Depth Map Upscaling for Three-Dimensional Television

The Edge-Weighted Optimization Concept

Sebastian Schwarz

Mittuniversitetet
MID SWEDEN UNIVERSITY

Department of Information Technology and Media
Mid Sweden University

Licentiate Thesis No. 92
Sundsvall, Sweden
2012

Mittuniversitetet
Informationsteknologi och medier
SE-851 70 Sundsvall
SWEDEN

ISBN 978-91-87103-41-4
ISNN 1652-8948

Akademisk avhandling som med tillstånd av Mittuniversitetet framlägges till offentlig granskning för avläggande av teknologie licentiatexamen torsdagen den 22 november 2012 i O111, Mittuniversitetet, Holmgatan 10, Sundsvall.

Tryck: Tryckeriet Mittuniversitetet

To Chris. Farewell, old friend!
And to Papitchaya. Welcome to my life.

Abstract

With the recent comeback of three-dimensional (3D) movies to the cinemas, there have been increasing efforts to spread the commercial success of 3D to new markets. The possibility of a 3D experience at home, such as three-dimensional television (3DTV), has generated a great deal of interest within the research and standardization community.

A central issue for 3DTV is the creation and representation of 3D content. Scene depth information plays a crucial role in all parts of the distribution chain from content capture via transmission to the actual 3D display. This depth information is transmitted in the form of depth maps and is accompanied by corresponding video frames, i.e. for Depth Image Based Rendering (DIBR) view synthesis. Nonetheless, scenarios do exist for which the original spatial resolutions of depth maps and video frames do not match, e.g. sensor driven depth capture or asymmetric 3D video coding. This resolution discrepancy is a problem, since DIBR requires accordance between the video frame and depth map. A considerable amount of research has been conducted into ways to match low-resolution depth maps to high resolution video frames. Many proposed solutions utilize corresponding texture information in the upscaling process, however they mostly fail to review this information for validity.

In the strive for better 3DTV quality, this thesis presents the Edge-Weighted Optimization Concept (EWOC), a novel texture-guided depth upscaling application that addresses the lack of information validation. EWOC uses edge information from video frames as guidance in the depth upscaling process and, additionally, confirms this information based on the original low resolution depth. Over the course of four publications, EWOC is applied in 3D content creation and distribution. Various guidance sources, such as different color spaces or texture pre-processing, are investigated. An alternative depth compression scheme, based on depth map upscaling, is proposed and extensions for increased visual quality and computational performance are presented in this thesis. EWOC was evaluated and compared with competing approaches, with the main focus was consistently on the visual quality of rendered 3D views. The results show an increase in both objective and subjective visual quality to state-of-the-art depth map upscaling methods. This quality gain motivates the choice of EWOC in applications affected by low resolution depth.

In the end, EWOC can improve 3D content generation and distribution, enhancing the 3D experience to boost the commercial success of 3DTV.

Acknowledgements

The work presented in this thesis is a compilation of over two years of my life. During this time there have been countless encounters that helped me settle in the scientific community and contributed to the ideas presented in this thesis. There have also been a great many of new friends who helped me to settle in this new country. To all of you, I'm thankful.

My special thanks to my supervisors Mårten Sjöström and Roger Olsson for their outstanding work introducing me to the Swedish culture filled with surströmming and fikapaus. And even more for their excellent support in all research matters. Their trust and confidence in me and my work allowed me to achieve this thesis. Thanks also to all members of our Realistic 3D research group for their inspiring exchanges and a pleasant working environment. Especially to Sylvain Tourancheau, Ulf Jennehag, and again Mårten and Roger for going the extra mile and providing tailored courses for our individual studies. Furthermore I thank all members of the Division of Information and Communication Systems at Mid Sweden University, in particular Annika Berggren, for organizational support in countless occasions and Jamie Walters for showing me around the place.

Moreover I want to thank all other parties involved in my work. Especially our project partners at Ericsson AB for their many ideas and suggestions within the field of depth map upscaling, and the people at Fotonic and Optronic for their technical advice with time-of-flight cameras.

To my friends, thank you so much for dragging me away from work every now and then. To my family, thanks for all the love and support over the years (decades!), I could have never done this work without you. And finally, thank you Papitchaya for giving all of this a reason.

Table of Contents

List of Papers

This thesis is based on the following papers, herein referred to by their Roman numerals:

I Sebastian Schwarz, Mårten Sjöström, and Roger Olsson. Depth map upscaling through edge weighted optimization. In *2012 Electronic Imaging - 3D Image Processing and Applications, IS&T/SPIE, Burlingame, CA, USA*, 2012.

II Sebastian Schwarz, Roger Olsson, Mårten Sjöström, and Sylvain Tourancheau. Adaptive depth filtering for HEVC 3D video coding. In *2012 Picture Coding Symposium, IEEE/EURASIP, Kraków, Poland*, 2012.

III Sebastian Schwarz, Mårten Sjöström, and Roger Olsson. Improved edge detection for EWOC depth upscaling. In *2012 International Conference on Systems, Signals and Processing, IEEE/EURASIP, Vienna, Austria*, 2012.

IV Sebastian Schwarz, Mårten Sjöström, and Roger Olsson. Incremental depth upscaling using an edge weighted optimization concept. In *Proceedings of 3DTV-Conference, IEEE/EURASIP/MPEG-IF, Zürich, Switzerland*, 2012.

Terminology

Abbreviations and Acronyms

2D	Two-Dimensional
3D	Three-Dimensional
3DTV	Three-Dimensional Television
3DV	Three-Dimensional Video
ATTEST	Advanced Three-Dimensional Television System Technologies (EU project)
AVC	Advanced Video Coding
CDR	Clean Decoding Refresh
CfP	Call for Proposal
CGI	Computer Generated Imagery
CI	Confidence Interval
DES	Depth Enhanced Stereo
DIBR	Depth Image Based Rendering
EU	European Union
EWOC	Edge-Weighted Optimization Concept
FPA	Focal Plane Array
GOP	Group Of Pictures
HD	High Definition
HDR	High Dynamic Range
HEVC	High Efficiency Video Coding
HSV	Hue, Saturation, and Value of brightness
HVS	Human Visual System
IST	Information Society Technologies
JBU	Joint Bilateral Upscaling
kbps	kilobits per second
MOS	Mean Opinion Score
MPEG	Motion Picture Expert Group
MRF	Markov Random Field
MSE	Mean Square Error
MVD	Multiview Video plus Depth

MVC	Multiview Video Coding
MVP	MPEG-2 Multiview Profile
NAFDU	Noise Aware Filter for Depth Upsampling
PSNR	Peak Signal-to-Noise Ratio
PWAS	Pixel Weighted Average Strategy
QP	Quantization Parameter
RGB	Red, Green, Blue
SSIM	Structural Similarity
ToF	Time-of-Flight
TV	Television
VSRS	View Synthesis Reference Software
V+D	Video plus Depth

Mathematical Notation

Capital bold symbols represent matrices, lower case bold values denote vectors. Individual matrix and vector items are represented with the corresponding non-bold letter and indices.

An example for image frame **I**, represented by a matrix of size $X \times Y$:

$$\mathbf{I} = \begin{pmatrix} I(1,1) & I(2,1) & ... & I(X,1) \\ I(1,2) & I(2,2) & ... & I(X,2) \\ ... & ... & ... & ... \\ I(1,Y) & I(2,Y) & ... & I(X,Y) \end{pmatrix}$$

$\mathbf{1}$	identity matrix
$\mathbf{a}$	point in 3D space
$\mathbf{b}$	point on 2D plane
$\mathbf{b}'$	point on 2D plane, corresponding to **b**
B	stereo baseline
$\mathbf{C}$	camera central point
$\mathbf{D}$	depth map
$\mathbf{D}_{adapt}$	adaptively filtered depth map
$\mathbf{D}_{low}$	low resolution depth map
$\hat{\mathbf{D}}_{low}$	decoded low resolution depth map
$\hat{\mathbf{D}}_{EWOC}$	EWOC upscaled decoded low resolution depth map
$\mathbf{E}_{blur}$	depth blur map
$\mathbf{E}_D$	depth edge map
$\mathbf{E}_I$	texture edge map
f	focal length
g	range filter kernel
G	Gaussian filter kernel

h	spatial filter kernel
$\mathbf{I}$	image or video frame
$\hat{\mathbf{I}}$	decoded image or video frame
$\tilde{\mathbf{I}}$	guidance image
$\mathbf{I}'$	low resolution image
$\mathbf{J}$	bilateral filtered image $\mathbf{I}$
$\tilde{\mathbf{J}}$	joint bilateral upscaled image $\mathbf{I}'$
k	normalization factor
$\mathbf{K}_I$	video camera calibration matrix
$\mathbf{K}_T$	depth camera calibration matrix
m	horizontal pixel position
m'	horizontal position in 3D space
$\mathbf{M}_C$	credibility map
n	vertical pixel position
n'	vertical position in 3D space
$\mathbf{P}$	projection matrix
Q_H	horizontal error energy
Q_S	spatial error energy
Q_V	vertical error energy
r	received signal
$\mathbf{R}$	rotation matrix
s	sent signal
t	temporal instance
$\mathbf{t}$	translation vector
u	horizontal pixel offset on image plane
W_E	edge weighting function
x	horizontal image index
X	maximum horizontal image index
y	vertical image index
Y	maximum vertical image index
z	depth position in 3D space
Z_{far}	maximum scene depth
Z_{near}	minimum scene depth
δ	disparity
ϵ_h	horizontal smoothness error
ϵ_v	vertical smoothness error
$\mathbf{\Omega}$	spatial neighborhood (filter window)
σ^2	variance

Chapter 1

Introduction

Television is probably the most important visual information and entertainment system of the last century. While modern trends, such as the increase in web-based applications and the spread of hand-held devices, move the focus more to the internet, TV still occupies a stable place in present society. Nevertheless, constant effort has to be made to remain competitive and to provide interesting entertainment for the viewer.

In recent years, three-dimensional television (3DTV) and its applications has received considerable attention within the research community [Onu11]. Moving the well established 3D cinema into the living room brings many new challenges. The explication of the three-dimensional (3D) scene geometry and acquisition is a classic problem in this field of research. An important task in this area is to match low resolution scene geometry data, i.e. depth maps, with a higher target resolution for three-dimensional video (3DV).

This thesis presents a novel depth upscaling approach, utilizing corresponding edge information from video, and the evaluation and application of this approach.

1.1 Background

This thesis addresses depth map upscaling for 3DTV. It is important to understand the background and motivation behind this task. This section will present the motivation behind the requirement for depth map upscaling and will provide a brief overview with regards to important aspects of 3DTV.

1.1.1 Motivation

The continuous success of 3D movies in the cinema is the driving force behind the idea of 3DTV for our living rooms. However, while some viewing limitations might be acceptable in the movie theater, restrictions such as glass-aided view separation and the limited viewing angle hinder the commercial success of 3DTV.

Recently, researchers have become increasingly interested in autostereoscopic multiview displays. Such displays can provide a three-dimensional viewing experience without additional eye-wear and a larger viewing angle. Scene geometry information, i.e. depth maps, is utilized to increase the number of available views without a dramatic increase of inputs. Fig. 1.1 shows the concept of a multiview displays with reduced input views. The necessary depth information must be created by some means. A standard procedure for capturing depth is stereo analysis from two or more viewpoints. Such view matching methods are often criticized for poor performance in low texturized or occluded areas. In recent years, depth from dedicated range sensors has gained a great deal of interest in this context. It is commonly suggested that such dedicated range sensors can deliver more accurate depth readings than stereo matching. In particular, there has been significant attention given to the Microsoft Kinect, a low-cost, easily accessible structural lighting sensor and to Time-of-Flight (ToF) cameras. However, these sensors suffer from limited spatial resolution compared to modern high definition (HD) video. This lack of spatial resolution motivates the search for sophisticated depth upscaling algorithms.

Another scenario, where spatial resolution of depth and texture sequences might not match, is the transmission of 3D video. The special characteristics of depth maps allow for some spatial downsampling to increase the overall coding efficiency [KWD09]. Well-conceived depth upscaling at the receiver's side can then reconstruct the full resolution depth map.

1.1.2 Three-Dimensional Television

The term 3DTV stands for the efforts to provide a more immersive viewing experience to the people in their every day media consumption. Since the latest breakthrough of 3D movies in cinemas, there has been increased activity in research, product development and marketing to spread the success of 3D entertainment into new markets. In this context, 3DTV not only involves living rooms but many more aspects of daily life, such as mobile devices [GAC+11], communication and telepresence [KS05], advertising and signage [RBV+06], and also medical applications

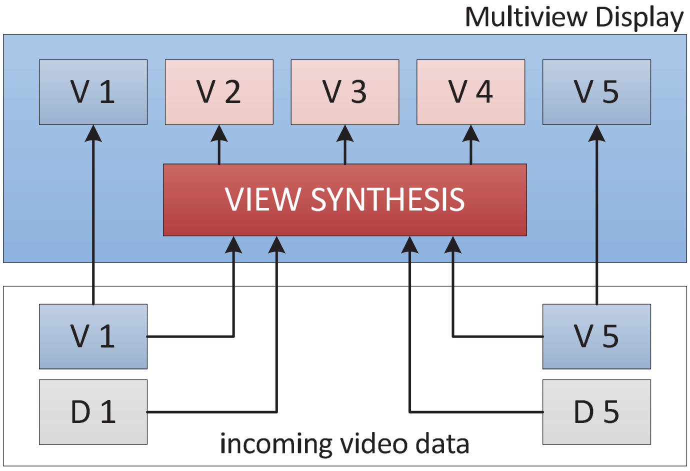


Figure 1.1: View generation for multiview displays from a small set of view inputs V and corresponding depth maps D.

[NBM$^+$12]. The following paragraphs offer a brief summary of the idea, history and technology behind 3DTV.

Human Depth Perception

For the first-time reader, the idea of 3DTV might prove to be rather confusing. After all, there is no actual three-dimensional picture, just an illusion of depth. To understand how this illusion is created, it is important to understand how depth is perceived: Our human visual system (HVS) perceives depth based on a variety of information. This information, or cues, can be categorized in monocular and binocular cues. Binocular cues require both eyes, while monocular cues can be perceived with a single eye. The different cues are:

Monocular Depth Cues

- Perspective: Parallel lines merge in a single vanishing point on the horizon. Points closer to the vanishing point are more distant.
- Occlusion: Closer objects occlude objects behind them.
- Relative size: Close objects appear bigger than distant objects. Memory about standard sizes, e.g. trees, people, allows a distance estimation.
- Accommodation: Changes in focus give feedback about absolute and relative distances.
- Motion Parallax: Objects closer to the viewer cover a bigger visual angle when moving at the same speed as objects further away.

Binocular Depth Cues

- Stereopsis: Object points in 3D space are projected on different positions in the left and right eye. The distance between the two projections forms a depth cue.
- Vergence: Left and right eyes are trimmed on the same point in 3D space. The vergence angle between the two eyes gives a cue for depth.

While monocular cues are already used in traditional, two-dimensional television, 3DTV introduces stereopsis to the viewing experience. To perceive depth, it is important to maintain all the depth cues to be as consistent as possible [CV95]. This is particularly true in relation to the conflict between accommodation and vergence which might lead to problems when combined with stereopsis in 3DTV [IO90]. Since a detailed discussion about aspects and effects of depth cues is beyond the purpose of this thesis, interested readers are referred to [CV95] and [Sed08].

History

The development of 3DTV is strongly interlinked with the history of stereoscopic movies and cannot be described separately.

Early research in binocular vision lead to Wheatstone's ”stereoscope” for still photographs in 1838 [Whe38]. With the rise of motion pictures, the idea was transferred to early concepts of stereoscopic cinema. In 1903, the Lumière brothers showed the first 3D short movie and in 1922 the first 3D feature film was released. The first experiments in 3DTV followed quickly in 1928, still based on mechanical TV. Despite these successful experiments in stereoscopic cinema, the first big impact was between 1952 and 1954, with over sixty-five Hollywood 3D movies. Unfortunately, the lack of stereographic experience hindered the commercial success and 3D movies were more or less forgotten, except for a small revival between 1981 and 1983. It took a hundred years, from the introduction by the Lumière brothers, to the final breakthrough of 3D movies. In 2008 the new 3D boom hit the cinemas. With the support of new digital movie cameras and a deeper understanding of the human depth perception, the viewing experience was dramatically increased. Since then, the number of new 3D movie productions has risen year by year, with over 150 feature-length movies in 2011 alone [Pro12].

Broadcast television had a much more difficult start into the 3D era. While the first experiments were conducted in 1953 and the first ”non-experimental” broadcast was aired in 1980, analogue TV provided only poor quality. In the early 1990s, the upcoming transition to digital services lead to increased research efforts in 3DTV [IJs03]. Soon the Moving Pictures Expert Group (MPEG) picked up on that trend and started working on compression standards for stereoscopic video, resulting in the multiview profile (MVP) for the MPEG-2 standard. In 1998 stereoscopic broadcast started with the transmission of the Winter Olympics in Nagano to special viewing venues. In 2010 the world's first stereoscopic 3D channel started in South Korea. Since then, 3DTV has slowly made its way in the living room. As of July 2012 there are currently thirty-four running stereoscopic 3D channels in the world.

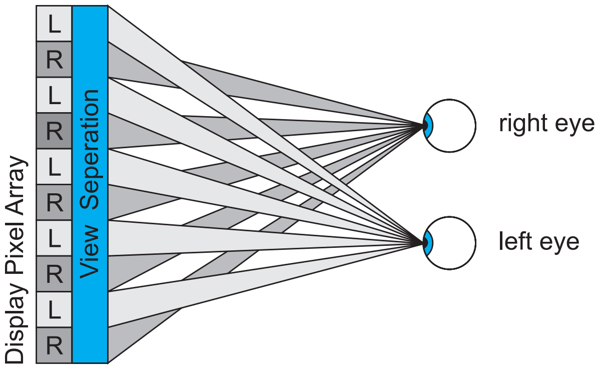


Figure 1.2: Autostereoscopic view separation for 3D display technology.

This summary is only a short excerpt, a more comprehensive history of stereoscopic cinema and 3DTV can be found in [Zon07] and [Feh05].

3D Display Technology

The key idea behind 3DTV is to provide stereopsis, two different views for each eye. That means that the views for the left and right eye must somehow be separated. While Wheatstone's stereoscope achieved this view separation mechanically, solutions for more than one viewer were necessary. Early experiments with anaglyph glasses provided only limited quality. Most modern 3D movie systems nowadays use active shutter glasses or polarization. Current available TV displays, marketed for 3DTV at home, adopt these view separation techniques from the movie theater and bring them into our living rooms. The common threads for all systems is the requirement for special eye-wear for the view separation and are therefore classified as ”aided-viewing” 3D display technologies. They provide only a single stereo view pair from a single viewpoint. Since no additional depth information is required, these viewing setups are of no further interest for this thesis.

In recent years, a new kind of 3DTV display has emerged, providing multiple stereo view pairs without the need to wear glasses. These autostereoscopic displays separate the different views for the left and right eye with lenticular lenses or parallax barriers in front of the display's pixel array, as shown in Fig. 1.2. Autostereoscopic multiview displays are predicted to be the future for the consumer 3DTV market, nonetheless a great deal of research still remains.

Discussing all different kinds of 3D display technologies lies beyond the scope of this thesis. A comprehensive overview is given in [Pas05], also covering more futuristic 3D technologies, such as electroholography and volumetric displays. A detailed survey of current stereoscopic and autostereoscopic displays can be found in [UCES11] and the next section explains how such displays can be provided with 3DTV content.

It is important to note, that this procedure requires per-pixel depth information, i.e. video and depth sequences must be of the same spatio-temporal resolution.

In this thesis, view synthesis is used as a tool for evaluation. Describing the different synthesis concepts falls outside of the scope of this thesis. A detailed description of the underlying geometry is found in [HZ03] and a good overview of different synthesis algorithms in [KES05].

1.2 Overall Research Aim

The intention of this work is to increase the quality of the 3DTV experience by overcoming today's shortcomings in the 3DTV distribution chain, specifically in the content creation and distribution.

3D content creation is still a complex and challenging task, especially for photographic content. New approaches to scene depth capture can simplify the capturing process and increase the quality of 3D content. The coding and transmission of 3DTV is still in an early stage and can profit largely from new compression concepts to limit the necessary bandwidth for the increased amount of data. Finally, good quality 3D content will provide benefits for the DIBR view synthesis and will deliver a more convincing experience on the viewer's 3D display.

In the end a simplified capturing process will increase the amount of 3D content available and efficient 3D coding allows for more 3DTV channels on the same bandwidth. Together these two factors will bring a larger content variety to the customer. In addition, combined with the increase in content quality, this will boost the commercial success of 3DTV.

1.3 Scope

The work presented here addresses the spatial upscaling of depth maps for DIBR view synthesis. For this thesis the depth map source is not further defined and standardized material is used [DGK$^+$09, SS02]. The only precondition for the upscaling process is the existence of a corresponding texture frame, at the same time instance, with the desired target resolution. Such scenarios can, for example, originate from downsampled depth map sequences to save bandwidth for transmission or dedicated range sensors such as time-of-flight (ToF) cameras.

The understanding of depth acquisition and its special characteristics is an important background for this thesis, but the actual capture process itself is not part of this work. Only pre-existing, open available video and depth test sequences are used and considered as reference in the upscaling evaluation. The necessary downscaled depth sequences were generated by downsampling original depth map sequences which have been provided. Parts of the evaluation are based on DIBR view synthesis and were carried out with the MPEG View Synthesis Reference Software (VSRS, [vsr10]). The references were generated using original full resolution depth with identical VSRS settings, since the perceived quality of virtual views is difficult to measure. The upscaling quality itself was assessed objectively, based on the peak-signal-to-noise ratio (PSNR) metric. This metric is very common and widely used in the field of video coding, although it is not without criticism for its poor correspondence to the HVS. Therefore additional metrics, such as the Structural Similarity (SSIM, [WBSS04]) index and subjective evaluations were also applied. More details concerning the used evaluation methodology are given in Chapter 5.

1.4 Concrete and Verifiable Goals

In order to achieve a better 3DTV experience, a sophisticated algorithm for fusing low resolution depth maps with high resolution texture video is necessary. Such an algorithm can provide depth maps suitable for DIBR view synthesis in scenarios where there is a mismatch between texture frame and depth map resolution. Depth map upscaling can provide a more accurately captured scene depth due to the use of dedicated range sensors and higher coding efficiency due to asymmetric 3D video coding. Affecting the entire distribution chain for 3DTV, depth map upscaling can finally lead to a better 3DTV experience.

A possible algorithm should utilize information from the corresponding texture frames, and it should be examined how this information can be best processed for good visual quality in virtual views. While the origin of the low resolution depth can vary, e.g. acquired from range sensors or downsampled for transmission, the upscaling procedure and outcome should not be affected.

In the strive for a better 3DTV experience, the following three goals are defined for this thesis:

I Investigate the upscaling of limited range information, utilizing additionally available data to improve the 3DTV quality, and propose an alternative concept to depth map upscaling.

II Investigate the utility of depth map upscaling for 3D video coding and propose an alternative compression scheme for 3DTV distribution utilizing depth map upscaling.

III Investigate the relationship between the visual quality and computational complexity for depth map upscaling and propose enhancements to the introduced depth map upscaling concept.

1.5 Outline

The remainder of this thesis is structured as follows: Chapter 2 discusses the characteristics and capture procedures for scene depth information. Different upscaling scenarios for this depth information are presented in Chapter 3. The main contribution of this thesis, the Edge-Weighted Optimization Concept (EWOC) for depth upscaling, is introduced in Chapter 4. Chapter 5 investigates the author's contribution in more detail. Finally, the thesis is summarized in Chapter 6, concluding the presented work and giving an outlook on future research.

1.6 Contributions

The author's contributions for this thesis are presented in the previously listed papers. The author is responsible for the concept and ideas, evaluation criteria and design, result analysis and presentation in all four papers. The co-authors have contributed to the definition of evaluation criteria and the analysis. The contributions in this thesis are:

I The introduction of the Edge Weighted Optimization Concept (EWOC) for depth map upscaling based on texture information.

II An application of EWOC for depth map compression in 3D video coding.

III Deeper analysis of the edge weighting process for increased upscaling performance.

IV Quality and performance improvement by an incremental upscaling approach.

More details to each single contribution are presented in Chapter 5.

Chapter 2

Depth Map Acquisition

The ideas for 3DTV distribution, mentioned in Sec. 1.1.2, are based upon depth information accompanying the traditional 2D video. This depth information usually comes in the form of depth maps, a grayscale representation of the video scene. For Computer-Generated Imagery (CGI) content, such depth maps can be easily be obtained via the 3D rendering software. For photographic content, the depth extraction is more complex. There are two widespread methods used to extract depth from photographic scenes: Stereo analysis and sensor-based depth capture, each of which has different advantages and drawbacks.

Both, the characteristics of depth maps and the differences in depth capture are important in order to understand the ideas behind this thesis and are described in this chapter.

2.1 Depth Map Characteristics

Before providing more detail in realtion to how depth maps are generated, it is important to understand depth map characteristics. There are many different ways to describe a 3D scene. Depth maps are the fundamental depth representation for 3DTV. They can be classified as ”Dense Depth Representations” [AYG$^+$07]. For each pixel in the 2D video sequence a corresponding pixel exists in the depth map. Each depth map pixel value corresponds to the distance between a point in 3D space, captured by the corresponding video pixel, and the camera. Together with additional metadata, i.e. camera parameters, this depth value enables a video pixel to be projected into 3D space. From there, the pixel is then re-projected onto a virtual viewing plane, generating a new virtual view.

Traditionally depth maps are represented as 8-bit grayscale images. This representation allows for 255 depth steps between the minimum and maximum distance given in the metadata. It is important that depth maps are dense, i.e. every video pixel has a depth pixel associated with it, otherwise missing depth values will lead to errors in the view synthesis.

Since depth maps represent the scene geometry and not the texture, they usually consist of several regions with smooth gradual value changes and sharp value transitions at the region borders. Fig. 2.1 shows an example for a video frame and the corresponding depth map.

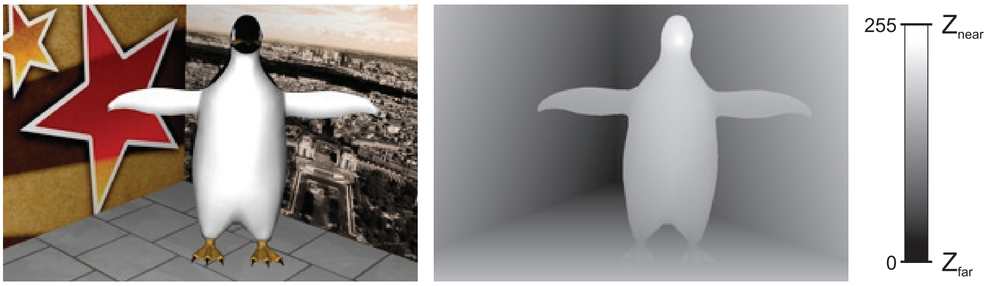


Figure 2.1: Example of a video frame with corresponding depth map.

2.2 Depth from Stereo Analysis

Image analysis from two or more cameras is the traditional depth extraction approach for view synthesis of photographic content [Sch99]. The idea is to find corresponding picture points between the different views. Together with information about the capturing process, i.e. extrinsic and intrinsic camera parameters, the correspondences allow for extracting depth from a 3D space. Extrinsic camera parameters describe the camera position and orientation in 3D space. The intrinsic parameters describe the internal camera configuration, i.e. focal length, resolution and pixel

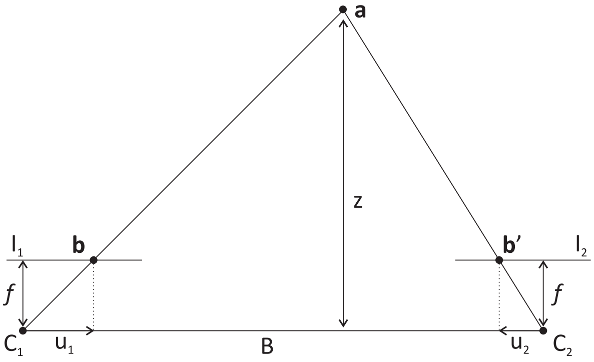


Figure 2.2: Concept of disparity.

aspect ratio. Because the correspondence matching between the views is based on likelihood calculations, this process is also called "depth estimation".

2.2.1 Image Disparity

The relationship between points in corresponding views is described by an offset in pixels, called "Disparity". Fig. 2.2 shows the concept of disparity: A point **a** in 3D space, seen from camera $\mathbf{C}_1$ at focal length f, will be projected at position **b** image plane I_1, with distance u_1 to the central point. From a different camera position $\mathbf{C}_2$, the point will be projected at $\mathbf{b}'$ on I_2, with a different distance u_2. The difference between u_1 and u_2 is the disparity δ.

$$\delta = u_1 - u_2 \tag{2.1}$$

Using the intercept theorem, the depth z of point **a** is given by the disparity δ, the focal length f and the distance B between the two cameras, also called "baseline":

$$z = \frac{Bf}{\delta} \tag{2.2}$$

In other literature, the terms positive, or uncrossed, and negative, or crossed, disparity might occur. These terms describe a depth relationship with respect to a plane, usually the display plane of a 3D source. Positive values describe points behind the relation plane, e.g. in front of the display whereas negative values lie behind the plane, e.g. "inside" the display [ISM05].

2.2.2 Stereo Matching

Stereo analysis or stereo matching relies on image and feature correspondences between two views. Based on the knowledge about the capturing system and the relation between the corresponding image points, it is possible to estimate the scene depth. Determining the corresponding image points between two or more views is critical for the quality of the estimated depth maps and the resulting view synthesis. There are a multitude of different approaches for stereo matching, this thesis is only able to mention the two main categories:

- Area-based approaches consider a window around the pixels to consider similarity and to handle ambiguity between two views. They deliver dense depth maps, but fail in low texturized areas and occluded regions in one or more views.
- Feature-based approaches use edge, line and corner correspondences in two views for sparse but robust depth maps.

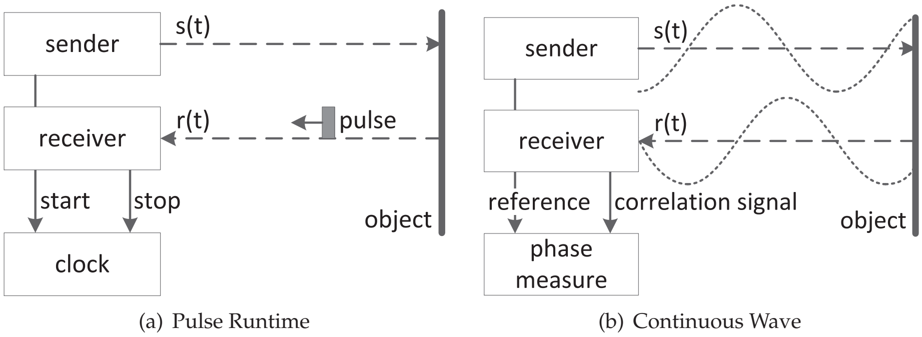
State-of-the-art stereo matching algorithms rely on a combination of both approaches [SB12], however, they are still unable to compensate for the main shortcomings of stereo matching:

- Occlusions: If parts of the scenery are not visible in one of the stereo views, it is impossible to establish a correspondence and therefore no depth value can be estimated.
- Low texture: Feature- and area-based correspondence relies on texture information. Low texturized areas do not deliver sufficient information to create robust correspondences and result in erroneous depth estimations.
- Repetitive texture: Again, stereo matching relies on texture information. Repetitive texture will lead to ambiguous correspondences, resulting in inaccurately estimated depth values.

Stereo matching is a complex and time consuming process and is usually performed in post processing. It is not necessary for this thesis to understand the deeper concepts of stereo matching. It is only important to understand the basic idea and the shortcomings which have been presented in this section. A comprehensive overview can be found in [AM05], a list of up-to-date stereo matching algorithms and their evaluation is found in [SB12].

2.3 Depth from Range Sensors

Sensor based solutions for scene depth extraction have been around for some time, it is only recently that they have started to be used in depth map creation. Traditional range sensors could only generate a single depth value at any time instance and had to scan models line by line. Model scanning is not feasible for moving scenery as intended for 3DTV. Nowadays, new sensor arrays, capable of measuring a whole scene in real-time, are entering the market. These Focal Plane Array (FPA) sensors for range sensing are the focus of this section. Additionally, structured lighting is addressed briefly, which is another depth capture approach without point-wise scanning. A wider description of alternative range sensors can be found in [GS05].



(a) Pulse Runtime (b) Continuous Wave

Figure 2.3: Classification of ToF systems according [KI06].

2.3.1 Time-of-Flight Principle

FPA sensors measure distance based on the signal travel time between sender, object and detector. This principle is called time-of-flight (ToF) and those sensors are usually referred to as ToF cameras. There are two basic categories for ToF measurements:

- Pulse Runtime Sensors: A pulsed wave is sent out and a clock measures the time which has passed until it has again been received (Fig. 2.3.1). Such sensors can deliver depth accuracy between 10-20mm for distances of up to a few hundred meters, but have low temporal resolution due to their pulsed nature [GS05].
- Continuous Wave Sensors: A modulated wave is sent out. When this wave hits an object, the reflected signal will be phase-shifted. This shift allows for the measurement of the travel distance (Fig. 2.3.1). In an FPA setup, such sensors have an accuracy of around ten millimeters and a maximum distance of about ten meters [KI06]. They can capture in real-time with sixty frames per second and more [Fot12].

Due to their real-time capabilities, continuous wave ToF cameras are predestined for 3DTV capture. They can deliver accurate depth maps at the point of capture, without any time intensive stereo matching in post production. Also, unlike stereo analysis, they deliver reliable and accurate depth information in low or repetitively texturized areas and do not suffer from occlusions, as there is no requirement for multiple viewpoints . Nevertheless, ToF cameras are not without their shortcomings:

- Spatial resolution: Current ToF sensors are still unable to deliver suitable spatial resolution to match HD video material [BOL$^+$05]. Some kind of depth upscaling is necessary.
- Ambiguity problem: The sensor cannot distinguish between multiples of the modulation frequency, reducing the usable range coverage [BOL$^+$05].

- Sensor noise: The ToF signal is not without noise affecting the depth accuracy. Noise reduction and filtering can improve the result [FPR+09].

Again, a deeper understanding of the ToF principle is not necessary for this thesis. The important aspects have been presented in this section. A detailed explanation of the ToF principle can be found in [LS01].

2.3.2 Structured Lighting

Structured lighting refers to scene geometry analysis based on a structured light source [YA88]. A scene is lit with a known pattern, e.g. a grid. This pattern is then captured by a camera. The relationship between the distorted pattern on the scene and the known pattern allows for the reconstruction of a 3D model. Modern structured light scanners send the pattern in a non-visible spectrum, so the depth capture does not interfere with the video capture. A very popular example is the Microsoft Kinect system [Zha12], which recently initiated significant attention to structural lighting due to the low cost and open availability of the Kinect. The downsides of depth from structural lighting involve the complicated geometric distortions from the pattern processing and the strong inaccuracies at object boundaries [CKH12].

Depth from structured light is not part of this thesis and is only mentioned in sake of completeness due to the recent boom of the Kinect within the consumer market. However, it is important to mention that the current Microsoft Kinect system does not deliver full resolution depth maps compared to the video resolution and therefore requires some sort of depth map upscaling.

2.4 Depth Acquisition Comparison

This chapter presented two common approaches for scene depth extraction, namely, stereo matching and range sensors. For range sensors, two approaches currently form the focus of the research community: 1) ToF cameras and 2) structural lighting with the Microsoft Kinect system. The different characteristics between these three different depth capture approaches are presented in Tab. 2.1.

In many areas ToF cameras and the Kinect system are superior to depth estimated from stereo analysis. The capture setup is simpler and occlusions or texture do not affect the outcome. Since there is only one capturing camera, no problems occur in relation to color differences between the two cameras. The limited operating range of the Kinect system restricts its use to close-up scenarios. ToF cameras can be operated in a more versatile manner. Together with their real-time capabilities of around 60 frames per second, ToF cameras are very interesting for 3DTV capture, especially for live broadcasting. The main drawback is the low spatial resolution of depth values, compared to the corresponding video frame. This lack of spatial resolution has been the motivation behind the search for sophisticated depth upscaling algorithms and will be addressed in the following chapter.

Table 2.1: Different characteristics for depth map capture.

	Stereo matching	ToF cameras	Microsoft Kinect
Features			
Temporal resolution	Extensive post-processing, no real-time application.	Up to 60 frames/second.	Up to 30 frames/second.
Spatial resolution	Same as video. Requires extensive post-processing.	Bad, around 1/8th of video.	Medium, 1/2th of video.
Range	Depends on capture baseline and camera resolution.	Depends on modulation frequency, typically 7-15m.	Limited to indoor application, min. 1,2 - max. 3,5m range.
Setup	High complexity, two or more calibrated cameras.	Low complexity, single depth camera.	Low complexity, single depth camera.
Problems			
Color differences	Low correspondence, erroneous depth values.	Accurate depth values.	Accurate depth values.
Occlusions	No correspondence, no depth values.	Accurate depth values.	Accurate depth values.
Low texture	Low correspondence, erroneous depth values.	Accurate depth values.	Accurate depth values.
Repetitive texture	Ambiguous correspondences, erroneous depth values.	Accurate depth values.	Accurate depth values.
Non-lambertian reflectance	Low correspondence, erroneous depth values.	Sensor noise, erroneous depth values.	Low structure correspondence, erroneous depth values.

Chapter 3

Depth Map Upscaling

The previous chapter dealt with the idea of depth maps and explained their characteristics. It also explained different approaches in relation to extracting depth maps during or after the capture process and the difficulties that might arise. One key problem was a possible discrepancy between the video and depth resolution. There are possible solutions for this problem if the depth maps are accompanied with corresponding video texture sequences, a typical scenario in 3DTV.

For this scenario, it is not essential to know where the depth maps have come from. They could be captured from a ToF camera, or be a downscaled version of previously recorded depth maps in order to save transmission bandwidth. All that is important is that there exist corresponding color video frames, in the desired target resolution, for each depth map at exactly the same time instance.

This chapter will address the basic idea of depth map upscaling with combined video data, depth upscaling application for 3DTV and will summarize the existing proposals.

3.1 Basic Idea

Depth maps can be seen as simple grayscale 2D images, and the idea of image upscaling is quite old. The pixel grid of a low resolution image is expanded to a higher resolution. Missing values are filled using different approaches, the most common of which involves interpolation between close known values. For the major part of a depth map this works in an acceptable manner, since it mainly consists of smooth areas with gradual value changes, as described in Sec. 2.1. The problems occur at the hard depth transitions at object borders. These sharp depth changes determine the foreground and background objects in the view synthesis. They must be preserved since the HVS is most sensitive to such depth changes [DRE+11]. Interpolation approaches would smooth out these transitions in depth. Other edge-preserving approaches such as ”nearest-neighbor” or median filtering do not work in an adequate manner for higher upscaling factors where they might separate depth from texture edges [KCLU07].

However, for a 3D video signal, more information exists which can then be utilized in the depth upscaling process. Depth maps are merely additional information added to the target resolution 2D video stream. With respect to the depth map, these high resolution images can deliver important data to the upscaling process. Combining a low resolution depth map with a high resolution video texture frame can assist in containing the edge-blurring interpolation effects and generate high resolution depth maps of good quality.

3.2 Applications

There are many different scenarios for which low resolution information from one source, e.g. depth, is required to be fused with a high resolution source, such as texture video. This thesis addresses the two major scenarios for depth map upscaling, namely, ToF scene capture and depth map compression for 3D video coding. Other possible applications lie in stereo matching algorithms and other complex computer vision tasks [KCLU07].

3.2.1 ToF Scene Capture

As mentioned in Sec. 2.3.1, ToF cameras can deliver reliable depth in scenarios where stereo analysis fails. Their real-time capabilities make ToF cameras interesting for 3DTV capture. However, due to their low spatial resolution the raw depth signal is not feasible for DIBR view synthesis and some sort of upscaling must be applied.

A typical ToF capture scenario for 3DTV consists of one or more standard cameras for 2D video and one or more ToF cameras for depth. Video and depth cameras should be synchronized, i.e. capturing at the same temporal instance. It is also important to know the spatial relationship between the cameras, to enable the different viewing angles to be merged to a single point of view. Once video and depth are aligned, the video texture can be used to assist in upscaling the ToF depth to the desired target resolution.

The most popular approach for guided depth map upscaling is probably Joint Bilateral Upscaling (JBU), introduced in 2007 by Kopf et al. in 2007 [KCLU07]. A slightly earlier proposal, based on a Markov Random field, came 2006 from Diebel and Thrun [DT05] and is explained below in Sec. 3.3.1.

3.2.2 3D Video Coding

Another application for the joint upscaling of depth maps is 3D video compression. A typical transmission format for 3DTV is the multiview video plus depth format (MVD): Multiple video streams together with matching depth map sequences allow for a DIBR view synthesis at the receiver side [KAF+07]. Due to their special characteristics, explained previously in Sec. 2.1, depth maps can be compressed very efficiently. The European ATTEST project [FKB+02] has shown that depth maps can be compressed at 10-20% of the overall bit rate budget [SMS+07]. Klimaszewski et al. [KWD09] showed that coding efficiency can be further increased by transmitting downscaled depth maps.

Several proposals have been made to increase coding efficiency by low resolution depth maps, upscaling to the desired target resolution with the corresponding texture information at the receiver's side, e.g. utilizing texture edges [EMWK09] or weighting color similarity [WYT+10, LS10].

3.2.3 Stereo Matching Algorithms

As presented in 2.2, scene depth can also be generated from two or more matching camera views. Stereo correspondence search and analysis is a very complex task and is often carried out over downsampled versions of the original source [SS02]. The low resolution representations can offer more reliable correspondences and speed up the overall disparity matching process. Guided upscaling with the original source provides a full resolution depth map.

This upscaling idea is not only limited to stereo depth applications. It is a common approach in computer vision to apply complex tasks on downsampled sources to increase computational performance. Popular examples are tone mapping for High Dynamic Range (HDR) imaging [LFUS06], colorization or recoloring [LLW04] and graph-cut based image operations [BVZ01], to name but a few.

3.3 Existing Approaches

The idea of joint depth upscaling is not new and several proposals have been made previously. This section provides a brief chronological overview of existing solutions, with a special focus on Joint Bilateral Upscaling. JBU has received a great deal aligned, the video texture can be used to assist in upscaling the ToF depth to the desired target resolution.

The most popular approach for guided depth map upscaling is probably Joint Bilateral Upscaling (JBU), introduced in 2007 by Kopf et al. in 2007 [KCLU07]. A slightly earlier proposal, based on a Markov Random field, came 2006 from Diebel and Thrun [DT05] and is explained below in Sec. 3.3.1.

3.2.2 3D Video Coding

Another application for the joint upscaling of depth maps is 3D video compression. A typical transmission format for 3DTV is the multiview video plus depth format (MVD): Multiple video streams together with matching depth map sequences allow for a DIBR view synthesis at the receiver side [KAF$^+$07]. Due to their special characteristics, explained previously in Sec. 2.1, depth maps can be compressed very efficiently. The European ATTEST project [FKB$^+$02] has shown that depth maps can be compressed at 10-20% of the overall bit rate budget [SMS$^+$07]. Klimaszewski et al. [KWD09] showed that coding efficiency can be further increased by transmitting downscaled depth maps.

Several proposals have been made to increase coding efficiency by low resolution depth maps, upscaling to the desired target resolution with the corresponding texture information at the receiver's side, e.g. utilizing texture edges [EMWK09] or weighting color similarity [WYT$^+$10, LS10].

3.2.3 Stereo Matching Algorithms

As presented in Sec. 2.2, scene depth can also be generated from two or more matching camera views. Stereo correspondence search and analysis is a very complex task and is often carried out over downsampled versions of the original source [SS02]. The low resolution representations can offer more reliable correspondences and speed up the overall disparity matching process. Guided upscaling with the original source provides a full resolution depth map.

This upscaling idea is not only limited to stereo depth applications. It is a common approach in computer vision to apply complex tasks on downsampled sources to increase computational performance. Popular examples are tone mapping for High Dynamic Range (HDR) imaging [LFUS06], colorization or recoloring [LLW04] and graph-cut based image operations [BVZ01], to name but a few.

3.3 Existing Approaches

The idea of joint depth upscaling is not new and several proposals have been made previously. This section provides a brief chronological overview of existing solutions, with a special focus on Joint Bilateral Upscaling. JBU has received a great deal

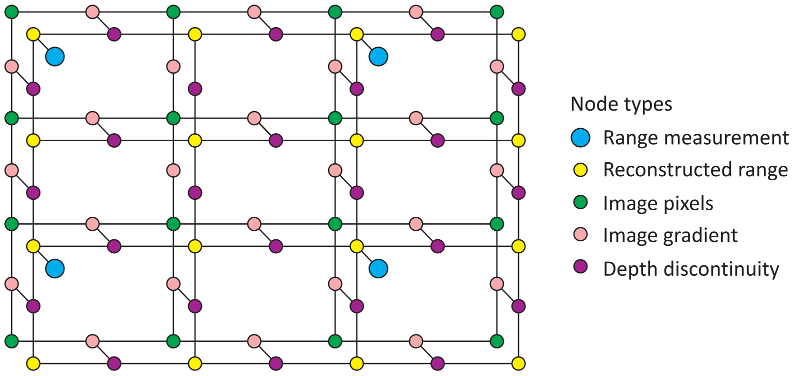


Figure 3.1: The MRF for range measurement upscaling according to [DT05].

of attention in this research field and is the basis for many subsequent proposals. It is also the main competing reference in the contributions I, III and IV of this thesis and therefore deserves more notice.

3.3.1 Markov Random Field Approach

In 2006 Diebel and Thrun presented a Markov Random Field (MRF) to fuse high-resolution texture data with low-resolution depth data, which was particularly aimed at range sensors, e.g. laser range measurement or ToF cameras. Exploiting the fact that discontinuities in depth and texture tend to co-align, this approach scales the spatial depth map resolution to the texture image.

Their MRF contains five types of nodes in two layers, namely, the range measurement, the reconstructed range, the image pixel, the image gradient and the depth discontinuity nodes. The interconnections between the nodes are shown in Fig. 3.1. The reconstructed range nodes have the same density as the image pixel nodes, whereas the original range measurement nodes are of lower density. The auxiliary nodes for image gradient and depth discontinuity mediate texture and depth information for the reconstructed range nodes.

Diebel and Thrun claim this work to be the first application of MRF in relation to multi-modal data integration and show that the use of MRF can substantially improve existing range imaging technology (in 2005), generating high-resolution, low-noise range images [DT05].

3.3.2 Joint Bilateral Upscaling

Kopf et al. [KCLU07] investigated into the upscaling of low-resolution images for different image analysis applications, such as tone mapping, colorization, graph-cut image operations and stereo depth. Such tasks are often run over a downsampled input image in order to reduce computational complexity. The solution is then upscaled to the original resolution. However, since traditional upscaling methods assume a smoothness-prior for the interpolation, a new upscaling method was proposed, using the original high resolution input image as a prior for a joint bilateral upsampling procedure based on bilateral image filtering introduced by Tomasi and Manduchi [TM98].

A bilateral filter is an edge-preserving smoothing filter, based on a nonlinear combination of the surrounding pixel values $I(x, y)$ in image $\mathbf{I}$. The filter blends pixel values based on geometric distance (spatial) and photometric similarity (range). In this context, images are represented as two dimensional matrices of pixel values, e.g. $\mathbf{I} = \{I(x, y); x = 1, ..., X; y = 1, ..., Y\}$ with X and Y as the maximum indices. The bilateral filter has a symmetric spatial filter kernel $h(\cdot)$ with support $\boldsymbol{\Omega}$ and a symmetric range filter kernel $g(\cdot)$. $h : \Re \to \Re$ uses the Euclidean distance and $g : \Re \to \Re$ the absolute value difference between two pixels as input. For a pixel at position $(x, y)^T$, the filtered result $J(x, y)$ of pixel $I(x, y)$ from the image $\mathbf{I}$ is:

$$J(x,y) = \frac{1}{k} \sum_{\binom{x'}{y'} \in \boldsymbol{\Omega}} I(x', y') h \left(\left\| \begin{pmatrix} x \\ y \end{pmatrix} - \begin{pmatrix} x' \\ y' \end{pmatrix} \right\|_2 \right) g\left(|I(x,y) - I(x', y')|\right) \tag{3.1}$$

where $\boldsymbol{\Omega}$ is the spatial support of the kernel, centered at $(x, y)^T$, and k is the number of all pixels in $\boldsymbol{\Omega}$. Edges are preserved, since the filter outputs smaller values when range or spatial differences increase.

It is not necessary that the filter kernel inputs come from the same source. If the range kernel is applied to a second image, this process is called a joint or cross bilateral filter. The second image can be used as guidance for an upscaling process. Applying the spatial filter kernel $h(\cdot)$ on pixel $I'(m, n)$ at position $(m, n)^T = \left(\frac{x}{\gamma}, \frac{y}{\gamma}\right)^T$ of low resolution source $\mathbf{I}'$ and the range filter kernel $g(\cdot)$ on pixel $\tilde{I}(x, y)$ of a full resolution guidance $\tilde{\mathbf{I}}$, yields the joint bilateral upscaling result $\tilde{J}(x, y)$:

$$\tilde{J}(x,y) = \frac{1}{k} \sum_{\binom{x'}{y'} \in \boldsymbol{\Omega}} I'(x', y') h \left(\left\| \begin{pmatrix} \frac{x}{\gamma} \\ \frac{y}{\gamma} \end{pmatrix} - \begin{pmatrix} x' \\ y' \end{pmatrix} \right\|_2 \right) g\left(\left| \tilde{I}(x,y) - \tilde{I}(\gamma x', \gamma y') \right| \right) \tag{3.2}$$

where γ is the upscaling factor between $\mathbf{I}'$ and $\tilde{\mathbf{I}}$. Since $I'(m, n)$ takes only integer coordinates, the guidance image $\tilde{\mathbf{I}}$ is only sparsely sampled.

Kopf et al. demonstrated the benefit of JBU for upscaling low resolution depth maps with high resolution texture guidance. Their results show high resolution depth maps with accurate, sharp edges at object boundaries. However, solving the edge smoothing problem with a range filter kernel introduced a new problem, namely, texture copying. Highly structured texture, especially letters, will be transferred into the depth map, since they are regarded as edges which should be preserved. This problem motivated Garcia et al. to introduce the Pixel Weighted Average Strategy (PWAS) for depth upscaling [GMO$^+$10] which will be addressed in the next section.

3.3.3 Pixel Weighted Average Strategy

The PWAS can be seen as an extension to JBU, particularly applied for ToF upscaling. Garcia et al. aimed to reduce sensor noise by joint bilateral filtering, while avoiding the effects of texture copying. Assuming that range values at depth transitions are less reliable, they introduce a two-dimensional credibility map $\mathbf{M}_C$, generated from the absolute gradient of the low resolution source $\mathbf{I}'$,

$$\mathbf{M}_C = G_{\sigma_c}\left(\left|\frac{\partial^2 \mathbf{I}'}{\partial x \partial y}\right|\right) \tag{3.3}$$

where G_{σ_c} is a Gaussian kernel with variance σ_c^2. The credibility map $\mathbf{M}_C$ extends the traditional bilateral JBU filter and provides a trilateral version of the JBU equation given in Eq. 3.2:

$$\tilde{J}(x,y) = \frac{1}{k}\sum_{\binom{x'}{y'}\in\mathbf{\Omega}} I'(x',y')h\left(\left\|\begin{pmatrix}\frac{x}{\gamma}\\ \frac{y}{\gamma}\end{pmatrix} - \begin{pmatrix}x'\\ y'\end{pmatrix}\right\|_2\right) g\left(\left|\tilde{I}(x,y) - \tilde{I}(\gamma x', \gamma y')\right|\right) M_C(x',y') \tag{3.4}$$

The objective evaluation shows a reduction in texture copying, thus not completely eliminating it. However, compared to JBU and MRF, PWAS yields better results.

3.3.4 Depth Upscaling Classification

Depth map upscaling approaches can be classified into two major groups: Guided algorithms with additional information, i.e. texture, and unguided algorithms based on the depth map alone. In other literature the term ”assisted” is also used to describe texture guidance. In recent years, there have been a multitude of proposals. Many are particularly aimed at ToF capture or 3D video coding. It is often the case that they are some sort of variation or extension to the JBU principle, e.g. the Noise Aware Filter for Depth Upsampling (NAFDU, [CBTT08]) or [KYY11]. MRF-based approaches are presented in [PKT+11] and [CLK+12].

The chart in Fig. 3.2 provides a graphical classification of different proposals and their relationships, but it does not claim to be a complete list. Covering all depth upscaling proposals of the last years would prove to be a work in itself. Nevertheless, it is apparent that many proposals share similar foundations. The next chapter will introduce a novel depth upscaling method, which is not based on JBU or MRF: The Edge-Weighted Optimization Concept (EWOC).

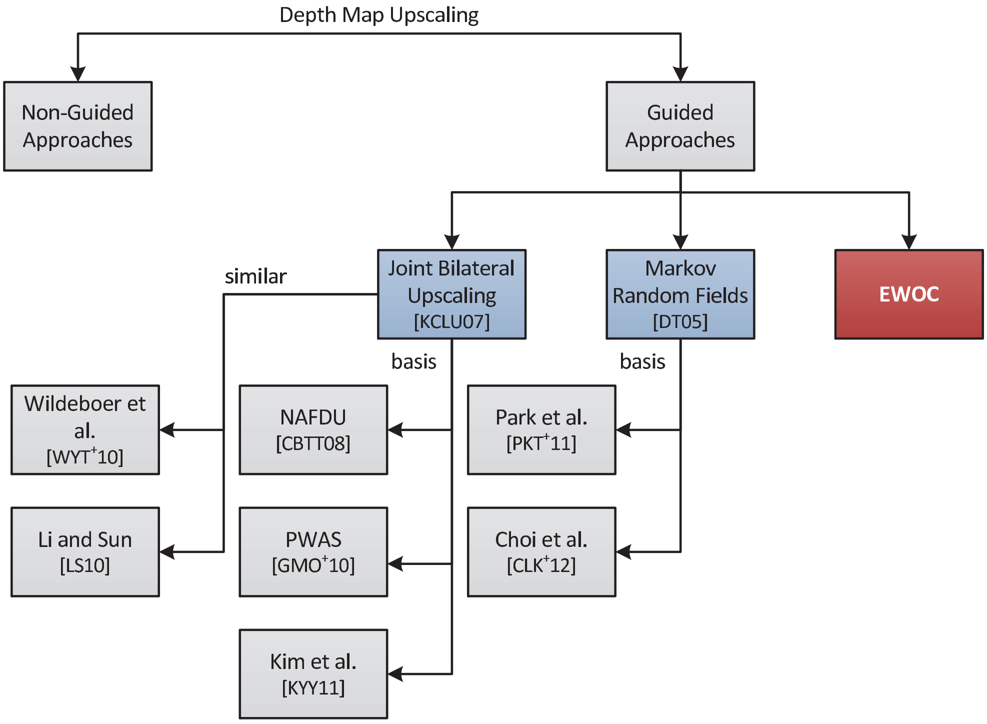


Figure 3.2: Classification of depth upscaling approaches.

Chapter 4

The Edge-Weighted Optimization Concept

In the previous chapter, the requirement for depth map upscaling and the idea of combined upscaling with texture information were introduced. Existing approaches, such as JBU, deliver already sophisticated results and form the basis for many variations and extensions.

The Edge-Weighted Optimization Concept, or EWOC, opens up a new path: Using the same principles as Guttmann et al. [GWCO09], low resolution depth is seen as a sparse representation of the target resolution depth. Missing values are filled by diffusion in an optimization process weighted with edges from the high resolution video frame. Additionally, these edges are validated with the low resolution depth to accentuate correlated data.

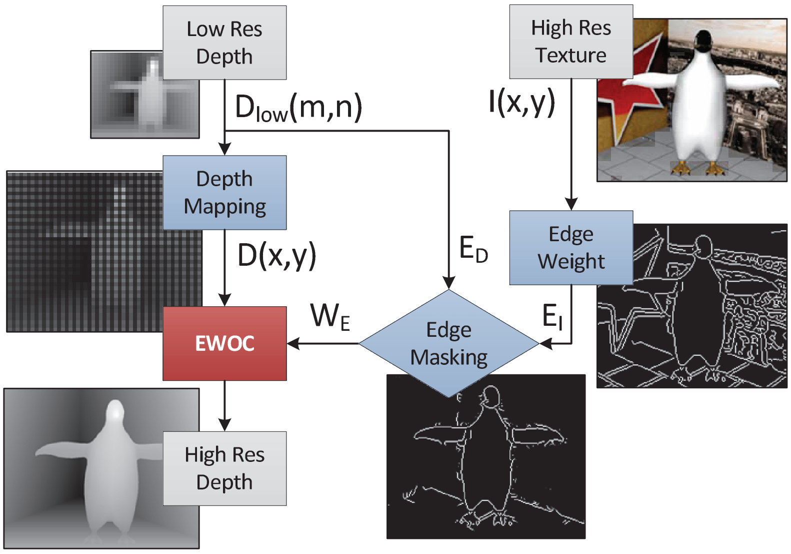


Figure 4.1: EWOC depth upscaling workflow.

4.1 Background

In [GWCO09], Guttmann et al. presented a semi-automatic system to convert conventional video into a stereoscopic video pair. With only a few user generated depth value inputs on one frame, so called ”scribbles”, a whole video sequence can be converted from 2D to stereoscopic 3D. Together with other constraints, such as spatial and temporal smoothness, the scribbles formed an over-determined linear equation system for a dense depth map. This system was solved for a lower (one quarter) resolution of the video input and then upscaled using JBU. Finally a simple forward warping was applied to generate the left and right view for stereoscopic 3D.

The underlying idea of EWOC is an adaption of the Guttmann approach to depth map upscaling for ToF capture or 3D video coding. In these scenarios, the temporal smoothness can be ignored. Unlike the scribbles, reliable depth values are available for every time instance. In addition, the low resolution depth still provides a great deal of data to improve the upscaling process. Previous depth upscaling proposals tend to ignore this data. Together, these ideas lead to the Edge-Weighted Optimization Concept for depth upscaling.

4.2 Upscaling Principle

Fig. 4.1 shows the basic principle of EWOC depth upscaling. The low resolution depth input is treated as a sparse depth representation at the video input resolution. Edge information from the corresponding video frame is used in an weighted optimization to fill out missing values for a dense, high resolution depth map. The single steps are now addressed in more detail.

4.2.1 Depth Value Mapping

The first step is to establish a spatial correspondence between the low resolution depth map and the high resolution guidance frame. In a 3D video coding scenario this mapping is straight-forward. Video and depth sequences are taken from the same viewpoint and the downsampling scheme should be known from the encoder. Known depth values are simply mapped to their original position prior to the downsampling, all other positions are left blank.

For ToF depth upscaling, the depth value mapping requires slightly more attention. It is important to know the relationship between the video and ToF camera in order to merge the two viewing angles into one. This relationship is expressed in the projection matrix **P**, a 3×4 full rank matrix containing the rotation matrix **R**, translation vector **t** between the two cameras, and the intrinsic parameters of the video camera in the calibration matrix $\mathbf{K}_I$,

$$\mathbf{P} = \mathbf{K}_I[\mathbf{R}|\mathbf{t}]. \tag{4.1}$$

A combined video plus ToF recording setup delivers a high resolution texture frame **I**, with the pixel values $I(x, y)$, and a low resolution depth map $\mathbf{D}_{low}$, with the depth values $D_{low}(m, n)$. The coordinates x, m and y, n are Euclidean pixel coordinates in 2D space with max(m) < max(x) and max(n) < max(y). The homogeneous pixel coordinates $(m, n, 1)^T$ can be translated to world coordinates by means of the ToF camera calibration matrix $\mathbf{K}_D$. Together with $D_{low}(m, n)$ for the depth value z, they give the world coordinates of point **a** in 3D space.

$$\mathbf{a} = (m', n', z, 1)^T = \mathbf{K}_D(m, n, D_{low}(m, n), 1)^T \tag{4.2}$$

With the projection matrix **P** from Eq. 4.1, the depth value $D_{low}(m, n)$ is mapped on the corresponding pixel coordinates $(x, y)^T$ for point **b**:

$$z \cdot \mathbf{b} = z \cdot \left(\frac{x}{z}, \frac{y}{z}, 1\right)^T = \mathbf{Pa} \tag{4.3}$$

Performing the projection for every known value in $\mathbf{D}_{low}$ on an empty frame with an equal size as **I** gives the depth values $D(x, y)$ of the depth map **D** from the same viewing angle of as the video camera:

$$D(x, y) = \begin{cases} D_{low}(m, n), & \forall \mathbf{b} \text{ from Eq. 4.3} \\ \text{not defined}, & \text{otherwise} \end{cases} \tag{4.4}$$

Fig. 4.2 illustrates the mapping process for a single point: A ToF camera at central point $\mathbf{C}_{ToF}$ captures the depth z of 3D point **a** as depth value $D_{low}(m, n)$. With the

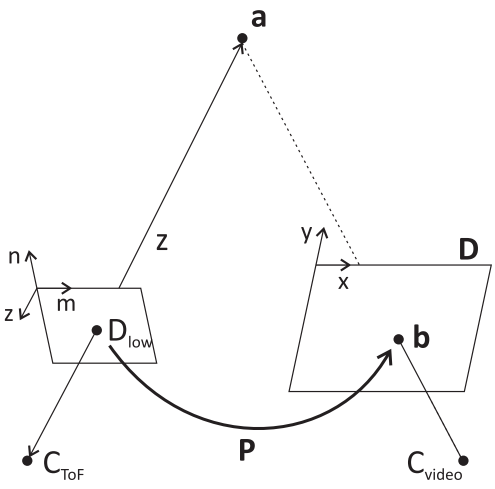


Figure 4.2: Mapping ToF depth values on video camera viewing angle and resolution.

world coordinates from Eq. 4.2, the value $D_{low}(m, n)$ is projected onto pixel position **b** using Eq. 4.3. The sum of all projected points results in the depth map **D**, as seen from a video camera at central point $\mathbf{C}_{video}$.

In both cases, 3D video coding and ToF upscaling, the mapping results in a depth map **D** with the desired target resolution, but a sparse, possibly irregular, value distribution. For DIBR view synthesis it is important to have per-pixel depth values, therefore it is essential to have some means of filling in the missing values.

4.2.2 Spatial Smoothness

For the majority of cases, the characteristics of the depth maps enables the assumption that there is a similarity between a depth pixel $D(x, y)$ at position $(x, y)^T$ and its spatial neighbors. The similarity is represented by the horizontal error ϵ_h and the vertical error ϵ_v:

$$\epsilon_h(x, y) = D(x, y) - D(x + 1, y) \tag{4.5}$$

$$\epsilon_v(x, y) = D(x, y) - D(x, y + 1) \tag{4.6}$$

One exception for the similarity assumption is the sharp depth transition at the borders of objects. Relying on the spatial smoothness errors alone would smooth out these important transitions. Therefore a weighting function $W_E(x, y)$ is introduced, relaxing the spatial smoothness constraints of Eq. 4.5 and 4.6 at the object boundaries. This weighting allows for sharp depth transitions between objects since there is a reduced requirement for the neighboring pixels to be similar. The spatial smoothness errors from Eq. 4.5 and 4.6 are converted into energy terms and weighted by means of W_E to form the horizontal and vertical error energies Q_H and Q_V:

$$Q_H = \sum_x \sum_y W_E(x,y)\epsilon_h^2(x,y) \tag{4.7}$$

$$Q_V = \sum_x \sum_y W_E(x,y)\epsilon_v^2(x,y) \tag{4.8}$$

$$Q_S = Q_H + Q_V \tag{4.9}$$

The sum of Q_H and Q_V provides the overall spatial error energy Q_S, which is then minimized using a block-active method [PJV94], implemented in MATLAB by Adlers [Adl98]. The optimization solution provides the missing values for $\mathbf{D}$, which are combined with the known values from $\mathbf{D}_{low}$ to form a pixel-dense depth map.

4.2.3 Edge Weighting

The weighting function $W_E(x, y)$ is important in order to obtain sharp object boundaries in depth maps. Since depth maps describe the scene geometry for a video sequence, it is possible to extract object boundaries from the corresponding video frame. Assuming that the texture edges correspond to the object boundaries, which then correspond to depth transitions, the weighting function $W_E(x, y)$ can be gained from an edge detection function $E_I(x, y)$ on a video frame $\mathbf{I}$:

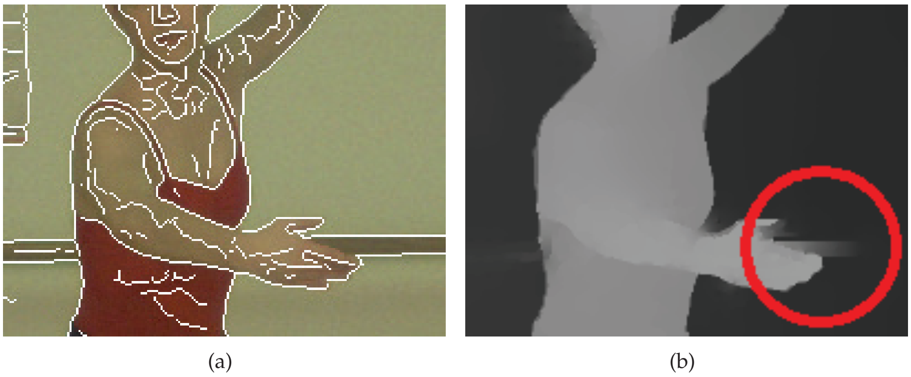
$$W_E(x,y) = 1 - E_I(x,y) \tag{4.10}$$

Accurate and cohesive edges are the key to providing an adequate depth upscaling. Missing or porous edges can lead to “depth leakage” where erroneous depth values spread into the wrong areas as shown in Fig. 4.3. Different edge detectors, pre-processing steps and color spaces can be utilized to influence the edge map accuracy. For the sake of simplicity, a standard Canny edge detector [Can86] on the video luminance channel is used at the present time.

4.2.4 Edge Validation

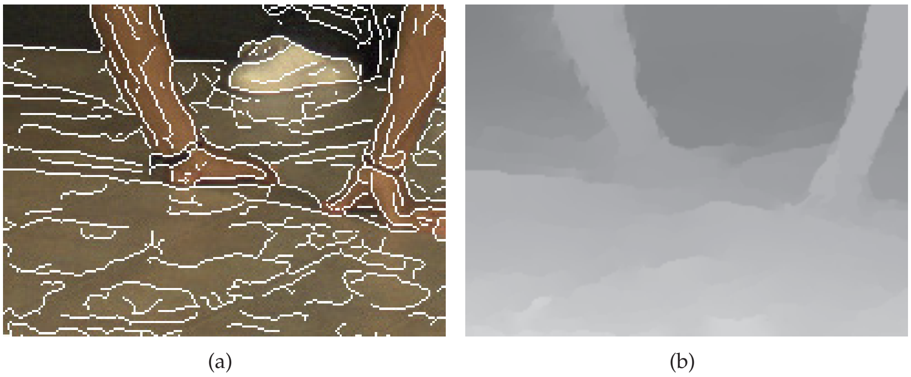
Thorough edge detection on a video frame will result in many more edges than there are actual objects. Many edges do not comply with actual depth transitions and will lead to an unwanted structurization effect in the upscaled depth map as shown in Fig. 4.4. A higher threshold for the edge detector reduces the amount of unnecessary edges, however it increases the risk of ”depth leakage”. Finding the correct edge threshold for each sequence is difficult and often impossible. Therefore it is more practical to use a lower edge detector threshold and validate the resulting edge map with actual depth transitions.

With the original low resolution depth available, it is possible to reduce depth structurization artifacts without increasing the risk of ”depth leakage”. The idea is, that only edges in the depth map correspond to object boundaries. To remove redundant edges in areas with uniform depth, another Canny edge detector is applied



(a) (b)

Figure 4.3: Depth leakage in 'Ballet' [ZKU$^+$04]: Porous edges (white) in (a) can lead to depth leakage marked in (b).



(a) (b)

Figure 4.4: Depth structurization in 'Breakdancing' [ZKU$^+$04]: Too many edges (white) in (a) can lead to depth structurization seen in (b).

on the low resolution depth map $\mathbf{D}_{low}$. The resulting edge map $\mathbf{E}_D$ is upscaled to the target resolution and used to mask out unnecessary edges in $\mathbf{E}_I$, giving the new weighting function W_E:

$$W_E(x,y) = 1 - E_I(x,y) \cdot E_D(x,y) \tag{4.11}$$

4.3 Multistep Upscaling

Typical ToF upscaling scenarios require upscaling factors of 8 or higher, in the horizontal and vertical directions respectively. While it is possible to upscale depth by means of EWOC in one step, an incremental approach can increase the upscaling performance. Multistep upscaling has already been presented to reduce computational complexity for JBU in [RGBB09]. For EWOC it also offers a gain in upscaling quality: As mentioned in Sec. 4.2.3, a major factor for the upscaling quality is the edge detection from the video frame. Downscaled texture versions can deliver more coherent edge maps, preventing erroneous depth values from spreading too far in the consecutive upscaling steps. Fig. 4.5 shows the concept for incremental EWOC depth upscaling with a factor of 8 in three steps. The downsampled texture versions are gained using standard bilinear filtering. The edge weight block combines both edge weighting and validation.

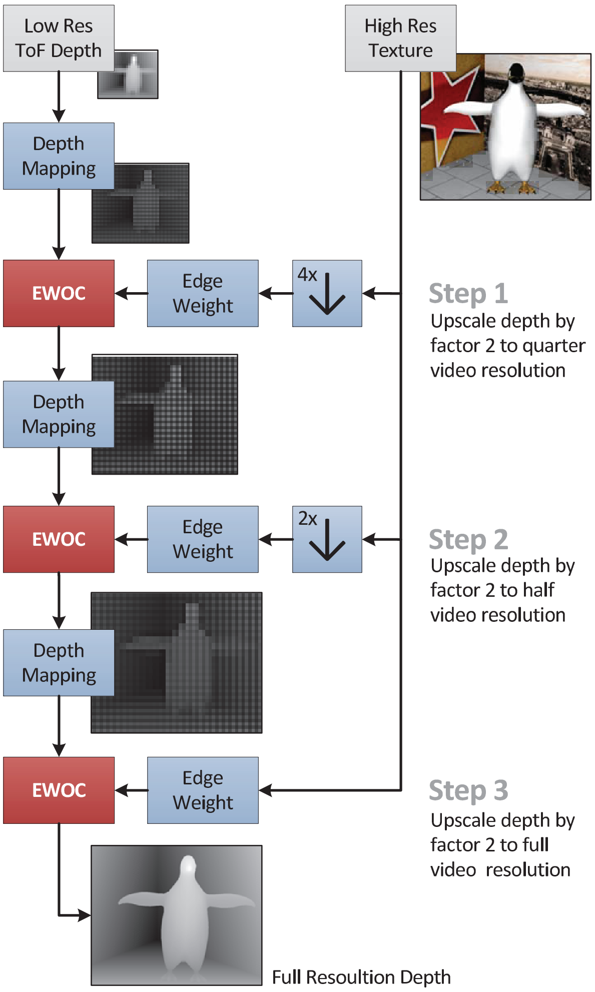


Figure 4.5: Different steps of incremental EWOC depth upscaling for a typical ToF depth upscaling by factor 8. Images not according to scale.

EWOC depth upscaling was introduced in [SSO12a] (Paper I) and an improved edge detection was presented in [SSO12b] (Paper III). The multistep enhancement was presented in [SSO12c] (Paper IV). The next chapter pools the separate contributions of this thesis.

Chapter 5

Contributions

This chapter presents a condensed analysis of the four papers included in this thesis. After the concepts behind EWOC depth upscaling were addressed in Chapter 4, the focus now lies on the novelty of each paper and its evaluation.

The addressed contributions are:

- Paper I: ”Depth map upscaling through edge weighted optimization”.
- Paper II: ”Adaptive depth filtering for HEVC 3D video coding”.
- Paper III: ”Improved edge detection for EWOC depth upscaling”.
- Paper IV: ”Incremental depth upscaling using an edge weighted optimization concept”.

5.1 Paper I: Introducing EWOC

The first paper introduced EWOC for depth map upscaling, fusing low resolution depth maps with high resolution texture frames, and evaluated it against a number of competing proposals [SSO12a].

Instead of widely-used ideas such as JBU or MRF, the upscaling process is considered as an energy minimization problem. By combining edge detectors on texture and by cross-verifying the detector results with low resolution depth, upscaling artifacts were reduced. Objective tests verify EWOC as a valid addition to the wide field of depth upscaling approaches with improvement to previous proposals.

5.1.1 Novelty

There are three major novelties in this contribution: Firstly, while many depth upscaling proposals are only extensions to existing solutions such as JBU, the EWOC transfers a stereo-extraction approach [GWCO09] to a depth map upscaling scenario. Secondly, in a similar manner to other approaches, EWOC uses information from a texture frame, but further validates this information with the low resolution depth map. Thirdly, while the stereo extraction approach still relied on JBU for the final upscaling step, EWOC is capable of performing the whole upscaling in one optimization step.

5.1.2 Evaluation and Results

Since this was the introductory paper for EWOC, an extensive evaluation was performed. Two test scenarios were presented: The first investigated into depth distortions introduced by the upscaling process and the second dealt with the actual view synthesis quality using the upscaled depth.

For the first test, the Middlebury Stereo Vision data sets were used [SS02]. The sets provide high quality depth maps as a reference, so the low resolution depth inputs can be easily achieved by subsampling the provided depth maps. Further, these data sets are widely used in the scientific community and this thus enables there to be an easy comparison between different proposals. Upscaling factors of 2, 4 and 8 were compared to the full resolution reference based on the mean square error (MSE) between pixel values. The results in Fig. 5.1 show that EWOC performs especially well for higher upscaling factors, outperforming other proposals by a factor of 2 or higher in MSE. The benefits of EWOC depth upscaling, particularly the more accurate depth transitions, are shown in Fig. 5.2.

In the second test, VSRS virtual views were generated based on depth maps upscaled by a factor of 8 using EWOC, JBU and the combined ”2-step” approach from [GWCO09]. Again, low resolution depth was generated by subsampling provided depth maps. To remove possible synthesis artifacts from the evaluation, the results were then compared with a reference synthesis based on the original depth maps.

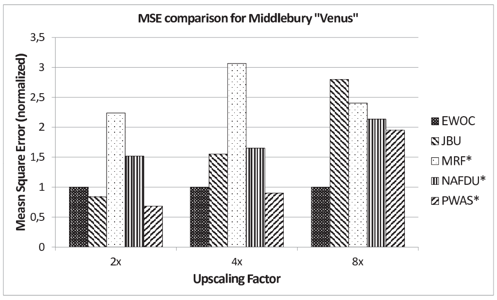


Figure 5.1: Comparison on depth map MSE between several approaches for the Middlebury "Venus" set. Note that values marked with "*" are taken from [GMO+10].

The evaluation criteria applied were PSNR, as a standard in image quality assessment, and SSIM, with a more close resemblance to the HVS experience. Fig. 5.3, shows the objective results for sequence "Street" by the Poznań University of Technology [DGK+09]. Corresponding view synthesis examples are shown in 5.4, together with the difference compared to syntheses with the original depth maps. The increase in visual quality is explicitly visible here around the side-view mirror of the car in front.

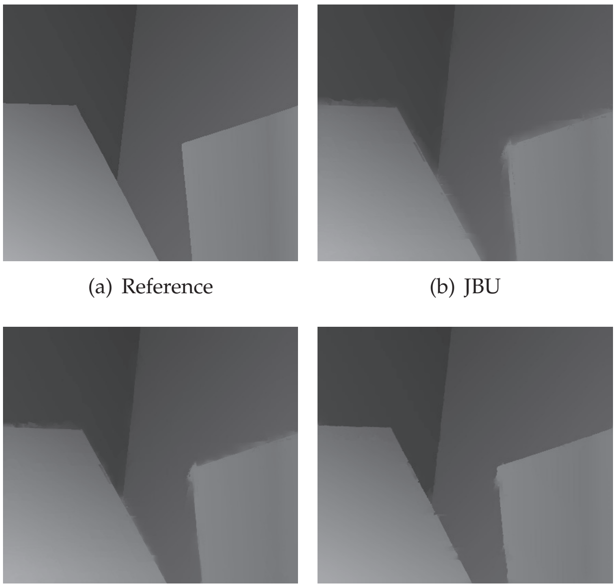
5.2 Paper II: MVD Coding

While the previous paper saw EWOC more aimed at ToF depth upscaling, this paper investigated possible applications in 3D video coding [SOST12].

The transmission of spatially downscaled depth maps is a common idea for MVD coding [KWD09]. It was shown that EWOC is a valid upscaling approach at the decoder side, even if the corresponding video sequences are highly compressed and guidance edges are harder to find. In addition, the special characteristics of EWOC motivated an adaptive pre-filtering for depth maps, further increasing coding efficiency.

5.2.1 Novelty

The idea of transmitting downscaled depth is not new and was already presented in Sec. 3.2.2. This paper expands this compression idea with an adaptive low-pass filter to reduce high energy parts in the depth maps prior to subsampling and compression. Fig. 5.5 shows the proposed coding scheme. Video coding algorithms are very good in compressing uniform areas by removing redundancy, and depth maps consist of large areas with redundant information. It is possible to increase redundancy further, as long as the transitions at object boundaries stay untouched. Based on the blur map $\mathbf{E}_{blur}$ from a edge detector on the texture frame $\mathbf{I}$, all parts not corresponding to object transitions are smoothed out. The original depth map $\mathbf{D}$ is convoluted



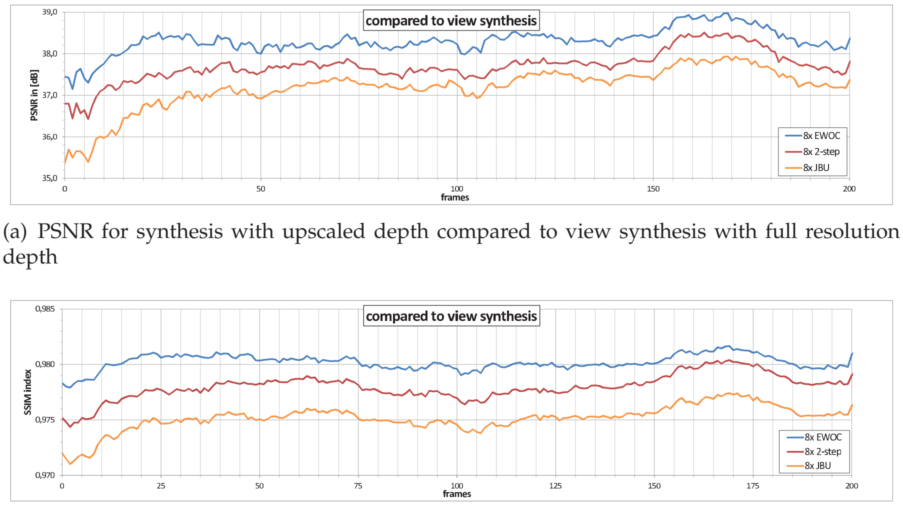
(a) Reference

(b) JBU

(c) PWAS

(d) EWOC

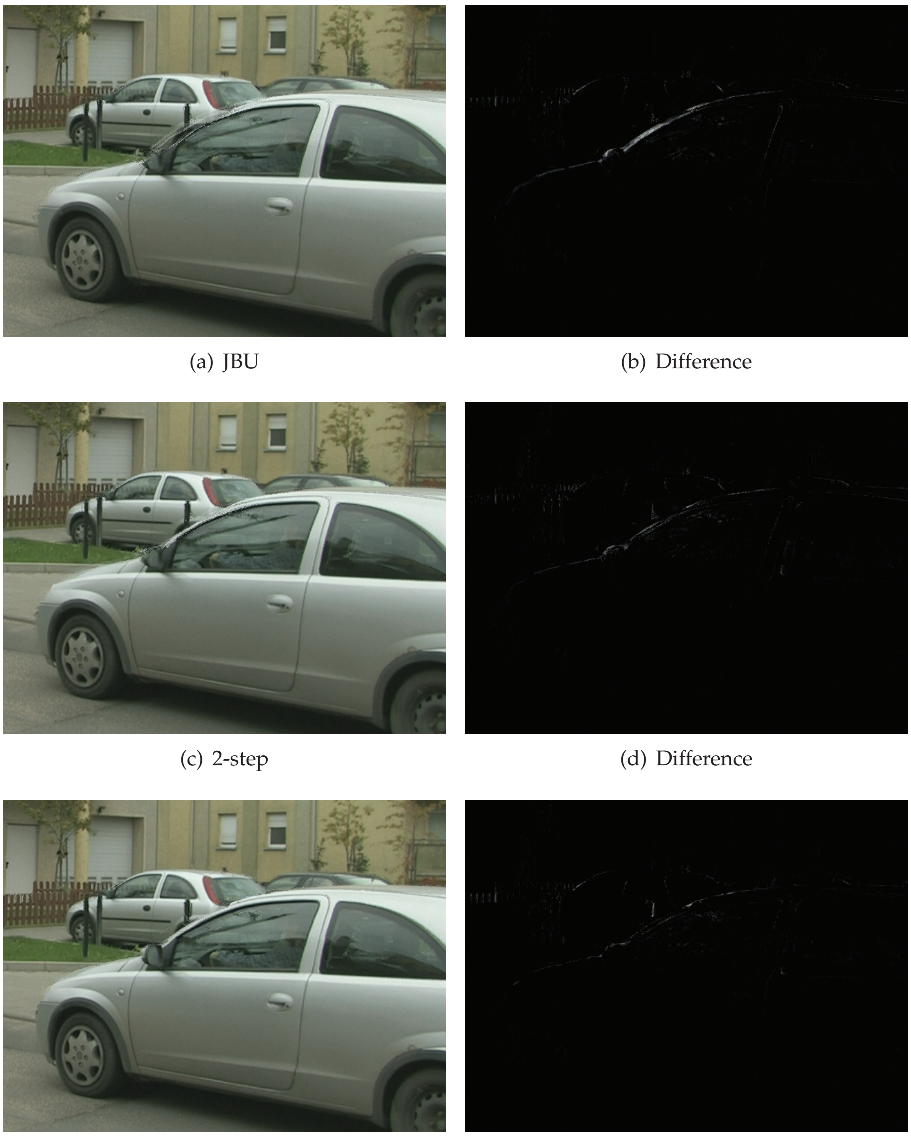
Figure 5.2: Depth reference and upscaling examples from Fig. 5.1 for 8x upscaling.



(a) PSNR for synthesis with upscaled depth compared to view synthesis with full resolution depth

(b) SSIM index for synthesis with upscaled depth compared to view synthesis with full resolution depth

Figure 5.3: PSNR & SSIM index comparison for view 4 of test sequence ”Poznań Street” in Paper I. Upscaling factor 8.



(a) JBU

(b) Difference

(c) 2-step

(d) Difference

(e) EWOC

(f) Difference

Figure 5.4: Details of view synthesis with depth upscaling with different approaches by factor 8: Left column shows the results using upscaled depth maps, right column shows the differences to syntheses with original depth map. Frame 1 of test sequence ”Poznań Street”.

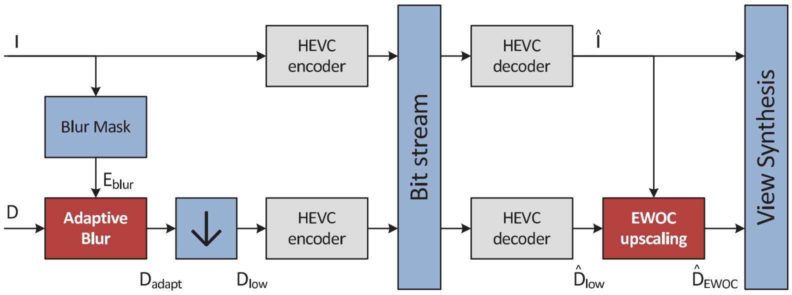


Figure 5.5: Proposed depth coding scheme in Paper II.

with a Gaussian kernel **G** and multiplied with the blur map $\mathbf{E}_{blur}$. The original depth map **D** is multiplied with the inverted blur map ($\mathbf{E}_{blur}$ subtracted from the identity matrix **1**). Both multiplications are element-wise and represented by the symbol "$\otimes$". The two depth maps are added together and form the adaptively filtered depth map $\mathbf{D}_{adapt}$.

$$\mathbf{D}_{adapt} = (\mathbf{D} * \mathbf{G}) \otimes \mathbf{E}_{blur} + \mathbf{D} \otimes (\mathbf{1} - \mathbf{E}_{blur}) \qquad (5.1)$$

Then $\mathbf{D}_{adapt}$ is downscaled by factor 2, encoded using High Efficiency Video Coding (HEVC, [JV11b]), decoded and upscaled using EWOC.

5.2.2 Evaluation and Results

The idea for this paper was highly motivated by the MPEG call for proposals (CfP) on 3D video coding technology [mpe11]. Therefore the evaluation can be related to the requirements given in the CfP. The test sequences used were "Poznań Street" and "Poznań Hall2". Video and depth sequences were coded using the HEVC test model HM-4.0 [JV11a], with a group of pictures (GOP) size of 12 and clean decoding refresh (CDR) for random access points every 0.5 seconds. Virtual views for evaluation were synthesized using VSRS.

In a first test, the feasibility of the adaptive filter was assessed. A comparison between three different approaches was performed: The proposed adaptive filter approach with HEVC encoded, smoothed and downscaled depth maps, upscaled with EWOC at the receiver's side, HEVC encoded downscaled depth maps, upscaled with JBU, and HEVC encoded full-scale depth maps. The different depth map sequences were coded with QPs of 16 to 44 with a constant video encoding at a quantization parameter (QP) of 32. The rate-distortion curves for "Poznań Street" and "Poznań Hall2" are shown in Fig. 5.6 and favor the proposed approach at low depth bit rates. A second test evaluated the processing for MVD coding at the MPEG CfP bit rate anchors. Fig. 5.7 shows the resulting rate-distortion curves, while Fig. 5.8 shows the result of a subjective evaluation with 20 test subjects. Again the proposed approach was favored.

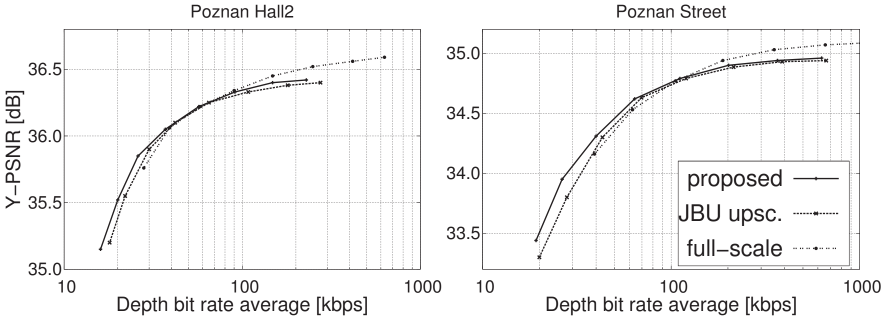


Figure 5.6: Rate-distortion curves at different depth compressions.

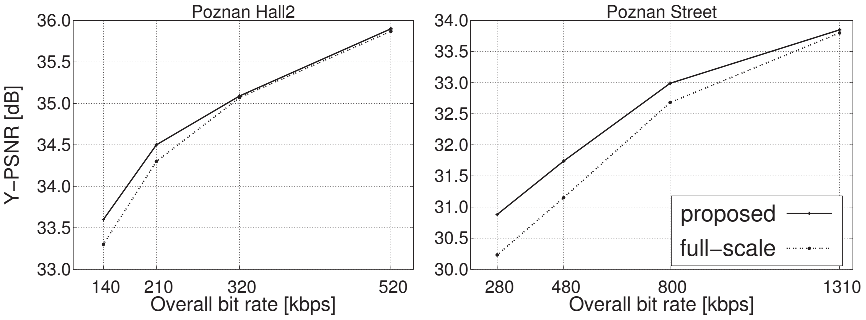


Figure 5.7: Rate-distortion curves at MPEG CfP bit rate anchors [mpe11].

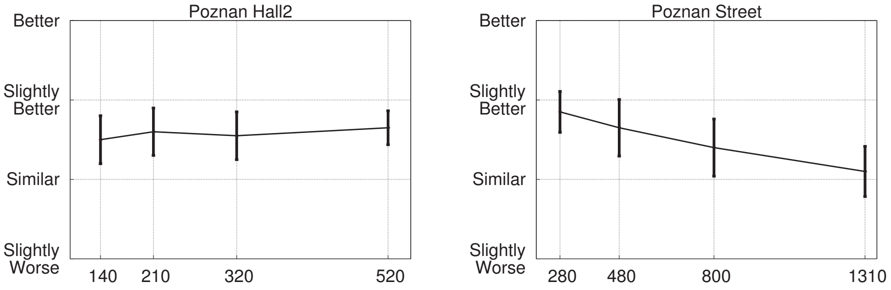


Figure 5.8: Mean opinion score (MOS) and 95% confidence intervals comparing sequences from Fig. 5.7.

5.3 Paper III: Improved Edge Detection

The third paper concentrated on the edge weight extraction from the corresponding texture frame. Different combinations of color spaces and pre-processing steps were evaluated [SSO12b].

The outcome of EWOC is highly dependent on accurate edge detection, both in texture and depth maps. Investigating several sources and pre-processing steps showed that intensive edge detection can increase the upscaling quality. However, this gain is achieved at the cost of computational complexity. Therefore the ideas for intensive edge detection were not continued for the sake of a future real-time implementation of EWOC. However, they can be easily integrated for scenarios requiring especially high depth quality with relaxed time restrictions.

5.3.1 Novelty

In the original EWOC proposal, the edge weight was generated from a combination of different edge detectors and color spaces and had a continuous value range of [0, 1]. This paper showed that a combination of sources based on the HVS can slightly improve the upscaling results. On the other hand, such improvements are only achieved at the cost of high computational complexity.

5.3.2 Evaluation and Results

Similar to Paper I, the evaluation was based on different test sequences and view synthesis with VSRS. Nine different sources for texture edge detectors were examined: Single channel grayscale and luminance, channel combinations for the RGB, YUV and HSV color spaces, as well as mean-shift filtered [CM02] versions of the luminance and the RGB color space, finally the CIE2000 color difference, which is supposed to resemble the human color perception [LCR01]. The results in Tab. 5.1 show that a combination of edges from luminance and CIE2000 color difference provided the best view synthesis quality, but increased the processing time on edge detection by a factor of 150 as compared to a simple Canny edge detector on the luminance channel only. The quality difference between the approaches did not provide the motivation for such an increase in complexity. It was also shown that mean-shift filtering, although a popular approach in image segmentation, is not feasible in this application. Based on the findings in this paper, the ideas of color space combinations and a continuous value range were not continued. Instead of the intended increase in quality, a decrease in complexity was found.

5.4 Paper IV: Incremental Upscaling

The last paper included in this thesis introduced the incremental upscaling enhancement presented in Sec. 4.3 [SSO12c].

Table 5.1: Comparison of different texture edge detector sources in Paper III. Mean PSNR and SSIM for 20 frames of test sequence “Poznań Street”.

Texture information	PSNR [dB]	SSIM
Graylevel	38.230	0.980
Luminance (Y)	38.234	0.980
RGB	37.267	0.978
YUV	37.843	0.979
HSV	37.843	0.979
CIE2000	38.078	0.980
meanshift Y	33.006	0.960
meanshift RGB	33.005	0.960
Y + CIE2000	38.437	0.980

Dividing the upscaling process into several steps decreases complexity and causes EWOC to be closer to real-time capability. Incremental upscaling allows for more coherent edges in lower resolutions, restricting the spread of erroneous depth in the following stages. It was also shown that incremental upscaling for JBU leads to a drop in quality, due to the lack of texture information in the early upscaling steps.

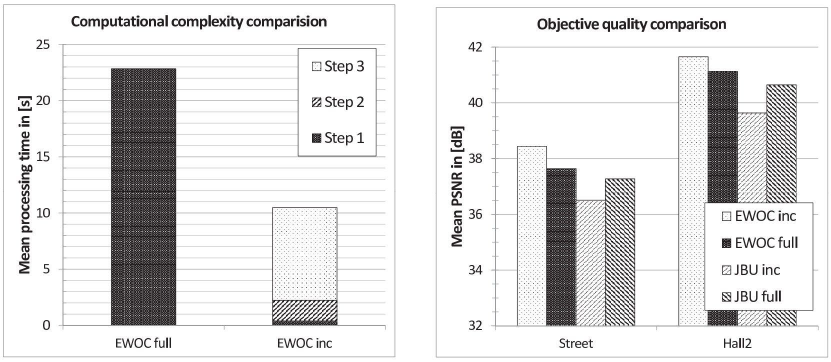
5.4.1 Novelty

The novelty for this paper lies in dividing up the upscaling process into several smaller steps instead of one big step. EWOC performance is improved by reducing computational complexity and increasing upscaling quality.

5.4.2 Evaluation and Results

Two factors were interesting for the evaluation of this approach. The first was the decrease in processing time compared to EWOC depth upscaling in one step, and the second was the increase in synthesis quality.

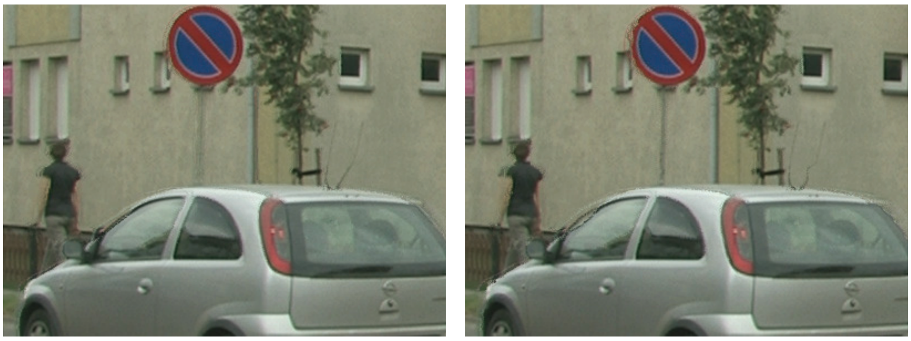
The first factor was assessed by a simple comparison of the new incremental implementation compared to a previous EWOC setup. It was possible to show that upscaling by a factor of 2 in three consecutive steps reduces the mean processing time per frame to less than half, compared to a single upscaling step by a factor of 8 (Fig. 5.9 (a)). For the second factor, the increase of view synthesis quality, again the same evaluation methodology from Paper I was applied. Comparison partners were the incremental JBU approach presented in [RGBB09], as well as full upscaling in a single step using EWOC and JBU. The objective results are presented in Fig. 5.9 (b). Incremental EWOC depth upscaling provides a difference of about 1dB in PSNR as compared to the second best approach, full EWOC. Two synthesis results, representing the difference between the incremental and full EWOC depth upscaling, are shown in 5.10. The gain in quality for incremental EWOC is especially visible at the traffic sign and the side-view mirror.



(a) Mean processing time

(b) Mean PSNR

Figure 5.9: Processing time comparison (a) over both sequences (400 frames) between single-step EWOC and the proposed incremental implementation. Mean PSNR (b) for ”Street” and ”Hall2” (200 frames each) with upscaled depth from different approaches.



(a) EWOC inc

(b) EWOC full

Figure 5.10: Details of view synthesis with depth upscaling factor 8 for incremental and single-step EWOC. Frame 40 of test sequence ”Poznań Street”.

It is interesting to note that the incremental upscaling with JBU leads to a quality loss. It is assumed that this drop in quality occurs at the lower upscaling steps, where a lot of filter information is missing due to the downscaled texture frame. These errors are then inherited to the higher upscaling steps.

Chapter 6

Conclusions

The previous chapter provided a condensed summary of the contributions of this thesis. It pointed out the novelty for each paper, recapped the evaluation process and presented important results. Starting from the overall aim, this final chapter offers a comprehensive conclusion in relation to all four publications, covering the whole aspect of EWOC depth upscaling. The compliance to the given goals is analyzed and the contribution of this work to the research community is presented. Finally, this work will conclude with an outlook into future work: Enhancements, applications and possible new related fields of research.

6.1 Overview

The work presented here set out to achieve a better 3DTV experience. In this context, improved depth map capture and encoding are crucial for the whole 3DTV distribution chain.

A new algorithm for texture guided depth map upscaling was proposed: The Edge-Weighted Optimization Concept (EWOC). The algorithm occupies a new zone in the field of guided depth upscaling. While other approaches often rely on (joint-) bilateral filters or Markov Random fields, EWOC introduces the idea of energy minimization to depth map upscaling. With simple edge information from texture, validated by the low resolution depth and a step-by-step process, convincing upscaling results are achieved. EWOC outperforms popular approaches in terms of objective and subjective visual quality, in both, depth maps as well as DIBR view synthesis.

EWOC can be utilized for depth map upscaling in scenarios in which the spatial resolutions of the texture frame and the depth map do not match. The application of EWOC allows for better 3D content creation using dedicated range sensors, i.e. ToF cameras, and more efficient 3D video coding with asymmetric depth compression. Thereby the introduction of EWOC fulfills the aim of this research for a better 3DTV experience. The following section will discuss the achievements of this thesis in more detail.

6.2 Outcome

Sec. 1.4 defined three concrete and verifiable goals in order to validate the outcome of the presented research. This section discusses the results for each goal with the respect to the overall aim for an improved 3DTV experience.

- **Goal I:** *"Investigate the upscaling of limited range information, utilizing additionally available data to improve the 3DTV quality, and propose an alternative concept to depth map upscaling."*
  Paper I investigated into texture guided depth map upscaling and introduced EWOC as an alternative to established concepts. EWOC interprets the low resolution depth map as a sparse representation of the target resolution depth map. Missing depth values are filled by edge weighted optimization. The necessary edge weights are taken from the corresponding texture frame. Unlike other guided depth upscaling solutions, EWOC additionally validates this texture edge information with the low resolution depth map for object consistency. This validation leads to improved depth upscaling results. Objective evaluations show increased quality in upscaled depth maps and resulting DIBR view syntheses, compared to competing proposals.

- **Goal II:** *"Investigate the utility of depth map upscaling for 3D video coding and propose an alternative compression scheme for 3DTV distribution utilizing depth map upscaling."*

In Paper II the idea of EWOC depth upscaling was applied to depth upscaling for 3D video coding, together with a new adaptive filtering concept. Subjective and objective evaluations show improvements compared to state-of-the-art HEVC video coding, especially at lower bit rates. The proposed depth coding scheme leads to a higher coding efficiency in the 3DTV distribution chain.

- **Goal III:** *"Investigate the relationship between the visual quality and computational complexity for depth map upscaling and propose enhancements to the introduced depth map upscaling concept."*
  Paper III and IV addressed the trade off between complexity and quality. The former paper investigated different sources for texture and depth alignment and proposed a simplified edge detection with higher real-time potential. Paper IV improved visual quality and reduced computational complexity with an incremental approach to EWOC depth upscaling. Together, the enhancements introduced in these two publications lead to an improved relationship between visual quality and computational complexity.

With regards to these three goals, the combination of all four publications fulfills the set aim of this work. Objective and subjective evaluation proves that the introduction of EWOC indeed improves the quality of the 3DTV experience.

6.3 Impact

The introduction of EWOC allows for improved visual quality in all scenarios related to low resolution depth. The research presented here leads to more efficient 3D video coding for autostereoscopic displays, allowing for a more efficient 3DTV distribution chain. EWOC also opens up the potential for full-scale scene depth capture with ToF cameras without restrictions of stereo capture. Decreasing the capture complexity and increasing the depth map quality is supposed to have a major impact on the content creation for 3DTV, lowering production and post-processing costs. More available content and improved distribution will lead to a wider variety and better quality of offered 3DTV services and will finally result in the commercial success of 3DTV.

Furthermore EWOC is not only limited to 3DTV applications. ToF cameras are also applied in different scenarios, often already combined with video cameras. Introducing EWOC can increase the spatial resolution of captured depth and can lead to improvements in many different areas such as manufacturing, product quality control, security and surveillance.

6.4 Future Research

As the name EWOC, or "Edge-Weighted Optimization Concept", states, the current implementation merely utilizes edges in the upscaling process. However, additional weights can be easily integrated. Possible ideas for such weights are:

- Temporal weights: Either for smoothing out depth flickering over time or allowing temporal subsampling in 3D video coding.
- Edge credibility weights: Defining certain edges as more reliable than others.
- Depth credibility weights: Defining certain depth values as more reliable than others.
- Iterative weights: Refining depth values in an iterative process.

Depth credibility is particularly interesting for ToF upscaling, where additional information about the received signal can be used to determine the depth accuracy. This information should increase the upscaling quality and error robustness by removing outliers and reducing sensor noise.

Possible future applications, apart from video coding and ToF upscaling, could include full multiview capture from a single view point by combining a video plus range capture system with EWOC and DIBR. Also integrated solutions to improve the spatial resolution of ToF cameras in general are conceivable. As well as applying EWOC to upscale other depth sensors, e.g. the Microsoft Kinect system. Finally, EWOC upscaled depth maps could be utilized for real-time eye gaze correction in teleconferencing systems.

Bibliography

[Adl98] M. Adlers. *Sparse Least Squares Problems with Box Constraints*. Licentiat thesis, Department of Mathematics, Linköping University, Linköping, Sweden, 1998.

[AM05] N. Atzpadin and J. Mulligan. Stereo analysis. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 115–132. John Wiley & Sons, 2005.

[AYG$^+$07] A.A. Alatan, Y. Yemez, U. Gudukbay, X. Zabulis, K. Muller, C.E. Erdem, C. Weigel, and A. Smolic. Scene representation technologies for 3DTV - a survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11):1587–1605, November 2007.

[BOL$^+$05] B. Büttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger. CCD/CMOS lock-in pixel for range imaging : Challenges, limitations and state-of-the-art. *Measurement*, 103, 2005.

[BVZ01] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, November 2001.

[Can86] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8:679–698, November 1986.

[CBTT08] D. Chan, H. Buisman, C. Theobalt, and S. Thrun. A noiseaware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[CKH12] D. Herrera C., J. Kannala, and J. Heikkila. Joint depth and color camera calibration with distortion correction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34:2058–2064, 2012.

[CLK$^+$12] O. Choi, H. Lim, B. Kang, Y. S. Kim, K. Lee, J. D. K. Kim, and C.-Y. Kim. Discrete and continuous optimizations for depth image super-resolution. In *Proceedings of the SPIE, vol 8290: Conference on 3D Image Processing (3DIP) and Applications*, 2012.

[CM02] D. Comaniciu and P. Meer. Mean shift: a robust approach toward feature space analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(5):603–619, May 2002.

[CV95] J. E. Cutting and P. M. Vishton. Perceiving layout and knowing distances: the integration, relative potency and contextual use of different information about depth. In W. Epstein and S. Rogers, editors, *Handbook of perception and Cognition.*, volume 5: Perception of Space and Motion, pages 69–117. Academic Press, San Diego, CA, USA, 1995.

[DGK$^+$09] M. Domañski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, and K. Wegner. Poznań multiview video test sequences and camera parameters. ISO/IEC JTC1/SC29/WG11 MPEG2009/M17050, October 2009. Xian, China.

[DRE$^+$11] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel. A perceptual model for disparity. *ACM Transactions on Graphics (Proc. of SIGGRAPH)*, 30(4), 2011.

[DT05] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *Proceedings of Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2005. MIT Press.

[EMWK09] E. Ekmekcioglu, M. Mrak, S. Worrall, and A. Kondoz. Utilisation of edge adaptive upsampling in compression of depth map videos for enhanced free-viewpoint rendering. In *Image Processing (ICIP), 2009 16th IEEE International Conference on*, pages 733–736, 2009.

[Feh05] C. Fehn. 3D TV broadcasting. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 23–38. John Wiley & Sons, 2005.

[FKB$^+$02] C. Fehn, P. Kauff, M. Op De Beeck, F. Ernst, W. A. IJsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, and I. Sexton. An evolutionary and optimised approach on 3D-TV. In *Proceedings of International Broadcast Conference (IBC)*, pages 357–365, 2002.

[Fot12] Fotonic. C70 time-of-flight camera. [online] `http://www.fotonic.com/assets/documents/fotonic_c70_highres.pdf`, August 2012.

[FPR$^+$09] M. Frank, M. Plaue, H. Rapp, U. Köthe, B. Jähne, and F. A. Hamprecht. Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras. *Optical Engineering*, 48(1), 2009.

[GAC$^+$11] A. Gotchev, G.B. Akar, T. Capin, D. Strohmeier, and A. Boev. Three-dimensional media for mobile devices. *Proceedings of the IEEE*, 99(4):708–741, April 2011.

[GMO$^+$10] F. Garcia, B. Mirbach, B. Ottersten, F. Grandidier, and A. Cuesta. Pixel weighted average strategy for depth sensor data fusion. In *IEEE 17th International Conference on Image Processing*, 2010.

[GS05] J.G.M. Gonçalves and V. Sequeira. Sensor-based depth capturing. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 299–314. John Wiley & Sons, 2005.

[GWCO09] M. Guttmann, L. Wolf, and D. Cohen-Or. Semi-automatic stereo extraction from video footage. In *IEEE 12th International Conference on Computer Vision*, 2009.

[HZ03] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2003.

[IJs03] W. A. IJsselsteijn. Presence in the past: What can we learn from media history. In G. Riva, F. Davide, and W. A. IJsselsteijn, editors, *Being there: Concepts, Effects and Measurement of User Presence in Synthetic Enviroments*, pages 18–40. Ios Press, 2003.

[IO90] T. Ionoue and H. Ohzu. Accommodation and convergence when looking at binocular 3D images. In K. Noro and O. Brown, editors, *Human Factors in Organizational Design and Management - III*, pages 249–252. Elsevier Science Inc., New York, NY, USA, 1990.

[ISM05] W. A. IJsselsteijn, P.J.H. Seuntiës, and L.M.J. Meesters. Human factors of 3D displays. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 219–233. John Wiley & Sons, 2005.

[JV11a] JCT-VC. HM4: High efficiency video coding HEVC test model 4. [online] `https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-4.0/`, 2011.

[JV11b] JCT-VC. HM4: High efficiency video coding (HEVC) test model 4 encoder description. JCTVC-F802, July 2011. Torino, Italy.

[KAF$^+$07] P. Kauff, N. Atzpadin, C. Fehn, M. Müller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability. *Signal Processing: Image Communication*, 22:217–234, February 2007.

[KCLU07] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM Transactions on Graphics*, 26(3), 2007.

[KES05] R. Koch and J.-F. Evers-Senne. View synthesis and rendering methods. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 235–260. John Wiley & Sons, 2005.

[KI06] T. Kahlmann and H. Ingensand. Calibration of the fast range imaging camera swissranger for use in the surveillance of the environment. *Electro-Optical Remote Sensing II*, 6396(1), 2006.

[KS05] P. Kauff and O. Schreer. Immersive videoconferencing. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 75–114. John Wiley & Sons, 2005.

[KWD09] K. Klimaszewski, K. Wegner, and M. Domañski. Influence of views and depth compression onto quality of synthesized views. ISO/IEC JTC1/SC29/WG11 MPEG2009/M16758, July 2009. London, UK.

[KYY11] C. Kim, H. Yu, and G. Yang. Depth super resolution using bilateral filter. In *Image and Signal Processing (CISP), 2011 4th International Congress on*, volume 2, pages 1067–1071, 2011.

[LCR01] M. R. Luo, G. Cui, and B. Rigg. The development of the CIE2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5):340–350, 2001.

[LFUS06] D. Lischinski, Z. Farbman, M. Uyttendaele, and R. Szeliski. Interactive local adjustment of tonal values. *ACM Transactions on Graphics*, 25(3):646–653, July 2006.

[LLW04] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. *ACM Transactions on Graphics*, 23(3):689–694, August 2004.

[LS01] R. Lange and P. Seitz. Solid-state time-of-flight range camera. *IEEE Journal of Quantum Electronics*, 37:390–397, March 2001.

[LS10] Y. Li and L. Sun. A novel upsampling scheme for depth map compression in 3DTV system. In *Picture Coding Symposium (PCS), 2010*, pages 186–189, 2010.

[McM97] L. McMillan. *An Image-Based Approach to Three-Dimensional Computer Graphics*. PhD thesis, University of North Carolina, Chapel Hill, NC, USA, 1997.

[mpe11] Call for proposals on 3D video coding technology. ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, March 2011. Geneva, Switzerland.

[NBM$^+$12] M.R. Neuman, G.D. Baura, S. Meldrum, O. Soykan, M.E. Valentinuzzi, R.S. Leder, S. Micera, and Yuan-Ting Zhang. Advances in medical devices and medical electronics. *Proceedings of the IEEE*, 100(Special Centennial Issue):1537–1550, 2012.

[Onu11] L. Onural. *3D video technologies: An overview in research trends*. SPIE, 2011.

[Pas05] S. Pastoorm. 3d displays. In O. Schreer, P. Kauff, and T. Sikora, editors, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, pages 235–260. John Wiley & Sons, 2005.

[PJV94] L. F. Portugal, J. J. Júdice, and L. N. Vicente. A comparison of block pivoting and interior-point algorithms for linear least squares problems with nonnegative variables. *Mathematics of Computation*, 63:625–643, October 1994.

[PKT$^+$11] J. Park, H. Kim, Y.-W. Tai, M. S. Brown, and I. Kweon. High quality depth map upsampling for 3D-ToF cameras. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1623–1630, 2011.

[Pro12] 3-D Revolution Productions. 3-D film list: The most comprehensive, concise and up-to-date 3-D filmography on the web. [online] `http://www.the3drevolution.com/3dlist.html`, July 2012.

[RBV$^+$06] A. Redert, R.-P. Berretty, C. Varekamp, O. Willemsen, J. Swillens, and H. Driessen. Philips 3d solutions: From content creation to visualization. In *3D Data Processing, Visualization, and Transmission, Third International Symposium on*, pages 429–431, 2006.

[RdBF$^+$02] A. Redert, M.O. de Beeck, C. Fehn, W. A. Ijsselsteijn, M. Pollefeys, L. Van Gool, E. Ofek, I. Sexton, and P. Surman. Advanced three-dimensional television system technologies. In *3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on*, pages 313–319, 2002.

[RGBB09] A. K. Riemens, O. P. Gangwal, B. Barenbrug, and R.-P. M. Berretty. Multistep joint bilateral depth upsampling. *Visual Communications and Image Processing 2009*, 7257(1):72570M, 2009.

[SB12] D. Scharstein and A. Blasiak. Middlebury stereo evaluation - version 2. [online] `http://vision.middlebury.edu/stereo/eval/`, August 2012.

[Sch99] D. Scharstein. View synthesis using stereo vision. In G. Goos, J. Hartmanis, and J. van Leeuwen, editors, *Lecture notes in computer science*, volume 1583. Springer, 1999.

[Sed08] H. A. Sedgwick. Visual space perception. In Goldstein E., editor, *Blackwell Handbook of Sensation and Perception*, pages 128–167. Blackwell Publishing Ltd, 2008.

[SMM$^+$09] A. Smolic, K. Mueller, P. Merkle, P. Kauff, and T. Wiegand. An overview of available and emerging 3d video formats and depth enhanced stereo as efficient generic solution. In *Picture Coding Symposium, 2009. PCS 2009*, pages 1 –4, 2009.

[SMS$^+$07] A. Smolic, K. Müller, N. Stefanoski, J. Ostermann, A. Gotchev, G.B. Akar, G. Triantafyllidis, and A. Koz. Coding algorithms for 3DTV - a survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(11):1606–1621, 2007.

[SOST12] S. Schwarz, R. Olsson, M. Sjöström, and S. Tourancheau. Adaptive depth filtering for HEVC 3D video coding. In *Picture Coding Symposium (PCS), 2012*, pages 49–52, 2012.

[SS02] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1-3):7–42, 2002.

[SSO12a] S. Schwarz, M. Sjöström, and R. Olsson. Depth map upscaling through edge weighted optimization. In *Proceedings of the SPIE, vol 8290: Conference on 3D Image Processing (3DIP) and Applications*, 2012.

[SSO12b] S. Schwarz, M. Sjöström, and R. Olsson. Improved edge detection for EWOC depth upscaling. In *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on*, 2012.

[SSO12c] S. Schwarz, M. Sjöström, and R. Olsson. Incremental depth upscaling using an edge weighted optimization concept. In *Proceedings of 3DTV-Conference*, 2012.

[TM98] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Computer Vision, 1998. Sixth International Conference on*, pages 839–846, 1998.

[UCES11] H. Urey, K.V. Chellappan, E. Erden, and P. Surman. State of the art in stereoscopic and autostereoscopic displays. *Proceedings of the IEEE*, 99(4):540 –555, april 2011.

[vsr10] Report on experimental framework for 3D video coding. ISO/IEC JTC1/SC29/WG11 MPEG2010/N11631, October 2010. Guangzhou, China.

[WBSS04] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[Whe38] C. Wheatstone. On some remarkable, and hitherto unobserved, phenomena of binocular vision. *Philosophical Transactions*, 128:371–394, 1838.

[WYT$^+$10] M. O. Wildeboer, T. Yendo, M.P. Tehrani, T. Fujii, and M. Tanimoto. Color Based Depth Up-sampling for Depth Compression. In *Picture Coding Symposium (PCS)*, 2010.

[YA88] Y.F. Yang and J.K. Aggarwal. An overview of geometric modeling using active sensing. *Control Systems Magazine, IEEE*, 8(3):5–13, June 1988.

[Zha12] Z. Zhang. Microsoft kinect sensor and its effect. *Multimedia, IEEE*, 19(2):4–10, February 2012.

[ZKU$^+$04] L. C. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics*, 23(3), 2004.

[Zon07] R. Zone. *Stereoscopic Cinema and the Origins of 3-D Film*. The University Press of Kentucky, 2007.

Biography

Sebastian Schwarz was born on the 27th of July 1980 in Trostberg, Germany. In 2009 he received his engineering diploma (Dipl.-Ing.) in Media Technology at the Technical University of Ilmenau, Germany. After his graduation he started as a research fellow in the Audio-Visual Technology Department at the Institute for Media Technology of the Technical University Ilmenau. Here, Sebastian worked on camera calibration and drafting new lectures on 3D video and television. In 2010 he started his PhD studies at the department of Information Technology and Media at Mid Sweden University in Sundsvall, Sweden. Sebastian's main research interests are scene depth capture, time-of-flight cameras and texture guided depth map upscaling.