Department of Applied informatics
Faculty of Mathematics, Physics and informatics
Comenius University, Bratislava
study program: Computer Science
branch of study: 9.2.1 Computer Science

Augmented Reality in Cultural Heritage Applications

(Dissertation thesis)

RNDr. Zuzana Haladová

Advisor: doc. RNDr. Milan Ftáčnik, CSc.

Bratislava, 2014

AFFIRMATION

I declare that I have written the dissertation thesis with title
Augmented Reality in Cultural Heritage Applications
on my own using only the literature and sources alleged in the Bibliography.

Bratislava, April 2014

....

Acknowledgment

I would like to thank my supervisor doc. RNDr. Milan Ftáčnik, CSc. for his time spent reviewing and discussing my thesis despite his enormously tight schedule.

I would also like to thank RNDr. Martina Bátorová, PhD. and doc. RNDr. Andrej Ferko, PhD. for their help with proofreading and correcting this thesis.

I am very grateful to Peter, my parents and all my friends for their encouragement and loving support.

Most of all I thank RNDr. Elena Šikudová, PhD. for her supervision. Since my Master's thesis, she has guided me through my studies and research with her inspirational comments and ideas. She helped me enormously in both the research and the writing period of this thesis and without her, this thesis would be a complete mess.

Abstrakt

Nedávny technologický vývoj umožnil vznik nových aplikácií rozšírenej reality v oblasti kultúrneho dedičstva umožňujúcich jeho popularizáciu a vizuálne príťažlivú prezentáciu. Kľúčovým aspektom aplikácií rozšírenej reality je registrácia reálneho a virtuálneho sveta a rozpoznávanie objektov, ktoré majú byť augmentované. V tejto práci navrhujeme nové metódy a techniky na detekciu a registráciu objektov využívajúce kombináciu lokálnych a globálnych príznakov a RGB a RGBD dát. Náš prístup je založený na dôslednej analýze obmedzení a problémov doteraz publikovaných metód. Hlavným prínosom tejto práce je nová metóda na detekciu viacerých inštancií objektov. Náš prístup je robustnejší a menej obmedzujúci ako iné prístupy. Predstavená metóda prekonáva obmedzenia predchádzajúcich metód, konkrétne: časovo náročné predspracovanie, detekciu objektov ležiacich výlučne v rovine kolmej na os kamery alebo obmedzenie na jedinú škálu objektov.

Kľúčové slová: Rozšírená realita, Kultúrne dedičstvo, Múzejný sprievodca, Rozpoznávanie malieb, Detekcia objektov, Registrácia objektov, Lokálne príznaky, Globálne príznaky, Detetkcia viacerých inštancií objektov

Abstract

Recent technology development gives rise to new augmented reality applications in the area of cultural heritage, enabling its popularization and visually attractive presentation. The key aspect of augmented reality applications is registration of real and virtual worlds and recognition of augmented-to-be objects. In this thesis, we propose novel methods and techniques for detection and registration of objects that utilize the combination of local and global features and RGB and RGBD data. Our approach is based on a thorough analysis of constraints and problems occurring in existing works. The main contribution of the thesis is the newly developed method for multiple instance object detection. Our technique is more versatile and robust than the surveyed methods. The presented method overcomes the limitations of other available approaches, namely: time consuming preprocessing phase, lack of support of off-plane rotations and problems in scale variations.

Key words: Augmented reality, Cultural heritage, Museum guide, Classification of paintings, Object detection, Object recognition, Local features, Global features, Multiple instance detection

Contents

List of Figures

Preface

For me personally, a PhD. study was an adventure. During my studies, I have visited about 10 conferences, 2 summer schools and many events focused on the research in computer science or its popularization in the public.

The highlights were the *Eurographics* conference, the *ICVSS summer school* and the *Heidelberg laureate forum*. I had the possibility to meet some of the great researchers in the field of computer vision, graphics and computer science as a whole.

My research in the area of augmented reality was a great journey, with stops in many neighboring problems of computer vision. It also included a lot of cooperation on different projects, for example *Biennale of Architecture* in Venice 2012 or *TEDxBratislava* 2013.

I am deeply grateful for being given the opportunity to explore the world of computer science.

(a) With Ivan Sutherland at *Heidelberg Laureate Forum 2013.*

(b) At *Biennale of Architecture 2012.*

(c) At TU Vienna.

Introduction

Nowadays, there is a growing interest in the augmented reality (AR) as a field of research and at the same time as a domain for developing popular applications. Since the coining of the phrase "augmented reality" in 1990, the area has come a long way from research laboratories and big international companies to pockets of millions of users all over the world. The popularity of the field among young generation also evokes an effort to utilize AR as a tool for education or presentation of art and cultural heritage. Therefore we decided to focus on the augmented reality in the cultural heritage applications.

This PhD. thesis is organized as follows. In the chapter *Introduction to augmented reality* the field of augmented reality is introduced in a context with virtual reality, the key milestones are presented and important application areas are mentioned. In the next chapter *Augmented reality systems and approaches* the aspects of augmented reality systems are briefly introduced with their three important parts: inputs, outputs and accessories. Different classifications of the AR systems are surveyed and a new classification based on the perception of the reality is proposed. The following chapter *Registration in augmented reality* deals with one of the key problems of augmented reality — the registration of the virtual and real worlds. Existing works in areas of visual, mechanical, outside-in/inside-out and dead reckoning approaches are surveyed.

In the chapter *Multiple instances detection* we propose two new methods for the detection and registration of multiple instances of objects utilizing local features which were created to overcome the limitations of previously published methods. Our methods are developed for 2D RGB images and RGBD data, which can be acquired from depth sensors such as Kinect. The method described in the next chapter *Classification and registration of paintings* overcomes another important problem of current object detection and registration methods — the problem of matching of the local features with the big databases of objects which are not final and can be extended consecutively. We propose a combination of local and global features for efficient classification (detection, registration) of fine art paintings which is necessary in museum/gallery guides. In the chapter *Related results* we mention our works and installations which were created utilizing our methods proposed in the previous chapters to present cultural heritage. In the *Conclusions* we summarize our contribution to the field of registration and detection in augmented reality applied to the area of the cultural heritage presentation and sketch some future work.

Chapter 1

Introduction to augmented reality

Augment reality is a popular computer graphics related field, which enables us to combine our reality with the limitless (limited only by our imagination) possibilities of virtual reality (VR). In this section we introduce augmented reality as a research field in the context of the related field of virtual reality, its history (prehistory) and future visions, published papers and different applications.

1.1 Augmented reality vs. virtual reality

The virtual and augmented reality are very close research fields and in spite of the clear definition of both terms, it is sometimes hard for the public to distinguish them. In the reality-virtuality continuum (figure 1.1, defined by Milgram et al. in [84]) we can see that the augmented reality is the variation of the mixed reality which lies between the real and completely virtual environment. Azuma in his paper [5] defines AR as the system that has the following three characteristics:

1. it combines real and virtual,

2. it is interactive in real time,

3. it is registered in 3D.

When defining the virtual reality we have to enclose the 3rd and the 2nd point from the AR definition. Another important aspect of virtual reality is the immersion in the

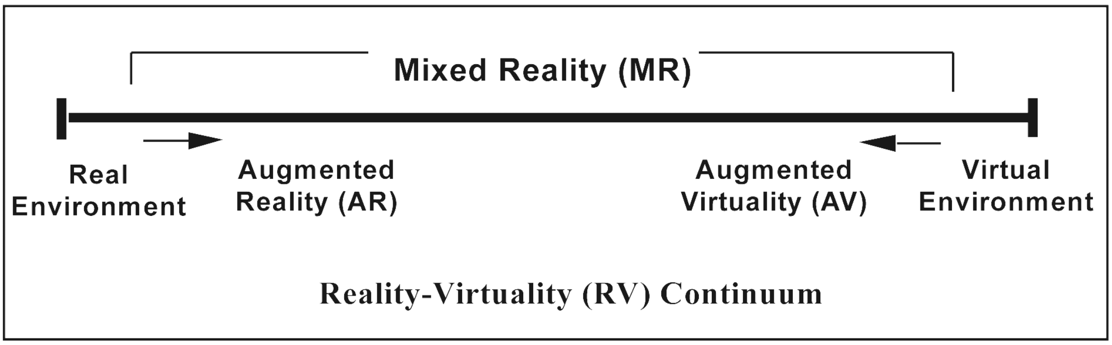


Figure 1.1: Reality-virtuality continuum. Figure is taken from [84].

virtual environment, but when defining the AR, we use the term ultimate immersion because there is nothing more immersive than the reality itself.

Despite the differences, these two fields have connected history. For example Sutherland's Ultimate display [116] or head-mounted display [117] are important milestones in both AR and VR history. The term virtual reality has been used since the '40s of the last century to describe different things, for example theater, but in the '80s, the term virtual reality was coined and popularized by Jaron Lanier (as referring to immersive environments created by applications with visual and 3D effects [74]) and the boom started at the beginning of the '90s. Not long after, the phrase augmented reality was coined by Tom Caudell in 1990 [23] and the boom started with the beginning of the new century.

1.2 Prehistory of the AR

Ivan Sutherland's head-mounted display(HMD) [117] is commonly known as the first AR act, but as we go back in the history we can find (with a little imagination) AR in the magicians' performances in the early 20th century (e.g. Pepper's ghost configuration [20]). They also merged real and virtual, creating ghosts and other effects directly on the stage. The core of the Pepper's Ghost was the actor (dressed like a ghost) whose image was projected through the 45-degree angled semi transparent mirror (beam splitter) upward and on to the stage. The projectionist operated from beneath the stage along with the actor(s). Other characters on the stage would interact with the ghost(s). The sketch of the possible setup in 1863 can be seen on figure 1.2. Although the possibilities of augmented reality are nowadays very broad the Pepper's ghost effect is still popular (the tutorial on Pepper's ghost video on Youtube [10] has more than 101 thousands view's by the 29.1.2014) and different installations utilizing beam splitters can be seen (for example in combination with leap motion device [2]).

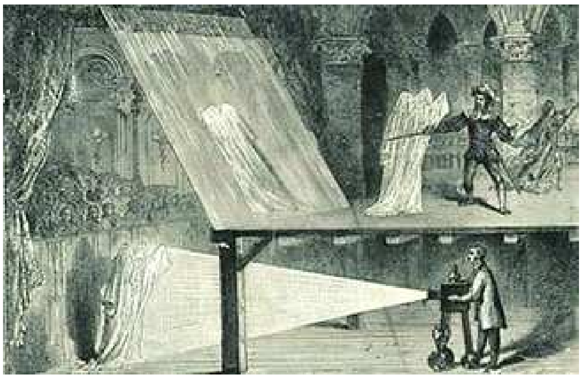


Figure 1.2: Pepper's Ghost as it would have looked in 1863. Figure is taken from [20].

As a significant augmented reality act we can mention Mark II Gyro Gunsight, used in the world war II by the Royal Air Force (RAF). It was first tested in late 1943 with production examples becoming available later in the same year. The Mark II was also subsequently produced in the United States as the K-14 (USAAF) and Mk18 (Navy) [122]. The device was used to project a small shape (cross or circle) in the pilot's field of view. The position of the shape indicated the position of the fire arm's target. The projection of the shape was based on the same principle as the Pepper's ghost with usage of the semi-transparent mirror. For the patent sketch of the device see figure 1.3. A similar device has been patented in USA by Woodson in 1944 (US patent: 2360298) and been further developed in 1960 by American Aviation in US Patent:2950340.

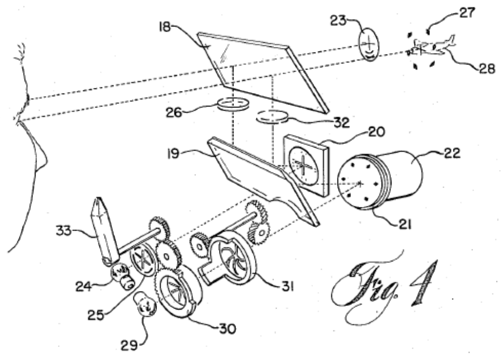


Figure 1.3: Aircraft monitoring system from US Patent:2950340.

1.3 History

After few (pre)historical instances we will focus on the milestones in the history of the augmented reality research field. We will also demonstrate the popularity of the field by showing the evolution of articles in the last 20 years.

1.3.1 Milestones

1966 Ivan Sutherland presented his concept of the ultimate display. His idea however goes further from the virtual and augmented reality we know today. In his paper [116] he marked that: "The ultimate display would, of course, be a room within which the computer can control the existence of matter. A chair displayed in such a room would be good enough to sit in. Handcuffs displayed in such a room would be confining, and a bullet displayed in such a room would be fatal". This act is considered the first AR interface. In the 1968 Sutherland presented his popular head-mounted display [117].

1975 Myron Krueger experimented with computer generated art and interaction. In the Video place project, a computer responded to the gestures and interpreted them into actions. Audience could interact with their own silhouettes generated from the video [69].

1978 Professor Steve Mann at the Department of Electrical and Computer Engineering at the University of Toronto is wearing the HMD (or HUD) since 1978. In 2001 Peter Lynch shot about him the film called Cyberman[1]. Much of the film was created by Mann himself with his EyeTap [80]. EyeTap is the HUD (heads-up display mounted in glasses) which records the reality with the camera, creates an virtual information and merges the reality seen by the user with a virtual information using beam splitter.

1990 Tom Caudell, the researcher who developed the augmented reality system supporting the aircraft manufacturing in the Boeing factory [23], coined the phrase augmented reality.

1991 The concept of the ubiquitous computing was presented by Weiser [130] in the beginning of the '90s. The goal of the ubiquitous computing is to provide computer interface which is natural for the users, to develop the computers which are not visible but "omnipresent", the computers indistinguishable from everyday life. This concept is closely connected to the possibilities and techniques of the augmented reality and the fusion of the fields is known as the ubiquitous augmented reality.

1993 The CAVE: Audio Visual Experience Automatic Virtual Environment was presented to the public.

1993 Steven Feiner, Blair MacIntyre et al. published two major AR papers, one in the Communications of ACM and the other in UIST. The first paper [35] presents the KARMA (knowledge based augmented reality for maintenance assistance) system which uses the optical see through head-mounted display that "explains simple end-user laser printer maintenance". The second paper [34], presents 2D information windows in the AR, a technique which is nowadays broadly used in smartphone (pseudo) AR systems (see figure 1.4. For an example of modern windows on the world application Metro Paris Subway [96]).

[1] wearcam.org/cyberman.htm

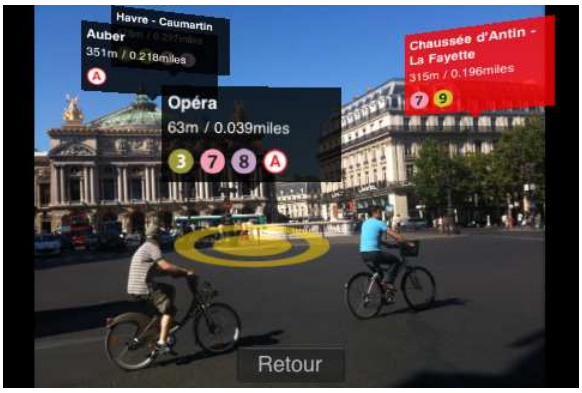


Figure 1.4: Metro Paris Subway iPhone and iPod Touch Application. The modern example of the Window on the world system, proposed in [34]. Figure is taken from [96].

1997 Ronald T. Azuma published the first survey [5] on AR. In the paper he gave the definition of augmented reality which is considered the most relevant. He also named the biggest problems of AR as the registration and the sensing errors. The paper presents a broad survey of different applications of AR in medical, manufacturing, visualization, path planning, entertainment and military fields.

1998 The first augmented reality conference International Workshop on Augmented Reality (IWAR 98) was held in San Francisco [131]. After 2 years the IWAR conference was replaced by the International Symposium on Mixed Reality (ISMR) and the International Symposium on Augmented Reality (ISAR) conferences. In 2002 the ISMAR conference has substituted the forerunners.

around 1998 Sport's augmentation starts in the television broadcasting. For more information see subsection 1.5.6.

1999 ARToolkit was developed by H. Kato in the Nara Institute of Science and Technology. In 1999 Kato and Billinghurst published their paper [61] about using HMD and markers for the conferencing system, based on the method proposed by Rekimoto [100]. ARToolkit is a computer library for the tracking of the visual markers and their registration in the camera space. With the ARToolkit[2] one can easily develop Augmented reality application with the virtual models assigned to different markers. For an example of the application built with the ARToolkit see figure 1.5. Since the release of this toolkit, many different augmented reality toolkits for different programming languages and with different features have emerged (FlarToolkit[3], NyARToolkit[4], Mixed Reality toolkit[5]...).

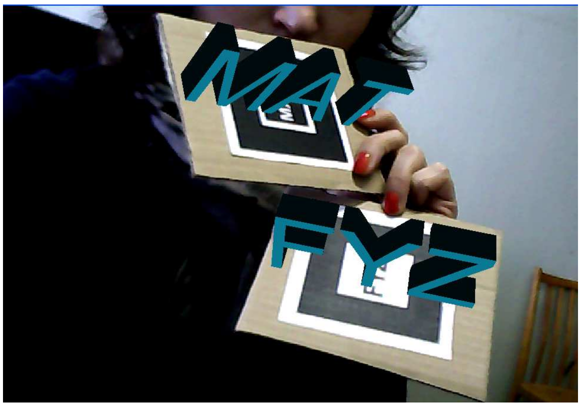


Figure 1.5: An example of the Augmented reality application built with the ARToolkit.

2002 Bruce Tomas developed the first augmented reality outdoor game called ARQuake [123]. It was an AR version of the computer game Quake. Different versions of the system (2000 – 2002) used the optical see through head-mounted display, mobile computer stored in the backpack, haptic gun or handheld device with button, head tracker, digital compass, GPS system and/or markers. It allowed the user to walk around in the real world and shoot virtual enemies from the Quake game. The equipped ARQuake player is shown in the figure 1.6.

2005 Oliver Bimber and Ramesh Raskar published the first book on the spatial AR [15]. In the book the authors describe and categorize augmented reality

[2] http://www.hitl.washington.edu/artoolkit/

[3] http://www.libspark.org/wiki/saqoosha/FLARToolKit/en

[4] http://nyatla.jp/nyartoolkit/wp/

[5] http://www0.cs.ucl.ac.uk/staff/rfreeman/

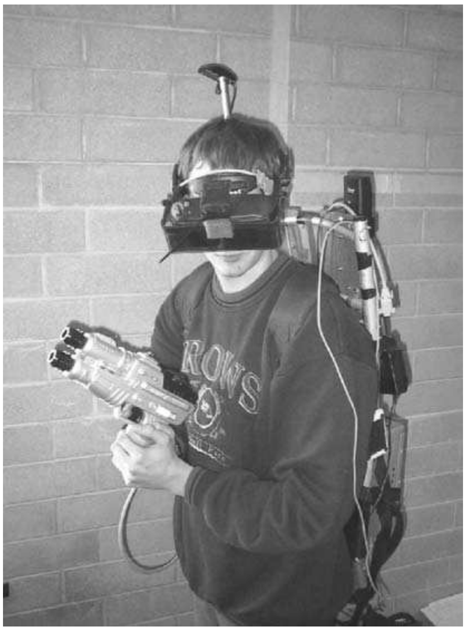


Figure 1.6: The ARQuake player. With HMD, haptic gun, backpack with computer and head tracker. Figure is taken from [123].

systems. They form 3 categories: head-mounted, handheld and spatial and then focus on the spatial systems (SAR). The main difference between spatial AR and other categories is that in the SAR the display is separated from the users of the system and so is suitable for bigger groups of users. SAR systems usually consist of digital projectors which display graphical information directly onto physical objects. Since 2007 the book is available to download free of charge and has been downloaded over 9000 times. In the book authors describe the technique of calibration of several projectors which compensate the inequality and the color of the surface.

2007 Klein and Murray in their paper [64] proposed a method for a markerless tracking for small-workspace augmented reality applications. They track a calibrated handheld camera in a previously unknown scene without any known objects or initialization target, while building a map of this environment.

2009 The Esquire magazine added the AR marker on the first page of their magazine with the hidden virtual Robert Downey Jr.

2009 Although the spatial augmented reality (and the projection mapping techniques) was introduced several years before, the biggest boom in the urban projection mapping was in the 2009/2010. As the most famous examples we have to mention the projection mapping during the 600th years anniversary of Orloj — the astronomical tower clock situated at Old Town Square in center of Prague — in 2010 [120], or the 2009 – 2011 NuFormer Projections in the Netherlands [88].

2010 When Microsoft released Kinect (see figure 1.7), the motion sensing input device for Xbox 360 console, it was expected to be "the birth of the next generation of home entertainment" [119] but not the milestone in the augmented reality history. Kinect sensor developed by PrimeSense company became a really cheap (150 $) source for the depth information for augmented reality applications. The sensor itself consists of the rgb camera, the infrared projector which projects a pattern of dots and the detector which establishes the parallax shift of the dot pattern for each pixel. Kinect holds the Guinness World Record of being the "fastest selling consumer electronics device" (8 million units in its first 60 days). When the first hackers brake into the device and found the way how to control the sensors it took only 2 months and hundreds of augmented reality application using Kinect sensor appeared on the internet. For the best examples see 12 best Kinect hacks [128].

2011 Qualcomm presented Vuforia — the software development platform for augmented reality. Vuforia enables the usage of real-world image markers and development of native applications with support for iOS, Android, and Unity 3D [97].

2012 Czecho-Slovak pavilion on Biennale of Architecture in Venice displayed the first AR installation called Asking Architecture. All the artworks were presented only as a virtual models through augmented reality.

Figure 1.7: The Kinect sensor.

2013 Google introduces Project Glass: a wearable computer combined with an optical head-mounted display (see figure 1.8). Although the original applications developed by Google do not include AR, there are several 3rd party companies and scientists working on augmented reality applications for the Google Glass, for example Open shades [38].

Figure 1.8: Google Glass.

2014 Google launched project Tango[6]. They have created a prototype Android smartphone capable of tracking the full 3D motion of the device and creating a

[6] https://www.google.com/atap/projecttango/

map of the environment. The device like this will allow a precise indoor tracking and registration which can be a breakthrough in the augmented reality.

1.3.2 Evolution of AR articles

As we mentioned in the previous section, the research in the field of the augmented reality has started to grew with the beginning of the new century. We want to demonstrate this expansion with the technique proposed in the Data et al. (2005) [31] to demonstrate the evolution of the articles about the CBIR (content based image retrieval). In figure 1.9 we can see the evolution of the articles focused on the AR since the 1990 until 2013.

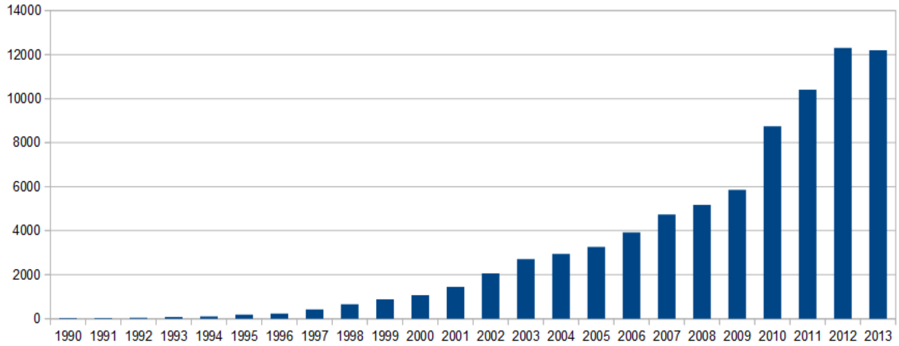


Figure 1.9: The evolution of articles focused on AR since 1990 until 2013. The blue columns represent the articles with the exact phrase 'augmented reality' presented anywhere within the article. The articles were retrieved using Google Scholar search engine on 19.3.2013.

1.4 Future

It was said by Niels Bohr that "Prediction is very difficult, especially about the future". However based on the books, talks and articles by famous researchers in the field of augmented reality we would like to state some of their predictions.

In the beginning of the new century, companies like Information in Place estimated that by 2014, 30% of mobile workers will be using augmented reality. However it is not yet true and AR is more popular in the entertainment and advertisement domain and the emerging of the applications and the growth of the field is visible. Except this prediction, most of the scientists predict the usage of head-mounted devices in favor to handheld devices which are nowadays most popular for AR. Rolf Hainich in his book [41] stated that "we need to eliminate the screen in favor of a near-eye projector, glasses with tiny add-on that could finally weight less than 20g." Oliver Bimber and Rolf Hainich in their book [40] predict the most new displays for AR to be head-mounted (near-eye displays) and spatial displays (printable displays, e-paper, true 3D displays) and they also predict increase in brain-computer interfaces BCI. BCI is a direct communication pathway between the brain and an external device.

Steve Mann, who wears the HMD since 1987 stated in the article about the risks of wearing the HMD everyday for IEEE Spectrum in 2013 "...there is a darker side: Instead of acting as a counterweight to Big Brother, could this technology just turn us into so many Little Brothers, as some commentators have suggested?... I believe that like it or not, video cameras will soon be everywhere: You already find them in many television sets, automatic faucets, smoke alarms, and energy-saving light bulbs. No doubt, authorities will have access to the recordings they make, expanding an already large surveillance capability."

Based on these predictions it seems the future lies in the ubiquitous reality which will be mostly carried out by near-eye displays.

1.5 Applications

In the following section we want to present examples of most common or popular AR applications.

1.5.1 Entertainment

In the area of entertainment we can recognize several basic types of applications. The first one is the handheld (or smartphone games) usually using printed visual markers, 3D registration using PTAM (see section 3.1.2) or pseudo-AR combination of GPS locators and compass for the registration of the virtual reality content position. When talking about smartphones we have to mention IOS by Apple and the Android OS as the most popular platforms. In the 3$^{\text{rd}}$ quarter of 2013 Android achieved the market share of 81% of sold smartphones and IOS 12,9% [17]. Among others there is Windows Mobile, Symbian, Blackberry and Java ME. Augmented reality games on these platforms are nowadays very popular (more than 240 results on the phrase augmented reality game on the Android market and more than 500 application on the App store on 9.1.2013). Second type of AR games encloses the multiplayer HMD games, among them ARQuake [123]. The third type includes games on spatial displays, usually computer stations with monitor and webcam (designed for one user), but sometimes also spatial setups for more users.

1.5.2 Education

Education is very promising application area for augmented reality. The potential of the collaboration was recognized and today there is also an annual international conference on „Virtual and Augmented reality in Education" called VARE. Great attention in the education field is paid to serious games [115, 81] and collaborative AR environments (for example in the geometry learning [62]).

1.5.3 Cultural heritage

Cultural heritage is one of the areas were augmented reality makes a big contribution. Virtual museums were known since the '90s however the augmented reality brings new possibilities into the field. Instead of touring the virtual museum at home on your computer or in the big kiosk in the museum, you can walk through the museum or gallery and watch the real and the virtual exponates side by side, or augmented together (see figure 1.10). The greatest advantage of the virtual/augmented museum experience lies in the possibility of exposition of the lost, damaged or never constructed cultural heritage objects and scenes.

The Museum of London has created an AR project called Streetmuseum, which allows you to see the old photographs of London, augmented in the 3D world, right on the spot where they have been photographed [121].

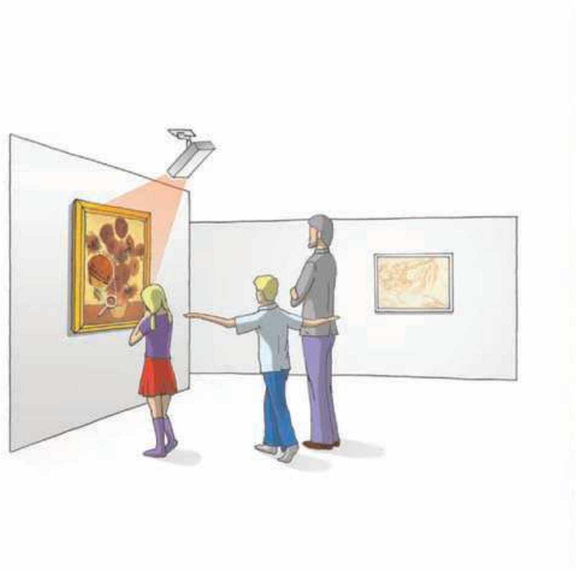


Figure 1.10: Concept sketch of Augmented painting. Figure is taken from [13].

In this category we have to mention also museum guides [19, 113, 85, 72], augmented exponates [14, 13] and urban projection mapping [120, 88].

Museum guides

The field of augmented reality museum applications is mostly focused on extending the information about exhibits with virtual textual or visual information. There are two different methods on extending the common exponates. The first technique uses the head-mounted or handheld display (smartphone or tablet) to provide the individual visitor with the museum guide, offering augmented content on paintings or exponates [8], [37]. The second method uses a spatial device (the projector, monitor or hologram) to provide spatial museum guide. However the first method allows the user to watch the augmentation on the different exponates based on his taste, it is not suitable for the interaction of more users. The second method on the other hand is adjusted for the collaboration of more persons but it is also more expensive than a mobile device. The interactive guide system proposed in [72] in 2002 is one of the first augmented reality spatial museum guides. The system consists of a sensing board (Reactable) capable of recognition of multiple objects (equipped with markers) and gestures. It uses the augmented reality (visual and audio) to immerse kids in the exhibition at the museum or gallery.

Museum guides on handheld devices are known since the 1997 and the Cyberguide [1], a mobile context-aware tour guide suitable for the indoor and outdoor environments. The system automatically updated the user's position according to the GPS position (outdoors) or the ID of the nearest infrared sensor (indoors).

In the paper [129] the authors propose an AR system on the PDA with webcam recognizing the ARToolkit tags distributed in the building and thus producing the augmented experience.

On the other hand the head-mounted museum guide was developed by Sparacino in [113]. Sparacino describes the Museum wearable as "a wearable computer which orchestrates an audiovisual narration as a function of the visitor's interests gathered from his/her physical path in the museum and length of stops." The system consists of a lightweight eye-piece display attached to conventional headphones, a small computer inside a shoulder backpack and a custom built infrared location sensors distributed in the museum space. See figure 1.11 for the scheme of the Museum wearable system.

An augmented reality museum guide [85] has been created for exhibition on Islamic art in the Musée du Louvre. The system uses the RFID chips for recognizing the area of the augmentation and the markerless inside-out method utilizing the rotational sensor mounted on the handheld guide for proper registration. After proper registration the virtual objects or animations (created in VRML97) are added.

In [19] the authors propose a system for large- scale museum guidance. The system acquires rough user position utilizing the Bluetooth emitters and receivers instead of commonly used RFID tags and instead of computationally intensive image processing tasks on remote servers or on high-end mobile devices (such as tablet PCs). All computations are carried out directly on mobile phones. The global and local feature vectors are used for identification of the object.

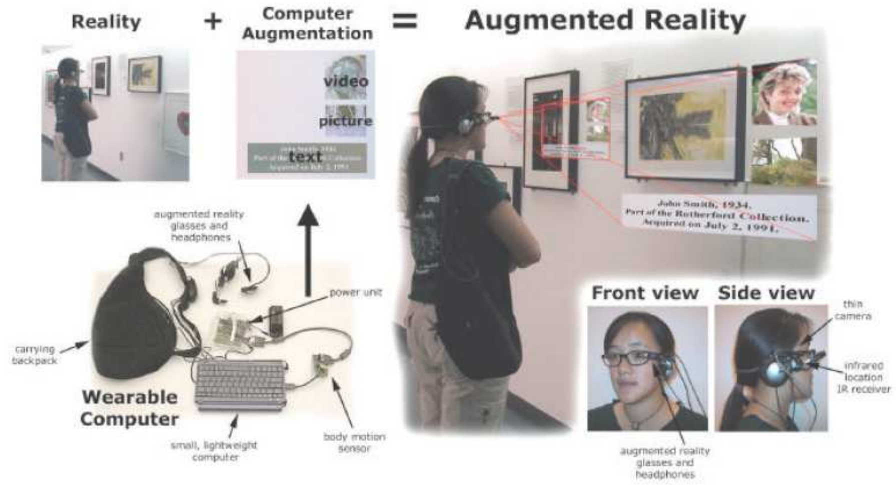

Reality
+
Computer Augmentation
=
Augmented Reality
video
picture
text
John Smith, 1934.
Part of the Rutherford Collection.
Acquired on July 2, 1991.
augmented reality glasses and headphones
power unit
carrying backpack
Wearable Computer
small, lightweight computer
body motion sensor
Front view
Side view
thin camera
infrared location IR receiver
augmented reality glasses and headphones

Figure 1.11: The museum wearable: explanation of concept and application. Figure is taken from [113].

Bay et al. [8] uses a tablet PC as a tool to provide the user with the museum guide with textual information about exhibits detected by affine transformation invariant local features (in this case SURF [9]). System has to deal with a database consisting of 20 exponates. For the real world museum exponates database (thousands) the authors recommend the usage of the Bluetooth locators.

Föckler et al. [37] uses neural networks for the recognition of exhibits and camera equipped smartphone to provide user with the textual information. However the system was tested on 60 objects only and it is not suitable for large scale museum applications.

Exponates Augmentation

In the category of spatial installations we have to mention the work of Oliver Bimber who augments the 2D visual information to the exhibits [13] (in this case Michelangelo's drawings) or extend the exhibit with 3D information in [14] (see Virtual Showcase system on figure 1.12).

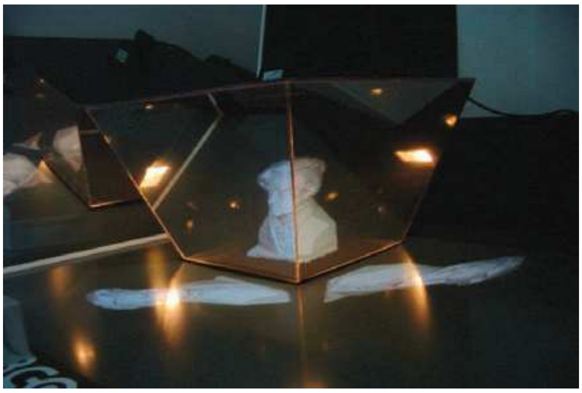


Figure 1.12: Virtul Showcase. Figure is taken from [14].

1.5.4 Sightseeing

The applications of AR in the area of sightseeing are closely connected with the cultural heritage applications. AR tour guides are similar to the museum guides, yet operating in the outdoor space, usually using GPS coordinates for the position estimation (in combination with visual registration to ensure AR experience). The most popular platform for these applications are mobile phones [115, 96]. The information provided by the tour guides usually involve 2D windows with textual and pictorial information about nearest restaurants, subway stations, shops or historical sites.

1.5.5 Design, construction and maintenance

The AR applications in the areas like design, construction and maintenance are about as old as the field itself. The first paper about AR in maintenance was the same in which the phrase augmented reality was coined [23]. There is a lot of research in these areas since. For example the attention is paid on the collaborative AR environments for designers [100], or the utilization of handheld projectors [98] in the maintenance and construction application. At TU Vienna, Kinect is utilized in the setup proposed for fire fighters for monitoring the fire in the buildings[7].

[7] http://augmentedblog.wordpress.com/tag/firefighting/

1.5.6 Sport

The augmented reality information in the sports broadcasting has started around 1998. The typical example are the competitor's national flags placed in the swimming lanes (see figure 1.13) during Olympic games (for the first time in Sydney 2000), or the yellow line in the American football games. The area of augmentation in the sports broadcasting is also analyzed in the papers of Jungong Han [53], [52]. He focused on the analysis of court-net sports.

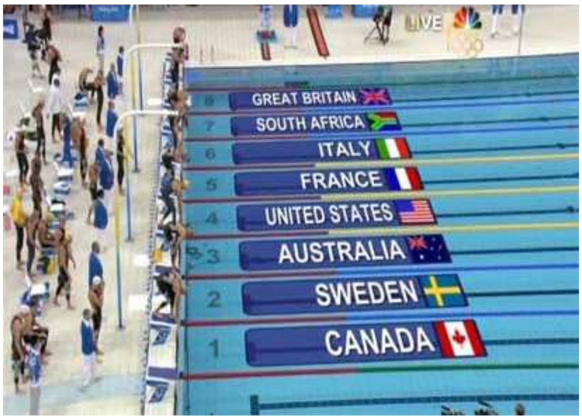


Figure 1.13: Swimming pool with augmented national flags. Figure is taken from [126].

1.5.7 Commercial

The commercial sphere is always engaged into everything new and cool. The augmented reality with its popularity among young generation became one of the ways to present and sell products. Most of these commercial applications take advantage of the augmented reality, just to show their product in the new attractive way. For example there were AR advertising campaigns on trying Rayban glasses [99], or Robert Downey Jr. on the cover of the Esquire magazine [33] and other campaigns created for the companies like Burger King, Mini Cooper, Nestle, Tatrabanka etc.

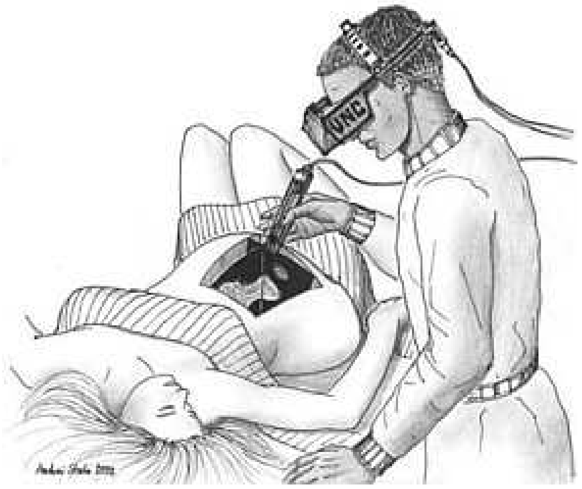


Figure 1.14: Augmented reality medical visualization. Figure is taken from [114].

1.5.8 Medical

A typical AR medical application takes advantage of the 3D model of the inside of human body created from the data acquired with the CT or MRI scanner. These data are then displayed on the human body using the projector, HMD [12], or the monitor with web camera (in the German Cancer Research center the tablet is used for the semi-sterile surgeries[8]) or RGBD camera such as the Kinect (in the Magic mirror project [16] at the Technical University in Munich). A concept sketch of possible AR medical application can be seen in figure 1.14.

There are many teams at universities and hospitals working on the augmented reality research for medical application and since 2001 there is an international workshop on Medical Imaging and Augmented Reality. The review of augmented reality in medicine, can be found in Sielhorst et al. [111].

[8] http://www.medgadget.com/2012/01/intraoperative-ipad-app-shows-where-the-internal-organs-are.html

Chapter 2

Augmented reality systems and approaches

In the first chapter we have defined the augmented reality according to Azuma's definition. There is a discussion in the AR community whether the definition created in '90s still suffices the requirements of the users. Especially in the commercial sphere there exist many applications which are categorized as AR applications, but don't fulfill the second, third or both Azuma's rules. These applications usually lie within the reality-virtuality continuum, but cannot be considered as augmented reality. This lack of true commercial AR leads to misclassification also in some scientific publications.

For example in the big survey [89] published in the proceedings of ISMAR authors decided to include 2 kinds of applications: AR browsers which they defined as: "...usually includes the delivery of points of interest (POI), user-created annotations, or graphics based on the GPS location of the device and orientation of the built-in magnetometer" and image recognition based AR which was defined as: "based on connecting surrounding objects, products, and other physical targets with digital information with the help of visual recognition. By identifying quick response (QR) codes, bar-codes, other graphic markers, or the objects themselves..." In this thesis, we decided to strictly follow Azuma's definition and to call the systems not fulfilling these rules the pseudo-AR.

2.1 AR system

Augmented reality system can consist of many different elements, depending on the type of the application. We can divide these elements in to four categories: *inputs (sensors)*, *outputs (projectors, displays)*, *computers* and *accessories*. It is necessary for every AR system to have at least one sensor for the estimation of the user's position (camera, GPS receiver), one device to display the augmented reality or to add virtual objects into user's view frustum (display, projector) and some device capable of processing of the data (computer).

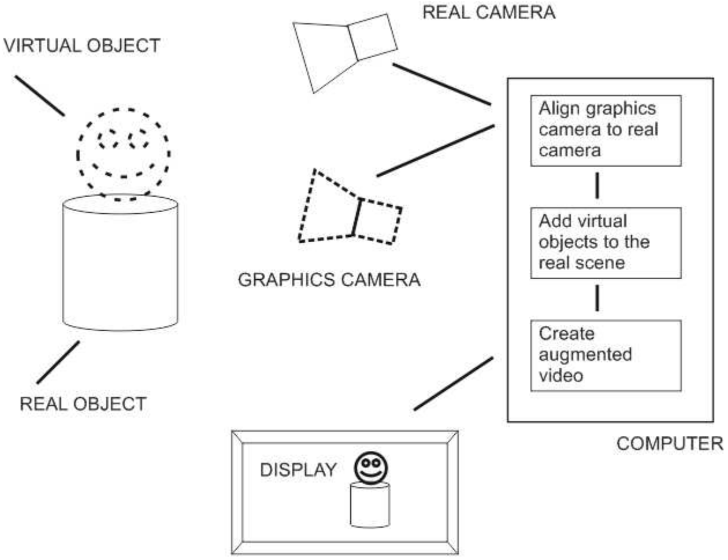


Figure 2.1: A scheme of augmented reality system.

In figure 2.1, we can see the scheme of the common AR system equipped with a camera, a computer and a display. As the first step, the position of the real camera in space has to be estimated and the alignment (registration) of the real camera to the graphics camera has to be done. Visual (or other types of) markers, pattern matching or local features matching are usually used for the estimation of the rotation and translation of the camera to the object to be augmented (we will focus on the registration of the virtual and real camera in the next chapter). The virtual objects are then merged with the real scene and the augmented video is created and displayed.

All the different components necessary for the AR system can be incorporated in one device, for example smartphone, tablet or notebook with a build in webcamera. In the following sections we want to focus on three categories of elements in AR system (inputs, outputs and accessories) and describe members of each category.

2.1.1 Inputs (Sensors)

Sensors used in the AR environments could be of different types, for example optical, acoustic, electric, magnetic, radio, positional and so on. They are mostly used for two main goals: the estimation of the users' and real objects' position in the real environment and the recording of the scene for the purpose of displaying it. The classification of sensors proposed in [103], states these 10 types of sensors: acoustic, biological, chemical, electric, magnetic, mechanical, optical, radiation, thermal and other. Optical sensors typically used in the augmented reality applications are the infra-red cameras, RGB cameras (equipped with the charge-coupled device (CCD) or complementary metal–oxide–semiconductor (CMOS) sensors), monochromatic cameras and the RGBD (RGB camera+ infrared projector and sensor) cameras such as Kinect. As an acoustic sensor the microphone is usually used. The magnetic sensors in AR are the magnetometers. The exemplary mechanical (positional) sensor is the gyroscope and the electric sensors are used in the radio frequency identification (RFID) chip readers.

2.1.2 Outputs

In the domain of AR system outputs we can create the experience for all 5 senses[1]. This concept is also present in the area of VR as a concept called real virtuality presented by Alan Chalmers as: "real virtuality is defined as a true high-fidelity multi-sensory virtual environment that evokes the same perceptual response from a viewer

[1] Classification inspired by www.ted.com/talks/jinsop_lee_design_for_all_5_senses.html

as if he/she was actually present, or "there", in the real scene being depicted. Also known as "there-reality", such environments are interactive and based on physics. All five senses are concurrently stimulated to deliver real world modalities naturally and in real time" [24]. Although in the AR we still perceive the visual output as the most important, we have to also deal with other senses.

Audio

Although, in VR the audio is taken as a key part (games, virtual environments etc.) it is not present in many AR application. Although in the domain of augmented reality, the audio is usually complementary to the visual output like in [46], there are some applications where the audio is the dominant or the only output. An example is the Audio museum guide [133], or the Memento — Google Glass application for visually impaired users [90].

Tactile

In AR environments a touch sense is usually stimulated by the properties of the real objects presented within the scene, however there are efforts to bring the virtual tactile experience to both VR and AR. In [7] the methods which simulate the tactile feedback are divided to: force feedback, actuation of the environment, tangible interfaces and wearable haptics. These methods can be intrinsic (augment the user, altering his tactile perception) vs. extrinsic (integrated in the environment). Authors from Disney's research designed the Revel device which injects a weak electrical signal to the user's body and creates an oscillating electrical field around the user's fingers which is perceived as highly distinctive tactile textures augmenting the physical object. By tracking of the user's fingers and the physical objects dynamic tactile sensations can be associated to the interaction context [7].

Gustatory and olfactory

The research in the field of gustatory and olfactory sensations in augmented or virtual reality is very sparse. This could be due to the complexity of the flavor which is defined by International standards organization as a complex combination of olfactory, gustatory, and trigeminal sensations perceived during tasting [25]. However in [87] the authors have developed the augmented reality display which combines the visual, gustatory and olfactory sensation. In their system the user is tasting the real cookies (with no particular flavor) which serve as eatable markers. While the user is tasting the cookie, the coating (chocolate, strawberry, mushroom) is displayed on the top of the cookie in HMD and the corresponding smell is generated. The research proved that most of the users perceived the difference in taste of different augmented cookies.

Another research utilizing olfactory sensations in the field of computer graphics was done in [18]. The authors investigated the cross-modal effect on the perception of users of computer generated field of grass in the presence of the smell of freshly cut grass. Their research proved that the viewer is not aware of the quality difference of lower quality rendering compared to high quality in presence of the smell of grass.

Visual

The visual output is usually the most important aspect of users' augmented reality experience. We divide visual output devices into two basic categories: projectors and displays. We can classify both, displays and projectors, based on different parameters, the common parameters being the size, the displaying technology or the technology of image production. However for the purpose of this experience based section we will divide all the imaging devices, into stereoscopic and non-stereoscopic devices. Based on these categories we can further divide the stereoscopic displays into autostereoscopic displays and goggle bound displays. The four classes of autostereoscopic displays proposed in [15] are: re-imaging displays, volumetric displays, parallax displays and holographic displays. Two classes of goggle bound spatial displays are: surround screen and embedded screen.

In the domain of goggle bound spatial displays the user has to be equipped with the field-sequential (LCD shutter glasses known as the active stereo technology) or light-filtering (passive stereo technology) goggles. Both methods need the images for left and right eye to appear on the same screen. This technique is known as shutter-

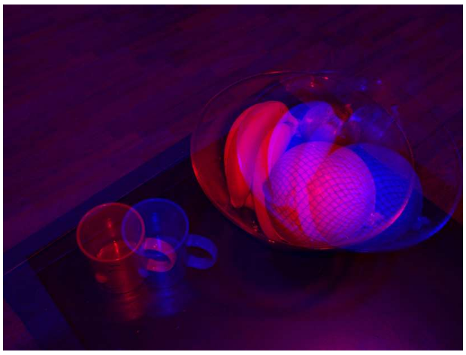


Figure 2.2: An example of anaglyph constructed from images from stereo camera pair.

ing. Depending on the type of the display, images are displayed either sequentially (active) or simultaneously (passive). The passive domain encloses techniques like anaglyphs (see figure 2.2), ChromaDepth, Pulfrich effect or polarization. The active stereo techniques utilize the display, sequentially providing the left and right image synchronized with the LCD goggles sequentially shading right and left glass. For the complete survey of displays, with different categorizations techniques and examples see the book [40].

2.1.3 Accessories

By accessories of the AR system we mean every component which is not enclosed in any previous sections. Additional elements necessary for the projection of the AR (projection screen, half silvered mirrors, beam splitters), components for the tagging of the objects to be augmented (visual paper markers, RFID chips, infrared light-emitting diodes (LED), bluetooth devices) or devices for the interaction (mouse, keyboard, wii remote, touch screen, etc.) are only three important types of accessories. As these accessories are very application-dependent we are not going to describe them in more detail.

2.2 Classification of AR approaches

The augmented reality system can be categorized by different factors, including the application area, the possibility of more persons collaboration or the size of the full system. In the following section we present two different classification schemes of the AR applications. The first one was developed by Bimber and Raskar in [15] and it presents a device based categorization. The second scheme is our own classification based on the way of augmentation of virtual and real world. The user's immersion is the key aspect of the augmented reality systems. Our classification is inspired by the survey from Azuma [5].

2.2.1 Device based classification

The categories proposed in [15] are based on the way how the output device is connected with the user. If the user wears the device on his head we talk about the head-mounted devices. The systems designed to be carried in hand belong to the handheld category and systems fixed within the space and not with the user are included in the spatial group.

Head-mounted devices

The head-mounted category consists of five main types of devices: Optical see through HMD, Video see through HMD, HMProjectors, HMProjective display and retinal displays. For more information about HMDs see [21].

Optical see through head-mounted display In Azuma's survey [5] the author states that: "Optical see-through HMDs work by placing optical 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. The combiners are also partially reflective, so that the user sees virtual images bounced off the combiners from head-mounted monitors. The optical combiners usually reduce the amount of light that the user sees from the real world. Since the combiners act like half-silvered mirrors, they only let in some of the light from the real world, so that they can reflect some

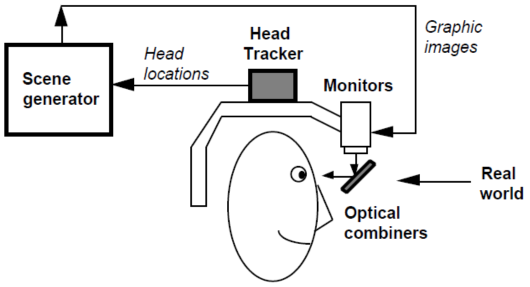


Figure 2.3: A scheme of optical see through head-mounted display. Figure is taken from [5].

of the light from the monitors into the user's eyes." For the scheme of the device see figure 2.3.

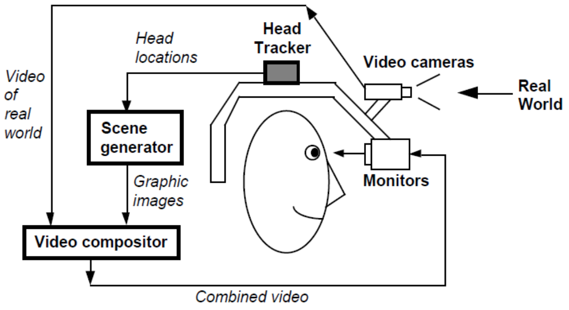


Figure 2.4: A scheme of video see through head-mounted display. Figure is taken from [5].

Video see through head-mounted display This type of HMD was defined in [5] as: "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 from these cameras is 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." For the scheme of the device see figure 2.4.

Head-mounted projectors beam the generated images onto the ceiling and use two half-silvered mirrors to integrate the projected stereo image in front of the user.

Head-mounted projective display redirects the image created by miniature projectors with mirror beam combiners so the images are beamed onto retro-reflective surfaces in front of the users eyes.

Retinal display uses low-power semiconductor lasers to project modulated light directly onto the retina of human eye. Main disadvantage of this technique is that it provide only nonstereoscopic monochromatic image [15].

Handheld devices

Handheld devices are nowadays the most popular platforms for the augmented reality applications. These devices usually incorporate all the necessary sensors, computer and display (or projector) in one portable gadget. Commonly known handheld devices are smartphones, tablets, palmtops or notebooks. Although most of the published papers in the area of mobile augmented reality focus on these particular devices, there were also some efforts to build special handheld devices, for example iLamps [98]. In iLamps Raskar et al. presented object augmentation with a handheld projector utilizing a new technique for adaptive projection on non-planar surfaces using conformal texture mapping.

Spatial devices

The spatial category encloses different solutions designed to be fixed within the environment (not to be worn in the hand or on the head). An example of the spatial solutions are: the PC station with the webcamera, the CAVE (cave automatic virtual environment) [28], Projection mappings [120, 88], Virtual showcase [14].

The Fish tank is the title of the system consisting of the computer station equipped with the webcamera and the monitor which are usually used for browsing of the augmented reality at home. The CAVE is an immersive virtual reality/scientific visualization system, which lies between virtual and augmented reality. The CAVE is a room-sized cube where three to six of the walls are used as projections screens.

The Virtual Showcase developed by Bimber et al. [14] presents the projection-based multiviewer augmented reality display device which consists of half silvered mirrors and the graphical display (for the overall look of the device see figure 1.12). In this device the user can see real objects inside the showcase (through the half-silvered mirrors) merged with the virtual objects or layers displayed on the projection screen under the showcase. This technique makes use of the concept of the Pepper's ghost developed in the 1862 [20] (described in the section 1.2).

2.2.2 Perception of the reality based classification

In our classification we start from Azuma's work [5] and we divide the augmented reality systems based on the way how they create the augmented experience. The further category encloses the applications which create augmented reality by adding the virtual information (3D models, images, text) to the record of reality. Later category encloses systems which create augmented reality by displaying/projecting the virtual information directly in front of our site of reality. A table 2.1 relates the device based classification and perception of the reality based classification.

The record of the reality mixed with virtual information (added to record)

All kinds of the video-see through approaches belong to this category. The video see through device basically consists of the camera which records the reality and the display (or a projector with a projection screen) which provides the user with the reality mixed with the virtual information (the augmented experience). This category enclose video see through head-mounted display, most of the existing handheld devices (smartphones, tablets, palmtops, netbooks) and the Fish tank solutions.

The reality mixed with virtual information (added to reality)

This category includes all the applications in which the virtual information is projected directly on the real world objects, or onto the optical see through device. The typical representatives of these approaches are the projection mapping applications, for example the projection on the astronomical tower clock Orloj situated in the center of Prague [120]. Other systems which belong into this category are optical see through head-mounted displays, retinal displays, head-mounted projectors, head-mounted projective display, CAVE [28], Virtual showcase [14], and also some handheld solutions (for example iLamps [98], as described in the section 2.2.1).

Table 2.1: Table relates the device based classification and perception of the reality based classification.

	added to record	added to reality
head-mounted	video see-through HMD (Museum wearable [113])	optical see-through HMD (Sutherland's HMD [117])
handheld	mobile/tablet AR (e.g. museum guides [19, 85, 72, 8, 37])	optical see-through handheld displays handheld projections (iLamps [98])
spatial	fish tank mirror projections [59].	Pepper's ghost [2] projection mappings [120] holographic displays [14]

Chapter 3

Registration in augmented reality

In the previous chapter we have committed to strictly follow Azuma's definition of augmented reality 1.1. One of the three basic rules states that the virtual objects should be registered in 3D with the real environment. There are several strategies to achieve this kind of registration and they will be described in the following chapter.

3.1 Visual registration

The most common tool for registration of the real and virtual world for the purpose of augmented reality is a visual sensor (camera). The area of visual registration belongs to the intersection of the augmented reality and computer vision. As the registration is one of the key problems of computer vision there is ongoing research since the creation of the field. In the following section we will present the most influential registration methods for the field of augmented reality.

3.1.1 Visual Markers

The first system which used visual black and white markers to identify the rotation and translation of the camera was developed by Rekimoto [101] (and lately added to the ARoolkit [61]). Rekimoto in his paper proposes a registration method based on detecting and identifying square black and white markers (quad-tangles) in the camera frames (for the scheme of the process see figure 3.1). The transformation

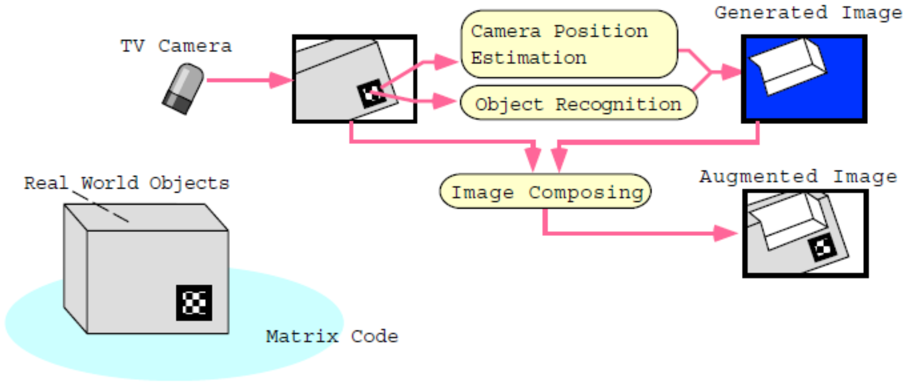


Figure 3.1: An overview of the registration process developed by Rekimoto. Figure is taken from [101].

parameters are then calculated based on the positions of the 4 corners of this quadtangle. Let $(x_i; y_i; 0)$ be a point on the plane of the square marker and $(X_i; Y_i)$ be a corresponding point on the image plane of the camera. These two points are related as follows

$$\begin{aligned} X_i &= \frac{a_1 x_i + a_2 y_i + a_3}{a_7 x_i + a_8 y_i + 1} \\ Y_i &= \frac{a_4 x_i + a_5 y_i + a_6}{a_7 x_i + a_8 y_i + 1}, \end{aligned} \tag{3.1}$$

where a_1 to a_8 represent the intrinsic (focal length, skew, coordinates of the principal point) and extrinsic (translation and rotation) camera parameters. If we have four pairs of $(x_i; y_i; 0)$ and $(X_i; Y_i)$, we can determine these parameters $(a_1, .., a_8)$ by solving

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} x_1 & y_1 & 1 & 0 & 0 & 0 & -X_1x_1 & -X_1y_1 \\ x_2 & y_2 & 1 & 0 & 0 & 0 & -X_2x_2 & -X_2y_2 \\ x_3 & y_3 & 1 & 0 & 0 & 0 & -X_3x_3 & -X_3y_3 \\ x_4 & y_4 & 1 & 0 & 0 & 0 & -X_4x_4 & -X_4y_4 \\ 0 & 0 & 0 & x_1 & y_1 & 1 & -Y_1x_1 & -Y_1y_1 \\ 0 & 0 & 0 & x_2 & y_2 & 1 & -Y_2x_2 & -Y_2y_2 \\ 0 & 0 & 0 & x_3 & y_3 & 1 & -Y_3x_3 & -Y_3y_3 \\ 0 & 0 & 0 & x_4 & y_4 & 1 & -Y_4x_4 & -Y_4y_4 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \\ a_7 \\ a_8 \end{pmatrix}. \tag{3.2}$$

The parameters $(a_1, \dots, a_8)$ store the effects of rotation, translation and the perspective transformation of the camera. Based on them we can map the distorted code image on the camera plane to the normalized matrix code space. In the next step we recognize the image in the middle of the code. The coordinate system is than created with the origin in the center of the code and axes x, y parallel with the sides of the code and z axis parallel to the normal vector of the plane of the code. Then we know which 3D object to augment (based on the ID of the code) and where and we can create the augmented video.

Since ARToolkit many different opensource and commercial marker-based AR system have emerged (Studierstube [73], Wuforia by Qualcomm [97], Flartoolkit [4]).

A different approach was proposed in [113] where the authors used infrared emitters an infrared camera for registration. The main advantage of markers constructed of infrared diodes is, that they are easily recognizable on the image from IR camera and do not disturb the users. The main disadvantage is, that in case of added to record applications another (not infra-red) camera is necessary to display the augmented reality.

Marker fields

The marker fields are special types of markers which are suitable for applications where the wide area needs to be covered. The presence of occluders prevents from using many unique markers. The marker field is composed of mutually overlapping checker-board like markers, aperiodic 4-orientable binary square-window arrays (De Bruijn tori). This type of markers was proposed in [118]. An example of the marker field can be seen in figure 3.2

3.1.2 Object detection and registration — markerless AR

Since 2000 there is a strong interest in the markerless augmented reality. To achieve visual markerless tracking, we can recognize and track the real world objects as if they were markers or we can estimate the 3D structure of the scene and augment virtual objects in it using SLAM, PTAM or utilize the depth sensors like Kinect.

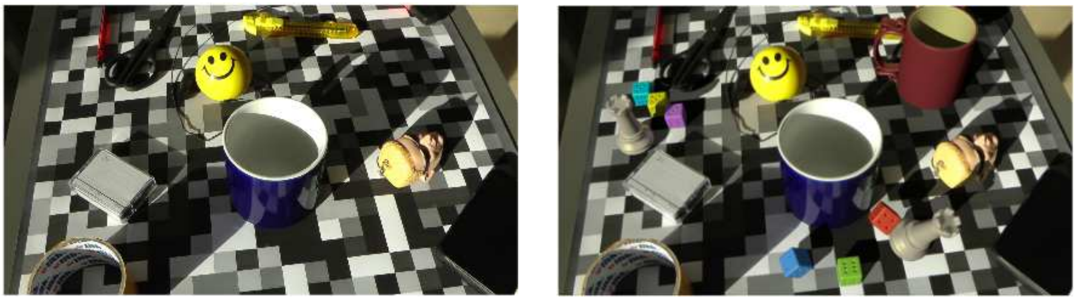


Figure 3.2: Left: input image with marker field. Right: recognized camera location and augmented scene. Figures were taken from [56].

In this section we will describe the first approach. Based on Zitova et al. [134] the registration has following 4 steps:

1. Feature detection
2. Feature matching
3. Transform model estimation
4. Image resampling and transformation

In the feature detection stage local features are mostly used. We provide a short overview of different methods. Local features extract information from the parts of the image, which are interesting, i.e. the intensity varies in their neighborhood. To extract local features, firstly the interesting points are detected, then the features are computed for all detected points and finally feature vectors (descriptors) are created. There are many methods how to detect interesting points. Two most basic approaches are to choose points uniformly or randomly from the whole image. However this will not ensure that the selected points are interesting. Another method is to detect blobs instead of points using for example MSER detector [82].

The methods which detect interesting points are called interest points detectors and three of them are used the most: The Harris corner detector [55] computes the eigenvalues of the second moment matrix of an image at some point. Harris method was boosted in [109] where the authors proposed taking the minimum of the eigenvalues and compare it to a given threshold. If it is bigger, the point is considered a corner.

The second method uses the approximation of Laplacian of Gaussian with the difference of Gaussians (DoG) and looks for the local extrema in the scale-space pyramid. Scale-space pyramid consists of consecutive image scales, so called octaves (scales of the image are $1, 1/4, 1/16$ etc.), with each octave containing the image progressively smoothed with a Gaussian kernel. This methods is used in the well-known SIFT and SURF detectors [9], [79].

The third method is based on the accelerated segment test (AST). This approach examines the neighborhood of every point of the size of the Bressenham's circle with diameter $d = 7$. The points are concerned as interesting if there is a set of $n = 12$ continuous pixels in the neighborhood that fulfill the following criterium. The intensity difference between the examined pixel and the neighborhood pixel must be larger than a given threshold. We can find this method in the FAST detector [105]. There are modifications of the method with different values of d and n.

As for the description methods, we can identify two types of most popular descriptors: integer and binary. The main advantage of binary descriptors is that two binary strings can be compared using the Hamming distance instead of the Euclidean distance. Hamming distance can be computed very fast and it saves the matching time. If $p = (p_1, p_2, \ldots, p_n)$ and $q = (q_1, q_2, \ldots, q_n)$ are two binary strings we can define their Hamming distance as follows

$$d_h(p, q) = \sum_{i=1}^{n} \frac{\delta(p_i, q_i)}{n}, \tag{3.3}$$

where

$$\delta(x, y) = \begin{cases} 0 & \text{if } \ (x = 1 \wedge y = 1) \vee (x = 0 \wedge y = 0) \\ 1 & \text{otherwise.} \end{cases} \tag{3.4}$$

Integer description methods typically compute the histogram of gradients (HoG) in the patches placed around the interesting point (for example the SIFT, SURF or DAISY descriptors [124]). On the other hand, binary methods use the binary intensity tests which compare the line endings in the mikado like patch (for example

BRIEF [22] or ORB [106] descriptors) or in the human visual system inspired patches (BRISK [76] or FREAK [3]).

Another important issue of the local feature detectors and descriptors is their invariance. The ideal local feature will be invariant to affine and projective transformations, however in real-world features this is not the case. The most important invariance is to scale, rotation and translation. The invariance to translation is achieved through the way how the local features are extracted. Since the descriptor is computed in a small neighborhood of the point, the actual position of the point in the image is irrelevant. The scale invariance is in [79, 9] achieved in the detector phase, where the interesting points are retrieved in the scale-space pyramid. The scale of the feature is then estimated as its level within the pyramid. The rotation invariance is in [79, 9, 106] achieved by rotating the neighborhood of the interesting point in the direction of the highest gradient in the neighborhood.

As we can see in Zitova et al. [134] the first step in the registration process is the matching of the features. Brute force matching can be very time consuming and the matching time grows with the number of objects in the database. Therefore there exist several strategies for speeding up of this process.

When the features are matched we can determine the best match from the database for every feature in the image (input frame). However the matched feature can be false positive so another filtering of the matches is usually necessary. The basic strategy is to use some previously trained/estimated threshold and keep only matches with distance smaller than the threshold. The second strategy proposed in [79] called second nearest neighbor is to compare the distance d of the closest match to the distance d_2 of the 2^{nd} closest match, if $d < 0.3d_2$ the match is taken as correct. The third approach is to take P' as the closest match of feature point P only if P is also the closest match of the P'. The more global approach to feature matching and filtering is to use the Bag of visual words (or Bags of visual features) approach [29]. The main idea behind the bag of visual features is to cluster similar features (using for example K-means clustering) in the feature space and represent them with the centroid of the cluster, so called Visual word. Then the object is represented by the histogram of the visual words present in the object. During the matching phase we can only compare the histograms of the objects from the database with that of the image. The extracted features are often further examined using their geometry consistency. To achieve this, the RANSAC approach is used to compute the homography. Assuming a pinhole camera model any two images of the same planar surface are related by homography. We have 2 cameras A and B, looking at points P_i on the plane π. Let ${}^{A}\mathrm{p}_i$ and ${}^{B}\mathrm{p}_i$ be the projections of the point P_i in the images of cameras A and B (see figure 3.3), then

$$ {}^{A}\mathrm{p}_i = K_a \cdot H_{ba} \cdot K_b^{-1} \cdot {}^{B}\mathrm{p}_i, \tag{3.5}$$

where K_a and K_b are the intrinsic parameters of A and B and H_{ba} is the homography matrix

$$ H_{ba} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix}, \tag{3.6}$$

which can be expressed as

$$ H_{ba} = R - \frac{tn^T}{d}, \tag{3.7}$$

where R is the relative rotation of B with respect to A, t is the translation vector, n is the normal vector of the plane and d is the distance from the plane. We need at least 4 point correspondences from A and B to compute the homography (or 3 in case of affine homography).

Homography is computed between the feature points from the database and the feature points which seem to belong to the same object in the image. We can then estimate the outliers of the most stable homography.

Most of these methods are effective in case, when there is only one object present in the image, or several objects of different types, or when segmentation was performed in the previous step. However when we have multiple instances of one object present in the image, we have to utilize different approaches. We propose a new approach of multiple instances detection in this thesis. It is explained and evaluated in chapter 4. By computing the homography we achieved the transform model estimation and can proceed to image resampling and transformation if needed. In case of registration for the purpose of augmented reality we only need to estimate the center of the object,

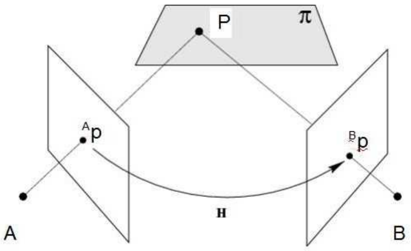


Figure 3.3: We have 2 cameras A and B, looking at points P_i on the plane π. Let $^A\mathrm{p}_i$ and $^B\mathrm{p}_i$ be the projections of the point P_i in the images of cameras A and B. Then the images of the plane π in A and B are connected by homography H.

the normal vector in that point and two orthogonal vectors in the plane of the object to set up a coordinate system.

3.1.3 SLAM and PTAM

If we think little out of the box, we can formulate the problem of the estimation of the user position in the indoor environment as the problem of the robot's position estimation in the space (with or without a map stored in the memory). This kind of problem is solved in robotics with different strategies, depending on the robot's sensors. Usually the SLAM (simultaneous localization and mapping) approach is used. The augmented reality researchers make use of this approach and in their paper [64] Klein and Murray propose PTAM (parallel tracking and mapping). PTAM uses a camera and does not need any additional markers, or sensors to track the relative user's (camera) position and to build a room-sized maps of the unknown environment. The system is based on systematic keypoints detection and tracking in the camera images and creating a map (model) of the environment based on the triangulation between the corresponding keypoints position in different keyframes. The authors develop the PTAM system for both PC [64] and the smartphone (IPhone3) [65].

For our purpose the IPhone PTAM is relevant since it can be used for the tracking of the mobile user in the indoor environment. The implementation of the PTAM on IPhone was challenging because of two main limitations, the lack of processing power and the camera parameters (low frame rate, rolling shutter and the narrow field-of-view). The authors provide several modifications in the keypoints detection, tracking and system initialization to overcome the limitations and to create the system capable of generating and augmenting small maps in real time and full frame rate. The keypoints detector developed for the PC based on the FAST corners detecton and tracking with 4 image pyramid levels was replaced with the Shi-Thomasi corners with 5 pyramid levels. Also, the culling of the keypoints and the culling of the redundant keyframes was done.

3.1.4 RGBD and 3D

There is a growing research in the domain of registration in RGBD (RGB image + depth) or 3D (point cloud representation). We will mention some of the most interesting works published in this domain. The registration of the RGBD image in the model of the scene using regression forests was proposed by Jamie Shotton et al. in [110]. The first important research in object matching and registration in the point cloud representations is [58]. The authors proposed so called spin images. For every point in the point cloud they estimate the reference frame (the coordinate system with origin in the point and the cylindrical coordinates) and create 2D accumulator called the spin image. Different approach to matching of the point clouds was proposed in [32]. For every point pair in the point cloud the descriptor of five features is estimated and hashed in the table. The SHOT (signatures of histograms) descriptor was published by Tombari et al. [125]. For every point, its spherical neighborhood is divided into 32 bins and local surface signatures are accumulated in the histogram for every bin. On the other hand there is also a research on registration of meshes, two key papers are [83] and [107]. In [83] the authors proposed a method where the point cloud is decimated and triangulated to form a mesh. The descriptor is then computed as follows. For every vertex pair, the coordinate system is created with the origin in the center of the line joining the vertices and the directional vectors

$\mathbf{v}_x, \mathbf{v}_y, \mathbf{v}_z$ of the three axes x, y, z are defined as follows: $\mathbf{v}_x$ is given by the cross product of normals, the $\mathbf{v}_z$ is the average of normals and $\mathbf{v}_y$ is the cross product of $\mathbf{v}_x$ and $\mathbf{v}_z$. The 3D accumulator (tensor) with dimensions 10x10x10 is then created at the origin. An element of each grid bin of the tensor is equal to the surface area of the mesh intersecting the grid bin.

In [107] the authors proposed so called FPFH (fast point feature histograms). For every point (vertex) P of the mesh, for all points p_i, p_j in the spherical neighborhood of P we define a Darboux frame uvw as follows

$$\begin{aligned} u &= n_i \\ v &= (p_j - p_i)xu \\ w &= uxv, \end{aligned} \tag{3.8}$$

where n_i is a normal vector at point p_i. The descriptor α, ϕ, θ is then computed as follows

$$\begin{aligned} \alpha &= v \cdot n_j \\ \phi &= \frac{(u \cdot (p_j - p_i))}{\|p_j - p_i\|} \\ \theta &= \arctan\left(\frac{w \cdot n_j}{u \cdot n_j}\right), \end{aligned} \tag{3.9}$$

where n_j is a normal vector at point p_j.

3.2 Outside-In Inside-Out systems

In [15] the authors define two types of tracking systems: outside-in and inside-out tracking, where the first type refers to "the system that applies fixed sensors within the environment that track emitters on a moving targets". The inside-out system is composed of the sensors (capable of determining their relative position to the emitters) attached to the moving target and the fixed emitters. Both types of systems can be implemented in two basic ways. One sensor gets information from the closest emitter and estimates the distance from the sensor or the information about the distance is acquired from several sensors and the position of the user in the scene is estimated using triangulation. In the area of augmented reality this kind of systems is commonly implemented using WiFi [102], RFID [85] or bluetooth [19] systems. We have to mention GPS, which is very popular representative of this registration category (especially for pseudo-AR applications).

3.2.1 GPS

Presently the GPS is fully operational and meets the criteria established in the '60s for an optimal positioning system. The system provides accurate, continuous, worldwide, three dimensional position and velocity of the user with the appropriate receiving equipment. GPS also disseminates a form of Coordinate Universal Time (UTC). The satellite array nominally consists of 24 satellites arranged in 6 orbital planes with 4 satellites per plane [60].

In the outdoor AR applications the information from the GPS (global positioning system) is a great help in the estimation of location of the user anywhere in the world. If we want to provide a user with the information about the monument he is looking at, we may first want to know the continent, state, the city and the street where he is located. This information will then help to narrow the search in the database of the monuments.

When talking about GPS we have in mind that this information is not sufficient for the correct augmentation of the object. For this, we need not only the rough position of the user, but his precise position and orientation (7D) relative to the object. However many pseudo-AR applications utilize only the information gathered from the GPS receiver sometimes fused with the information from devices like magnetometers or compass. The problem is that the presence of metal objects and electronic devices usually causes incorrect output of these devices.

3.3 Dead reckoning indoor positioning system

In the paper [66] authors proposed the indoor user dead reckoning tracking system composed of an accelerometer, gyroscope, magnetometer, camera and a head-tracker. The dead reckoning system is based on calculating current position using a previously determined position. The system is not dependent on any external markers, chips or sensors and determines the user's relative position from the variation of the vertical and horizontal acceleration caused by human walking locomotion. To estimated absolute position authors used additional method of matching the camera stream with the database of images utilizing the Kalman filter framework [59]. "The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error."

To achieve information about the acceleration vector and the angular velocity vector the authors attached sensors (accelerometer, gyroscope and magnetometer) to the user's torso. The approach is based on the clinical studies of human movement which claim that the pattern of the movement and the forces applied to the user's center of gravity (torso) are almost unaffected by the individual characteristics and so they did not introduce any individual walking learning mechanisms. The three basic types of the locomotion were introduced (walking on the flat floor, going up and down stairs and taking an elevator) and the data from the sensors were analysed to detect and measure the unit cycle of walking locomotion and direction and then to identify the one of three locomotion types. The demonstration of the relation between different stages of the unit cycle of walking on the flat floor and the change in the horizontal and vertical acceleration can be found in figure 3.4. The first step in the process of estimating the acceleration vector is the determination of the direction of the gravity and the forward direction, which has to be calibrated after the sensors are attached to the user. Then the decomposition of the acceleration vector into each component (vertical and horizontal acceleration) is done. In the next phase, the relation between acceleration vectors and angular velocity is analyzed and the locomotion type is recognized.

In 2006, the system proposed in [66] was further extended with the RFID reader, the GPS and the embedded computer [67].

In the area of pseudo-AR applications, there exist many applications which make use only of the dead-reckoning user position estimation based on the sensors embedded in almost all new mobile devices (smartphones, tablets). The biggest problem

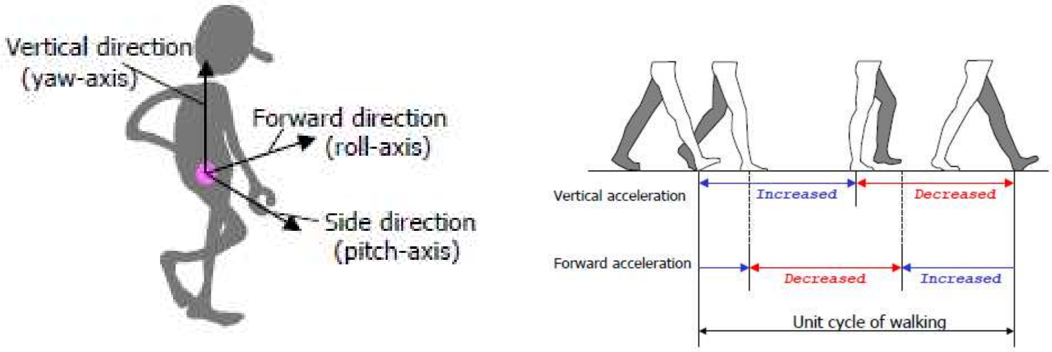


Figure 3.4: Left: Definition of each axis. Right: Definition of the unit walking cycle and relationship between the change in acceleration and the cycle stage. Figures were taken from [66].

of these applications is the weak accuracy of their sensors (accelerometers) and the accumulating error. The accumulation of the error of position is caused by the nature of the process, where the actual position is estimated from the previous one.

3.4 Others

3.4.1 Mechanical tracking

The first example of the use of tracking for augmented reality was the so called "sword of Damokles" developed by Ivan Sutherland for his first HMD.

3.4.2 Audio

Out of the box solution was proposed by the researchers at the McCormick School of Engineering and Applied Science. They have created the IPhone application called Batphone [104] which allows the user to record ambient noise in a room and tag it with an acoustic fingerprint. This allows to determine future approximate location of the user by the matching the actual ambient noise with the database of acoustic fingerprints. However this method can estimate only very approximate position not accurate for the purpose of the augmented reality application.

Chapter 4

Multiple instances detection

Since the beginning of the new century the growing popularity of marker-less augmented reality applications inspired the research in the area of object instance detection, registration and tracking. The usage of common daily objects or specially developed fliers or magazines (e.g. IKEA[1]) as AR markers became more popular than traditional ARtoolkit like black and white patterns. Although there are many different methods for object instance detection emerging every year, very little attention is paid to the case where multiple instances of the same object are present in the scene and need to be augmented (e.g. a table full of fliers, several exemplars of historical coins in the museum, etc.). In this chapter we review existing methods of multiple instance detection and propose a new method for RGB images and RGBD images overcoming the limitations of previous methods. For a scheme of the proposed process see figure 4.1.

We propose a new method for multiple instance detection of objects in cluttered scenes using local features and Hough-based voting. For the purpose of correct object detection and registration in augmented reality it is necessary to correctly register objects even when several instances of the augmented object are present in the image. When dealing with the visual (ARToolkit like) markers or when we add non-visual markers, the correspondence is not an issue. On the other hand in the case of marker-

[1] https://www.youtube.com/watch?v=vDNzTasuYEw

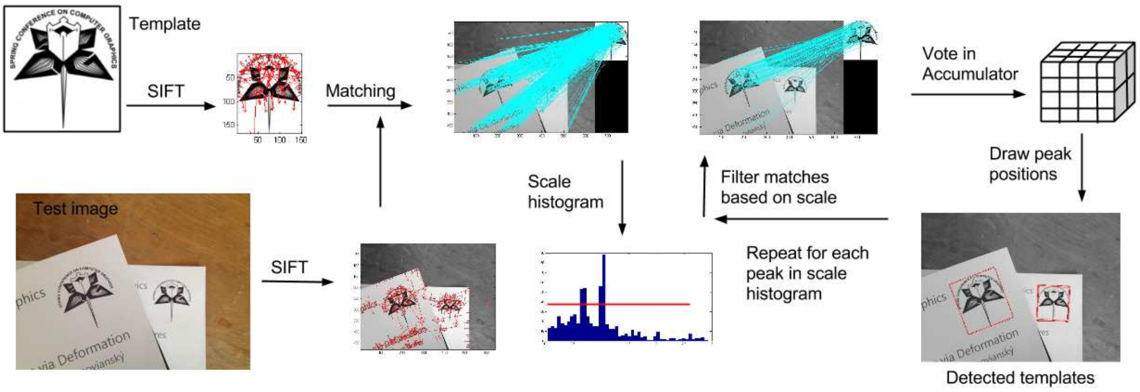


Figure 4.1: Overview of the registration process. In the first step SIFT features are computed for both template and test images. Features are matched and the best match is estimated for every interesting point in the test image. Then the histogram of the ratios of the scales of the interesting points and their matches is created. For every peak in the histogram corresponding interesting points pair votes in the 3D accumulator for the (x, y)-position of the object center and the object's rotation. Then the peaks of the accumulator correspond to the objects on the test image.

less object detection utilizing local features, it is not easy to determine how many instances of one object are present in the image.

In the area of multiple object instances detection we can identify 2 basic strategies. One approach is to first segment the image and then recognize and register all the segmented objects. The segmentation can be done in the image domain or in case of the RGBD data in the depth data. This method however deals with the not yet fully solved problem of the segmentation of objects in cluttered scenes. The problem of segmentation in the depth domain can occur when e.g. the objects lie side by side on the same plane. The second method estimates the number and the position of the object instances based on the clustering of the detected points, the iterative RANSAC and/or Hough based voting [6]. The basic clustering of features approach is inefficient in case of more complex scenes with very close or overlapping objects. Although, the greedy method which iteratively finds the best-matching instance, remove the corresponding features and find another instance using RANSAC approach to compute homography is extensively used, it proved to be inefficient for robust multiple instance object detection in [95]. We focus on more approaches in the following sections.

4.1 Important previous approaches

The most significant work in this area was done in [94, 93, 95, 27] and [132].

The authors of [94, 93, 95] focus on the segmentation of multiple instances of a low-textured object on a conveyor belt. In the first step they extract the SIFT features and divide the image into P regions and then use second-nearest neighbor method for matching of the features. On every region they search for instances of one object with marked control points (the vertices of the approximation of the object contour by a polygon). For every matched feature they find the positions of the control points in the 2D space. Then they use the mean-shift algorithm to cluster the control points positions and estimate the final positions as the center of the clusters. The authors utilize the color similarity measure to distinguish overlaying segmented objects with the overlap larger than 30% percent. The main limitation of this method is the fact that all instances have to be of one object of the same scale without perspective deformations.

In [132] the authors proposed a method for detection and localization of multiple objects and multiple instances of objects using PCA-SIFT [63] and agglomerative clustering of the features. In the training phase they acquire a video sequence containing the objects to be recognized and annotate and segment them manually. Then they use PCA-SIFT (PCA is used to estimate 20 main components in the 128-dimensional SIFT [79] space) to find the keypoints and store their descriptors and relative location towards the annotated object's center. In the recognition phase, they compute the PCA-SIFT keypoints in the image and match them using linear nearest neighbor search with estimated threshold for maximal matching distance. Then every keypoint correctly matched with the object votes for the object center, based on the rotation and scale of the keypoint estimated during SIFT detection. Afterwards these votes are clustered with agglomerative clustering and small clusters are discarded. The main limitation of this work is that the scale of the objects must be known and all objects are supposed to be of same size.

The authors of [27] use a sparse object model created from different views using SIFT features in the preprocessing step and a bundle adjustment. For every feature in the database its position on the sparse model is stored. In the recognition phase the process iterates over every object in database:

1. Extract the SIFT features and match them with the database.

2. Cluster the SIFT features locations using the mean-shift algorithm.

3. For each cluster choose a subset of points and estimate a hypothesis about the pose of the object according to these points. If the number of consistent points is bigger than a threshold, create a new object instance and refine its pose using all consistent points. Repeat until not enough points left or the number of iterations reached.

4. Merge all instances from different clusters with similar pose.

5. Estimate the position and orientation of the camera given a set of 2D $\Leftrightarrow$ 3D correspondences using orthogonal Procrustes decomposition. The scale is corrected using the ratio of standard deviation of all pairwise distances within the test and train images and the rotation is done aligning the principal components of test and train data.

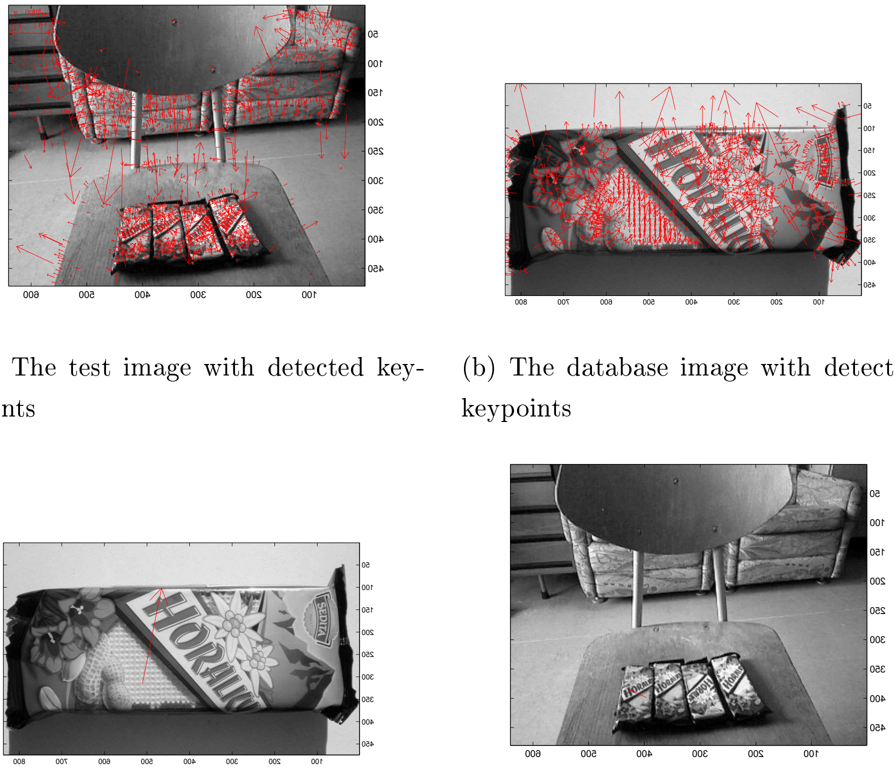
The main limitation of this work is that the preprocessing phase is time consuming and there is a necessity to photograph the object from different positions.

Based on these previous works and their limitations we decided to construct 2 methods for multiple objects detection. One utilizes only the 2D image, and the second one also the information from the RGBD sensor.

4.2 2D approach

4.2.1 Objects of the same scale

As a first step we have decided to develop the method for detection of the multiple instances of objects of the same scale in grayscale images. All objects are placed approximately perpendicularly to the camera axis and no off-plane rotations are allowed.



(a) The test image with detected keypoints

(b) The database image with detected keypoints

(c) The test image with marked keypoint

(d) The database image with marked coresponding keypoint

Figure 4.2: Local features matching.

We utilize the SIFT features, because they posses, together with the position of the feature, the information about the scale ς and rotation θ. The scale information is derived from the detector's scale-space pyramid as the octave in which the keypoint is detected. The rotation of the feature is estimated as dominant orientation of the gradient in the neighborhood of the interesting point. The size of the neighborhood is determined by the previously estimated scale.

In the training phase of our method we build the database of the objects to be recognized. For now we assume all objects are planar and rectangular. We mark (or compute) the center points of the objects in the images. We then compute the SIFT features (interesting points, IP) and store their descriptors, scale and rotation in the database.

The recognition is performed as follows. We extract the SIFT features from a test image and store their descriptors, rotation, scale and position. For every extracted SIFT feature in the test image we compute the closest features in all objects from the database based on the Euclidean distance of the descriptors.

There exist several strategies to filter the SIFT matches.

Threshold The matches are filtered based on some threshold which has to be estimated in the training phase (e.g. using the K-fold cross validation).

Second nearest neighbor The distance of the closest match d is compared to the distance of the 2nd closest match d_2. If $d < 0.3d_2$, the match is considered as correct.

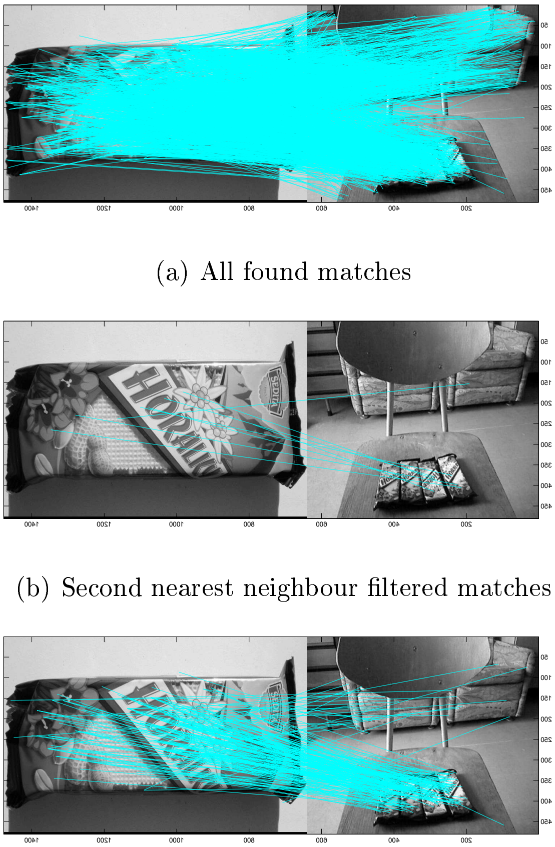
Double check The match of the descriptor A of an IP from the image I_1 and B of an IP from the image I_2 is correct if they are mutually the closest. That means B is the closest to A from all the descriptors of IPs from I_2 and A is the closest to B from all the descriptors of IPs from I_1.

We propose a new criterion called *the scale ratio* – r. To filter the matches we firstly compute the ratios of the scale of test image features ς and the scale of their corresponding database image (template) features ς' as follows

$$r = \frac{\varsigma}{\varsigma'}, \tag{4.1}$$

Then we create the histogram of these scale ratios.

We choose the highest peak of the histogram as the correct scale ratio. Then we preserve only the features having this correct scale ratio. The comparison of all found



(a) All found matches

(b) Second nearest neighbour filtered matches

(c) Scale ratio criterion filtered matches

Figure 4.3: Example of matches satisfying different filtering criteria.

matches, second nearest neighbor filtered matches and scale ratio criterion filtered matches can be seen in figure 4.3.

We have compared the previously mentioned criterion for the purpose of multiple instance detection utilizing our voting scheme. The corresponding precision/recall curves for second nearest neighbor, double check and scale ratio methods can be seen on figure 4.4. These methods were evaluated on 30 images from a database Test 1. Our new scale ratio criterion proved to work with 100% precision and 100% recall.

The next step is to create a 3D accumulator, where we store the center points of the objects in the test images. The 3 dimensions of the accumulator are the x and y coordinates of the image scaled to 1⁄10 and the rotation α of the object sampled to 60

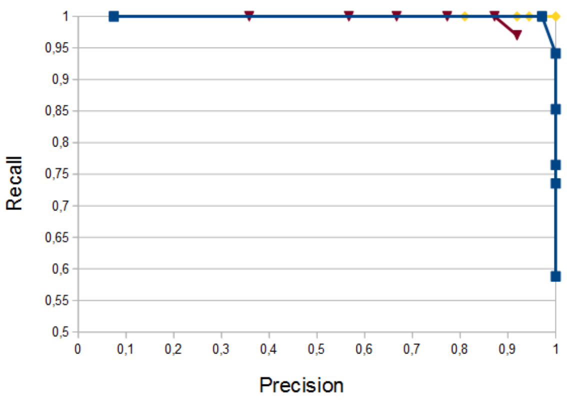


Figure 4.4: Precision/recall curves for different match filtering criteria (blue – double check, yellow – scale ratio, red – second nearest neighbor).

bins. The coordinates of the center point are estimated from the SIFT orientations of the matched IPs A and A' as follows

$$S = A + r \cdot \mathbf{M}_{rot}(\alpha) \cdot \mathbf{v}, \tag{4.2}$$

where r is the scale ratio,

$$\mathbf{v} = A' - S' \tag{4.3}$$

is the vector from the IP A' to the center point S' of the template. The matrix

$$\mathbf{M}_{rot}(\alpha) = \begin{bmatrix} \cos\alpha & \sin\alpha \\ -\sin\alpha & \cos\alpha \end{bmatrix} \tag{4.4}$$

is the rotation matrix and α is the rotation of the object with

$$\alpha = (\theta - \theta') \mod 2\pi, \tag{4.5}$$

where θ and θ' are the SIFT orientations of the matched IPs A and A'.

The peaks of the accumulator represent the position of centers and the rotation of the objects from the database found in the test image. To determine the detections in the accumulator we use the threshold p on the height of the peak. We have tested different values of p which we discuss in section 4.4.1. The image and the corresponding accumulator (converted to 2D — x, y and the highest value from different orientation) can be seen in figure 4.6

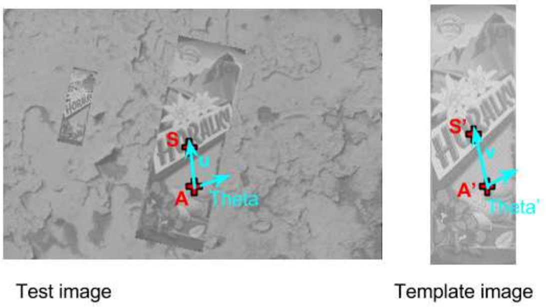
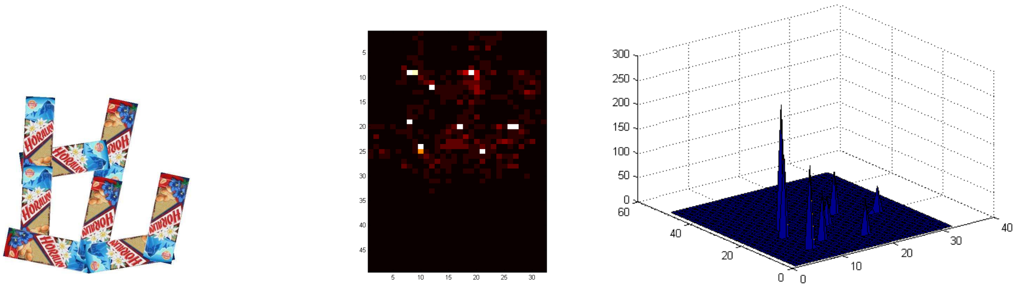


Figure 4.5: The illustration of the vectors, points and orientations on template and test image.

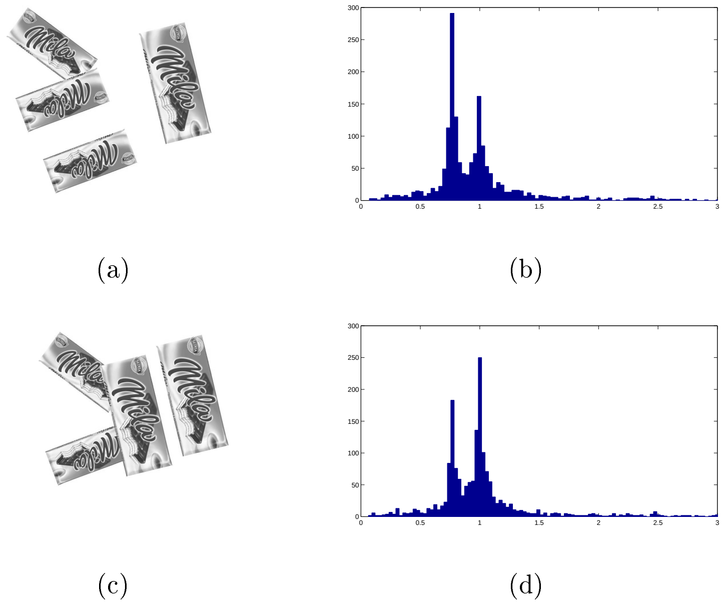


(a) (b) (c)

Figure 4.6: Image (a) and corresponding 2D version of accumulator (only the highest value for all orientations is taken) displayed as a heat map (b) and a surface (c).

4.2.2 Objects of different scales

If we want to generalize the previously proposed approach to images with objects of different scales, we have to extend the voting approach proposed in the previous section. We not only look for the highest peak in the histogram of the scale ratios, but we identify all the peaks. A peak in the histogram has both neighboring bin values lower and its height is bigger than the threshold $t_h = 15$. An example of images and their corresponding histograms can be seen in figure 4.7.



(a) (b)

(c) (d)

Figure 4.7: (a), (c) test images, (b), (d) corresponding scale ratio histograms.

4.2.3 Perspective distortions

In the previous sections we described the process of detecting multiple instances of objects placed perpendicular to the camera axis based on the voting in the accumulator. The main limitation of this approach lies in the fact that we cannot identify the out of a plain rotation of the objects present in the image. To achieve this we can compute the homography transformation between the matched points from the test image and the database objects. We need at least 3 corresponding pairs to compute the affine homography.

When we extend our previously defined accumulator with a list of points that vote in the corresponding bin, we can compute the homography from these points and their corresponding pairs in the database image. The process of computing the homography transformation is described in section 3.1.2.

We can compute the homography using two approaches. We can use all point pairs assuming that all matches were correct or we can utilize the RANSAC approach [36]. Then the homography can be used to compute the correct positions of the object corners (4 in the case of rectangle). To check the correctness of the computed corner positions we compare the area of the detected object to the area of the template object scaled according to scale ratio. If the difference is within the threshold we draw the final contour of the object.

Although the RANSAC approach is extensively used in the registration of object instances in the image, it often fails to detect more than one instance even if it is used in iterative manner (see [95]). In our work we utilize it only for estimation of the transformation of pairs belonging to one instance. This is ensured by the voting in the accumulator.

4.3 RGBD approach

Since an emerge of the Kinect sensor in 2010, RGBD sensors became affordable and very popular in computer vision tasks. We decided to utilize the depth information produced by this sensor for our multiple instance detection approach. The main advantage of this approach is, that we do not deal with different scales of objects because the "scale" in 2D image is defined based on the distance of the object from sensor (we assume that we have constructed our database with a similar sensor) and the objects are stored together with their size. The first steps are similar to the previous approach, we extract the interesting points and compute their descriptors using SIFT approach and match them with the database. We can then filter the matches based on their scale ratio, similarly to previous approach or proceed without filtering.

Another filtering can be done by checking the consistency of the scale ratio and the ratio of the depth of the point from test image and the template image.

Then we will iterate over the point pairs and vote in the 3D accumulator. We decided to use the 3D accumulator (2D subsampled image space + rotation) instead of 4D accumulator (3D subsampled space + rotation). In 3D accumulator we will vote for the projections of the objects centers (S) on to the image space. To determine the correct vote for each point pair, we compute the vector $\mathbf{v}$ from the point A to the center S' in the template space, similar to section 4.2.1. Then we estimate the normal vector $\mathbf{n}_p$ at the point A from the depth data as follows. First, the points in the sphere neighborhood of the A with diameter $d = 10$ are extracted. In the next step the Principal component analysis (PCA) [91] of the extracted points is computed. Next, the eigenvectors (principal components) are sorted by the corresponding eigenvalues and the cross product of first two principal components is computed. The cross product is then the normal vector of the plane of the object in the 3D space.

In the next step we want to estimate the position of the center of object in 3D S'' (on the object plane P defined by the computed normal vector and point A). The center of the object S'' will be defined as

$$S'' = A + r \cdot \mathbf{K}_{rot}(\beta, \mathbf{s}) \cdot \mathbf{M}_{rot}(\alpha) \cdot \mathbf{v}, \tag{4.6}$$

where

$$\mathbf{K}_{rot}(\beta, \mathbf{s}) =$$
$$= \begin{bmatrix} \cos\beta + s_x^2\,(1-\cos\beta) & s_x s_y\,(1-\cos\beta) - s_z\sin\beta & s_x s_z\,(1-\cos\beta) + s_y\sin\beta \\ s_y s_x\,(1-\cos\beta) + s_z\sin\beta & \cos\beta + s_y^2\,(1-\cos\beta) & s_y s_z\,(1-\cos\beta) - s_x\sin\beta \\ s_z s_x\,(1-\cos\beta) - s_y\sin\beta & s_z s_y\,(1-\cos\beta) + s_x\sin\beta & \cos\beta + s_z^2\,(1-\cos\beta) \end{bmatrix}$$

is the rotation matrix, the angle

$$\beta = \arccos\left(\frac{\mathbf{n}_p \cdot \mathbf{n}_l}{\|\mathbf{n_p}\| \cdot \|\mathbf{n}_l\|}\right) \tag{4.7}$$

is the angle between $\mathbf{n}_p$ and normal vector of the image plane $\mathbf{n}_l = [0, 0, 1]$ and

$$\mathbf{s} = \mathbf{n}_p \times \mathbf{n}_l \tag{4.8}$$

is the direction vector of the line defined as the intersection of the object and image plane.

Then the new center S is the projection of the center S'' to the image space.The scheme of described situation can be seen on figure 4.8.

We have to have in mind that our 3D accumulator do not posses the information about the z-coordinate of the S' point nor the normal vector of the plane. To properly register the detected object on the image we need to have at least the information about the normal vector. We decided to store the average normal vectors of the points which voted in the corresponding accumulator bin.

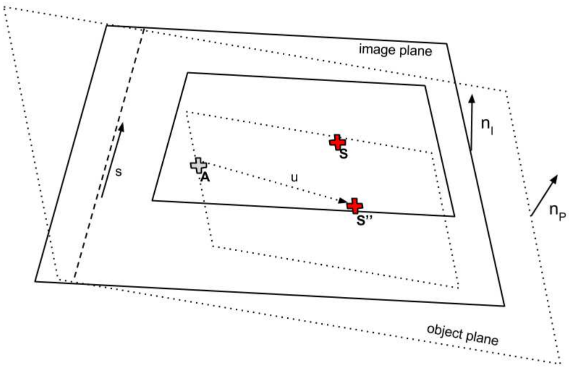


Figure 4.8: The scheme of the object center estimation in RGBD case.

4.4 Results

We decided to validate our method in 3 steps. In the first step we test our 2D approach on artificially created images containing one type of object in up to 8 instances of varying scales. In the second step we test the 2D approach using real world RGB images (converted to grayscale) acquired by camera from the RGBD sensor from Primesense and in the third step the RGBD approach has been tested on real world RGBD data (RGB image + depth values).

4.4.1 Test 1 – Artificial images

For this test we created 30 artificial images containing instances of one object. To ensure that our method is independent of the template object the test was repeated on 30 artificial images of another object. Templates of the used objects can be seen in figure 4.9. Each image contains at most 8 instances of the object of different scales and in-plane rotations. The images were created by placing scaled and rotated versions of the template object onto an uniform and cluttered backgrounds. Some objects are occluded by each other and some are only partially visible (at least ⅓ of the object is visible). The examples of the images can be seen in figure 4.10.

To evaluate our method we have manually annotated the objects on all created images. We marked the 4 corners of the objects and saved their (ground truth) coordinates. In the evaluation process we have tested the distance of the detected

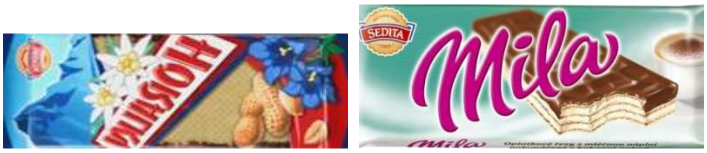
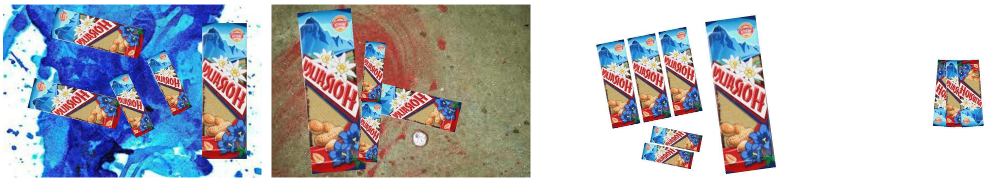


Figure 4.9: Templates of the objects.



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

Figure 4.10: Examples of images in Test 1.

and true positions of each corner of the object. The detected object is considered a correct detection (CD) if there exists a ground truth object whose corners are within a given threshold from the detected corner positions

$$CD = \begin{cases} 1 & \exists G \; \forall i \in \{1, \dots, 4\} \colon d(G_i, D_i) < T_{pix} \\ 0 & \text{otherwise,} \end{cases} \tag{4.9}$$

where G_i is the i-th corner of the ground truth object, D_i is the i-th corner of the detected object and $T_{pix} = 5$ is the threshold.

The method proposed in section 4.2.1 was tested for different values of threshold p. The precision/recall curve displaying the relation of the precision and recall values for different values of p can be seen in figure 4.11. The examples of the processed images with correct detections can be seen in figure 4.12.

4.4.2 Test 2 – Real world images

The second test was carried out on 30 real world images captured by Primesense RD1.09 sensor in a cluttered environment. RGB images were used for the second test. To acquire images from the device we used the OpenNI 2 SDK[2]. Images contain

[2] http://www.openni.org

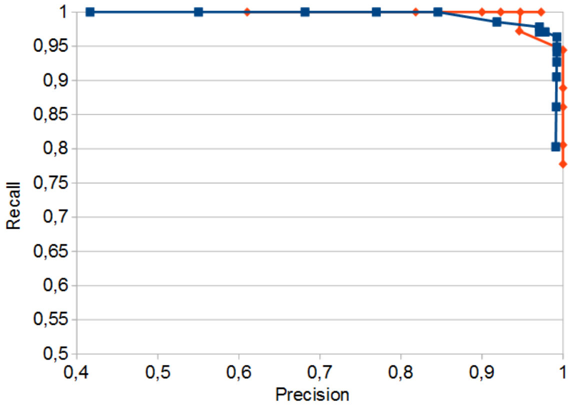


Figure 4.11: Precision/recall curves in Test 1. Blue curve represent data for template 1 and red for template 2.

p	6	8	10	12	14	16	18	20	22
Precision	0,416	0,550	0,682	0,770	0,846	0,918	0,971	0,971	0,978
Recall	1	1	1	1	1	0,985	0,978	0,971	0,971
p	24	26	28	30	32	34	36	38	
Precision	0,992	0,992	0,992	0,992	0,922	0,992	0,992	0,992	
Recall	0,964	0,964	0,949	0,942	0,927	0,905	0,861	0,803	

Table 4.1: The precision values for different threshold levels in Test 1.

up to 5 instances of the object and were taken as a keyframes of the video sequence acquired in the cluttered office environment. The examples of the images can be seen in figure 4.13.

The method was evaluated in the same manner as the first test. The four corners of the object instances in the test images were marked manually and the threshold T_{pix} was set to 15 pixels.

The proposed method was tested for different values of threshold p and the values of precision and recall were computed for each value of p. The precision/recall curve (blue) displaying the relation of the precision and recall values for different values of p can be seen in figure 4.14.

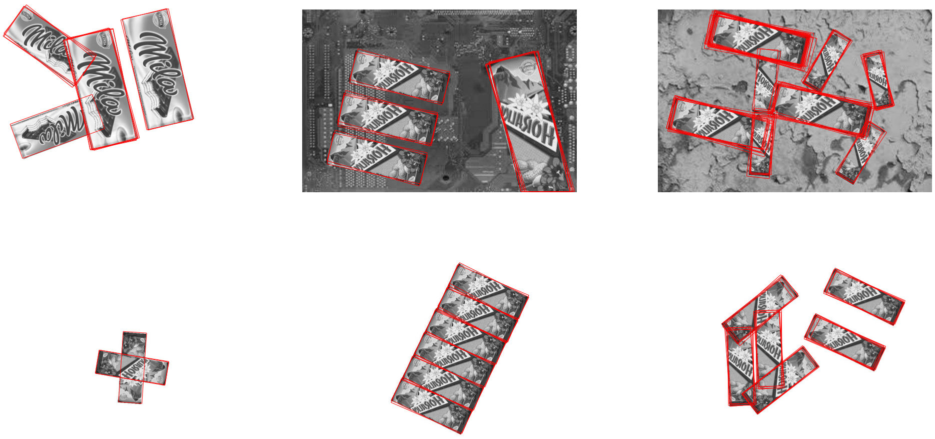


Figure 4.12: Examples of correct detections in Test 1.

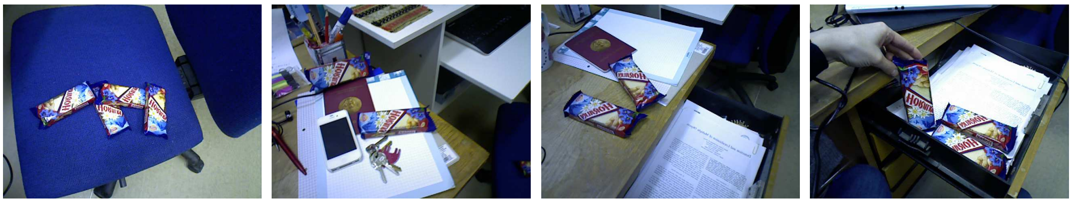


Figure 4.13: Examples of images in Test 2.

The examples of the processed images with correct detections can be seen on figure 4.15.

4.4.3 Test 3 – RGBD samples

For evaluation of the proposed RGBD method 30 images from test 2 dataset with the corresponding depth information (16-bit grayscale images) were used. Examples of images with corresponding depth data can be seen on figure 4.16.

The method was tested with the ground true data created for Test 2 for different values of threshold p and the values of precision and recall were computed for each value of p. The precision/recall curve (red) displaying the relation of the precision and recall values for different values of p can be seen in figure 4.14.

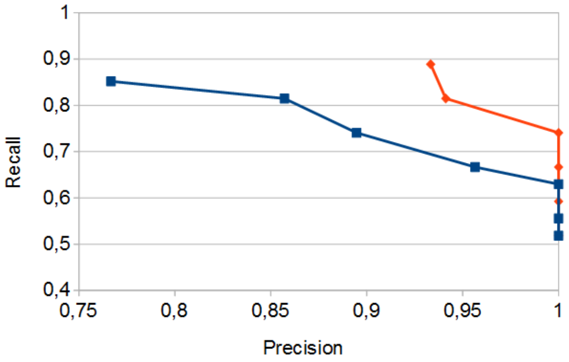


Figure 4.14: Precision/recall curve for Test 2 (blue) and Test 3 (red).

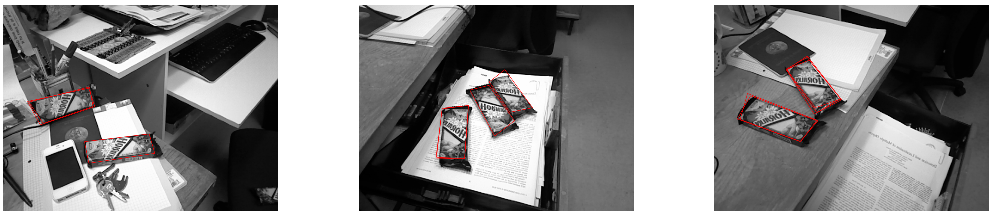


Figure 4.15: Examples of correct detections in Test 2.

4.4.4 Incorrect detections

We recognize four types of detections in the validation process: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). In this section we discuss the incorrect detections which were classified as FP or FN. The FN detection occurs when the object instance is not detected on the image. The FN value grows with the growing threshold. The FP detection is the false alarm detection. It grows with the lowering threshold.

There are several reasons for incorrect detections. The FN are caused by the low number (under threshold) of correct matches between the points in the test and template image. The low number of matches naturally occurs when the object is occluded or only partially present. The FP are caused by incorrect matches between the points in the test and template images. They occur if the neighborhoods of two

p	4	5	6	7	8	9	10
Precision T2	0,7667	0,8571	0,8947	0,9565	1	1	1
Recall T2	0,8519	0,8148	0,7407	0,6667	0,6296	0,5556	0,5185
Precision T3	0,9333	0,9412	1	1	1	1	1
Recall T3	0,8889	0,8148	0,7407	0,7407	0,6667	0,5926	0,5185

Table 4.2: Precision for different values of threshold in Test 2 and Test 3.

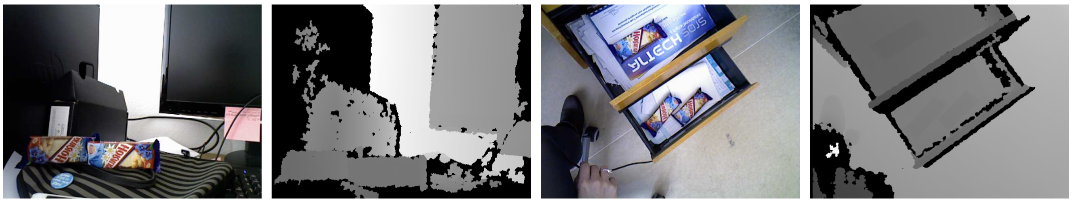


Figure 4.16: Examples of images (RGB and corresponding depth) in Test 3.

not corresponding points are too similar and can be reduced using filtering (e.g. scale ratio method).

These problems are also connected to the usage of local feature methods to pair the points, and their problems with blur, big off-plane rotations, specular reflections and shadows.

In our approach we allow multiple detections of one object instance. You can see multiple detections marked by rectangles on images 4.15, 4.17 and 4.12. We decided not to filter these detections to check if all of them are correct. However,

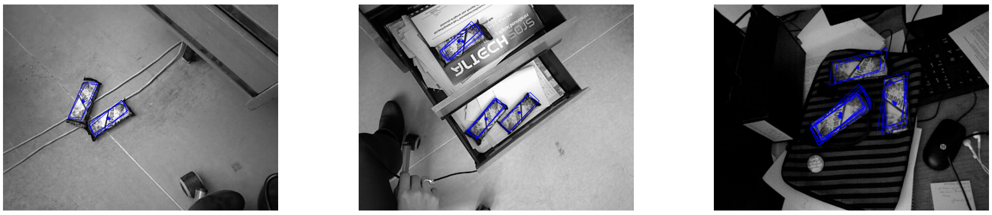


Figure 4.17: Examples of correct detections in Test 3.

they can be easily avoided in three ways. The accumulator can be made sparser, which would not be a problem as far as we compute the correct corner positions utilizing RANSAC style homography transformation between the template and the points which voted for the corresponding accumulator bin. Instead of the homography computation we can estimate the more precise center coordinates and the rotation angle by computing the median of all points (positions, angles) that voted in the corresponding bin. Another possibility is to perform a (agglomerative) clustering on the accumulator or on the marked corners in the image. The straight approach is to look for the local maxima in the accumulator instead of taking all bins satisfying the threshold.

4.5 Conclusions

We have proposed a new approach for multiple instance detection in images of cluttered scenes. We have decided to overcome the limitations of previous state of the art methods — the time consuming preprocessing phase and the fact that all objects has to be of the same known scale without perspective distortions.

In our methods we use the SIFT local features, but any scale and rotationally invariant methods that can extract the scale and the rotation of the features (like SURF [9]) can be utilized. SIFT extracts the scale of the feature in the detection phase where a scale-space pyramid is used. We can use the same scale-space approach to extract the scale from detectors like e.g. FAST [105]. To extract the rotation SIFT determines the dominant orientation of the gradient in the neighborhood of the interesting point when creating the feature descriptor. Similar approach was used to change the BRIEF [22] detector to rotational invariant detector ORB [106]. SIFT was chosen as it provided the highest precision in the tests in our previous work. To speed up the process it is possible to utilize the GPU version of the SIFT [112].

Our method was tested on rectangular template objects however the generalization to polygonal objects can be easily performed, since the features vote for the central point of the object and its orientation in the accumulator. Generalization for non-planar 3D objects will be possible with the extension of the template database with photographs of the objects from different sides.

Our approach was tested on artificial and real-world images. Our method was evaluated on the images with instances of just one template object. The extension to instances of multiple template objects in one image is in iterative manner processing every object from the template database.

Based on the validation we can state that our 2D method works with 98% precision (97% recall) on artificial images and 85% precision (81% recall) on real images. The proposed 3D method works with 93% precision (89% recall) on real world RGBD data.

Chapter 5

Classification and registration of paintings

The following chapter presents our approach to the efficient classification and registration of fine art paintings using local features and our approach to feature matching utilizing global features. Since the '80s the computer graphics and vision community is focusing on the cultural heritage preservation issues. This big mission includes the restoration and the classification of fine art paintings. In this area the most significant assignments are the digital restoration of the paintings, classification of the author's style and categorization of paintings based on the style [57], distinguishing paintings from real scene photographs [30] and determination of new features for paintings classification (e.g. description of paintings' textures analyzing brush strokes [108]). For the relatively complete overview see [78].

On the other hand the strong trend in augmented reality and museum guides induced the research on recognition (and registration) of museum artifacts. We can divide current augmented and visual museum guide solutions based on the used method of exponate recognition into following groups: Visually based systems, Outside-in inside-out systems, Dead-reckoning systems, Combination of systems and the User input based systems. In this thesis, we deal with the visually based methods only. For further information on museum guides see [44].

Visually based methods utilize images from the camera to recognize exponates and estimate their exact 3D position. The exponates can be recognized in different ways. In the first case, binary markers (ARToolkit tags) printed and placed (registered) near the exponate are used [129]. The second approach is based on matching of the local features in the camera frame with the local features of the database of photographs of exponates computed in advance [8, 42]. The third approach consists of recognition of the exponates using global features (for example color histograms, histograms of gradients, [37, 39]). As the representative of this approach we can mention [37], where the authors use global features and neural networks for the recognition of museum exponates (both 2D and 3D). The fourth type of methods uses the bags of the visual words approach and their combination with the local (global) features for the recognition of paintings [54].

Another important aspect of the object recognition using local features is the matching of the feature vectors. In the matching phase, the feature vectors extracted from the unknown object are matched with the database of the feature vectors extracted from the labeled objects. The unknown object is labeled with the same label as the object with the most matches. This phase can be time consuming when performing all-to-all brute-force matching. Different methods for organizing of the database of features for faster search and match have been published. They are based on, for example, kd-trees (in the later implementation of SIFT), random trees [75], spectral hashing [127] or bag of visual words [26, 29].

The main advantage of our approach compared to the previously mentioned methods is, that our approach is not dependent of the construction of the database in the preprocessing, in contrary, the paintings can be added to the database on the fly, with only computing the global feature and inserting it to 1D vector.

The method presented in this chapter uses a global feature value to organize the database. We present the results obtained using different global features and different local feature descriptors. We first describe the dataset used, then the algorithm consisting of segmentation, feature extraction and matching. The result of our experiments are detailed in section 5.4.1. The overview of the method can be seen on figure 5.1.

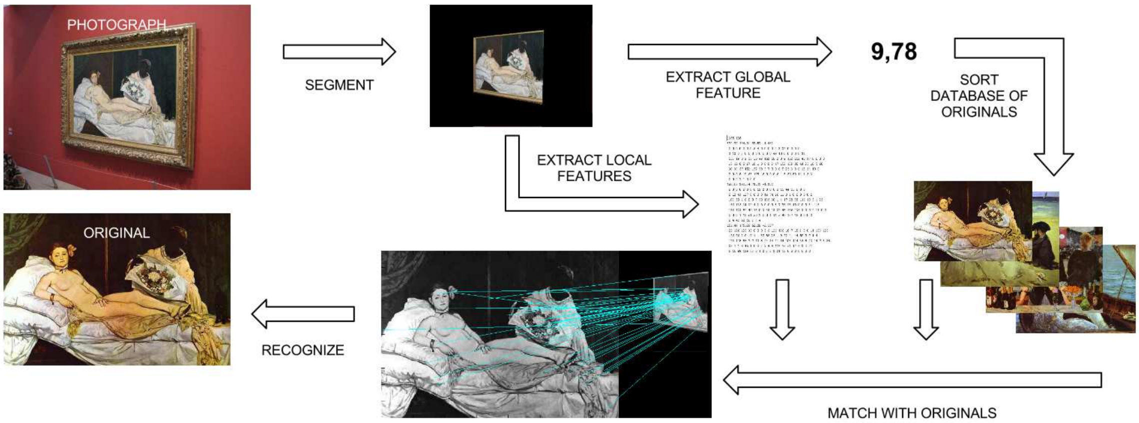


Figure 5.1: Overview of the recognition process. In the first step painting is segmented from the photograph using the method proposed in the Segmentation section. Then the global feature and local features are extracted from the segmented painting. In the next step the database of the Originals is sorted based on the global feature value. Local features extracted from the painting are then matched with the sorted database and the painting is recognized as the first painting from the database of Originals with matches exceeding the threshold.

5.1 Dataset

Dataset used in our work consists of two parts — the Photographs and the database of Originals. The Photographs dataset contains 500 photographs of the paintings created by 5 painters: Leonardo Da Vinci, Rembrandt Van Rijn, Vincent Van Gogh, Eduard Manet and Gustav Klimt. These photographs were taken in galleries by various unspecified digital cameras. Photographs from the collection of the authors of this paper, from the initiative on their website and from the Flickr web portal[1] are used. Photographs have different resolutions, miscellaneous scales and were taken under varying lighting. An example of photographs from Photographs dataset can be seen in figure 5.2.

[1] http://www.flickr.com

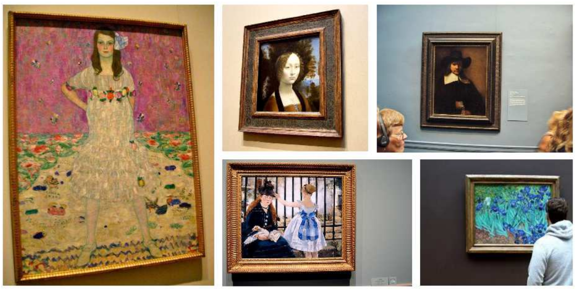


Figure 5.2: Sample images from the Photographs dataset.

The database of Originals consists of 59 (10–15 from each painter) ground truth paintings taken from Olga's web gallery[2]. Paintings from the database of Originals corresponding to the photographs in figure 5.2 can be found in figure 5.3.

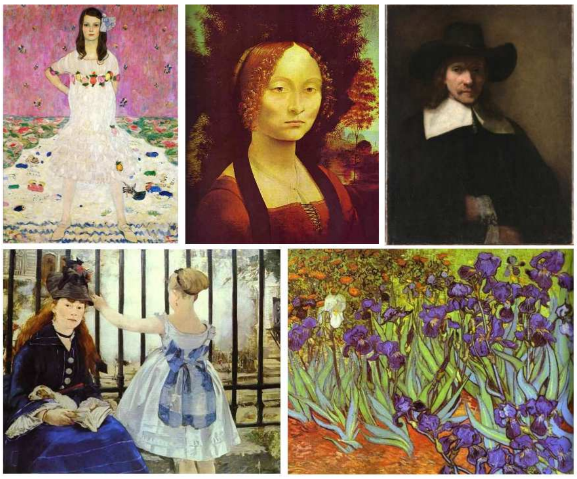


Figure 5.3: Sample images from the database of Originals.

Figure 5.4 displays the number of photographs of paintings by different painters, with the corresponding painting present in the database of Originals (blue) or not (red).

[2] http://www.abcgallery.com/index.html

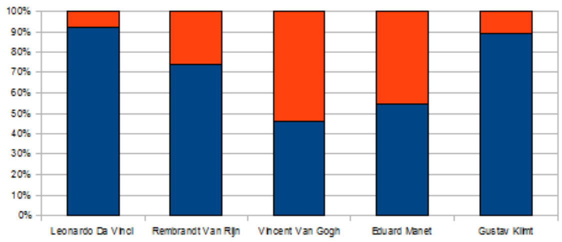


Figure 5.4: Distribution of the photographs in the Photographs dataset. Blue means that the photographs have the corresponding painting in the database of Originals.

5.2 Algorithm

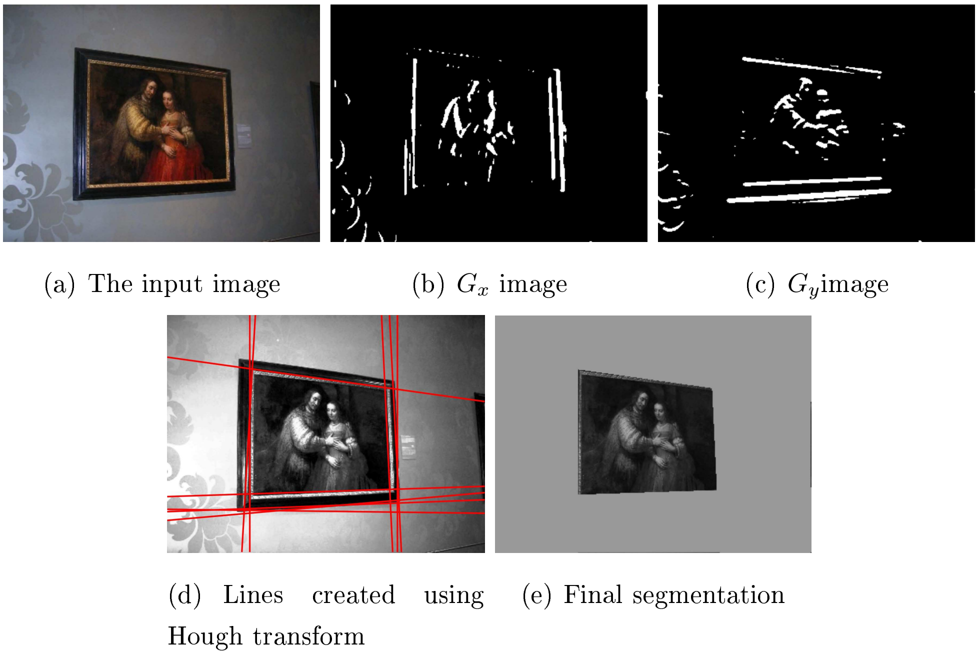
The work flow of our method is graphically depicted in figure 5.1. In the first step the painting is segmented from the photograph using the method described in the following section. Then the global feature and local features are extracted from the segmented painting. In the next step the database of the Originals is sorted based on the global feature value. Local features extracted from the painting are then matched with the sorted database and the painting is recognized as the first painting from the database of Originals with number of matches exceeding a given threshold. The details of individual steps can be found in following sections.

5.2.1 Segmentation

The goal of the segmentation phase is the segmentation of the painting and its frame in the input image (from the Photographs dataset). Three different techniques are used. The basic one uses the Gauss gradient method, in the improved method the anisotropic diffusion [92] is applied and the final method is based on the watershed transformation [11].

Gauss gradient method

In the method the image is processed using Gauss gradient function which computes the gradient using first order derivative of the Gaussian. It outputs the gradient images G_x and G_y of the input image using convolution with a 2D Gaussian kernel.


(a) The input image

(b) G_x image

(c) $G_y$image

(d) Lines created using Hough transform

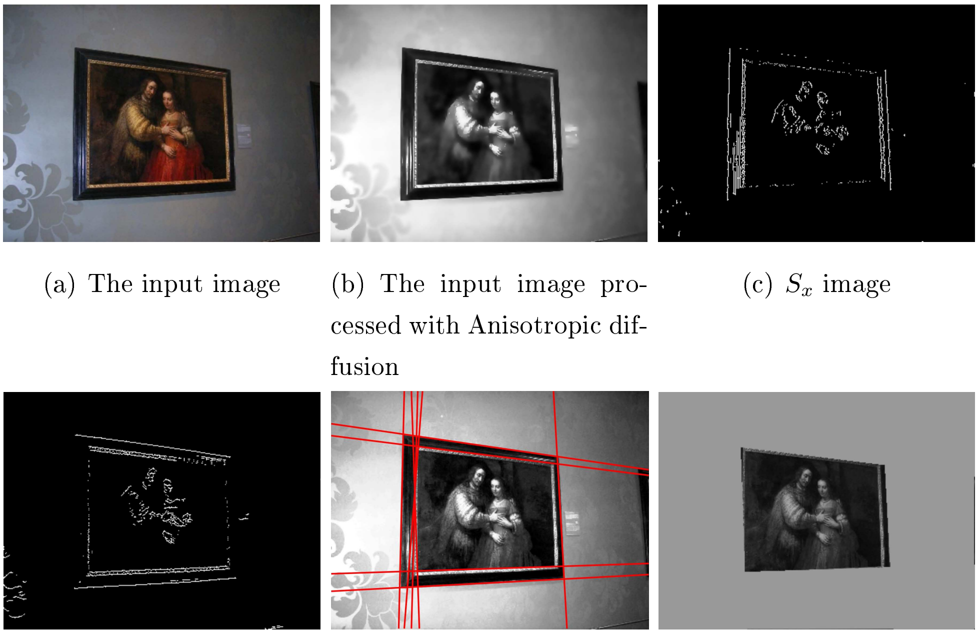
(e) Final segmentation

Figure 5.5: The process of the Gauss gradient method of the segmentation.

In the next phase the G_x and G_y gradient images are sent as an input to the Hough transform. The Matlab implementation of the Hough transform is used, since it enables to count the lines from the Hough peaks directly and to connect or trim them based on their length. Lines created in the previous step are then expanded to the borders of the image and lines with big slope are filtered out. Lines are then divided into four groups, one for upper, lower, left and right edges. Consecutively, the painting is segmented as the smallest quadrilateral created from the lines. Gauss gradient method is depicted in figure 5.5.

Anisotropic diffusion method

In this approach firstly the histogram equalization is done. Then the image is processed using anisotropic diffusion, the technique which smooths the image, but preserves the edges. The function is used with the following parameters: number of iterations = 10, $\kappa = 30$, $\lambda = 0.25$ and option = 1 (κ controls conduction as a function of gradient, λ controls speed of diffusion, it is 0.25 for maximal stability, option = 1



(a) The input image

(b) The input image processed with Anisotropic diffusion

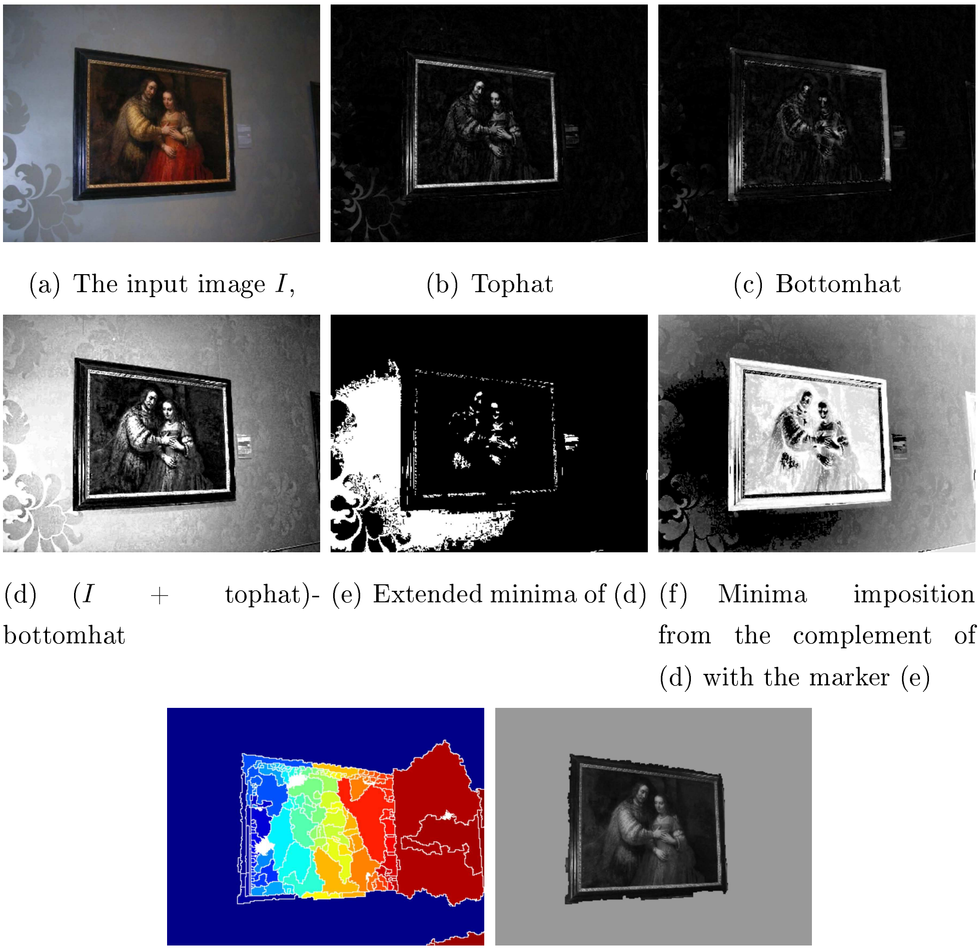
(c) S_x image

(d) S_y image

(e) Lines created using Hough transform

(f) Final segmentation

Figure 5.6: The process of the anisotropic diffusion method of the segmentation.



(a) The input image I,

(b) Tophat

(c) Bottomhat

(d) (I + tophat)-bottomhat

(e) Extended minima of (d)

(f) Minima imposition from the complement of (d) with the marker (e)

(g) Clusters created with watershed transform

(h) Final segmentation

Figure 5.7: The process of the watershed method of the segmentation.

means the diffusion equation, which favors high contrast edges over low contrast ones). The output image of the function is then convolved with the horizontal and vertical Sobel edge filters, resulting in two binary images S_x, S_y. The images S_x, S_y are processed equally to G_x and G_y images in the first method, and the input image is also segmented in the same way. Anisotropic diffusion method is presented in figure 5.6.

Watershed method

In the third approach the input image is firstly preprocessed to enhance edges of the painting's frame. Afterwards the watershed transform is applied. The preprocessing phase consists of 4 steps:

1. Create tophat I_t and bottomhat I_b of the input image.
2. Create image $I_2 = (I + I_t) - I_b$.
3. Create I_3 as extended minima (regional minima of the H-minima transform) of I_2.
4. I_4 is created as the minima imposition from the complement of I_2 $(1 - I_2)$ with the marker I_3. In the next step, clusters are created with watershed transform applied on the I_4 image. In the last phase the final segmentation is made by growing the background from the corners in the clustered image.

Watershed method is presented in figure 5.7.

Results

In the segmentation phase the methods were tested on the smaller dataset which consisted of 100 Rembrandt paintings. This dataset consisted of the photographs taken by tourists in different galleries, under different lighting condition and with different cameras. During the segmentation phase most problems were caused by the low contrast of the photographs, which was eliminated in the anisotropic diffusion method by the equalization of the histogram. Table 5.2.1 summarizes the percentage of the correctly segmented versus oversegmented and undersegmented paintings.

Method	Gauss gradient	Anisotr. Diffusion	Watershed
Correct segmentation	73%	89%	49%
Oversegmentation	6%	3%	1%
Undersegmentation	21%	8%	50%

Table 5.1: Percentage of paintings properly segmented by different methods.

Oversegmentation, mostly in the Gauss gradient method was induced by the strong edge responds in the paintings, especially in the painting Night Watch (Rijksmuseum, Amsterdam) where the pale flags and spears have very strong color edges in the black background. Other problem with the Night Watch was the low contrast of the black frame of the painting to the dark gray wall paint. In the watershed method, oversegmentation occurred in one image, where the shadow in the upper right side of the image blends with the black upper right corner of the painting. Undersegmentation arised mostly in the following cases: the paintings frame is mostly covered, the paintings frame is in low contrast with the wall, the background of the painting contains strong edges (wall corner or cartouch presented on the photograph).

The problems with oversegmentation and undersegmentation were partly eliminated by using the anisotropic diffusion in the second method, which smoothed the color edges in the painting and also the edges in the background, but preserves the edges of the frame. The basic method, the Gauss gradient uses smoothing with the Gaussian kernel, which smoothed all edges uniformly. The watershed method presents a different approach to the segmentation, but the results indicated that it is not efficient for this purpose. Finally, as expected the best results were achieved with the anisotropic diffusion method (see table 5.2.1) and we have decided to use this method for the segmentation of the paintings.

5.2.2 Global features

A feature that is computed statistically over all pixels of the image is defined as global feature. Global features help to detect similar images in global view. Usually the low-level global features describe color, intensity or texture of an image. We decided to test several chosen global features in order to find the best discriminative feature for our method. Since the photographs of paintings usually have low quality, we did not work with textural descriptions, only with color and intensity features. In [77] 40 features for describing painting were presented, only 12 of them were global.

The global features we have compared were

- A. average intensity
- B. percentage of light pixels
- C. normalized intensity histogram
- D. entropy, E. normalized hue histogram
- F. number of pixels that belong to the most frequent hue (f4)
- G. most populated hue, H. hue contrast (f5)
- I. hue count (f3)

The labels in parentheses correspond with the labeling of features in [77]. The difference between features was computed as a distance for A, B, D, F, G, H, I and as a Kullback − Leibler distance (KLD) [71] in case of histogram features C, E.

Intensity features

The grayscale image is computed as

$$I = 0.2989R + 0.5870G + 0.1140B. \tag{5.1}$$

A. The average intensity is computed as

$$A = \frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h} I_{i,j}, \tag{5.2}$$

where w and h are the width and height of the image respectively.

B. The percentage of light pixels is the number of pixels with the intensity above the average intensity A, divided by the total number of pixels

$$B = \frac{1}{wh} \sum_{i=1}^{w} \sum_{j=1}^{h} \begin{cases} 1 & \text{if } (I_{i,j} > A) \\ 0 & \text{otherwise.} \end{cases} \tag{5.3}$$

C. Normalized intensity histogram is computed as follows

$$C(i) = \frac{N(i)}{wh}, \tag{5.4}$$

where $N(i)$ is the number of pixels with intensity i.
D. The entropy of the image is calculated as

$$D = -\sum (p \log_2(p)), \tag{5.5}$$

where p is the image intensity histogram with 256 bins.

Color features

For these features the image is transformed to CIE L*a*b* color space and hue is calculated as the four-quadrant arcus tangens of $^{b}/_{a}$.
E. The normalized hue histogram. The hue histogram has 90 bins and is calculated as

$$E(i) = \frac{N_H(i)}{wh}, \tag{5.6}$$

where $N_H(i)$ is the number of pixels with the hue i.
F. The percentage of pixels that belong to the most frequent hue is computed as

$$F = \frac{N_H(p)}{wh}, \tag{5.7}$$

where

$$p = \arg \max_{i=\{0,\dots,90\}} N_H(i). \tag{5.8}$$

G. The most populated hue G is

$$G = p. \tag{5.9}$$

H. Hue contrast is computed as the arc-length distance of two most populated hues as

$$H = (S_1 - S_2) \mod 45, \tag{5.10}$$

where

$$\begin{aligned} S_1 &= \arg \max_{i=\{0,\dots,90\}} N_H(i) \\ S_2 &= \arg \max_{i=\{0,\dots,S_1-5,S_1+5,\dots,90\}} N_H(i). \end{aligned} \tag{5.11}$$

I. Hue count feature is the number of local maxima in the hue histogram above a preset threshold t and is calculated as

$$I = \sum_{N_H(i)>t} i. \tag{5.12}$$

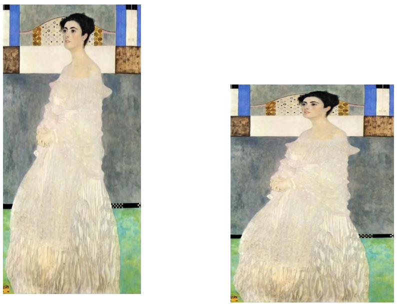
5.2.3 Normalization

We use 2 types of features, one describing the intensity and one describing the color. In the case of the features computed in the images of different sizes, shapes and acquired under different lighting conditions with different cameras we always have to normalize the features to achieve comparable results.

One way of dealing with the different scales is to normalize using the division of the feature values with the number of pixels in the image as was used in the feature computation.

Another way of achieving the scale and shape normalization is to scale all the paintings to the same resolution, for example (VGA) resolution 800 × 600. We resample all images in the databse of Originals in preprocessing phase, and we also resample all paintings segmented from the images in the recognition phase. We can also loose the texture information in kind of high frequency textures with frequency close to one pixel.

If we want to normalize the deformation caused by the perspective transformation we need to compute the homography between images. To compute the homography between 2 views of the same plane taken by same camera we need at least 3 correspondence of point pairs. In our case the painting segmented from the photograph taken



(a) Preserving the aspect ratio (b) Resampled to 800×600

Figure 5.8: Comparison of original and re-sampled painting.

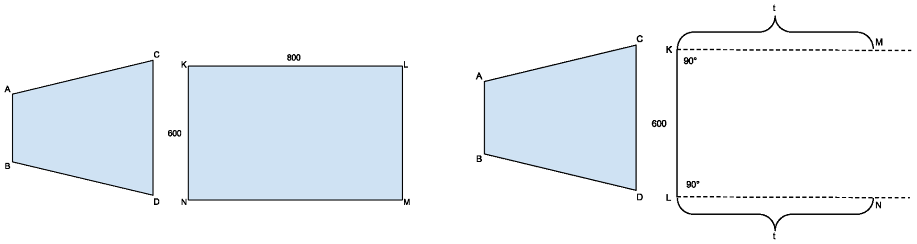
by tourist and the image of the painting from the database. As mentioned before we don't have any information about intrinsic camera parameters or the distortion of the lens used. As far as we can't estimate these parameters from one photograph with just have 4 known points (the corners of the segmented painting) with only partially known correspondences, we will assume that all images were taken with the same camera. To compute the affine homography we need 3 correspondences of point pairs, however in this part of our pipeline, we don't know which paintings we are dealing with and therefore we don't know the exact position of the corner points on the painting form the database. We propose 2 ways of solving of this problem.

The first one was described in the previous paragraph and it converts all the paintings to the 800×600 resolution and estimate the correspondence between corners in the way the longer distance is fit to 800 and shorter to 600. The image 5.9(a) shows the correspondence between 4 corner points on the images A, B, C, D and K, L, M, N.

The second approach preserves the aspect ratio of the original image. In the first step we find the first 2 points of the rectangle, K and L, then we can estimate 2 lines which lie perpendicular to the line connecting the K and L, then the M and N points lie on these lines in the same distance from K and L, which can be expressed by the parameter t. Then we can express 8 equations for 4 point pairs with 7 unknown parameters (6 for affine homography and one for t)

$$
\begin{aligned}
x_K &= h_{11}x_A + h_{12}y_A + h_{13}\\
y_K &= h_{21}x_A + h_{22}y_A + h_{23}\\
x_L &= h_{11}x_B + h_{12}y_B + h_{13}\\
y_L &= h_{21}x_B + h_{22}y_B + h_{23}\\
x_K + t(y_K - y_L) &= h_{11}x_C + h_{12}y_C + h_{13}\\
y_K + t(x_L - x_L) &= h_{21}x_C + h_{22}y_C + h_{23}\\
x_L + t(y_K - y_L) &= h_{11}x_D + h_{12}y_D + h_{13}\\
y_L + t(x_L - x_L) &= h_{21}x_D + h_{22}y_D + h_{23}.
\end{aligned}
\tag{5.13}
$$

When the transformation is estimated, we have to remap the painting on the rectangular image. We use backward mapping and the bilinear interpolation to compute the final pixel values.



(a) Resampled to 800 × 600

(b) Resampled with preserved aspect ratio

Figure 5.9: Two ways of resampling.

Until now we have only discussed the normalization of the shape and scale of the images, however the radiometric properties are also key for the normalization. As mentioned before the cameras used to capture our database are of unknown parameters, and we have to mention that also the images are of unknown previous compression. On the other hand, as far as we are dealing with the fine art galleries, the photographs of the same painting are photographs under the same or very similar lighting conditions (except for using the flash, which is usually forbidden to use). The problem is, that many parameters such as the white balance, the length of the exposition and the size of the aperture changes the produced intensities and color values and they are unknown for our images. For this reason we decided to normalize each color feature values, instead of normalizing the color values overall the image.

5.2.4 Local features

We have tested three different local feature methods: SIFT, SURF and ORB. We have chosen these methods because SIFT is a standard method, SURF is faster and almost equally precise, and ORB has a fast binary descriptor. All of them have their own detector and descriptor methods. SIFT and SURF detectors were described in the section 3.1.2 and ORB uses the modification of the FAST (also described in section 3.1.2) detector called oFAST.

The descriptor of SIFT forms a 128-valued vector — histogram of gradient orientations of the 16 areas in the neighborhood. SURF descriptor uses the values of Haar wavelet responses to form 64-valued vector. ORB uses modified BRIEF as a descriptor and forms 128-valued binary descriptor which stores the results of binary intensity tests.

5.3 Classification

The process of the classification consists of two phases: the preprocessing and the run time. In the preprocessing phase, the global and local features are extracted from the database of the Originals (see section 5.1) and stored.

The stored scalar global features consist of $8m$ bits, where m is the number of paintings in the database of Originals. In case of histogram features, the number was multiplied by the number of bins (256, resp. 90 for intensity and hue histograms). On the other hand, the size of local features dataset was $128 \cdot 8mn$ bits in the case of SIFT, $64 \cdot 8mn$ for SURF and $510mn$ in the case of ORB, where m is the number of paintings in the database of Originals and n the number of detected keypoints.

combination	number of bits
scalar + SIFT	$8m(1 + 128n)$
scalar + SURF	$8m(1 + 64n)$
scalar + ORB	$m(8 + 510n)$
int. hist. + SIFT	$1024m(2 + n)$
int. hist. + SURF	$512m(4 + n)$
int. hist. + ORB	$m(2048 + 510n)$
hue hist. + SIFT	$8m(90 + 128n)$
hue hist. + SURF	$8m(90 + 64n)$
hue hist. + ORB	$m(720 + 510n)$

Table 5.2: Number of bits needed to store extracted features.

Table 5.2 summarizes the number of bits need for storing different combination of features.

In the run time the segmented painting is classified as follows. In the first step, the global and local features are extracted from the segmented painting. The precomputed database of features is then sorted using the dissimilarity measure between the precomputed and extracted global feature values. The next step consists of matching the extracted local feature descriptors with the sorted database of precomputed local features. The match is found using the K-nearest neighbor classifier and the selected norm Euclidean distance or Hamming distance.

As the dissimilarity between the feature values of the paintings in the database of Originals and Photographs we use for scalar values the absolute value of the difference. The degree of dissimilarity between two histograms p and q was given by the symmetric Kullback-Leibler distance [71]

$$KLD(p, q) = \sum_{x \in I} (p(x) - q(x)) \log \left(\frac{p(x)}{q(x)} \right). \tag{5.14}$$

In order to select the best global feature, we sorted the differences in ascending order and found the position of corresponding paintings in the database of Originals. Figure 5.10 shows the cumulative histogram of the positions. We can see that with feature B, 90% of photographs (304) had the corresponding paintings in the database of Originals at position 35 or less, whereas with the second best feature only 82% (287) photographs had the corresponding paintings in the database of Originals at position 35 or less and 90% has the corresponding paintings in the database of Originals at place 45 or less. For this reason, the global feature used in our subsequent experiments is the percentage of light pixels in the painting.

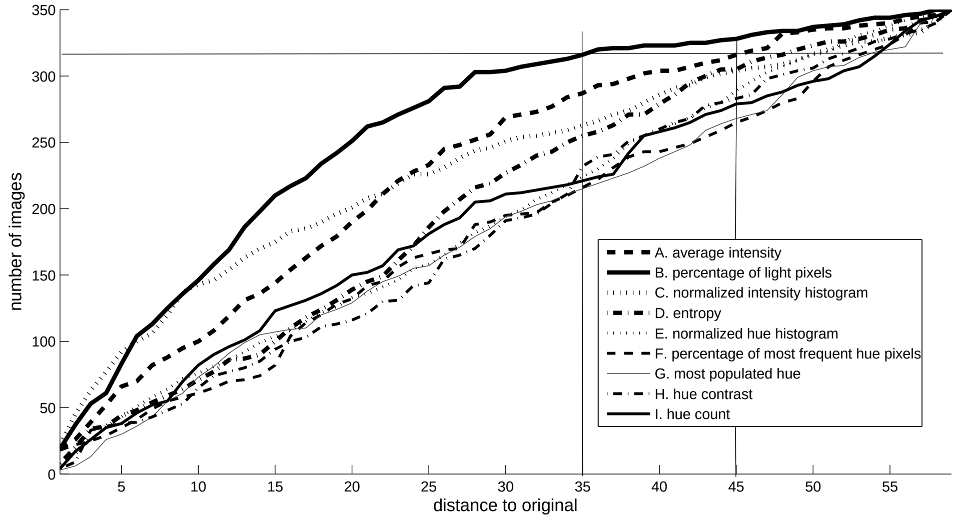


Figure 5.10: Cumulative histogram of positions for examined global features.

In our previous work [42] we have tested the performance of SIFT and SURF methods in classification of 100 paintings not included in the Photograph and Originals sets. Now we have conducted additional testing on the same data using the ORB method. Table 5.3 presents the result of the classification showing the precision and the threshold used to achieve this precision in classification of 100 paintings using SIFT, SURF and ORB method. Based on our evaluations we have decided to use the ORB method for the classification, as far as it is considerably faster and lacks only the 5% of precision of SIFT and SURF.

method	ideal threshold	precision
SIFT	12-14	90%
SURF	6	90%
ORB	5	85%

Table 5.3: Classification precision of SIFT, SURF and ORB method using ideal threshold on number of matches.

Method	SIFT	SURF	ORB
Time	0,8125 s	0,32025 s	0,01966

Table 5.4: Duration of one descriptor file creation using SIFT, SURF and ORB method.

5.4 Experiments and Results

As reasoned in the previous section, we use the difference in the percentage of light pixels between a photograph and the painting from the database of Originals to sort the database of local features. The local features are represented using the ORB descriptor.

In order to see if using the global feature sorting is beneficial, we conducted three experiments.

In the first one we found the distribution-based sorting of the database of precomputed features. The most photographed paintings from the database of Originals were at the top of the list. When not utilizing other features, this sorting minimizes the average position of keypoint matchings, so it optimal for a given distribution of the photographs.

The second experiment used randomly sorted database. We used several permutations of the positions, to get an average performance.

In the third experiment the database was sorted using the extracted global feature.

	sorting		
	global feature	distribution-based	random
mean	14.2925	19.3805	32.3089
std	11.4679	16.1783	17.0755

Table 5.5: The mean and standard deviation of position for three types of database sorting.

In all experiments we classified the images from database of photographs into 60 categories. For each image we also computed the the position of the corresponding original in the sorted database of originals.

5.4.1 Results

The results are summarized in table 5.5 showing the average position of a corresponding original. We can see that when the originals are placed randomly in the database, we match keypoints in average of 32 images in order to get the closes match (by finding the original). The sorting based on distribution of images in the Photographs dataset gave much better results. But distribution-based sorting can be done only when we know the distribution of the images in the database of photographs before the evaluation. Our approach generated the best average position of the corresponding painting from the database of Originals. The unpaired t-test for testing the difference of two means showed that the difference was statistically significant with P value less than 0.001.

Paintings from the Photographs dataset were classified into one of 60 categories (59 paintings from the database of Originals and 1 not in the database). The precision achieved by using different thresholds is visualized in figure 5.11. The threshold determines the number of keypoints matches needed to classify the photograph as a given painting from the database of Originals. The precision is computed as an average precision over all classes. Precision within class ω_i is the number of correctly classified ω_i paintings divided by the number of all paintings classified as ω_i.

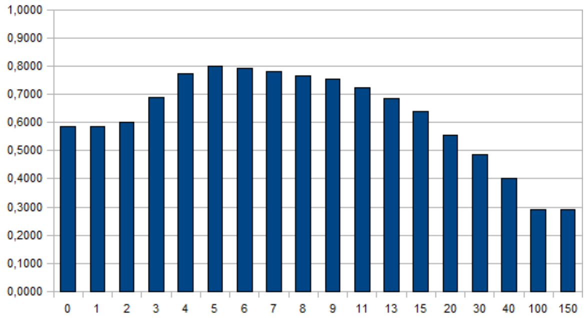


Figure 5.11: Average classification precision of ORB method for different thresholds on number of correct matches. The highest precision (80%) was achieved with the value 5. The average was computed over 10-fold cross validation on 1000 images.

The average case complexity evaluation shows that our approach is less complex than the approach with unsorted database. Since in a common case we do not know the distribution of photographs, we cannot sort the database to achieve globally ideal sorting. We have to assume that the average position of the corresponding paintings from the database of Originals is in the middle. In that case the complexity is $O(p0,5mn)$, where p is the number of photographs, $0,5m$ is the average position, i.e. number of comparisons and n is the average number of keypoints. The complexity of our approach is $O(p(m\log(m)+0,25mn))$, where $O(m\log(m))$ is the complexity of the sorting and $0,25m$ is the average position computed in our experiments. If we assume that the average number of keypoints is 500, then our approach is less computationally complex for up to 10^{125} paintings in the database of Originals. We can see that main computational load lies in the matching of vast number of keypoints. That is why the proper sorting of the database speeds up the process of classification.

5.5 Conclusions

In this chapter we propose a new method of classification of the special objects, fine art paintings. It combines the local and global features approaches. We have tested 3 types of local features and 9 global features and evaluated the method on 500 paintings.

Chapter 6

Related results

The presentation and preservation of the cultural heritage are the key tasks which are in last few years also supported by European Union through projects like Europeana, or Comeniana. Augmented reality plays a key role in the presentation of the cultural heritage in two base areas: the museum (gallery) guide and spatial installations. During the PhD. research we have contributed to both areas.

Our work on museum guides

In the area of museum (gallery) guides, we have cooperated on creation of a concept of the augmented guide like virtual installation in Czecho-Slovak pavilion at Biennale

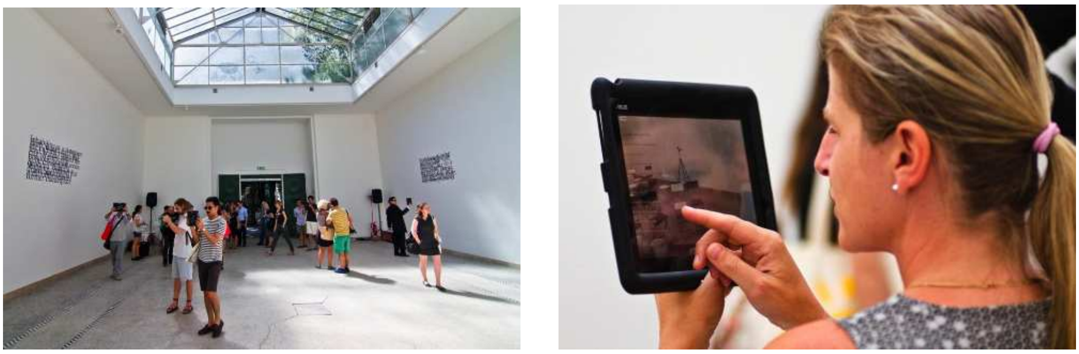


Figure 6.1: Photos of the pavilion and the installation.

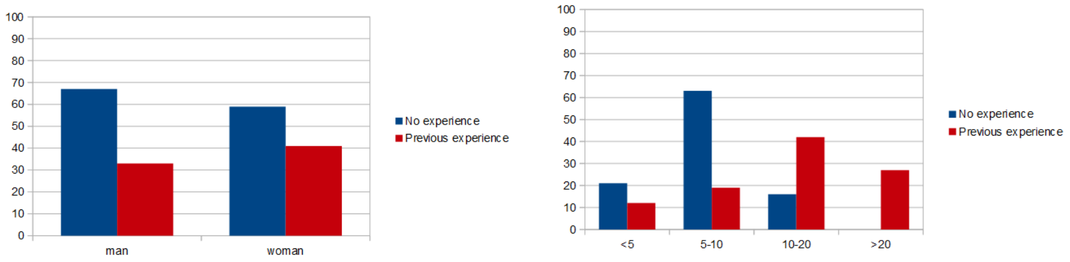


Figure 6.2: User study from Bienale 2012. Left: graph displaying the distribution of the users with and without previous experience with augmented reality among men and women. Right: graph displaying the dependence between previous experience with AR and the time spend exploring the installation in minutes.

of Architecture in Venice, Italy in 2012 (photos of the pavilion and the installation can be seen in figure 6.1).

We have carried out a user study focused on the previous experience with AR and it's correlation with the time spent on the installation. The user study was performed on 100 people 50 men and 50 women. Figure 6.2 shows two graphs. First one displays the distribution of the users with and without previous experience with augmented reality (theoretical or practical) among both men and women. Second graph presents the dependence between previous experience with AR and the time spent exploring the installation in minutes.

We have also created a concept of augmented reality gallery guide and published it in [45], [46]. Our guide is designed for the mobile devices (smartphones, tablets) and it uses the information from the camera of the device to detect and register the paintings in the gallery. Then it displays the augmented reality — the virtual information (image, video, text or 3D object) overlaying the real world stream. Our guide also provides audio with the comments about paintings and about the displayed virtual footage. We have also designed the virtual reality mode, which display the virtual footage without the necessity of pointing your device on the painting.

As explained in the previous chapters the key problem of the augmented reality guides and augmented reality applications generally is the registration of the real and virtual world. In our concept we have suggested the registration using local features.

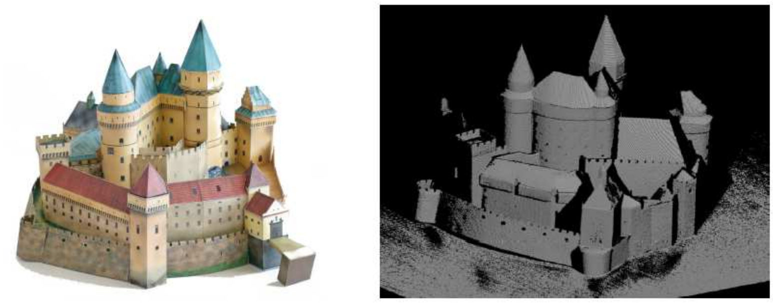


Figure 6.3: The photograph and the point cloud representation of the paper model of Bojnice castle.

In the area of object detection using local features we have surveyed, tested and designed different methods, problems and applications in our previous works since the master thesis [50], [42], [70].

Our work on spatial installations

Apart from the area of museum guides we have also proposed and implemented two spatial installation. The first one was published as a poster at Eurographics [43].

Reconstruction of cultural heritage object utilizing its paper model for augmented reality

In this work we have developed a method for augmenting the real paper model of the historical site with virtual animation. We have implemented our method for the augmented reality reconstruction of the fire in the Bojnice castle. The whole process can be seen in the figure 6.4. The process can be divided into preprocessing and run-time phases. In the preprocessing phase we scan the paper model using the SMISS structured light scanner developed by Tomáš Kovačovský and Jan Žižka [68]. The model is scanned from different positions and the final point cloud is created by merging of the partial point clouds. The photograph of the paper model and the corresponding point cloud can be seen in figure 6.3 For the registration of the real (paper) and virtual (3D model) models the paper markers similar to [61] are used and

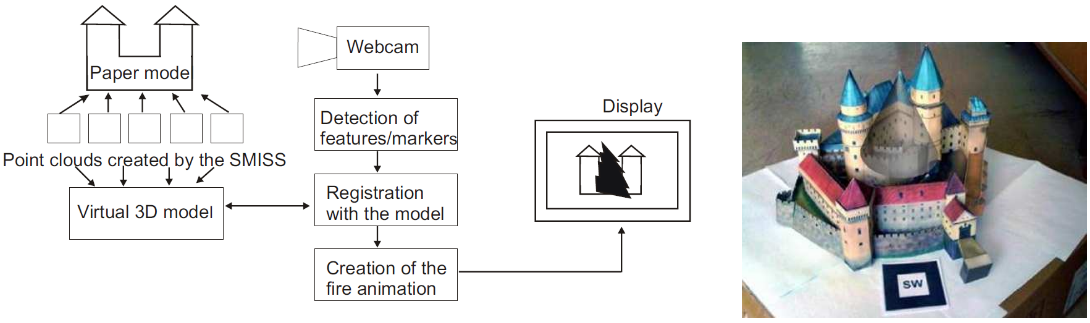

Paper mode
Point clouds created by the SMISS
Virtual 3D model
Webcam
Detection of features/markers
Registration with the model
Creation of the fire animation
Display
SW

Figure 6.4: Left: Scheme of the Augmented paper model installation. Right: Printscreen of the Augmented paper model installation.

their centers are manually marked in the virtual model. Then the virtual animation is created and registered with the virtual model. In the run-time phase the markers are detected and registered in the camera frame. Then the virtual animation is rendered on the video with the virtual model as an occluder. The resulting video then contains the properly registered animation — the parts occluded by the paper model are not visible.

Augmented map presentation of cultural heritage sites

Second created spatial installation combines augmented reality, cultural heritage and education. The application Slovak Augmented Reality uses a drawn floor map of Slovakia with several points of interests marked on the map. Kinect sensor is used to display and interact with 3D virtual models of cultural heritage objects using gestures. Our installation can be used for entertainment and as a learning tool in geography or history classes. In the area of AR there have been several works on interaction with the physical map, however it was usually a small paper map combined with the handheld or head-mounted device [86]. We have decided to create the context where the user can actually stand on the map directly on the location of the cultural heritage site. To display the 3D models we created mirror-like installation, where user sees himself standing on the map with 3D model in the front. Mirror-like installations are popular within the AR community because they allow the user to control his position and his gestures without refocusing from augmented to real environment.

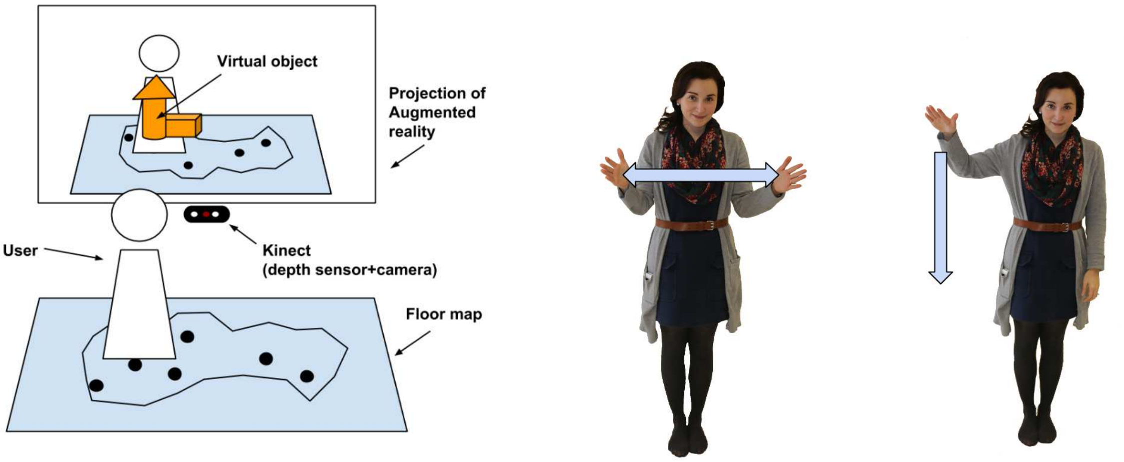


Figure 6.5: Left: Scheme of the Slovak Augmented Reality, Center and Right: Example of gestures for scaling and rotating.

Our system consists of a floor map of Slovakia, a projector and a projection screen (or a big display) and the Kinect device (or another RGBD sensor). The scheme of our proposed system can be seen in figure 6.5 and the photographs of the setup can be seen in figure 6.6. The initialization of our system is done as follows. The four corners of the map are marked on the RGB image acquired from Kinect. We compute the homography transformation between the four corners of the image of the real map and of the precreated model map with known positions (x, y) of the cultural heritage sites. Then the positions of the heritage sites in 2D (x, y) are computed and the depth

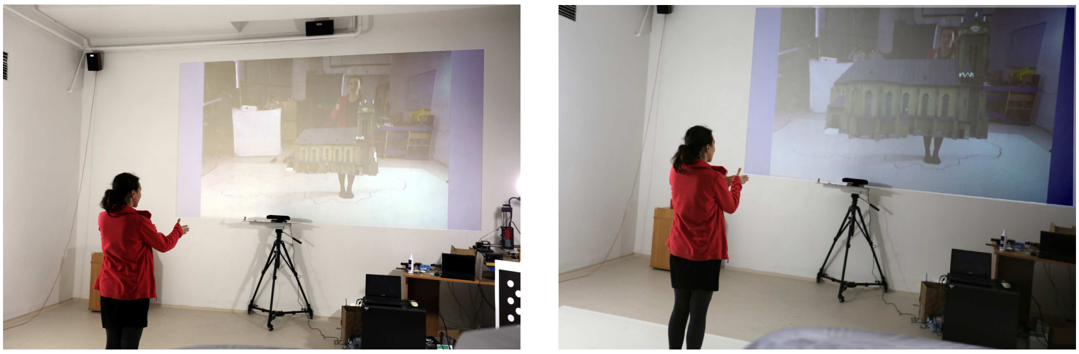


Figure 6.6: Scaling a 3D model using gestures.

is extracted from the initialization frame to achieve the 3D position information. In the runtime we track the skeleton of the user standing on the map extracted using the Kinect SDK. When the user steps on the marked heritage site, we display the 3D model of the corresponding cultural heritage object in front of the user on the projection screen. Then the virtual object can be transformed using defined gestures. We decided to perform two types of transformations: scaling and rotating along two axis. Our application runs on the PC (AMD Phenom II X4 3,4Ghz, 4GB Ram, ATI Radeon HD 5700 1GB) with 57 frames per second. The database of cultural heritage sites used in the installation consist of 8 virtual 3D models e.g. Slovak National Theater, New castle of Banská Štiavnica or mountain hut "chata pri Zelenom plese" which were created by students of our faculty and Martin Samuelčík.

Conclusions

This thesis has contributed to different areas of object detection and registration, augmented reality and cultural heritage presentation.

The main contribution was done in the area of multiple object detection and registration in both RGB and RGBD images. The detection and registration of multiple instances of objects is a problem closely connected with markerless registration for the purpose of augmented reality. In our work on detection of the multiple instances of the objects we proposed 2 new methods for 2D images and RGBD data in which we have overcome the limitations of previous state-of-the-art methods. Our methods do not suffer from the time consuming preprocessing phase [27], the scale constrains (all objects have to be of the same previously known scale) [132] and the limitation induced by the non-perspective deformations [95]. Our methods are based on local features and Hough based voting for the center of the object in 3D accumulator. The 3 dimensions of the accumulator are the x and y coordinates of the image and the rotation α of the object. We use SIFT features as they provide information about scale and rotation and are the most robust (based on our tests in [51]). In the RGBD method the depth information was utilized to compute the normal vectors in the feature points and to check the correct scale of the feature. We have developed a new method for filtering of the local feature matches using the scale ratio criterion, which outrun in precision and recall the previously used filtering methods (second nearest neighbor, both ways, threshold) in case of detection of multiple instances. Both our approaches were tested on 2 datasets in 3 tests. The 2D method works with 98% precision (97% recall) on artificial images and 85% precision (81% recall) on real images. The proposed 3D method works with 93% precision (89% recall) on real world

RGBD data. The 2D version of this method was accepted to SCCG conference [48] and the full version was submitted to ICCVG conference [49].

The second contribution of this thesis is in optimization of the recognition of fine art paintings. We dealt with the problem of matching of the local features with the big database of objects which is not final and can be extended consecutively. It is time consuming to utilize commonly used methods as random trees [75], spectral hashing [127] or bag of visual words [26, 29] as far as the data structures used for storing the local features need to be reconstructed every time the database is extended. We have overcome this problem with our method for classification of fine art paintings. Our method combines the segmentation, local and global feature approaches for efficient classification in big database. We have tested 3 types of segmentation methods (anisotropic diffusion, gauss gradient and watershed), 3 types of local features (SIFT, SURF and ORB) and 9 global features (average intensity, percentage of light pixels, normalized intensity histogram, entropy, normalized hue histogram, number of pixels that belong to the most frequent hue, most populated hue, hue contrast and hue count). The novelty of the method is in the utilizing of the global feature to reorder the database of paintings and therefore speed up the process of the local feature matching. The efficiency of the method was tested on 500 real world photographs of fine art paintings. The best results — 90% precision — were achieved by using the method combining anisotropic diffusion, SIFT or SURF features and percentage of light pixels global feature. The method also showed to be less computationally complex then previous methods for up to 10^{125} paintings in the database. This method was partially published in CESCG conference [42] and Computer Graphics and Geometry journal [50] and the full method was published in SCCG conference [51].

Within the scope of this thesis we have utilized our developed registration methods in the implementation of several installations which helped to present cultural heritage of Slovak republic on national and international events (Deň otvorených dverí FMFI 2014, TEDxBratislava 2013, Biennale of architecture in Venice 2012, Virtuálny svet v Avione 2012, Virtuálna realita bez hraníc 2012). The most important were the *Reconstruction of cultural heritage object utilizing its paper model for* *augmented reality* which was published as a poster at Eurographics [43] and *Slovak Augmented Reality* which is accepted to CISRGW [47].

In the future the speed of our multiple instances detection method can be improved utilizing the paralelization on GPU. The possible improvement of RGBD method can be done using data from the new version of the Kinect device (Kinect 2) which is based on time-of-flight approach and therefore does not have the problem caused by the fixed baseline and can display objects closer to the device. The quality of the depth map is also improved on the Kinect 2 based on the test carried out by the Photoneo company[1]. This will improve the computation of normals in our RGBD multiple instances detection and so further improve the performance of the method.

[1] http://www.photoneo.com/

Projects and publications

Selected publications (8 of 23)

- Haladová, Z. – Šikudová, E.: *Multiple instances detection.* In: Spring Conference on Computer Graphics SCCG'2013: Conference Proceedings. Bratislava: Comenius University, 2014. (accepted)

- Haladová, Z. – Samuelčík, M. – Varhaníková, I.: *Augmented map presentation of cultural heritage sites.* In: Current Issues of Science and Research in the Global World: Conference Proceedings. Vienna: Technical University, 2014. (accepted)

- Haladová, Z. – Šikudová, E.: *Combination of global and local features for efficient classification of paitings.* In: Spring Conference on Computer Graphics SCCG'2013: Conference Proceedings. Bratislava: Comenius University, 2013. pp. 21-27. ISBN 978-80-223-3377-1.

- Haladová, Z. – Šikudová, E.: *Limitations of the SIFT/SURF based methods in the classifications of fine art paintings.* In: Computer Graphics and Geometry: Journal. Vol. 12, No. 1, 2010. pp. 40-50. ISSN 1811-8992.

- Haladová, Z.: *Reconstruction of cultural heritage object utilizing its paper model for augmented reality.* In: Eurographics 2011: Conference Proceedings. Eurographics Association, 2011. pp. 7-8.

- Haladová, Z.: *Segmentation and classification of fine art paintings.* In: Central European Seminar on Computer Graphics: Conference Proceedings. Vienna: Institute of Computer Graphics and Algorithms, 2010. pp. 59-65.
- Haladová, Z. – Bolyós, Cs.: *Augmented gallery guide.* The 11th International Conference on Modeling and Applied Simulation: Conference Proceedings. Genoa, 2012. pp. 254-259. ISBN 978-88-97999-02-7.
- Haladová, Z. – Bolyós, Cs. – Šikudová, E.: *Concept of portable gallery guide.* 6th International Conference on Digital Arts: Conference Proceedings. Faro, 2012. pp. 321-324. ISBN 978-972-98464-7-2.

Monograph

- Šikudová, E. – Černeková, Z. – Benešová, W. – Haladová, Z. – Kučerová, J.: *Počítačové videnie: Detekcia a rozpoznávanie objektov.* - 1. vyd. - Praha: Wikina, 2014. - 378 s. ISBN 978-80-87925-06-5.

Invited talks

- Haladová, Z.: *Augmented reality and computer vision in context with new media.* Media lab. Academy of fine arts and design. Bratislava. 2013.
- Haladová, Z.: *Augmented reality and cultural heritage.* Invisible Štiavnica Workshop. Banská Štiavnica. 2013.
- Haladová, Z.: *Počítačové videnie, rozšírená realita a iné zázraky.* Krakatoa talk club. Bratislava. 2012.
- Haladová, Z.: *Rekonštrukcia historickej udalosti s využitǐm rozšǐrenej reality.* Vision and graphics group seminar. FIIT. Slovak Technical University. 2011.
- Haladová, Z.: *Rozšírená realita a jednoduché nástroje na jej vytváranie.* Mini-conference for pupils – Virtual reality without borders. Bratislava. 2011.

Research projects and travel grands

- *Comeniana – metódy a prostriedky digitalizácie a prezentácie 3D objektov kultúrneho dedičstva.* APVV 26240220077.
  Principal investigator: RNDr. David Běhal, PhD.

- *Intergration of visual information studies and creation of comprehensive multimedia study materials.* KEGA 068UK-4/2011.
  Principal investigator: RNDr. Zuzana Černeková, PhD.

- *Virtualizer: 3D Scanner for Complete Reconstruction.* 2012. Tatrabanka Foundation grant Etalent.
  Principal investigator: Mgr. Tomáš Kovačovský

- *Snímanie pohybu, interakcia a kooperácia ľudi a avatarov v 3D rozšírenej a virtuálnej realite (SPINKLAR-3D).* KEGA 068UK-4/2011.
  Principal investigator: RNDr. Stanislav Stanek, PhD.

- *Nová metóda detekcie a registrácie viacerých inštancií objektov.* UK/164/14. 2014. Comenius University grant.
  Principal investigator: RNDr. Zuzana Haladová.

- *Nová metóda registrácie multimodálnych dát s využitím lokálnych príznakov.* UK/228/13. 2013. Comenius University grant.
  Principal investigator: RNDr. Zuzana Haladová.

- *Travel grant for ICVSS summer school.* 2013. Tatrabanka Foundation grant Študenti do sveta.
  Principal investigator: RNDr. Zuzana Haladová.

- *Travel grant for SSIP summer school.* 2012. SPP Foundation grant Hlavička.
  Principal investigator: RNDr. Zuzana Haladová.

- *Travel grant for Eurographics conference.* 2011. Literary fund.
  Principal investigator: RNDr. Zuzana Haladová.

Awards

- Attendee of the 1. Heidelberg Laureate Forum 2013
- Best Presentation Award (1st place, based on public voting). Spring Conference on Computer Graphics 2013.
- ŠVK (Student Scientific Conference), Slovak literary fund award, 2013.
- 3rd best team project, SSIP summer school, 2012.
- Best Presentation Award (3rd place, based on public voting). Central European Seminar on Computer Graphics 2010.
- ŠVOČ (Student Scientific Conference, Czechoslovak international round), Winner 3rd place, 2010.
- ŠVK (Student Scientific Conference), Winner, Slovak literary fund award, 2010.

Tutoring

- Vozny M.: *CBIR system for microscopy images.* ŠVK (Student Scientific Conference), Winner, 2014.
- Franta R.: *Spatial Super Resolution.* ŠVOČ (Student Scientific Conference, Czechoslovak international round), Winner 1st place, 2013.
- Bolyós Cs.: *Scarlet – Fast Mobile Augmented Reality Library.* Winner of the Kunii prize 2013.
- Rjabinin I.: *Lowii – symbiosis of the music, painting and the computer generated art.* ŠVK (Student Scientific Conference), Winner, 2012.

Bibliography

[1] G. Abowd, C. Atkeson, J. Hong, S. Long, R. Kooper, and M. Pinkerton. Cyberguide: A mobile context aware tour guide. *Wireless Networks*, 3:421–433, 1997. 10.1023/A:1019194325861.

[2] M. Adrien and B. Claire. eMotion with the Leap Motion - Pepper's Ghost technique @ONLINE. http://vimeo.com/71216887, 2013.

[3] A. Alahi, R. Ortiz, and P. Vandergheynst. Freak: Fast retina keypoint. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 510–517. IEEE, 2012.

[4] ARToolworks. Flartoolkit @ONLINE. http://www.artoolworks.com/products /web/flartoolkit-2/, 2014.

[5] R. Azuma. A survey of Augmented reality. *Presence: Teleoperators and Virtual Environments*, 6(4):355–385, 1997.

[6] D. H. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern recognition*, 13(2):111–122, 1981.

[7] O. Bau and I. Poupyrev. REVEL: tactile feedback technology for Augmented reality. *ACM Trans. Graph.*, 31(4):89:1–89:11, July 2012.

[8] H. Bay, B. Fasel, and L. Van Gool. Gool. Interactive museum guide: Fast and robust recognition of museum objects. In *Proc. Int. Workshop on Mobile Vision*, 2006.

[9] H. Bay and *et al.* Speeded-up robust features (SURF). *Computer Vision and Image Understanding*, 110(3):346 – 359, 2008. Similarity Matching in Computer Vision and Multimedia.

[10] beefcakejim. Peppers ghost illusion tutorial @ONLINE. http://www.youtube.com/watch?v=xCYWEwtrxAk, 2008.

[11] S. Beucher and C. Lantuejoul. Use of watersheds in contour detection. In *International Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation,*, 1979.

[12] C. Bichlmeier, T. Sielhorst, S. M. Heining, and N. Navab. Improving depth perception in medical ar. In *Bildverarbeitung fur die Medizin 2007*, Informatik aktuell, pages 217–221. Springer Berlin Heidelberg, 2007. 10.1007/978-3-540-71091-2_44.

[13] O. Bimber, F. Coriand, A. Kleppe, E. Bruns, S. Zollmann, and T. Langlotz. Superimposing Pictorial Artwork with Projected Imagery. *IEEE MultiMedia*, 12(1):16–26, 2005.

[14] O. Bimber, B. Fröhlich, D. Schmalstieg, and L. M. Encarnação. The virtual showcase. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Courses*, USA, 2006. ACM.

[15] O. Bimber and R. Raskar. *Spatial Augmented Reality: Merging Real and Virtual Worlds.* A K Peters/CRC Press, July 2005.

[16] T. Blum. Augmented reality magic mirror using the Kinect @ONLINE. http://campar.in.tum.de/Chair/ProjectKinectMagicMirror, 2011.

[17] T. Bradley. Android Dominates Market Share, But Apple Makes All The Money @ONLINE. http://www.forbes.com/sites/tonybradley/2013/11/15/android-dominates-market-share-but-apple-makes-all-the-money/, 2013.

[18] B. R. Brkic, A. Chalmers, K. Boulanger, S. Pattanaik, and J. Covington. Cross-modal affects of smell on the real-time rendering of grass. In *Proceedings of the*

25th Spring Conference on Computer Graphics, SCCG '09, pages 161–166, New York, NY, USA, 2009. ACM.

[19] E. Bruns, B. Brombach, T. Zeidler, and O. Bimber. Enabling mobile phones to support large-scale museum guidance. *Multimedia, IEEE*, 14(2):16 –25, 2007.

[20] P. Burns. The history of the discovery of cinematography – 1860 – 1869 @ONLINE. http://www.precinemahistory.net/1860.htm, 2010.

[21] O. Cakmakci and J. Rolland. Head-worn displays: A review. *J. Display Technol.*, 2(3):199–216, Sep 2006.

[22] M. Calonder, V. Lepetit, C. Strecha, and P. Fua. Brief: Binary robust independent elementary features. In *Computer Vision–ECCV 2010*, pages 778–792. Springer, 2010.

[23] T. Caudell and D. Mizell. Augmented reality: an application of heads-up display technology to manual manufacturing processes. In *System Sciences, 1992. Proceedings of the Twenty-Fifth Hawaii International Conference on*, pages 659 – 669, 1992.

[24] A. Chalmers, D. Howard, and C. Moir. Real virtuality: a step change from virtual reality. In *Proceedings of the 25th Spring Conference on Computer Graphics*, SCCG '09, pages 9–16, New York, NY, USA, 2009. ACM.

[25] J. Chandrashekar, M. A. Hoon, N. J. P. Ryba, and C. S. Zuker. The receptors and cells for mammalian taste @ONLINE, 2006.

[26] Y. Chen, A. Dick, X. Li, and A. van den Hengel. Spatially aware feature selection and weighting for object retrieval. *Image and Vision Computing*, 31(12):935–948, 2013.

[27] A. Collet, D. Berenson, S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *IEEE International Conference on Robotics and Automation, 2009. ICRA '09.*, pages 48–55, 2009.

[28] C. Cruz-Neira, D. J. Sandin, and T. A. DeFanti. Surround-screen projection-based virtual reality: the design and implementation of the CAVE. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, SIGGRAPH '93, pages 135–142, New York, NY, USA, 1993. ACM.

[29] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, page 22, 2004.

[30] F. Cutzu, R. Hammoud, and A. Leykin. Distinguishing paintings from photographs. *Comput. Vis. Image Underst.*, 100(3):249–273, Dec. 2005.

[31] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[32] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3D object recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 998–1005, 2010.

[33] Esquire. Behind the scenes of augmented esquire @ONLINE. http://www.esquire.com/the-side/feature/augmented-reality-technology-110909, 2009.

[34] S. Feiner, B. MacIntyre, M. Haupt, and E. Solomon. Windows on the world: 2D windows for 3D augmented reality. In *Proceedings of the 6th annual ACM symposium on User interface software and technology*, UIST '93, pages 145–155, New York, NY, USA, 1993. ACM.

[35] S. Feiner, B. Macintyre, and D. Seligmann. Knowledge-based augmented reality. *Commun. ACM*, 36:53–62, 1993.

[36] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[37] P. Föckler and *et al.* PhoneGuide: museum guidance supported by on-device object recogn. on mob. phones. In *MUM '05: Proc. of the 4th intern. conf.*, pages 3–10, USA, 2005. ACM.

[38] E. Gorman. OpenGlass gives Google Glass real-time augmented reality ONLINE. http://www.engadget.com/2013/08/21/openglass-google-glass-real-time-augmented-reality, 2013.

[39] B. Gunsel, S. Sariel, and O. Icoglu. Content-based access to art paintings. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 2, pages 558–61, 2005.

[40] R. Hainich and O. Bimber. *DISPLAYS: FUNDAMENTALS & APPLICATIONS*. A K Peters/CRC Press, 2011.

[41] R. R. Hainich. *The End of Hardware: A Novel Approach to Augmented Reality.* BookSurge Publishing, 2006.

[42] Z. Haladová. Segmentation and classification of fine art paintings. *Central European Seminar on Computer Graphics for students (CESCG) , TU Wien*, 2010.

[43] Z. Haladová. Reconstruction of cultural heritage object utilizing its paper model for augmented reality. In *Eurographics 2011-Posters*, pages 7–8. The Eurographics Association, 2011.

[44] Z. Haladová. Would it work in Louvre? Evaluation of augmented museum guides. In *Proceedings of the Student Science Conference 2012*, pages 297 –305, 2012.

[45] Z. Haladová and C. Bolyós. Augmented gallery guide. In *The 11th International Conference on Modeling and Applied Simulation*, pages 254–259, 2012.

[46] Z. Haladová, C. Boylós, and E. Šikudová. Concept of portable gallery guide. In *6th International Conference on Digital Arts*, pages 321–324, 2012.

[47] Z. Haladová, M. Samuelčík, and I. Varhaníková. Augmented map presentation of cultural heritage sites. In *Accepted to: Current Issues of Science and Research in the Global World*, 2014.

[48] Z. Haladová and E. Šikudová. Multiple instances object detection. In *(accepted to) Spring Conference on Computer Graphics SCCG 2014*.

[49] Z. Haladová and E. Šikudová. Multiple instances object detection in rgb and rgbd images. In *(submitted to) International Conference on Computer Vision and Graphics ICCVG 2014*.

[50] Z. Haladová and E. Šikudová. Limitations of the SIFT/SURF based methods in the classifications of fine art paintings. *Computer Graphics & Geometry*, 12(1):40–50, 2010.

[51] Z. Haladová and E. Šikudová. Combination of global and local features for efficient classification of paitings. In *Spring Conference on Computer Graphics SCCG 2013*, pages 21–27, 2013.

[52] J. Han, D. Farin, and P. de With. Generic 3D modeling for content analysis of court-net sports sequences. In *Advances in Multimedia Modeling*, volume 4352 of *Lecture Notes in Computer Science*, pages 279–288. Springer Berlin / Heidelberg, 2006. 10.1007/978-3-540-69429-8_28.

[53] J. Han, D. Farin, and P. de With. Broadcast court-net sports video analysis using fast 3D camera modeling. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1628 –1638, nov. 2008.

[54] J. S. Hare and P. H. Lewis. Content-based image retrieval using a mobile device as a novel interface. In *Storage and Retrieval Methods and Applications for Multimedia 2005, 18 January 2005, San Jose, CA, USA*, volume 5682 of *SPIE Proceedings*, pages 64–75. SPIE, 2005.

[55] C. Harris and M. Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

[56] A. Herout, I. Szentandrási, M. Zachariáš, M. Dubská, and R. Kajan. Five shades of grey for fast and reliable camera pose estimation. In *Proceedings of CVPR*, 2013.

[57] S. Jiang, Q. Huang, Q. Ye, and W. Gao. An effective method to detect and categorize digitized traditional chinese paintings. *Pattern Recogn. Lett.*, 27(7):734–746, May 2006.

[58] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 21(5):433–449, 1999.

[59] R. E. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME – Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[60] E. D. Kaplan and C. J. Hegarty. *Understanding GPS: principles and applications.* 2006.

[61] H. Kato and M. Billinghurst. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *Augmented Reality, 1999. (IWAR '99) Proceedings. 2nd IEEE and ACM International Workshop on*, pages 85 –94, 1999.

[62] H. Kaufmann. Collaborative augmented reality in education. In *Proc. Imagina 2003 Conf. (Imagina03)*, 2003.

[63] Y. Ke and R. Sukthankar. PCA – SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.

[64] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on*, pages 225 –234, nov. 2007.

[65] G. Klein and D. Murray. Parallel tracking and mapping on a camera phone. In *Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on*, pages 83 –86, 2009.

[66] M. Kourogi and T. Kurata. Personal positioning based on walking locomotion analysis with self-contained sensors and a wearable camera. In *Proc. of the Second IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 103–112, 2003.

[67] M. Kourogi, N. Sakata, T. Okuma, and T. Kurata. Indoor/ outdoor pedestrian navigation with an embedded GPS/RFID/self-contained sensor system. In *Proceedings of the 16th International conference on Artificial Reality and Telexistence (ICAT 2006)*, pages 1310–1321.

[68] T. Kovaćovský. HDR SMISS -– fast high dynamic range 3D scanner. *Central European Seminar on Computer Graphics for students (CESCG), TU Wien*, pages 125–132, 2012.

[69] M. W. Krueger, T. Gionfriddo, and K. Hinrichsen. Videoplace -— an artificial reality. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '85, pages 35–40, New York, NY, USA, 1985. ACM.

[70] J. Kučerová and Z. Haladová. Towards semantic visual attention models. *Perception*, 42:219, 2013.

[71] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

[72] F. Kusunoki, M. Sugimoto, and H. Hashizume. Toward an interactive museum guide system with sensing and wireless network technologies. In *Wireless and Mobile Technologies in Education, 2002. Proceedings. IEEE International Workshop on*, pages 99 – 102, 2002.

[73] T. Langlotz. Studierstube tracker @ONLINE. http://handheldar.icg.tugraz.at/stbtracker.php, 2011.

[74] J. Lanier. Brief biography of Jaron Lanier @ONLINE. http://www.jaronlanier.com/general.html, 2010.

[75] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1465–1479, 2006.

[76] S. Leutenegger, M. Chli, and R. Y. Siegwart. Brisk: Binary robust invariant scalable keypoints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2548–2555. IEEE, 2011.

[77] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *J. Sel. Topics Signal Processing*, 3(2):236–252, 2009.

[78] T. Lombardi. *The Classification of Style in Painting*. VDM Publishing, 2008.

[79] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[80] S. Mann. Continuous lifelong capture of personal experience with eyetap. In *Proceedings of the the 1st ACM workshop on Continuous archival and retrieval of personal experiences*, CARPE'04, pages 1–21, New York, NY, USA, 2004. ACM.

[81] J. Martin-Gutierrez, M. Contero, and M. Alcaniz. Training spatial ability with augmented reality. In *Proceedings of Annual International Conference "Virtual and augmented reality in education" (VARE 2011)*, pages 49–53, 2011.

[82] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and vision computing*, 22(10):761–767, 2004.

[83] A. Mian, M. Bennamoun, and R. Owens. Three-dimensional model-based object recognition and segmentation in cluttered scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1584–1601, 2006.

[84] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino. Augmented reality: a class of displays on the reality-virtuality continuum. In *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, volume 2351 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pages 282–292, 1995.

[85] T. Miyashita, P. Meier, T. Tachikawa, S. Orlic, T. Eble, V. Scholz, A. Gapel, O. Gerl, S. Arnaudov, and S. Lieberknecht. An augmented reality museum guide. In *Proceedings of the 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, ISMAR '08, pages 103–106, Washington, DC, USA, 2008. IEEE Computer Society.

[86] A. Morrison, A. Oulasvirta, P. Peltonen, S. Lemmela, G. Jacucci, G. Reitmayr, J. Näsänen, and A. Juustila. Like bees around the hive: a comparative study of a mobile augmented reality map. In *Proceedings of the SIGCHI*, pages 1889–1898, 2009.

[87] T. Narumi, S. Nishizaka, T. Kajinami, T. Tanikawa, and M. Hirose. Augmented reality flavors: gustatory display based on edible marker and cross-modal interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 93–102, New York, NY, USA, 2011. ACM.

[88] NuFormer. First 3D video mapping projection zierikzee @ONLINE. http://www.nuformer.com/, 2011.

[89] T. Olsson and M. Salo. Online user survey on current mobile augmented reality applications. In *Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on*, pages 75–84, 2011.

[90] Open Shades. Testing wearscript with visually impaired users @ONLINE. http://www.youtube.com/watch?v=CEDg0k1HsH8/, 2013.

[91] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[92] P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. Technical report, Berkeley, CA, USA, 1988.

[93] P. Piccinini, A. Prati, and R. Cucchiara. A fast multi-model approach for object duplicate extraction. In *Applications of Computer Vision (WACV), 2009 Workshop on*, pages 1–6, 2009.

[94] P. Piccinini, A. Prati, and R. Cucchiara. SIFT – based segmentation of multiple instances of low-textured objects. In *Proceedings of international conference on machine vision (ICMV), Hong Kong*, pages 28–30, 2010.

[95] P. Piccinini, A. Prati, and R. Cucchiara. Real-time object detection and localization with SIFT – based clustering. *Image and Vision Computing*, 30(8):573 – 587, 2012. Special Section: Opinion Papers.

[96] Presselite. Metro Paris Subway iPhone and iPod Touch Application @ONLINE. http://www.metroparisiphone.com, 2010.

[97] Qualcomm. Vuforia @ONLINE. https://www.vuforia.com/, 2013.

[98] R. Raskar, J. van Baar, P. Beardsley, T. Willwacher, S. Rao, and C. Forlines. iLamps: geometrically aware and self-configuring projectors. In *ACM SIGGRAPH 2005 Courses*, SIGGRAPH '05, New York, NY, USA, 2005. ACM.

[99] Rayban. Ray-ban - virtual mirror @ONLINE. http://www.ray-ban.com/usa/science/virtual-mirror, 2011.

[100] J. Rekimoto. Transvision: A hand-held augmented reality system for collaborative design @ONLINE, 1996.

[101] J. Rekimoto. Matrix: a realtime object identification and registration method for augmented reality. In *Computer Human Interaction, 1998. Proceedings. 3rd Asia Pacific*, pages 63 –68, 1998.

[102] J. Rekimoto, A. Shionozaki, T. Sueyoshi, and T. Miyaki. PlaceEngine: a WiFi location platform based on realworld folksonomy. In *Internet conference*, volume 2006, pages 95–104, 2006.

[103] M. Richard. A sensor classification scheme. *IEEE Transactions On ultrasonic, ferroelectrics, and frequency control*, 34(2):124–126, 1987.

[104] Robert R. McCormick School of Engineering and Applied Science, Northwestern University. New 'batphone' app uses acoustics to determine location @ONLINE. http://www.mccormick.northwestern.edu/news/articles/article_935.html, 2011.

[105] E. Rosten and T. Drummond. Machine learning for high-speed corner detection. In *Computer Vision–ECCV 2006*, pages 430–443. Springer, 2006.

[106] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. ORB: an efficient alternative to SIFT or SURF. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2564–2571. IEEE, 2011.

[107] R. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (FPFH) for 3D registration. In *Robotics and Automation, 2009. ICRA '09. IEEE International Conference on*, pages 3212–3217, 2009.

[108] R. Sablatnig, P. Kammerer, and E. Zolda. Hierarchical classification of paintings using face – and brush stroke models. In *Proceedings of the 14th International Conference on Pattern Recognition-Volume 1 - Volume 1*, ICPR '98, pages 172–174, Washington, DC, USA, 1998. IEEE Computer Society.

[109] J. Shi and C. Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[110] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in RGBD images.

[111] T. Sielhorst, M. Feuerstein, and N. Navab. Advanced medical displays: A literature review of augmented reality. *Display Technology, Journal of*, 4(4):451 –467, dec. 2008.

[112] S. N. Sinha, J.-M. Frahm, M. Pollefeys, and Y. Genc. Feature tracking and matching in video using programmable graphics hardware. *Machine Vision and Applications*, 22(1):207–217, 2011.

[113] F. Sparacino. The museum wearable: real-time sensor-driven understanding of visitors' interests for personalized visually-augmented museum experiences. In *In: Proceedings of Museums and the Web (MW2002*, pages 17–20, 2002.

[114] A. State. Unc ultrasound/medical augmented reality research @ONLINE. http://www.cs.unc.edu/Research/us/, 2000.

[115] M. Stier and D. Grüntjens. Factors for knowledge transfer in mobile gamebased city tours on smartphones. In *Proceedings of Annual International Conference "Virtual and augmented reality in education" (VARE 2011)*, pages 106–113, 2011.

[116] I. E. Sutherland. The ultimate display. In *Proceedings of the IFIP Congress*, pages 506–508, 1965.

[117] I. E. Sutherland. A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I*, AFIPS '68 (Fall, part I), pages 757–764, New York, NY, USA, 1968. ACM.

[118] I. Szentandrási, M. Zachariáš, J. Havel, A. Herout, M. Dubská, and R. Kajan. Uniform marker fields: Camera localization by orientable de bruijn tori. In *Mixed and Augmented Reality (ISMAR), 2012 IEEE International Symposium on*, pages 319–320, 2012.

[119] D. Takahashi. Microsoft games exec details how project Natal was born @ONLINE. http://venturebeat.com/2009/06/02/microsoft-games-executive-describes-origins-of-project-natal-game-controls/, 2009.

[120] the macula. The 600 years @ONLINE. http://vimeo.com/15749093, 2010.

[121] The Museum of London. The streetmuseum @ONLINE. http://www.museumoflondon.org.uk/ Resources/app/you-are-here-app/index.html, 2010.

[122] The Spitfire Site. Mark II Gyro Gunsight @ONLINE. http://spitfiresite.com/2007/11/mark-ii-gyro-gunsight.html, 2007.

[123] B. Thomas, B. Close, J. Donoghue, J. Squires, P. De Bondi, M. Morris, and W. Piekarski. Arquake: an outdoor/indoor augmented reality first person application. In *Wearable Computers, 2000. The Fourth International Symposium on*, pages 139 –146, 2000.

[124] E. Tola, V. Lepetit, and P. Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010.

[125] F. Tombari, S. Salti, and L. Stefano. Unique signatures of histograms for local surface description. In *Computer Vision – ECCV 2010*, volume 6313 of *Lecture Notes in Computer Science*, pages 356–369. Springer Berlin Heidelberg, 2010.

[126] TV Worth Watching. NBC's olympics coverage – so far, so great @ONLINE. http://www.tvworthwatching.com/blog/2008/08/, 2008.

[127] J. Ventura and T. Höllerer. Fast and scalable keypoint recognition and image retrieval using binary codes. In *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, pages 697 –702, jan. 2011.

[128] Vsauce. 12 best Kinect hacks 2010 @ONLINE. http://www.youtube.com/watch?v=ho8KVOe _y08, 2010.

[129] D. Wagner and D. Schmalstieg. First steps towards handheld augmented reality. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers*, ISWC '03, pages 127–135, Washington, DC, USA, 2003. IEEE Computer Society.

[130] M. Weiser. The computer for the 21st century. *Scientific American*, 1991.

[131] F. Zhou, H. B.-L. Duh, and M. Billinghurst. Trends in augmented reality tracking, interaction and display: A review of ten years of ISMAR. *Mixed and Augmented Reality, IEEE / ACM International Symposium on*, 0:193–202, 2008.

[132] S. Zickler and M. Veloso. Detection and localization of multiple objects. In *Humanoid Robots, 2006 6th IEEE-RAS International Conference on*, pages 20–25, 2006.

[133] A. Zimmermann and A. Lorenz. LISTEN: a user-adaptive audio-augmented museum guide. *User Modeling and User-Adapted Interaction*, 18(5):389–416, 2008.

[134] B. Zitová and J. Flusser. Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000, 2003.