Using RGB-D cameras
in live action visual effects

Jan Kruse

A dissertation submitted to

Auckland University of Technology

in partial fulfillment of the requirements for the degree

of

Postgraduate Diploma in Communication Studies (PgDipCS)

Dedication

This dissertation is dedicated to my little brother Matthias Kruse (July 30, 1990 – March 3, 2012).

You will be dearly missed. Rest in peace!

Attestation of Authorship

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgments), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning.

Acknowledgements

I wish to thank my wife, Jeannine. Your support and motivating encouragement carried me through this amazing time. Without your help and presence this would not have been possible.

I wish to thank my supervisors Abhishek Kala and Gudrun Frommherz for their invaluable advice, support and endless patience. Thank you for stopping me to try to do too much, and thank you for pushing me further, helping me to achieve more and get a deeper understanding of this fascinating topic. Thank you for helping to keep my work in balance.

I also wish to thank James Charlton and Charles Walker along with Gudrun Frommherz for making this work possible and for opening up a pathway at University for me.

Thank you all for your help and for making this great journey possible.

Table of Contents

List of Figures and Tables

Abstract

Film production greatly depends on digital visual effects to combine live action images and computer generated elements. The introduction of *Deep Compositing* streamlines the integration workflow of rendered computer images. This new technique relies on depth information, which is the distance between any objects in the scene and the camera itself. While depth data is available as a byproduct in computer generated images, live action camera files lack this kind of information.

Only the recent arrival of a new generation of devices called RGB-D cameras, has enabled researchers and filmmakers to conduct first experiments with depth data in conjunction with live action images. The consistent acquisition of distance information and color pictures might allow for workflow improvements in live action visual effects similar to what has been demonstrated with Deep Compositing in conjunction with rendered computer images.

This research is investigating the impact of RGB-D devices on established live action visual effects workflows. RGB-D images are used for effective CG placement, and a proof of concept workflow has been implemented. Findings related to quality, resolution and range of acquired images have been documented and discussed, and suggestions for future improvements of devices and workflows have been made.

1 Introduction and Overview

1.1 Introduction

In recent years, Computer Graphics (CG) has become a key element of modern film making. Visual effects are not only used to create imaginary worlds or artificial characters, but have many applications in image correction, enhancement and reparation as well. If for instance undesirable objects like power lines or water towers in a historic movie were visible at the time of filming and could not be avoided at reasonable cost to the producers, removal of these objects would be necessary at a later stage during post-production. Another popular application are invisible extensions of set pieces which would be too expensive to build practically, aging of landscapes or cities to match contemporary settings to historic story contexts or simple digital cosmetics to change the appearance of actors to support the storytelling.

Any of these techniques have contributed to a measureable reduction in production costs and shorter timelines, which in turn has led to the increased popularity of visual effects in film and television. But the visual effects process itself has undergone some important improvements in efficiency too. Most recently access to depth information in CG rendered images has accelerated image compositing and improved workflows, which lead to higher quality output and more sophistication in the resulting images. Depth information is the distance of objects like set pieces or actors to the observing camera. Based on this distance information, placement of CG objects in the scene is simplified and more accurate including position and scale. This technique has been introduced as Deep Compositing (Hollander, 2011) and provides artists with an advanced set of tools for CG integration. However, it is based on CG rendered per pixel depth data and has not been applied to practical live action shots yet. This research aims to prove that depth data captured on set with a physical camera provides similar benefits as the CG rendered counterpart. If depth information is collected while color images are recorded, distance information for all visible points would be available at a later stage during visual effects creation.

Currently there are two common methods used to capture depth information in moving images, structured light pattern (Guan, Hassebrook, & Lau, 2003) and Time-Of-Flight (*The Making-of "House of Cards" video*, 2008) measurements. Structured light pattern projects a grid of invisible infrared laser points across the scene, analyses the returned structure of the grid and reconstructs the scene by calculating the distortion of the pattern through obstacles in frame. TOF sends pulses of light out and measures the time it takes for the light to return to the sensor (to be reflected by objects). Given the know speed of light, the distance to points in the scene is computed. Due to the high entry price point of time-of-flight devices (above $70,000) and the limitations of the technology, especially the very low resolution of 320x240 pixels for any of the available devices, this research is going to focus on structured light pattern sensors, which are considerably more affordable ($200) and offer at least four times the resolution compared to time-of-flight devices.

Using an depth information in addition to a RGB color images, this study examines the use of RGB-D cameras in live action based visual effects shots and compares the findings to established workflows.

1.2 Structure of this paper

This Exegesis is divided into two main parts. The first covers a generic overview of the topic of RGB-D capture and existing technology as well as Deep Compositing. It also discusses the existing literature, the importance of this research and introduces the Methodology used for this study.

The second part introduces the Proof of concept workflow, which is the main outcome of this research. It also provides a final conclusion and suggests topics for future research.

1.3 Definitions

RGB	Red Green Blue image data, usually refers to what is commonly known as color images
RGB-D	Red Green Blue plus depth image data, this is the output of new devices such as Microsoft Kinect or Asus Xtion, which add the distance to camera for each RGB image pixel
CG	Computer graphics, which are artificially generated images through digital painting inside the computer or more often digital rendering of modeled and animated objects
Compositing	Image integration or combination. This can be live action images captured by a camera, computer generated images or a mix of both

Deep Compositing

Refers to a new compositing technique, which utilizes per pixel depth data in CG images. The depth data (or deep data) is a byproduct of the CG rendering process.

Shot

In visual effects a shot refers to a short piece of film from cut to cut and is typically the unit used to assign visual effects tasks to artists.

OpenNI

Open Source Kinect driver and SDK (software development kit), which offers access to camera images, tilt motor and skeleton detection.

GUI

Graphical user interface, the part of a software project that relates to user interaction.

Cartesian Coordinates

In a three dimensional Cartesian coordinate system, each point is described as a triplet of values X, Y and Z. Each value represents the signed perpendicular distance from the corresponding axis of the coordinate system.

SDK

Software development kit

Fps

frames per second, the running speed of video and film, indicating the number of frames, that are projected or displayed per second. 25fps and 30fps are common for TV, whereas film usually uses 24fps in cinemas

2 Research Questions

The research is driven by the main research question:

“How do RGB-D cameras impact on established live action visual effects workflows?”

The problem set has been broken down into following sub-questions:

- How do existing visual effects workflows need to be adjusted to suit the additional depth data component?
- Do any advantages justify the additional overhead of capturing and storing depth data?
- What are the technical requirements from a workflow perspective towards RGB-D cameras?
- Is the resolution and accuracy sufficient for visual effects work?
- Is the range of the depth acquisition unit, in this case the pattern based depth measurement device sufficient for visual effects work?

3 Significance of the Research

Visual effects are constantly evolving, getting more sophisticated and technologically more complex. In order to keep sufficient space for creative freedom, the tools that support visual effects artists in their work need to be improved at a similar pace and new technologies have to be adapted in order to keep up with the demand of production companies and ultimately the consumer.

This study examines the potential of expanding a key technology that has been established in the area of CG renders into the field of live action visual effects. Deep compositing, while still being a very young concept, has proven to increase efficiency and effectiveness in visual effects production within a short period of time and has widely been adapted by software development companies, visual effects companies and artists alike (Hollander, 2011). This research is conducted under the assumption that these advantages will prove to be similarly significant for live action visual effects. The study investigates a proof of concept workflow and attempts to expose any issues of state of the art technology in order to produce suggestions for future improvements and future research as well. It lays the foundation for a prototype workflow that might be adapted by visual effects houses and artists.

Specifically, the issues around image acquisition, CG element placement and edge treatment in compositing will be examined in the light of added depth information. These areas have been identified as the most problematic by Okun and Zwerman (2010), but also the most promising in terms of potential improvements through adaption of Deep Compositing.

4 Literature and Technology review

This section reviews the existing literature and discusses workflows in visual effects as well as the RGB-D technology currently available. It also provides an overview of the device chosen for this research and investigates the advantages and issues of the device sensor. Further, some alternatives are discussed and an explanation is provided why these were not as suitable as the selected sensor in the context of this research. Finally and introduction to Deep Compositing is provided as it is the core workflow that this study expands on.

4.1 Visual effects workflows

The two main literature pieces providing an overview of existing visual effects workflows are *"The VES Handbook of Visual Effects: Industry Standard VFX Practices and Procedures"* by Okun and Zwerman (2010) and the Master's Thesis *"Creating a Workflow for Integrating Live-action and CG in Low-cost Stereoscopic Film Production*" (Kala, 2010).

Okun and Zwerman (2010) were the first authors to establish a standard for visual effects work covering a wide area from principal photography, visual effects on-set work, image preparation and processing to CG integration and post-production processes. They found a definition for all processes involved in visual effects generation and called it a "visual effects pipeline" (Okun & Zwerman, 2010). This definition is derived from Computer Science, where a serial set of processes is usually defined as a 'pipeline'. It is important to note that despite two decades of visual effects production, their work was a milestone as it looks into a much wider area than previous papers and books such as "The Art and Science of Digital Compositing" by Ron Brinkman (1999).

The key aspects relevant to this research are image acquisition or generation, which could include camera-based image capture or CG rendered images. Okun and Zwerman (2010) argue that both are found in nearly every visual effects pipeline, independent of the size of the facility or production and are therefore crucial center pieces, which have to be considered when efficiency and effectiveness of tools and resources are being improved. Both workflows are important for this study as the proposed RGB-D based workflow draws on camera-based conventions and techniques, but expands on CG rendering workflows as well. The proof of concept workflow increases the overlap between 2D camera captured images and 3D point clouds, which have so far only been produced through CG generation.

In the third chapter of their book, Okun and Zwerman (2010) emphasize on the importance of best practices applied to image and on-set data acquisition and lay a foundation for this study in that they introduce the use of LIDAR (light detection and ranging) technology, which in principle provides the depth component similar to our RGB-D capture device. While the authors acknowledge the significance of depth data acquisition, they fail to put it in context with RGB color images. Okun and Zwerman (2010) see depth data more as an addition to CG modeling and animation, but not as an integral part of the entire visual effects pipeline or at least compositing workflows.

4.2 Depth information in computer graphics

The idea to combine RGB images and depth information is based on the concept of *Deep Shadow Maps* presented by Lokovic and Veach at Siggraph in 2000. Lokovic and Veach stored additional information for the CG rendered shadows of each image, in form of a point in 3D, independent of the visibility of the corresponding RGB pixel. This leads effectively to the storage of all shadows that are cast in a CG rendered scene, even if the objects and shadows are not visible through the chosen virtual camera. While this approach creates a seemingly unnecessary overhead in file size, it nevertheless offers huge benefits in form of time savings, whenever the scene has to be re-rendered, for instance for creative reasons such as camera animation changes. In case a changed camera move reveals objects that were previously not visible, this previously rendered shadow map is used to significantly reduce the computationally costly raytraced shadow renderings (Lokovic & Veach, 2000).

Shadow Maps represent an important change in visual effects workflows as they add an overhead for the benefit of higher efficiency in the creative process of visual effect creation and the iterative process of improving every shot by re-rendering several times. This is a significant improvement as it allows to add additional information to files, without an immediate benefit, but long term savings in other related workflows within the visual effects pipeline. Lokovic and Veach (2000) lay the foundation for depth based compositing by providing a theoretical framework for RGB image data combined with additional per pixel information.

Heckenberg, Saam, Doncaster and Cooper (2010) expanded on this concept by adding depth data to the RGB information of each pixel in a CG rendered frame. Their *Deep Images* store information of every volume and object surface in a CG scene. They are composed of color, opacity and distance values and are effectively a point cloud of all objects in a scene, regardless of their actual visibility at render time (Heckenberg, Saam, Doncaster, & Cooper, 2010). *Deep images* allow access to any of the objects after rendering, which enables artists to manipulate lighting, shading or per pixel color information without raytracing again. In addition, *Deep Images* provide distance information in relation to the virtual camera, which is used by post-rendering and compositing tools to alter defocus, atmospheric integration and colorization. These processes usually require costly additional rendering, but implemented as post filters, the rendering cost is marginal.

Building on *Deep Images* as introduced by Heckenberg et al, a collection of compositing tools has been produced by The Foundry for their Nuke compositing software. The file format used for *Deep Images* is currently Dtex, introduced and implemented by Pixar as part of their industry standard rendering software Renderman. This powerful combination of CG rendering and compositing showcases the potential of deep data based workflows in context of computer generated images. This research will expand on this concept and introduce depth based compositing derived from live action images, in order to apply the same benefits to filmed footage. 'Deep data' as introduced by Hollander (2011) is an integral part of the proof of concept workflow of this study. Hollander provides an overview of tools, formats and other aspects involved in depth based compositing in his talk "*Deep Compositing in Rise of the Planet of the Apes"* (Hollander, 2011). Deep compositing is a novel approach to CG integration using depth information, co-developed by Weta Digital (a visual effects company) and The Foundry (a visual effects software development company). Hollander argues that this new technique offers some benefits over traditional CG compositing workflows. He examines a range of issues in rotoscoping, color grading and depth layering, and offers simple solutions using depth data rendered in conjunction with the CG elements that he integrates. Hollander points out, that especially complex visual effects shots, composed of several CG layers, benefit from deep compositing the most. Instead of having to consider the correct placement of each element layer by layer, this process can be automated using deep data and a custom software plugin called deep merge. CG layers had to be positioned in a certain order from the farthest to the closest to camera and extra care had to be taken not to substitute any layers by accident. Deep compositing has no such requirements. The artist simple drag and drops the layers into the shot, connects them to the deep merge nodes and the correct layering is done based on the deep data information. Hollander (2011) also addresses the issue of rotoscoping, which is normally a laborious and time consuming process (Okun & Zwerman, 2010) and demonstrates a simplified approach, which uses a single roto matte and deep data to generate hold out mattes for a whole sequence at once. While automated rotoscoping shortens CG integration in moving shots significantly, the benefits for stereoscopic 2D-to-3D conversions as discussed by Okun and Zwerman (2010) are even more relevant in context of this study, applying the same principles to live action based material. Considering that a movie usually contains more than a hundred thousand frames, the manual stereoscopic conversion process is very costly and consumes a lot of time. Automated tools based on deep data could reduce the amount of labor significantly and therefore help minimizing overall production costs of 3D converted films.

4.3 Depth sensing technology

Acquisition or filming of depth data is not a new concept and has been used for many years in visual effects. In contrast, the combination of concurrent RGB (color) image and depth image capture is fairly young and has only been made available at reasonable cost with the introduction of Microsoft Kinect and the PrimeSense technology behind it (Freedman, Shpunt, Machline, & Arieli, 2011).

At present, there are two fundamental principles used for depth data acquisition, structured light pattern and Time-Of-Flight (TOF) measurements. While Microsoft Kinect uses structured light pattern to generate the depth image, TOF offers an interesting alternative with its longer range, higher accuracy and high resolution images. The main obstacle for TOF applications in visual effects is the significantly higher entry price point. Cost of producing such sensor is high, because the unit needs to be able to distinguish between light pulse signals, which are only a few nanoseconds apart. This requires very fast processing and increases in turn the production cost of the device. Also, there is currently no company that produces a combined RGB and TOF based depth camera. All devices used in the past were custom-built combinations of cameras and depth sensors. Following Okun and Zwermans (2010) discussion of data consistency and easy handling on busy film sets, an integrated approach seems beneficial to the successful introduction of new technology for visual effects production. A further indication that standard technology is preferable over custom solutions is what Kala (2010) identifies as the difference between a custom made and a professional rig, with the latter being able to deliver a better footage quality, therefore reducing subsequent costs in post-production. Kala argues that even with careful alignment and set up of the rig, quality easily suffers and subsequently requirements in post-processing increase.

4.4 Kinect and Alternatives

Microsoft Kinect is the main capture device for this research. Kinect offers simultaneous acquisition of RGB (color) images and depth information with a resolution of 640x480 pixels per frame. According to Khoshelham (2011), the resulting point cloud has just over 300,000 individual 3D points, which are exported as Cartesian XYZ information. The Kinect sensor is based on an invention by PrimeSense Ltd (see Freedman et al., 2011), which uses a structured light pattern approach utilizing invisible infrared laser light and a dedicated camera for depth acquisition. Due to a triangulation process, the sensor is reasonably accurate within a few meters, but has limited range and a larger quadratic error at distance. The quadratic error or quadratic deviation represents an exponential increase of inaccuracy (deviation from the expected measurement) over distance, which renders measurements beyond a few meters unusable as they become too inaccurate. Khoshelham (2011) explains that the sensors original use was for interactive computer game play, but due to the low cost has become a popular device for many applications in robotics and other fields outside of computer gaming. The Kinect sensor is supported by an extensive SDK, which offers direct access to the depth information and color images.

Since early 2012, there is a new version of Kinect, called ‘Kinect for Windows’ available, but only in some countries. It seems to be fairly similar to the original Kinect (or ‘Kinect for Xbox’) in terms of hardware, but dissimilar in a few aspects. Kinect for Windows is not suitable to be used with an Xbox, but exclusively with Windows PCs, while the original Kinect (for Xbox) can be used with both Windows and Xbox. Secondly, Kinect for Windows has a near mode and is able to measure from about 40 cm in front of the sensor as opposed to 80 cm with the original Kinect (for Xbox). Finally, Kinect for Windows has a shorter USB cable, which is supposed to be more suitable to work with Windows PCs (Kolakowski, 2011).

There are a few alternatives to the selected capture device, which could be considered to capture depth information in addition to RGB color images. This section of the dissertation provides a brief overview of such alternatives, and concludes that these alternative technologies are either not low cost and therefore not necessarily viable within the usually tight budget constraints of visual effects production (Okun & Zwerman, 2010), or that they are technologically not suitable for the proposed application in film production, despite their seemingly similar intended use as game controllers or distance measurement devices.

3DV Systems developed a camera add-on called ‘ZCam’ (Iddan & Yahav, 2009), which had to be attached between the camera body and the lens, and used a Time-of-Flight (TOF) principle to measure the camera distance from the objects in the field of view. The TOF unit used pulsed near infrared light, which was captured at a rate of 60Hz and at a resolution of 320x240 pixels. This allowed for a 1-2 centimeter resolution of the depth channel, at a maximum range of about 10 meters. In 2009, just after 3DV Systems announced a gaming sensor based on the same technology, Microsoft bought the company and its assets, probably to keep competition off the market prior to the release of their own Kinect in 2010.

While the ZCam technology is interesting for visual effects, as it does not use a triangulation process to determine the distance of objects from the camera, which avoids the exponential increase of resolution errors at distance, it faces similar limitations and also had a depth resolution that was significantly lower than the Kinects 640x480 pixels. The use of near infrared light poses the same range issues as Kinect, resulting in a similarly low, but acceptable range of about 10 meters. The implications surrounding the reasonably low range will be discussed in a later section of this dissertation.

PMDTechnologies announced the market availability of their sensor PMD PhotonIC 19K-S3, which is a very small device, the size of a LED. The sensor uses the TOF principle and delivers up to 90fps (Buxbaum, 2012). While it offers low cost depth perception at a high frame rate, the device seems not to be suitable for visual effects work for the following reasons. The pixel array is only 160x120 pixels, which is too small to provide and acceptable match for film or HDTV. Secondly, the range is about 2 meters, which is too short to avoid the issues of parallax, as discussed in section 6.4.1. Therefore, this device has not been considered as part of this study.

4.5 Deep Compositing

4.5.1 Introduction to Deep Compositing

The process of using CG rendered depth data in compositing, originally developed by Weta Digital Ltd. has been named Deep Compositing and has become the quasi standard for this workflow (Foundry, 2012). It is important to note that Deep Compositing is not only a technique that requires a different workflow or approach to the compositing process, but also combines an additional set of data, which has to be specifically rendered for this purpose and a set of software tools in CG and image compositing software, which allow to utilize the new type of data. This part of the dissertation will take a look at this very young technique, examine some of the existing tools and look at the possibility to expand on this concept by using camera captured depth data in addition to the CG rendered deep data.

4.5.2 History of Deep Compositing

The idea of using depth based information in visual effects is not new. Just over a decade ago, which is a seemingly long time in the rapidly moving and overall quite young field of digital visual effects, the use of *Shadow Maps* has been promoted by Lokovic and Veach (2000) as method to improve efficiency in visual effects creation. As we have seen from Okun and Zwerman (2010), nearly every visual effects shot undergoes several iterations during the initial creation process, but also due to client review and approval processes and sometimes technical difficulties, which require a full or partial rework of the CG elements. Often very long render times are the consequence. Shadow calculations are among the highest cost items with regards to time and required resource. But being an essential part of a final photorealistic image (Brinkman, 1999), CG generated shadows pose a necessity and could not be avoided. Working under the same assumption, Lokovic and Veach derived the shadow information for all objects in the scene, which was simply a byproduct of the render software, stored this information in a separate data file and reused it in any subsequent additional rendering of the same scene. By adding additional time and disk space requirements to the initial rendering, they saved significant time for any iteration.

4.5.3 Tools and Techniques

Nuke offers a range of tools to support Deep Compositing as of version 6.3 and above. These include conversion tools to visualize deep data in 2D or 3D viewers (DeepToImage, DeepToPoints) or to create deep data (DeepFromImage, DeepFromFrames). Further, there are tools that manipulate deep data, as it cannot be treated with the existing image processing functions. These include transform, reformat, crop and expression based nodes. In principle a basic set of deep data treatment tools is provided and it is expected to be expanded with the upcoming release of Nuke 7.0.

4.5.4 File Formats

Currently, the most common file format to import deep data into Nuke is using Pixars DTex file format (Heckenberg et al., 2010). This file format is going to be replaced by OpenEXR 2.0 with its capability to carry an ArrayList, which is a suitable data structure widely used in computer science for large arrays of complex data types. These ArrayLists will be composed of individual per pixel depth samples in the form of Cartesian 3D points. This allows for a convenient integration of deep data into an established file format and guarantees data consistency between 2D image channels and 3D deep data (Kainz & Bogart, 2012).

5 Methodology and Research Design

The proposed methodology utilizes a practice-led approach as suggested by Candy (2006). While practice-led research is often overlapping and interlinking with practice-based methods (Dean & Smith, 2009), it is mainly focused on advancing the knowledge about processes and practices and to a lesser extent on the outcome or product of the practice. The main objective is to improve workflows or techniques, instead of creating an artifact. While a product may be produced during practice-led research, this product is not part of the outcome of the research. The practice is the data to be examined. It is the foundation for a better understanding of practice.

This is important to note as this research does not aim to create specific visual effects shots and produce a certain aesthetic or style. This study simply tries to create a new type of workflow in visual effects and examines whether the creation of such workflow is successful in terms of efficiency improvements and quality advancements.

The second part of this study is leaning on an experimental type of research as described by Walliman (2011). A very static setup in a controlled (indoor) environment with artificial lighting is incorporated to evaluate the technological implications of RGB-D image acquisition using low cost sensors. It compares Microsoft Kinect to higher quality DSLR cameras in order to establish in which way future improvements of the depth sensor should be made. By using an affordable consumer grade camera (Canon 60D) it will also investigate the likelihood of such advancements by drawing a parallel between digital camera technology and RGB-D sensing cameras. It is based on the Hypothesis that RGB-D sensing will improve at a similar rate as digital cameras did over the past few years, in order to project how much potential the depth technology might have in a few years.

The two main components of this study, the proof of concept workflow and the evaluation of existing RGB-D technology in the context of visual effects creation are going to be evaluated based on different measures. The proof of concept workflow will be assessed based on whether it is possible to incorporate RGB-D technology into visual effects workflows or not. It is simply a matter of identifying the issues around this idea and laying a foundation for future research, e.g. a deeper look into a consistent prototype workflow, which could be readily adapted by visual effects companies and artists. The second part of this study will test the quality of RGB-D data produced by Microsoft Kinect and hold it against existing standards as defined by Brinkman (1999). These standards include noise on the edges of objects in frame as well as pixilation as an indicator of reduced image quality and potential image resolution issues.

In order to be able to quantify noise, which poses a problem by just looking at it, the researcher is going to modify an approach normally used to detect flicker in images, called difference keying and luminance averaging. Two identical pictures of a static setting, with a locked off (static) camera will be taken. These will be subtracted from each other, pixel by pixel. This is going to result in a mainly black image, as the result of the subtraction of two nearly identical images results in zero values (black). Everything that is not perfectly identical, for instance sensor noise or exposure artifacts, will show as a difference value, slightly larger or smaller than zero as the result of the subtraction. By turning every negative values of the resulting image into positives, the difference between both originals will show as grey values.

After having extracted the difference of both nearly identical images, a reformat is applied, which scales all pixels of the difference image down to 1 pixel by 1 pixel. The values are effectively averaged and reduce into one single pixel value. This value can be quantified and shows in Nuke as a code value between 0.0 and 1.0. Comparing these 1 pixel image values of different sensors, including the Canon 60D, Kinect color and Kinect depth sensor, gives an indication of the true sensor noise and other quality degrading artifacts.

Figure 1 (*/figures/01_script.jpg*) shows the corresponding script in Nuke. The difference of both read nodes is calculated and reformatted to 1 pixel.

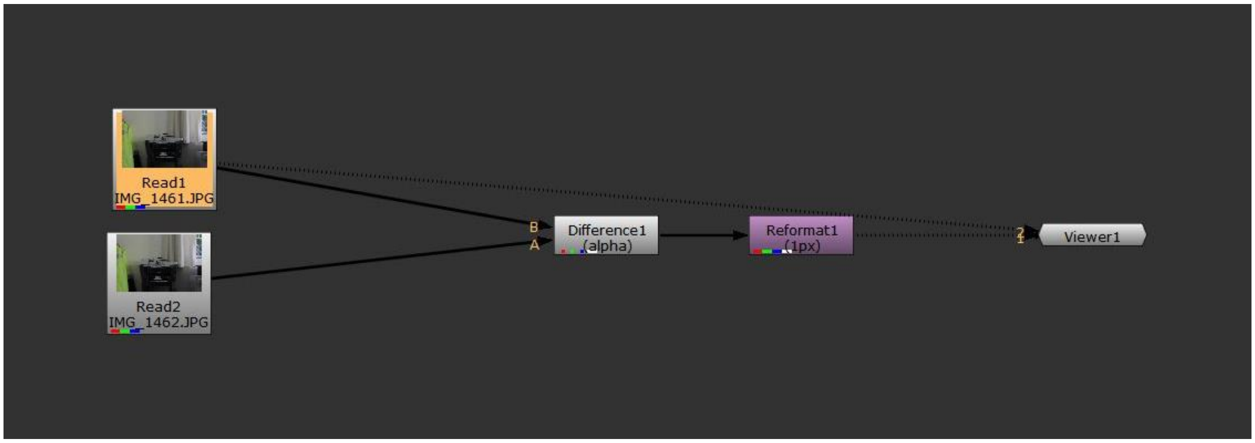


Figure 1 – Nuke difference script

Finally, this research will draw on the experience of the researcher, who has more than 15 years of experience in the visual effects industry and has worked on many award winning projects, but also helped to establish some core technologies of contemporary visual effects pipelines. This experience is reflected in the ability to provide quality control to visual effects projects, which is a standard requirement of production companies and therefore serves as a measurement for the success of RGB-D integration into current visual effects production environments.

6 Proof of concept workflow

6.1 RGB-D cameras in visual effects workflows

The release of low cost dense RGB-D sensors such as Microsoft Kinect in late 2010 and Asus Xtion Pro a few months later, has led to a wide interest of using the technology in different fields from robotics, computer graphics to medical applications. The sensor is based on a PrimeSense module (Freedman et al., 2011) and utilizes infrared structured light pattern acquisition, which has limitations in range and precision (Khoshelham, 2011), but offers unmatched affordability and relative robustness.

At the moment there are no cameras with interchangeable lens mounts available that integrate depth acquisition into the image capture process. Another known limitation is the resolution of the cameras, regarding the RGB part of the image as well as the Depth information. Finally, the noise to signal ratio, that is the amount of artifacts from one captured frame to another (noise) compared to the actual image information (signal), is relatively high compared to digital video cameras. All three factors contribute to fair image quality, which is expected for low cost devices. As with many other young camera systems, for instance digital photo cameras, which have been improved significantly since their introduction in the 1990s, it seems reasonable to assume that in the near future, RGB-D cameras with higher resolution and interchangeable lenses will be available to the consumer market.

Therefore the Microsoft Kinect has been selected as the RGB-D capture device for the proof of concept workflow under the assumption that certain parameters will be improved by hardware manufacturers, given enough time.

6.2 Image acquisition

This section discusses the hardware requirements towards capture software, a selection of available software tools to capture RGB-D images, as well as the development of custom capture software.

6.2.1 Hardware properties

The RGB-D camera used in this proof of concept workflow, Microsoft Kinect, has a resolution of 640x480 pixels for the RGB components as well as the Depth image, running at 30 frames per second (fps). The device features a USB 2.0 connection and is powered by a separate power supply. It is possible to use batteries (12V) to power Kinect, but only with a custom made wire harness. Mobility is therefore limited.

6.2.2 OpenEXR 2.0

As for the software, there is no standardized software interface or file format for RGB-D data available yet. An attempt to close this gap is the new OpenEXR 2.0 file format. OpenEXR is an image file format originally introduced by Industrial, Light & Magic (ILM) in 2000, which has become a standard in visual effects software in recent years. ILM has released OpenEXR as an open source C++ library in early 2003 (Crabtree, 2003). It has been actively developed by individuals and visual effects facilities, and undergone multiple iterations and improvements to facilitate the needs of different parts of the visual effects industry. Most large scale software companies have incorporated OpenEXR in their applications, including Adobe After Effects, The Foundry Nuke and Autodesk Maya. The file format features up to 64 channels, which can be used for RGB information, alpha channels, depth channels and many other pixel based image data. OpenEXR is capable of storing 8bit, 10bit or 16bit (half) and 32bit floating point color information per pixel. It also offers image compression algorithms, most of which are lossless and able to achieve up to 2:1 compression ratios.

The latest OpenEXR 2.0 update, which is currently in beta, has a number of added features , with deep data handling being the most significant addition from a RGB-D workflow perspective. Deep data has been implemented as a list data type, which has an arbitrary length at each pixel location. This differs from multichannel images, which have a fixed length of data per pixel (Kainz & Bogart, 2012). Deep data storage allows consistent handling of RGB color information as well as per pixel depth data at each location. Effectively, storage of variable sized point clouds along with color values in one file format has been established and standardized. In future applications, this is going to allow image capture software to store RGB-D data in a file format, which then can be read by existing visual effects software without any additional conversion. With its release, OpenEXR 2.0 is also going to enable camera manufacturers to integrate RGB-D storage in their devices, to streamline the image acquisition and conversion process even further.

6.2.3 Color bit depth

Another important aspect is bit depth or bit resolution. Different image formats store color information as separate RGB channels with a specific bit depth per channel. Many traditional image formats such as TIF or DPX use a maximum of 8bit (a maximum of 256 values) or 10bit (a maximum of 1024 values) per color channel. This leads to unwanted effects like Banding, which can be described as visible steps within a color gradient as shown in Figure 2 (*/figures/02_banding.jpg*), visible in the dark part of the left image). But for Depth images (the D-channel in RGB-D images) this limitation is even more significant. Kinect provides a resolution of 13bit for the depth channel, which equals 8192 different depth values. A conversion into a 10bit DPX file format would incur a significant loss of data and reduce the quality of the depth channel to 1/8th of its original resolution. Therefore a higher bit depth of at least 13bit is desirable. OpenEXR accounts for this and therefore qualifies as a suitable format from the quality stand point as well.

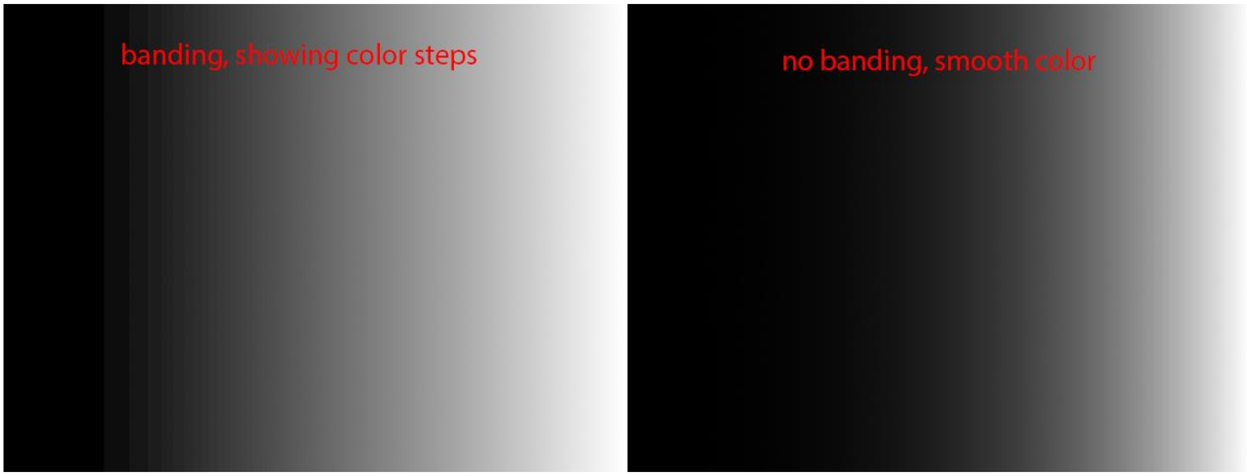
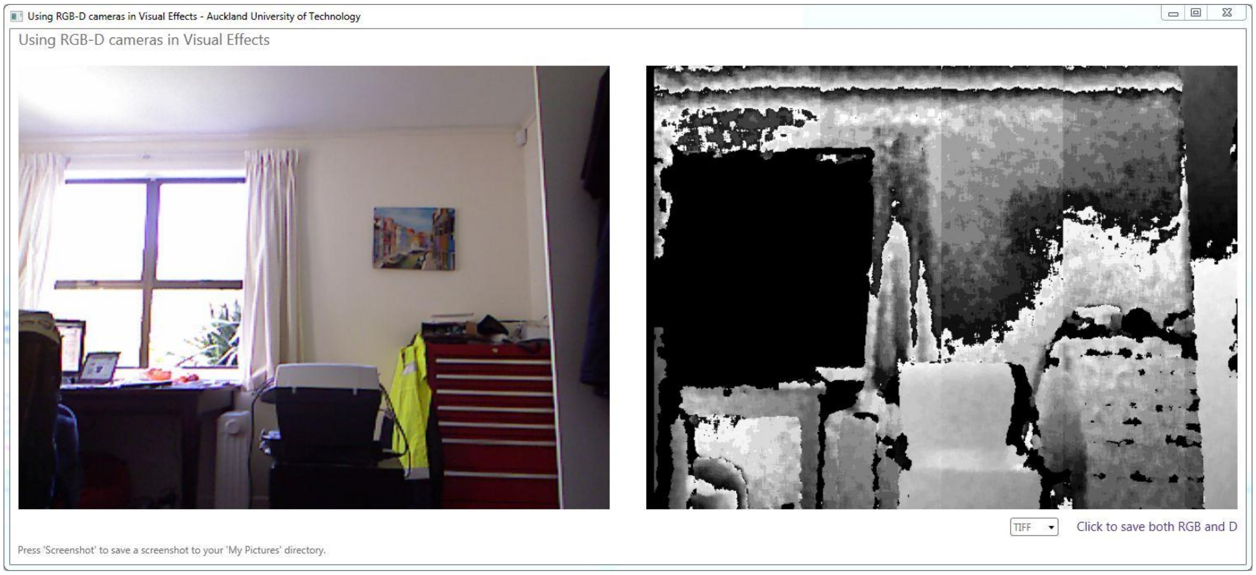


Figure 2 - Color banding

6.2.4 *Custom capture tool development*

For this research, a custom software tool has been created, which is able to display RGB and depth images concurrently (Figure 3, */figures/03_sdk.jpg*). It also serves to save images to Tiff files, as OpenEXR 2.0 is not yet available to the public. The foundation for the custom RGB-D capture tool used in this research is the Microsoft Kinect Software Development Kit (SDK) (Kean et al., 2011). The SDK provides a range of libraries dealing with image capture, skeleton tracking and voice recognition. In this case the image capture has been utilized to capture RGB color images as well as a separate stream of synchronized depth images. The depth images are represented as gray scale images, with white being the closest to camera and black being far distant.


Using RGB-D cameras in Visual Effects - Auckland University of Technology
Using RGB-D cameras in Visual Effects
TIFF
Click to save both RGB and D

Press 'Screenshot' to save a screenshot to your 'My Pictures' directory.

Figure 3 - Software tool to capture consistent RGB and D images

The software is written in C# and based on Microsoft .NET 4.0 using Windows Presentation Foundation (WPF) for the GUI elements. An event based approach has been selected to ensure that both image streams from RGB and depth sensors are captured at exactly the same moment in time. This is a crucial requirement for further experiments, which rely on consistent RGB-D data. The two image streams are displayed in two GUI pictures and saved to disk into two separate image files using TIFF format. Microsoft SDK does not support OpenEXR at all and therefore TIFF was selected to allow for uncompressed image storage. A future OpenEXR integration might be possible with the upcomping OpenEXR 2.0 file format, which has a C# wrapper available. This software tool could then serve as a foundation for future prototype workflows as it ensures full control over source code.

6.2.5 Brekel Kinect

A second alternative for the image acquisition process, Brekel Kinect in conjunction with the Kinect sensor has been successfully tested (Brekelmans, 2012). This software is based on the open source OpenNI framework by PrimeSense and offers similar capability as the Microsoft SDK.

The Brekel Kinect toolkit has the option to export point clouds, RGB color images as TIFF and Depth images as OpenEXR. It is possible to capture an image stream at 30fps or single images. The resolution of the resulting TIFF and OpenEXR images is 640x480 pixel. This toolkit offers an easy way to capture images quickly, but is limited to this functionality only and does not allow any modification of the capture process or final images. It is therefore not suitable to capture RGB-D images in a single file nor is there a way to adjust the image quality or resolution. For future research, a custom made tool, which allows RGB-D capture into the new OpenEXR 2.0 format is desirable, in order to ensure consistency of all four image channels including deep data.

The last capture tool that has been evaluated is the RGBD toolkit from Carnegie Mellon University (George, 2012). This tool is still under development (pre-release 0031) and has proven to be not very stable. It has to be considered experimental. The concept uses just the depth sensor of Kinect and captures the RGB portion of the image with a DSLR camera. This offers superior RGB image quality combined with the relatively low resolution of the Kinect depth capture. The images have to be aligned, which is done through a tool that is part of this project. The RGBD toolkit webpage provides some video information about the aesthetics and ideas behind this project, as well as a range of rapid prototyping blueprints, which can be used to make a mount for Kinect and a DSLR camera. While the concept is interesting, it still seems to be very clumsy to separate both RGB and D components by sourcing them from different devices. This requires not only careful optical calibration, but also a solid mechanical structure to ensure consistency throughout the shoot. And the calibration process has to be repeated, should the cameras become misaligned due to physical force. The concept could be an interesting option as soon as the image resolution of depth camera can be increased, but at this stage the discrepancy between DSLR and Kinect is very high (about 1:8 pixel ratio). Therefore, while the aesthetic effects are intriguing, the applicability of the depth separation for visual effects shots seems questionable.

6.2.6 Kinect USB connection

Microsoft Kinect (for Xbox) has proven to be problematic with some laptops during this research. The USB connection did not work reliably and showed some erratic connection/disconnection behavior, which was initially confusing. With the release of 'Kinect for Windows' and the accompanying feature list, pointing out the shortened USB cable "to ensure reliability across a broad range of computers" (Kolakowski, 2011) the connection issue was solved. The original Kinect (for Xbox), which is the selected device for this study, does not guarantee a working USB connection. Therefore, it may be beneficial to use 'Kinect for Windows' for any future work. This improved device was not available in New Zealand during the time this research was conducted.

6.3 Image conversion

The image acquisition using the Microsoft Kinect RGB-D camera provides two different outputs as a result of the two cameras (RGB and Depth). Based on the Microsoft Kinect SDK both output streams could be captured separately and treated as two different images during the visual effects integration. But this would effectively double the number of files, which might seem insignificant looking at a single test shot, but would have notable effects in a large scale visual effects company, dealing with thousands of shots concurrently and several hundred terabytes of data. The second main issue with two different files would be consistency. Keeping each RGB color image consistent with the related depth data, while using separate files, might create an unnecessary challenge for visual effects artists and network administrators. If only a single error occurred, which put RGB and D image out of synchronization, a whole movie sequence could be affected and restoring consistency could be time consuming and costly in economic terms. Combining both streams into one single file for each movie frame helps to keep data consistent and avoids larger consequences, should a file get deleted by human error or computer system failure. The single file would have to be restored, but any subsequent movie frames would keep their synchronization with the interrelated depth information.

While this is not possible for the proof of concept workflow due to the time constraints of this dissertation and the proposed release date of the new OpenEXR 2.0 file format, it would be desirable to implement this in a prototype workflow, subject to future research.

For this proof of concept, the two image streams are kept separately, while extra care is taken to ensure consistency between color and depth image. The RGB color image is converted into an OpenEXR (1.0 standard) file, while the depth data is handled as a XYZ point cloud in a text file. This process has been used in the past in image compositing, pre *Deep Compositing* or more specifically deep data times according to Hollander (2011). The XYZ point cloud file contains three columns of values, one for each component of a three dimensional Cartesian coordinate system, x, y and z. These are read into Nuke using a PositionToPoints node (as shown in Figure 4, */figures/04_script.jpg*), which converts XYZ point data into depth information. This depth information can then be visualized as a gradient black and white image in the 2D viewer, where the luminance values (or black and white color values) reflect the distance from camera as provided by Kinect. Alternatively, the output of the PositionToPoints node can be visualized in the 3D viewer using a PointCloud node (see Figure 5, */figures/05_3dview.jpg*).

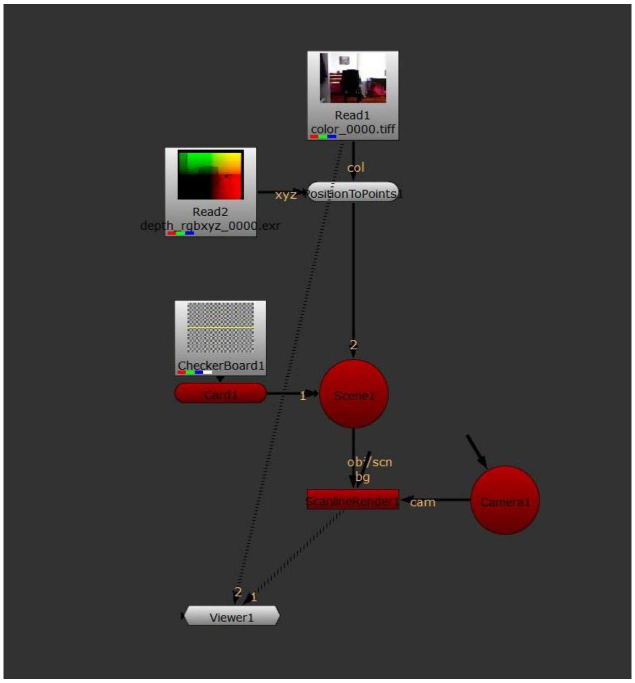


Figure 4 - PositionToPoints Nuke script

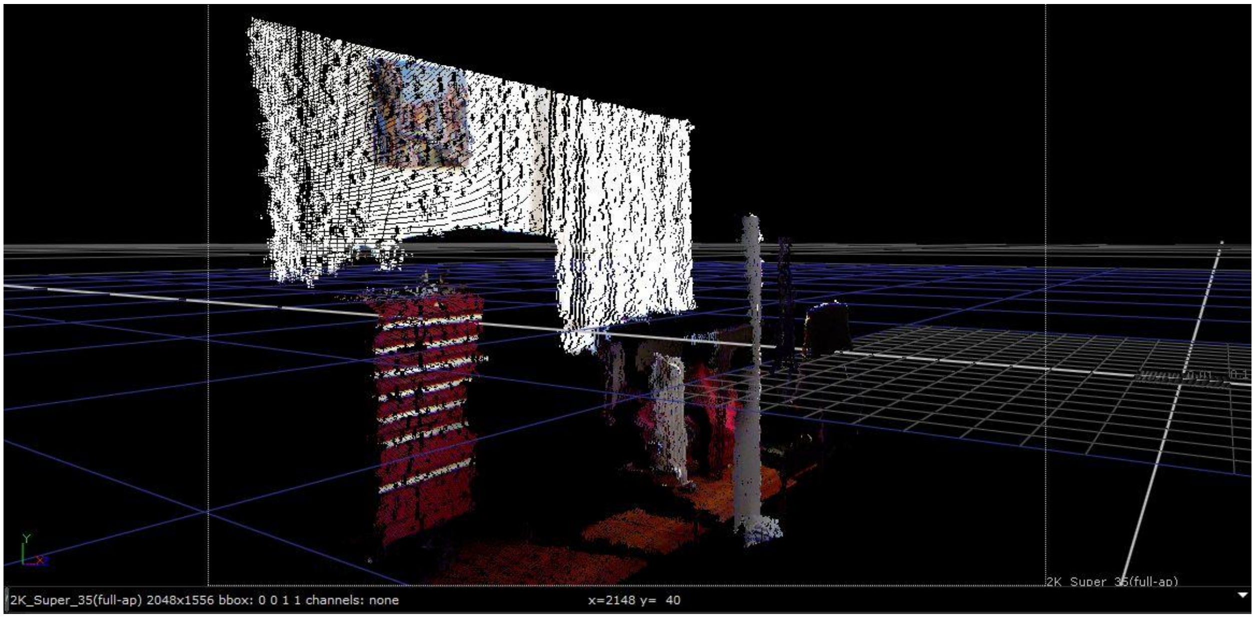


Figure 5 - PositionToPoints visualised as point cloud (colorized using the corresponding RGB image)

6.4 CG placement

6.4.1 Overview of CG placement

The correct placement of CG elements in visual effects shots is one of the most important yet challenging tasks. Particularly, with a moving camera instead of a static ("locked off") position, this can be time consuming and problematic as it requires careful preparation on set and several manual laborious steps during post production. To reconstruct the physical camera movement, visual feature tracking is required, no matter which specific technique is employed for the actual camera reconstruction process. Tracking visual features, that is following specific points or shapes from frame to frame to retrieve the camera move, has inherent inaccuracies as it is based on the resolution of the digital frames and therefore always only as good as the digitized image, which is often interpolated and anti-aliased. Furthermore, the algorithms used to calculate (or reconstruct) the camera position are limited by the data types used in computers. As insignificant the rounding error during computation might seem, it can easily reach a few centimeters difference, depending on the scale of the physical set, the velocity of the physical camera and the distance of any objects in frame. Parallax of a few centimeters between CG positioned objects and the original live action set is easily observed even by untrained eyes and often described as "jittering CG". Parallax is the spatial shifting of objects against each other, when the camera is moving from side to side. It describes the effect when foreground objects seem to move or shift faster than more distant objects. For visual effects work it simply means that anything that is closer to camera needs more attention to reduce jitter, color imperfections and increase spatial placement accuracy. Foreground objects need to be treated more accurately than distant objects. Often, when backgrounds need to be replaced, the distant part of an image is covered by one static matte painting (a single image), whereas the imminent foreground needs to be rendered CG, made of individual objects. The foreground objects also require a higher level of detail in terms of modeling, texturing and rendering as they are closer to the camera and therefore any imperfections are more visible. This means that anything in the foreground requires a higher level of attention than anything in the distant background. Concluding, CG placement poses a few challenges, which are mainly introduced by the technique used to reconstruct the conditions of the live action set and the camera including its motion at the time of image acquisition and the resulting parallax.

With the addition of depth data using RGB-D cameras, a significant portion of these issues can be neglected. The known distance of any objects in frame allow for a pixel or sub-pixel (if the depth channel uses an oversampled higher resolution then the RGB sensor) accurate placement of CG objects. The camera reconstruction would save the intermediate step of visual feature recognition to reconstruct the scene, by using the available depth data. This allows for an instant result, immediate visual feedback and is not depending on the resolution used in post-production, but entirely on the quality and resolution of the depth measurement sensor.

Furthermore, during the creative process of placing objects and integrating them in terms of color, shadows, motion blur and other quality defining factors, the artist would be able to use the provided depth data visualized as a point cloud to simply select specific points in the scene and snap objects to that position. A lot of convenient tools would be made possible by providing accurate depth information.

6.4.2 Using depth data for CG placement

For this research, depth data has successfully been imported into Nuke as point cloud data. The chosen file format as XYZ point data is a temporary workaround until deep data is officially supported by OpenEXR 2.0 and subsequently part of Nuke’s supported import formats. The XYZ file contains Cartesian coordinates for each point in form of comma separated values in three columns for each component (x, y and z coordinates). Nuke is able to visualize the point cloud in the 3D view as a guide for the artist, but offers a limited set of tools at this stage. It is possible to snap a 3D card to point, which is done by selecting a point and placing a rendered element (CG or live action) at that point in space. This "card" serves as the transformation layer for the CG or live action content. Snap-to-point allows the effective placement of CG elements at specific points in a very quick and efficient manner, but requires the element to be scaled and rotated to match the scene. The card basically just snaps into position, but does not necessarily face the camera or have the right proportions. This is up to the artist to decide.

In the context of *Deep Compositing*, some more advanced concepts have been demonstrated (Hollander, 2011), which go beyond the simple positioning in relation to 3D points. If three or more points are selected, *Deep Compositing* nodes allow to create an image plane that intersects all these points and therefore to establish the correct or intended rotation of the 3D card (or layer). Currently, the right scale has to be manually chosen, but this could be automated in future software updates. If at least two points are selected, the scale of the layer could be based on these. But there is also the possibility to take any two points of the live action set, which have a known distance to each other and use them to scale the element. Even if these two points are not at the same distance to camera, the right scale can still be derived based on the depth point cloud and the known distances to camera. This concept could even be more simplified, if the depth camera would be calibrated at the start of shooting the live action elements. This would give not only relative distances of the points in the captured point cloud, but allow for absolute accuracy. In case of the Kinect device the absolute distance of the points in the point cloud is only approximately known, as there is no facility to calibrate the depth sensor. While this could be done with custom calibration software, the development of such software tool is beyond the scope of this research and subject to future investigation. Khoshelham (2011) has provided the foundation for such work by examining the sensor, but did not use a custom software tool at the time.

6.4.3 Evaluation of CG placement

It is important to note, that the placement of the elements based on depth data captured by an RGB-D camera is only as good as the quality of the data. Diminishing factors are sensor noise, lack of resolution and possibly increased triangulation errors at larger distances.

Sensor noise has in fact proven to be an issue with the Kinect sensor. For the size, resolution and intended purpose it delivers an acceptable quality and shows no unexpected artifacts on the edges of captured objects. This confirms what Khoshelham (2011) suggested in his conference paper. The noise of the depth sensor is not significantly different from the RGB sensor, probably as both have a similar resolution of the pixel array (the sensor surface).

Looking at a specific example, the captured shot shows a static scene with a static camera to reduce any artifact introduced by motion blur. Without any movement the noise that is visible on the edges of sharp objects from frame to frame will indicate the level of noise introduced by the sensor. This can be visualized by subtracting both frames from each other, which results in only the difference between them. The shot taken for this research shows a lot of noise compared to established digital film cameras like RED or ARRI. Comparing it to Canon DSLR cameras, Kinect shows significant quality issues which is not unexpected. The DSLR image has very clean edges and very low inter-frame noise. Given the price difference, which should be reflected in component quality of the devices, it can be assumed that a high quality and slightly more costly RGB-D chip would provide a similar performance as current DSLR cameras.

Image resolution, the density and number of pixels per frame is an important factor in deriving high quality images as well. Most visual effects shots are delivered (and shot) in 2k film resolution (2048x1536 pixels) or recently even in 4k (4096x3072) in case of ‘The Hobbit’ as confirmed by Peter Jackson and his production company (Giardina, 2012). The 2k and 4k refer to the number of horizontal pixels that compose a film frame and are a common acronym in the visual effects industry. The Kinect RGB and D sensor would be rated to be a VGA (or 0.6k resolution) frame, which is significantly smaller than the common standard. It is therefore expected that a direct comparison between CG based depth information, which is rendered at the 2k industry standard, and a Kinect captured 0.6k frame does not lead to a consistent quality. At this point in time, the VGA resolution of Kinect is state of the art, but a new, updated version with a near 2k resolution (Full HD or 1920x1080 pixels) has been announced by Microsoft. It will be available in 2013 and confirms that the there is no technological obstacle hindering the development of a device with sufficient resolution for visual effects work. Testing this new Kinect device and implementing a visual effects workflow with it will be subject to future research.

The final factor of importance for quality is the resolution in depth or along the z axis. This has been discussed by Khoshelham (2011) as well and he argues that the error is expected, but somewhat significant, with a quadratic increase over distance. This might be of importance for other applications such as robotic navigation or interactive gameplay, but in case of visual effects work is has a surprisingly small effect due to parallax. Although Microsoft’s Kinect depth sensor has some inherent issues not being able to measure as accurately at distance, than right in front of the camera, it is reasonable to assume that this weakness is not going to have a strong negative impact on image quality.

7 Evaluation of Image Quality

This section provides an overview of the two chosen methods to investigate the RGB-D workflow and evaluates the quality of the resulting images. The two parts show a scenario based on a real world indoor set (the study, where part of this research commenced) and secondly an experimental setting in a studio, where controlled lighting, known distances and objects with known properties help to create a controlled environment.

7.1 Real world set piece

Following the chosen methodology, several different image pairs have been taken with different cameras sensors, including Canon 60D, Kinect RGB and depth sensors. The Canon 60D represents a high quality, low noise sensor and serves as the reference in this evaluation process. The Kinect RGB color sensor and the depth sensor are the devices being tested against the reference. Given the low image resolution of Kinect compared to the Canon camera, it has been decided to reduce the significantly larger image of the Canon to 640x480 to match the Kinect. This is necessary as there is no high quality sensor available with such low resolution to match Kinect. While this is not an ideal way to achieve a comparison, it is still possible to make a statement about the Kinects quality performance. The chosen process even degrades the Canon sensor a bit more than Kinects measurement, as the noise of the larger Canon sensor surface is quantified against the smaller sensor surface of Kinect. The results show a clear picture of what the main issue of Kinect is. The RGB sensor is very noisy and the depth measurement is even more problematic.

Figure 6 (*/figures/06_canondiff.jpg*) shows the difference image of the Canon 60D sensor. Aside from the superimposed technical overlays, there are a few grey lines visible on the upper right hand side, plus a few very faint lines across the whole image. These lines represent the difference between both original static images. They are effectively the sensor noise and some minor exposure artifacts.

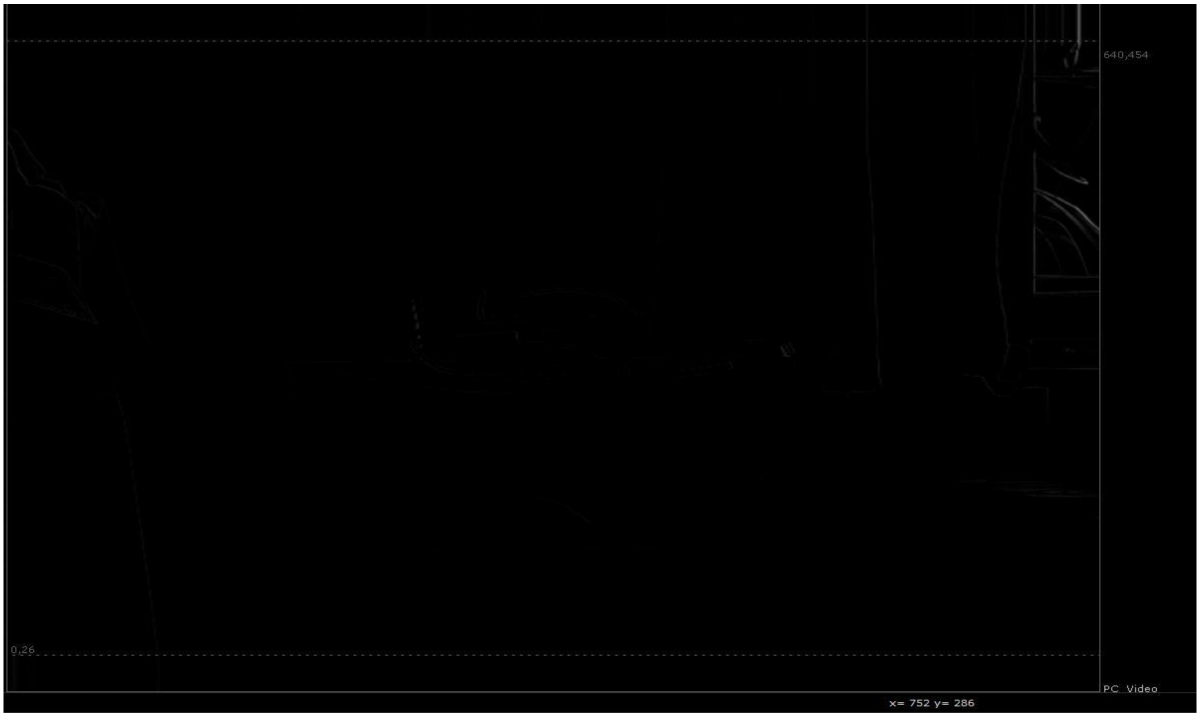


Figure 6 – difference image of Canon 60D capture

Figure 7 (*/figures/07_canononepixel.*jpg) is the result of the reformatting process to 1 pixel. The image (or pixel) is effectively black, which means that there is very little noise present in the two original

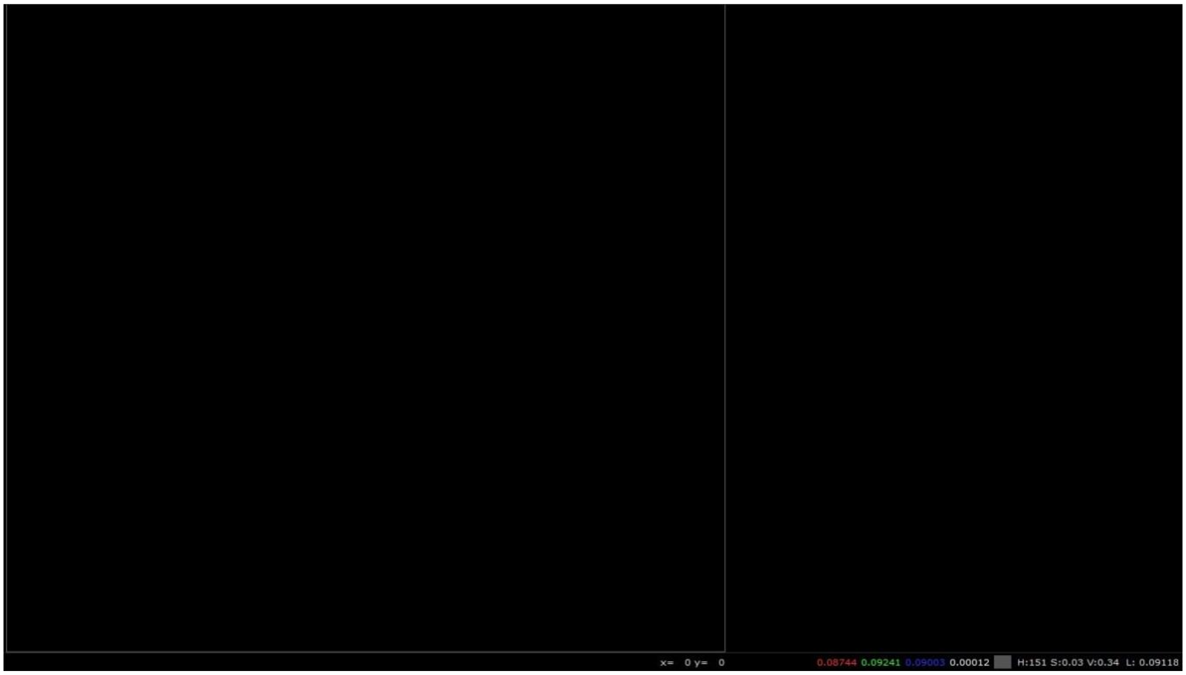

x= 0 y= 0
0.08744 0.09241 0.09003 0.00012 H:151 S:0.03 V:0.34 L: 0.09118

Figure 7 – One pixel image quantifying the noise

images taken with the Canon 60D. The value is 0.00012 which is a code value near black (or 0.0). This sets the context for the comparison with the Microsoft Kinect sensor. Looking at the color sensor of Microsoft's Kinect, we evaluate the noise in the same way. The difference shows notably more artifacts as visible in Figure 8 (*/figures/08_rgbdiff.jpg*).

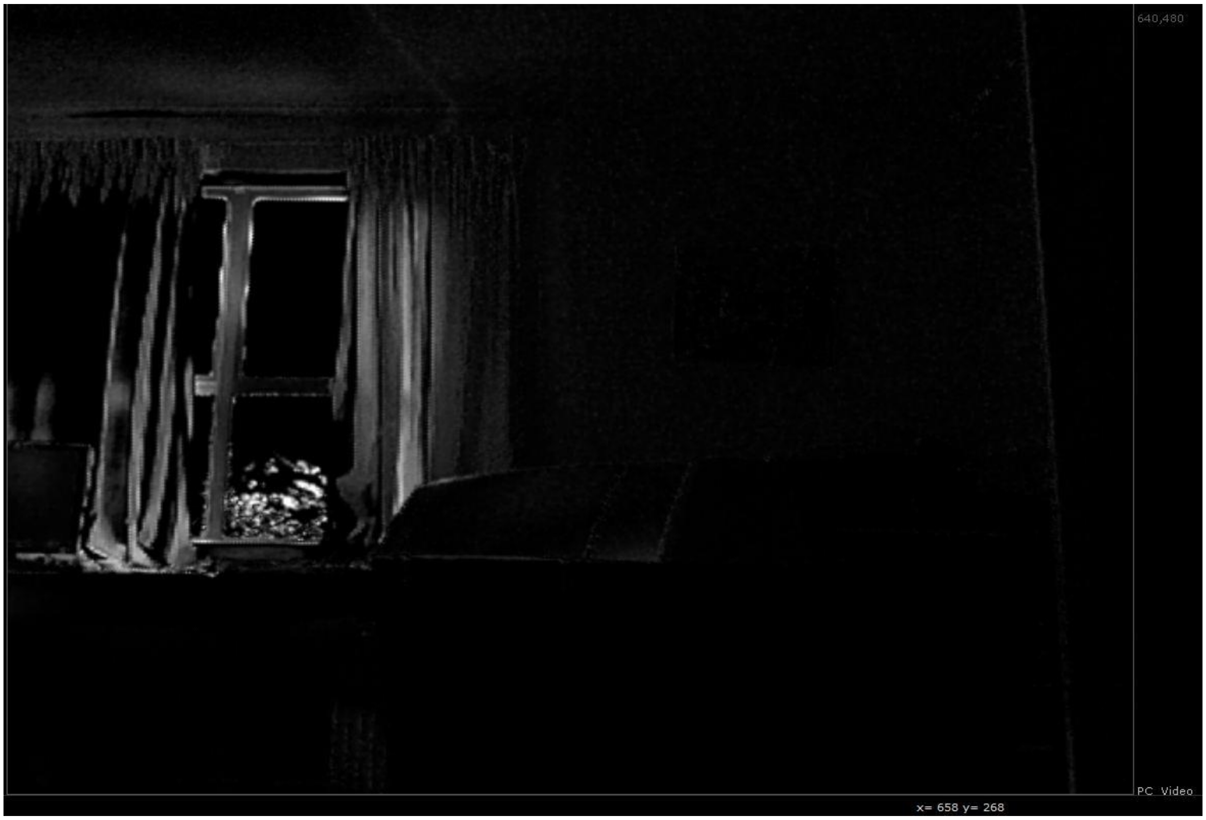


Figure 8 - Difference image from Kinect RGB color sensor

The difference exposes an interesting problem though. The brighter parts of the image surrounding the window have the biggest difference, whereas the wall and most of the interior of the room show very low noise. This leads to the conclusion that the exposure control, which is most affected by the bright daylight through the window is not very stable and therefore changes from frame to frame. Effectively, this is a slight flicker in the brighter parts of the image. The outside sky appears black, simply because this is at fully saturated maximum exposure or overexposed and has a value of 1.0 (maximum white). Accordingly, the difference outside is 0.0. But any values just under full white (or 1.0) show a huge variation between frames. The relatively cheap device is probably not manufactured to the highest standards and does not deliver a very stable exposure.

While the exposure artifacts are quite significant, the noise in the RGB image seems to be very low, as most of the frame does not show any strong noise artifacts but is almost black (just over 0.0).

Finally, holding the depth sensor image against both previous color images, a strong noise pattern is visible and confirms the visual impression of the original depth images. They are very noisy, not only on the edges of objects, but distributed across the whole image, which leads to wide areas of difference with fairly high values (closer to white, than black) as shown in Figure 9 (*/figures/09_depthdiff.jpg*).

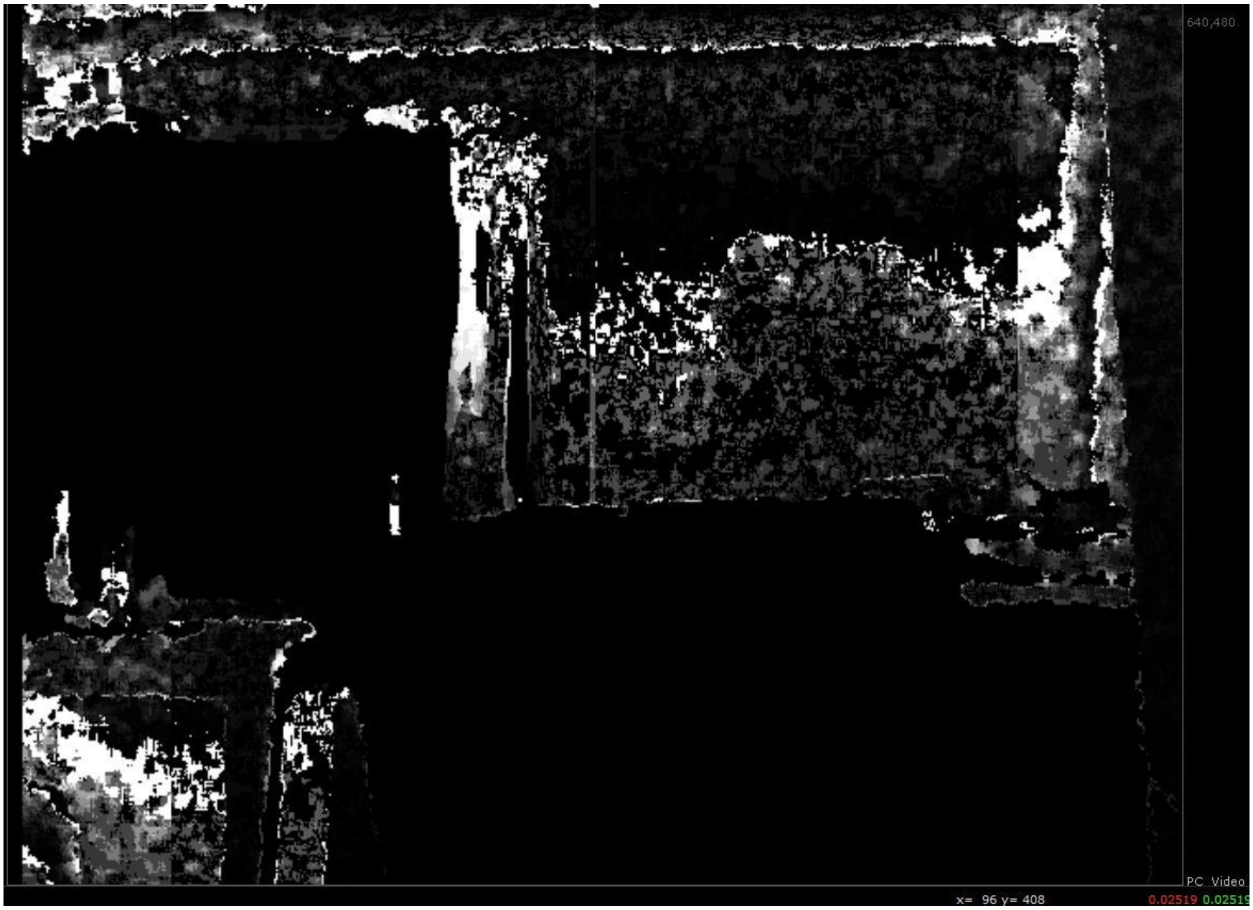


Figure 9 - Depth difference image showing a lot of noise

The noise problems around the edges of objects are even stronger as shown by zooming into two detail sections in Figure 10 (*/figures/10a_noisedetail.jpg*), which shows brighter (whiter) edges of these two objects in the shot. This is also notable in Figure 9 as a white edge around most of the noise patterns.

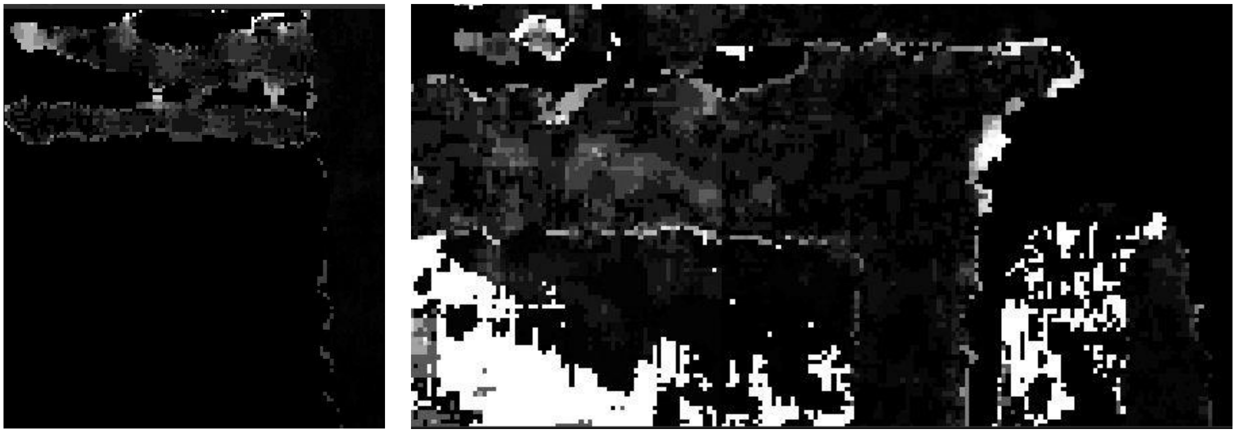
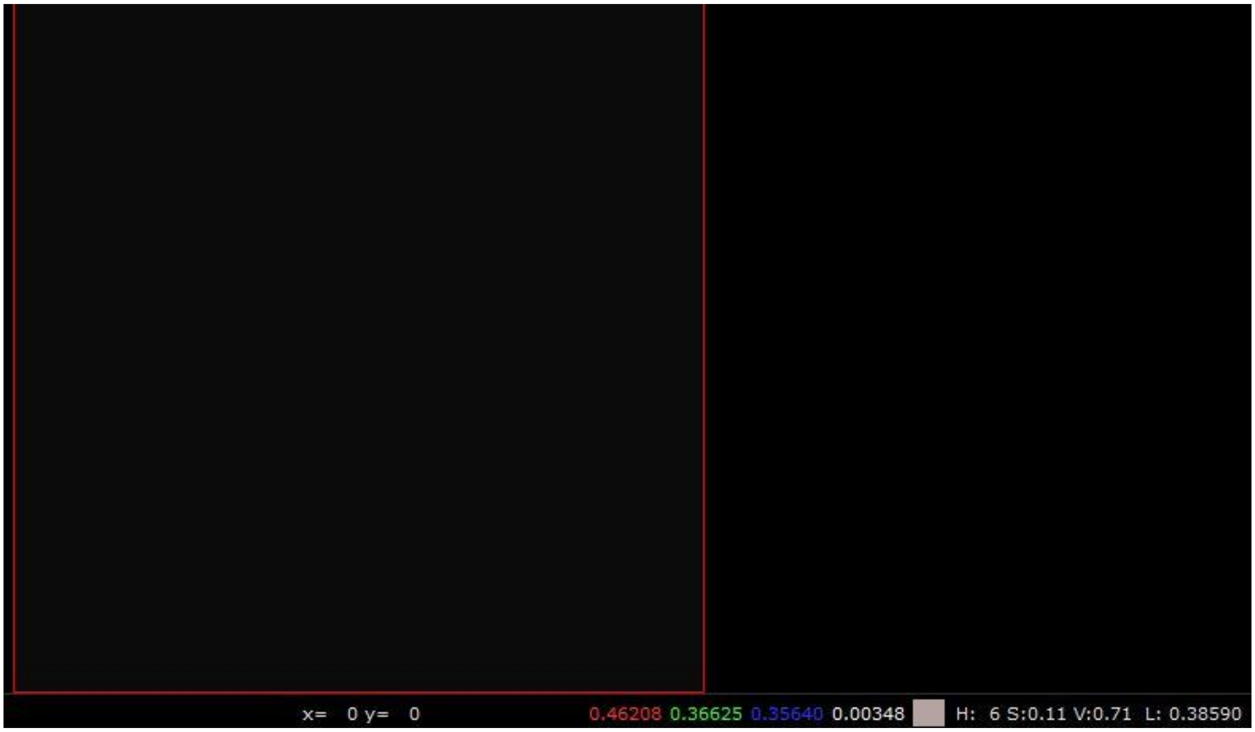


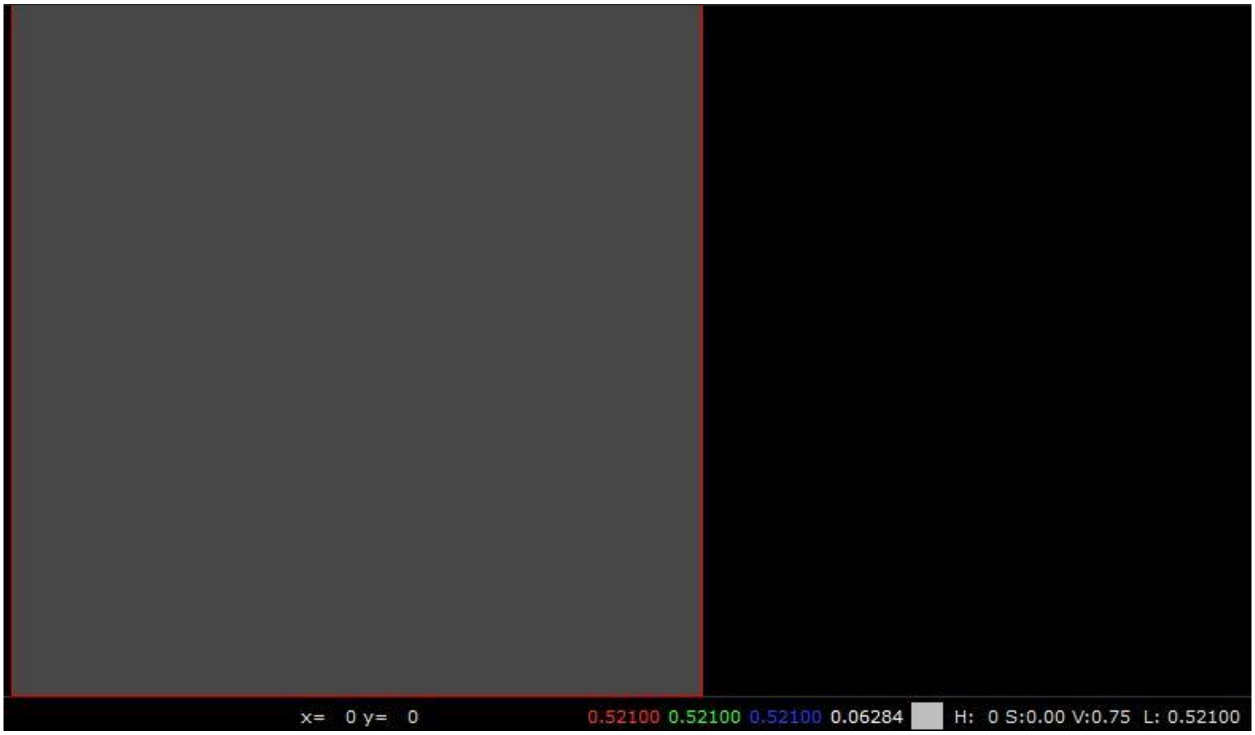
Figure 10 - Detailed noise edges

To quantify this rather significant error, the same technique as before has been applied and the image was reduced to 1 pixel. This is shown in Figure 11 (*/figures/11_rgbonepixel.*jpg) and Figure 12 (*/figures/12_depthonepixel.*jpg) for the Kinect RGB and depth sensor respectively.



```
x=   0 y=   0          0.46208 0.36625 0.35640 0.00348    H:  6 S:0.11 V:0.71  L: 0.38590
```

Figure 11 - One pixel image of Kinect RGB color sensor



```
x=   0 y=   0          0.52100 0.52100 0.52100 0.06284    H:  0 S:0.00 V:0.75  L: 0.52100
```

Figure 12 - One pixel image of Kinect depth sensor

A comparison of all three sensors (Table 1) shows an increase of overall image noise, with the Canon sensor being of the highest quality with just under 0.2% of noise that the Kinect depth sensor introduces into the image. In other words, the Kinect depth sensor produces 500 times more noise than the reference sensor in the Canon 60D. This is a very significant amount and not acceptable for visual effects work as it leads to problems separating elements and leads to jitter of integrated elements (Brinkman, 1999).

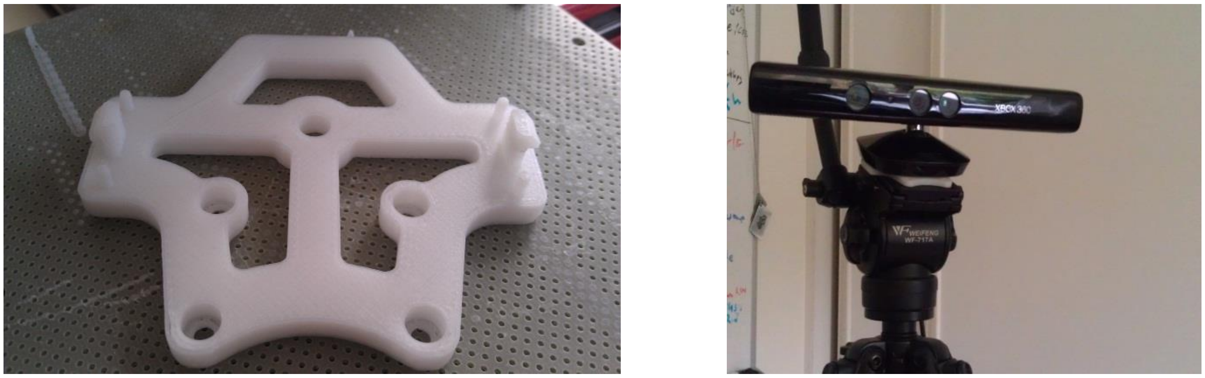
Table 1 – Sensor noise comparison

Device	Value	Percent
Canon 60D	0.00012	0.19%
Kinect RGB sensor	0.00348	5%
Kinect Depth sensor	0.06284	100%

7.2 Lab experiment in a controlled environment

The second setup utilizes a controlled lighting situation with KinoFlo™ lights, producing a stable and repeatable setting. These are normally used for greenscreen work and provide the highest standard in studio lighting.

The sensor was placed on a tripod using a custom made mount (Figures 13, */figures/13a_mount.jpg, /figures/13b_mount.jpg*), which was drawn in Solidworks and then 3D printed on a rapid prototyping machine. It is based on a part from Henkka (2011), which was modified to suit a single screw tripod quick release plate.



Figures 13a and 13b - Custom made Kinect mount detail and on tripod

Further, a custom laser cut object was used as an object to be captured. These laser cut panels have a 30x30mm square hole, which is used to evaluate the level of detail the depth sensor is able to capture. According to Khoshelham (2011), the minimum resolution of the structured light pattern is just under 30mm and the hole in the panels will allow to verify of this claim.

The three panels used in this experiment are set 1.0m apart along the centerline of the sensors viewing direction. The first panel is 1.2m away from the sensor to ensure it is well within the requirement of 0.8m minimum distance to the camera (Kolakowski, 2011). A *Laser Disto*, which is a TOF based laser range finder, is used to ensure accurate placement of each object. Figure 14 (*/figures/14_studiosetup*.jpg) shows the setup in the studio including the Kinect device on a tripod and the three panels (two green and one red) with their respective distances.

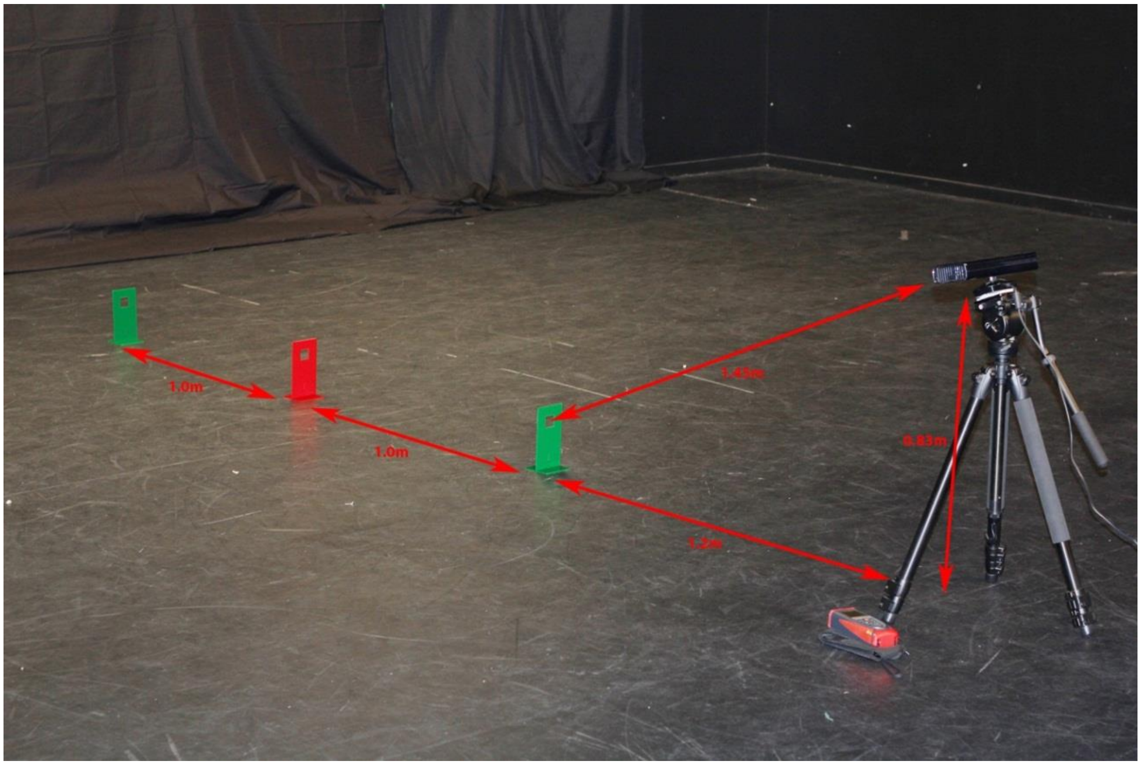


Figure 14 - Studio setup showing distances between objects and camera

In this experiment, three different image capture tools were used, similar to the real world setting as discussed in section 7.1. All three tools showed the same behavior and performed the capture task flawlessly. Kinects performance was rather disappointing in that the depth resolution is very poor and the noise is very strong, clearly visible in large parts of the depth images (Figure 15, */figures/15_studiocapture.jpg*).

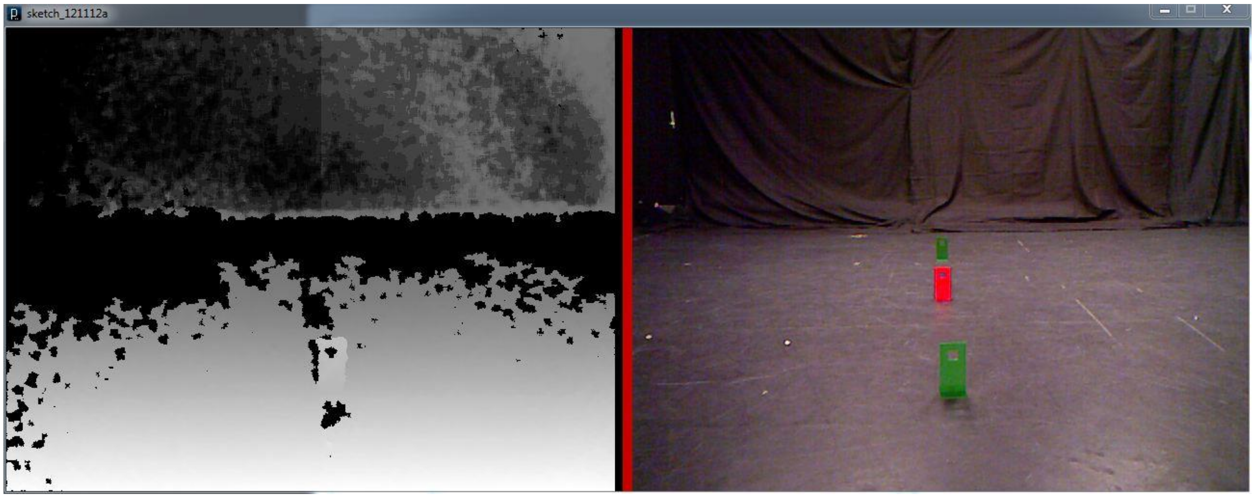


Figure 15 - Image capture showing major issues in the depth channel

Part of the floor has not been successfully captured and the edges around the backdrop and on the floor and very patchy and show a strong degradation towards the edge between backdrop and ground. The depth sensor exposes resolution issues with the ground plane at a flat angle, whereas the backdrop, which is almost perpendicular to the scanning sensor, is captured with acceptable quality, given that it is on the edge of the specified range of Kinect.

Unfortunately, the same poor performance regarding the range of the sensor affects the panels on the ground as well. The first green panel is shown with expected detail, even depicting the square hole at the top, but both more distant objects are not visible in the depth channel at all. Combining both RGB and D channels into a colored 3D model (Figure 16, */figures/16_studio3d.jpg*) shows that there is in fact some information of the red panel present, but not enough to identify the square hole in the panel. Figure 16 also shows the missing ground and the strong noise pattern around the edges very clearly.

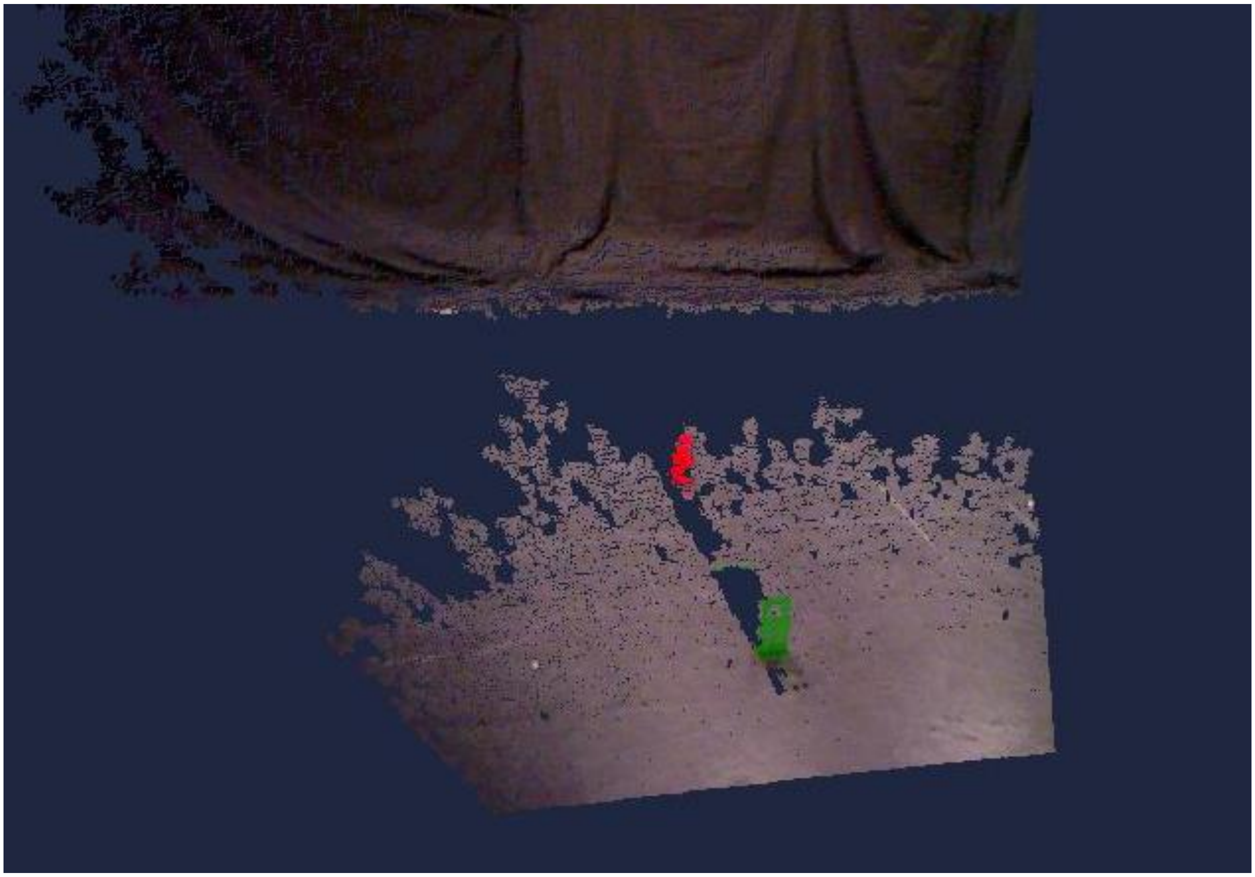


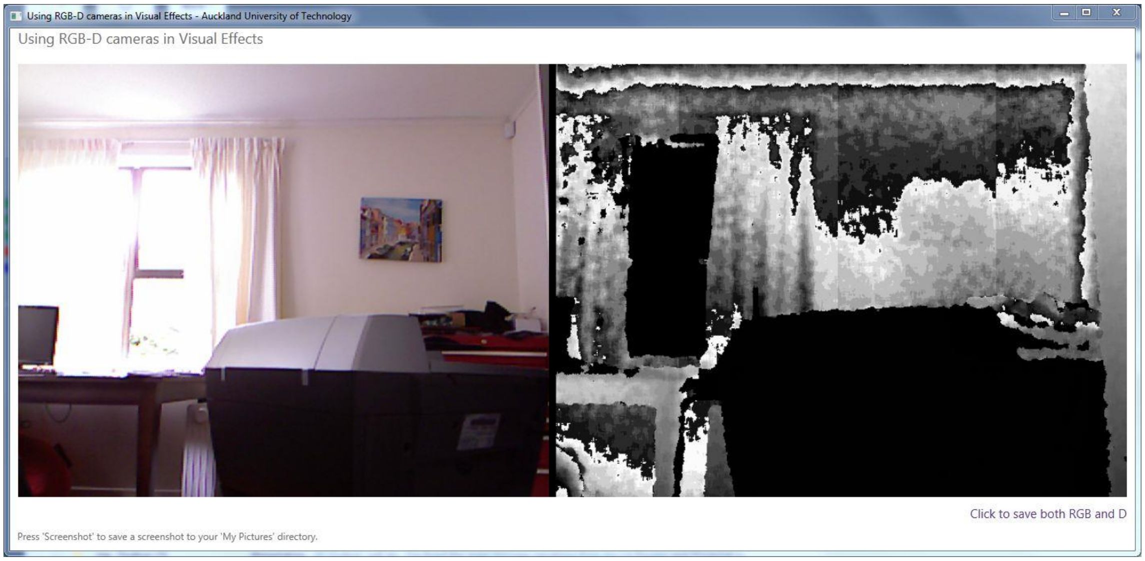
Figure 16 - Combined RGB and D channels

Concluding, it is apparent that the Kinect sensor barely meets its own specification, when examined under controlled conditions in a predefined environment. The usable range judging by the red and green panels seems to end after about 3 meters. While the backdrop in at around 5 meters distance is visible to a certain extent, the smaller objects like our test panels get consumed by noise and are not identifiable beyond 2.5m – 3 m. This range and resolution is not sufficient for visual effects work, as phenomena like parallax would affect any compositing attempt and render it unusable.

7.3 Additional findings

In addition to the above mentioned quality defining factors, the research has uncovered a few unexpected issues. If a highly reflective surface close to camera is being captured, for instance a piece of polished stainless steel, the depth sensor gets erroneous readings due to the structure of the light pattern being disturbed. The resulting chaotic values between black and white for the depth image, randomly scattered across the frame represent failed readings of some depth pixels. Removing the object instantly restored correct measurement behavior.

Another peculiar property of the depth unit produces black pixels for objects that are too close to be correctly read from the depth camera. Black pixels normally represent infinitely far distant points in space, as they don't return any infrared light from the depth measurement unit. While it is possible for a human user to distinguish between far distance and near view, the computer interprets both as infinite. This could lead to problems only if care is not taken to prevent anything in frame that is closer than about 80cm from the sensor. An example is shown in Figure 17 (*/figures/17_captureproblem.jpg*) where the printer in the foreground shows black (infinite distance) in the depth image, despite the closer proximity to the sensor.



Using RGB-D cameras in Visual Effects - Auckland University of Technology

Using RGB-D cameras in Visual Effects

Click to save both RGB and D

Press 'Screenshot' to save a screenshot to your 'My Pictures' directory.

Figure 17 – Printer very close to camera shows black (infinite distance) in the depth image

8 Conclusion

This section summarizes and discusses the findings and limitations of this study. Some suggestions for future work to address implications of this research are given in section 8.344.

8.1 Summary of Findings

This research has examined the feasibility and viability of using RGB-D cameras for visual effects production. The main outcome of this study is a proof of concept workflow which has been setup utilizing the Microsoft Kinect RGB-D sensor, the Microsoft Kinect SDK and Nuke compositing software. It has been demonstrated that a workflow using depth data is possible and that some of the expected positive effects such as depth placement are available to the compositing artist even at this very early stage. The ability to select points inside Nuke that have a defined depth relationship and known distance to camera makes positioning of elements much easier and straight forward. Many of the *Deep Compositing* tools available for CG renders will be usable for live action footage, which promises huge efficiency and quality improvements accordingly. With the future introduction of OpenEXR 2.0 there will be a common file format to carry depth data through the visual effects pipeline consistently. Future updates of Nuke (version 7.0) will support this file format and use deep data in one convenient format without any custom tools. Assuming that RGB-D chips will experience a similar rise in quality and resolution like DSLR cameras and digital film cameras like RED or ARRI did, another set of applications will be possible. Most importantly, per pixel depth data will allow to separate foreground and background information similar to green/blue screen keying. This is going to save a lot of time and effort on set, but also guarantee high quality results in post-production, which is not subject to human skill, budget and effort anymore, but simply a matter of using the right workflow and a suitable technology.

But this study has also revealed that the noise of the sensor is too strong and the range is too poor for visual effects compositing. While the basic principle has been confirmed working, a practical application of RBG-D based compositing with the current quality of the depth sensor seems unlikely. The edge quality is not sufficient to allow matting foreground objects, nor is the resolution adequate to separate fine details. The biggest advantage at the current state is clearly the ability to place CG elements by use of depth data. The resolution and noise is in the way of more advanced techniques like automated stereoscopic conversions.

The research questions have been addressed and some conclusive answers have been found. It seems advisable to slightly adjust the workflow, in that the new OpenEXR file format will help to maintain consistency of RGB and D data. The additional overhead of depth information is justified as the benefits of CG placement, overall time-savings and potentially semi-automated matting operations outweigh the extra storage requirements. Suggestions for requirements towards RGB-D cameras have been made and discussed. Potentially desirable features of future cameras have been included in 8.3.1. And finally, range, resolution and noise of the Kinect sensor have been examined as part of a proof of concept workflow and the findings have been discussed as well.

Overall it has been successfully demonstrated that the use of depth data in live action visual effects offers significant advantages, while only adding a moderate amount of additional data. The most important factor for its future success will be the availability of suitable devices. The purpose of this study is to confirm the potential of RGB-D data in live action visual effects work and this has been successfully achieved.

While the choice of Microsoft Kinect has been a step in the right direction towards depth based visual effects work, as the sensor offers a reliable solution at a very low price point, it is not a visual effects camera and lacks in resolution, interchangeable lenses and other features which digital film cameras and DSLR have. For its intended purpose as a gaming device it is very successful though and it allows for first experiments with RGB-D cameras in visual effects workflows.

8.2 Limitations

This study is one of the first research projects examining the possibility of using RGB-D cameras in live action visual effects workflows. The findings have to be treated with caution though, as the results are based on an experimental technology in a very specific environment, using a certain set of tools. Therefore the conclusions may only be generalized to a certain extend. Prerequisites may vary in professional visual effects production environments. The setup of shots in live action visual effects production is potentially more complex and the study may not be representative of all possible scenarios.

The capture tool is still very limited in function and does not consider OpenEXR based workflows yet. This might introduce additional problems, which are not foreseeable without further examination. In this context, it is also unknown, whether the established Dtex file format will be replaced by the upcoming OpenEXR 2.0 or if they will coexists in future visual effects pipelines.

This study is based on the original Kinect (for Xbox) sensor, as it is the only Kinect version available in New Zealand at the moment. The Kinect for Windows sensor has not been evaluated and it is unknown whether it offers the same image quality as the original Kinect for Xbox.

Finally, the *Deep Compositing* tools in Nuke have not been tested due to the lack of a suitable exchange format between RGB-D camera and compositing software. Some assumptions have been made based on rendered CG images and *Deep Compositing*, but advantages and issues of live action footage using *Deep Compositing* tools remain unknown.

8.3 Future Work

8.3.1 Wish-list for a future RGB-D visual effects camera

This research has shown that Microsoft Kinect is not a perfectly suitable RGB-D camera for visual effects purposes. But it offers an insight into the technology and allows establishing a list of desirable features for a RGB-D visual effects camera. The following collection of features is based on the findings of this research and might not cover additional important factors, but could serve as a starting point for future developments.

- Interchangeable lenses being able to creatively frame and compose shots.
- HD or 2k resolution to match established standards
- 12bit color depth per RGB channel to match existing quality standards
- 16bit resolution for the depth channel to gain fine resolution in z-depth
- Battery power to be independent of computers and cables
- Wireless data transfer or recording on media cards

8.3.2 OpenEXR 2.0

OpenEXR 2.0 is currently in Beta-testing and will be released by the end of this year (Kainz & Bogart, 2012). It will be supported by Nuke 7.0 which is due at the same time. One of the main components for a prospective prototype workflow, which could be readily adapted by visual effects facilities and individual visual effects artists will be a capture tool that exports this new file format. It will be subject to future research to build this connection and also determine, whether the export of the depth information into the deep channel of OpenEXR 2.0 or the export into a newly defined channel is more desirable. While it seems logical to use the deep data channel, there may be implications that only practice will be able to identify. Additionally, there is a gap in knowledge with regards to the established Dtex file format. A detailed comparison of Dtex and OpenEXR 2.0 might be valuable to understand storage of RGB-D images better.

8.3.3 Deep Compositing with live action RGB-D images

One of the driving factors behind the original idea for this study is the use of live action RGB-D images with Nukes *Deep Compositing* tools. While the evaluation of image quality in section 7 has identified that the sensors will require a higher resolution and less noise, the idea to do live action based *Deep Compositing* still stands. Future research into further integration of RGB-D cameras and *Deep Compositing* software should probably commence after the OpenEXR 2.0 and Nuke 7.0 release to be able to use the advanced features. But in general it seems to be valuable to produce a *proof of concept* workflow with Kinect, Nuke 7.0 and OpenEXR 2.0 as a first step towards integrated live action *Deep Compositing*. In the long run, a *prototype workflow* which could be adapted by smaller studios and visual effects artists, could be worth to be investigated.

8.3.4 RGBD Toolkit

The RGBD Toolkit is a collection of software tools, which propose a different solution to the integrated RGB-D capture approach. It proposes to use a DSLR camera for the RGB information and a Microsoft Kinect for the depth measurement. While this seems to be an interesting alternative to both existing technologies based on structured light pattern and time-of-flight, it is beyond the scope of this research to establish whether the RGBD Toolkit is a viable solution for this approach. The Toolkit provides tools for RGB-D capture and processing, which cover a range of utilities and applications from camera calibration to video editing. The RGBD Toolkit is still in Beta and only available on OSX at this point in time.

9 References

Brekelmans, J. (2012, June 29). Brekel Kinect. *www.brekel.com*. blog. Retrieved from http://www.brekel.com/?page_id=155

Brinkman, R. (1999). *The Art and Science of Digital Compositing*. San Diego: Morgan Kaufmann.

Buxbaum, S. (2012, January 10). PMDTechnologies ramps 3D CMOS imager for mass market usage. *PMDTechnologies ramps 3D CMOS imager for mass market usage*. Retrieved October 13, 2012, from http://www.pressebox.de/pressemeldungen/pmdtechnologies-gmbh/boxid/474030

Candy, L. (2006, November). Practice Based Research Guide. Retrieved May 26, 2012, from http://www.scribd.com/doc/72480138/Practice-Based-Research-Guide

Crabtree, S. (2003). ILM aims to unify digital imaging. *Hollywood Reporter*, *377*, 8.

Dean, R. T., & Smith, H. (2009). *Practice-led Research, Research-led Practice in the Creative Arts* (1st ed.). Edinburgh: Edinburgh University Press.

Foundry, T. (2012, October 4). The Foundry & Weta Digital go DEEP. *News*. Retrieved from http://www.thefoundry.co.uk/articles/2011/05/24/250/the-foundry-weta-digital-go-deep/

Freedman, B., Shpunt, A., Machline, M., & Arieli, Y. (2011, November 1). Depth mapping using projected patterns. Retrieved from http://www.google.com/patents/US20080240502

George, J. (2012). *RGBD toolkit*. Pittsburg, USA: Carnegie Mellon University. Retrieved from https://github.com/downloads/obviousjim/RGBDToolkit/RGBD_preRelease_0031_osx.zip

Giardina, C. (2012, April 28). Peter Jackson Responds to "Hobbit" Footage Critics, Explains 48-Frames Strategy. *The Hollywood Reporter*. Retrieved October 20, 2012, from http://www.hollywoodreporter.com/news/peter-jackson-the-hobbit-cinemacon-317755

Guan, C., Hassebrook, L., & Lau, D. (2003). Composite structured light pattern for three-dimensional video. *Optics Express*, *11*(5), 406–417. doi:10.1364/OE.11.000406

Heckenberg, D., Saam, J., Doncaster, C., & Cooper, C. (2010). Deep Compositing (p. 2). Presented at the Siggraph. Retrieved from http://www.johannessaam.com/deepImage.pdf

Henkka. (2011). *Kinect mount to camera tripod*. Retrieved from http://www.thingiverse.com/thing:5601

Hollander, R. (2011). *Deep Compositing in Rise of the Planet of the Apes*. Twentieth Century Fox Film Corporation. Retrieved from http://vimeo.com/37310443

Iddan, G. J., & Yahav, G. (2009, December 6). 3D IMAGING IN THE STUDIO. 3DV Systems. Retrieved from http://web.archive.org/web/20090612071500/http://www.3dvsystems.com/technology/3D%20Imaging%20in%20the%20studio.pdf

Kainz, F., & Bogart, R. (2012, August 5). Technical Introduction to OpenEXR. Industrial, Light & Magic. Retrieved from https://github.com/openexr/openexr/blob/master/OpenEXR/doc/TechnicalIntroduction_2.0.pdf

Kala, A. (2010, November). *Creating a Workflow for Integrating Live-action and CG in Low-cost Stereoscopic Film Production*. AUT, Auckland, New Zealand.

Kean, S., Hall, J. C., Perry, P., Kean, S., Hall, J. C., & Perry, P. (2011). Microsoft's Kinect SDK. *Meet the Kinect* (pp. 151–173). Apress. Retrieved from http://www.springerlink.com.ezproxy.aut.ac.nz/content/h6u47k47x5505073/abstract/

Khoshelham, K. (2011). Accuracy analysis of Kinect Depth Data. *ISPRS workshop laser scanning 2011* (p. 6). Presented at the ISPRS, Calgary, Canada: ISPRS. Retrieved from http://www.isprs.org/proceedings/XXXVIII/5-W12/Papers/ls2011_submission_40.pdf

Kolakowski, N. (2011, November 23). Microsoft Preps Kinect for Windows Hardware 744086. *eWeek*.

Lokovic, T., & Veach, E. (2000). Deep Shadow Maps. Presented at the Siggraph, Los Angeles: Addison-Wesley. Retrieved from http://graphics.stanford.edu/papers/deepshadows/

Okun, J. A., & Zwerman, S. (Eds.). (2010). *The VES Handbook of Visual Effects: Industry Standard VFX Practices and Procedures* (1st ed.). Burlington, MA: Focal Press.

The Making-of "House of Cards" video. (2008). Retrieved from http://www.youtube.com/watch?v=cyQoTGdQywY&feature=youtube_gdata_player

Walliman, N. S. R. (2011). *Research Methods the Basics*. The basics. London ; New York: Routledge. Retrieved from http://www.AUT.eblib.com.au/patron/FullRecord.aspx?p=667795