Federal University of Bahia
State University of Feira de Santana

MASTER THESIS

A multi-view environment for markerless augmented reality

Caio Sacramento de Britto Almeida

Master in Computer Science – MMCC

Salvador
December 15th, 2014

MMCC-Msc-0007

CAIO SACRAMENTO DE BRITTO ALMEIDA

A MULTI-VIEW ENVIRONMENT FOR MARKERLESS AUGMENTED REALITY

Thesis submitted to the Master Program in Computer Science from Federal University of Bahia and State University of Feira de Santana, in partial fullfillment of the requirements for the degree of Master in Computer Science.

Advisor: Antônio Lopes Apolinário Júnior

Salvador
December 15th, 2014

Index card.

Sacramento de Britto Almeida, Caio

A multi-view environment for markerless augmented reality/ Caio Sacramento de Britto Almeida– Salvador, December 15th, 2014.

21p.: il.

Advisor: Antônio Lopes Apolinário Júnior.
Thesis (master)– Federal University of Bahia, Institute of Mathematics, December 15th, 2014.

1. Multi-view environment. 2. Augmented reality. 3. Computer graphics.
I. Apolinário Jr., Antônio Lopes. II. Federal University of Bahia. Institute of Mathematics. III A multi-view environment for markerless augmented reality.

CDD 20.ed. 123.45

APPROVAL SHEET

CAIO SACRAMENTO DE BRITTO ALMEIDA

A MULTI-VIEW ENVIRONMENT FOR MARKERLESS AUGMENTED REALITY

This thesis was considered worthy of acceptance of the master's degree in Computer Science and approved on its final form by the UFBA-UEFS Master Program in Computer Science.

Salvador, December 15th, 2014

Prof. Dr. Antônio Lopes Apolinário Júnior 1
Federal University of Bahia - Brazil

Prof. Dr. Michelle Ângelo 2
State University of Feira de Santana - Brazil

Prof. Dr. Rodrigo Silva 3
Federal University of Juiz de Fora - Brazil

ACKNOWLEDGEMENTS

I would like to thank my family for all the support during all those years, not only the time spent on the master program, but also during all my graduation. Namely, first I would like to thank my brother, Rodrigo, for being my inspiration on following this career; my twin sister, Thalita, for being my support of all times; my parents, for the great education that was given to me; my grandmother Lourdes; my girlfriend Jéssica; my friends; and specially my advisor Antônio Apolinário for relying on me to make this work happen.

ABSTRACT

Augmented reality is a technology which allows 2D and 3D computer graphics to be aligned or registered with scenes of the real-world in real-time. This projection of virtual images requires a reference in the captured real image, which is often achieved by using one or more markers. But, there are situations where using markers can be unsuitable, like medical applications, for example. In this work, we present a multi-view environment, composed by augmented reality glasses and two Kinect devices, which doesn't use fiducial markers in order to run augmented reality applications. All devices are calibrated according to a common reference system, and then the virtual models are transformed accordingly too. In order to achieve that, two approaches were specified and implemented: one based on one Kinect plus optical flow and accelerometer data from augmented reality glasses, and another one based purely on two Kinect devices. The results regarding quality and performance achieved by these two approaches are presented and discussed, as well as a comparison between them.

Keywords: augmented reality, augmented reality glasses, kinect, transformation, optical flow, markerless

CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter

1

In this chapter I present the motivation, objectives and overview of this work.

INTRODUCTION

1.1 MOTIVATION

Augmented reality has taken advantage from the progresses on the fields of multimedia and virtual reality, making feasible new forms of interaction between humans and machines. Differently from virtual reality, that takes the user to a virtual environment, the augmented reality keeps the user on his physical environment and takes the virtual environment to the user's space, allowing the interaction with the virtual world, in a more natural manner and without needing training or adaptation (**??**). Many times, this interaction means merging virtual images with images captured from a real environment.

One of the greatest challenges on the field of augmented reality is to determine, in real time, which virtual image to be displayed, in which position and how it should be represented. In order to obtain an integration illusion between real objects and virtual objects, the generated object should be aligned with the tridimensional position and orientation of the real objects (**??**). This can be achieved by estimating the camera position.

On many situations, fiducial markers[1] are used (often this is due to the real time requirements of the augmented reality applications) (**??**) and are drawn in a way that they can be easily identified. Those markers need to be placed on the target scene and can achieve great results using just a few computational resources. Figure 1.1 shows the usage of a fiducial marker and a tridimensional object being projected over it.

However, besides requiring human interference on the scene, there situations where the usage of fiducial markers is not possible, feasible or comfortable for the target model. That is the case, for example, of medical applications on which this model is a patient. It's also possible to cite other limitations of fiducial markers, like, for example, occlusion (a virtual image could be not projected if the marker is not completely visible) and illumination (the intensity of light reflected by the marker could make it hard to be

[1] A fiducial marker or fiducial is an object placed in the field of view of an imaging system which appears in the image produced, for use as a point of reference or a measure.

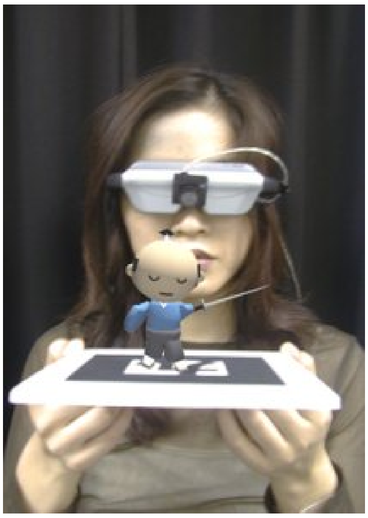

Figure 1.1 Fiducial marker used to represent a tridimensional model over it (??)

identified). Less common, there are approaches that replace fiducial markers (??) (??) by GPS, gyroscopes, accelerometers, cameras, among others (??) (??). These approaches have the advantage of not requiring human interference on the scene (to put a marker or to move it around).

Depending on the way that a user sees the mixed world, augmented reality can be classified on two ways. When the user sees the mixed world pointing his eyes straight to the real positions with optical scene or video, this augmented reality is called *immersive* or of *direct vision*. On the other hand, when the user sees the mixed world by some device, like a monitor screen or projector, not aligned with the real positions, this augmented reality is *non-immersive* or of *indirect vision* (??).

This work proposes a multi-view environment for augmented reality, of direct vision, composed by two Kinects (??) and augmented reality glasses, that allows a watcher visualize, in real time, virtual images merged with real images from the target model. In this approach, it's not intended to use any fiducial marker. Instead, it will be used a geometric approach based on the data captured by each Kinect. This proposed environment can be used, for example, on the medical field (real situations, education and training) or in other situation where a multi-view environment for markerless augmented reality is applicable.

1.2 OBJECTIVES

The environment proposed on this work aims to contribute to augmented reality applications where virtual images need to be merged with real images in real time, without using fiducial markers, and considering the viewing angle of the observer and the position of the target object.

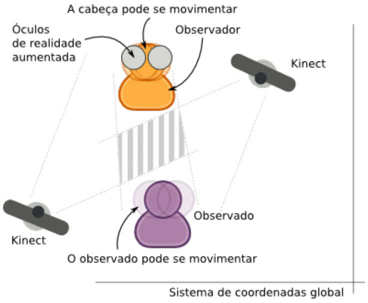
A cabeça pode se movimentar
Óculos de realidade aumentada
Observador
Kinect
Observado
Kinect
O observado pode se movimentar
Sistema de coordenadas global

Figure 1.2 Global vision

1.2.1 Features

Based on the study of related works, it was defined the following features that define the scope of the first version of the augmented reality environment, graphically represented on Figure (**??**):

- There are two main elements on the environment, the observer and the target model;
- Observer and target model are positioned one in front of the other, with some restriction of minimum and maximum distance;
- The observer sees the combination of a real image with a pre-defined virtual image over the target model, through the augmented reality glasses;
- The observer and target model can move their heads, in a way that the virtual image adapts itself in real time to fit the new viewing angles;
- No fiducial marker is used.

The elements presented on this environment are the following:

- Observer - He is the user of the system that wears a pair of augmented reality glasses and is positioned in front of the target model (which can be a human being, for example, or a static object). On a medical context, the observer would be a medical specialist, responsible for observing the patient and responsible for analyzing the combination of the real image (part of the patient body) with a virtual image (from magnetic resonance imaging or computed tomography). The observer is able to move his head.
- Augmented reality glasses - The augmented reality glasses are wore by the observer and have two cameras. The images captured from the target model will be merged with virtual images and displayed on the two lenses of these glasses. On one of the approches implemented by this work, the observer movimentation is determined from sensors present on the glasses: acceletometer, magnetometer and gyroscopes. Based on sensor data, it's possible to determine the variation on the orientation of the glasses, and so it's possible to know how much the observer has moved his

head. This calculation returns values that define moviments on longitudinal axis, vertical axis and lateral axis. The virtual image should be reprojected in real time according to those movements.

- Target model - The target model (a human being, for example), is placed in front of the observer and doesn't use any kind of fiducial marker. The main goal is that the virtual image is placed over the target model. In order to calculate where and how the virtual image should be displayed, it's necessary to identify the position and orientation of the real object relative to the observer. This is done based on two sensors placed on the environment, where the first one captures data from the observer and the second one captures data from the target model.

- Sensors - Two sensors are placed on the environment and capture data from the observer and the target model (one for each). The information captured by the target model's sensor contains its model, that will be merged with the virtual image.

Each device presented on this multi-view (glasses and sensors) environment has its own coordinate system, but all information must be converted to a global coordinate system.

Since there are two sensors and one pair of augmented reality glasses, another objective is to implement two different approaches: one that uses the augmented reality glasses to determine the observer's pose (based on data from accelerometer and magnetometer) and another one that uses a second Kinect device to determine the observer's pose based on a reconstruction of his model.

1.3 THESIS OVERVIEW

The next chapter "Conceptual primer" describes some basic theory needed. It also describes the hardwares that are used by this work and how to calibrate them. After that, the "Related work" chapter presents some works related to this one, divided by subject. The third chapter, "Solution architecture", after the theory and related works were presented, explains the steps performed in order to implement the objectives of this work. The results of this implementation are presented on the chapter later, "Results", and finally the conclusions about those results are presented on the last chapter, "Conclusions", where possible future works are also listed.

Chapter

2

In this chapter I present the main concepts behind this work.

CONCEPTUAL PRIMER

2.1 AUGMENTED REALITY

Augmented reality was born on the decade of 1990, to merge a virtual image or virtual environment with a real image or real environment. But only from the 2000s it became more popular, due to lower costs of hardware and software devices, and ready to be used on tangible and multimodal (voice, touch, gesture, etc.) applications (**??**). It can be considered the mixing of real and virtual worlds at some point of the continuum reality-virtuality, that connects completely virtual environments to completely real environments (**??**), like shown on Figure 2.1. It can also be considered a system that completes the real world with virtual objects, in a way that they seem to exist on the same space, respecting the following features:

- Real objects are mixed with virtual objects;
- Execution is interactive and on real time;
- Virtual and real objects are aligned;
- Applicable to all human sensory systems, including auditory, olfactory and somatosensory (**??**).

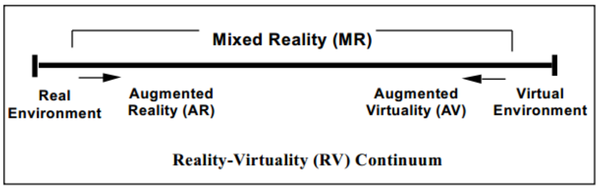

Figure 2.1 The augmented reality is localized between the extremes of the reality-virtuality continuum

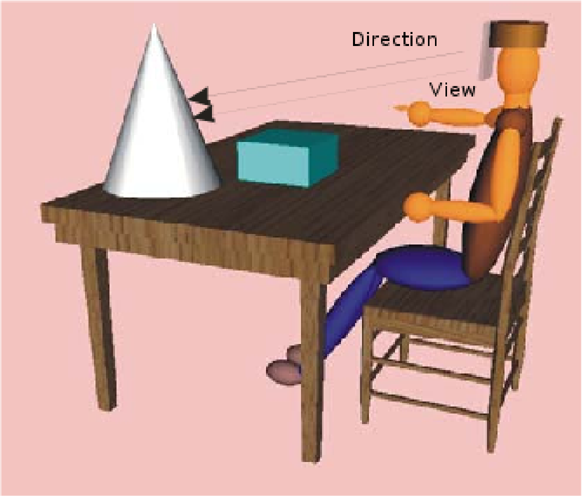

Figure 2.2 Immersive augmented reality

From a human-computer interaction perspective, the augmented reality can be considered a new way of interaction between humans and computers, and, on this aspect, it can be classified as direct vision or indirect vision.

2.1.1 Direct or indirect vision

Augmented reality is classified as *direct vision* or *immersive* when the user sees the mixed world pointing his eyes straight to the real position of the objects of interest, like shown on Figure 2.1.1.

On the direct vision, images from the real world can be seen with the naked eye ou brought by video, while the generated virtual images can be projected on the eyes, on the real scenario or mixed with the real world video. Immersive augmented reality can be implemented with optical helmets, for example (**??**).

On the other hand, non-immersive augmented reality (indirect vision), happens when the user sees the mixed world through other output device, like a monitor screen, for example, not aligned with the real positions, as shown on Figure 2.1.1. On this kind of vision, real and virtual images are merged and displayed as video to the user. It can be achieved by using cameras and projectors (**??**).

2.1.2 Markers

In order to identify the position where a virtual image should be rendered, two main approaches can be applied by augmented reality applications: one is to use fiducial markers, the other is not use them.

2.1.2.1 Fiducial markers Fiducial markers are often implemented as square white cards with a black symbol printed (or drawn) on it, easy to be recognized, working like a barcode or QR code. Computer vision techniques are used to calculate the position of

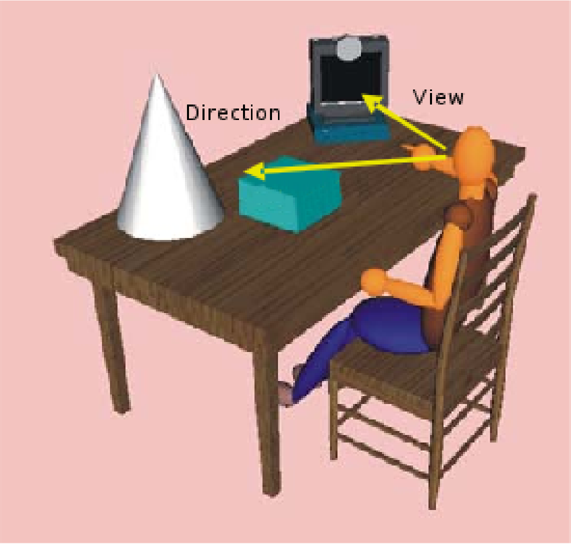

Figure 2.3 Non-immersive augmented reality

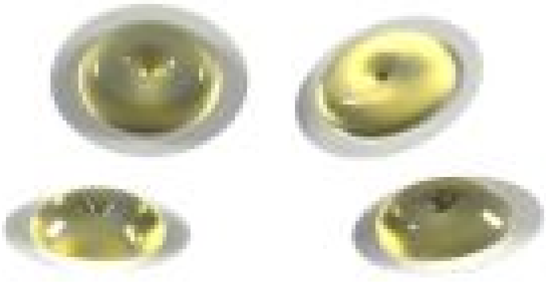

Figure 2.4 Example of fiducial marker used by medical applications

the real camera and its orientation relative to the markers, in a way that virtual objects can be projected over them.

Fiducial markers can assume other shapes besides a square card. The Figure 2.1.2.1 shows an example of a fiducial marker used on medical applications, similar to a sticker, which is fixed on the patient's skin. On these applications, augmented reality can be used for visualization and training on surgeries. It's possible to collect patient's data in real time, by using non-invasive sensors as the ones used for magnetic ressonance imaging and computed tomography. This dataset can be merged in real time with the real image of the patient (**??**).

On motion capture systems, other kinds of markers can also be used, like, for example, the one described on (**??**) and represented on Figure 2.1.2.1. In order to estimate the pose, it uses a single camera and three spherical fiducial markers, which are also reflexive. Two of those markers are placed at the left and right sides of the actor, perpendicular to his ears. The third marker is placed on the same height as the other two, but in front of

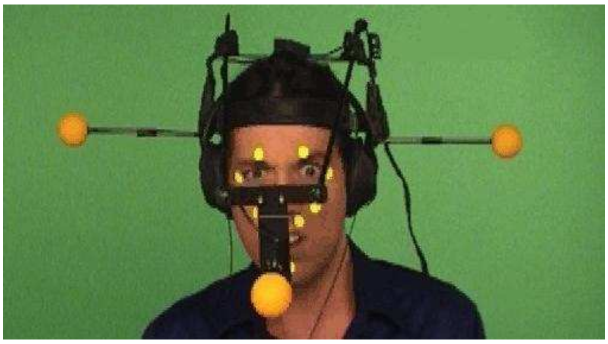

Figure 2.5 Example of fiducial marker for motion capture

the actor. The markers have a color close to orange for high reflectance (in order to be easily recognized) and are supported by a structure made of carbon fiber (which is light).

Disadvantages of fiducial markers are listed on (**??**), for example, they are invasive (need to be placed on the scene or fixed on the target object), they have limited interaction (can't be moved around so much because it needs to still be visible and recognizable) and need to be printed or drawn before using and stored for future usage.

2.1.2.2 Markerless Markerless augmented reality means that fiducial markers are not used on the scene. Instead, features from the target scene need to be used in order to estimate the camera pose and objects orientations.

Target detection based on computer vision has been extensively studied e successfully applied on augmented reality applications. On related computer vision literature, geometric primitives can be used for pose estimation, on most cases, points, segments, lines, edges of edge points, cones, cylinders, or a combination of two or more of those features.

Markerless augmented reality has the advantage of using parts of the real environment as targets and can even extract information from the environment to be used by the augmented reality system (**??**).

Avoiding markers leads to a much more effective augmented reality experience, but requires the implementation of several image processing or sensor function techniques, resulting in more complex algorithms and in higher computational resources (Shumaker, 2011).

2.2 CAMERAS

Many tasks on augmented reality deal with an imaging device. Usually this imaging device is a camera, which performs a mapping from a 3D world to a 2D image (Hanning, 2011). The problem called *camera calibration* is the one handles the determination of the parameters for this mapping.

2.2.1 Calibration

Camera calibration means to determine the camera model parameters which fit best to the observed behavior of the actual camera (Hanning, 2011). So, it's necessary to measure the the distance of an observation to a given camera model. The determination of the optimal camera mapping concerning each of these distance functions defines a non-linear optimization problem, and as such, it depends on the initial value.

Ordinary cameras are very often modelled as pinhole cameras. A pinhole is an imaginary wall with a tiny hole in the center that blocks all rays except the ones that pass through this tiny center hole (Bradski; Kaehler, 2008). Unfortunately, a real pinhole is not a good way to make images because it does not gather enough light for rapid exposure. This is why human eyes and cameras use lenses to gather more light than what would be available at a single point. This way, the simple geometry of the pinhole camera model is not enough and it also introduces distortion of the lens itself.

The camera coordinate system (CCS) is a Cartesian coordinate system defined by the principal plane: The x-axis and y-axis of the CCS determine the principal plane, the z-axis is given by the optical axis. The optical center of the lens determines the origin (0, 0, 0) of the CCS. Thus, in the camera coordinate system, the principal plane becomes $z = 0$ (Hanning, 2011).

Camera calibration is usually split up into two distinct parts, the intrinsic and extrinsic parameters, which will be covered on the following section.

2.2.2 Parameters

The process of camera calibration gives us both a model of the camera's geometry and a distortion model of the lens. Both compose the intrinsic parameters of the camera.

2.2.2.1 Intrinsic parameters The intrinsic parameters are those that describe the internal workings of the camera and consist of the focal length, the principal point, the skew coefficients and radial and tangential distortions (Tillapaugh; Engineering, 2008). These values are needed to help to describe imperfections in the lens of the camera and give a mapping from camera reference frame to the image plane. The intrinsic parameters depend only on the camera itself, and so they just need to be found once, regardless the environment changes or not.

2.2.2.2 Extrinsic parameters The extrinsic parameters, on the other hand, represent the viewpoint of the camera by a rigid transformation, which describes its position and orientation (Bajramovic, 2010). This part of the model is independent of the camera itself. The according parameters are called extrinsic, as they describe the relation between the camera and the world.

According to (Tillapaugh; Engineering, 2008), the extrinsic parameters are those that are dependent on the environment. To relate an object's coordinate system to the world's coordinate system, a translation and a rotation matrix are needed. Therefore, the extrinsic parameters consist of these two matrices, so that a mapping from the world coordinate system to the camera reference frame can be found. Since the extrinsic parameters for the camera explain how the camera relates to the environment, if the camera changes position, the parameters have to be recalculated (differently from the intrinsic parameters as explained on the previous section).

The relation between world coordinate system and the camera reference frame can be described by the equation:

$$X_c = R_c \times X + T_c$$

Where X is a 3x1 vector that represents a point on the world coordinate space, R_c is the extrinsic rotation matrix, T_c is the extrinsic translation matrix and X_c is a 3x1 vector in the camera reference frame. The rotation matrix is build from the combination of three single-axis rotation matrices:

$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos(\theta) & sin(\theta) \\ 0 & -sin(\theta) & cos(\theta) \end{bmatrix}$$

$$R_y(\phi) = \begin{bmatrix} cos(\phi) & 0 & -sin(\phi) \\ 0 & 1 & 0 \\ sin(\phi) & 0 & cos(\phi) \end{bmatrix}$$

$$R_z(\omega) = \begin{bmatrix} cos(\omega) & sin(\omega) & 0 \\ -sin(\omega) & cos(\omega) & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The extrinsic parameters can be represented by a final matrix called *homegenous matrix*. It looks as follows:

$$\begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix}$$

So, in order to convert from a world origin coordinate to a camera orientation coordinate system, only a single homogenous matrix needs to be used.

2.2.3 Multi-view

Calibrating a multi-camera system accordingly means estimating the intrinsic and extrinsic parameters of all cameras (Bajramovic, 2010). Actually it is mainly concerned with estimating extrinsic parameters, since most approaches first compute intrinsic parameters individually for each camera. Extrinsic parameters are subsequently estimated given the intrinsic parameters.

2.3 SENSOR-BASED REGISTRATION APPROACH

2.3.1 Kinect

2.3.2 Registration

2.3.3 Transformation

2.4 VISION-BASED REGISTRATION APPROACH

2.4.1 Optical flow

2.4.2 Lucas-Kanade algorithm

Chapter

3

In this chapter I present some related works.

RELATED WORK

Cover all the related works, with multiple Kinects, optical flow, markerless augmented reality, medical applications, multi-view environment, reconstruction, etc.

3.1 MARKERLESS AUGMENTED REALITY

3.2 MULTIPLE KINECTS

3.3 MULTI-VIEW ENVIRONMENT

Chapter

4

In this chapter I explain in details the steps performed in order to implement the objective of this work.

SOLUTION ARCHITECTURE

4.1 ENVIRONMENT

Technologies, machines, SOs, etc.

4.2 CALIBRATION

4.2.1 Augmented reality glasses calibration

4.2.2 Initial calibration between Kinects

4.2.3 Initial calibration between Kinect and glasses

4.3 COMMUNICATION

Network, sockets, etc.

4.4 TRANSFORMATIONS

4.5 METHOD 1: GLASSES ACCELEROMETER AND ONE KINECT

Cover Lucas-Kanade algorithm, etc.

4.6 METHOD 2: TWO KINECTS

Talk about performance of two Kinfus fighting for a single GPU.

4.7 HYBRID APPROACH

When optical flow has just a few feature points, we switch to the second Kinect.

Chapter 5

In this chaper I present the results of the procedure explained in the previous chapter.

RESULT

5.1 SCOPE

Talk about error propagation.

5.2 ANALYSIS

Talk about performance and alignment results.

5.3 COMPARISON

Compare methods 1 and 2 with regards to performance and quality.

Chapter 6

In this chapter I discuss the conclusions of this work and list some possibilities of future works.

CONCLUSIONS

6.1 FUTURE WORK

BIBLIOGRAPHY

Bajramovic, F. *Self-Calibration of Multi-Camera Systems.* [S.l.]: Logos Verlag Berlin, 2010. ISBN 9783832527365.

Bradski, G.; Kaehler, A. *Learning OpenCV: Computer Vision with the OpenCV Library.* [S.l.]: O'Reilly Media, 2008. ISBN 9780596554040.

Hanning, T. *High Precision Camera Calibration.* [S.l.]: Vieweg+Teubner Verlag / Springer Fachmedien Wiesbaden GmbH, Wiesbaden, 2011. (Vieweg + Teubner research). ISBN 9783834898302.

Shumaker, R. *Virtual and Mixed Reality - New Trends, Part I: International Conference, Virtual and Mixed Reality 2011, Held as Part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings.* [S.l.]: Springer, 2011. (Information Systems and Applications, incl. Internet/Web, and HCI). ISBN 9783642220203.

Tillapaugh, B.; Engineering, R. I. of T. C. *Indirect Camera Calibration in a Medical Environment.* [S.l.]: Rochester Institute of Technology, 2008. ISBN 9780549934479.