(5), 572–582. doi: 10.1002/adma.200902148

Kanade, T., & Bajcsy, R. (1993). Computational Sensors. Pennsylvania, USA: DARPA

Kansaku, K., Hata, N., & Takano, K. (2010). My thoughts through a robot's eyes: An augmented reality-brain–machine interface. *Neuroscience Research 66*(219–222).

Kavanagh, B. (2008). *Surveying: Principles and Applications (8th Edition)*: Prentice Hall; 8 edition.

Kazhdan, M., Bolitho, M., & Hoppe, H. (2006). Poisson Surface Reconstruction. *Eurographics Symposium on Geometry Processing*.

Kennedy, R., Berbaum, K., & Lilienthal, M. (1992). *Human operator discomfort in virtual reality systems: simulator sickness – causes and cures*. Paper presented at the Advances in Industrial Ergonomics and Safety IV, London, England.

Kern, P., & Riedel, O. (1996). *Ergonomic issues of head mounted displays*. Paper presented at the Advances in Occupational Ergonomics and Safety I, Germany.

Kiyokawa, K., Billinghurst, M., Campbell, B., & Woods, E. (2003). *An Occlusion-Capable Optical See-through Head Mount Display for Supporting Co-located Collaboration* Paper presented at the

International Symposium on Mixed and Augmented Reality, Tokyo, Japan.

Klein, G., & Murray, D. (2007). *Parallel Tracking and Mapping for Small AR Workspaces*. Paper presented at the International Symposium on Mixed and Augmented Reality, Nara.

Knight, J., & Baber, C. (2004). Neck Muscle Activity and Perceived Pain and Discomfort due to Variations of Head Load and Posture. *Aviation, Space, and Environmental Medicine, 75*(2), 123-131.

Knight, J., Williams, D., Arvanitis, T., Baber, C., Wittkaemper, M., Herbst, I., & Sotiriou, S. (2005). *Wearability Assessment of a Mobile Augmented Reality System* Paper presented at the Virtual Systems and MultiMedia Ghent, Belgium.

Koelsch, M., Bane, R., Hoellerer, T., & Turk, M. (2006). *Multimodal interaction with a wearable augmented reality system.* Paper presented at the Computer Graphics and Applications.

Kolb, C., Mitchell, D., & Hanrahan, P. (1995). *A realistic camera model for computer graphics.* Paper presented at the SIGGRAPH.

Kollin, J., & Tidwell, M. (1995). *Optical engineering challenges of the virtual retinal display* Paper presented at the Optical Systems Design and Optimization.

Kopeika, N. (1998). *A system engineering approach to imaging*: International Society for Optical Engineering.

Kosonocky, S., & Collins, F. (2013). ISSCC Trends. *International Solid-State Circuits Conference*.

Kot, K., Chernikov, A., & Chrisochoides, N. (2006). *Effective out-of-core parallel delaunay mesh refinement using off-the-shelf software*. Paper presented at the International Parallel and Distributed Processing Symposium, Rhodes Island.

Kraut, R., Miller, M., & Siegel, J. (1996). *Collaboration in Performance of Physical Tasks: Effects on Outcomes and Communication.* Paper presented at the Computer Supported Cooperative Work.

Kraut, R., Siegel, J., Hanson, J., & Lerch, J. (2000). *Coordination of communication: Effects of other's presence and visual information on human performance.* Paper presented at the Computer Supported Cooperative Work.

Krevelen, R., & Poelman, R. (2010). A Survey of Augmented Reality Technologies, Applications and Limitations. *International Journal of Virtual Reality, 9*(2), 1-20.

Kritzenberger, H., Winkler, T., & Herczeg, M. (2002). Collaborative and Constructive Learning of Elementary School Children in Experiental

Learning Spaces along the Virtuality Continuum. In M. Herczeg, W. Prinz & H. Oberquelle (Eds.), *Mensch & Computer* (pp. 115--124). Stuttgart, Germany: B. G. Teubner.

Kruijff, E. (2006). *Unconventional 3D User Interfaces for Virtual Environments* PHD, Graz University of Technology, Graz.

Kruijff, E., Swan, J., & Feiner, S. (2010). *Perceptual Issues in Augmented Reality Revisited.* Paper presented at the International Symposium on Mixed and Augmented Reality.

Kurata, T., Sakata, N., Kourogi, M., Kuzuoka, H., & Billinghurst, M. (2004). *Remote Collaboration using a Shoulder-Worn Active Camera/Laser.* Paper presented at the International Semantic Web Conference Arlington, USA.

Kuzuoka, H. (1992). *Spatial workspace collaboration: a sharedview video support system for remote collaboration capability.* Paper presented at the Human Factors in Computing Systems.

Kuzuoka, H., Kosuge, T., & Tanaka, M. (1994). *GestureCam: Platform for Augmented Reality Based Collaboration* Paper presented at the Artificial Reality and Telexistence.

Laine, S., & Karras, T. (2010). Efficient Sparse Voxel Octrees. *SIGGRAPH*

Lamond, B., Peers, P., Ghosh, A., & Debevec, P. (2009). *Image-based Separation of Diffuse and Specular Reflections using Environmental Structured Illumination.* Paper presented at the International Conference on Computational Photography.

Lee, M., Green, R., & Billinghurst, M. (2008). *3D natural hand interaction for AR applications.* Paper presented at the Image and Vision Computing New Zealand.

Lee, T., & Hollerer, T. (2007). *Handy AR: Markerless inspection of augmented reality objects using fingertip tracking.* Paper presented at the International Symposium on Wearable Computers, Washington DC, USA.

Lee, T., & Hollerer, T. (2008). *Hybrid Feature Tracking and User Interaction for Markerless Augmented Reality.* Paper presented at the Virtual Reality Reno, Nevada, USA.

Levoy, M. (2010). *Experimental Platforms for Computational Photography*

Lincoln, J. (2010). The latest video projectors can fit inside tiny cameras or cellphones yet still produce big pictures. *Spectrum 47*(5), 41-45. doi: 10.1109/MSPEC.2010.5453140

Liu, N., Lu, Z., Zhao, J., McDowell, M., Lee, H., Zhao, W., & Cui, Y. (2014). A pomegranate-inspired nanoscale design for large-volume-change

lithium battery anodes. *Nature Nanotechnology, 9*, 187-192. doi: 10.1038/nnano.2014.6

Loomis, J., Golledge, R., & Klatzky, R. (1993). *Personal guidance system for the visually impaired using GPS, GIS, and VR technologies.* Paper presented at the Virtual Reality and Persons with Disabilities, Millbrae, CA.

Lothridge, K., & Fitzpatrick, F. (2013). *Crime Scene Investigation: A guide for law enforcement* (Vol. 1). USA, Largo: National Forensic Science Technology Center.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91-110.

Lui, S., & Cooper, D. (2011). A Complete Statistical Inverse Ray Tracing Approach to Multi-View Stereo. *Computer Vision and Pattern Recognition*.

Lukosch, S., Billinghurst, M., Alem, L., & Kiyokawa, K. (2015). *Collaboration in Augmented Reality.* Paper presented at the Computer Supported Cooperative Work.

Lukosch, S., Lukosch, H., Datcu, D., & Cidota, M. (2015). *Providing Information on the Spot: Using Augmented Reality for Situational Awareness in the Security Domain.* Paper presented at the Computer Supported Cooperative Work.

Lukosch, S., Poelman, R., Akman, O., & Jonker, P. (2012). *A Novel Gesture-based Interface for Crime Scene Investigationin Mediated Reality.* Paper presented at the Computer Supported Cooperative Workshop.

Maloney, A., Allen, B., Bardell, B., Collings, S., Gradkowski, A., Maloney, K., & Ritter, C. (2009). HemoSpat Validation *Forident*

Mann, S. (2003). Introduction to Mediated Reality. *Human-Computer Interaction, 15*(2), 205-208. doi: 10.1207/S15327590IJHC1502_1

Manning, P. (2008). *Crime Mapping, Information Technology, and the Rationality of Crime Control*: New York University Press.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology *Descision Suppert Systems, 15*, 551-266.

Marwah, K., Wetzstein, G., Bando, Y., & Raskar, R. (2013). *Compressive Light Field Photography using Overcomplete Dictionaries and Optimized Projections.* Paper presented at the SIGGRAPH.

Meager, D. (1982). Geometric modeling using octree encoding. *Computer Graphics & Image Processing 14*(2), 129-147.

Mei, C., Sibley, G., Cummins, M., Newman, P., & Reid, I. (2010). RSLAM: A System for Large-Scale Mapping in Constant-Time using Stereo. *International Journal of Computer Vision*, 1-17.

Milgram, P. (2006). *Some Human Factors Considerations for Designing Mixed Reality Interfaces.* Paper presented at the Virtual Media for Military Applications, Neuilly-sur-Seine, France.

Milgram, P., & Colquhoun, H. (1999). A Taxonomy of Real and Virtual World Display Integration. *International Symposium of Mixed Reality, 1*(Mixed Reality - Merging Real and Virtual Worlds), 5-30.

Milgram, P., & Kishino, F. (1994). A Taxonomy of Mixed Reality Visual Displays. *Information and Systems, E77-D*(12), 13--21.

Milner, A. (1998). *The Visual Brain in Action* Oxford University Press.

Mistry, P., Maes, P., & Chang, L. (2009). *WUW - Wear Ur World - A Wearable Gestural Interface.* Paper presented at the Human-Computer Interaction, Boston, USA.

Mutual, L. (2004). Manual Materials Handling Guidelines.

Newcombe, R. A., Lovegrove, S. J., & Davison, A. J. (2011). DTAM: Dense Tracking and Mapping in Real-Time. *International Conference on Computer Vision*.

Ng, R. (2006). *Digital light field photography.* PHD, Stanford University.

Nilsson, S., Johansson, B., & Jonsson, A. (2009). *A co-located collaborative Augmented Reality application.* Paper presented at the Virtual-Reality Continuum and its Applications in Industry, Yokohama, Japan.

Norretranders, T. (1999). *The User Illusion: Cutting Consciousness Down to Size*

Nuvan, L. (2009). Time of Flight Camera Technology Zurich: Centre suisse d`electronique et de microtechnique.

O'Sullivan, D., & Igoe, T. (2004). *Physical Computing: Sensing and Controlling the Physical World with Computers*.

Pan, Z., Zhigeng, A., Yang, H., Zhu, J., & Shi, J. (2006). Virtual reality and mixed reality for virtual learning environments. *Computers & Graphics, 30*(1), 20--28.

Patterson, R., Winterbottom, M., Pierce, B., & Fox, R. (2007). Binocular Rivalry and Head-Worn Displays. *Human Factors, 69*(6), 1083-1096.

Pekkola, S. (2002). *Critical approach to 3D virtual realities for group work.* Paper presented at the Nordic conference on Human-computer interaction, New York, USA.

Peng, T., & Gupta, S. (2007). Model and algorithms for point cloud construction using digital projection patterns. *Journal of Computing and Information Science in Engineering, 7*(4), 372-381.

Persa, S., & Jonker, P. (2000). *Human-computer Interaction using Real Time 3D Hand Tracking.* Paper presented at the Information Theory in the Benelux, Wassenaar, Netherlands.

Piekarski, W., & Thomas, B. (2001). *Tinmith-Metro: New outdoor techniques for creating city models with an augmented reality wearable computer.* Paper presented at the International Symposium on Wearable Computers, Zurich, Switzerland.

Piekarski, W., & Thomas, B. (2002). *Using artoolkit for 3d hand position tracking in mobile outdoor environments.* Paper presented at the Augmented Reality Toolkit, Workshop.

Poelman, R., Akman, O., Lukosch, S., & Jonker, P. (2012). *As if being there: mediated reality for crime scene investigation.* Paper presented at the Computer Supported Cooperative Work, NY.

Poelman, R., & Fumarola, M. (2009). *Introducing a selection method of game engines for computer supported serious games.* Paper presented at the International Simulation And Gaming Association, Singapore.

Poelman, R., Rusak, Z., Verbraeck, A., & Alcubilla, L. S. (2010). The effect of visual feedback on learnability and usability of design methods. *Tools and Methods of Competitive Engineering*, 1487-1496.

Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.-F., Kirbas, C., . . . Ansari, R. (2002). Multimodal human discourse: gesture and speech. *Computer-Human Interaction*(9), 171–193.

Rekimoto, J. (1996). *Transvision: A Hand-Held Augmented Reality System for Collaborative Design.* Paper presented at the Virtual Systems and Multimedia, Gifu, Japan.

Remondino, F., & El-Hakim, S. (2006). Image-based 3D modelling: A review. *The Photogrammetric Record, 21*(115), 268-291.

Rocha, R., & Araújo, R. (2010). *Selecting the Best Open Source 3D Games Engines* Paper presented at the SBGames

Rohrer, M. (2000). *Seeing is believing: the importance of visualization in manufacturing simulation.* Paper presented at the Winter Simulation Conference, Orlando.

Rolland, J., Biocca, F., Hamza-Lub, F., Ha, Y., & Martins, R. (2005). Development of Head-Mounted Projection Displays for Distributed, Collaborative, Augmented Reality Applications. *Presence, 14*(5), 528-549.

Rolland, J., Ha, Y., & Fidopiastis, C. (2004). Albertian errors in head-mounted displays: I. Choice of eye-point location for a near- or far-field task visualization. *Optical Society, 21*, 901-912.

Rosten, E., Porter, R., & Drummond, T. (2010). FASTER and better: A machine learning approach to corner detection. *Pattern Analysis and Machine Intelligence 32*, 105-119.

Rusinkiewicz, S., & Levoy, M. (2001). *Streaming QSplat: a viewer for networked visualization of large, dense models.* Paper presented at the symposium on Interactive 3D graphics, New York, USA.

Rutter, D. (1987). *Communicating by telephone*. New York: Pergamon Press.

Santos, P., Gierlinger, T., Stork, A., & McIntyre, D. (2007). *Display and rendering technologies for virtual and mixed reality design review* Paper presented at the International Conference on Construction Applications of Virtual Reality, Penn State, PA, USA.

Saxena, A., Sun, M., & Ng, A. (2007). Learning 3-D Scene Structure from a Single Still Image. *International Journal of Computer Vision*.

Scharstein, D., & Hirschmüller, H. (2007). *Evaluation of cost functions for stereo matching.* Paper presented at the Computer Vision and Pattern Recognition, Minneapolis, MN.

Scharstein, D., & Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision, 47*(3), 7-42.

Scheer, F., Abert, O., & Muller, S. (2007). *Towards Using Realistic Ray Tracing in Augmented Reality Applications with Natural Lighting.* Paper presented at the Workshop Virtual and Augmented Reality

Schlattmann, M., & Klein, R. (2007). *Simultaneous 4 gestures 6 DOF real-time two-hand tracking without any markers.* Paper presented at the Virtual Reality Software and Technology.

Schlattmann, M., Nakorn, T., & Klein, R. (2009). *3D Interaction Techniques for 6 DOF Markerless Hand-Tracking.* Paper presented at the International Conference on Computer Graphics, Visualization and Computer Vision.

Schnieders, D., Fu, X., & Wong, K. (2010). *Reconstruction of Display and Eyes from a Single Image.* Paper presented at the Computer Vision and Pattern Recognition, San Francisco, CA

Shankland, S. (2007). Adobe shows off 3D camera tech. *CNET*.

Silva, C., Chiang, Y., Corrêa, W., El-sana, J., & Lindstrom, P. (2002). Out-of-core algorithms for scientific visualization and computer graphics *Visualization*

Spada, H., Meier, A., Rummel, N., & Hauser, S. (2005). *A new method to assess the quality of collaborative process.* Paper presented at the Computer Supported Collaborative Learning.

Stenger, B., Mendonc, P., & Cipolla, R. (2001). Model-Based 3D Tracking of an Articulated Hand. *Computer Vision and Pattern Recognition, 2*, 310-315.

Stevens, R., & Harvey, T. (2002). Lens arrays for a three-dimensional imaging system. *Optics, 4*, 17-21.

Strecha, C., Bronstein, A., Bronstein, M., & Fua, P. (2010). LDAHash: Improved matching with smaller descriptors. *Pattern Analysis and Machine Intelligence*.

Stricker, D., Vigueras-Gomez, J., Gibson, S., & Ledda, P. (2004). *Photorealistic Augmented Reality.* Paper presented at the International Symposium on Mixed and Augmented Reality, Arlington, VA, USA.

Sutherland, I. (1965). *The Ultimate Display.* Paper presented at the Information processing.

Sutherland, I. (1968). *A head-mounted three dimensional display*. Paper presented at the Joint Computer Conference, San Francisco, California

Szalavari, Z., Schmalstieg, D., Fuhrmann, A., & Gervautz, M. (1998). Studierstube: An Environment for Collaboration in Augmented Reality. *Virtual Reality, 3*(1), 37--49.

Tait, M., & Billinghurst, M. (2015). *The Effect of View Independence in a Collaborative AR System.* Paper presented at the Computer Supported Cooperative Work.

Takeuchi, T., Wada, T., Mukobaru, M., & Doi, S. (2007). *A Training System for Myoelectric Prosthetic Hand in Virtual Environment.* Paper presented at the Complex Medical Engineering, Beijing, China.

Tola, E., Lepetit, V., & Fua, P. (2010). *DAISY: An Efficient Dense Descriptor Applied to Wide Baseline Stereo.* Paper presented at the Pattern Analysis and Machine Intelligence.

Venman, T. (2015). The wearables report: Growth trends. Web: Business Insider.

Vogel, D., & Balakrishnan, R. (2005). *Distant Freehand Pointing and Clicking on Very Large, High Resolution Displays*. Paper presented at the Symposium on User Interface Software and Technology Seattle, Washington, USA.

Vosinakis, S., Koutsabasis, P., Stavrakis, M., Viorres, N., & Darzentas, J. (2008). Virtual environments for collaborative design: requirements and guidelines from a social action perspective. *CoDesign International Journal of CoCreation in Design and the Arts 4* (3), 133–150.

Wachs, J. P., Kölsch, M., Stern, H., & Edan, Y. (2011). Vision-Based Hand-Gesture Applications *Communications of the ACM 54*(2), 60-71.

Wagner, D. (2007). *Handheld Augmented Reality.* PHD, Graz, Austria.

Wagner, D., & Schmalstieg, D. (2003). *First Steps Towards Handheld Augmented Reality.* Paper presented at the International Symposium on Wearable Computers, Washington, DC, USA.

Walden, D., & Roedler, G. (2015). INCOSE Systems Engineering Handbook (pp. 304).

Walker, B., Stanley, R., Iyer, N., Simpson, B., & Brungart, D. (2005). *Evaluation of bone-conduction headsets for use in multitalker communication environments.* Paper presented at the Human Factors and Ergonomics Society.

Wand, M., Berner, A., Bokeloh, M., Jenke, P., Fleck, A., Hoffmann, M., . . . Seidel, H. (2008). Processing and Interactive Editing of Huge Point Clouds from 3D Scanners. *Computers and Graphics, 32*(2), 204-220.

Wandell, B. (1995). *Foundations of Vision* Sinauer Associates.

Wang, R., & Popovic, J. (2009). *Real-Time Hand-Tracking with a Color Glove.* Paper presented at the Graphics.

Wang, Y., & MacKenzie, C. (2000). *The role of contextual haptic and visual constraints on object manipulation in virtual environments.* Paper presented at the Human-Computer Interaction.

Wann, J., Rushton, S., & Mon-Williams, M. (1994). Natural Problems for Stereoscopic Depth Perception in Virtual Environments. *Vision Research, 35*(19), 2731-2736.

Watanabe, M., Nayar, S., & Noguchi, M. (1996). *Real-time computation of depth from defocus.* Paper presented at the The International Society for Optical Engineering.

Watson, A. (1986). *Temporal sensitivity, Handbook of Perception and Human Performance*: Wiley.

Wehr, A., & Lohr, U. (1999). Airborne laser scanning - an introduction and overview *Journal of Photogrammetry and Remote Sensing, 54*(2-3), 68-82.

Weiser, M. (1991). The Computer for the 21st Century. *Scientific American, 265*(3), 94--104.

Weiss, S. (2008). *Forensic Photography: Importance of Accuracy* Prentice Hall.

Welten, B. (2004). Spelverdeler in de opsporing: Projectgroep Forensische Opsporing, Raad van Hoofdcommissarissen.

Wexelblat, A. (1998). *Research challenges in gesture: Open issues and unsolved problems.* Paper presented at the Gesture and Sign Language in Human-Computer Interaction.

Whelan, t., Kaess, M., Fallon, M., Johannsson, H., Leonard, J., & McDonald, J. (2012). *Kintinuous: Spatially Extended KinectFusion.* Paper presented at the RGB-D Workshop Massachusetts.

Winkler, T., Kritzenberger, H., & Herczeg, M. (2003). *Mixed Reality Environments as Collaborative and Constructive Learning Spaces for Elementary School Children*.

Woyke, E. (2011). Identifying The Tech Leaders In LTE Wireless Patents: Forbes.

Wu, C., Wilburn, B., Matsushita, Y., & Theobalt, C. (2011). High-quality shape from multi-view stereo and shading under general illumination. *Computer Vision and Pattern Recognition*.

Yoganandan, N., Pintar, F., Zhang, Z., & Baisden, J. (2009). Physical properties of the human head: Mass center of gravity and moment of inertia. *Journal of Biomechanics, 42*, 1177–1192. doi: 10.1016/j.jbiomech.2009.03.029

Zhang, R., Tsai, P., Cryer, J., & Sha, M. (1999). Shape from Shading: A Survey. *Pattern Analysis and Machine Intelligence, 21*(8), 690-706.

Zhang, Z., Liu, Z., Sinclair, M., Acero, A., Deng, L., Droppo, J., . . . Zheng, Y. (2004). *Multi-sensory microphones for robust speech detection, enhancement and recognition.* Paper presented at the Acoustics, Speech and Signal Processing.

Zwern, A. (1995). *Virtual Reality: State-Of-The-Art and Key Challenges*. Paper presented at the Wescon Conference Record California (Horváth).

SUMMARY

This thesis is concerned with the creation of a hard- and software artifact that allows professionals to collaborate in a digitally augmented physical space, otherwise known as meditated reality or physical computing. On-premise professionals are supported with sensors and human computer interaction capabilities while freely walking around. A significant amount of time was devoted to the creation of this artifact, but the main goal is the development of a new way of collaboration.

The use case is introduced in the first chapter, namely crime scene investigation. The use case allows mediated reality to show its potential. A crime scene is a unique pristine environment. Although 3D models of the environments may exist, they do not reflect reality. Many spatial related tasks take place at crime scenes; line of sight verifications, bullet trajectory analysis and blood pattern analysis. Preferably, research on crime scenes should be contactless, as the contamination of a crime scene should be avoided at all cost. Furthermore, a shared understanding of a crime scene is important, due to the number of people and types of expertise involved in crime scene investigation. This chapter introduces the domain disciplines that flank mediated reality research, such as computer vision, real-time rendering, human computer interaction and computer supported collaborative work. The chapter concludes with the research philosophy and the research questions. The main research question to be answered is:*How can we support collaborative spatial interaction in a pristine environment applying mediated reality?*

Chapter two primarily deals with gathering the requirements to build the system, based on the actual work conducted by experts today. After receiving demonstrations with preliminary research results, the professionals in the field were interviewed. The interview results are supplemented with literature research and requirements are abstracted. The requirements fall into different buckets, such as: ergonomic (a weight limit), safety (non-contact measurements) and socio-technical (collaboration capabilities). Important requirements include the sharing of 3D acquired data, interaction paradigms for collaboration and logging.

In chapter three, it is explored how to narrow down the solution space between the overlapping research domains. In this chapter, the requirements and research question are analyzed to focus the literature research. An abstract architecture is developed that summarizes the major components of the artifact, see figure (A)

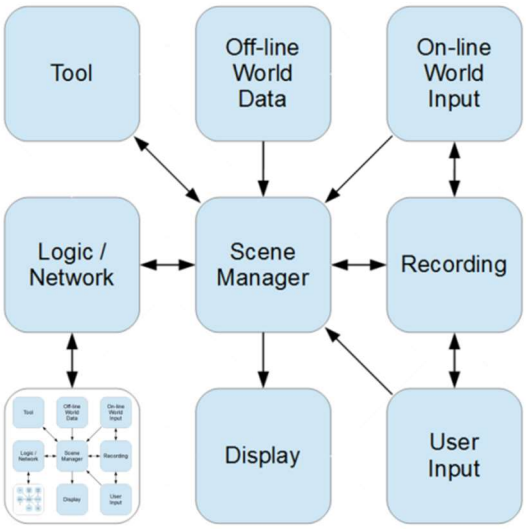


FIGURE (A) MEDIATED REALITY SYSTEM HIGH-LEVEL ARCHITECTURE

Figure (A) shows a comparison between systems that are similar to the intended artifact in chapter four. Analyzed systems include SIXTHSENSE, ARTHUR and SharedView. From researching these systems, it became apparent that the preferred method for spatial interaction is based on head mounted displays and that none of the systems included real-time reconstruction. Furthermore, the collaborative capabilities and human computer interaction functionality still have many unknowns. The follow-up literature study addresses four areas of relevance. Computer vision research has a level of maturity that allows it to be a viable alternative to traditional laser scanning based acquisition methods. Collaborative capabilities in 3D shared spaces are mature, but it is unknown whether the lessons learned translate to mediated reality. The hardware in head mounted displays is also becoming significantly better. The field of view, resolution, refresh rate and device weight have improved considerably, negating much of the previously "negative publicity" garnered by these devices. And last but not least, the human-computer interaction also benefited from the improvements in hardware. An impressive list of recently developed gesture recognition software demonstrates the feasibility of gesture-controlled 3D interaction.

Chapter five starts with the validation of the researched capabilities with the experts. Preliminary prototypes in the four designated research areas were created; a real-time mapping system based on PTAM, a hand tracker for 3D interaction based on Handy AR, an optical see through head mounted display

from our TU Delft and a remote session with the PTAM tracker. Experts validated the research direction without implications for the prototype. Being able to overlay the physical scene with digital content was the base subsystem addressed in the creation of the prototype. An iterative design approach was used to improve on the designs. First, the head mounted display hardware was created in three iterations resulting in a lightweight high resolution head set. Secondly, the subsystems for 3D mapping were addressed. A real-time mapping system was created that improves the scene reconstruction while the user walks around. Thirdly, the interaction paradigm was created and validated with the peer group. Users could use hand gestures to communicate and interact with the 3D scene. Finally, the interaction between peers was validated, to allow a remote investigator and an on-premise expert to collaborate. The result was an, open source, lightweight and high resolution video-see through head mounted device with overlay and interaction capabilities. The chapter concludes with mapping the results of the experiments to the requirements from chapter two.

Chapter six is used to answer sub- research questions. The software architectures of prior systems are compared to the artifact from chapter 5 and the completeness of the system is illustrated by being able to map all prior systems to our architecture. This architecture allows us to design many collaborative interaction paradigms and integrates well with modern hardware. Next, the interaction of a user with an augmented overlay is addressed. The artifact allows users to manipulate 3D elements within the augmented scene by using gesture control. 3D elements can be placed in the scene, moved and removed while wearing the head mounted display. Even complex interactions such as adding elements beyond arms' reach are made possible. The experiment is also valid for participants not wearing the head mounted display. Lastly, both the remote and on- premise participants were asked to work on the same scene and add, move and remove elements to demonstrate the collaborative aspect of the artifact. A dummy crime scene was created that allowed for a "realistic" experiment in which they were jointly required to fulfill a technical collaborative objective.

In chapter 7, the key research question is addressed. To validate whether the created artifact satisfies its intended purpose, a setup was created in the forensic field lab. In the setup, an on-premise novice was required to investigate a crime scene aided by a remote expert. The artifact was used to conduct the investigation, see Figure (B).

FIGURE (B) EXPERIMENT SETUP FORENSIC FIELD LAP

The remote experts could see what the novice was seeing on a large screen in either 2D video or as a 3D scene. A webcam on the screen could detect hand gestures for interacting with the novice and scene. The novice was wearing the head mounted display and mapped the environment. His hand gestures could be detected by the cameras on the head mounted display. They had to fulfill a few 3D tasks, during which the expert guided the novice through the process. While the system functioned as expected and answered the research questions, many unanswered questions remain for further research.

In chapter eight, the epilogue, the implications and further research are discussed. The idea for such a system is by no means new: it was tried by Sutherland as early on as in 1968. However, the technology has progressed significantly and remote assistance is the norm for many things nowadays. Researchers and developers might have cracked the 2D interface, but the 3D augmented interface is still pretty much in its infancy: even simple things like intuitive selecting are still hard. With the artifact created in this thesis, a holistic system is created that demonstrates an integration of many components and pushes boundaries on remote collaboration.

Ronald Poelman

CURRICULUM VITAE

Ronald received his bachelor's degree (2000) in Industrial Design Engineering from The Hague University of Applied Sciences. He earned a master's degree from the Open University of the Netherlands (Prof. Dr. Ir. Jack Gerrissen, 2002), located in the University of Utrecht.

He started his professional career by working for a remote sensing startup in Delft. As the CTO, he launched a successful software department in laser scan processing. He subsequently worked for the Technical University Delft as assistant Professor in Mixed Reality. Most of his publications and research are related to technologies empowering computer supported collaborative work.

During his time as an assistant professor, he founded a company in reality capture software together with several others, where he fulfilled the role of CTO. His company was acquired by Autodesk Inc. and he moved to San Francisco to work as a software architect for the reality capture team.

While at Autodesk, he worked on streaming technologies for meshes and point clouds, simultaneous localization and mapping, structure-from-motion and more recently machine learning. In his leadership role, he was technically responsible for multiple products used by millions of customers.

APPENDIX I - QUESTIONNAIRE EXPERT FORM

Expert questionnaire form, reserved time per expert ~90 minutes Introduction provided by discussing the thesis question and preliminary research material.	
Interviewer: Phone/Visit	Name participant:

(A) Participant background questions

A1	For how many years are you involved with 3D crime scene reconstructions?
A2	Is 3D reconstruction a well know technique in your investigation discipline?
A3	What kind of 3D reconstructions are happening most frequently?
A4	How often are 3D reconstructions conducted in your district/city/county?
A5	Who decides if a 3D reconstruction takes place?
A6	Are the prospects who decide on a 3D reconstruction all aware of the potential?

(B) Tool specific open ended questions

B1	What are the strongest motivations to conduct a 3D reconstruction?
B2	Is the entire scene reconstructed or only what's relevant?
B3	How often is the same crime scene measured/3D reconstructed?
B4	What are the biggest inhibitors to conduct a full or part 3D reconstruction? (time/cost/usability/applicability)
B5	Can you explain what you like and don't like about current 3D reconstruction equipment?
B6	What's missing or can be improved with respect to collaboration in 3D reconstruction?
B7	What would you like to see improved on interaction with 3D reconstruction?
B8	Can you think of cased were the 3D reconstruction failed to get the desired outcome?
B9	What are the dominant reasons not to 3D reconstruct, aka with respect to other techniques?
B10	Any remarks that you would like to make?

APPENDIX II – QUESTIONNAIRE FOR 3D INTERACTION

Questionnaire 3D user interaction "CSI The Hague" version 0.2	
Name: ..... Age:.....	
Circle correct occupation:	*Academic / project partner / Investigator*
Generic questions:	
Do you have a positive attitude towards new technologies?	Yes / Sometimes / No
How often do you work with 3D software tools?	Daily / Weekly / Monthly / Yearly / Never
Do you play 3D games on console or personal computer?	Daily / Weekly / Monthly / Yearly / Never
Did you ever experience virtual reality or immersive augmented reality?	Yes / No
Questions about the experiment: (Rate the answers and circle most applicable)	
(1a) Do you consider the menu navigation easy/hard to learn?	Easy -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Hard
(1b) Do you consider the menu navigation easy/hard?	Easy -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Hard
(1c) Do you think you can confidently use a tool based on this menu navigation technology?	Likely -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Unlikely
(1d) Did you achieve what you intended to select?	Yes / No
(1e) Do you consider this a suitable mechanism for augmented menu navigation	Yes / No
(2a) Do you consider the placement of 3D objects easy/hard to learn?	Easy -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Hard
(2b) Do you consider the placement of 3D objects easy/hard?	Easy -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Hard
(2c) Do you think you can confidently use a tool based on this augmented 3D placement technology?	Likely -> 1 – 2 – 3 – 4 – 5 – 6 – 7 <- Unlikely
(2d) Did you achieve what you intended to place in 3D?	Yes / No
(2e) Do you consider this a suitable mechanism for augmented 3D placement navigation	Yes / No

APPENDIX III – QUESTIONNAIRE FOR COLLABORATION

		Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Fluidity of collaboration	The team members knew when to speak and when not to speak, during collaborative analysis.					
	During the analysis the team members had a shared focus..					
	There were misunderstandings during the collaboration because it was unclear who leads the conversation.					
	The team members were confused during the analysis as they did not have a shared focus.					
Sustaining mutual understanding	The team members achieved a clear mutual understanding.					
	The team members did not have a shared understanding of the current situation.					
	The team members had a wrong perception of the system state, what tool was active, etc.					
	The team members had different perceptions of the state of the spatial analysis.					
Information exchange for problem solving	The team members generated shared ideas to solve the spatial analysis.					
	The team members jointly followed up on their ideas.					
	The team members faced misinterpretations during the refinement of ideas.					
	The team members were confused and could not follow up their ideas.					

		Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
Argumentation and reaching consensus	The team members followed a constructive argumentation.					
	The team members jointly and inteligibly checked the constraints of the spatial analysis.					
	The spatial analysis confused the team and prevented joint decisions.					
	The team members did not jointly discuss their tasks during the spatial analysis.					
Task and time management	The team members had a clear work pan, i.e. it was clear to them on how to time wise solve the problem.					
	The team members had a clear task division for the joint spatial analysis.					
	The team members were confused about their task divisions when solving the spatial tasks.					
	The team members showed wrong time management.					
Cooperative orientation	The communication between the team members showed that they had a shared goal.					
	While using the tools for the spatial analysis the team members showed that they have a shared goal.					
	During the spatial analysis process the team members lacked of a clear cooperative orientation.					
	The team members suffered from misunderstanding and an unclear cooperative orientation.					
Individual task orientation	The team members encouraged each other.					
	Each team member was clearly committed to solve the spatial analysis task.					
	Each team members was not motivated to solve the spatial analysis task as they suffered from misunderstandings.					
	The team members provided wrong signals and feedback that disrupted each other's task orientation.					