

University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

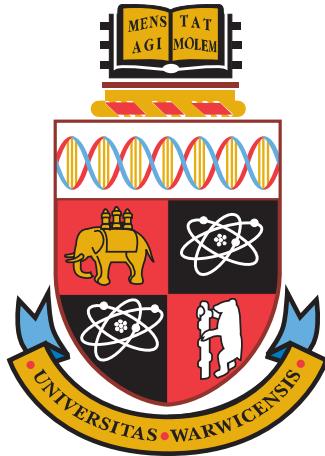
A Thesis Submitted for the Degree of PhD at the University of Warwick

<http://go.warwick.ac.uk/wrap/58125>

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.



Stereoscopic High Dynamic Range Imaging

Elmedin Selmanović
BSc (Hons)

*A thesis submitted for the degree of
Doctor of Philosophy in Engineering*

*School of Engineering
University of Warwick
2013*

Contents

Acknowledgements	x
Declaration	xi
Abstract	xii
1 Introduction	1
1.1 Digital Imaging Pipeline	1
1.1.1 Capture	2
1.1.2 Storage	3
1.1.3 Display	4
1.1.4 Other Techniques	4
1.2 Stereoscopic Imaging	4
1.3 High Dynamic Range Imaging	7
1.4 Research Objectives	10
1.5 Thesis Outline	11
2 Stereoscopic Imaging	13
2.1 A Brief History of Stereoscopy	14
2.2 Theory of Stereo Vision	18
2.2.1 Epipolar Geometry	18
2.2.2 Fundamental Matrix	20
2.2.3 Image Rectification	21
2.2.4 Disparity Maps	21
2.2.5 Stereo Correspondence Algorithms	25
2.3 Visual Discomfort in Stereo Vision	31
2.3.1 Individual Differences	31
2.3.2 Excessive Screen Disparity	32
2.3.3 Vergence - Accommodation Decoupling	32
2.3.4 Zone of Comfortable Viewing	33
2.3.5 Stereoscopic Impairments	34
2.4 Stereoscopic Capture	36
2.4.1 Capturing Two Symmetric Views	37

2.4.2	Asymmetric Stereoscopic Capture	39
2.4.3	Single View Plus Depth	44
2.4.4	Stereoscopic Rendering	48
2.5	Stereoscopic Content Storage	49
2.5.1	Animated Graphics Interchange Format (GIF)	50
2.5.2	Stereoscopic Portable Network Graphics Format (PNS) . .	51
2.5.3	Stereoscopic JPEG (JPS)	51
2.5.4	Multiview Video Format (MVC)	52
2.6	Stereoscopic Displays	53
2.6.1	Direct-View Stereoscopic Displays	53
2.6.2	Head-Mounted Displays	55
2.6.3	Autostereoscopic Displays	56
2.7	Summary	58
3	High Dynamic Range Imaging	59
3.1	Theory of High Dynamic Range Imaging	60
3.2	High Dynamic Range Capture	62
3.2.1	Multiple Exposures	63
3.2.2	Native HDR Capture	68
3.2.3	Expansion Operators (EOs)	71
3.3	High Dynamic Range Content Storage	78
3.3.1	File Formats	78
3.3.2	High Dynamic Range Image Coding	84
3.3.3	High Dynamic Range Video Coding	90
3.4	HDR Display	96
3.4.1	Native HDR Displays	96
3.4.2	Tone Mapping	101
3.5	Summary	106
4	Stereoscopic High Dynamic Range Pipeline	107
4.1	SHDR Capture	107
4.2	SHDR Storage	110
4.3	SHDR Display	111
4.4	Summary	112
5	Stereoscopic High Dynamic Range Images	113
5.1	LDR to HDR methods	114
5.1.1	Expansion Operators	116
5.1.2	Stereo Correspondence	117
5.2	The Experiment	121
5.2.1	Methodology	122
5.2.2	Participants	125
5.2.3	Materials	125
5.2.4	Procedure	127
5.3	Results	128

5.3.1	User Study Results	128
5.3.2	Objective Measures	131
5.3.3	Image Quality Comparison	132
5.4	Discussion	132
5.4.1	Limitations	136
5.5	Summary	137
6	Stereoscopic High Dynamic Range Video	138
6.1	LDR to HDR Methods	139
6.1.1	Stereo Correspondence	139
6.1.2	Expansion Operator	141
6.1.3	Hybrid Method	143
6.2	Results	147
6.2.1	Materials	147
6.2.2	Objective Quality Measurements	149
6.2.3	Qualitative Results	151
6.3	Discussion	151
6.3.1	Limitations	154
6.4	Summary	154
7	Stereoscopic High Dynamic Range Compression	159
7.1	JPEG SHDR Methods	160
7.1.1	Side-by-side (SBS) Method	161
7.1.2	Half Side-by-side (HSBS) Method	162
7.1.3	Image Plus Disparity (IPD) Method	163
7.1.4	Image Plus Disparity with Corrections (IPDC) Method	164
7.1.5	Motion Compensation (MC) Method	166
7.2	Results and Analysis	167
7.3	Summary	171
8	Conclusions	174
8.1	Capture of Stereoscopic High Dynamic Range Images	174
8.2	Capture of Stereoscopic High Dynamic Range Video	175
8.3	Compression of Stereoscopic High Dynamic Range Images	176
8.4	Contributions	178
8.5	Impact	178
8.6	Future Work	179
8.6.1	Extended User Studies	179
8.6.2	Additional Operators	180
8.6.3	SHDR Display	180
8.6.4	Beyond SHDR Imaging	181
8.7	Final Remarks	182
References		183

List of Figures

1.1	Digital Imaging Pipeline	2
1.2	Stereoscopic Imaging	5
1.3	High Dynamic Range Imaging	8
2.1	Wheatsone's Mirror Stereoscope	16
2.2	Random-dot Stereogram	17
2.3	Illustration of Stereo Capture Setup	19
2.4	Epipolar Geometry	20
2.5	Disparity Map	22
2.6	Keystone Distortion	35
2.7	Single Lens Stereo Camera Setup	38
2.8	Hybrid Stereoscopic Camera	39
2.9	Using Photographs to Enhance Videos	42
2.10	Kinect	45
2.11	Multiview Video Coding	52
2.12	Stereoscopic Glasses	54
2.13	Two-View Autostereoscopic Displays	57
3.1	Dynamic Range	61
3.2	Multiple Exposures	64
3.3	Multiple Exposure Video	67
3.4	HDR Video Camera	70
3.5	Banding Artefact	74
3.6	Overview of the Expand Map Algorithm	76
3.7	Radiance File Format	80
3.8	LogLuv File Format	81
3.9	OpenEXR File Format	83
3.10	JPEG-HDR Encoding Pipeline	85

3.11	JPEG-HDR Tone Mapped and Residual Images	86
3.12	HDR-MPEG Video Encoding Pipeline	91
3.13	Rate-Distortion Optimised HDR Video Encoding Pipeline	93
3.14	HDR Stereoscopic Viewer	97
3.15	SHDR Viewer Transparency Generation	98
3.16	HDR Projector Based Display	99
3.17	HDR LED Display	100
5.1	Expansion Operator SHDR Method for Images	114
5.2	Stereo Correspondence SHDR Method for Images	115
5.3	COGC Algorithm Artifacts	120
5.4	SHDR Images Used for Testing	126
5.5	Experimental Setup	127
5.6	Preferences of SHDR Methods for Images	130
5.7	Example Images Generated Using SHDR Methods	133
5.8	Comparison of Generated SHDR Images	135
6.1	Generation of SHDR video from HDR-LDR video pair	138
6.2	HDR-LDR Stereo Correspondence Pipeline	140
6.3	HDR-LDR Expansion Operator Pipeline	142
6.4	HDR-LDR Hybrid Pipeline	143
6.5	Disparity Map Generation for SHDR Video	144
6.6	Interpolated Disparity Map	145
6.7	SAD Frame Warping Artifacts	146
6.8	SHDR Video Scenes	148
6.9	SAD Video Method Artifacts	152
6.10	Example Frames Generated Using SHDR Video Methods	155
6.11	Quality Comparison of SHDR Video Frame	156
6.12	PSNR Results for SHDR Video	157
6.13	TQ Results for SHDR Video	158
7.1	SBS Encoding	161
7.2	SBS Decoding	161
7.3	HSBS Encoding	162
7.4	HSBS Decoding	162
7.5	IPD Encoding	163
7.6	IPD Decoding	163
7.7	IPDC Disparity Example	165
7.8	IPDC Disparity Generation	165

7.9	MC Encoding	166
7.10	MC Decoding	166
7.11	An Example Decoded Image	168
7.12	SHDR Scenes Used for Compression Evaluation	172
7.13	SHDR Scenes Used for Compression Evaluation Continued	173

List of Tables

5.1	Example Preference Table	123
5.2	Ranking of SHDR Method for Images and the Summary of Results	129
5.3	PSNR Results for SHDR Image Generation	131
5.4	RMSEL Results for SHDR Image Generation	132
6.1	SHDR Video Data	149
6.2	Summary of PSNR Results for SHDR Video	150
6.3	Summary of TQ Results for SHDR Video	151
7.1	SHDR Compression Compatibility	167
7.2	Sizes of Compressed SHDR Images	169
7.3	PSNR Results for SHDR Image Compression	170
7.4	RMSEL Results for SHDR Image Compression	171
7.5	Compression Results Summary	171

Acknowledgements

Firstly, I would like to thank my supervisors Alan and Kurt. Alan provided me with the opportunity to do a PhD, and his constant support, advice, encouragement and enthusiasm are sincerely appreciated. Kurt was a great mentor who patiently guided me on my PhD path and helped me avoid the many perils of research. Moreover, Kurt was also a true friend and ensured my stay in the UK was enjoyable.

Jasmina and Selma recommended me to Alan, and for this I am very grateful.

The members of the Visualisation Group at Warwick University have made the PhD experience exciting and interesting, both professionally but also as friends. Vedad made the transition from Bosnia to England much easier and set me on the right footing. Tom took the role of a third mentor and we had many useful discussions. Carlo, Jass and Vibhor provided a continuous support and offered helpful advice. It was great working with Alena, Alessandro, Ali, Belma, Jon, Josh, Keith, Louis Paulo, Martinho, Miguel, Piotr, Ratnajit, Remi, Sam, Sandro, Silvester, Stratos, Tim and Vasu.

I am also grateful to people outside the group who made life during the PhD fun: Ado, Adnan and Tarik, Alex, Anna, Ammar, Džidžo, Jo, Jason, Kate, Mike, Mirza, Mosh and Nedim. I would like to say thank you to everyone I have missed or not mentioned, as there were many others who helped me in numerous ways during the past four years.

My family was there for me throughout the whole time. My grandmas were especially inspiring . Majka Sena demonstrated how persistence and courage can overcome any obstacle while majka Zilha and me shared some of the toughest moments of our lives.

I would like to thank Silvija for providing extra motivation to finish the thesis and for being so loving, patient and supportive. Volim te!

My eternal gratitude goes to my parents Azra and Senad who always made sure I was on the right path and this thesis is the result of such efforts. Hvala!

Declaration

The work in this thesis is original and no portion of this work has been submitted in support of an application for another degree or qualification at this university or at another university or institution of learning.

Elmedin Selmanović

Abstract

Two modern technologies show promise to dramatically increase immersion in virtual environments. Stereoscopic imaging captures two images representing the views of both eyes and allows for better depth perception. High dynamic range (HDR) imaging accurately represents real world lighting as opposed to traditional low dynamic range (LDR) imaging. HDR provides a better contrast and more natural looking scenes. The combination of the two technologies in order to gain advantages of both has been, until now, mostly unexplored due to the current limitations in the imaging pipeline. This thesis reviews both fields, proposes stereoscopic high dynamic range (SHDR) imaging pipeline outlining the challenges that need to be resolved to enable SHDR and focuses on capture and compression aspects of that pipeline.

The problems of capturing SHDR images that would potentially require two HDR cameras and introduce ghosting, are mitigated by capturing an HDR and LDR pair and using it to generate SHDR images. A detailed user study compared four different methods of generating SHDR images. Results demonstrated that one of the methods may produce images perceptually indistinguishable from the ground truth.

Insights obtained while developing static image operators guided the design of SHDR video techniques. Three methods for generating SHDR video from an HDR-LDR video pair are proposed and compared to the ground truth SHDR videos. Results showed little overall error and identified a method with the least error.

Once captured, SHDR content needs to be efficiently compressed. Five SHDR compression methods that are backward compatible are presented. The proposed methods can encode SHDR content to little more than that of a traditional single LDR image (18% larger for one method) and the backward compatibility property encourages early adoption of the format.

The work presented in this thesis has introduced and advanced capture and compression methods for the adoption of SHDR imaging. In general, this research paves the way for a novel field of SHDR imaging which should lead to improved and more realistic representation of captured scenes.

Keywords: Stereoscopy, High Dynamic Range

CHAPTER 1

Introduction

Humans rely extensively on visual information to interpret the surrounding environment. Presenting the human visual system (HVS) with captured or created data which is indistinguishable from reality remains one of the main goals of digital imaging. Established techniques are able to convey the impression of a real scene, to an extent, but are still limited in a number of aspects including accurate light and depth reproduction. More recent digital techniques such as high dynamic range (HDR) imaging and stereoscopic imaging are able to overcome some of these limitations. More specifically, HDR imaging is able to preserve the full range of light available in the scene, thereby improving representation of light in an image, while stereoscopic imaging improves depth perception by simulating binocular vision (observing the world with both eyes). Each of these methods comes with a number of challenges which, so far, have been explored in isolation. This thesis brings HDR imaging and stereoscopic imaging together, and tackles some of the problems that arise as a result of this integration. It proposes a novel imaging method which advances current imaging technology.

1.1 Digital Imaging Pipeline

As this thesis is dealing with a new imaging method, the general imaging pipeline needs to be introduced. Any image or video passes through a number of manipulations in its lifetime. In general, this can be seen to consist of three key stages: capture, storage and display. They constitute the *digital imaging processing pipeline*, as is shown in Figure 1.1. Each stage of the traditional pipeline contains a set of established functions and operations which modify the content depending on the intended application. Each stage is examined in more detail

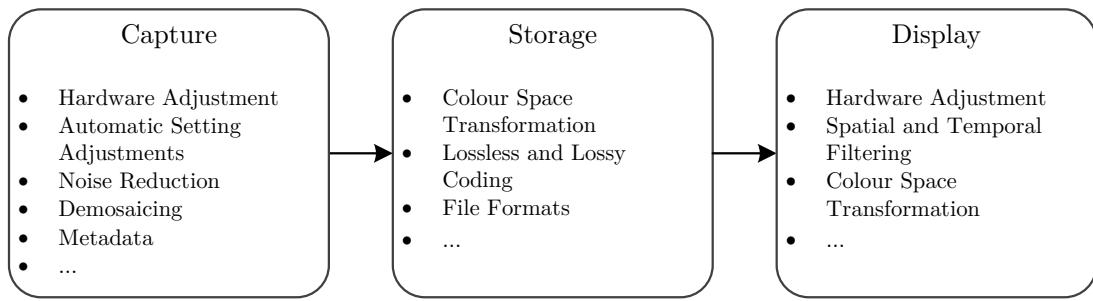


Figure 1.1: The digital imaging pipeline represents the stages through which images and videos pass from capture to display. In each stage different processing steps may be applied depending on the intended application.

in the next sections. Focus is placed on capture and compression, as the thesis contributes to these two stages.

1.1.1 Capture

Capture is concerned with recording images and videos. One aspect of this is the specification and choice of the right hardware which includes optics, sensors and lights. The skill of the device operator may be crucial, as visual data not recorded properly may be lost. The next important aspect is quality of the sensor as it determines key properties of the captured image or video: the resolution, amount of noise, quality of colour representation, dynamic range, sharpness and number of recorded frames.

On the software side, capture is concerned with automatic setting adjustments which try to take the load off the operator. Algorithms try to find optimal focus, white balance, shutter speed, aperture size and flash usage (Lee *et al.*, 2001). Recently, computer vision techniques have enabled more advanced techniques such as facial feature detection (Eckhardt *et al.*, 2009) which try to avoid capturing images of people blinking or wait for everyone to smile (Whitehill *et al.*, 2009). More fundamental problems involve demosaicing and noise reduction. One of the techniques to generate a colour image is to put a colour filter array over the image sensor. This is applied such that each sensing element encodes one of the three colour channels. Demosaicing is the process which combines neighbouring colour elements into a single pixel. When increasing the sensor sensitivity the amount of noise increases, especially in dark regions. Multiple filtering techniques can be applied to try and mitigate the problem (e.g. median filter). Captured content

may also contain metadata, which gives image or video properties. This can be useful in later stages of the pipeline (e.g. using exposure value metadata for creating HDR images).

Finally, content may also be generated using computer graphics. Setting up a virtual scene may require significant user input. Objects need to be modeled, textured, animated and rendered. While it may be time consuming, the output is usually of high quality.

1.1.2 Storage

It is common for colour images to consist of three colour channels: red, green and blue. Each of the channels is typically represented by a single byte which totals approximately 6.2 MB for a full high definition (HD) image (resolution of 1920×1080) or around 186 MB for each second of HD video captured at 30 frames per second (fps). However, most of the pixels are highly correlated so it can be expected that neighbouring pixels are similar to each other. In video, the changes between consecutive frames are usually small, meaning that pixels at the same spatial position are likely correlated. These facts allow for compression (encoding) methods which can reduce file sizes significantly.

Two general types of encoding exist: lossless and lossy. Lossless uses compression techniques which reduce redundancy and which are, also, fully reversible. Decoded content is identical to the original as all the data is preserved. Higher compression rates may be achieved by sacrificing some information resulting in lossy coding. The characteristics of the HVS and human perception are utilised in these approaches and data which is deemed less perceptually important is eliminated. This results in images or videos which are different from the original, but these differences - depending on the amount of compression - should not be noticeable to an observer. At this stage, the image may be converted to a different colour space which decorrelates colours better, improving compression. Also, improvements may be achieved by downsampling less influential channels. Both encoding and decoding may require substantial processing time which is not acceptable for real-time applications. To overcome this hardware implementations of standard techniques exist.

While many methods may be used to reduce the file size of images or videos, a standard is required if compressed content is expected to be widely accessible. To this end, file formats define a structure of the encoded data. A decoder is able

to open and display any content that follows an agreed specification irrespective of the encoding technique used.

1.1.3 Display

The final stage of the imaging pipeline is display and it relies heavily on hardware. Many manufacturers use custom processing methods to improve video reproduction on their devices. This includes contrast boosts and spatial filtering to improve appearance. Devices themselves may use different colour spaces requiring appropriate transformations. The screen resolution is typically fixed while it can receive content of many different sizes. Upsampling and downsampling may be required in such cases ensuring the image is displayed properly on the screen.

1.1.4 Other Techniques

In addition to these stages, the image may be manipulated in other ways by different applications. Compositing is a technique used in visual effects where parts of multiple images and/or videos are combined into a single one. Different filters may be applied to change the look and feel of the image, frequently for artistic purposes. For instance, an image may be turned from colour to greyscale and noise might be added to achieve an antique appearance. In-painting allows unwanted objects to be removed from the images and videos, by filling the empty regions with surrounding content.

1.2 Stereoscopic Imaging

Stereoscopic imaging enhances the perception of depth using two binocular depth cues: stereopsis and vergence (see Figure 1.2). Stereopsis is the phenomenon in which points in the observed environment project to different locations on the retina based on their distance from the observer. Vergence is the movement of the eyes to or from each other which ensures that the object of interest is projected to the centre of the retina. The distance of the focused object determines the angle of vergence which is used to infer depth.

In order to simulate this effect, the digital imaging pipeline needs to be modified to allow for two images, one for each eye. This means that two views need

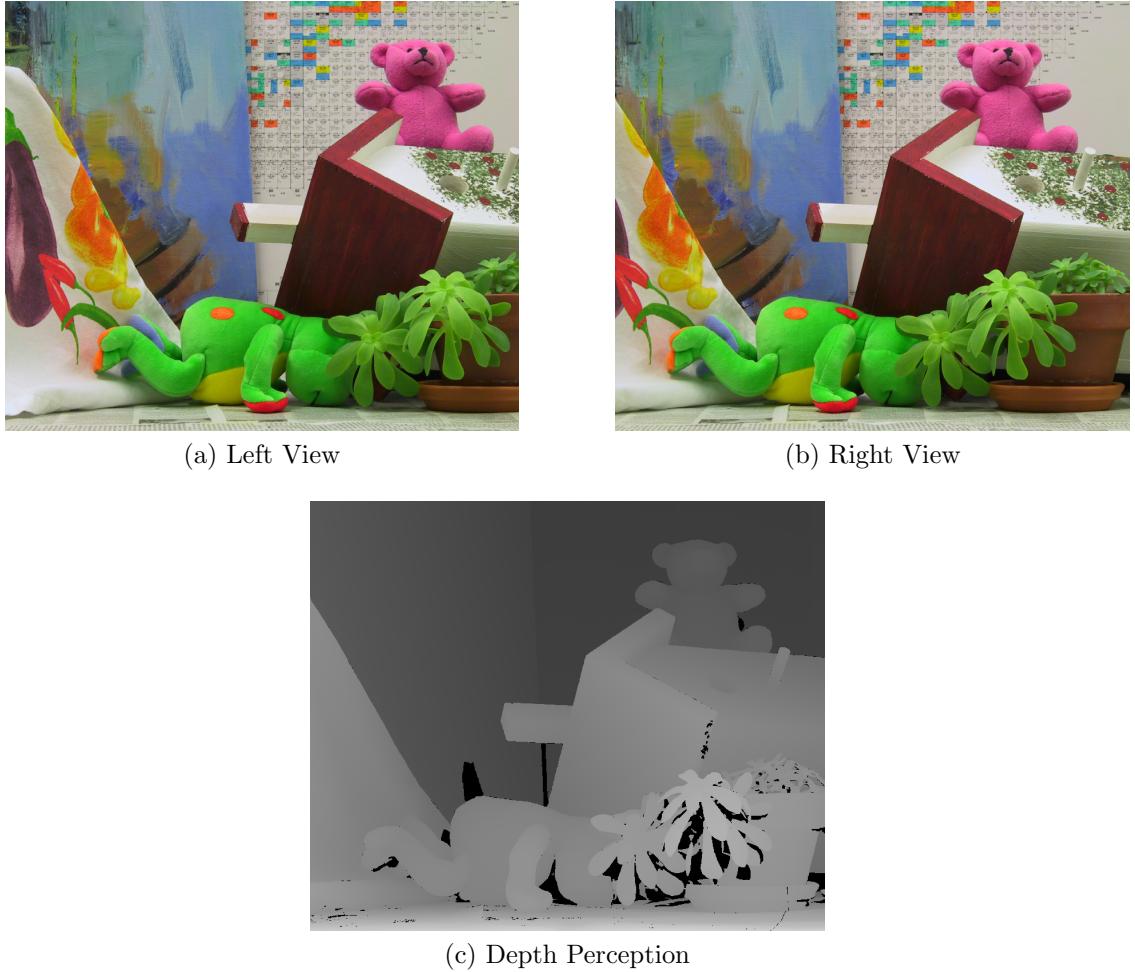


Figure 1.2: In binocular vision each eye gets slightly different view of the world. HVS uses discrepancies between the views to infer depth of the scene. Depth sensation is visualised in image (c) where brighter values represent closer objects. Completely black regions are occluded and cannot be seen by one of the eyes. Images courtesy of Scharstein & Szeliski (2003).

to be captured requiring two cameras positioned to emulate the human eyes, or by using a single camera and a depth sensor. The data captured by such a sensor allows for generation of the other view in subsequent steps. Alternatively, legacy content may be converted but may require significant manual input for high quality images. The two images required for stereoscopy double the amount of storage required, which current infrastructure may struggle with. Fortunately the images of both views are highly correlated allowing for efficient compression techniques. Displaying stereoscopic content is challenging and relies on observers wearing glasses which split the two views or an observer viewing the screen from

specific positions so that each eye gets its intended view.

A vast amount of research has compared stereoscopic to traditional two-dimensional viewing and a number of literature surveys are available. McIntire *et al.* (2012) reviewed and classified 71 objective experiments which examined human performance of 2D versus 3D displays. Overall results showed that 58% of the experiments found that 3D benefited measured human performance, 14% offered mixed results, while 28% showed no improvement in performance. Therefore 72% of the experiments demonstrated at least some benefit of stereo displays. Experiments were categorised into 6 groups based on performance type: (1) position and distance judgement, (2) identifying objects, (3) spatial manipulation of objects, (4) navigation, (5) spatial understanding and (6) learning. The major benefit of stereo displays was apparent for spatial understanding and spatial manipulation tasks (92% and 85% respectively, showed at least some benefit). Only two studies were examined in the learning category and they showed a lack of improvement. The other three categories had approximately half the experiments showing some benefit. This review excluded medical related literature because satisfactory summaries already exist. For instance, Held & Hui (2011) examined how 3D displays could improve diagnostic, surgery and training in medicine. They concluded that displays could: help doctors in detecting diagnostically relevant anatomical features, help novice surgeons in navigating the surgical landscape and aid in performing complicated tasks, and help students with anatomical understanding. Similarly, Getty & Green (2007) focused on medical applications and presented five cases where stereo displays were successfully used: teaching anatomy, digital mammography, tomography, diabetic retinopathy and minimal invasive surgery.

A literature review for military applications was provided by Dixon *et al.* (2009) who were concerned with displaying complex information using stereo and perspective representations, and they summarised 75 relevant papers. They concluded that 3D technology was the most useful in representing qualitative information, providing a quick overview of data, facilitating mission rehearsal, helping in route planning, visualising network attacks and providing realistic simulator training.

Other application areas include entertainment, industrial computer aided design (Brown & Gallimore, 1995) and photogrammetry (Kraus, 2007). Stereoscopy also found esoteric applications in the food industry where Lin *et al.* (2006) used it to identify edible birds' nests and Quevedo & Aguilera (2008) used it for esti-

mating firmness of salmon fillets.

Chapter 2 provides an overview of stereoscopic imaging. It provides a short history of the field, describes important theoretical concepts and examines how challenges at each stage of the imaging pipeline are being tackled.

1.3 High Dynamic Range Imaging

Current imaging techniques - termed *low dynamic range (LDR)* - are able to capture only a limited dynamic range. This causes overexposed (white) and underexposed (black) regions in an LDR image at the places which would otherwise contain details. For example, when taking photograph of a sunny day, one has to choose between capturing either the background sky and highlights or capturing the details in the shadow, even though both are visible to the observer, as illustrated in Figure 1.3. High dynamic range (HDR) imaging overcomes this limitation and is able to preserve the full range of light available in a scene.

High dynamic range imaging enhances the quality of image colour reproduction (Reinhard *et al.*, 2010), and as such it inherently provides benefits to many fields which rely on digital images and which are not hindered by image size. HDR imaging is also useful in a number of other more direct application areas.

Physically based rendering was one of the first applications for HDR imaging. In order to simulate light transport accurately, rendering programs need to store and process the full dynamic range of light (Ward, 1994a) and the final output may be scaled to match traditional formats using tone-mapping techniques (Banterle *et al.*, 2011). Increased precision is another useful property of HDR relevant for physically based rendering as it prevents error accumulation in multi-stage, multi-pass algorithms. HDR environment maps are used for image-based lighting, where HDR captured real-world lighting is used to light a virtual scene, thereby improving the realism and quality of the generated image.

Remote sensing acquires information about a phenomenon or an object without physical contact. Frequently, it refers to an aerial sensor detecting and classifying objects on Earth. Images captured using such an approach usually contain data outside human visible range (Lillesand *et al.*, 2004), making HDR of special importance for this application area.

In *digital photography*, cameras capture images in a specific RAW format developed by the manufacturer. These formats are not standardised so any program

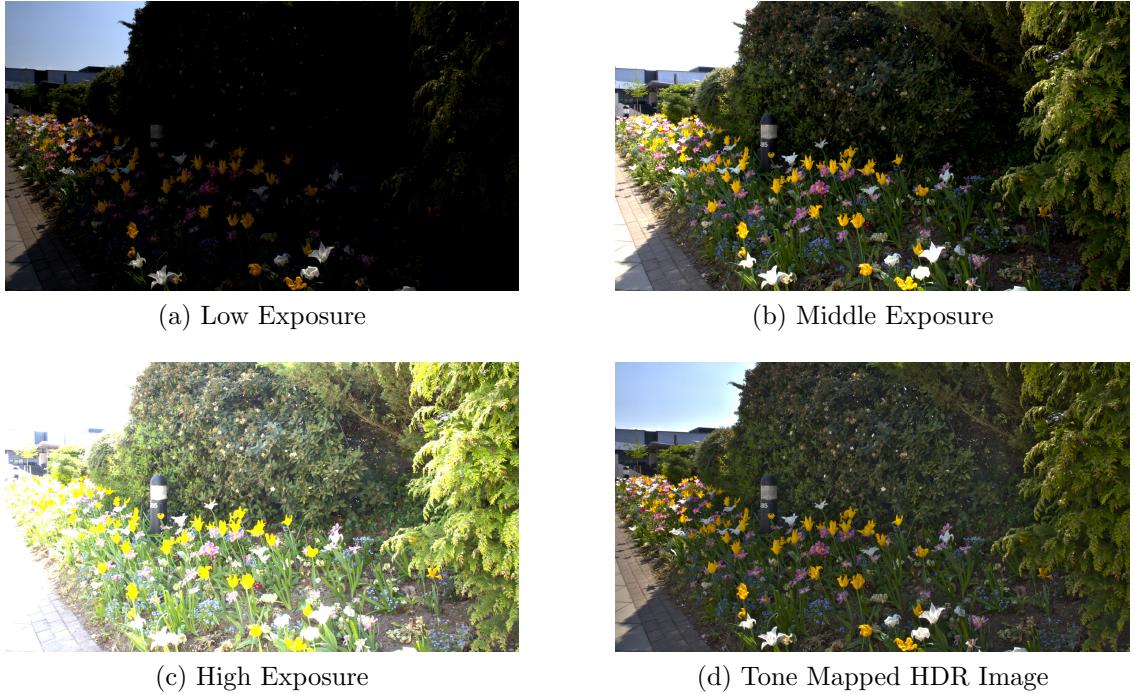


Figure 1.3: When capturing a scene which contains a wide range of light with current hardware one needs to chose whether to lose details in highlights or in shadows. (a) When capturing low exposure it is possible to see information in highlights, but the rest of the detail is lost. For example, the colour of the sky is visible. (b) Middle exposure provides a tradeoff by clipping the image both in the highlights and shadows but it preserves most of the data. (c) High exposure contains all shadow details but the rest of the image is overexposed. (d) The combination of the three exposures into a single HDR image which preserves all the information visible to the human eye.

which wants to use them needs to implement each one individually. This process is cumbersome, inconvenient and reduces compatibility. An elegant solution would use HDR as the standard encoding, which will likely become the case in the future. A downside of such an approach might be the file size, but top end consumer cameras already use 16 bits per channel for the RAW representation which could be used to represent HDR data instead.

Image and video editing benefits from HDR imaging as well. Most of the major editing packages enable a 32 bit workflow, including: Adobe Photoshop, Adobe Premiere and Foundry Nuke. HDR is of special importance for non-destructive rendering where operations that change intensities could run over or under low dynamic range limits and lose data. Also, operations on integers frequently generate quantisation errors which can accumulate over a number of

steps - a problem avoided by using HDR.

The *entertainment industry* is starting to recognise the potential of HDR. The current trend in digital cinema and the movie industry is turning towards medium dynamic range for digital film distribution. The main challenge which prevents HDR entering the mainstream is file-size. The development of robust and efficient compression algorithms will likely overcome this problem and HDR video may eventually reach the home screen. Many computer game engines render in HDR and tone-map output for increased realism. HDR texture compression is also becoming a critical element in the game rendering pipeline (Munkberg *et al.*, 2008).

Virtual reality (VR) strives to provide a realistic rendition of the real world which requires a true light representation and simulation. The HDR requirement for VR applications is highlighted by the fact that users may move around a virtual space that contains sharp light changes (e.g. walking from inside to the outside of a house).

Computer vision extracts information from photographs or video which can help solve a number of tasks including robot navigation, event detection, information organization and automatic inspection. HDR images provide more data and may improve performance of many computer vision algorithms. For example, Cui *et al.* (2011) suggest a method for generating reliable material colour images by subtracting highlights and shadows, which are recognised with the the help of HDR data. If the algorithm is required to run in real time, the HDR data size might become an issue.

Security systems which rely on video monitoring may benefit from HDR. Security cameras are frequently positioned to observe entrances of buildings or windows. Such a setup is especially susceptible to sharp differences in contrast, where the inside of the building is expected to be dark with strong shadows while outside may be bright and sunny. Traditional cameras would have to choose which of the two extremes will be filmed, or worse they might auto adjust to the middle range and risk capturing everything poorly. In addition, traditional cameras have problems with dark environments which are usually recorded with an increased amount of noise - and which are, in terms of security, usually more important. Future HDR cameras are expected to perform well in both setups.

Current trends in *education and training* are turning towards online delivery of lectures (Koller, 2011). In order to transfer the complete visual sensation to students and improve understanding, HDR data might be required. This is

especially the case for professions such as medical surgery. Here operations are performed under strong lights which focus on specific regions thereby creating sharp contrast.

Switching from traditional LDR to HDR imaging requires alterations to and optimisation of the imaging pipeline. Sensors which capture such a range with all the colours and with satisfactory resolution have not been designed yet. To overcome this, multiple images of the same scene may be taken at different exposures and combined. Such an approach is not feasible for video due to a speed requirement. Instead multiple sensors and a beam splitter (optical elements that let some amount of light pass and reflect the rest) are aligned so that a different amount of light arrives at each of the sensors, after which the images are combined. The amount of HDR data is quadrupled compared to LDR imaging as data is represented by floating point numbers which use 4 bytes per colour channel instead of a single byte as is the case with LDR data. Lossy methods which achieve high compression rates have been proposed and try to quantise the image so that errors are not perceivable by the HVS. The problem of constructing displays which would enable natural visualisation of HDR content has been tackled as well. The current solutions rely on a combination of a low resolution LED array which boosts the range and a standard LCD display which reproduces details. Alternatively, the range of HDR content may be scaled down and displayed on LDR screens - a technique termed *tone mapping*.

High dynamic range imaging is an active research area and is reviewed in more detail in Chapter 3. Theoretical concepts behind HDR imaging are described, and proposed methods and algorithms that tackle problems at each stage of the pipeline are examined.

1.4 Research Objectives

The combination of stereoscopic and high dynamic range imaging has been, until now, mostly unexplored. This thesis aims to merge these techniques in order to create a powerful visualisation method, building upon the existing knowledge in both areas. The main research objectives of this thesis are:

- to provide a comprehensive literature review of the stereoscopic and HDR imaging fields

- to recognize the potential challenges that arise from the merge of stereoscopic and HDR technologies
- to devise and validate an approach which facilitates capturing of SHDR images and overcomes limitations posed by using two static HDR cameras
- to propose and validate a temporally robust approach for capturing SHDR videos using insights gained while developing techniques for single images
- to devise and evaluate techniques for compressing SHDR images which are backwards compatible with traditional imaging techniques and thereby facilitate the adoption of SHDR imaging

1.5 Thesis Outline

This thesis consists of eight chapters, which are arranged as follows:

Chapter 2: Stereoscopic Imaging: Provides an overview of stereoscopy with a focus on concepts relevant to this thesis. This chapter presents a brief history of stereoscopy, explores important theoretical aspects such as epipolar geometry and disparity maps, examines visual comfort, and covers the stereoscopic imaging pipeline (capture, storage and display).

Chapter 3: High Dynamic Range Imaging: Examines the main concepts of high dynamic range imaging which are relevant for this thesis. It focuses on the HDR imaging pipeline (capture, storage and display).

Chapter 4: Stereoscopic High Dynamic Range Imaging Pipeline: Explains the different stages of SHDR image and video processing. It presents challenges which need to be addressed to enable SHDR.

Chapter 5: Generating SHDR Images using an HDR-LDR Camera Pair: Explores two different approaches for the creation of SHDR images captured with a combination of HDR and LDR cameras. The first approach is based on expansion operators while the second one uses disparity maps, which maps corresponding pixels of the two stereo views, to transfer high dynamic range data.

Chapter 6: Generating SHDR Video using HDR-LDR Camera Pair: The approach which performed the best for obtaining SHDR images is extended to enable video capture. Achieving temporal constancy is challenging requiring techniques which prevent flickering. The proposed techniques are compared to the best static one.

Chapter 7: SHDR Coding: Proposes and compares five image coding methods. Two methods format images so that existing coders can be used, two methods exploit disparity maps while the last one relies on motion compensation to compress differences.

Chapter 8: Conclusions: Summarises the work, and describes directions for future research.

CHAPTER 2

Stereoscopic Imaging

The ultimate goal of computer graphics is to obtain a realistic representation of the imaged scene to be viewed by people. The main purpose of the human visual system (HVS) is the inverse - to infer the geometric properties of the observed environment and extract information about events, objects and location. One key ability of the HVS is to perceive depth, as this allows understanding of the world in three dimensions. It is therefore desirable that captured video and images are able to correctly represent such characteristics.

Depth cues can be divided into two groups: monocular and binocular (Goldstein, 2009). Monocular cues are observed by one eye and are used in traditional media (e.g. paintings and photographs). They include: perspective, relative size, familiar size, occlusion, texture gradient, defocus blur and accommodation. Binocular depth cues, which use both eyes, are stereopsis and vergence.

Objects in the environment project to different positions on a retina based on how far they are from the observer. The HVS uses this information to infer depth, in a phenomenon termed *stereopsis* or *binocular vision*. When inspecting the object of interest, eyes move to or from each other, to project it onto the centre of the retina. This simultaneous horizontal movement is called *vergence*.

Stereoscopy is any imaging technique which enhances or enables depth perception using the binocular vision cues. The following subsections examine a brief history and theory of stereoscopy, discuss potential viewing discomfort, and present its imaging pipeline with special focus on disparity map generation.

2.1 A Brief History of Stereoscopy

It is interesting that mechanisms of stereopsis were understood only around 170 years ago while the required theory was attainable since ancient times. Greek philosophers were the first to recognize the problem of perceiving a single image while observing the world with two eyes - a phenomenon termed *singleness of vision*. Aristotle (384 - 322 BC) noticed how the image of the viewed object doubled, by pressing an eye. In about 300 BC, Euclid wrote the earliest known book on optics simply titled *Optics* which contained over 60 theorems relevant to this day. He observed that the world appears somewhat different when viewed by each eye. Euclid, like many Greek philosophers, adhered to the *emission theory* of light - suggested by Empedocles (5th century BC) - which stated that rays of light left an eye in a cone shape and sensed the surface on which they fell, like fingers. Ptolemy (127 - 165) thoroughly investigated binocular vision but was concerned with the singleness of vision, and not depth perception. He also followed emission theory and stated that axes of ray cones, projected from an eye, are the lines of fixation. When the lines meet on an object it is seen as one. Then he mistakenly concluded that the single vision in periphery is achieved for the points lying on a frontal plane passing through the fixation point. Had he chosen the circle passing through the fixation point and centre of the eyes instead (which he had the idea of), he would have come to the modern definition of horopter: a set of points in space that get fused into a single vision. Also he mistakenly believed that the eye could sense distance from the length of the ray projected from the eye to the object. Galen (AD 129 - 201), too, tackled the problem of single vision. He suggested that it occurs at the crossing of optic nerves (chiasm) where two images unite. Galen also noticed that near objects during binocular vision appeared at two different places in the background.

Research on optics was continued centuries later by Arab scholars (Lindberg, 1996) such as: Abu Yusuf Ya'quib ibn Ishaq al-Kindi, who stated that everything in the world shoots rays in all directions, but still supported emission theory; and Avicenna, whose important goal was rebutting the emission theory. In this period, Alhazen (965) contributed to optics and vision the most. He, similar to Avicenna, rejected emission theory and described how light emanates from its sources, reflects and refracts when hitting an object, and enters the eye. This is the foundation of modern optics. He also referred to the *pinhole camera* when comparing the eye to a dark chamber with a small hole through which light

enters and forms an inverted image. Following Ptolemy, Alhazen explained that objects near intersection of lines of fixation appear single, but also added that other objects to the side of the frontal plane double. Even though he realised that horopter was not corresponding to the frontal plane, he did not go a step further and show it was actually a circle. In terms of depth perception Alhazen was focused on monocular cues such as apparent size, aerial perspective and parallax during head motion; but he also made an important contribution to binocular depth perception: recognising that we could sense a degree of convergence of the eyes.

Further major advances in knowledge about binocular vision occurred in seventeenth century. Before this, Leonardo da Vinci and Giovanni Battista della Porta (1535 - 1615) also made contributions; Leonardo analyzed partial and total occlusions and noted their role in depth perception while Giovanni reported the phenomenon of binocular rivalry when observing different images with each eye. In 1613, Franciscus Aguilonius published a book on optics, synthesizing the works of Euclid, Alhazen, Vitello and others. He understood that binocular vision improved depth perception, and coined the term horopter but used it differently to its modern meaning. For him the horopter was the frontal plane passing through the convergence point where all vision rays ended, leading to single or double vision. Aguilonius also rejected Alhazen's proposal of convergence as a cue for distance, but introduced a new important idea: the length of the ray of one eye can be judged by the other.

Johannes Kepler (1571 - 1630) set the first geometrical principles of image formation in the eye. He noted that an image of the outer world projected on the retina is inverted and flat. This discovery made the problem of stereopsis harder by posing the question of how depth can be seen from flat images. Kepler's solution was the same as Alhazen's: we perceive distance by sensing the rotation of the eyes. In 1667, Christiaan Huygens defined corresponding retinal points - the points in each eye that have the same location in relation to the centre of the retina. Isaac Newton, incorrectly, presumed that fibres from corresponding retinal points fuse in chiasma (the place where optic nerves partially cross), and one nerve from each pair proceeds into the brain. This hypothesis would explain singleness of vision but not depth perception.

At this point scientists were at the verge of discovering the theory of binocular vision, but it took another 150 years before that happened, due to the influence of new philosophical thought. John Locke (1632 - 1704), forerunner of the En-

lightenment and Empiricism, maintained that man had no innate ideas and was born as a *tabula rasa*, a blank slate. This meant that knowledge was formed only by experience derived from sense. George Berkeley (1685 - 1753) applied Locke's teaching to vision and stated that distance cannot be seen by itself and cannot be seen immediately. He suggested that the empirical connection between vision and touch needs to be established before seeing depth or other spatial relations.

Gerhard Vieth (1763 - 1836) offered the modern definition of horopter in 1818. He clearly presented geometry of corresponding points and horopter as the locus of points producing the single image, a concept which had eluded Ptolemy, Alhazen and Aguilonius. A few years later Johannes Müller (1801 - 1858) made a similar analysis, so today, theoretical horizontal horopter is termed *Vieth-Müller circle*.

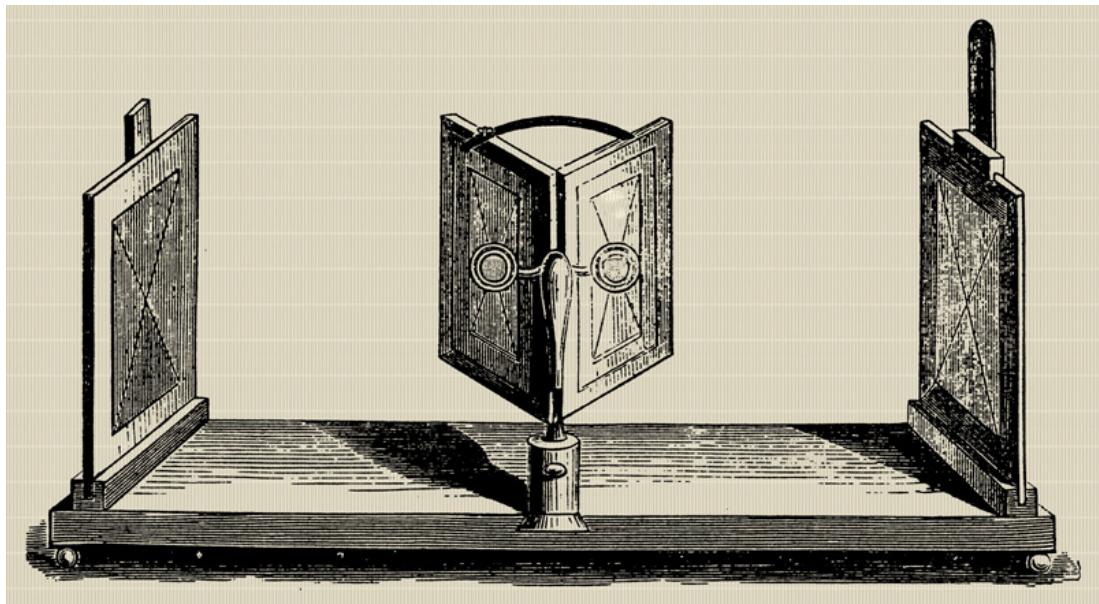


Figure 2.1: Wheatstone's mirror stereoscope. In the middle are two mirrors at the right angle and on the sides are vertical picture holders.

Charles Wheatstone (1802 - 1875) was an English scientist from the Victorian era who contributed to many fields including telegraphy, cryptography, chronometry and optics. He invented the stereoscope (Wheatstone, 1838) and had two of them made by the end of 1832: one using mirrors (Figure 2.1), the other a prism. He immediately understood the theoretical background of this finding. The stereoscope showed the relationship between depth perception and binocular vision. Also, adjusting the arms on this new instrument changed the eye convergence while keeping disparity constant, thereby demonstrating that depth



Figure 2.2: Random-dot stereogram. Both images lack any monocular cues or any familiar 3D object. When viewed in stereo, a square appears in the middle and is perceived closer than the rest of the image. Courtesy of Julio M. Otuyama

perception depended on more than disparity alone.

William Shaw (1861) made the first experimental moving picture by combining the thaumatrope (display device which contained revolving drum with sequence of images) and a mirror stereoscope. David Brewster (1781 - 1868) created his own version of prism stereoscope. A number of these devices were shown at the Great Exhibition of 1851 in London and one was specially made for Queen Victoria who showed a great interest. Three months later, close to a quarter of a million had been sold and the stereoscope became the optical wonder of the age only later overtaken by the advent of cinema.

In 1960, Bela Julesz demonstrated that depth perception is an innate visual ability which did not arise from high-level cognition. His random-dot stereogram (Figure 2.2) elicited the sensation of depth without any monocular cues. The finding was confirmed by Barlow *et al.* (1967) who showed that disparity-sensitive cells existed in the visual cortex, and for the first time, stereopsis entered the domain of physiology.

Other publications provide a more detailed historical account of binocular vision. Wade (1987) and Crone (1992) examined problems which delayed development of stereoscopy. Howard & Rogers (1995) give an overview of the field, with a chapter dedicated to the history of vision in general, with a focus on stereoscopy. Lindberg (1996) covers the history of visual theory from its beginning to the seventeenth century.

2.2 Theory of Stereo Vision

Computer vision poses two main questions in relation to stereoscopy. The first is: how do pixels between two stereo images correspond? For example, in Figure 2.3, one might be interested in finding the relation between the left and right image pixels of the apex of the cone. Recognising and correlating the cone's apex is easy for both humans and computers, because of its very distinct structure. However, this is not the case for a pixel located in proximity to the centre of the cone, as all the neighbouring pixels look very similar and might even have the same value. Finding correspondences between the left and right image for such a pixel is very challenging. Even the best algorithms tackling the problem only estimate the solution and are error prone for very difficult pixels.

The second question of stereo vision is: how to determine distance (depth) of a point - represented by corresponding pixels - in 3D space? For example, in Figure 2.3, one might inquire about how far the cone apex is from the camera. The distance might be expressed in relative (e.g. is the cone in front of the ball) or in absolute terms (e.g. how many metres is the apex from the camera). Absolute distance requires information of camera properties, which are not always available. The problem of calculating depth is related to the one of correspondence and depends on its correctness.

To understand why these two questions are challenging, how they are related, how they are solved, and how we can interpret results, a number of concepts need to be explained. Epipolar geometry explains image formation from two views - independent of scene structure, it also constrains the correspondence problem and enables calculation of point depth. The fundamental matrix algebraically relates correspondences. Rectification of images transforms them so the matching pixels lie on a horizontal line. These concepts are discussed in more detail below.

2.2.1 Epipolar Geometry

Epipolar geometry is the geometry of stereo vision. It relates 3D points and their 2D projections. To explore epipolar geometry the scenario presented in Figure 2.3 is examined. In this typical stereo setup, two cameras capture a scene and each one views the scene slightly differently. A theoretical analysis enables a study without the issues that arise in practice: lenses often introduce blurring and geometric distortions and the imaging plane is behind the aperture which inverts

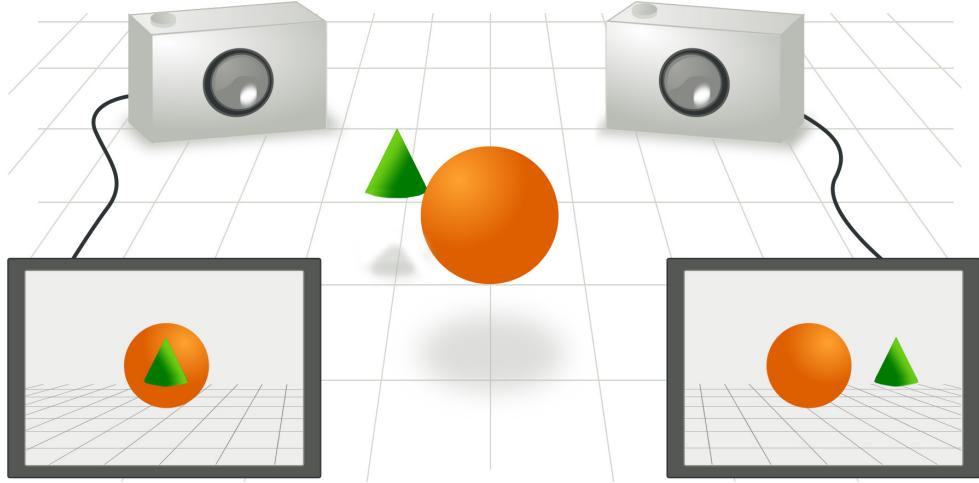


Figure 2.3: Illustration of stereo capture setup. Two cameras are imaging the scene. Two objects take different positions in each of the views depending on their distance from the camera. Courtesy of Arne Nordmann

the image. In the theoretical analysis, cameras get approximated by the *pinhole camera model* which uses no lenses and represents the aperture as a point (the point is termed the centre of projection or camera centre). In addition, projection is simplified by moving the image planes in front of the cameras resulting in a non-inverted image. Such a theoretical setup is shown in Figure 2.4.

Points \mathbf{C}_L and \mathbf{C}_R are the centres of projection for the corresponding left and right cameras. \mathbf{X} is a point of interest in 3D space which projects onto the two image planes by emanating two rays from each centre of projection which travel to the point \mathbf{X} . They intersect image planes at points \mathbf{x}_L and \mathbf{x}_R . A ray emitted from the left camera centre ($\mathbf{C}_L - \mathbf{X}$) corresponds to a single point on the left imaging plane (\mathbf{x}_L). Importantly however, the right camera observes this ray as a line - defined by two points: \mathbf{e}_R , projection of left camera centre (\mathbf{C}_L); and \mathbf{x}_R , projection of point of interest (\mathbf{X}).

The observation can be generalised for both views. Any ray projected from the camera centre to a point in 3D space is projected to a single point on its image plane and to a line on other camera's image plane. This projected line is termed the *epipolar line*. In Figure 2.4 lines $\mathbf{x}_L - \mathbf{e}_L$ and $\mathbf{x}_R - \mathbf{e}_R$ are epipolar lines. Since all rays emanate from camera centre their epipolar lines pass through a point that the camera centre is projected to - *epipoles*. Points \mathbf{e}_L and \mathbf{e}_R are

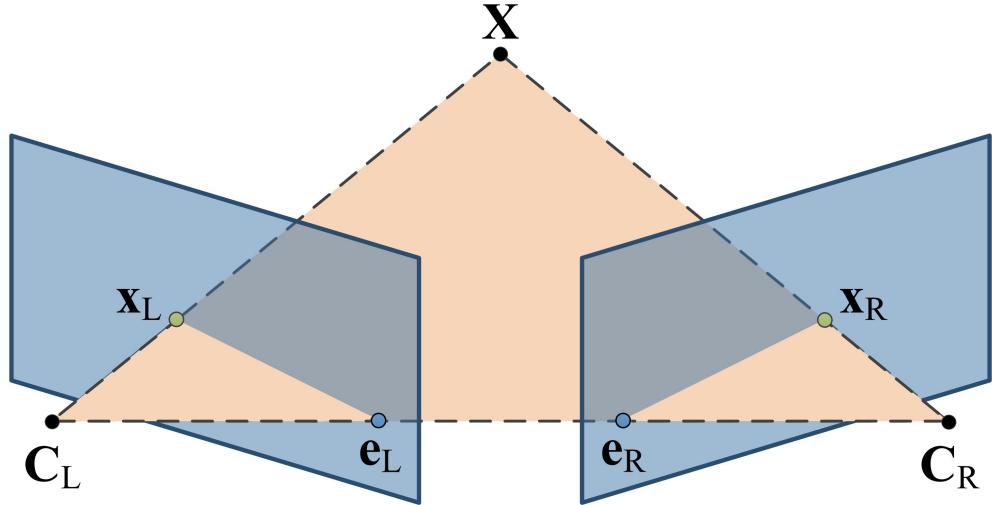


Figure 2.4: Epipolar geometry. C_L and C_R are the centres of projection, X is the point of interest, x_L and x_R are projections of point X onto the image planes and e_L and e_R are left and right epipoles.

epipoles of the left and right images.

Epipolar geometry does not provide direct correspondence between stereo image pixels, because the projection of a point to a line is a one-to-many mapping. However, it reduces the search for a matching pixel to a single line. Going back to the example in Figure 2.3, finding the cone apex becomes straightforward because on its epipolar line there is only one pixel of such a colour. Pixels close to the centre of the cone are still a challenge, as their epipolar lines contain pixels of the same intensity. Still, the epipolar line constrains the search significantly as many other same or similar pixels in the image are discarded. This has a major impact on speed and correctness of correspondence search algorithms.

Correct matches between pixels enable calculating the position of an imaged point in 3D space using triangulation. If focal length and camera separation are provided it is possible to get accurate measurements of the distance, otherwise relative measures are obtained.

2.2.2 Fundamental Matrix

The fundamental matrix captures an arithmetic mapping between an image point and its epipolar line. Given the point x_L in the left stereo image and a corresponding point x_R in the right image located on the epipolar line l_R , there is a fundamental matrix \mathbf{F} for which Equations 2.1 and 2.2 hold.

$$\mathbf{l}_R = \mathbf{F}\mathbf{x}_L \quad (2.1)$$

$$\mathbf{x}_R^\top \mathbf{F} \mathbf{x}_L = 0 \quad (2.2)$$

The fundamental matrix is a 3×3 homogeneous matrix of rank 2 with 7 degrees of freedom. Seven correct matches are required for computing this matrix (Hartley & Zisserman, 2004b). They can be obtained using a robust sparse correspondence algorithm based on SIFT (Se *et al.*, 2002) and further refined using RANSAC algorithm (Fischler & Bolles, 1981). For more properties of the matrix, its derivation and calculation please refer to Hartley & Zisserman (2004b).

2.2.3 Image Rectification

A special case of epipolar geometry arises when planes of two stereo cameras coincide. Here, the epipolar lines also coincide ($\mathbf{e}_L - \mathbf{x}_L = \mathbf{e}_R - \mathbf{x}_R$) and are parallel to the line connecting two camera centres ($\mathbf{C}_L - \mathbf{C}_R$). This means that corresponding pixels of stereo images are on the same horizontal line which simplifies matching even further. In practice, aligning two cameras so that they are perfectly parallel is difficult. However, captured images can be transformed afterwards using the process of *image rectification*.

The image rectification method starts by computing the fundamental matrix, after which a projective transformation maps the epipole of one of the images to the infinity point. Then the algorithm finds optimal matching transformation for the other image. Finally, it resamples two images each using its corresponding transformation. For detailed implementation please refer to Hartley & Zisserman (2004a).

2.2.4 Disparity Maps

As mentioned above, epipolar geometry and rectification simplify the search for matching pixels to a single horizontal line. A pixel representing a point in 3D space on the left image, may have a matching pixel in the right image which represents the same 3D point. These pixels are in the same row on both images, but they are offset horizontally (in different columns) by some amount, depending on depth. So, for all the pixels in the left image it is possible to use a set of horizontal offset values to map to the corresponding pixels in the right image.

Such a mapping can be represented by an image whose pixels symbolise the left image pixels, and whose values are horizontal offsets. This image is called the disparity map, and it maps stereoscopic disparities of a stereo image pair. An example of the ground truth disparity map (containing no errors) is shown in Figure 2.5.

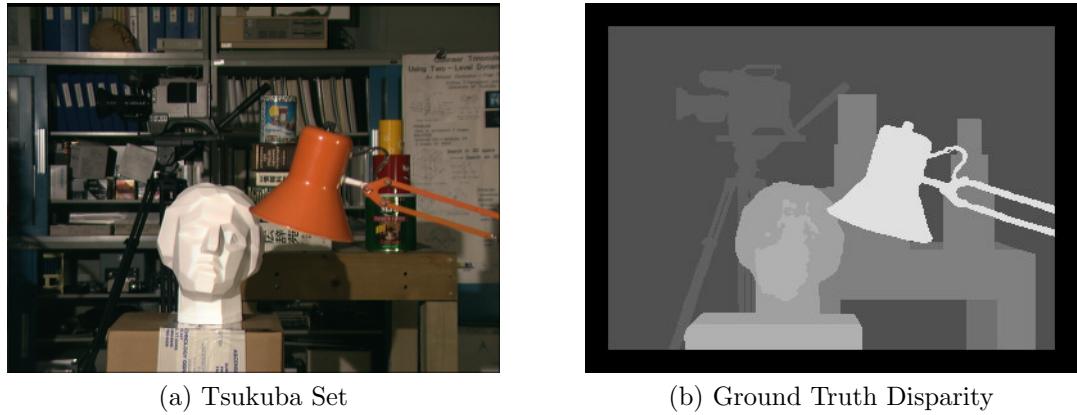


Figure 2.5: Example of disparity map. (a) The *Tsukuba* scene is frequently used for testing stereo correspondence algorithms. (b) The ground truth disparity map where closer points are brighter while further away points are darker. Courtesy of the University of Tsukuba.

While epipolar geometry simplifies the problem of calculating the disparity map, the problem still remains a challenge as it is not constrained and ambiguities can occur. This is especially the case for the pixels in regions of similar colour where multiple equally good matches are possible. In addition, close objects cover the ones behind them causing occlusions. One of the cameras in a stereo capturing system can capture parts of the scene that are occluded to the other (Nakayama & Shimojo, 1990), causing the lack of pixel correspondence. The same problem is created by the surfaces whose appearance depends on a viewing point - for example specular or refractive ones. Finally, stereo cameras may not be perfectly colour calibrated reducing the robustness of matches.

The main goal of stereo vision, and one of the main goals of computer vision in general, is generating accurate disparity maps and overcoming the mentioned challenges. This active research area provides novel and improved solutions each year. In general, the disparity calculation can be separated into dense and sparse methods. Dense methods provide disparity values for all the pixels in an image while sparse methods provide highly robust matches for a much smaller set of

points. Sparse methods are more suitable for applications such as robot navigation, where the quick and robust matches are needed but their quantity is not crucial. On the other hand, dense methods are more often used in applications such as image based rendering where the disparity value for each pixel is needed to generate a novel view. The work presented in this thesis uses dense maps so they are examined in more detail.

In 2001, Scharstein *et al.* (2001) contributed to the field by providing a survey and evaluation of existing dense stereo matching methods. More importantly, they provided a test bed which enabled new algorithms to compete against the existing ones, and compare their performance with the ground truth. This has made it possible to test algorithms quickly and conveniently in a fair environment, which has accelerated their development. All the submitted algorithms ran on the data set of four scenes, and were ranked based on the number of the correct matches. Currently, over 130 methods have been evaluated.

Since the first publication, additional data sets have been added, but the tested algorithms were still ranked based on the four scenes. While these images provided challenging matching regions they could not account for many situations which arise in other circumstances. For instance, all the scenes represented closed environments, limiting the disparity range, and they portrayed inanimate objects avoiding high frequency depth changes present in nature (e.g. tree branches). The spatial resolution was limited as well (less than 450×383) and results do not necessarily hold for larger images. Algorithm running time was not taken into account when calculating ranking so some submissions ran in real-time while others could take hours to compute.

Taxonomy of Stereo Matching Algorithms

Scharstein *et al.* (2001) also provided a taxonomy of the stereo matching algorithms. They recognised four steps that stereo algorithms generally performed: computation of matching cost, aggregation of the cost, optimisation or computation of disparity, and refinement of disparity. For instance, the current top ranked algorithm (Mei *et al.*, 2011) clearly follows each stage. Below, each step is discussed further.

Computation of matching cost uses a measure to decide how similar two pixels are. Cyganek & Siebert (2009) analysed common matching measures and explained how a scalar value could describe the quality of a match. For example,

squared or averaged difference of pixel intensity values are frequently used. Cost for all the pixels and all the disparities generate an initial *disparity space* which is fed to the following steps.

The cost for a single pixel still produces ambiguities as many pixels have the same cost value especially for images with larger spatial resolution. To alleviate this problem, an **aggregation of the cost** step assumes disparity smoothness - that is neighbouring pixels should match as they are likely to be at the same depth. In this step each pixel uses a support region - consisting of matching costs of neighbours - to aggregate results and distinguish between similar matches. Frequently, the support region is a square block of fixed size, but more advanced solutions include shiftable windows or dynamic regions which are computed for each pixel depending on the frequency of its neighbourhood (Mei *et al.*, 2011).

The disparity computation or optimisation step examines costs and decides the disparity for all the pixels. This can consist of: local, global and dynamic programming. *Local methods* are focussed on two previous steps and compute the final disparity using a straightforward winner-takes-all (WTA) approach where each pixel selects disparity with the smallest cost. This approach guarantees uniqueness of matches for one image only, as multiple pixels can pick the same point from the other image. Unlike local operators, *global methods* perform the majority of work in this third phase often skipping cost aggregation. They go through possible disparity configurations (disparity maps) trying to find one which would minimise the global energy (cost) function. Algorithms formulate the function differently and use different techniques for solving the minimisation problem. A generic function can be expressed as in Equation (2.3):

$$E(d) = E_{\text{data}}(d) + \lambda E_{\text{smooth}}(d) \quad (2.3)$$

where d is a configuration of disparity values, $E(d)$ is the global cost of disparity configuration which is to be minimised, $E_{\text{data}}(d)$ is the data term, $E_{\text{smooth}}(d)$ is the smooth term, and λ weights influence of the two terms.

The data term measures the global matching cost of a disparity configuration d by summing intensity differences between pixels. The output from either the first or second step is used, so generally the data term can be expressed as:

$$E_{\text{data}}(d) = \sum_{(x,y)} C(x, y, d(x, y)) \quad (2.4)$$

where $C(x, y, d(x, y))$ is the matching cost for the pixel at the location x, y for the disparity at that pixel $d(x, y)$.

The smoothness term imposes a smoothness constraint by penalising jumps in the disparity map. Generally, this term can be expressed as:

$$E_{\text{smooth}}(d) = \sum_{(x,y)} \rho(d(x, y) - d(x + 1, y)) + \rho(d(x, y) - d(x, y + 1)) \quad (2.5)$$

where ρ is a monotonically increasing function which penalises differences depending on their range. For instance large jumps in depth can be penalised less to allow sharp disparity changes around object edges, while small jumps are penalised more enforcing smooth disparity change across an object. After defining the energy function, global minimisation can be performed using a number of methods including graph-cuts, max-flow and Markov Random Fields.

An alternative approach to global disparity computation is dynamic programming. While energy minimisation - as described in Equation 2.3 and for frequently used smoothness functions - is NP-complete, dynamic programming computes a global minimum for individual image lines in polynomial time. This method operates by first constructing a matrix for a matching pair of scanlines (e.g. the same row of left and right image). In the matrix, columns represent pixels in one scanline and rows represent them in the other, while the cells entries hold the matching costs. A disparity is calculated by finding the minimum cost path through the matrix (Mei *et al.*, 2011).

Refinement of disparity is the final step performed by most of the algorithms. Once disparities are computed, methods that refine them are available. The disparity map can be smoothed and sub-pixel precision can be achieved by fitting a curve to discrete disparity values. Cross-checking is a process where both left-right and right-left disparity maps are calculated and compared in order to determine inconsistencies and occlusions. Filtering can remove noise and smooth the map further. To this end many existing filters can be used, such as the median or edge-preserving bilateral filter (Elad, 2002). Identified occlusions can be filled using in-painting methods (Wang *et al.*, 2008).

2.2.5 Stereo Correspondence Algorithms

This section describes three stereo matching methods that were used in this thesis in more detail. The sum of absolute differences is a local, fast and straightforward

technique, but lacks precision (used in Chapters 5, 6 and 7). Kolmogorov & Zabih (2001) suggested a method for computing correspondences with occlusions using graph cuts (used in Chapter 5). This is a global technique which uses energy minimisation and produces smooth disparity maps. However, it requires long computation times. Mei *et al.* (2011) proposed another global method which calculates smooth disparity maps quickly using average differences and a census measure together with dynamic programming (used in Chapter 7). Also, it was developed to run on graphics hardware boosting its speed.

Sum of Averaged Differences (SAD)

Sum of averaged differences (SAD) is a basic method to decide if two pixel values correspond Cyganek & Siebert (2009). It follows three of the initial steps described above. The measure of matching cost is the average intensity difference between two pixels. For colour images, differences are summed across the channels. Costs are aggregated using a square window of a specified size, centred at the pixel for which disparity is to be calculated. Averaged differences of all pixels in this window are summed and represent the final matching measure (hence the name of the algorithm). The process can be described by the equation:

$$SAD(x, y) = \sum_{k \in R, G, B} \sum_{(i, j) \in W(x, y)} |I_{k,1}(x + i, y + j) - I_{k,2}(x + d_x + i, y + j)| \quad (2.6)$$

where $W(x, y)$ are point coordinates of a window located at (x, y) , $I_{k,l}(x, y)$ are the intensity values of k -th channel of l -th image at (x, y) , d_x is a horizontal image displacement, and $SAD(x, y)$ is the value representing the difference between the compared regions.

It is assumed that stereo images are rectified (Hartley & Zisserman, 2004a) reducing the search to a single horizontal line of pixels and the search is usually limited to a certain range. Final disparities are decided using a WTA method, where the matching pixel is the one with minimal aggregated cost. In case of two pixels having the same cost, the first checked pixel is selected.

Using SAD for stereo matching provides results quickly and can easily be implemented in real time, but the method provides erroneous disparity matches, with a low signal to noise ratio resulting in high frequency disparity maps. Also it is important to note the matching property of this algorithm, which was useful for solving problems in Chapters 5, 6 and 7: the SAD technique may be less

precise for disparity correspondence, but it still connects pixels of two stereo images which closely match in colour.

Correspondence with Occlusion via Graph Cuts (COGC)

Correspondence with occlusion via graph cuts (COGC) (Kolmogorov & Zabih, 2001) follows all the steps of the global stereo matching method, making it representative of those methods. Two specific aspects of COGC are: imposing uniqueness of matches - that is making sure that one pixel from the left image matches only one pixel in the right and vice versa; and occlusion handling - explicitly recognising occluded pixels. This method defines an energy function to be minimised, expanding on Equation 2.3 by adding an extra term for handling occlusions:

$$E(d) = E_{\text{data}}(d) + E_{\text{occ}}(d) + E_{\text{smooth}}(d) \quad (2.7)$$

The data term $E_{\text{data}}(d)$ measures intensity differences between matching pixels and is the same as in Equation 2.4. The measure used is squared difference which is defined as:

$$C(x, y, d_x) = \sum_{k \in R, G, B} ((I_{k,1}(x, y) - I_{k,2}(x + d_x, y))^2) \quad (2.8)$$

where $I_{k,l}(x, y)$ are the intensity values of k -th channel of l -th image at (x, y) , d_x is a horizontal image displacement, and $C(x, y, d_x)$ is the value representing squared difference between compared pixels.

The occlusion term $E_{\text{occ}}(d)$ penalises pixel occlusion. It adds a constant penalty value C_p for any pixel which is deemed occluded:

$$E_{\text{occ}}(d) = \sum_{p \in P} C_p \cdot T(|N_p(d)| = 0) \quad (2.9)$$

where p is a pixel in the set of all pixels in both images P , $|N_p(d)|$ is the number of found matches for pixel p and $T(\cdot)$ is truth function. A lack of matches for a pixel indicates that the pixel is occluded and a penalty is added.

The smoothness term $E_{\text{smooth}}(d)$ favours similar disparities across neighbouring pixels. It is the most challenging term to formulate because it can result in NP-completeness if not stated properly. Kolmogorov & Zabih (2001) suggest the

following solution:

$$E_{\text{smooth}}(d) = \sum_{n1, n2 \in N} \lambda \cdot |d(n1) - d(n2)| \cdot T(d(n1) \neq d(n2)) \quad (2.10)$$

where N is a neighborhood system in d consisting of pairs of neighbouring pixels $n1$ and $n2$, and λ is a constant controlling the strength of the smoothness penalty which is defined as the L_1 distance between disparities.

Once the energy function is formulated, it can be optimised using graph cuts and the high level algorithm can be stated as follows:

Algorithm 1 Find optimal disparity configuration d

```

1: choose unique configuration  $d$  randomly
2:  $finished = 0$ 
3: for all disparities  $\alpha$  do
4:   find  $\hat{d} = \arg \min E(d')$  among unique  $d'$  within single  $\alpha$ -expansion of  $d$ 
5:   if  $E(\hat{d}) < E(d)$  then
6:      $d = \hat{d}$ 
7:      $finished == 1$ 
8:   end if
9: end for
10: if  $finished = 1$  then
11:   goto 2
12: end if
13: return  $d$ 
```

In this algorithm α is a disparity value selected at random or in a fixed order. α -expansion is a process that assigns a disparity value α to some pixels thereby minimising the energy function. The graph cut algorithm decides this assignment and finds an optimal solution. Kolmogorov & Zabih (2001) detailed instructions on how to construct the graph and how to perform the cut.

COGC generates robust smooth disparity maps with labelled occlusions. The authors report 6.7% incorrect matches for unoccluded pixels on a tested image pair. The algorithm's running time is a drawback as for a full high definition (1920×1080) image, calculation may take hours.

Averaged Differences and Census (ADC)

The averaged differences and census (ADC) method, proposed by Mei *et al.* (2011), attempts to achieve a balance between precision and speed. Parts of the

solution can be parallelised on graphics hardware increasing performance. The algorithm follows the four steps of stereo matching closely.

Cost computation combines two measures: census transform and absolute differences. Census describes the local structure of a region surrounding a pixel. The intensity of the pixel is compared to its neighbour and a boolean value is used to represent the result. If a neighbouring pixel is smaller 0 is recorded and 1 otherwise. Comparing pixels to all the neighbours generates a bit sequence. For example, a 9×7 box neighbourhood generates 62 bit sequence. This sequence can be used as a measure between left and right image pixels and the Hamming distance (number of equal bits) is used to decide how well two of them correspond. While this measure shows the best results for local and global stereo matching, it may underperform in repetitive regions with similar structure. To alleviate this problem, the measure is combined with absolute differences, which is calculated as in Equation 2.6 but without neighbourhood aggregation:

$$AD(x, y) = \frac{1}{3} \sum_{k \in R, G, B} |I_{k,1}(x, y) - I_{k,2}(x + d_x, y)| \quad (2.11)$$

where $I_{k,l}(x, y)$ are the intensity values of k -th channel of l -th image at (x, y) , d_x is a horizontal image displacement, and $AD(x, y)$ is the value representing the difference between the compared regions.

The two measures are combined as shown in the Equation 2.12.

$$C(x, y, d) = \rho(C_{\text{census}}(x, y, d), \lambda_{\text{census}}) + \rho(C_{\text{AD}}(x, y, d), \lambda_{\text{AD}}) \quad (2.12)$$

where C_{census} and C_{AD} are matrices containing census transform values and absolute difference values, respectively, for all pixels and all disparities. ρ is defined as:

$$\rho(c, \lambda) = 1 - \exp(-\frac{c}{\lambda}) \quad (2.13)$$

The function ρ is used to scale measures to the $[0, 1]$ range and the parameter λ controls the influence of the measures on $C(x, y, d)$. When the right balance is achieved, this combined measure provides better results than they do individually.

Cost aggregation generates a custom support region (as opposed to using a square window) for each pixel using a cross based technique. This method proceeds in two steps. First the cross is constructed by selecting horizontal and vertical neighbouring pixels resulting in four arms. The colour difference

and spatial distance thresholds decide the length of each arm, where the colour difference takes priority. In the second step horizontal arms are created for each of the pixels on two vertical arms. This results in a support region whose values get aggregated. Mei *et al.* (2011) extend this method by comparing not only the pixel in question with the ones on the arm, but also comparing consecutive pixels on an arm. This prevents the region growing over an edge. In addition, two colour difference and spatial distance thresholds were used. When an arm exceeds the first spatial threshold, a more strict colour difference is used to stop growth. This allows for large regions in low frequency areas but limits their size in high frequency ones.

Disparity computation is performed using dynamic programming. More specifically, multi-direction scan-line optimisation performs semi-global matching. It runs along four directions (2 horizontal and 2 vertical) independently. The process is described as:

$$C_{\mathbf{r}}(\mathbf{p}, d) = C_1(\mathbf{p}, d) + \min(C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d), C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d \pm 1) + P_1, \min_k C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k) + P_2) - \min_i C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i) \quad (2.14)$$

where C_1 is the aggregated cost volume, $C_{\mathbf{r}}(\mathbf{p}, d)$ is the cost at pixel \mathbf{p} with disparity d in scanline direction \mathbf{r} , $\mathbf{p} - \mathbf{r}$ is the previous pixel in the direction \mathbf{r} and P_1 , P_2 ($P_1 \leq P_2$) are parameters that penalise disparity change between neighbours.

In other words this formula takes the pixel's disparity in disparity space $C_{\mathbf{r}}(\mathbf{p}, d)$ and adds a penalty for discontinuities. To determine the final cost, the minimum of the three following terms are used. $C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d)$ is the cost of selecting the same disparity value for the previous pixel. This means that neighbouring pixels have the same disparity so no penalty is added. $C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, d \pm 1)$ is the cost of a neighbouring pixel having a disparity value which differs by one. As the disparity difference is small, penalty P_1 is added. The third term $\min_k C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, k)$ finds a disparity value for a neighbouring pixel which differs by more than one but which has the smallest cost. Because disparity difference here is the largest, a large penalty P_2 is added. Once all three terms are available the smallest one is selected. Finally, term $\min_i C_{\mathbf{r}}(\mathbf{p} - \mathbf{r}, i)$ prevents growth of cost along the path - which may result in large values - by subtracting the smallest path cost of the previous pixel. Penalties P_1 and P_2 vary based on colour difference in the left and right image. Mei *et al.* (2011) use four simple conditions to set the two

parameters using an empirically obtained colour difference threshold.

Costs for all four scan directions are averaged:

$$C(\mathbf{p}, d) = \frac{1}{4} \sum_{\mathbf{r}} C_{\mathbf{r}}(\mathbf{p}, d) \quad (2.15)$$

and a WTA approach provides the final disparity for a pixel by selecting the disparity value with the smallest cost.

The final stage of the ACD algorithm **refines disparities** in multi-step process which tries to remove outliers in occluded regions and at depth discontinuities. The aim of each step is to eliminate errors caused by different factors. Cross-check *detects outliers* and the intersection (or lack thereof) of a point with its epipolar line further classifies outliers into occlusions and mismatches. *Iterative region voting* assigns an outlier with a reliable disparity which is the most prevalent in the neighbouring support region of that outlier. *Proper interpolation* fills the rest of the outliers using an interpolation strategy which looks for reliable pixels in 16 directions. *Depth discontinuity adjustment* reduces errors around the depth discontinuities by detecting edges and comparing two costs of two pixels from both sides of an edge. If one of them has the smaller cost than the edge pixel it is used instead. *Sub-pixel enhancement* reduces quantisation of disparities using polynomial interpolation. The authors provide, a more detailed explanation of each step and report error reduction of 3.8% for all regions.

2.3 Visual Discomfort in Stereo Vision

Viewing stereo videos or images may cause discomfort and fatigue if a specific set of guidelines and rules is not followed. Any algorithm that processes stereo data needs to take these into consideration. This section explores the main causes of discomfort and how to avoid them. More details are provided in the works by IJsselsteijn *et al.* (2000), Meesters *et al.* (2004), and Tam *et al.* (2011).

2.3.1 Individual Differences

The HVS has different characteristics between people, some of which determine how well one can perceive depth. A complete lack of the ability to perceive stereoscopic depth is termed stereo blindness and it affects around 10% of the population. The survey by Richards (1970) showed that about 4% of the participants

were unable to see a hidden figure in random-dot stereogram (for an example of random-dot stereogram, see Figure 2.2), while another 10% incorrectly reported the position of the figure in relation to the background (in front or behind). A more recent study by Laframboise *et al.* (2006) also used a random-dot stereogram but evaluated how aging affects stereoscopic interocular correlation. Older participants required more corresponding dots to perceive stereoscopic stimuli. The number of points slightly increased for the individuals between the age of 45 to 64 while it strongly increased for the participants above 65. They noticed that gender did not affect results. Ostrin & Glasser (2004) show that the ability of eye accommodation - the process of focussing on an object as it changes distance - decreases with age. Finally, distance between eyes, i.e. interpupillary distance (IPD), varies between people and depends on age, gender and race. Dodgson (2004) provides statistics on IPD which spans 63 mm on average in adults. The majority fall in the range of 50 - 75 mm and it is unlikely there is anyone outside the range of 45 - 80 mm. Children (five years old being the bottom limit) have the minimum IPD of around 40mm.

2.3.2 Excessive Screen Disparity

The human visual system cannot fuse images containing large disparities. Yeh & Silverstein (1990) found that without vergence movements for short stimulus duration (200 ms) the threshold for single vision is around $27'$ angle for crossed and $24'$ for uncrossed disparity and for long duration (2 s) threshold increases to around 4.93° for crossed and 1.57° for uncrossed disparity. Multiple factors influence fusion: image properties, eye movement, length of exposure, illuminance and individual differences. The fusion limit increases when observing bigger, moving objects and by adding peripheral objects to the fixation object (Schor *et al.*, 1984; Yeh & Silverstein, 1990; Howard & Rogers, 2002).

2.3.3 Vergence - Accommodation Decoupling

Vergence and accommodation respond to proximity cues and they are connected via cross-links (Hung, 2001). However, synchronisation between the two of them can break when accommodation is focused on the stereoscopic screen while the vergence changes based on the displayed disparity. As this rarely happens when viewing real scenes, such decoupling was often considered a significant factor in

causing visual discomfort (Emoto *et al.*, 2005; Okada *et al.*, 2006; Hoffman *et al.*, 2008). However, some results show that accommodation does not stay focused on the screen constantly and moves towards the object of interest instead (Inoue & Ohzu, 1997; Ukai & Howarth, 2008). It is unclear if vergence drives this shift or if it is caused by observing the screen from close range (Goss & Zhai, 1994). In any case, as long as the shift is small enough to enable sharp viewing it should not result in discomfort but when defocus occurs the HVS tries to correct for it and accommodation conflicts with vergence (Hiruma & Fukuda, 1993). The HVS is able to cope with such a problem to some degree, but during prolonged viewing, discomfort might increase and may lead to blurred vision due to loss of accommodation and double vision due to loss of fusion (Lambooij *et al.*, 2009).

2.3.4 Zone of Comfortable Viewing

The range in which accommodation and vergence are achieved without major errors is termed *the zone of clear single binocular vision*. It constrains disparity of displayed images and needs to be considered when generating stereo content. The retinal disparity of under 1° can be regarded as a zone of comfortable viewing, under natural viewing conditions (Wopking, 1995; Speranza *et al.*, 2006). The threshold of 1° is widely accepted, even though lower recommendations have been suggested (Woods *et al.*, 1993; Jones *et al.*, 2001). The limit was calculated from the properties of depth of focus (Wopking, 1995), which is the length in front or behind the focal point for which the image is focused without reducing sharpness beyond a tolerable threshold (Millodot, 2008). Ukai & Kato (2002) show that once the limit is breached, effort to provide fusion and preserve image sharpness increasingly stresses the HVS.

Even though the zone of comfortable viewing is a relevant guideline for processing stereo content, other considerations need to be taken into account. For instance, within this zone, discomfort might occur due to large oscillations in screen disparity (Nojiri *et al.*, 2003; 2006). Yano *et al.* (2002) performed a study in which participants evaluated stereoscopic sequences. The degree of visual fatigue was rated high for sequences containing large disparity and high amounts of motion, while it was rated low for cases of large disparity but small amounts of motion. Speranza *et al.* (2006) also performed a user study which examined viewing discomfort of stereo images, but they focused on objects moving in depth. They concluded that the magnitude of disparity change could impact visual com-

fort more than the overall video disparity, which is also the case with frequent changes between uncrossed and crossed disparities.

2.3.5 Stereoscopic Impairments

Stereoscopic distortions occur in different stages of the stereo image pipeline (capture, compression and display) and can depend on camera choice and configuration, the 2D to 3D conversion process, compression, display type, viewer's position and ambient light level (IJsselsteijn *et al.*, 2005). Distortions differ in the amount of visual discomfort caused, and when combined their effect can interact non-linearly. The literature examines various types of distortions and here a short overview of the most relevant ones is provided. For a more detailed account please refer to Meesters *et al.* (2004) and Boev *et al.* (2008)

Keystone Distortion and Depth Plane Curvature

Keystone distortion occurs when two cameras converge and so each sensor is directed at an angle towards the scene (Figure 2.6a). Such a geometrical setup results in trapezoidal image projections between left and right view and erroneous vertical and horizontal disparities (Figure 2.6b). Keystone distortion grows when distance or angle between cameras increases, as lens focal length decreases. Distortion is most prominent at the image corners and causes objects in these areas to appear further away - which is another impairment termed depth plane curvature. IJsselsteijn *et al.* (2000) advised avoiding converged camera setup, based on a study which examined naturalness and quality of stereo images. Also, Woods *et al.* (1993) suggested that large vertical disparities may cause eye-strain.

The Cardboard Effect

The cardboard effect is the phenomenon which affects the perception of depth and results with objects looking unnaturally flat in the stereo images, as if they are cardboard cut-outs. The imaged scene appears as being divided into discrete depth planes, similar to a scenery in a theatre. It occurs when the perceived size and depth of an object do not correspond and a spectator underestimates the distance to the object (Howard & Rogers, 1995). This happens for a distant object captured with converging cameras with narrow lenses. Yamanoue *et al.* (2006) evaluated the effect objectively and subjectively and showed that it is

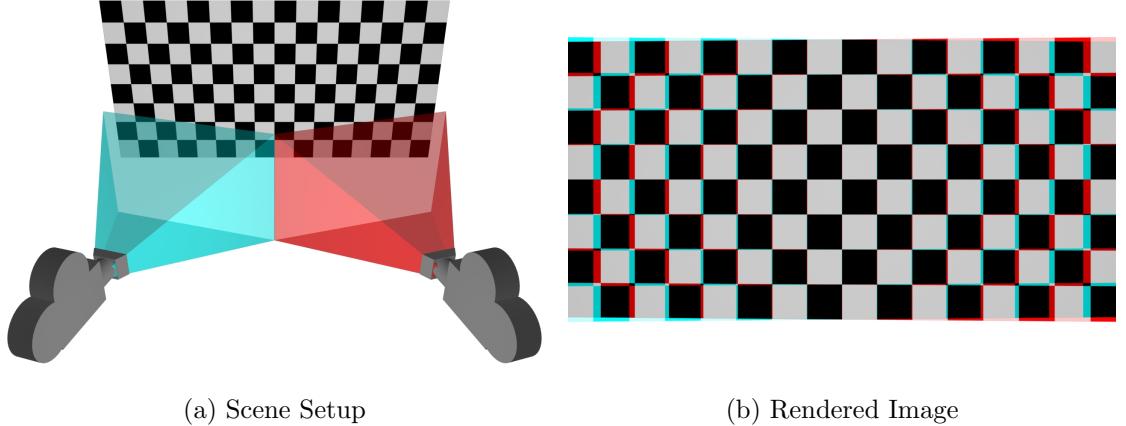


Figure 2.6: Keystone distortion. (a) Converging camera setup shooting a flat plane and causing keystone distortion. (b) The image with distortion which is prominent at the corners of the plane. Disparity should be equal across the whole plane, as depth does not change.

affected by inter-camera, object and viewing distance. This effect can also be caused by high quantization of disparity maps during compression or 2D-to-3D conversion (Boev *et al.*, 2008). In such cases an object can seem torn-up as it gets mapped to different depth planes and is perceived to be disjointed. In videos this change can occur in time so objects appear to oscillate in depth.

The Puppet Theatre Effect

The puppet theatre effect is also caused by mismatch of size and depth of an object (Pastoor, 1995). Perspective tells us that two objects of the same absolute size will have a different relative size in an image if they are at different depths - the closer object will be bigger. When perspective and stereo cues collide, the latter takes priority and the perceived object appears miniaturised (Boev *et al.*, 2008). For example, when viewing an image of a real ship on a 3D screen, disparity cues might suggest that the ship is only few centimetres away from the observer while perspective cue would suggest that, at such a short distance, the observer would only be able to see a small part of the ship on the screen. Since the whole ship is visible and disparity takes precedence, the brain interprets it as a small model of a ship. Yamanoue *et al.* (2006) showed that puppet theatre effect is correlated to prior knowledge of the object size - more familiar objects yield more effect. They also demonstrated that the effect is avoided when the scene

with a parallel camera setup where the distance between cameras corresponds to viewer's eyes. On the display side Hopf (2000) proposed an autostereoscopic technique which would minimise this effect.

Sheer Distortion

Sheer distortions occurs with displays which provide only one ideal viewing position - "the sweet spot". Almost all currently available consumer 3D displays fall into this category (Woods *et al.*, 1993; Pastoor, 1993). When the spectator moves his head away from the sweet spot the image gets perspectively distorted and as parts of the image seem to shift depending on disparity. Objects which appear behind the screen move in the opposite direction of the observer while the ones in front move in the same direction. This feels unnatural as it does not occur in reality. The effect can be counteracted by headtracking and updating the image based on viewpoint, but the amount of data and processing is increased when obtaining the multiple views.

Crosstalk

Crosstalk arises from an inadequate separation of left and right views during display. The image from the other view leaks to the current one and is perceived as a shadow, ghost or a double contour. Imperfect separation is caused by different issues in display hardware which include: poor timing and efficiency of shutter glasses, low polarisation quality and slow pixel response rate. Crosstalk has major impact on visual comfort as even a small amount can lead to headache (Pastoor, 1995). Konrad *et al.* (2000) concluded that suppression of crosstalk using his proposed algorithm improves visual comfort, while Seuntiens *et al.* (2005) showed that participants report an increase in image distortions as crosstalk increases. For more details please refer to Woods (2010) who describes mechanisms which cause this impairment, methods to measure and characterise it, and solutions which reduce and cancel the crosstalk.

2.4 Stereoscopic Capture

Stereoscopic data can be obtained in three general ways. Firstly, two stereo views can be captured by one or two cameras. This requires special attention to avoid the problems mentioned in Section 2.3. The two camera sensors can have

same characteristics (e.g. spatial resolution), or they can be different resulting in symmetric and asymmetric capture respectively. The second method involves a single image and an additional depth map which can be used to generate the other view. The depth map can be captured simultaneously with the image or it can be generated later (potentially requiring significant user input for satisfactory results) allowing for the conversion of legacy content. Finally, geometrically accurate stereo images can be rendered using computer graphics. This section explores these approaches in more detail.

2.4.1 Capturing Two Symmetric Views

The most intuitive way of capturing stereo images involves positioning two cameras so they emulate two eyes. The current state of the art in consumer products with focus on the entertainment industry are provided by Mendiburu (Mendiburu, 2009; 2011). He points out a number of technical challenges that current cameras have: size, synchronisation and control. Lenses and camera bodies can be large preventing small or normal separation. This can be overcome by positioning cameras at different angles (usually 90°) and using a beam splitter - an optical device that splits light so that half goes to each of the cameras (Heinzle *et al.*, 2011). Synchronisation of camera parameters requires special attention. Frame timings, focus and zoom control need to be handled digitally, because they need to be changed simultaneously and precisely on both cameras. Finally, in some cases, control of the camera convergence and movement is required so the camera rig needs to be robust and steady to minimise image distortions (Mendiburu, 2011). A design of a system which tries to overcome all of these problems is presented by Heinze *et al.* (2011).

The problem of distortions between images and the requirement of two camera bodies can be overcome by using a single imaging sensor. The two views are separated using optics before they reach the sensor. An early example is provided by Nishimoto & Shirai (1987) who designed a stereo camera with a single lens and a specialised glass plate which was placed in front of it, as shown in Figure 2.7a. The plate had two positions each transferring a slightly different view of the scene. While this setup simulated a coplanar camera pair and facilitated correspondence calculation, it captured only coarse disparities. A similar system was proposed by Teoh & Zhang (1984) who improved the disparity range by adding two fixed mirrors and replacing the rotating glass with a rotating mirror (see Figure 2.7b).

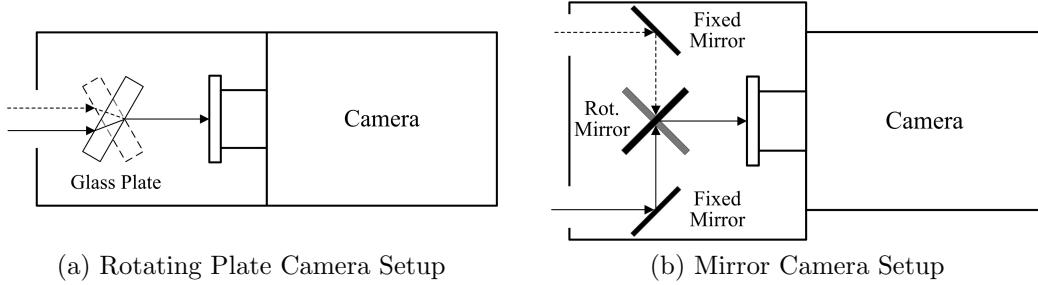


Figure 2.7: Single Lens Stereo Camera Setup. (a) In setup proposed by Nishimoto & Shirai (1987) rotating glass plate shifts the optical axis obtaining two views (b) In setup proposed by Teoh & Zhang (1984) glass is replaced with rotating mirror and two fixed mirrors are added allowing stereo capture.

Mirrors were positioned so that the setup simulated two cameras with parallel optical axes. Both systems had moving parts - a major design issue - and both needed two shots, requiring the scene to stay static.

The issue of being able to capture only a static scene was overcome by Gosttasby & Gruver (1993) by using two static mirrors and enabling the capture of the scene with a single shot. The two cameras that this setup simulates would be converged, and the angle between the mirrors could be changed to adjust the fixation point. Images were reversed so they needed to be transformed before further processing. An alternative approach of capturing stereo images with a single camera uses prisms to separate images (Lee & Kweon, 2000; Xiao & Lim, 2007). However, since these methods use only a single sensor to capture two views simultaneously, the resolution of stereo images is halved.

Some research stereo cameras have been designed for specific tasks in mind. For instance Neukum & Jaumann (2004) and Jaumann *et al.* (2007) describe the design of an advanced camera used for scanning the surface of the Mars during the European Space Agency's Mars Express mission. The camera possesses stereo colour scanner capable of capturing high definition images quasi-simultaneously using nine CCD sensors. Images are captured at five phase angles and in four colours.

Finally, the number of cameras that capture the scene can be larger than two (Kubota *et al.*, 2007). Such setups are the topic of the multiview imaging field which is beyond the scope of this thesis.

2.4.2 Asymmetric Stereoscopic Capture

The methods discussed in the previous section assume that both captured views are of the same or very similar quality and properties. However, there has been work which use sensors with different attributes to capture left and right images. Such approaches are especially relevant for the work presented in Chapters 5 and 6.

Hybrid Stereo Camera

Sawhney *et al.* (2001) proposed a method in which high spatial resolution stereoscopic video (at least 6,000 horizontal pixels) was generated from a pair of videos with asymmetric resolution. One video was of the target resolution (e.g. 6,000 pixels) while the other was typically a quarter of the size (Figure 2.8). This allowed for reduced image rendering times because the computational complexity of generating one of the views was independent of scene complexity and depended only on the number of pixels. In addition, the low resolution view was captured using a digital camera resulting in a smaller rig and savings in film costs.

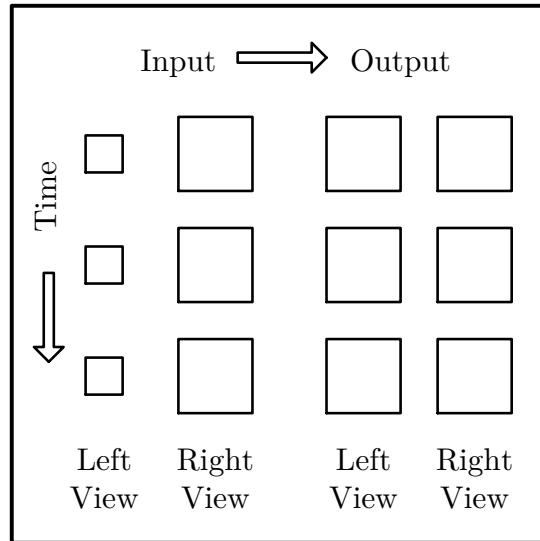


Figure 2.8: In the hybrid camera setup one of the recorded views is four times smaller than the other. The low resolution view then guides the generation of a novel view which is of high resolution.

The proposed approach for increasing the resolution of one of the views consisted of three steps: analysis, test and synthesis. The analysis step searched for a relationship both between the left and right frames, and also between consecutive

frames in a video sequence. To this end, the high resolution view was downsampled making both views the same size. To calculate dense stereo correspondences between views, the *plane-plus-parallax* method was used (Kumar *et al.*, 1994). This method modeled pixel disparities as the result of camera motion, and the motion of the planar surfaces present in the scene. Correspondences between pixels in consecutive frames (the previous two and the next two frames) were estimated using an optical flow algorithm (Horn & Schunck, 1981).

The analysis step produced multiple correspondence maps (e.g. one between views and four between consecutive frames). Any of these maps could have been upsampled and used to warp (transfer) high resolution data onto a novel view, but to increase robustness, an optimal combination was sought in the test step. A correspondence quality measure, termed the alignment map, was proposed and it represented the colour similarity between the original and pixels warped using the disparity map. A composite low resolution image was then created by combining individual warped images, where the contribution of each was determined using their alignment map as a weight. To guide the synthesis at a higher resolution, the alignment map was also obtained for the composite image.

The final synthesis step moved to high resolution by upsampling original low resolution frame, disparity maps and alignment maps. A novel high resolution frame was generated by a warping the original high resolution frame using disparity maps and compositing the results using alignment maps. The composite alignment map was used to detect mismatched pixels via thresholding. These were then replaced with upsampled pixels of original low resolution frame.

Two video sequences both lasting ten seconds were used to test the approach. The ground truth videos had 6,000 horizontal pixels for each view. For testing, one of the views was reduced to a quarter of the size and reconstructed using the proposed technique. The number of misaligned pixels was used as the objective performance measure. Results were given for a single frame from each sequence and less than 1.4% misaligned pixels were reported for both. The authors reported that it took between 20 and 30 minutes to generate a novel view for each frame, but that once the code is optimised, the required time was expected to drop to 5 minutes.

Large Camera Arrays

A multiple sensor approach was proposed by Wilburn *et al.* (2005). Large arrays of low quality video cameras were set-up to produce high quality video. The authors argued that with multiple low cost cameras it is feasible to generate similar or better video than by using high end consumer products, but engineering an entire system may require considerable assembly.

The system consisted of 100 CMOS camera sensors, 100 lenses, 100 processing boards and four PCs. Individual sensors had a spatial resolution of 640×480 pixels. Two different sets of lenses were used depending on the scene configuration. To record close up objects, 6.1 mm focal lens were used, while for far objects, 50 mm lenses were used. The processing boards allowed the cameras to be configuring and for local processing. Video frames were compressed using MPEG before sending them to the host computer which facilitated data transfer and reduced the number of required computers. The PCs either stored the collected data or enabled displaying the video data in real time. The system recorded videos at 30 fps and allowed storage of 2 GB of data (equivalent to 2.5 minutes of video) before reaching full capacity.

The authors explored different applications for the proposed arrays including increasing the resolution, frame-rate and dynamic range of the video. They also investigated simulating camera motion and large camera aperture. HDR high resolution video was achieved by arranging the cameras in a dense 12×8 array so that frames overlapped by approximately 50%. Most of the points in the scene were viewed by four of the cameras which allowed each point to be captured at four preset exposure times. The use of the autostitch algorithm (Brown & Lowe, 2003), which assembles multiple images into a single panorama, enabled HDR video generation. It was modified so that exposure times of an individual frame were considered during image assembly, and the values in the middle of the range (of each frame) were assigned a larger weight.

Results were illustrated by providing a single example image which was compared to the image captured using a consumer Canon D20 camera. The authors reported that the contrast of the generated HDR image was noticeably worse than that of the image captured using the D20. This was attributed to light leakage and lens aberration. Objective measures (e.g. achieved dynamic range and image differences) were not provided.

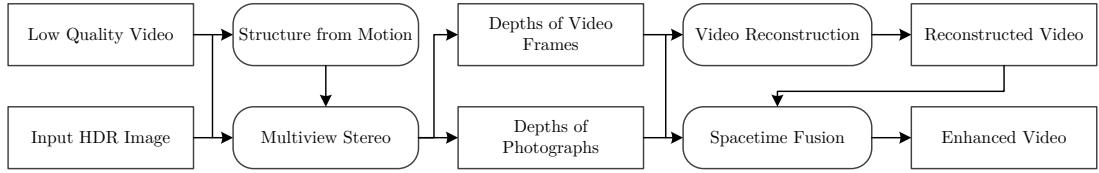


Figure 2.9: The technique proposed by Bhat *et al.* (2007) enhances the quality of captured video using photographs of a static scene. To this end SFM, multiview stereo, MRF formulation (video reconstruction) and gradient domain compositing (spacetime fusion) were used.

Using Photographs to Enhance Videos

Bhat *et al.* (2007) noticed that reasonably priced digital still cameras are capable of capturing scenes in higher resolution and better quality than similarly priced video cameras which recorded noisier video of lower resolution, so they proposed a generic approach for improving low quality video using high quality images. The suggested framework had a wide range of applications, including the improvement of spatial resolution, dynamic range and lighting, noise reduction, removal of objects and camera shake from the video and more efficient video editing. The approach was restricted only to scenes with static geometry in order to avoid the challenging problem of finding pixel correspondence in the presence of scene motion.

The pipeline of the technique is shown in Figure 2.9. To improve the photographic qualities of video (e.g. dynamic range) the proposed technique took a video and a set of images of the same scene as an input. The framework consisted of two main stages. In the first stage geometry of the imaged scene was reconstructed. The structure from motion (SFM) algorithm of Snavely *et al.* (2006) was used to generate camera poses and sparse cloud of 3D scene points. The output of SFM technique was used as a priori for modified multiview stereo algorithm Zitnick & Kang (2007) which generated dense depth maps for photographs and video. These maps identified pixel correspondences between photographs and video frames but also between consecutive video frames. In the second stage, correspondences were utilised to enhance video quality while preserving its temporal dynamics. MRF formulation was used to select a photograph to be used for transferring colour to pixels of each video frame. Enhancing frames in this manner resulted in visible spatial and temporal seams when large exposure variations in photographs were present. Also holes appeared in regions of reconstructed video which lacked corresponding photograph patches. To overcome this, compositing

regions from several photographs was performed in the gradient domain where motion-compensated temporal gradients were obtained from video, and spatial gradients were obtained from the photographs. HDR videos were generated in this manner by using LDR video as an input and a set of HDR photographs.

Quantitative results were lacking and only a single qualitative example for each application was presented. The authors reported that under careful examination some artefacts can be observed in most of the results and that they were caused by the errors in the output of the computer vision algorithms. Finally, very slow computation speeds (five minutes per single low resolution image) were reported.

HDR from Stereoscopy

Lin & Chang (2009) suggested a form of creating HDR images using a stereo setup where each image of the pair was taken at a different exposure level and subsequently combined. As the images were captured at the same time, it was possible to avoid ghosting artefact due to motion. However, another type of ghosting was introduced by the parallax of the separated cameras.

The proposed algorithm took as an input a stereo pair where each image was captured with different exposures by using different shutter speeds. To allow accurate HDR merging and stereo correspondence generation the CRF was required. As the images were not overlapping SIFT matching was performed to find corresponding sample points used for calculating the CRF. Dense correspondences were calculated by first scaling images so they were in the same exposure range using CRF. Then the stereo matching algorithm proposed by Sun *et al.* (2003) was applied in a straightforward manner. A disparity map was used to transform one image so it was aligned with the other and an HDR image was generated via a weighted sum of these two images.

The method was tested on the Middlebury stereo dataset, but only two qualitative results were presented and relative quantitative data were lacking. The disadvantage of the proposed method was that in order to generate a satisfactory disparity map used for calculating HDR values, over and under-exposed regions need to be avoided in both images. This forced the image pair to be captured using very close exposure values significantly limiting the potential dynamic range of generated images. This was seen from the examples provided by the authors.

Evaluation of the Asymmetric Approach

The effects of a stereo pair with mixed resolution images were explored by Stelmach *et al.* (2000) and Lo *et al.* (2009). However, instead of seeking to improve the low quality image they tested how the HVS copes with the dissimilarity.

Stelmach *et al.* (2000) wanted to low-pass filter and thereby compress one of the views while preserving the perceptual quality. Filtering was performed in both the spatial and the temporal domains. Spatially, images were filtered at $1/2$ and $1/4$ resolution. Temporal filtering was performed in two variants: averaging pixels of consecutive frames, and replacing every other frame with the previous one. Video filtering was performed in real time. In a user study, participants rated the overall quality, sharpness and sensation of depth. Two video sequences of ten seconds each were used for testing. Results suggested that low-pass filtering did not affect the perception of depth, while it did strongly bias the weighting of quality and sharpness towards the eye observing the high quality image. Temporal filtering resulted in the perception of images with poor quality and sharpness while the perception of depth was mostly unaffected.

In a similar experiment, Lo *et al.* (2009) tested whether rendering times of stereo images could be decreased by reducing the resolution of one of the views while still preserving the same image appearance. The approach utilised binocular fusion - a process in which the HVS fuses two percepts presented to each eye into a single view. A single rendered scene was used for testing. In the stereoscopic asymmetric test condition, the resolution of one image in the pair was decremented in 10 steps generating 10 novel images while the other was kept at the maximum. Such images were compared to the ground truth (GT) - a stereo image where both views had the maximum resolution of 800×800 pixels. Results showed that less than 15% of participants could differentiate between the GT and the image pair with one image at the reduced resolution of 640×640 . Once the image was reduced to 320×320 , more than 50% of participants could detect the difference.

2.4.3 Single View Plus Depth

Stereo images can be generated from a single image and a depth map (see Section 2.2.4) which is utilised to obtain the second view using image based rendering (IBR) techniques. Based on the way the depth map is obtained we can distinguish between the two general methods: *direct*, where the map is captured using



Figure 2.10: Kinect is capable of capturing depth map and colour map at resolution of 640×480 pixels. Courtesy of Evan Amos.

specialised hardware; and *indirect*, where the map is absent and depth needs to be estimated from a single image. The second method is more frequently termed 2D-to-3D conversion and is of specific importance for making legacy content available in stereoscopic format.

Direct Depth Map Capture

Depth maps can be captured using cameras, which in addition to an imaging sensor also have depth sensors. For example, ZCam (Iddan & Yahav, 2001) measured the time that projected infra-red light took to reflect back to the sensor. This measurement inferred distance of objects to the camera. Depth images were captured at the same frame rate as the corresponding video. Similarly, Kawakita *et al.* (2002) used an infra-red LED array for capturing depth with their HDTV Axi-vision camera at a resolution of more than 920,000 pixels at 30 frames per second. Microsoft (who bought the company that produced ZCam) developed the Kinect - an input device for the game console XBox which is capable of capturing depth (Khoshelham, 2011). Kinect (see Figure 2.10) has been used by the research community for different applications (Henry *et al.*, 2010; Oikonomidis *et al.*, 2011). Alternative methods described by Scharstein & Szeliski (2003) project structured light pattern onto a scene. Shapes and distances of the objects caused distortions in the pattern which was analysed to obtain depth. A limitation of all the mentioned methods is the range and the precision of captured depth data. ZCam had range from 1 to 10 m with resolution of 0.5 cm (for distances of 1 m), Axi-vision had unreported range with resolution of 1.7 cm (for distance of 2 m), Kinect has a range of 0.7 to 6 m with resolution of 640×480 pixels.

2D-to-3D Conversion

As stereoscopic imaging is just entering the main-stream market, there is a lack of content natively filmed in 3D, so there is a need to generate 3D content from currently available images and video using 2D-to-3D conversion. The conversion also enables moving legacy 2D content to the new medium. In the entertainment industry conversion is sometimes considered as a cheaper option than shooting a movie in native stereoscopy, but still requires a lot of manual input in post-production and can produce lower quality video (Mendiburu, 2009). An automatic approach would reduce the cost of conversion significantly, but it is a challenging ill-posed problem. Currently there are a number of solutions, most of which generate satisfactory results for a specific set of scenes and conditions. Most of the methods are based on the fact that depth can be inferred from monoscopic cues such as blur, texture gradient, geometric perspective, and shading. This section describes representative methods for automatic 2D-to-3D conversion, which can be divided into four groups: depth from blur, depth from geometric constraints, depth from motion, and other methods.

Depth From Blur

Depth from blur method is inspired by the relationship between an object's depth and the object's blurriness in the image for a given focal length of lens. The depth can be obtained by modelling the effect of varying focal parameters on the image (Lai *et al.*, 1992; Yokota *et al.*, 2005). Inverse filtering determines a defocus operator and estimates scene depth (Ens & Lawrence, 1993). Valencia & Rodriguez-Dagnino (2003) proposed a method, which did not rely on camera lens parameters, but analysed a local region using multi-resolution wavelets. Blurred regions had a larger number of zero wavelet coefficients compared to sharp regions which enabled depth calculation. Tam *et al.* (2005) suggested using luminance intensity gradients which contained edges and textures for generating depth maps. Their reasoning was that increasingly sharp edges would produce increasingly sharp gradings, but regions which lacked edges were problematic.

An inherent problem for all depth from blur methods is that blur can be absent or, more importantly, can be caused by other factors, such as: lens aberration, fuzzy objects, motion and atmospheric interference. Even though there are methods which try to counteract some of the challenges, the application of this method is still constrained to specific cases (Chaudhuri & Rajagopalan, 1999).

Depth From Geometric Constraints

Geometrically based methods for recovering depth from single images use gradient and linear perspective cues. In the method proposed by Battiato, Curti & Cascia (2004), colour segmentation first classified an image in the categories heuristically (e.g. indoor, outdoor). Category and identified straight lines helped in estimating vanishing points and vanishing lines. The vanishing “point” was the region where most straight lines intersected, while vanishing lines were the main straight lines passing close to vanishing points. Slopes of the vanishing lines determined depth gradient planes which were combined with segmented data to produce convincing depth maps (Battiato, Capra, Curti & La Cascia, 2004). This approach was limited to specific scene types and could not be generally used.

Depth From Motion

Image sequences provide an additional depth cue of motion parallax - a phenomenon where objects at different distances from the camera, moving at the same speed get displaced differently on a recorded image. Objects closer to the camera get more displaced. Different techniques can be used to analyse motion and infer depth, such as occlusions (Ogale *et al.*, 2005), optical flow (Gautama & Van Hulle, 2002) and motion vectors (Moustakas *et al.*, 2005). However, these methods are able to provide ordinal depth only for moving objects and can rarely be used to produce full depth map for a frame.

Okino *et al.* (1996) proposed a hardware implementation of the *Modified Time Difference* method which allowed real-time conversion of 2D content and was based on object motion detection. Results showed that such an approach worked satisfactorily only for simple cases. Other motion based approaches were proposed by Zhang *et al.* (2005), Kim *et al.* (2008) and Pourazad *et al.* (2009). Motion between two frames is usually not large enough to produce consistent and satisfactory results. In addition layered motions and occlusion are difficult to handle which limits the application of depth from motion techniques.

Other Methods

Other methods for automatic 2D-to-3D conversion have also been proposed. Tam *et al.* (2005) suggested using sparse depth maps which they termed *surrogate depth maps*. The maps were generated using edge information so disparities were only present in regions corresponding to object boundaries in the 2D images and

were missing from large low frequency regions. In a user study the perceived depth of the generated images was deemed adequate. Saxena *et al.* (2007) proposed a supervised machine learning approach which they trained using a set of monocular images and ground-truth depth maps. Results showed the potential of such an approach for generating depth maps. However, these might lack precision for satisfactory image based rendering. Feng *et al.* (2011) proposed a method which converted 2D videos to 3D by combining the techniques of optical flow, occlusion reasoning, object segmentation, content based matching and depth-ordinal based regularisation. The authors validated the effectiveness of their approach through objective and subjective evaluations of the qualitative and quantitative performance of their algorithm.

2.4.4 Stereoscopic Rendering

Computer graphics (CG) techniques allow rendering of geometrically accurate images without any distortions. Many commercially available 3D computer graphics software packages support virtual stereo cameras (Autodesk Maya, Autodesk 3ds Max, NewTek LightWave 3D) while Nvidia with its 3D Vision line enables stereoscopic video gaming. One of the main challenges of rendering is improving performance. When rendering two views for stereoscopy, the time required could potentially double so a number of optimisation strategies have been proposed.

Adelson & Hedges (1993) optimised the ray tracing rendering algorithm so it computed the second view with as little as 5% overhead compared to a single view. Their method used reprojection which moved pixels from one image to the inferred position in the other and recalculated unknown or potentially erroneous pixels. Es & Isler (2007) adapted this approach to work on graphic processors in real time. More recently Andersson *et al.* (2011) suggested an efficient ray tracing algorithm for multiview rendering from a camera line that was based on multi-dimensional adaptive sampling light field reconstruction.

Hasselgren & Akenine-Möller (2006) proposed a novel multiview rasterisation architecture. The authors assumed that texturing consumes most of the memory bandwidth, so they suggested rasterising a triangle to all views before proceeding with the next one. As the same texture is applied to a triangle for all the views, more hits were expected in the texture cache. They also modified a pixel shader so that exact evaluation of the pixel shader was done for a single view while for the others the evaluation was reused if possible, thereby avoiding many per-pixel

shader instructions. A final optimisation was enabling all views to access a single colour buffer while keeping their own depth and stencil buffers, so that effects such as depth of field could be rendered faster. Results showed that the algorithm reduced bandwidth usage compared to simple rendering of two views and that executing the pixel shader could be avoided for up to 95% of the fragments.

An optimised rendering pipeline for stereoscopic or multiview images was proposed by Kalaiah & Capin (2007). In their single-pass solution, the vertex shader was split into two parts: view-independent and view-dependent. The first part held information such as the light sources and the surface material while the second held the projection matrix, the viewport matrix and the eye position. This allowed the shader to execute expensive view-independent computations only once. The authors emulated this architecture on a standard graphics card and the results showed that for a single view, performance was worse than the conventional method, while for the stereo case the average frame rate across six tested scenes improved by around 35%.

Bulbul *et al.* (2010) suggested a perceptually based rendering optimisation technique. The method relied on the effect of binocular suppression - a phenomenon where perception of a stereo image in a region is determined by the dominant image in that region. This meant that it would be feasible to render the less dominant view in lower quality so the authors proposed a measure of a view strength, named intensity contrast. This method identified which view is more dominant and decided if decreasing quality would affect perception. A user study evaluated quality, depth, comfort and sharpness of the stereo pair where one image was rendered with low quality attributes. Results suggested that the perceived quality could be preserved while improving performance when changing framebuffer upsampling, specular highlight, shading and texture resampling, while perceived quality decreased when changing antialiasing and shadowing.

2.5 Stereoscopic Content Storage

Stereoscopic content contains two images or streams, one for each eye, hence, doubling the amount of data compared to a traditional 2D representation. As current infrastructure (e.g. bandwidth, medium sizes) was primarily aimed at 2D images and videos, stereoscopic data should be compressed in order to facilitate its processing and transmission. This would also help the transition from 2D to

stereo technology and speed up its adoption by the community. While both the left and right views can be compressed separately, a good coding algorithm should exploit the fact that both views are highly correlated with small differences which result from camera separation. This section describes the more popular formats for storing stereoscopic data. Gorley (2012) offers a similar overview in his thesis.

Digital stereoscopic imaging is relatively novel so file formats are still mostly unestablished and new stereo standards are still emerging. Current stereoscopic file formats are mostly based on existing image and video formats which get extended to store two or more views. As extensions, these formats inherit properties of the original formats, including the type of compression (e.g. lossy or lossless), coding methods (e.g. discrete cosine transform) and meta information (e.g. colour statistics). Another advantage of extending existing 2D file formats is backwards compatibility. Stereo images usually contain tags or flags in the file description which indicate to the application that data about to follow represents a second view. Legacy applications are unable to recognise these tags, which are ignored and only a single view gets displayed. However, while enabling compatibility and simple implementations, these file formats do not implicitly take into account similarity between stereo pairs.

2.5.1 Animated Graphics Interchange Format (GIF)

GIF (CompuServe Incorporated, 1990) is a bitmap image format which stores images using up to 8 bits per pixel (bpp). These values identify up to 256 colours which are stored in a separate colour map. GIF uses the lossless *LZW* coding algorithm to compress data. A small size is one of the main advantages of GIF, which made it popular with internet users. Also, it can store number of frames thereby enabling short animations (Animated GIF).

The animation feature is especially interesting for stereoscopy, where left and right views can be stored as consecutive frames which alternate when viewed. The technique, informally termed *Wobble GIF*, results in motion parallax and produces the sensation of depth. Standard web browsers or image viewers open animated GIFs thereby enabling delivery of stereo to a wide audience without the need for special hardware or software. However, binocular vision is not achieved so depth cannot be fully appreciated and animation can distract from the actual content of the image. The number of colours is limited and GIF is unable to reproduce more complex scenes with high quality

2.5.2 Stereoscopic Portable Network Graphics Format (PNS)

Portable Network Graphics (PNG) format was developed to improve upon GIF and circumvent its patent license. It supports 24 bpp (or 32 bpp when using an alpha channel) allowing for a much larger colour palette and usage of RGB colour space (Roelofs, 1999). PNG uses filtering and the *DEFLATE* coding algorithm to losslessly compress images. The format is separated into a header section which is followed by data segments named chunks. Chunks can be: critical - it needs to be read by any application opening a PNG file; and ancillary - it can be ignored if the application cannot read it. This enables backward compatibility. Chunks indicated by the letters “*sTER*” identify a stereo images which are saved in a side-by-side manner. One bit is used to indicate whether the left image comes first. When opened by a supported viewer, the image can be displayed in an appropriate manner, but when opened in a traditional viewer, both images will be displayed side by side. The same behavior, without internal file formatting, can be achieved by saving a PNG file, consisting of a side-by-side stereo pair, using the PNS file extension.

2.5.3 Stereoscopic JPEG (JPS)

Joint Photographic Experts Group (JPEG) format is the most commonly used image format. It uses a lossy method for storing images (lossless is also supported but not used as frequently), achieving higher compression ratios compared to a lossless method with a potential loss of quality (Ghanbari, 2010). Encoding can be implemented in different ways most of which perform the following operations: the image is converted to Y'C_BC_R colour space and chroma components (C_BC_R) are downsampled by 2; the image is separated into 8×8 pixel blocks which are transformed using discrete cosine transform. Frequency amplitudes are then quantised based on frequency and finally lossless *Huffman coding* is used to further compress data. Similar to PNG, JPEG is separated into segments identified by markers. Stereoscopic JPEG images have the JPS file extension and use an *APP3 (application specific)* marker to indicate the format used to save the stereo pair: interleaved, side-by-side, over-under and anaglyph (Siragusa *et al.*, 1997). Viewing behaviour is also similar to PNS, so supported applications will show the image appropriately while traditional viewers will show both images the way they were stored.

2.5.4 Multiview Video Format (MVC)

Advanced Video Coding (AVC), also referred to as MPEG-4 Part 10 or H.264, is a video format which is currently the most popular format for storing videos. It is used by Blu-ray Discs, YouTube, Adobe Flash Player and HDTV broadcasters. AVC consists of a family of standards, which provide versatile functionalities, only a subset of which may be relevant for a specific use (Ghanbari, 2010; Richardson, 2010). It supports both lossless and lossy coding using techniques including: wavelets, discrete cosine transform, entropy coding and deblocking filters. Broadly, AVC can be described as a block-oriented motion-compensation coder. Block-oriented refers to the idea of separating the image into blocks which are compressed separately (similar to JPEG). Motion-compensation is a technique for predicting a frame from previous or subsequent frames, based the fact that consecutive frames in video are alike.

As the same fact holds for stereoscopic and multiview images, AVC was extended to include *Multiview Video Coding* - a specification which also uses motion-compensation to predict the second image of stereo pair. The process is illustrated in Figure 2.11. Frames from two stereo streams are coded using MVC specification. The frame used as a reference is termed an *I-frame (intra-coded picture)* and is coded independently from any other image. In general, *P-frames (predicted picture)* store differences from the previous frame, while *B-frames (bi-predictive picture)* store differences from previous and following frames. In the stereo case, the other view is also included in the prediction. I-frames are the largest in size as motion compensation is avoided, while B-frames are smallest.

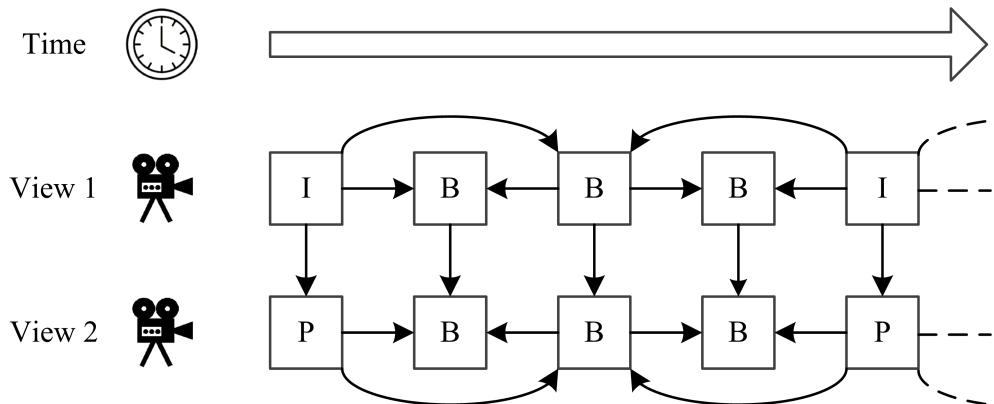


Figure 2.11: Multiview video coding relies on the motion compensation, where some frames are predicted from the previous and the following frames, and from the other view.

However, B-frames imply that sections of video need to be loaded before they are played to enable reconstruction.

The MVC approach is effective in the case of multiview and improves with the number of additional streams compared to coding each stream individually using MPEG-4. However, in the stereo case gains are rather limited, because temporal prediction performs well already and additional inter-view prediction does not significantly affect coding efficiency. A major improvement is achieved in case of I-frames of the second view which get coded as P-frames, but they typically occur every 0.5 - 1 s.

2.6 Stereoscopic Displays

The main problem of displaying stereo images is delivering left and right views to corresponding eyes while minimising interference. Additional challenges are created in cases where multiple observers watch stereo content or when a single observer changes position while watching content. Different techniques have been proposed for displaying stereoscopic images and videos, and all of them have some trade-offs. Devices can be broadly divided into three categories: stereoscopic direct-view displays, head-mounted displays and autostereoscopic displays. This section explains concepts and provides a brief overview of stereo displays. Detailed literature surveys, which describe many specific implementations, are provided by Holliman (2006), Benzie *et al.* (2007) and Urey *et al.* (2011).

2.6.1 Direct-View Stereoscopic Displays

Direct-view stereoscopic displays require specialised eye-ware to separate left and right views (see Figure 2.12). Wearing glasses is usually undesirable for users, which is the main drawback of this approach. Techniques that these displays use can be categorised into: colour separated, polarisation separated, and time separated.

Colour-Separated Displays

One of the earliest methods used to separate stereo images was using colour (anaglyph) glasses (Rollmann, 1853). In this case the colour of glass/film filters the appropriate view from the stereo image which is colour coded. Typically



Figure 2.12: Stereoscopic Glasses and Head-Mounted Display (a) Anaglyph glasses are the most affordable, but provide low quality view separation. (b) Polarised glasses are light, affordable, separate images well, but require specialised display hardware to generate polarised image. (c) Shutter glasses separate images well, but are more expensive and bulkier than the other two. (d) Head-mounted displays provide complete separation, but are expensive and heavy.

chromatically opposite colours such as red and cyan colours are used for coding. Anaglyph glasses are cheap, and anaglyph images can be displayed using traditional media including 2D screens and paper, making this approach easily accessible. However, it is also one of the worst methods to separate stereo images as it loses colour information, increases crosstalk (see Section 2.3.5) and can create visual discomfort during prolonged viewing especially if colour coding is not performed well (Dubois, 2001).

Polarisation-Separated Displays

Polarisation-separated displays use the state of polarisation of light to separate left and right images. A projector or screen polarises each of the images in one of the two mutually orthogonal directions while polarised filters on glasses

respond to a single view. Linear or circular types of polarisation can be used, with latter being more desirable as it enables larger head tilt before crosstalk becomes apparent. This method is frequently used in cinemas but is not limited to projectors only, and flat panel displays are also available (Wu *et al.*, 2008). The advantages of the method include: cheap glasses, high resolution images and good colour reproduction. However, a special screen is required to preserve the polarisation of the image, with front projection using silver screen and back projection using Fresnel-lenticular surfaces. Also two projectors might be needed (each having different polarisation filters), but single projection solutions are in use as well (Bogaert *et al.*, 2008). Inadequate choice of the projection lens or the screen can cause ghosting, hot spots and falloff of intensity.

Time-Separated Displays

Time-separated displays exploit the concept of vision persistence. Here left and right images are displayed consecutively on the screen at high frame rates (usually 120 frames per second) while active shutter glasses completely block the view to one of the eyes and are synchronised with the screen. An infrared emitter usually synchronises the glasses with a screen. This technology minimises crosstalk, enables viewing of images in full resolution and can use traditional, high refresh rate screens. On the other hand, the cost of glasses is increased as they are battery powered, and a higher frame rate requires a higher bandwidth. This technology is popular with PC and laptop users, because they are usually used by an individual requiring a single set of glasses. Nvidia's 3D vision technology, which uses shutter glasses (Boher *et al.*, 2010), enables playing most PC games in 3D.

2.6.2 Head-Mounted Displays

Head-mounted displays (HMD) deliver the image to two (or more) small screens mounted in front of the user's eyes. The left and right images are completely separated so crosstalk is avoided, and because the device is head-worn, the user is able to move without interrupting viewing. This improves the feeling of immersion which can be even further enhanced using HMDs with head tracking and see-through HMDs. Head tracking provides information about the device's position and orientation which can be used to update the view and avoid unwanted effects such as sheer distortion (Section 2.3.5). See-through HMDs are

especially useful for augmented reality applications because they can superimpose computer-generated images and information over real-world scenes (Ferrari *et al.*, 2009). Increased immersion has made HMDs a popular choice in training applications and simulations. However, they still come with a set of problems. A large field of view is hard to achieve, but this challenge was tackled using the method of optical tiling where eye tracking is used to combine a high resolution inset image and a low resolution background. The HMDs can be bulky and heavy which can be distracting during viewing, and they are also expensive.

2.6.3 Autostereoscopic Displays

Autostereoscopic displays separate left and right views without the observer wearing any special eye-wear. While this enables viewing stereo content in a more traditional and comfortable way, these displays usually have other undesirable properties, including decreased image resolution, limited viewing angles and increased crosstalk. Autostereoscopic displays can be categorised further into: two-view, multiview and head-tracked displays.

Two-View Autostereoscopic Displays

Two-view displays present a single left and right image which is independent of observer's position. Images can be separated using the parallax barrier system or the lenticular system. Both systems display left and right image columns alternatively (see Figure 2.13). A physical barrier made of vertical strips can be put in front of the screen to block one of the views for one eye resulting in the parallax barrier systems. Alternatively, cylindrical lenses can be used to direct light from pixels to specific points in space resulting in a lenticular system. These elements can be disabled, resulting in traditional 2D displays. Depending on the screen design, the stereo image can be observed correctly either from a single location (a "sweet spot") or from a number of predetermined locations. In the latter case, the same pair will be displayed resulting in sheer distortion (Section 2.3.5). If viewed from positions other than the intended viewing positions, the stereo effect can be lost and the number of artefacts can be large, increasing the chance of eye fatigue. Also, because the screen is displaying left and right images simultaneously, half the resolution is lost.

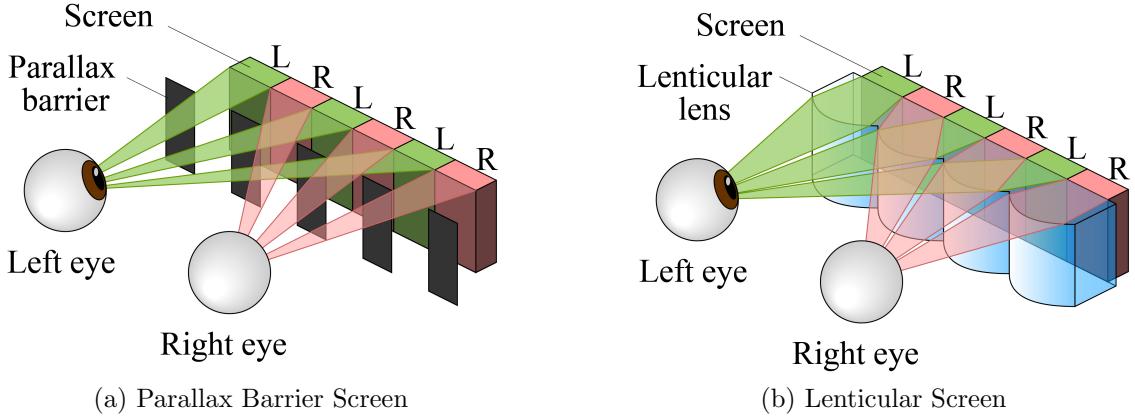


Figure 2.13: Two-View Autostereoscopic Displays (a) Parallax barrier screen blocks one of the views using vertical strips. (b) Lenticular system uses cylindrical lenses to deliver a view to its target.

Multiview Autostereoscopic Displays

Multiview displays are similar to the two-view ones, in that they provide a correct image at specific locations. However, a different stereo image is displayed to each of these locations corresponding to the projection of the scene to that point. In this way sheer distortion is avoided and the impression of depth is strengthened. Multiview displays can also use parallax barriers and lenticular technology to separate views or they can use time separation. However, the number of views they generate is small, so they are unable to provide continuous motion parallax. Multiview displays have found an application area in the mobile phone industry (Flack *et al.*, 2007), as mobile phones usually involve a single user capable of rotating a device and finding an optimal viewing position easily.

Head-Tracked Displays

Head-tracked displays are able to follow the position of the observer and project the stereo pair to the current position (Sexton *et al.*, 2006). Eye position can be tracked using computer vision algorithms which analyse video at high frame rates. The perception of the image is similar to direct-view displays, but glasses are not required when using this method. Technologies which deliver images to the intended positions include: traditional and Fresnel lens, projectors, parallax barrier, prisms and holography. Depending on the design, displays can support one or multiple users. A single viewer version is a commercially available, while

multi-user ones are still prototypes. Head-tracking using video is challenging for cases where user(s) are watching TV in extremely dark or bright environments.

2.7 Summary

This chapter has served as an introduction to many concepts related to stereoscopy, such as epipolar geometry, the fundamental matrix and disparity maps. In addition, it presented causes of visual discomfort and suggested how these can be avoided in order to produce acceptable stereoscopic content. Finally, the full stereoscopic imaging pipeline (capture, storage and display) was described and major techniques used by each of the steps were discussed. The next chapter aims to do the same for high dynamic range imaging. Special attention was paid to the concepts (and their implementations) that will be used later in the thesis. The next chapter aims to do the same for high dynamic range imaging.

CHAPTER 3

High Dynamic Range Imaging

The change from traditional digital low dynamic range imaging to high dynamic range imaging might seem deceptively straightforward: instead of 8-bit integers simply use floating point numbers to represent the full range of colour and luminance. However, such a switch brings a whole set of challenges which need to be overcome. For instance, representing the full colour range significantly increases the amount of data. To accommodate for this, the whole image architecture - designed and standardised for LDR data - needs to be extended and modified. Each step of the imaging pipeline is affected: capture, storage and display.

Theoretically, HDR imaging is able to represent the full range of light visible to the human eye (even though it might be expanded to the invisible part of the spectrum), but existing consumer devices are unable to capture this range in one go. Researchers have suggested techniques which allow HDR image generation from multiple LDR images and designed prototype cameras able to record HDR video. This increased amount of data pushes the boundaries of current transfer and storage techniques. Most of the existing coding algorithms are based on the perception of LDR images and cannot be naively used with HDR data, which is also true for the majority of image formats that only work with 8-bit data and integers. Once an image is captured it still cannot be displayed or printed in its native form, because current displays are limited in the amount of light they can emit. Similar to cameras, research prototypes of HDR displays are becoming available, but they are still not part of the main consumer market.

Backwards compatibility plays a major role when working with HDR data (especially for storage and display) as it is unrealistic to expect a sudden switch from LDR to HDR technologies. Instead, sensible solutions need to support both of them, facilitating the transition, thereby improving the chances of the public

embracing HDR.

This chapter examines the challenges HDR technology poses and reviews solutions proposed by the research community. It explores each section of the HDR imaging pipeline and focuses on concepts relevant for this thesis. Techniques to capture HDR data are presented first with a detailed overview of expansion operators. This is followed by a description of HDR storage file formats and HDR coding techniques. Finally, displays used to present HDR data are examined. Further details about these concepts can be found in the work by Reinhard *et al.* (2010) and Banterle *et al.* (2011), while a more practical and technical account of HDR imaging is provided by Bloch (2007).

3.1 Theory of High Dynamic Range Imaging

As an emerging field, HDR imaging is still establishing standard terminology. There are some inconsistencies between the meanings of terms, especially given that many have been borrowed from related fields of photography and signal processing. This section explains the key ideas of HDR imaging and defines terms used throughout this thesis.

Dynamic range is the difference between the brightest and the darkest luminance value present in a scene, image or on a display. Alternatively, it is termed contrast ratio and it can be represented using the ratio notation, e.g. 100 : 1. It is customary to scale this ratio making the lowest value one. Formally, this is expressed as:

$$CR = \frac{L_{\max}}{L_{\min}} : 1 \quad (3.1)$$

where CR is contrast ratio, and L_{\min} and L_{\max} are the minimum and maximum luminance values. Black is not considered the darkest colour, as the fraction would contain division by zero, therefore the next smallest value is used. This concept is illustrated in Figure 3.1 using an image with a contrast ratio of 10 : 1.

The human visual system (HVS) senses light nonlinearly. The Weber - Fechner law (Weber, 1834; Fechner, 1838) implies that the eye perceives brightness approximately logarithmically. This means that perceived intensity differences at lower light ranges are larger than at higher ones. For example, increasing the contrast of an image by 50 from 150 : 1 to 200 : 1 impacts perception significantly, but the same increase from 10,000 : 1 to 10,050 : 1 may pass unnoticed.

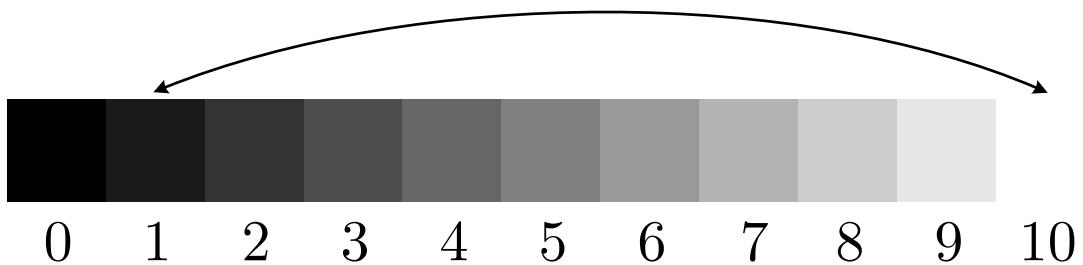


Figure 3.1: Example of 10 : 1 dynamic range. There are eleven different levels of brightness with corresponding values and the difference between the consecutive blocks represents the smallest change in brightness. Black is not included in the calculation.

This is why fractional representation, as shown in Equation 3.1, is not especially intuitive when discussing dynamic range. Instead, it can be expressed using the *exposure value* measure.

Exposure value (EV) is the term used in photography to represent the amount of light passing through the lens and hitting the sensor (Allen & Triantaphillidou, 2010). It is the relative amount of light (scene and camera dependent) and it is controlled using a camera's shutter speed and aperture size. An exposure value of zero is captured with aperture fully open and with an exposure of one second, as is defined by the International Organization for Standardization (ISO). The unit on this scale represents doubling or halving the amount of light. For instance, increase from EV zero to EV one can be achieved by halving the exposure time to half a second, by halving the aperture size or by changing both the exposure time and the aperture size so that the amount of light reaching the sensor is halved.

The EV measure corresponds to doubling and halving of light, such that EV follows a binary logarithmic scale. The HVS responds to light in approximately logarithmic manner as well, making the EV scale a more natural measure of dynamic range than contrast ratio. Conversion between contrast ratio and exposure value is straightforward: $EV = \log_2(CR)$ and $CR = 2^{EV} : 1$. The exposure value is also termed *stop*. While *f-stop* in photography is strictly used to denote the aperture size it is sometimes used synonymously to EV to represent dynamic range.

Another important question is what *high* dynamic range is. The human eye is capable to differentiate contrast of about 14 EV at any point. However, it can

adapt to lighting conditions which span approximately 30 EV. For example, sunlight can be 100,000,000 times brighter than starlight (Ferwerda, 2001). Current imaging technology does not even approach such ranges and is only able to cover about 8 EV. These images are termed *low dynamic range* images. A true high dynamic range image would capture all the intensities of light visible to the human eye, but sensors which would enable such capture in one go are not available at the moment. EVs of 30 and 8 are the two points on the dynamic range scale which can clearly be categorised. However, there is no defined dynamic range threshold which would split it into high and low, so it may depend on individual interpretation. Also it is worth noting that HDR imaging is not limited to the spectrum observable by humans and can capture data outside of it (e.g. infrared data).

Alternatively the dynamic range categorisation may be based on the data format used for storing the images. Currently most image formats dedicate eight unsigned bits per colour channel (bpc). For RGB images this totals 24 bits per pixel (bpp). Integers of 8 bits limit the maximum value to 255, so the dynamic range cannot be larger than 8 EV. Instead of integers, floating point numbers may be used to represent intensity values. This enables greater precision and does not give an upper bound, so, in theory, it can store any dynamic range present in the scene. Again, it is not possible to categorise images by just observing the data format used for storing an image. For example, capturing a scene of an overcast day is unlikely to produce sufficiently high dynamic range whether stored as floats or integers.

While 8 bit integers are unable to provide more than 8 EV, increasing the bit depth does not necessarily equate to increase in dynamic range. This is especially the case with consumer cameras which use proprietary RAW formats with increased bit depths but which are mostly used to increase precision (usually in the midrange) and not the dynamic range. If a RAW image is saved using 16 bits per channel it does not mean that camera is able to capture a dynamic range of 16 EV.

3.2 High Dynamic Range Capture

Current consumer products are limited in the range and resolution of light they record. Some expensive, high end models, support 16 bpc (Phase One IQ180)

or 14 bpc (Nikon D800), doubling what is usually available in digital images. However, manufacturers try to improve precision and reduce quantisation at the top of the intensity range, so an increase in dynamic range does not correspond to an increase in bpc. Video cameras face an even more difficult challenge. They are expected to record up to 30 frames per second or more generating large quantities of data and limiting the time available to expose sensor(s) to light.

The research community has proposed a number of different solutions for capturing and generating HDR images and videos. They can be categorised into four broad groups. Multiple exposure techniques combine LDR images captured at different exposures to generate HDR content. A more convenient way relies on devices that capture HDR natively, but currently they are in the realm of prototypes but are expected to become readily available in the near future. The third method converts the legacy LDR content to HDR using expansion operators. This task is challenging because the necessary data is lacking, but it can be estimated by imposing assumptions of scene configuration and light behaviour. Finally, HDR images and videos can be generated using existing rendering algorithms and saving the output in an appropriate HDR format. This section overviews the first three methods in more detail, while the fourth uses standard rendering algorithms, which are not examined in this thesis (Dutre *et al.*, 2006). More focus is put on LDR to HDR conversion as two such techniques are used in Chapter 5.

3.2.1 Multiple Exposures

When an LDR image of the scene with a high dynamic range is captured, some of the regions will end up underexposed (unsaturated) while the others will be overexposed (oversaturated). This is shown in Figure 3.2c. Those problematic regions can be captured by changing the exposure of the camera, but then details in other regions will be lost. When the same scene is captured a number of times using different exposures, each of the regions will be visible in one or more images and all the data present in the scene will be obtained.

The multiple exposures techniques (Mann & Picard, 1995) take an input consisting of a set of LDR images captured at a range of exposure times (ET) and combines them into a single HDR image. The LDR images cover the full dynamic range present in the scene by capturing all areas from the darkest to the brightest (Figure 3.2). Assuming the camera has a linear response - that is, the

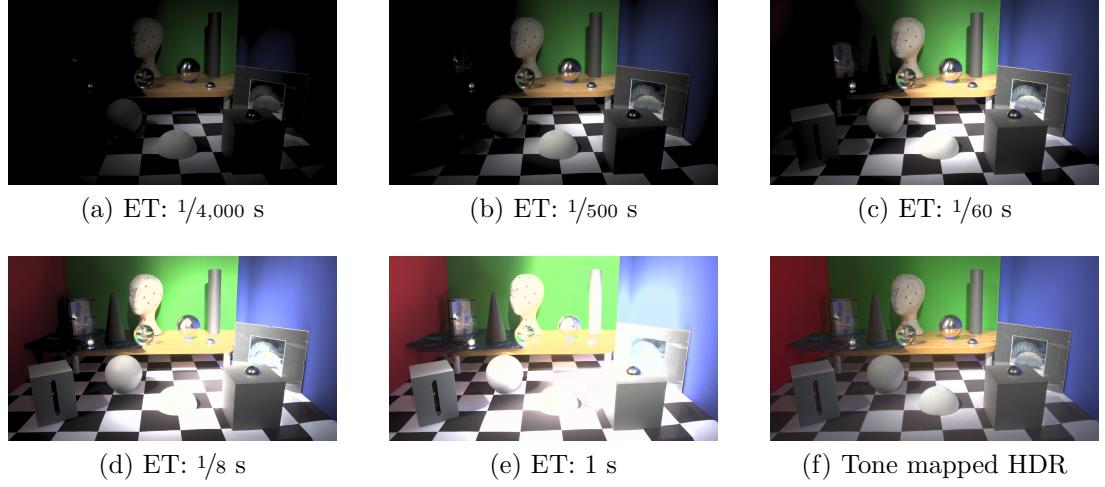


Figure 3.2: The multiple exposure technique merges LDR images to generate the HDR one. Images (a-e) are images captured using different exposure times (ET) each of which has some of the regions overexposed, underexposed or both. Image (f) is a tone mapped HDR image which resulted from the merge process.

amount of light hitting the sensor linearly corresponds to the output pixel value - each of the images can be scaled to the common domain by dividing its values by the exposure time. To obtain HDR values, the under- and overexposed pixels are excluded and the rest are averaged across the images (Figure 3.2f). This can be expressed as:

$$E(x, y) = \frac{\sum_{i=1}^N \frac{1}{\Delta t_i} \omega(I_i(x, y)) I_i(x, y)}{\sum_{i=1}^N \omega(I_i(x, y))} \quad (3.2)$$

where E is HDR irradiance, N is the number of different exposure images, Δt_i is the exposure time for the image I_i , I_i is the i -th exposure LDR image and $\omega(I_i(x, y))$ is a weight function used for removing outliers. The function $\omega(\cdot)$ favours middle range values more as they are less prone to noise.

In reality, cameras never have a linear response. Instead manufacturers boost image contrast using customised response functions to produce more vibrant images and also they modify the ends of a response function to reduce noise in dark regions and to smoothen the highlights. Camera makers rarely provide this function as they consider it proprietary. This curve can be inferred as long as the response stays the same across different exposures. Once the camera response function (CRF) is known it can be inverted and inserted into the Equation 3.2

to recover correct exposures:

$$E(x, y) = \frac{\sum_{i=1}^N \frac{1}{\Delta t_i} f^{-1} \omega(I_i(x, y)) I_i(x, y)}{\sum_{i=1}^N \omega(I_i(x, y))} \quad (3.3)$$

where f^{-1} is the inverted camera response function.

Mann & Picard (1995) suggested using a parametric function of the form: $f(x) = ax^\gamma + b$ but this solution is rather limited as it does not approximate most of the CRFs well. Debevec & Malik (1997) used linear optimisation to minimise the mean squared error produced by fitting a smooth curve to the pixel samples. This resulted in a lookup table for all 256 values and for each of the three RGB channels. Mitsunaga & Nayar (1999) calculated a polynomial estimation of the CRF. The added advantage of their method is the ability to find the correct exposure ratios, meaning that images lacking metadata stating exposures can be used to recover HDR content.

It might seem that using all of the pixels for CRF calculation would provide reliable and robust results, but not only would such approach require long computation times, it would also include noisy and misaligned pixels. Instead, sampling image patches is recommended (Reinhard *et al.*, 2010). Grossberg & Nayar (2002) suggested an alternative approach based on histogram evaluation which overcomes the problem of pixel alignment. They matched histogram percentiles of different exposures to infer the CRF.

The multiple exposure technique described above assume images are perfectly aligned, the scene is non changing and noise is absent. However, this rarely happens when capturing images in reality. Multiple techniques have been proposed which address all of these problems, mostly inspired by computer vision.

While many camera manufacturers now enable quick automatic bracketing (capturing sequence of images with different exposure), these images might end up misaligned due to camera movement in between the exposures, especially if a tripod is not used. Traditional approaches for image alignment find images with different exposures are challenging to align and are prone to errors. For example edge detection filters cannot maintain robust edges across exposures. Ward (2003) suggested a method which relies on a *median threshold bitmap* to translate images so they match. The algorithm generates bitmaps for each exposure by finding the median of the histogram, and setting all the pixels greater than the median to one and others to zero. Such bitmaps are robust to exposure change

and are matched using shift (translates image) and difference (counts errors) operations. Similar techniques based on a median threshold bitmap were proposed by Grosch (2006) and Jacobs *et al.* (2008).

Capturing multiple exposures somewhat limits control of the camera by imposing the usage of specific exposure times. This may result in noisy LDR images, especially for short exposures, which might persist in the merged HDR image. Moreover, Equation 3.3 shows that pixel values in short exposures influence the final result more, as they are divided by exposure time Δt_i . The traditional way of reducing the noise is *frame averaging* (Aggarwal & Ahuja, 2004) where a number of noisy, but otherwise equivalent images of the same scene are averaged. The multiple exposure techniques are a special case of this approach, as images differ in exposure. Frame averaging can be incorporated into the composition (Equation 3.3) to reduce noise in the final HDR image (Akyuz *et al.*, 2007). Noisy pixels are reduced by blending them with scaled less noisy ones from different exposures.

When capturing multiple exposures, scenes will often not be static; people and objects might be moving thereby altering the scene between each exposure frame. When a sequence is combined, this results in an artefact termed *ghosting* where the moving objects will appear blurred and semi-transparent. Khan *et al.* (2006) proposed a ghost removal scheme which assigns pixel weights during the merge process based on the probability of them belonging to the static part of an image. This approach works well when many exposures are available. A simpler technique suggested by Grosch (2006) exploits the fact that each image in the sequence is consistent, so different regions in an HDR image can be assigned data from a single LDR image to avoid ghosting. Such an approach might fail around edges of the regions so Gallo *et al.* (2009) perform the computation in the gradient domain to alleviate the problem. Sen *et al.* (2012) proposed a more advanced algorithm for combining LDR images and avoiding ghosting based on a patch-based energy-minimization function that merged both alignment and reconstruction in a single optimization. The method was able to combine exposures captured during large camera or scene motion, and generated less artefacts compared to other techniques.

The multiple exposure technique is difficult to apply to videos because they are recorded at approximately 25 frames per second (fps), not leaving much time to capture different exposures for each frame. The technique can be easily used to record time-lapse and stop-motion videos but for general applications different

solutions are needed.

Kang *et al.* (2003) proposed a method for capturing HDR videos using the multiple exposure technique. They recorded a video at 15 fps using a programmable camera, which allowed them to control shutter speed so that low and high exposure frames were captured interchangeably. This allowed them to combine the frames and generate HDR video. The main challenge was to align consecutive frames so that they could be merged, as both camera and object motions were present. In their elaborate solution, shown in Figure 3.3, they generated HDR frames by using data from the previous and the next frame.

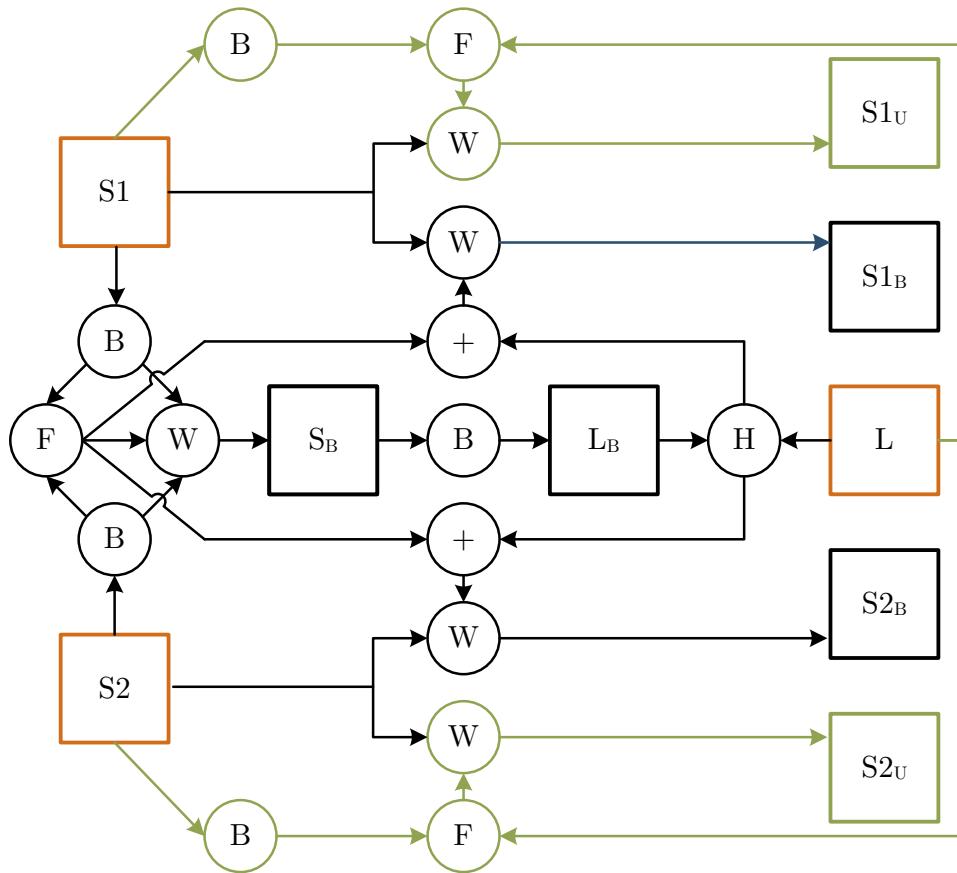


Figure 3.3: Three consecutive frames S1, L and S2 are used to generate HDR version of L. S1 and S2 are short while L is long exposed. Four additional images are created using luminance boosting (B), warping (W), homography (H) and optical flow (F). Two images are generated using unidirectional warping while two used bidirectional.

This can be best explained by examining the example they showed, where the current frame has a long exposure (L). The two dark surrounding frames (S1 and S2) were used to generate four novel images:

1. $S1_U$ - $S1$ unidirectionally warped to match L
2. $S2_U$ - $S2$ unidirectionally warped to match L
3. $S1_B$ - intermediate view between $S1$ and $S2$ bidirectionally warped to match L using $S1$ as the reference
4. $S2_B$ - intermediate view between $S1$ and $S2$ bidirectionally warped to match L using $S2$ as the reference

The brightness of short exposure images is boosted to match that of long exposure frames to facilitate warping during different stages in the process. Warping aligns pixels from consecutive frames so they can be merged. Optical flow - a motion estimation algorithm (Bergen *et al.*, 1992) - and homography are used to generate dense correspondence maps used for warping. Finally when these four images are aligned with the current long exposed L frame they are combined using the extended Mitsunaga & Nayar (1999) technique. They relaxed the requirement of perfectly aligned LDR images to tolerate errors in pixel registration which is explained in more detail in their work (Kang *et al.*, 2003). They have also applied their technique to static series alignment.

This technique does have several shortcomings. A frame rate of 15 fps is inadequate for smooth videos but also produces artefacts for faster moving objects, large geometrical differences (e.g. if applied to stereo), and large occlusions. Cameras with faster frame rates would help alleviate these problems, but would not completely solve them. An additional challenge is non-rigid objects which change between frames. Finally, the captured dynamic range is still limited by combining only two LDR images, which require significant overlap to reduce noise in the mid-tones. This could be overcome by using consecutive frames to capture more than two exposures but then the frame rate dependant problems would become more apparent.

3.2.2 Native HDR Capture

Consumer products are moving toward HDR imaging. A number of camera manufacturers, such as Canon, Nikon, Sony, Panasonic and Sigma, now support automatic multiple exposure bracketing and merging. Their high end products use up to 16 bits per colour channel, mostly for improving precision, but also for extending dynamic range. The situation is similar with the top of the range

video solutions - from companies such as Red Digital Cinema, ARRI, Vision Research and Panavision - which capture high quality, high speed, full high definition video, while claiming up to 14 stops of dynamic range (calculated using test charts which are not necessarily representative of the real world conditions). All of the high end solutions are expensive and aimed at professional film studios. The products which capture full dynamic range and which completely automate multiple exposure techniques for capturing spherical images where provided by companies including Spheron (26 stops) and Weiss (30 stops). While these cameras capture environment maps quickly and with high quality their application is specialised.

Researchers have proposed a number of solutions which enable true HDR video capture. One approach is to design novel camera sensors or modify existing ones to enable recording of HDR data. Nayar & Mitsunaga (2000) suggested placing an optical mask over the sensor which varied transmittance of light spatially using a predefined pattern. For example, a 2×2 pixel pattern could be repeated across the whole image where each pixel in the box had different exposure. The image was then demosaiced and combined into the HDR image. Drawbacks of such a method are a decreased resolution, wasting light entering the camera and possible misalignment. Nayar & Branzoi (2003) also masked the sensor in order to control the amount of light falling onto each pixel. However, they used a dynamic light modulator which would adapt in real-time based on the brightness of imaged scene point. The problem with any novel sensor proposal is cost of its development and the amount of time required until it becomes available to consumers.

Other approaches use optics to split the light and direct it to multiple sensors. Aggarwal & Ahuja (2004) and Wang *et al.* (2005) split the light with a three-sided pyramid shaped mirror and transmit it to three standard sensors. The amount of light for each of the images is controlled either by changing sensors' exposures or by splitting the light unevenly at the mirrors. The number of mirrors and sensors can potentially be increased. In this approach, each sensor captures the scene from a slightly different angle introducing parallax error and affecting the reconstructed HDR scene. The two methods are also wasteful of light (losing approximately 66%) as the sensor masks allow only a fraction of light through (approximately $1/3$).

A prototype of a true HDR video camera has been designed by the University of Warwick and Spheron (Chalmers *et al.*, 2009). The camera is capable of

capturing full high definition HDR videos (20 stops) at 30 frames per second. The raw data is streamed using five optical cables to a 24 terabyte HDD array and is then converted to more accessible OpenEXR format, accounting for spherical distortion, chromatic aberration, noise and lens vignetting. This camera was used to capture some of the videos used in Chapter 6.

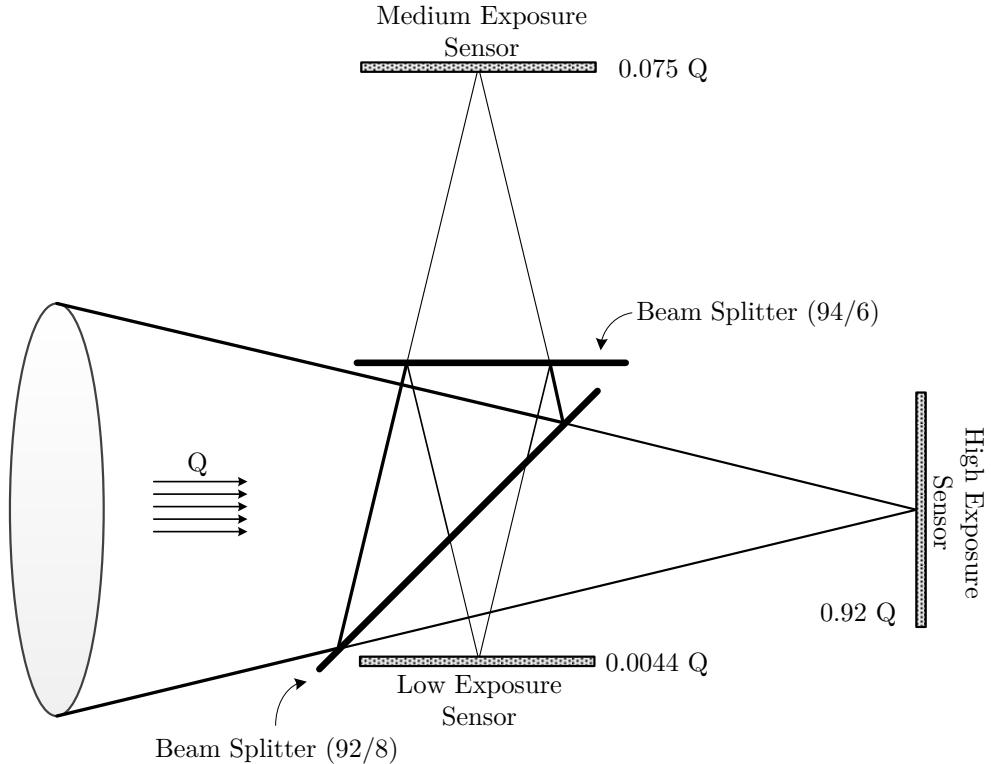


Figure 3.4: The HDR camera proposed by Tocci *et al.* (2011) consists of one lens, two partially-reflecting beam-splitters, and three standard sensors. The beam splitters have the different transmittance/reflectance (T/R) values, so each of the sensors gets to capture a different portion of the dynamic range (high, medium and low).

Tocci *et al.* (2011) proposed an efficient optical architecture which used a single lens, two partially-reflecting beam-splitters, and three standard sensors to capture HDR video, as shown in Figure 3.4. Such an arrangement of elements enabled well aligned capture of images without wasting light (losing about 0.04%). In addition Tocci *et al.* (2011) proposed a new HDR-merging algorithm which preferred data recorded by the brightest and the best exposed sensor. This allowed them to generate better images compared to the method of Debevec & Malik (1997) when separation between exposures was large (as in their case). The camera was able to capture full high definition (1920×1080) video at 30 fps with a dynamic range of 17 stops.

3.2.3 Expansion Operators (EOs)

Expansion operators (EOs), frequently called inverse tone mappers or reverse tone mappers, generate HDR content by converting LDR images and videos to HDR. LDR content lacks the data which gets lost during capture or storage due to the camera hardware limitations or lossy compression. Expansion operators use a priori knowledge of image formation and scene configuration to pose assumptions and constraints which help to recover lost data. This approach allows quick and cheap generation of HDR content compared to the other two methods (multiple exposures and native capture) as it is a software solution. In addition they allow conversion of abundant legacy LDR content which could help transition from LDR to HDR. However, generated results are estimated and do not necessarily correspond to the original scene.

This is a relatively novel field with only a handful of operators currently available. Banterle *et al.* (2009) provided psychophysical comparison of the five techniques that existed at the time. Two experiments were performed. The first tested which of the methods generated results most similar to the ground truth and the second evaluated the quality of image based lighting using the expanded images. They used the method of paired comparisons (David (1988)) to obtain the results - the same technique used in Chapter 5. In both cases, the operator (discussed below) proposed by Banterle *et al.* (2006) performed the best.

In general, the expansion of LDR images requires implementation of all or some of the four following processes: image linearisation, value expansion, over-exposed and underexposed region reconstruction, and artefact reduction. Below, each of the steps is discussed in more detail.

Linearisation is the process of relating real-world and pixel values linearly. The cameras apply a CRF, as discussed in Section 3.2.1, which needs to be removed to transfer the values into the linear space thereby providing more control and predictability of the expansion. However, in contrast to the multiple exposure case, the CRF needs to be recovered from the single image and a single sample from each pixel.

Farid (2001) assumed that images are stored with gamma correction applied and suggested a method to blindly estimate the inverse of that gamma function. He recognised that after applying gamma to the image, the frequency domain would have novel higher-order correlations which, once detected, could be minimised in order to estimate the inverse gamma function. The reconstructed

gamma differed from the ground truth with the errors between 5.3% and 7.5%. Lin & Yamazaki (2004) observed that the edge pixels were linear interpolation of the two colours: one from each side of the edge. Applying the CRF would transform this relation to the non-linear one - a property which was used to recover the transformation function and to linearise the image. The estimated function was obtained by minimising the inverse interpolation distance between three pixels: the edge and two neighbours. Lin (2005) extended the technique using histograms and applied it to grayscale images.

Value expansion is the main step of any EO and it converts an LDR image to HDR. Based on the way this is performed, EO can broadly be divided into the two categories: global and local. Global EOs (Landis, 2002; Akyuz *et al.*, 2007; Masia *et al.*, 2009) apply a single expansion function over all of the image pixels. The function selection and the method used for its calculation is the main difference between operators. Once the function is obtained it is applied to the whole image making these techniques efficient and easily implementable in hardware. Global operators introduce more image artefacts, such as contours and halos, and produce less precise colours and luminance values (Banterle *et al.*, 2011) compared to the local ones, making them less accurate. Local techniques (Banterle *et al.*, 2006; Meylan & Süssstrunk, 2006; Meylan *et al.*, 2007; Rempel *et al.*, 2007; Didyk *et al.*, 2008) expand images based on region level information. This allows for more control and better reconstruction as assumptions about the local content may be made. The cost of such improvement is the computation time, as local analysis and expansion requires more processing. Two representative approaches, one from each class are discussed further, later in the section.

If the amount of light reaching the camera is small it will not be registered by the camera sensor and corresponding pixels will be completely black; if it is too large the sensor will become oversaturated and the pixel will be white. For colour images this happens on a per channel level. This means that once the image is captured, data in those regions is lost. Local operators identify and then try to **reconstruct overexposed and underexposed regions**. Wang *et al.* (2007) required the user to mark the problematic regions using brush strokes and then to mark the regions with the content similar to the missing one. The algorithm then separated the image into high and low frequency layers. The dynamic range of the low frequency layer was boosted, while the high frequency layer was used to transfer textures. Fully automated approaches have also been proposed. Banterle *et al.* (2006) (discussed below) and Rempel *et al.* (2007) boosted overexposed

regions using a weight image that decided the extent to which pixel would be boosted. The image was termed *expand map* and was calculated by finding and smoothing the highlights in the original LDR image. Another automated method proposed by Meylan & Süstrunk (2006) classified an image into highlights and diffuse regions and expanded each using different functions.

Artefact reduction techniques try to identify and remove unwanted artefacts which may be introduced during the expansion operation. For example, by increasing the dynamic range, errors caused by the quantisation, which were not apparent previously, may become visible in the form of false contours. This effect is termed contouring, banding, haloing or posterization and is illustrated in Figure 3.5. It is mostly prominent in the gradient regions (e.g. clear sky, dawn, sunset), where a sudden intensity jump becomes noticeable. Daly and Feng tackled the problem in two ways, by trying to hide or remove contouring in medium dynamic range images. The first technique (Daly & Feng, 2003) relied on amplitude dithering (Figure 3.5c). Dithering is the technique that adds a noise pattern before quantizing the image and removes it after, thereby hiding structured artefacts such as contours. Daly & Feng (2003) extended the technique to work with HDR images. They suggested removing contours instead of masking. To detect them, the image was low-pass filtered and stored with increased bit-depth. Subsequently, it was quantized and any novel contours were deemed false and were subtracted from the expanded image.

The rest of this section describes the two expansion operators that were used in Chapter 5 in more detail. Linear scaling is a global, fast and straightforward operator, but it is artefact prone and the results are likely to deviate from the original scene significantly. The local expand map operator takes more time to compute the results but it is expected to produce better results as it analyses image regions and expands them accordingly.

Linear Scaling (LS)

Linear scaling (LS) is a global EO which scales all the luminance values of an LDR image using a linear function. The operator consists of the two steps: linearisation and value expansion. Any of the linearisation methods explained above can be used while the expansion is performed as in Equation (3.4).

LS was formally introduced as the part of the work by Akyuz *et al.* (2007) which explored three questions relating HDR and LDR content. They examined

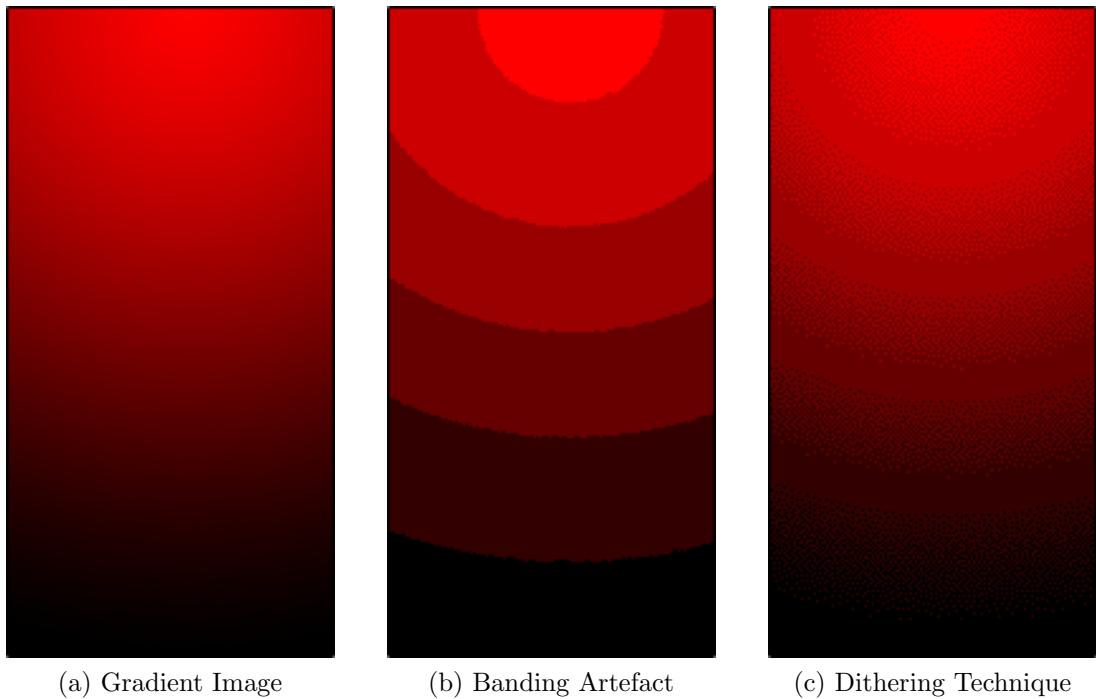


Figure 3.5: Banding artefacts may become prominent once the image is expanded. When the dynamic range is low, as in the case of an LDR image, quantised errors are hidden. Expansion stretches quantised values so difference between them become larger and banding occurs. This is illustrated in images (a) and (b). Smooth gradient, as in (a), after expansion would have visible contours, as in (b). Traditional technique to minimise the effect is to use dithering - adding noise which masks structured contours, as shown in (c).

user preference between HDR and LDR imaging pipelines, the effect of average luminance and contrast on visual appeal, and how does a method of inverse tone mapping compare to the display of original HDR content. To answer the last question the authors suggested LS - a straightforward way of expanding the dynamic range of an LDR image. The luminance values of pixels were normalised and then scaled so that the maximum luminance corresponds to the desired value (usually, the maximum screen brightness). Formally, value expansion can be written as in Equation (3.4):

$$L_o(\mathbf{x}) = k \left(\frac{L_i(\mathbf{x}) - L_{\min}}{L_{\max} - L_{\min}} \right)^{\gamma_A} \quad (3.4)$$

where k is the maximum luminance to be achieved, \mathbf{x} is the coordinates of the processed pixel, L_i is the luminance value of the input, L_{\min} and L_{\max} are minimum

and maximum luminances of the input image respectively, γ_A is the nonlinear scaling factor, and L_o is the output luminance value. Only when $\gamma_A = 1$ is the image expanded linearly.

The study compared the original HDR image with the expanded LDR image. The LDR image was selected in a pilot study where participants selected the best single exposure out of those used to create the HDR image (using the multiple exposure technique). The LDR image was expanded using Equation (3.4) and three γ_A values were tested: 0.45, 1 and 2.2. The experiment showed that linear scaling was the most favoured. Moreover, the linearly expanded image was overall preferred to the original HDR image. The authors also evaluated the effect on four specific visual attributes: naturalness, visual appeal, spaciousness, and visibility. Again, results showed that for all of the attributes LS was in the better or the same preference group as the HDR image. Other advantages of this operator include straightforward execution and speed making it suitable for hardware implementation.

However, an image expanded using LS in general does not represent the captured scene as well as the HDR image. While the participants find the expanded image appealing, its values do not necessarily correspond to the original HDR values. Overexposed and underexposed regions are not handled at all which may become a problem especially if they are substantial in size. Quantisation is present as no new values are introduced but instead existing ones (256 per colour channel) are scaled. In addition, the images used in the experiment were of high quality and were not compressed, so the results were not affected by the compression artefacts which would likely become apparent when using this technique.

Expand Maps (EM)

Banterle *et al.* (2006) suggested an EO which attempts to overcome some of the problems which global methods face (e.g. banding, compression artefacts, flat highlights). It performs all of the EO processes (discussed above). The key step is generation of the expand map - the weight image used for local, per-pixel luminance boost.

The algorithms consists of five consecutive steps, also shown in Figure 3.6:

1. LDR image linearisation;
2. Inverse tone mapping of the linearised image;

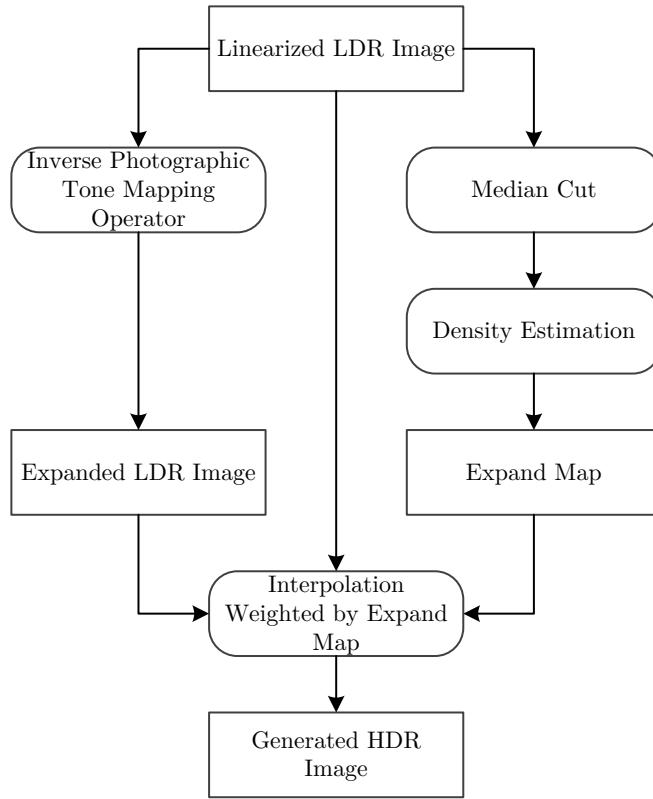


Figure 3.6: The diagram shows the overview of the EM algorithm. The linearised LDR image is inverse tone mapped and an expand map is generated using median cut and density estimation. The final image is generated by linearly interpolating the linearised with the expanded LDR image using the expand map as a weighting function. Image (a) courtesy of Paul Debevec.

3. Finding light sources;
4. Generation of an expand map using the found light sources; and,
5. Linearly interpolating the LDR image and inverse tone mapped HDR image using the expand map.

Similar to linear scaling expand maps are not dependant on the linearisation technique. Ideally, a known CRF should be used, but methods that recover the CRF may be used as well.

The second step uses the inverse of the tone mapping operator (TMO) to expand the image. The idea behind this step is that TMOs scale an HDR image to LDR (see Section 3.4.2), and so their inverse does the opposite - it converts an LDR image to the HDR one. This is the case even for LDR images which were not tone mapped.

In theory any invertible TMO may be used, but some are more complicated to invert than the others. Especially challenging are the local ones as they scale the image depending on the regional content. In their implementation, Banterle *et al.* (2006) inverted the global *Photographic TMO* (Reinhard *et al.*, 2002). They justified the operator choice by easy inversion, frequent use and good performance shown in TMOs' comparison studies (Ledda *et al.*, 2005; Smith *et al.*, 2006). In addition, the expansion function is smooth and avoids the noise caused by naive algorithms. The inverse of the Photographic TMO can be expressed as:

$$L_w(\mathbf{x}) = \frac{1}{2} L_{\max} L_{\text{white}} \left(L_d(\mathbf{x}) - 1 + \sqrt{(1 - L_d(\mathbf{x}))^2 + \frac{4}{L_{\text{white}}^2} L_i(\mathbf{x})} \right) \quad (3.5)$$

where L_{\max} is the maximum luminance to be achieved in cd/m^2 , L_{white} controls the shape of the expansion curve and is proportional to contrast, and L_w and L_d are HDR and LDR luminances respectively. The operator required two input parameters (L_{white} and L_{\max}) but the authors suggested that $L_{\text{white}} \approx L_{\max}$ so only a single value was needed.

The image resulting from this step has increased dynamic range but there is a limit as to how far it can be increased. If expanded excessively, blocky artefacts may be expected. To overcome this, the algorithm boosts high luminance regions additionally.

The bright regions are first identified using the median cut algorithm (Debevec, 2006) which attempts to find light sources in the image and clusters them near high luminance areas. Then, density estimation (Duda *et al.*, 2001) associates the amount of brightness boost, for each pixel, based on its proximity to and intensity of neighbouring light sources - thereby generating the expand map. Finally, the original LDR image and the image expanded using the inverse TMO are combined using the expand map as an interpolation weight.

The results showed that the EM algorithm performed better than the naive expansion of the dynamic range. Also HDR images generated in such a manner were successfully used for image based lighting applications. In a different user study Banterle *et al.* (2009) evaluated existing expansion operators where results of EM algorithm were deemed closest to the ground truth (a comparison with the original HDR image).

While the EM technique performs well there are number of shortcomings. Generated images have an expanded dynamic range but, same as with linear

scaling, the faithfulness to the original scene can not be guaranteed and expanded images usually do not correspond well to the actual HDR. To achieve adequate results, two parameters need to be properly adjusted using trial and error which may take a significant amount of time, especially given that the algorithm is not as fast as the straightforward methods. The authors reported a failure of the algorithm when reconstructing large overexposed areas (approximately 30% of the image) as they become smooth and gray. Underexposed areas are not handled explicitly by the algorithm so no or slight improvement can be expected there. When used for video this technique results in flickering, hence it was extended (Banterle *et al.*, 2011) to handle the problem using three-dimensional sampling and volume density estimation at the expense of performance.

3.3 High Dynamic Range Content Storage

Similar to stereoscopy, high dynamic range imaging increases the amount of data needed to represent captured content compared to traditional LDR imaging techniques. Compared to stereoscopy, which doubles the amount of data, HDR increases storage requirements even further as it quadruples data. Raw HDR image data are considered to be composed of three floating point values, one for each of the red, green and blue channels, for a total of 96 bpp, compared to 24 bpp for LDR. To illustrate the impact of such a size increase, the following facts may be considered - a single raw high definition HDR video frame would take 24 MB while 8 seconds of such video running at 25 fps would take 4.8 GB and could not fit on a standard single sided DVD.

Coding techniques which reduce the size of HDR content are required for integrating HDR with current ICT infrastructure and thereby facilitating its adoption. This challenging task has been tackled by researchers and standardised formats are beginning to emerge. Current HDR compression techniques mostly rely on existing LDR coding methods, which are modified and extended to enable HDR storage. This section first looks at the existing lossless HDR file formats and proceeds to examine a number of HDR data coding methods.

3.3.1 File Formats

A number of HDR image formats are available and they have been embraced by the computer graphics community. For instance, image editing packages such as

Adobe Photoshop and GNU Image Manipulation Program (GIMP) are capable of reading and writing most of the HDR formats and have extended some of their tools to enable HDR processing. Most of the file formats are extended versions of existing LDR formats which use more bpp to represent additional data. Not all of these use 96 bpp, so some loss of precision may be expected compared to raw images. However as long as the relative quantisation error is below 1% it is invisible to the HVS (Wyszecki & Stiles, 2000). All file formats apply only lossless compression methods internally.

HDR videos have no established file format. While some video editing softwares such as Adobe After Effects and The Foundry Nuke enable HDR pipelines they represent videos as image sequences of the standard HDR image formats. Efforts have been made to extend existing video standards such as MPEG to support HDR (as discussed further in the section) but this has yet to become a standard.

Three established HDR image formats exist. These are: Radiance, TIFF and EXR. In a comparison which tested how suitable these are for archiving HDR images, Reinhard *et al.* (2010) concluded that all three perform well and may be used for the task. The rest of the subsection discusses the formats in more detail.

Radiance (HDR)

Ward (1994b) presented the *Radiance Lighting Simulation and Rendering System* where he saved rendered image data using a custom image format which enabled storage of HDR. This was the first digital HDR image format and it became widespread in the graphics community. Radiance image files initially used the *.pic* and later *.hdr* file extension.

The file wrapper is divided in three parts: an ASCII header which defines the file type, the resolution string which states the size and orientation of the image, and the pixel data compressed using run-length encoding. Each pixel is saved using 32 bits (32 bpp) which are equally divided amongst the three colour channels and the exponent (8 bits each, as shown in Figure 3.7). Two colour spaces may be used to represent colour: RGB and XYZ. The novel feature of this representation is the usage of 8 bits for the exponent which enables significant colour intensity scaling. The actual scene colour values (E , R_W , G_W , B_W) are



Figure 3.7: Radiance file format is 32 bits long and split so that each of the RGB or XYZ channels gets 8 bits while the final 8 are used for the exponent which scales the dynamic range.

converted to the radiance format (*RGBE* or *XYZE*) as follows:

$$\begin{aligned} E &= \lceil \log_2(\max(R_w, G_w, B_w)) + 128 \rceil \\ C &= \left\lfloor \frac{256 C_w}{2^{E-128}} \right\rfloor \end{aligned} \quad (3.6)$$

where C is used for brevity to represent one of the three colour channels and is calculated for each. Both colour spaces (*RGBE* or *XYZE*) are converted using the same equation. To invert Radiance back to scene values the following Equation is applied:

$$C_w = \frac{C + 0.5}{256} 2^{E-128} \quad (3.7)$$

Again C is substituted with the appropriate colour channel.

The exponent component allows a major boost of the dynamic range which can span a range of more than $10^{76} : 1$. Some precision is lost compared to the raw floating point number representation because only 8 bits get allocated per colour. The relative quantisation error is 1% and is unnoticeable. Lossless run-length encoding is applied to the data and reduces the file size by 25% on average, making it approximately as large as a raw 24 bpp image.

Tagged Image File Format (TIFF)

The Tagged Image File Format (TIFF) was first proposed by Aldus Corporation in 1986 making it one of the oldest digital file formats. It has since been acquired by Adobe Systems who holds copyrights to the specification which was last updated in 1992 (Adobe Developers Association, 1992). TIFF is a flexible file wrapper which uses descriptive tags to identify image elements and supports different colour spaces, bit depths and compression techniques. As such it can support storage of HDR data in four major ways.

Firstly, it allows usage of 32 bit float point numbers to encode RGB channels and hence can represent raw HDR images enabling $10^{79} : 1$ range with excep-

tional precision. As such it is well-suited for an intermediate representation when processing HDR content. It allows quick reading from and writing to a frame buffer avoiding loss of data. For research purposes it can be used as a ground truth for different HDR image evaluations. While TIFF allows different methods of data encoding, in general, floating point numbers do not compress well and only about 10% size reduction may be expected. The main drawback of using TIFF in this manner is the file size.

Ward (1998) proposed a method - termed LogLuv - to tackle the size challenge and suggested two variants using TIFF, each using different bit depths (24 and 32 bpp). As luminance determines the dynamic range of an image Ward (1998) used *CIE 1976 (Luv) colour space* which separated luminance (L) from chroma components (u and v). The human visual system responds to brightness in a logarithmic manner (as discussed in Section 3.1) and so in both variants logarithm is applied to the luminance channel.

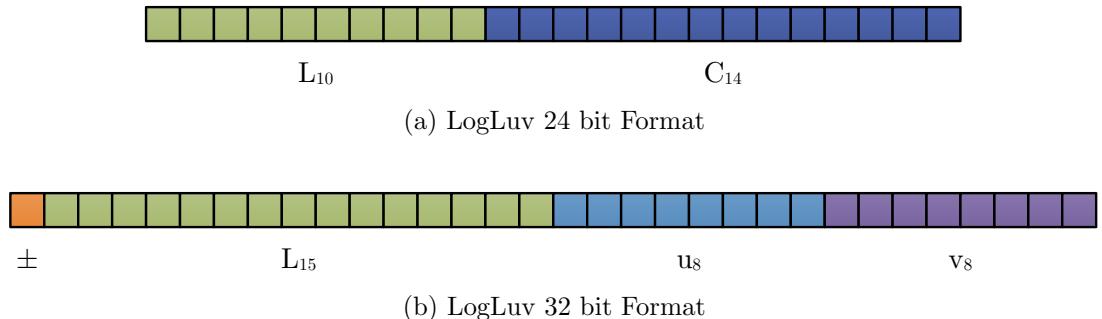


Figure 3.8: LogLuv file format bit partitioning.

The 24 bit version allocates 10 bits for the log-luminance channel (see Figure 3.8a) which is converted as follows:

$$\begin{aligned} L_{10} &= \lfloor 64(\log_2 Y_W + 12) \rfloor \\ Y_W &= 2^{(L_{10}+0.5)/64-12} \end{aligned} \tag{3.8}$$

where L_{10} is the quantised luminance value and Y_W is its real world counterpart. The achieved range is $10^{4.8} : 1$ with a relative error of 1.1%. In the cases that exceed the upper range bound, a scaling factor can be applied and stored in TIFF's auxiliary tag and used in the decoding. Chroma is stored using the remaining 14 bits, which refer to the values in a global (used by all images) look-up table. The table was originally calculated by uniformly sampling the visible colour

gamut expressed by u and v values. As Luv colour space is perceptually uniform, the constant step size (of 0.0035) used for sampling ensures equal perceptual difference between colour samples.

While this method manages to represent an HDR image using 24bpp - the same bit-depth as an LDR image - both luminance and the (u, v) step sizes are above visible thresholds and may result in visible artefacts. To alleviate the problem, dithering is used.

The 32 bit LogLuv technique tries to achieve a balance between size and quality. This time 15 bits are allocated for the luminance channel (see Figure 3.8b) which is quantised and recovered as follows:

$$\begin{aligned} L_{15} &= \lfloor 256(\log_2 Y_W + 64) \rfloor \\ Y_W &= 2^{(L_{15}+0.5)/256-64} \end{aligned} \quad (3.9)$$

The additional bits significantly increase the dynamic range to $10^{38} : 1$ and reduce relative error to 0.3% - well below the visible threshold. In this version the look-up table is avoided and actual u and v coordinates are stored using 8 bits for each providing sufficiently small colour step size. The chromaticity conversion is performed as follows:

$$\begin{aligned} u_8 &= \lfloor 410 u' \rfloor & u' &= \frac{u_8 + 0.5}{410} \\ v_8 &= \lfloor 410 v' \rfloor & v' &= \frac{v_8 + 0.5}{410} \end{aligned} \quad (3.10)$$

The extra bit is used as the sign allowing for negative values of luminance. This feature may be useful for compositing and visualisation of differences between images.

In the 32 bit variant of LogLuv dithering may be applied but it is not necessary or recommended as it might hinder compression performance. The size reduction achieved by the TIFF library is between 10 and 70% with an average of 40%. While this format seems to achieve a good balance between size and quality in practice, it is not always supported. As TIFF is so versatile many imaging applications do not implement all of its features, including LogLuv.

The fourth and the last incarnation of HDR images in TIFF involves representing the colour channels as 16 bit floating point numbers. This is a standard feature of the TIFF format which does not offer a significant dynamic range when

an image is encoded linearly. When a gamma curve is applied a larger range may be achieved, but the variable step size becomes an issue reaching relative errors of 5%.

Extended Range Format (EXR)

The OpenEXR is a format designed primarily for storing HDR data. It was developed by Industrial Light & Magic mainly for image editing and compositing purposes to avoid accumulation of precision errors in their production pipeline.

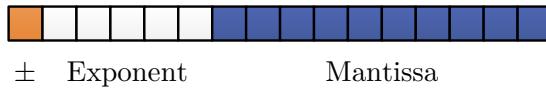


Figure 3.9: OpenEXR format bit allocation for one of the colour channels.

Similar to TIFF there are multiple variants of EXR available (Kainz & Bogart, 2009). These include 32 bit floating point numbers per channel, 24 bpc, or - the most frequent option - 16 bpc. The last is also simply termed *Half* referring to the half precision floating point number. The bit allocation for one of the channels is represented in Figure 3.9 where the first bit is used for the sign, the next 5 for the exponent and the 10 remaining bits for the mantissa. Formally, the value represented is calculated as follows:

$$v = \begin{cases} (-1)^S 2^{E-15} \left(1 + \frac{M}{1024}\right) & 1 \leq E \leq 30 \\ (-1)^S 2^{-14} \frac{M}{1024} & E = 30 \end{cases} \quad (3.11)$$

where v is the represented value, S is the sign bit, E is the exponent (0 - 31), and M is the mantissa (0 - 1023). When $E = 31$ then v is either not-a-number (NaN), or infinity - if $M = 0$.

The largest value a 16 bit EXR can represent is 65504 which allows for a $10^9 : 1$ dynamic range. This can be extended by using less accurate values below a threshold of 0.000061. For most of the range the relative quantization error is less than 0.1% which enables extensive image editing before the accumulated error becomes visible. NVidia's Cg language and CUDA support the Half data type allowing EXRs to easily be integrated on GPUs.

The OpenEXR library implements a number of lossless compression methods including ZIP deflate and their own PIZ technique. PIZ uses wavelet compression

and achieves reduction of approximately 60% on average (Reinhard *et al.*, 2010). The number of channels is not limited so additional ones may be included (e.g. alpha, shadow and motion vectors). Adding rich metadata is enabled; so besides the provided fields for defining colour space, pixel density, capture date and camera setting it is possible to include user-defined attributes.

3.3.2 High Dynamic Range Image Coding

The previous section described three popular HDR image formats. While some of them (e.g. OpenEXR) apply compression techniques they are in general, lossless and do not take into account the specific properties of HDR content. As the amount of HDR data is much larger compared to LDR, techniques which reduce the size significantly need to be developed if traditional media and distribution channels are to be used. In addition, a desirable feature of any coding algorithm is backwards compatibility, as this supports both LDR and HDR formats. Such a property could help the transition from LDR to HDR. This subsection examines some of the lossy compression efforts specifically designed for HDR images. Special focus is placed on the JPEG-HDR technique, as it is used in Chapter 7.

JPEG-HDR

Ward & Simmons (2004) and Ward (2005) introduced a backwards compatible lossy method for compressing HDR images. They argued that HDR technology should be developed in a similar vein to colour television when it first emerged (this introduced a chroma sub-band without interfering with the black-and-white signal). This means that a solution was required which would naturally support LDR technology but accommodate HDR data as well. To this end they chose to extend the JPEG image format because it was universally supported, encoding and decoding were fast, it had good compression performance and was implemented in freely available software libraries.

The method stored a tone mapped (TM) version of the HDR image as a JPEG and put restorative data in one of the metadata channels. This allowed traditional LDR image viewers to display the TM image, while HDR viewers looked for the metadata and used it to generate the HDR image and display it. Figure 3.10 provides a more detailed overview of the encoding process.

Using this method an HDR image was first tone mapped and the framework supported multiple TMOs. However, the authors suggested that a good operator

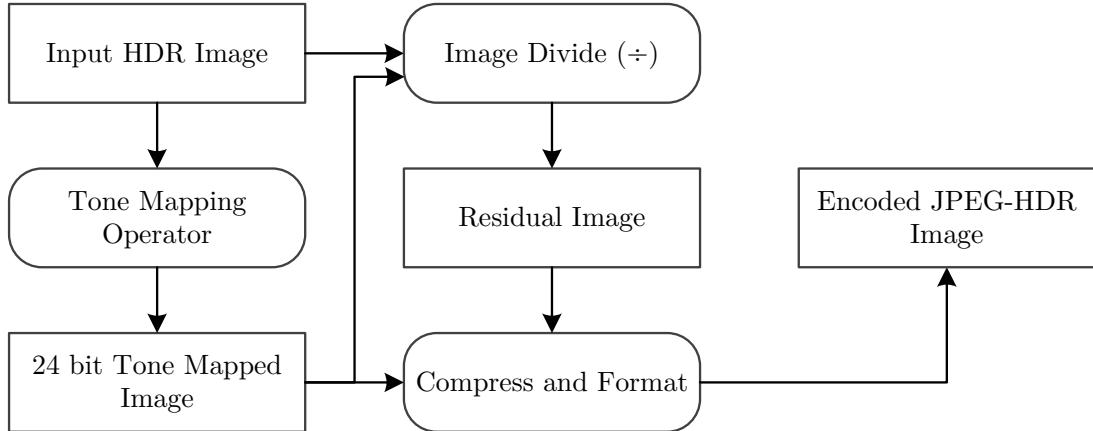


Figure 3.10: JPEG-HDR encoding process performs four major operations: tone mapping, image division, image compression and formatting.

should map the image smoothly into a 24 bit RGB domain without clamping, and it should maintain the hue at each pixel. The authors experimented with four global TM operators and found that the *Bilateral Filter* operator (Durand & Dorsey, 2002) performed the best, closely followed by the *Photographic* operator (Reinhard *et al.*, 2002). They set the latter as a default. The gamut of sRGB got vanishingly small near white values, even when using well performing operators. JPEG supports YC_bC_r colour space which has a larger gamut in the white region so it was used instead and alleviated the problem. In addition, a global desaturation was applied. It pulled all the colours towards gray so that the entire visible gamut was contained in the image and also enhanced the appearance of the TM image. Detailed accounts of colour space conversion and desaturation are provided in the original paper.

The two images (HDR and TM) were divided to produce the greyscale ratio image as follows:

$$RI(x, y) = \frac{\text{Lum}(HDR(x, y))}{\text{Lum}(TM(x, y))} \quad (3.12)$$

where $RI(x, y)$ is the resulting residual image pixel at (x, y) coordinates, $HDR(x, y)$ and $TM(x, y)$ are the pixels of HDR and TM images respectively, and Lum is a function that calculates luminance. The ratio image was logarithmically encoded and quantized.

The TM image was JPEG compressed using traditional 8×8 , 8 bit, discrete cosine transform (DCT) encoded blocks. To save space, the ratio image was downsampled based on the fact that the HVS was not very sensitive to sharp

changes in luminance. The authors suggested two methods to reduce error occurring due to downsampling. The first *precorrects* the image by inspecting how the restored image is going to appear and including error corrections in the TM image. In the second, during the decoding process, the TM image is utilised to guide the synthesis of high frequency details in the upsampled ratio image. Examples of the uncompressed TM and the ratio image are shown in Figure 3.11.

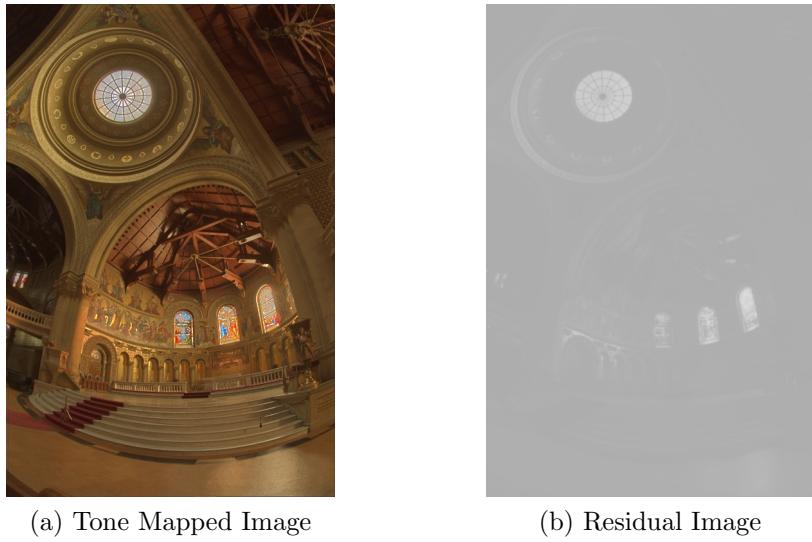


Figure 3.11: The TM image provides a preview of the HDR content when opened in an LDR viewer. The residual image is a one channel low frequency image which compresses well.

For storing the encoded image, the authors used JFIF - the file format for wrapping JPEG compressed images (mentioned in Section 2.5.3). Its definition contains 16 metadata storage channels termed *application markers* or *sub-bands*. Each marker may hold 64 Kb of data but the total limit may be overcome by reusing the same marker identifier. Ward (2005) used the image segment to store the compressed TM image. The application markers held the ratio image.

The decoding process is straightforward. The LDR viewer simply decompresses the TM in the usual manner ignoring the HDR metadata. The HDR viewer finds the ratio image in the JFIF sub-band and upsamples it. Optionally, it may perform error reduction after upsampling. Finally, the two images - the TM and the residual - are multiplied in order to produce an HDR image which is displayed.

The performance of JPEG-HDR was tested using more than 200 images of natural scenes. It was possible to change the JPEG compression quality between

0 to 100 but the quality degraded quickly below a value of 60 so those values were not considered. On average, the compressed images used 0.6 bits per pixel for a quality of 57 and 3.75 bpp for a quality of 99. In comparison the LogLuv TIFF format averaged 21.5 bpp for the same set meaning that the JPEG-HDR further compressed the image with a ratio of approximately 6 : 1 and 40 : 1 depending on the compression quality. The ratio image contributed $\frac{1}{4}$ of the total file size on average. However, its size varied substantially depending on the image. It ranged between 16% and 27% for a quality of 71 and between 24% and 37% for a quality of 85.

HDR-JPEG2000

JPEG 2000 is an image compression technique created in 2000 by the same group behind the JPEG standard. Their intention was for this format to succeed JPEG but the imaging community was reluctant to relinquish JPEG which was well-established. The compression performance of the new format was slightly better compared to JPEG, but it offered a more flexible codestream. For example, a truncated codestream would still allow displaying an images but at a lower resolution. The compression method was based on wavelets compared to JPEG's DCT. The feature - which Pattanaik & Hughes (2005) exploited for storing HDR images - is to use 16 bit unsigned integers for storing colour channels. In their solution Pattanaik & Hughes (2005) suggested a lossy HDR compression scheme which converted floating points into 16 bit integers.

The process initially transformed the raw HDR pixel colours R_w, G_w and B_w into the logarithm domain. Floating points were then quantised to unsigned short integers. This was done by normalising the value and multiplying it by the maximum integer which can be represented in JPEG 2000. Formally, quantisation was performed as shown below:

$$\begin{aligned} [\bar{R}, \bar{G}, \bar{B}] &= f([R', G', B'] : n) \\ f(x : n) &= (2^n - 1) \frac{x - x_{\min}}{x_{\max} - x_{\min}} \end{aligned} \tag{3.13}$$

where x_{\min} and x_{\max} were minimum and maximum values of the x colour channel, and n was the number of bits available for representing unsigned integers, which for JPEG 2000 was 16.

Once the image was discretised it was sent to the JPEG 2000 encoder which

proceeded to operate in the usual manner. The only modification was made to the wavelet domain sub-band quantisation where the perception-related factor was omitted because HDR images were scene referred as opposed to being display referred. The JPEG 2000 encoder allowed for both a lossless and a lossy mode of compression. In the lossless mode, the only errors were introduced by the float to integer conversion step, while lossy introduced additional errors due to wavelet coefficient quantisation and rate-distortion optimisation.

The decompressing process used the standard JPEG 2000 decoder. The original HDR values are obtained by inverting Equation 3.13.

The authors compared the performance of their method to the techniques available at the time. The lossless method was compared to Radiance, 32 bit LogLuv TIFF and OpenEXR while lossy was compared to JPEG-HDR and HDRV - the method by Mantiuk, Krawczyk, Myszkowski & Seidel (2004) described below. The testing used objective metrics: logarithmic root mean square error (RMSE) and VDM (Lubin, 1995). The lossless variant did not perform well and the file size was larger than the file sizes of all other competing methods. However, the lossy HDR-JPEG2000 was superior to both JPEG-HDR and HDRV, more so at the low bit rates for which competing methods produced artefacts.

One of the drawbacks of HDR-JPEG2000 was the time required for lookups which hindered real-time application. In addition, the method used the limited range of JPEG2000 compression capabilities and operated at a high level. For instance, more space could have been saved by separate processing of the luminance and chrominance channels.

Two-Layer HDR Coding

Okuda & Adami (2007) suggested a backward compatible method for compressing images. They classified it as a two-layer technique as it separated the image into the two parts: TM image displayable by an LDR viewer, and data used for reconstructing HDR. As such it closely resembled JPEG-HDR and HDRV (Mantiuk, Krawczyk, Myszkowski & Seidel, 2004) techniques. The key contribution of the method was approximating the tone mapping function and inverting it in order to expand the image. In addition residuals were compressed using wavelets.

The encoding process used the HDR image and its TM version as input. The TM image was to be inverse tone mapped for residuals to be calculated.

However, the authors assumed that the tone mapped image was given, instead of being computed which made the inverse tone mapping more challenging. TMs that could have been used to reduce images' dynamic range included global and local methods (see Section 3.4.2). Many local operators could be approximated using an s-shape function. This was exploited by the authors who used the Hill function to estimate any TM operator as follows:

$$L_d(\mathbf{x}) = f(L_w(\mathbf{x})) = \frac{L_w(\mathbf{x})^n}{L_w(\mathbf{x})^n + k^n} \quad (3.14)$$

where L_d and L_w are luminances of the HDR and the TM images respectively, and n and k are parameters that controlled the shape of the s-curve. The function was inverted resulting in:

$$L'_w(\mathbf{x}) = f^{-1}(L_d(\mathbf{x})) = k \left(\frac{L_d(\mathbf{x})}{1 - L_d(\mathbf{x})} \right)^{\frac{1}{n}} \quad (3.15)$$

Parameters n and k were estimated by minimising the mean square error (MSE) in logarithmic domain:

$$E = \frac{1}{M} \sum_{\mathbf{x} \in I} \left(\log(L_w(\mathbf{x})) - \log(L'_w(\mathbf{x})) \right)^2 \quad (3.16)$$

where M represents the number of pixels in the image I . Logarithmic scaling was used to match the brightness response of the HVS and to avoid outliers at high values.

The calculated parameters were used to expand the TM image which enabled the generation of the residual image using Equation 3.17:

$$R(\mathbf{x}) = \left(\frac{L_w(\mathbf{x})}{L'_w(\mathbf{x}) + \epsilon} \right)^\gamma \quad (3.17)$$

where γ was a constant less than one which weighted differences in high values more and prevented them from being undervalued; ϵ prevented the values from growing too big when $L'_w(\mathbf{x})$ approached zero.

The residual image was encoded using wavelets, the TM image was encoded using JPEG compression, and parameters n , k , γ and ϵ were preserved. The decoding process was quick as it had expansion parameters readily available. The TM image was expanded and multiplied with residuals restoring the original

HDR image.

The authors also suggested two colour compensation techniques used to reduce distortions caused by tone mapping. The first modified the method of Ward & Simmons (2004) by calculating parameters α and β using quadratic minimisation. The second technique applied a polynomial $P(x)$ to each TM colour channel based on the relationship between the TM and the HDR images. Polynomial coefficients were fitted using a Gaussian weighted difference.

Twelve HDR images were used to compare the method with JPEG-HDR and HDR-MPEG (Mantiuk, Efremov, Myszkowski & Seidel, 2006; described below). The two metrics used were the mean distance (MD) in CIELAB colour space and MSE with Daly's nonlinearity. The method performed better than the other two for both metrics. The quality was up to two times better for high bit rates (8-10 bits) and similar for low bit rates (1-4 bits).

3.3.3 High Dynamic Range Video Coding

Compression of HDR videos is of special importance as they are significantly larger than LDR videos and in RAW form they can take a sizable portion of a current hard-drive. Transferring them over a current network infrastructure, or using available media would be difficult and even reading and playing such uncompressed data from a local hard-drive is challenging. This section presents video coding algorithms and discusses backwards compatible algorithms in more detail.

Backward Compatible HDR MPEG Video Coding

Mantiuk, Efremov, Myszkowski & Seidel (2006) proposed a backwards compatible method for storing HDR videos. The technique was similar to the JPEG-HDR image compression. The input video stream was split into two streams: a tone mapped and a residual stream. The well-established MPEG-4 video encoder processed each of them separately. The TM stream was displayed on LDR devices while the HDR devices used residual data to generate an HDR image. The main differences from the JPEG-HDR were: usage of the inverse TMO to restore the original HDR, calculating residuals using difference and filtering the residual stream to remove noise invisible to HVS. The full pipeline of the encoder is shown in Figure 3.12.

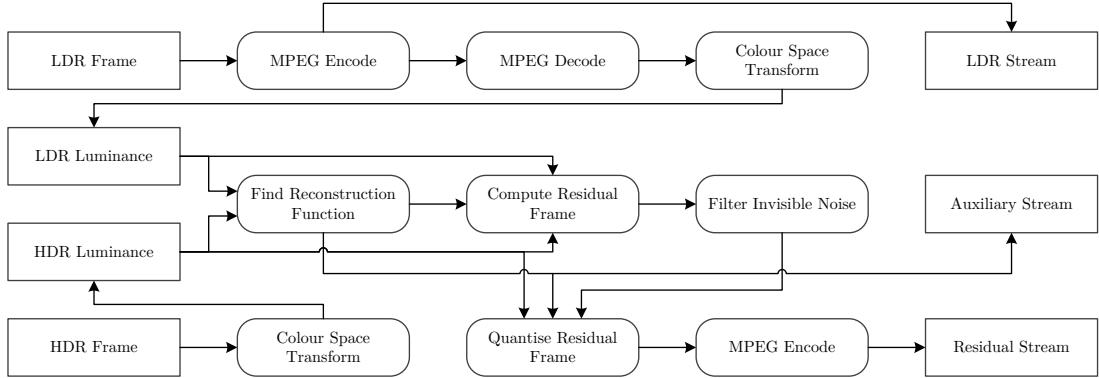


Figure 3.12: The coding process transforms images into the common colour space, allowing residuals to be calculated and subsequently filtered. Both the LDR stream and the residual stream are MPEG encoded.

The pipeline started by MPEG encoding the TM video which was stored - without further processing - as the final backwards compatible LDR stream to be displayed on an LDR device.

To calculate residuals the LDR video was decoded exposing all the errors due to compression. The next step involved the decoded tone mapped and the original HDR streams. They were transformed into a common colour space which decorrelated RGB (or XYZ) representations and allowed comparison of the two. Chromas of both were converted to the CIE 1976 uniform scale (u' , v' , similar to LogLuv) which could represent the full visible gamut. The luma of the TM frame was nonlinearly transformed using sRGB, which had linear and power function parts. The HDR used a different luma coding as the sRGB nonlinearity could not have been used for the values spanning the 10^{-5} to 10^{10} cd/m² range. The authors applied an encoding that was based on contrast detection measurements for the luminance range visible by the HVS. It ensured that the quantisation errors were invisible. The transformation is a piecewise function separately defined across the three dynamic ranges. For exact values please refer to the original work (Mantiuk, Efremov, Myszkowski & Seidel, 2006).

Once both images were in the perceptually similar colour space it enabled the authors to approximate the reconstruction function $RF(\cdot)$ which expanded the TM luma values (l_d) back to the original HDR luma (l_d). It was assumed that the original TMO was unknown. The HDR values were put into 256 bins based on their TM counterparts. The reconstruction function was then calculated by

finding the arithmetic mean of all pixels for each bin Ω_i :

$$RF(i) = \frac{1}{|\Omega_i|} \sum_{\mathbf{x} \in \Omega_i} l_W(\mathbf{x}) \quad (3.18)$$

where $|\Omega_i| = \mathbf{x}|l_d = i$ and $i \in [0, 255]$ was the bin index. The chromaticity reconstruction function was approximated by: $(u'_d, v'_d) = (u'_W, v'_W)$. Residual function data was computed for all frames, stored in the auxiliary stream and Huffman encoded.

After expanding the TM image, the residual frame (r_1) was calculated by a simple subtraction:

$$r_1(\mathbf{x}) = l_W(\mathbf{x}) - RF(l_d(\mathbf{x})) \quad (3.19)$$

The obtained values could range from -4095 to 4095 which, if left uncompressed, required 12 bits. As MPEG optimally encoded 8 bit data, results needed to be scaled and quantised or clamped. A solution was proposed which allowed a trade-off between the errors due to clamping and the errors due to quantisation. To enable even more control a quantisation factor (i.e. scaling value) was set for each bin. So, the final residual image was computed as follows:

$$\hat{r}_1(\mathbf{x}) = \left[\frac{r_1(\mathbf{x})}{q(m)} \right]_{-127}^{127}, \quad \text{where } x = k \Leftrightarrow i \subset \Omega_k \quad (3.20)$$

where $[\cdot]_{-127}^{127}$ was the rounding operator that clamped values below -127 and above 127, and the quantisation factor, $q(m)$, was selected separately for each bin Ω_l :

$$q(m) = \max \left(q_{\min}, \frac{\max_{\mathbf{x} \in \Omega_l} (|r_1(\mathbf{x})|)}{127} \right) \quad (3.21)$$

The scaled and quantised \hat{r}_1 contained high frequency values which hindered compression but could have been removed without losing perceptual quality. Hence, the residual frame was filtered using the original HDR frame as a guide. The operation was performed in the wavelet domain and was applied to the three finest scales as filtering at coarser scales could lead to noticeable artefacts.

The performance of HDR-MPEG was measured using three metrics: HDR VDP (Mantiuk, Myszkowski & Seidel, 2004; Mantiuk *et al.*, 2005), universal image quality index, UQI (Bovik, 2002), and signal-to-noise ratio (SNR). The first study evaluated the influence of a chosen TMO on quality and bit rate. Five operators were tested: time-dependent visual adoption (Pattanaik *et al.*, 2000),

fast bilateral filtering (Durand & Dorsey, 2002), photographic tone reproduction (Reinhard *et al.*, 2002), the gradient domain tone mapping (Fattal *et al.*, 2002), and adaptive logarithmic mapping (Drago *et al.*, 2003). Temporal coherence was preserved by modifying the operators and default parameters were used. All of them exhibited similar performance except the gradient domain one, for which the output was larger. Still, the latter generated images which appeared better during LDR playback, and were more suitable when backward compatibility was important. The second test compared the proposed method against HDRV and JPEG-HDR using the photographic tone reproduction TMO. HDR-MPEG performed better than JPEG-HDR but was similar to HDRV.

Rate-Distortion Optimised HDR Video Coding

Lee & Kim (2008) proposed another method which separates HDR into a TM stream and a residual stream making it backwards compatible. They added two new contributions to the method: temporal coherence was imposed which reduced flickering, and bits were allocated between TM and residual frames in a way which would optimise appearance of both TM and the restored HDR frames. The diagram of the encoder is presented in Figure 3.13.

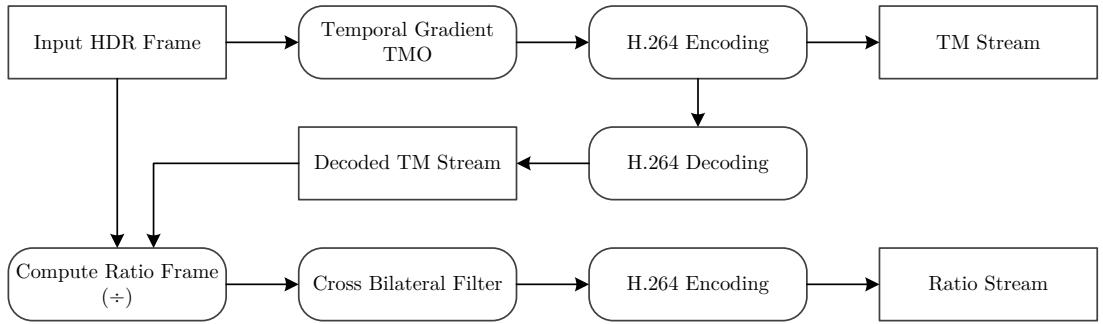


Figure 3.13: The pipeline showing encoding of HDR video using *rate-distortion optimised coding*. The key components are tone mapping that operates in both the spatial and temporal domains, and a technique for improving the quality of the TM image.

First, the input HDR stream was tone mapped using a temporally coherent version of the gradient domain TMO (Lee & Kim, 2007). The TMO reduced flickering by estimating a motion vector field between two consecutive HDR frames which were used to generate LDR pixel values. The TM stream was then encoded using the standard H.264 encoder allowing for backwards compatibility.

The compressed TM stream was then decoded so a ratio image could be calculated using the equation:

$$R(\mathbf{x}) = \log_2 \left(\frac{L_W(\mathbf{x})}{L_d(\mathbf{x}) + \epsilon} \right) \quad (3.22)$$

where L_W , L_d were the luminances of the HDR and LDR frames respectively and ϵ was a small constant which prevented division by zero. The residual values were normalised to the range from 0 to 255. The residuals contained high frequency components due to noise introduced by tone mapping and quantisation. To improve coding efficiency, the edge preserving cross bilateral filter (Eisemann & Durand, 2004) was used on residuals with the luminance of the HDR image as a guide. H.264 was used to encode the residual stream as well.

While previous methods sought to maximise the quality of the reconstructed image, here the authors were concerned with the quality of the TM image as well. To reduce distortions of both the TM and reconstructed HDR sequences (D_d and D_w) they controlled the quantisation parameters of the LDR and ratio sequences (QP_d and QP_{ratio}). The optimisation problem was solved by minimising the Lagrangian cost function:

$$J = D_d + \mu D_w + \lambda(R_d + R_{ratio}) \quad (3.23)$$

where R_d and R_{ratio} were the bit rates for TM and ratio sequences, μ controlled the importance of the HDR sequence and λ determined the trade-off between bit rates and the distortion. The authors analysed J and suggested an equation for controlling the quality of both the TM and ratio streams using only a single parameter QP_d : $QP_{ratio} = 0.77QP_d + 13.42$.

The method was evaluated against MPEG-HDR. Peak signal to noise ratio (PSNR) measured the quality of the tone mapped frames while the HDR visual difference predictor (HDR-VDP) compared the reconstructed frames. For TM frames, the proposed technique had better quality averaging more than a 10 dB difference. The generated HDR frames were better for low bit rates with 10% smaller VDP error, but for the rates above 1 bpp quality was worse with 2 to 5 % larger VDP error. In a final test authors concluded that the ratio stream constituted 10 to 30 % of the total file size.

The authors extended the method (Lee & Kim, 2012) to provide a more effective rate-distortion optimisation at the macro-block level in order to maximize

the quality of both the LDR and HDR streams given the limited bit depth.

Other Video Compression Techniques

Mantiuk, Krawczyk, Myszkowski & Seidel (2004) suggested one of the first methods for compressing HDR videos termed HDR video (HDRV). They used the well-established capabilities of the MPEG-4 video codec and extended it to work with HDR video data. The main characteristic of the proposed algorithm was quantisation of luminance where errors were kept below the just noticeable threshold values of the HVS. To facilitate HDR data, MPEG-4 data structures were expanded from 8 to 11 bits and an efficient coding scheme for DCT blocks was introduced. Three captured video sequences together with rendered videos were used to evaluate the approach. The achieved compression rates were between 0.09 and 0.53 bpp. This was compared to the performance of MPEG encoded TM data only, which was approximately half the size. OpenEXR, on the other hand, required between 16 to 28 bpp for the same sequences.

Mantiuk, Myszkowski & Seidel (2006) also suggested a novel colour space which allowed compression of HDR data while preserving the error below the visibility threshold of the HVS. This space was capable of representing the complete luminance range and full colour gamut visible to the human eye. The current coding algorithms required minor changes to support the proposed colour space. To validate the approach, the authors developed two lossy HDR compression algorithms (for static images and video). They claimed that image compression was “efficient and fast”, but did not provide any results.

Adaptive bit-depth transformation of HDR data was explored by Motra & Thoma (2010) and Zhang *et al.* (2011). Motra & Thoma (2010) transformed HDR images to LogLuv format, which they have optimised for 16 bit floating point numbers. Then quantisation errors were minimised by adaptively utilising levels which were left unused after transformation. Three video sequences were employed to test the approach. Non-adaptive and adaptive techniques were compared to GT using the VDP metric where the percentage of detected errors was significantly lower for the adaptive case. For example, at a 11,200 bit rate for one of the sequence, the VDP error percentage was 8.5 for non-adaptive and 0.01 for the adaptive method. Zhang *et al.* (2011) extended the method by optimising bit-depth quantisation via the Lloyd-Max algorithm (Max, 1960; Lloyd, 1982). In addition invisible high frequency noise was reduced by transforming frames

into the wavelet domain where a contrast sensitivity function weighed wavelet sub-bands. The proposed technique showed improvement over the technique of Motra & Thoma (2010) by achieving VDP results which were between 65% and 18% better.

3.4 HDR Display

While some captured HDR images and videos are aimed for specific applications, such as image based lighting, the ultimate goal is to enable their display to human observers. Display devices which support HDR content still have not penetrated the consumer market and are only available in research laboratories. The available prototypes are still expensive, bulky and lack software support. Still, the new robust SIM 2 HDR display and the increase of current TVs with LED back-lighting hint that HDR might reach homes soon.

Even LDR displays may benefit from the availability of HDR content. As was discussed in Section 3.2.2 many of the consumer cameras capture a dynamic range which is greater than what can currently be displayed. Tone mapping is performed on such data - the operation which reduces the dynamic range of an image. There are many possible ways in which reduction can be performed and it depends on what needs to be achieved (e.g. preserving as much information as possible).

This section first discusses devices which can natively visualise HDR content. Then tone mapping operators are classified and short overview is provided. The *photographic tone reproduction operator* (Reinhard *et al.*, 2002) is described in more detail as it is used in Chapters 5 and 7.

3.4.1 Native HDR Displays

The current way to display HDR content relies on technology termed *local-dimming displays* or *dual-modulation displays*. The idea behind this approach is to optically combine the two displaying devices so that their intensities multiply. For instance these could be two projectors or a projector and an LCD screen. If aligned properly the brightest value which can be displayed is the multiple of the maximum value of each display. In such a setup, the dynamic range is significantly increased and may be controlled on a pixel level. However, in practice it is difficult to achieve this theoretical increase in dynamic range because of optical

imperfections caused by light scatter.

Stereoscopic High Dynamic Range Viewer

Ward (2002) implemented the first device which utilised the dual-modulating technology. Not only did it allow native visualisation of HDR images but it was also a stereoscope (see Figure 3.14). The device consisted of the three main pairs of parts: two wide-field lenses, two bright uniform backlights - 50 watt lamps, and a pair of layered transparencies. The optics used was the Large Expanse Extra Perspective (LEEP) ARV-1 (Howlett, 1990). The two main challenges were mapping the view for the optics and designing a technique for layering transparencies in order to increase dynamic range.

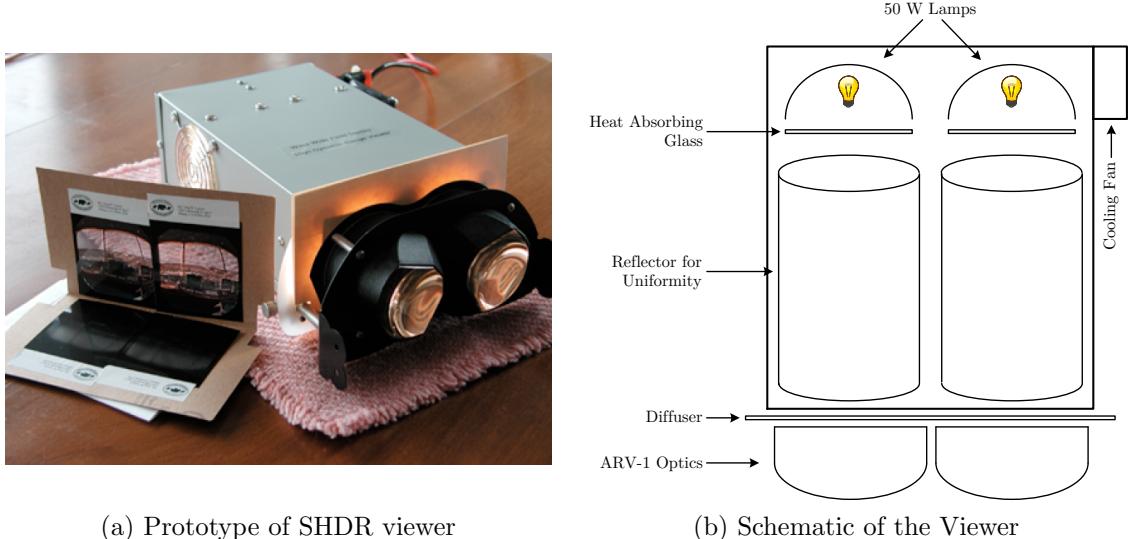


Figure 3.14: The stereoscopic viewer constitutes of three main parts: lenses, lamps and image transparencies.

The ARV-1 optics exhibited major chromatic aberrations resulting in coloured fringes at the view edges. This was offset by a camera with aberration in the opposite direction. To avoid the image appearing blurred, a high resolution was required. The images were printed using 800 dots per inch (dpi) and totaled 2048 × 2048 pixels.

Film transparencies were able to encode only 8 bit images so an extra image was needed to increase the dynamic range utilising the light source modulation (dual-modulation technique). The process of the pair creation is illustrated in Figure 3.15. The first image hit by light modulated it and encoded the global

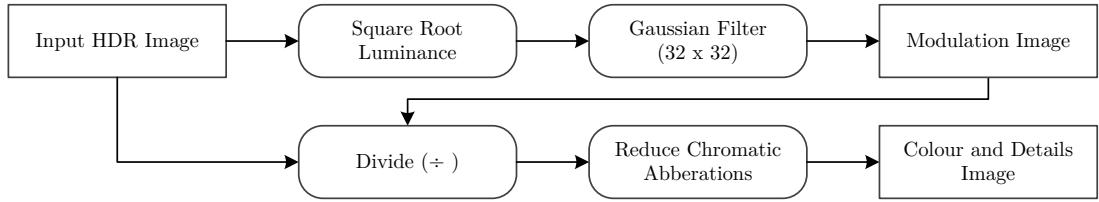


Figure 3.15: Pipeline showing how transparency images were generated for the stereoscopic HDR viewer. The input image is separated into a low frequency monochromatic modulation film and a coloured film.

luminance distribution. The image was generated by calculating the square root of the original luminance and filtering the result with a 32×32 Gaussian. The second image encoded details and colours and accounted for chromatic aberration. It was generated by dividing the input HDR image by the modulated one. Chromatic aberration was counteracted by scaling the red channel up by 1.5% compared to the blue one and positioning the green one halfway between the two.

The viewer provided around 120° field of view and a dynamic range of over $10,000 : 1$ with the maximum and minimum luminance of $5,000 \text{ cd/m}^2$ and 0.5 cd/m^2 . In a user study, Ledda *et al.* (2003) compared the device output with a real scene and the image on a CRT monitor tone mapped with the histogram adjustment operator (Ward *et al.*, 1997). The results suggested that the image presented on the SHDR viewer was closer to reality than to the TM reproduction.

While the system reproduced the scene appearance well it is only limited to static images which can be viewed by a single observer at the time. Moreover, the cost of printing one such image (consisting of four transparencies) was estimated to be approximately \$200 US.

Projector Based HDR Displays

Seetzen *et al.* (2004) developed the first display which was viewed in a conventional way like a TV screen. Similar to the SHDR viewer it was based on dual-modulation technology (as shown in Figure 3.16) and consisted of two devices: a digital light processing (DLP) projector and a transmissive liquid crystal display (LCD). The projector's modulated image increased the dynamic range image while the front LCD encoded colour and details.

While the images were processed in a similar manner to the stereoscopic HDR viewer there were a number of differences. A lack of optics meant no chromatic aberrations hence no counteracting was required. The square root of luminance

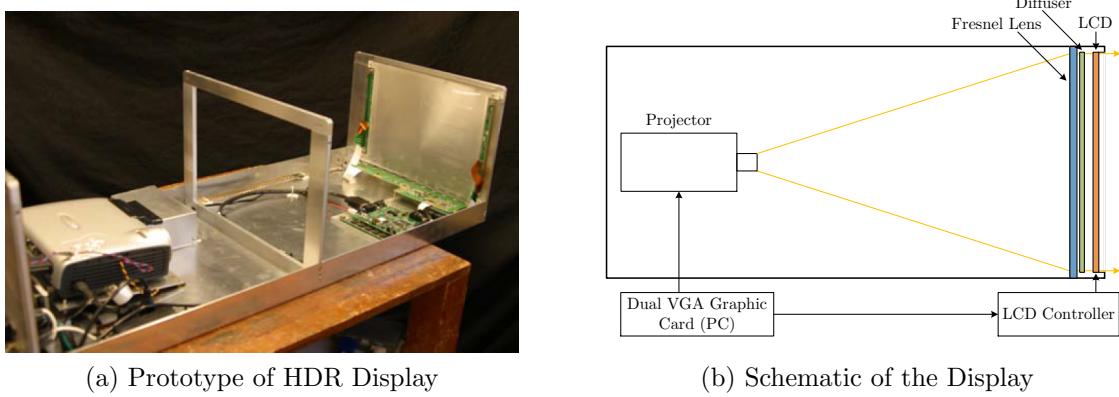


Figure 3.16: The HDR display by Seetzen *et al.* (2004) consists of: a projector providing strong modulated backlighting and an LCD panel providing colour and details.

was filtered based on the point spread function of the projector instead of a Gaussian. Lastly, the signal was linearised by measuring the response functions of both the LCD and the projector and applying their inverse to the modulation image.

The hardware used to build the prototype consisted of a Sharp LQ150X1DG0 15.1" LCD panel (300 : 1 DR) and an Optoma DLP EzPro737 (800 : 1 DR). A Fresnel lens was put immediately behind the LCD to collimate the light into a sharp angle. This helped to achieve a maximum brightness and avoid colour distortion. A standard LCD diffuser, put between LCD and Fresnel lens, redistributed the collimated light and achieved an acceptable viewing angle. The final prototype achieved a dynamic range of 50,000 : 1 with a peak luminance of 2,700 cd/m² and a minimum measured luminance of 0.054 cd/m².

One drawback of the design was the large optical path (approximately 1 m) that rendered the device impractical for everyday use. While the diffuser helped to increase the viewing angle it was still rather limited and the device had a strong falloff at wide viewing angles. Lastly, the projector generated very bright light, which resulted in significant power consumption and heat generation.

LED Based HDR Displays

To overcome practicality issues, Seetzen *et al.* (2004) built a second HDR display which resembled classic TV sets. The display was based on the concept of *veiling glare* in which local contrast on the retina is reduced because of light scattering in

the eye. This phenomenon impacts visibility of dark detail next to bright regions and meant that the HVS processes local contrast poorly, around edges, compared to the global contrast, which is present over a wider spatial range. Hence, it was not required to generate very dark pixels next to very bright ones as long as those intensities could be presented in different parts of the image.

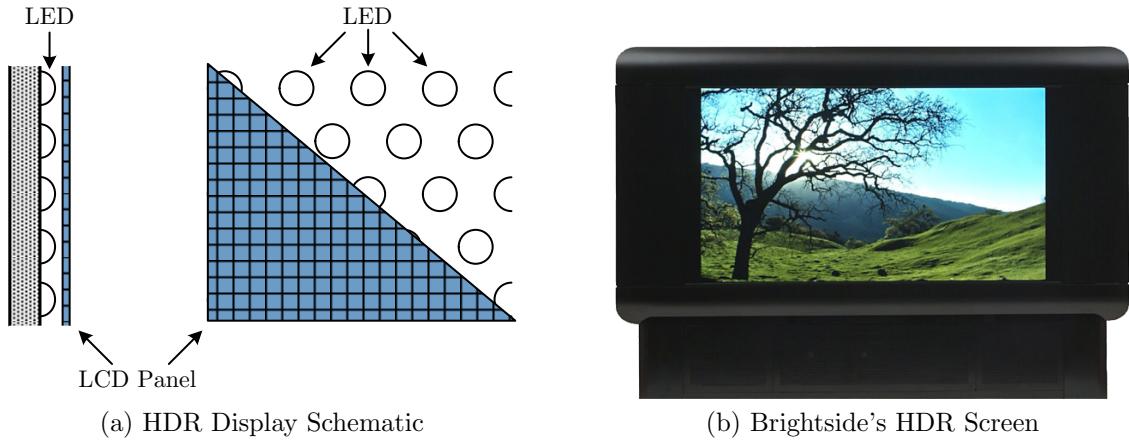


Figure 3.17: HDR LED displays use a low resolution LED array as backlighting and a standard LCD screen in the front. (a) A schematic shows the design of a display from the side and front. A single LED provides backlighting for multiple pixels. (b) One of the first HDR LED display prototypes.

The authors utilised this insight and used a low resolution LED panel to modulate the light and the LCD panel for details and colour (see Figure 3.17a). The way in which an input image was processed was similar to the projector-based displays, except for the step in which the luminance for a single LED was calculated. The square root of the luminance of the input image was downsampled to the resolution of the LED panel and approximately deconvolved by the LED point spread function.

The DR 37P (see Figure 3.17b) consisted of 1380 LEDs and a 37" LCD panel (300 : 1 DR). It was capable of a 200,000 : 1 dynamic range with a peak luminance of 3,000 cd/m² and a minimum luminance of 0.015 cd/m². While the dynamic range was improved, the image quality suffered due to the low resolution of LEDs. In addition it consumed a considerable amount of power (1680 W) and required an extensive and noisy cooling system (both fans and liquid cooling).

Dolby Labs acquired Brightside in 2007 and licenced it to SIM2 who in 2009 released the first commercial HDR display named the Solar 47. The display size was 47" with a full HD resolution of 1920 × 1080 and 2,206 LEDs. It was also

capable of processing 16 bpc images.

3.4.2 Tone Mapping

HDR displays are still not common while HDR data is available. The captured dynamic range of the image frequently cannot be displayed on an LDR screen. Two main options are possible when displaying HDR data on an LDR screen. Either the range outside displayable limits is discarded (cropped) or it is scaled in order to preserve the available data. The problem predates digital cameras and was troubling photographers that used analogue equipment. Camera films and negatives contained more data than what was possible to print. Skilled photographers used techniques of *dodging* and *burning* which would allow regions of the image to be shifted into the printable range. This process of scaling the dynamic range (digital or analog) is termed *tone mapping*.

The process of tone mapping is loosely defined and leaves plenty of space for interpretation. The dynamic range can be reduced in multiple ways and a ground truth does not exist. The main guide should be the intended function of the image and the medium it will be displayed on. For instance, the goal might be to use a TM image for compression so preserving available information is a priority. One might try to keep the appearance of a captured scene faithful to the original and even account for the ambient lighting on the fly. Alternatively, the purpose of the image might be artistic and it could appear abstract and unrealistic. Similarly, the content may be intended for different media. The image might be printed requiring a very limited dynamic range or it may be displayed on a traditional LDR screen. Even when displayed on an HDR screen, extreme values might exceed the capabilities of the device, demanding slight tone mapping adjustments or cropping. It is not surprising that the number of proposed tone mapping operators is substantial and that their creation and evaluation is one of the most active areas of HDR research. The one aim, frequently strived for, of tone mapping operator (TMO) is a faithful representation of the depicted scene on an 8 bpc LDR screen.

Formally, a tone mapping operator f performs a tone mapping process, described in Banterle *et al.* (2011) as:

$$f(I) : \mathbb{R}_i^{w \times h \times c} \rightarrow \mathbb{D}_o^{w \times h \times c} \quad (3.24)$$

where I is the image, $\mathbb{R}_i \subseteq \mathbb{R}$ are the real input values of an HDR image, $\mathbb{D}_o \subset \mathbb{R}_i$ are the output values of the tone mapped image, w , h and c are the width, height and number of channels in the image. Typically, $D_o = [0, 255]$ and $c = 3$ when images are mapped for the LDR monitors which use 8 bpc, RGB representation. However, the process is usually performed on the luminance channel only, disregarding the colours. The output image is calculated by dividing the input image's colours with its own luminance - and multiplying the result with the new tone mapped luminance. This operation is expressed as:

$$f(I) = \begin{cases} L_d = f_L(L_w) : \mathbb{R}_i^{w \times h} \rightarrow [0, 255] \\ [R_d, G_d, B_d] = L_d \left(\frac{1}{L_w} [R_w, G_w, B_w] \right)^s \end{cases} \quad (3.25)$$

where $s \in (0, 1]$ is the parameter that controls saturation which is usually increased during tone mapping. Once an image is tone mapped as in Equation 3.25, the output is linear so gamma correction is applied. The gamut of the HDR image is often greater than that of the output image, so the multiplication of *RGB* values with the luminance may result in the product exceeding the 255 range. In those cases values are clamped and methods that improve colour appearance have been proposed (Akyuz & Reinhard, 2006; Kuang *et al.*, 2007; Mantiuk *et al.*, 2009).

Tone mapping operators, in general, are separated in two groups based on the way in which function f is applied. *Global operators* apply a single function across all the pixels in the image, while *local operators* may apply different functions depending on local image content.

Global operators preserve global contrast because they treat all pixels equally. In some instances, the operator calculates image statistics in an initial pass (e.g. maximum and minimum luminance, logarithmic and arithmetic averages). The obtained data optimises the parameters used for dynamic range reduction. As the maximum and minimum values in the image are likely to be affected by noise (due to capturing sensor limitation), it is common to discard a predetermined percentile of these extreme values, thereby increasing robustness of the operator. Extending a global tone mapper to video can be performed by filtering the image statistics or computed parameters in the temporal domain. In such a manner, flickering caused by noise and sharp changes in consecutive frames are reduced. Temporal tone mapping techniques are a subject of current research (Banterle *et al.*, 2011; Boitard *et al.*, 2012). The major disadvantage of global operators is

the loss of fine details and local contrast due to strong value quantisation which disregards local region appearance. Global operators include the ones proposed by Schlick (1994), Ward (1994a), Ferwerda *et al.* (1996), Larson *et al.* (1997), Pattanaik *et al.* (1998), Tumblin *et al.* (1999), Reinhard *et al.* (2002) and Drago *et al.* (2003).

Local operators try to preserve both local and global contrast thereby improving the appearance of the tone mapped image. Instead of performing the operation over the whole image at once, in this approach the operator f considers the neighbourhood of each pixel. This is computationally more expensive and may result in image artefact such as prominent halos around edges. While these artefacts may be desired (e.g. when drawing attention to specific area) many operators introduce techniques to reduce them. Extending local operators to the temporal domain is more challenging compared to the global ones, as small local regions are less robust to noise and intensity changes and preserving constancy may require significant computation. Local operators include the ones proposed by Chiu *et al.* (1993), Pattanaik *et al.* (1998), Ashikhmin (2002), Reinhard *et al.* (2002) and Ledda *et al.* (2004).

Detailed descriptions of all the aforementioned operators are provided in the works of Banterle *et al.* (2011) and Reinhard *et al.* (2010). To illustrate both approaches, the *photographic tone reproduction* operator is described in more detail next. It comes in two versions - global and local - making it particularly suitable for the discussion. Moreover, this operator is used in Chapters 5 and 7.

Photographic Tone Reproduction

As mentioned above, the problem of tone mapping outdates digital media. Photographers have been faced with the issue for more than 150 years. The Ansel Adam's Zone System was a method where photographers utilised information obtained in the field to improve the appearance of the final print. They would try to predict the mapping of scene luminances to a print. The measure of a surface's luminance - perceived as a middle-grey - was typically mapped to the 18% reflectance of the print. For bright scenes the middle-grey would be mapped to a lower value and for the dark scenes to higher ones. When the captured dynamic range exceeded the printable one, problematic areas were mapped to pure black or white. Alternatively, the photographers applied dodging and burning techniques which locally modified the brightness. While the described process was

difficult to automate it provided the photographer with a few subjective controls.

Reinhard *et al.* (2002) proposed a tone mapper which was inspired by the Ansel Adam's Zone System. The log average luminance was used as to approximate brightness of a scene and was mapped to 0.18% of the display's range. If the whole scene was brighter or darker, an alternative value may be used as in the photographic process. The initial range reduction was given as:

$$L_m(\mathbf{x}) = \frac{a}{\bar{L}_w} L_w(\mathbf{x}) \quad (3.26)$$

where \bar{L}_w was the log average luminance and a was the value to which it was mapped (default 0.18). Typically, the majority of the captured image values fall in the middle of dynamic range with smaller regions in the highlights and shadows. Traditional photography emphasised this midrange by applying a sigmoid function to the image, which compressed high and low luminance values. In modern photography, only the high luminances were reduced which may be expressed as:

$$L_d(\mathbf{x}) = \frac{L_m(\mathbf{x})}{1 + L_m(\mathbf{x})} \quad (3.27)$$

This function brought all the luminances into the range [0, 1] which in some cases was not desired. For instance, extreme values sometimes needed to be discarded, clipping the range. To allow for this, Equation 3.27 was combined with linear mapping resulting in the *global photographic tone reproduction operator*:

$$L_d(\mathbf{x}) = \frac{L_m(\mathbf{x}) \left(1 + \frac{L_m(\mathbf{x})}{L_{white}^2} \right)}{1 + L_m(\mathbf{x})} \quad (3.28)$$

where L_{white}^2 was the clipping threshold that mapped any larger value to white. The global technique was quick to compute and produced images which preserved detail in low contrast areas and scaled high luminances to a displayable range. When the range was very high images lost important details. For these a local tone operator which simulated dodging and burning was proposed.

The local version was, essentially, the global operator applied to smaller image regions. These regions were found for each pixel by looking for a neighbourhood which did not contain any sharp contrast. The measure used to differentiate between high and low contrast was the traditional centre-surround difference. Two Gaussian-weighted averages of different sizes, centred at the same pixel

were computed and subtracted. If the difference was small, so was the contrast. However, the contrast edge present in the surround Gaussian but missing from the centre would cause a large difference. Given the scale s , a blurred image was calculated by convolving a Gaussian R_s with the corresponding image region L_m expressed as: $L_s^{\text{blur}}(\mathbf{x}) = L_m(\mathbf{x}) \otimes R_s(\mathbf{x})$. The centre-surround difference was then calculated as follows:

$$V_s(\mathbf{x}) = \frac{L_s^{\text{blur}} - L_{s+1}^{\text{blur}}}{2^\Phi a/s^2 + L_s^{\text{blur}}} \quad (3.29)$$

where the normalisation term $2^\Phi a/s^2$ enabled thresholding the result by a common value which was shared by all scales. This was required as V_s depended only on local luminance values. The control parameter Φ represented sharpening. When set too small then V_s was similar to the luminance L_m and the local operator reduced to the global one. When set too big, Φ caused rings around bright regions (i.e. halo artefacts) but when correctly chosen it preserved enough detail and contrast without introducing artefacts.

Equation 3.29 was calculated for increasingly large Gaussians around the pixel of interest. The largest area with low contrast was defined by the largest scale s_{\max} for which the difference of Gaussians was below a given threshold ε :

$$s_{\max} : |V_{s_{\max}}(\mathbf{x})| < \varepsilon \quad (3.30)$$

Finally, the local operator emulating dodging and burning techniques was given as:

$$L_d(\mathbf{x}) = \frac{L_m(\mathbf{x})}{1 + L_{s_{\max}}^{\text{blur}}(\mathbf{x})} \quad (3.31)$$

While the local operator was less computationally efficient than the global one its performance was increased by executing a scale selection mechanism on the fly. Also a Gaussian pyramid was computed in a preprocess and the most appropriate scale was selected during tone mapping.

Ledda *et al.* (2005) performed a user study in which they compared the similarity of different TMOs to the original HDR image. The *balanced paired comparisons* method was used and four cases were tested: colour TM image, greyscale TM image, similarity of dark regions, and similarity of bright regions. The local photographic tone reproduction operator performed well for all of the cases being placed 2nd, 1st, 3rd, and 1st respectively.

3.5 Summary

This chapter introduced high dynamic range imaging. It provided definitions of many underlying concepts including dynamic range, high dynamic range and exposure value. Methods of capturing HDR content (e.g. the multiple exposure technique) and devices which allowed native capture were described. The following sections discussed the main file formats and compression methods. Finally, display devices and tone mapping operators were covered. The next chapter brings the two presented techniques together - stereoscopic and high dynamic range imaging - recognising challenges which arise when they are combined and discusses possible solutions.

CHAPTER 4

Stereoscopic High Dynamic Range Pipeline

In the introduction the traditional imaging pipeline was discussed. The previous two chapters, explained how the pipeline requires major modifications to enable alterations, such as capturing two views or using floating point numbers instead of integers. Considerations and trade-offs were needed to facilitate capture, storage and display of stereoscopic and high dynamic range imaging. While researchers have made progress and proposed standards for all the stages, both pipelines still lack the robustness, accessibility and acceptance of traditional imaging. So far, no attempts have been made to combine stereoscopy and high dynamic range imaging. This chapter examines what such a merge requires and how it affects each of the processing stages. Challenges and problems which might arise are recognised and possible answers are discussed with advantages and disadvantages which they might entail. This chapter sets the stage for the upcoming ones which propose solutions to some of the problems discussed in this chapter and enable scenes to be captured using SHDR technology, compressed and displayed to the observer.

4.1 SHDR Capture

Currently, the most straightforward way to generate SHDR content is using computer graphics. The majority of renderers already use floating point numbers to represent data thereby providing HDR content. To obtain the second view for stereo another virtual camera is added. In most CG software packages these slight modifications can be achieved quickly and existing scenes may be ren-

dered in SHDR without significant effort. Moreover CG packages allow precise control of the camera settings, so many challenges encountered in real capture are avoided. These include alignment of stereo images, synchronisation of video frames, synchronisation of zoom and focus, colour calibration, increased noise in dark regions and limited dynamic range (even HDR cameras eventually get saturated). SHDR content produced using CG may achieve a quality hard to replicate in real scenes. However, adding a second view may double rendering time. This becomes a significant problem especially when rendering photo-realistic scenes using global illumination. Stereoscopic rendering techniques which optimise the process (Lo *et al.*, 2009), as described in Section 2.4.4, may be applied in such cases and improve the speed. Also, existing renderers may be used without modifications, where a single image and depth map are generated. The depth map could subsequently be used to generate the second view.

Capturing real scenes in SHDR is more challenging. The dynamic range of camera sensors is improving with each new generation. However, none of the existing photo cameras supports native HDR capture, instead they rely on the multiple-exposure technique to obtain HDR content with all of its limitations (Section 3.2.1). Prototypes of HDR video cameras already exist, but they are bulky and expensive. For example, University of Warwick's HDR video camera requires a 120 kg HDD array and customised fiber optic cables. It may be expected that these devices will move from the research to the consumer realm and in the process some of the issues will be tackled. Still, it is hard to predict when two affordable and practical HDR cameras mounted side-by-side in a stereo rig will become available. Alternative more accessible solutions are required which would enable the capture of SHDR content.

Two solutions (Kang *et al.*, 2003; Lin & Chang, 2009), whose primary goal was HDR capture, could with modifications be used and even combined to record SHDR video. Kang *et al.* (2003) alternated between two exposures while video was recorded. They tracked movement between consecutive frames which allowed them to generate HDR frames using the multiple exposure technique. By modifying the approach and adding a second camera, SHDR video could be recorded. Similarly, Lin & Chang (2009) captured two frames of different exposures but instead of separating them temporally they used spatial separation, i.e. a stereo camera pair. They matched the features between the two images, warped one of them to align it with the other and then combined the two into the HDR image. This process could be used for both images to generate an SHDR one, but was

not explored by Lin & Chang (2009). The main problem with both approaches is that they only capture two exposures which is not sufficient to cover HDR (Mann & Picard, 1995). In addition, matching pixels requires significant overlap in dynamic range reducing its final gain. This is even more the case for the spatially separated technique as differences between matching pixels are larger compared to small movements between the frames. Theoretically, the two techniques could be combined where two cameras across two frames would capture four different exposures. However, merging such data would require three pixel matching operations which would accumulate errors and many artefacts could be expected in the final video whose range could still not match the native HDR one.

A more robust technique should use at least one HDR sensor which would guarantee that the full dynamic range of the scene will be captured. This data could then be transferred to the second view to create SHDR content. One possible approach is to use a single HDR camera and a depth sensor which would generate a depth map. The map could be used to guide HDR information transfer to the second view. However, current depth sensors also have limitations. They are of low resolution (e.g. 640×480) and low precision which reduces with the distance from the objects. It is expected that this HDR plus depth technique would work well for close up scenes but for the majority of scenes depth map precision would be insufficient to produce satisfactory results.

An alternative approach is to combine an HDR and an LDR camera and use the latter to generate the second view. Here the full dynamic range is captured in one of the views while the other contains a subset of that range. This means that they fully overlap allowing for robust matching. In addition, the second image or video can have the same (or higher) spatial and temporal resolution as the HDR one allowing for a high quality disparity map generation. The map's precision becomes less affected by the depth of the scene compared to using a depth sensor. Another advantage of such an approach is that it provides film directors with a traditional camera allowing them to use filming methods they are accustomed to. This may play an important role in industry embracing this novel technology. Chapters 5 and 6 propose and examine in detail techniques which enable SHDR capture from HDR-LDR image and video pairs respectively.

4.2 SHDR Storage

Previous chapters discussed how raw stereoscopic data consumes twice the space while raw HDR data quadruples the required space compared to traditional imaging. This means that raw SHDR content would increase demands on space by 8 times if left uncompressed. To illustrate the problem, the following facts may be considered - a single raw high definition SHDR image would take approximately 50 MB, 4 seconds of SHDR video running at 25 fps would take around 4.8 GB and could not fit on a standard DVD while an hour and a half movie would require 6.7 terabytes (TB). The current media and transfer technologies cannot support SHDR content in its raw form.

If SHDR is to be accepted it should initially support backwards compatibility allowing all four possible combinations to be presented to the viewer: monoscopic LDR, monoscopic HDR, stereoscopic LDR and stereoscopic HDR. A truly accessible SHDR image would be the one that could be opened using any standard viewer and that would correctly display the appropriate content.

Compression techniques have been proposed for both stereoscopic and HDR imaging and were reviewed in previous chapters. Most HDR encoders separate the image into an LDR part and a residual part which are then encoded using traditional image encoders. The LDR part can be displayed using a standard viewer while the second is used by an HDR viewer to boost the dynamic range and compensate for errors.

In stereoscopic encoding the correlation between two views is utilised to reduce file size. Disparity maps may be calculated and stored together with a single view and residuals. The advantage of using disparity maps is that they are monochromatic, usually with a limited number of values and of low frequency, hence especially suitable for compression. Backwards compatibility is achieved in a similar manner to the HDR case. A single view is displayed in a traditional viewer, while the stereo one uses disparity and residual data to generate two views.

The main question is if both stereo and HDR coding techniques are compatible and how can they be combined. Chapter 7 proposes and compares five methods to compress SHDR images based on stereoscopic and HDR encoding techniques. Backward compatibility is achieved by all of them. The best performing method is inspired by video motion compensation techniques and while tested only for single images it can be easily extended to work with video data.

4.3 SHDR Display

This thesis is concerned with the software side of the SHDR imaging pipeline. The display of the SHDR content, on the other hand, is strongly coupled with the hardware technology.

The technique for displaying HDR content on a screen relies on a combination of LED and LCD panels, while with stereo there are number of competing technologies many of which could be merged with HDR. As LEDs can have high refresh rates it would be possible to use them with high frequency LCD and shutter glasses to provide SHDR. It is unknown how increased brightness would affect the view separation and if it would increase crosstalk. Another approach is putting a parallax barrier or a lenticular system in front of current HDR screens to enable stereoscopy. A possible challenge is caused by the LED panel which is of low resolution and cannot be controlled on a pixel level as required by these systems. As the differences between two images is small it might be possible to account for such changes using just the LCD panel, but for large disparities this might not be possible.

Large scale SHDR displays could be achieved by polarising light from two projectors thereby separating the two views. To preserve the dynamic range, the screen onto which the image is projected should be highly reflective. The effect such a setup would have on crosstalk is unknown and should be explored further.

The problem of tone-mapping is also relevant for SHDR images. A straightforward approach is to map each image of a stereo pair independently. However, many operators use parameters which are based on image statistics so the two images might look different after processing and hinder the comfort of stereo vision. For example, the maximum luminance value is frequently used as a TM parameter and it is affected by specular highlights which are view dependent. As this parameter influences brightness of the TM image, one stereo view might look darker and cause problems when fusing a stereo image by the HVS. A similar problem is encountered in tone mapping HDR videos where these differences may occur between consecutive frames. To overcome this problem for the video case Mantiuk, Efremov, Myszkowski & Seidel (2006) filtered parameters so that the difference between TM frames was kept below a visibility threshold. The same approach could be used for SHDR images while SHDR videos would require parameter filtering in both the spatial and temporal domain.

Yang *et al.* (2012) proposed a binocular tone mapping framework in which

they used stereoscopic displays to present a TM image. Two TM versions were generated each using either a different set of parameters or different TM operators. The goal was that one image preserves global contrast while the other preserves details. When displayed stereoscopically, the pair was fused into a single image and a user study confirmed that it increased visual richness compared to a single tone mapped image. The major challenge was generating a pair which would avoid binocular rivalry (the inability of the HVS to fuse the stereo pair) and visual fatigue. To this end Yang *et al.* (2012) developed a *binocular viewing comfort predictor (BVCP)* which together with a visual difference predictor (VDP) measurement identified image pixels that cause rivalry. This allowed for TM parameters to be optimised so that the two TM image versions contained most information when presented stereoscopically to the observer without causing discomfort. Yang *et al.* (2012) mentioned the applicability of this technique to stereoscopic data where TM parameters would be calculated using one view and then applied to both. This approach failed to consider a view dependent phenomena (e.g. specular highlights) or occluded regions.

4.4 Summary

Switching from traditional to SHDR imaging pipeline requires changes in each stage. SHDR images or videos can be generated using rendering which can be optimised using reprojection. Capturing two views using a pair of native HDR cameras is currently impractical. Alternatives include a single HDR camera combined with a depth sensor but the quality of such an approach decreases with distance. Combining an HDR with an LDR camera is more promising and such an approach is explored further in the following two chapters. Compression of the SHDR content is required if it is to reach the consumer market. Backwards compatibility allows delivery of all existing technologies: monoscopic LDR, stereoscopic LDR, monoscopic HDR and SHDR. Chapter 7 explores five backward compatible techniques for encoding SHDR data.

CHAPTER 5

Stereoscopic High Dynamic Range Images

This chapter proposes an approach for generating SHDR images using an HDR-LDR static camera setup. This consists of one native HDR camera and a standard LDR camera. As there is a significant overlap between the two captured views, HDR data from one viewpoint can be used to restore the missing HDR data in the other view using the LDR data as guidance. To this end, two general methods are proposed. These are based on the techniques of stereo correspondence and dynamic range expansion.

The benefits of the proposed approach include capturing of an SHDR image in one go, hence reducing or completely avoiding ghosting artefacts. Having an LDR image in the pair allows for capturing of one of the views in a traditional manner, to which users might be more accustomed, thereby helping the adoption of the technology. Moreover, methods developed for images can further be extended to videos. The generation of SHDR video using an HDR-LDR video pair is examined in Chapter 6, where insights obtained here guide the development of video techniques.

The viability of the HDR-LDR to SHDR approach was verified in a user study in which SHDR images generated using four specific techniques were compared to ground truth (GT) images (natively captured using an HDR-HDR stereo pair). To this end, a custom stereo rig with four mirrors was built and enabled SHDR content to be displayed using two HDR screens. A *Balanced pair comparison* method was used to evaluate how close the proposed techniques were to GT and allowed ranking of the techniques. Results confirmed the viability of the approach by identifying a technique which was statistically indistinguishable from GT for

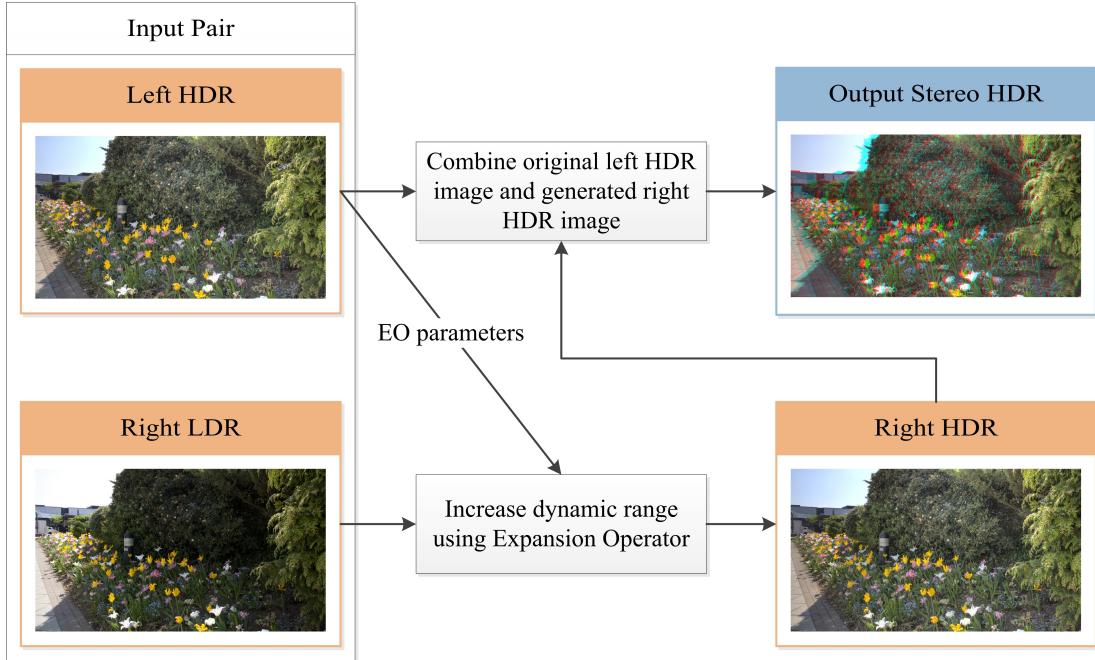


Figure 5.1: In the EO approach the dynamic range of the LDR image is increased by applying the EO which uses parameters obtained by analysing the available HDR image. Without loss of generality it is assumed the left image is of HDR while the right is an LDR image.

the tested scenes.

Section 5.1 describes four proposed methods for generating SHDR images from an HDR-LDR image pair. Section 5.2 provides the description of the user study that examined which method was closest to the ground truth. Finally, results are presented and discussed (Sections 5.3 and 5.4).

5.1 LDR to HDR methods

Two general methods for generating the HDR-HDR stereo image from an HDR-LDR stereo pair were considered. The first relied on methods for increasing the dynamic range of an image by applying expansion operators (EOs), as shown in Figure 5.1. Expansion was performed using parameters obtained from the available HDR image. The second method used stereo correspondence to transfer HDR data to an LDR image (Figure 5.2).

Generating an HDR image from an LDR one is an ill-posed problem for which an exact solution cannot be found, because the required data is missing and can only be estimated. The LDR image captures only a subset of the full range, and



Figure 5.2: In the stereo correspondence approach the missing HDR view is generated using a disparity map which guides the transfer of HDR data. The disparity map is obtained by performing stereo matching between the LDR image and its corresponding LDR slice of the HDR image.

is scaled and quantised appropriately into an 8-bit range (per channel). Overexposed and underexposed regions in the LDR image are outside the captured range, lack any information, and are difficult to reconstruct on their own. Out-of-range pixels may be recovered from the HDR image which may contain those missing regions. However, the two images are not aligned, so mapping between pixels in the left and the right view depends on the depth of the imaged object from the camera. This means that in most circumstances the reconstructed HDR image will not perfectly correspond to the captured scene. However, binocular fusion allows the HVS to cope with discrepancies in a stereo image to an extent, thus the proposed methods are able to provide a viable and efficient solution for generating SHDR.

The rest of this section describes the EO and stereo correspondence methods used to generate SHDR images.

5.1.1 Expansion Operators

EOs expand the dynamic range of an LDR image based on the input parameters. When an HDR-LDR stereo pair is provided, the parameters which drive the expansion can be obtained using the data from the available HDR image, to try to best match the appearance of the expanded image to that of the original HDR image. This results in an HDR-HDR stereo pair, i.e. an SHDR image.

Multiple EOs are available, so to evaluate how the choice of an operator influences the generated SHDR image, two representative operators are tested. The first is linear scaling (LS), a global operator, which applies a curve to all the pixel values in an image, while the second is based on the expand maps (EM) and was shown to perform best in a user study by Banterle *et al.* (2009). Both operators were overviewed in Chapter 3. Their application for SHDR image generation is explained next.

Linear Scaling (LS)

The LS operator normalises and scales the luminance of pixels so that the maximum luminance corresponds to the desired value. It only takes two parameters: a desired maximum luminance and an expansion gamma curve. The first was set to correspond to the maximum luminance of the available HDR image while the second was set to a value of one, as suggested by Akyuz *et al.* (2007).

The LS operator is fast and can be implemented in real time. However, it does not explicitly handle overexposed and underexposed regions. Having only two parameters limits ability to control the appearance of the output image, and the expansion process is guided by the extreme value (which is unreliable as it is likely to be the result of noise). As all the values are scaled linearly and the maximum value comes from the intensity which is overexposed in the LDR image, expanded values will be biased towards the highlights, and the final HDR output is expected to be brighter than the original HDR image. The process of binocular fusion is less affected by the differences in brightness (MacMillan *et al.*, 2007) compared to differences in shape, so the HVS might be able to combine two views so it appears indistinguishable from the ground truth (the native HDR-HDR stereo pair).

Expand Maps (EM)

The EM operator scales the luminance of an LDR image non-linearly (Banterle *et al.*, 2006). It can apply any provided expansion function to the LDR image, but the authors experimented with the inverse Photographic TMO (Reinhard *et al.*, 2002), which is also used here. The inverse Photographic TMO requires two parameters: one which controls the maximum luminance and the other which controls the shape of the expansion curve. The maximum luminance is obtained from the available HDR image. The second parameter is set to the half of the maximum luminance as suggested by Banterle *et al.* (2006). In addition, the EM operator tries to reconstruct the overexposed regions. Here, a single parameter influences estimation of light sources. The parameter can be provided or calculated automatically, and the latter option is selected.

The EM operator is slower than the LS one but the authors claim it can run in real time when implemented on a GPU. As the method which outperformed LS for monoscopic images (Banterle *et al.*, 2009), the EM approach is expected to repeat the result for stereoscopic images. However, control of the final output is still limited to a few parameters so an expanded image which matches the appearance of the existing HDR one is not guaranteed.

5.1.2 Stereo Correspondence

The second general method for generating SHDR images relies on stereo correspondence. Two pixels, one from each image in a stereo pair, are matched if they

represent the same point in 3D space. It is assumed that the colour values of the imaged 3D point are equal in both stereo views. Matches allow for the transfer of HDR data from the available HDR image (view) onto the corresponding position of the other view, thereby creating a novel HDR image.

Matching cannot be easily performed between the HDR and LDR image pair as their values do not correspond. To overcome the issue, a single exposure image is obtained from the HDR image, so it corresponds to the LDR one. This is achieved by traversing through different exposures of the HDR image and extracting LDR slices. The slice which minimises the histogram difference between the available LDR image and itself gets selected. Pixel matching between this exposure and the LDR image is performed in a standard manner and generates a disparity map (see Chapter 2).

Calculated disparities guide the HDR value transfer. The existing HDR image is warped by a straightforward transfer of intensity values using stereo matches as a map between left and right image pixels. A more advanced method could be employed here. For instance, a technique proposed by McMillan & Bishop (1995) tries to avoid a many-to-one mapping but takes more time to compute. Such mappings are unlikely to significantly affect the results so a faster method is used.

A large number of the stereo correspondence algorithms exist (more than 100) and they are divided into the global and local. To test how the choice of a matching algorithm influences a generated HDR image, a representative from each method is evaluated. The local method selected is the *sum of absolute differences (SAD)* algorithm which calculates the matching cost by subtracting the values of two pixels, one from each view. SAD is commonly used to quickly find stereo correspondence and is frequently used for cost calculation (Scharstein *et al.*, 2001) Global methods are represented by the *correspondence with occlusion via graph cuts (COGC)* algorithm which enforces smoothness of the disparity map and handles occlusions explicitly. It was selected due to ability to explicitly detect occlusions as discussed below. Both algorithms were overviewed in Chapter 2. The following sections explain how they are applied for SHDR image generation.

Sum of Absolute Differences (SAD) Correspondence

Once the LDR image is extracted from the HDR image, the SAD algorithm is used in a straightforward manner to obtain a disparity map between the extracted

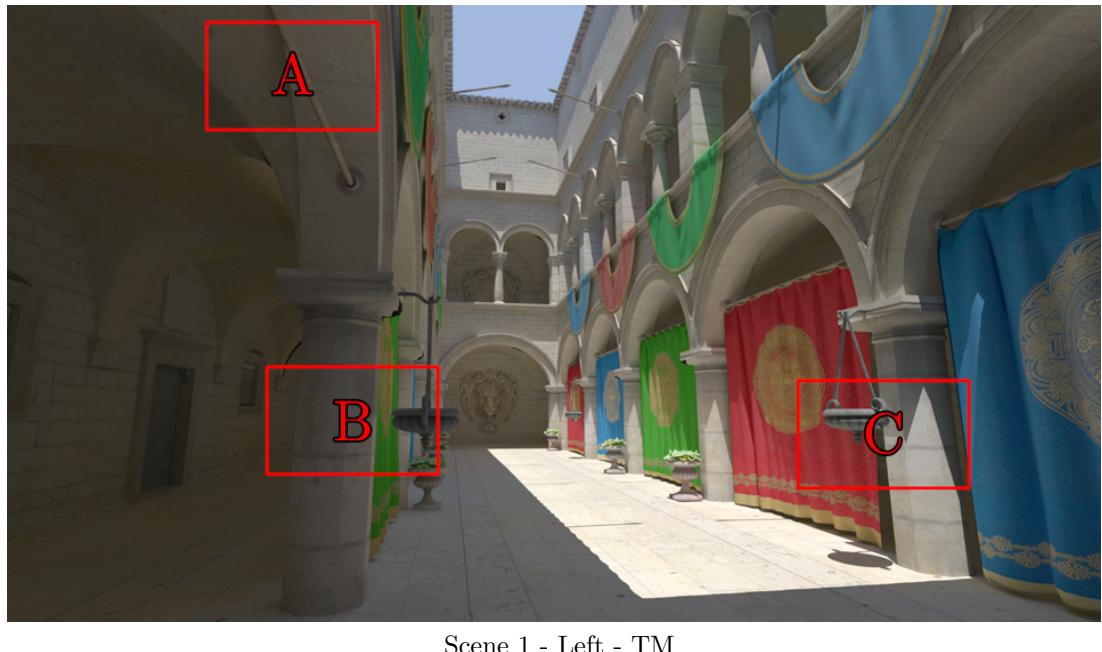
and the existing LDR image. The algorithm requires a single parameter which determines the size of the window used for matching cost aggregation. The parameter is affected by the resolution of an image (e.g. images of a larger resolution require a larger window size to achieve a similar result). For the purposes of HDR image generation the parameter is determined empirically and the window size is set to five for all the images (which were of equal resolution).

The SAD algorithm produces disparity maps quickly and can be implemented in real time. It finds correct disparities for high frequency regions but performs less well for smooth regions, especially if they are larger than the window size. For each pixel in one image, the algorithm finds the pixel in the other image which is closest in colour. While this does not always result in a correct disparity it is well suited for HDR image generation as the transferred values will be similar in intensity (even if they come from inaccurate disparity match). This property is important for the view dependent phenomena, such as occlusions and specular highlights, for which correct disparity matches are unavailable. The SAD algorithm finds intensities that are closest in value for these problematic pixels.

Correspondence with Occlusion via Graph Cuts (COGC)

COGC produces a more accurate disparity map compared to the SAD algorithm. It also recognizes occluded regions and fills them with the disparity of the neighbouring background object. This process is beneficial for depth estimation as occluded objects are likely to be at the same depth as neighbouring background objects. However, these estimated values cannot be used for transferring HDR data, because corresponding pixels are hidden (due to occlusion) in the other image, and disparity values point to the foreground object which is likely to be of a completely different texture and colour. If such a disparity map is used in a straightforward manner, foreground objects appear shifted to one side around the edges (see Figure 5.3). To overcome this, disparity values calculated by the SAD algorithm are used in the occluded regions which are identified by COGC.

The modified algorithm requires three parameters. The window size for occlusion correction performed by SAD is set to one in order to find the closes colour value disregarding the neighbouring pixels. For the graph cut part of the algorithm, the order of labels is randomised every iteration. The parameter λ controls smoothness and is calculated automatically using heuristics for estimating noise in images, as described by Kolmogorov (2004).



Scene 1 - Left - TM

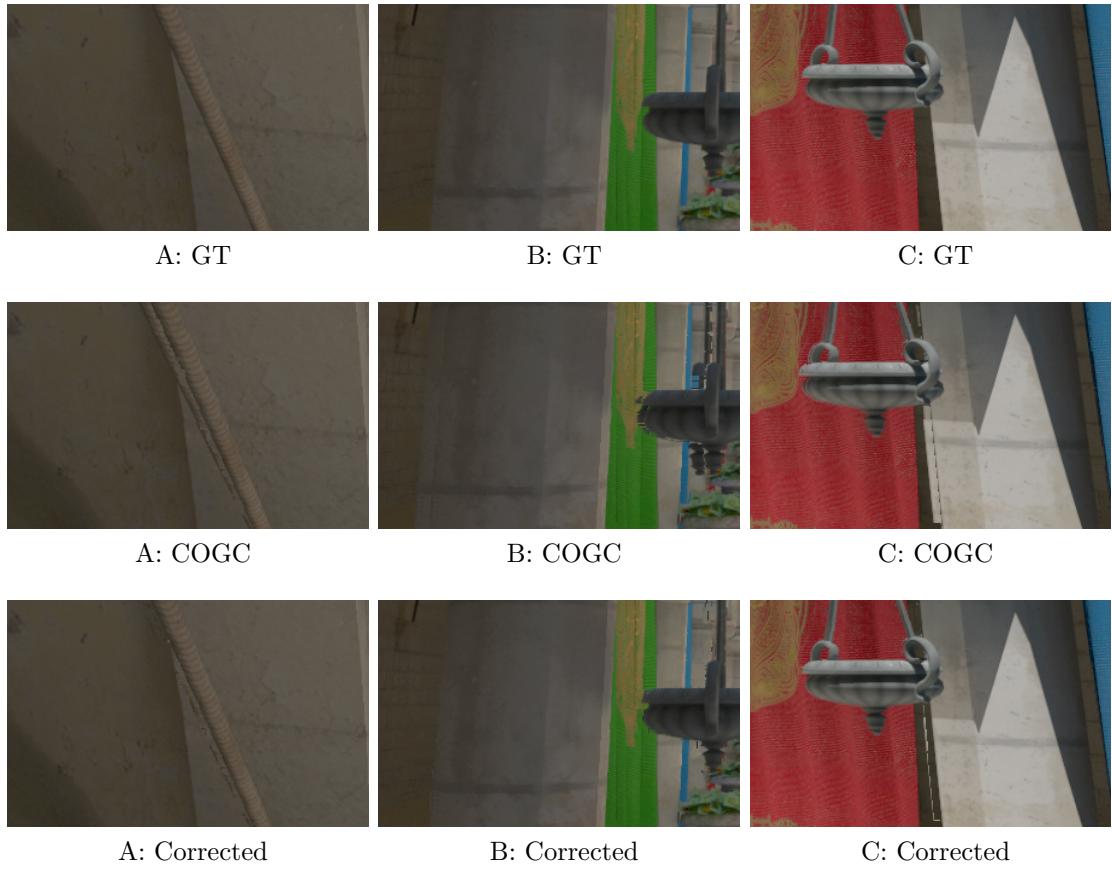


Figure 5.3: Disparities in the occluded regions generated by COGCC algorithm are not suitable for transferring HDR data. SAD disparities are used for such regions. The images are tone mapped.

Overexposed and underexposed regions benefit from the smoothness constraint imposed by the COGC algorithm. Here, disparities vary slowly and take values from the boundary of the out-of-range region. This is expected to generate good results unless out-of-range regions contain objects that are at different depths. COGC is still prone to errors and when they happen they tend to affect larger regions. Also, the function minimisation is time consuming and tens of minutes are needed to compute graph cuts for a full high definition image on a modern standard PC.

5.2 The Experiment

The purpose of this experiment is to identify a technique, from the four mentioned in the previous section, that is most appropriate to enable SHDR from an HDR and LDR pair and compare such methods with a fully captured SHDR pair. Methods available to conduct this study included rating, ranking and paired comparisons.

Even though rating might potentially answer how close to the GT each of the methods is, it could introduce problems as well. A very large number of trials and previously trained participants are needed to get credible data.

Ranking requires less participants. Compared to rating it provides more general but less precise results. Ranking images might prove challenging for the participants, and giving them a vague task, which is the case here, might affect their judgement.

The use of the paired comparisons method, on the other hand, simplifies the process by providing a straightforward choice between only two images at a time. This allows for small differences between compared objects to show. Additional statistics that reveal data consistency both within-participant and between-participant are available.

This method has also been successfully applied, in a manner similar to the one presented here, by Ledda *et al.* (2005), Banterle *et al.* (2009), Rubinstein *et al.* (2010) and Navarro *et al.* (2011). Ledda *et al.* (2005) examined which tone-mappers generated an LDR image that was perceptually closest to the original HDR image. Banterle *et al.* (2009) used the method to compare which of the five Expansion Operators (EOs) produced an image closest to the native HDR one. Rubinstein *et al.* (2010) applied the method of paired comparisons in a user

study comparing image retargeting operators. Navarro *et al.* (2011) explored how high level properties of rendering and different rendering parameters influence perceptual quality of motion blur in CG images.

5.2.1 Methodology

A forced choice paired comparisons method was used to present all of the HDR from LDR techniques in pairs, applied to a number of different scenes, to all the participants. The task was to choose one of the techniques that looks more similar to the GT. The chosen technique was said to be preferred.

More specifically, the *balanced paired comparison* method, which required every participant to perform every possible paired comparison, was used. For t techniques, n participants, and s scenes the total number of comparisons is $ns(t(t - 1)/2)$

In the experiment there were five different scenes and five methods. Besides the methods described in Section 5.1, GT was also compared. This work aimed to investigate if there is significant difference between GT and the proposed methods. With 26 participants doing 50 comparisons each, the total number of comparisons was 1300.

The choices for each participant and for each of the scenes were recorded using a two way preference table (e.g. Table 5.1). If *Technique 1* is preferred to *Technique 3* (written $T1 \rightarrow T3$) the value one is recorded in the row T1 and column T3 of the preference table and zero in the row T3 and column T1. In the example table, T1 is preferred to T3 and T5 but not T2 and T4. The table's principal diagonal is left empty as a method cannot be preferred to itself. Entries below the diagonal are redundant but are recorded nevertheless. The last column of the table gives the score of a technique (denoted a_i) which measures how many times a method has been preferred. The total score in the preference table then is:

$$\sum_{i=1}^t a_i = \frac{t(t - 1)}{2} \quad (5.1)$$

A preference table allows for the calculation of two measurements namely: coefficient of consistence and coefficient of agreement. If these two coefficients are significantly high it is possible to proceed and perform a test of equality and a range test, as explained below.

Table 5.1: Example preference table

Technique	T1	T2	T3	T4	T5	Score
T1	-	0	1	0	1	2
T2	1	-	1	0	1	3
T3	0	0	-	0	0	0
T4	1	1	1	-	1	4
T5	0	0	1	0	-	1

Coefficient of Consistence (ζ)

When comparing three techniques, there are eight possible outcomes. Six outcomes have one of the techniques scoring two wins, another scoring one, and the last having none. In two cases, however, each technique scores one win (e.g. $T1 \rightarrow T2$, $T2 \rightarrow T3$ and $T3 \rightarrow T1$). These are called *circular triads* and they express an inconsistency of the participant, possibly caused by a small difference between very similar methods which requires a guess. The number of circular triads for any number of techniques can be calculated using Equation (5.2), as proposed by Kendall & Smith (1940).

$$c = \frac{t}{24}(t^2 - 1) - \frac{1}{2}T \quad (5.2)$$

where $T = \sum(a_i - \bar{a})^2$, and $\bar{a} = \sum \frac{a_i}{t} = \frac{1}{2}(t - 1)$.

After computing the number of circular triads it is possible to obtain *coefficient of consistence* ζ which for an odd number of methods is defined by Kendall & Smith (1940) as:

$$\zeta = 1 - \frac{24c}{t(t^2 - 1)} \quad (5.3)$$

The coefficient ζ ranges from zero to one. A value of one means that no circular triads are present and preferences can be expressed as rankings, and as this value approaches zero the number of triads increases and so do the inconsistencies.

Coefficient of Agreement (u)

It is possible to test if participants performing comparisons make the same choices between themselves, that is, to calculate the *coefficient of agreement* u . Again, a preference table is used, but this time entries signify how many participants preferred each of the methods. If all the participants completely agree, half of

the entries have value n (number of the participants) while the other half are zero.

The sum of agreements between pairs of participants Σ defined in Equation (5.4) is calculated first.

$$\Sigma = \sum_{i \neq j} \binom{\alpha_{ij}}{2} \quad (5.4)$$

where α_{ij} is the number of times *Technique i* is preferred to *Technique j*. The summation extends over $t(t - 1)$ terms. Having found Σ , u is calculated as:

$$u = \frac{2\Sigma}{\binom{n}{2} \binom{t}{2}} - 1 \quad (5.5)$$

The maximum value u can take is one, which signifies all the participants completely agree. For complete disagreement u takes its minimum value of $-1/(n-1)$ for an even number of participants or $-1/n$ for an odd number, and all preference table entries are $\frac{1}{2}n$ or $\frac{1}{2}(n \pm 1)$ respectively.

Test of Equality and Range Test

The overall test of equality examines if the score differences are simply by chance or due to actual perceptual dissimilarities of methods. It verifies overall significance in data and assesses differences in preference scores (a_i) obtained by a specific method. This is similar in principle to the ANOVA test but specifically for use with ordinal data. Initially, a standardised sum of squares of the scores D_n using Equation (5.6) is calculated.

$$D_n = 4 \frac{\left[\sum_{i=1}^t a_i^2 - \frac{1}{4}tn^2(t-1)^2 \right]}{nt} \quad (5.6)$$

For a detailed derivation of Equation (5.6) see the book by David (1988). The null hypothesis $H_0 : \pi_i = \frac{1}{2}$, where π_i is average preference probability for method i , can be rejected if D_n exceeds or equals the critical value which is obtained from χ^2 tables using the desired significance level and $t - 1$ degrees of freedom.

The test of equality shows if there are statistical differences between methods but it cannot tell where these differences, if present, lie. To this end, the multiple comparison range test is used. It determines the significance of score differences between compared methods. This is the analogue of a post-hoc comparison test

such as Tukey's test used in ANOVA. Any pairwise difference in scores which exceeds or equals R can be declared significant. R is calculated using Equation (5.7).

$$R = \frac{1}{2}W_{t,\alpha}\sqrt{nt} + \frac{1}{4} \quad (5.7)$$

where $W_{t,\alpha}$ is the upper α significance point of the W_t distribution of a variance-normalised range.

5.2.2 Participants

The number of tested participants was 26 (21 males and 5 females) with an age range between 20 and 52 (mean 31). All the participants were volunteers. They all had normal or corrected to normal vision and were able to perceive stereoscopy.

5.2.3 Materials

Five different HDR scenes in *high definition* resolution (1920×1080 pixels) were used for the experiment (see Figure 5.4). The first was computer generated using a physically-based renderer. Images were rendered using path tracing. The other four were captured using a *Canon 1Ds Mark II* camera and the multiple exposures technique (Debevec & Malik, 1997). For each scene and each eye position, seven exposures, separated by two f-stops, were taken and combined. These images were used as the GT. In addition, for each of the scenes, four other SHDR images were generated using the methods described in Section 5.1. The middle exposure image was used as the input for the LDR image.

Capture was performed in a way which would avoid or minimise stereoscopic impairments. Keystone distortion and plane curvature were avoided by keeping cameras parallel. The cardboard effect was prevented by using a short lens (35 mm). The horizontal offset of cameras was set to 63 mm (mean interpupillary distance (Dodgson, 2004)) thereby minimising the puppet theatre effect.

Images were presented using two custom-made 46" HD HDR displays based on the Dolby DR-37P HDR displays, with a luminance range of $0.15cd/m^2$ to $3,000cd/m^2$. They were coplanar and separated by 4cm. The displays were calibrated using procedures suggested by Ruppertsberg *et al.* (2007).

A custom-built stereo rig with four highly reflective mirrors transferred images from the screens to the participants. The schematic and the real experiment setup are shown in Figure 5.5. The rig was positioned 1.6m from the screens and was

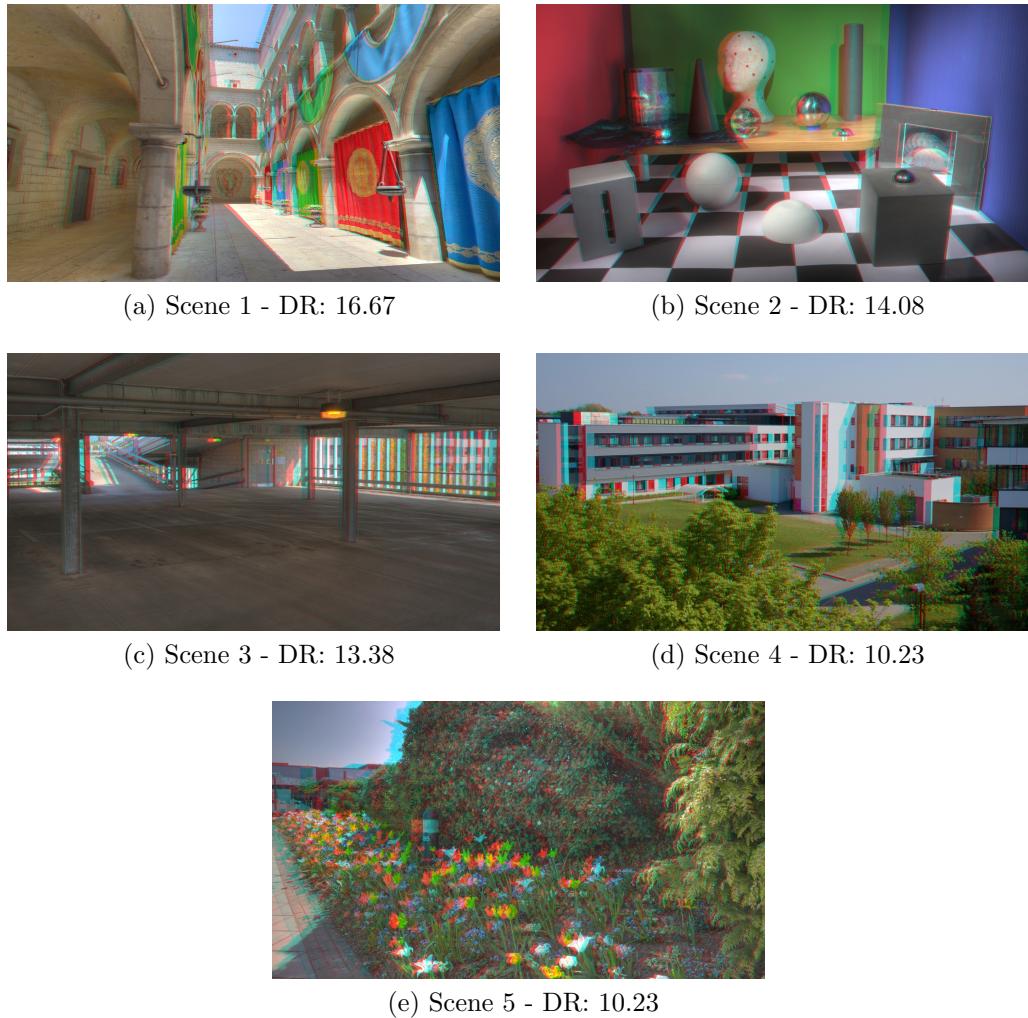


Figure 5.4: SHDR scenes (tone-mapped anaglyph) with corresponding dynamic range in stops displayed below.

centred both horizontally and vertically. Fine adjustments in image alignment were made by controlling the two front mirrors. The rig allowed stereo viewing at full resolution without any cross talk as each eye got a distinct view from one of the screens. Moreover, the rig included a head rest which immobilised head movement thereby avoiding the sheer distortion impairment, which could occur during stereo viewing.

Participants used a Microsoft XBox controller as the input device. Feedback information was provided using 2.1 stereo speakers and controller vibration.

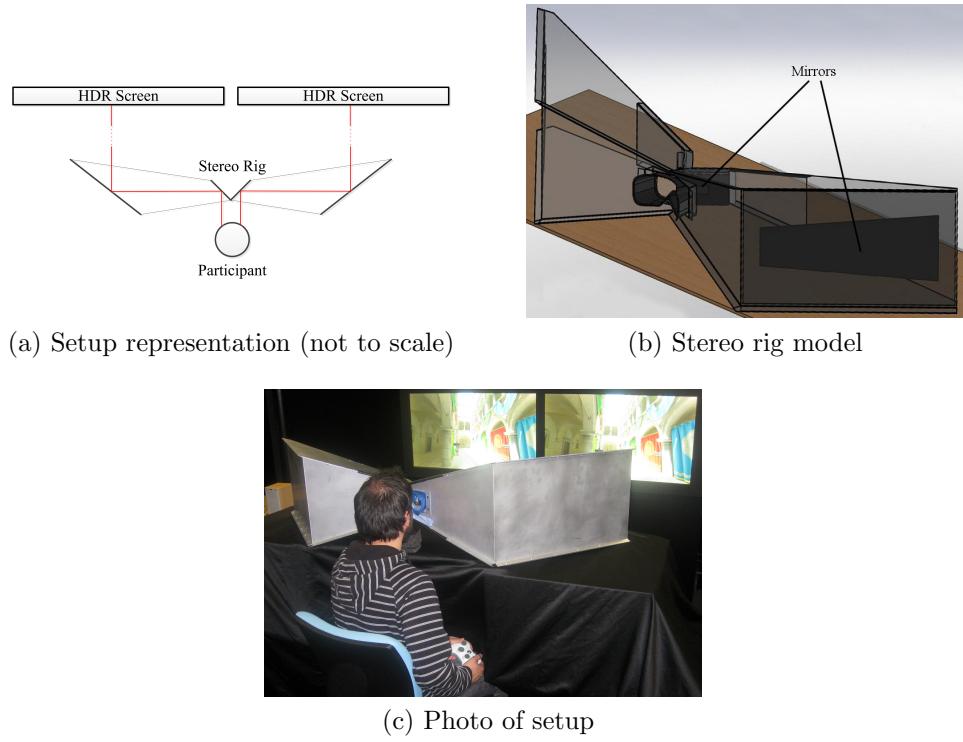


Figure 5.5: Experimental setup

5.2.4 Procedure

Each participant was presented with fifty randomised image sets, as five methods and five scenes were tested. All image sets contained three images: two random techniques to be compared (named A and B) and the GT. Participants were instructed to decide which of the two, A or B, appeared more similar to GT. Images were displayed for five seconds each in the following sequence: A, GT and B. During this period participants were only able to observe the displayed images ensuring that each image was seen at least once. Subsequently, the controller vibrated informing the participant that switching freely between any of the three images and selecting either image A or B was possible. Side by side comparison was not used as perspective would change the perception of the stimuli. Revisiting any of the image pairs was permissible giving the participant the opportunity of providing a better informed decision. Fifteen seconds were allocated for this part and after the time was up a short sound was played meaning that selection had to be made even if it required guessing. This would allow enough time for observing the images while still keeping the experiment length relatively short avoiding participant fatigue. Every time an image was displayed a sound notified

which of the image pairs A, B or GT was presented. Sound was also used to inform the participant that the next set had started. A uniformly grey colour was displayed for 0.3 seconds prior to each image pair for eye desensitisation. The screen showing the generated HDR image of the pair was chosen at random.

The experiment was conducted in a dark room to minimise ambient light. Participants were allowed five minutes to adjust to the environment. Initially, a training set was shown to the participants such that they could familiarise themselves with the task. The set consisted of ten images, unrelated to the sets used in the actual experiment. After a one minute pause, the experiment started and lasted between twenty and thirty minutes depending on how quickly participants made their choices.

5.3 Results

Results of the experiment are presented below. In addition, objective measurements were used to evaluate the quality of the generated images. It was also possible to test how objective measures correlated to the results obtained in the user study. Finally, the output of a single image produced by each method is presented to illustrate the difference between them. Regions deemed of special interest are zoomed in.

5.3.1 User Study Results

The results of the user study are shown in Table 5.2. The mean coefficient of consistence (ζ) values were high and statistically significant for the given degrees of freedom. This indicated that participants understood the task, that the difference between some of the methods was big enough to be detected, and that results were reliable.

Preference tables for each scene and all the participants were generated. For a better visualisation this data is represented as a graph in Figure 5.6. Each bar shows how many times a corresponding method was preferred. Aggregated data for all the scenes (labelled *Total*) is also presented.

The coefficient of agreement u was used to test the null hypothesis which states that *selections were made at random*. The large sample approximation to the sampling distribution (χ^2) was used to determine significance of u (see Table 5.2). Details of this test statistic are described by David (1988). At the

Table 5.2: Experiment results. Methods within the same circle cannot be considered perceptually different.

	mean ζ	u	χ^2	sig. u	D_n	sig. D_n	1st	2nd	3rd	4th	5th
Scene 1	0.769	0.279	79.692	p < .05	78.215	p < .05	GT	SAD	COGC	EM	LS
Scene 2	0.846	0.596	158.923	p < .05	141.107	p < .05	GT	COGC	SAD	EM	LS
Scene 3	0.938	0.602	160.615	p < .05	145.723	p < .05	GT	COGC	SAD	EM	LS
Scene 4	0.885	0.513	138.154	p < .05	131.815	p < .05	GT	SAD	COGC	EM	LS
Scene 5	0.931	0.562	150.462	p < .05	147.630	p < .05	SAD	COGC	GT	EM	LS
Average	0.874	0.510	137.569	p < .05	128.898	p < .05	GT	SAD	COGC	EM	LS

standard $\alpha = 0.05$ level and for $\binom{t}{2} = 10$ degrees of freedom, the null hypothesis was rejected and it was concluded that there is agreement between participants.

The sum of squares of the scores, D_n , was used for the overall test of equality. D_n was compared to a critical value of 9.45 (for $\alpha = 0.05$ significance level and $t = 5$ different methods). D_n exceeded the critical value for all the scenes causing rejection of the null hypothesis $H_0 : \pi_i = \frac{1}{2}$. This implied that there is a statistical difference between some of the methods.

The multiple comparison range test was used to determine which methods were statistically different from the others. The range, R , was calculated using Equation (5.7). For $W_{t,\alpha} = 3.86$ (from Table 22 by Peason & Haetlet (1976)), $\alpha = 0.05$ and $t = 5$ value of range R was 23. To calculate R for the aggregated result, the total number of comparisons between two methods had to be accounted for. The number of participants n in Equation (5.7) was multiplied by the number of scenes s and to give $R = 50$. R is represented in Figure 5.6 with the black range bars. The preference score of each method was compared with scores of all the other methods. If the difference was greater than R it was declared significant. Visually, in Table 5.2 methods which were not statistically different from each other were circled to represent a group. If a method is not grouped it means it is significantly different by itself. The final five columns of Table 5.2 represent

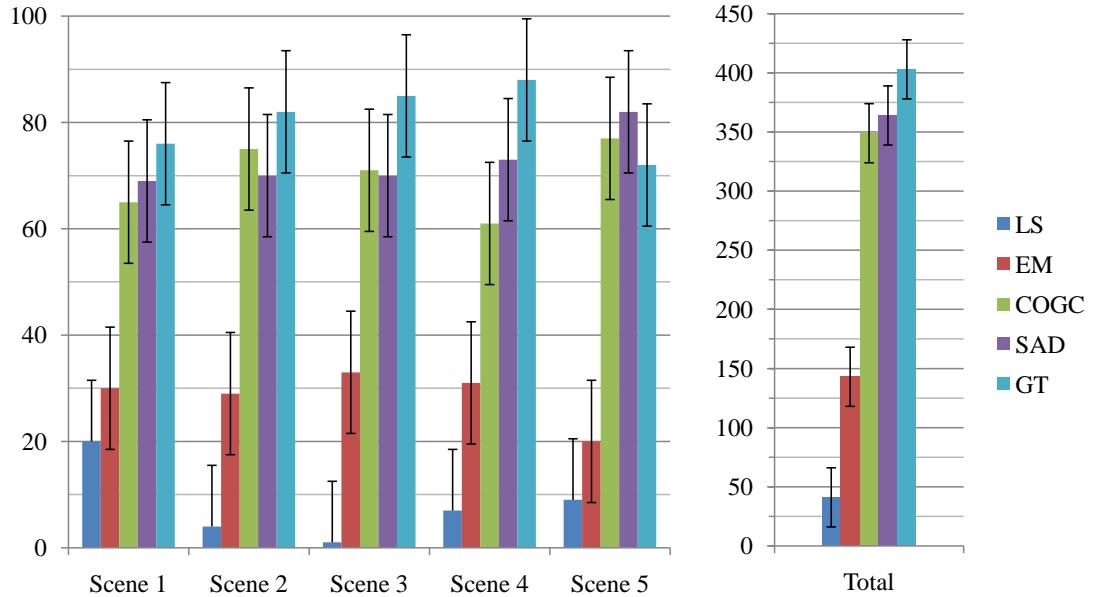


Figure 5.6: Method preference. Black bars represent range R from Eq. (5.7). The value of R is 23 for single scenes and 50 for total.

method rankings.

5.3.2 Objective Measures

In addition to perceptual evaluation of methods, two objective metrics evaluated the quality of the generated images: Peak Signal-to-Noise Ratio (PSNR) and *Root Mean Square Error* measure on logarithmically scaled HDR values (RMSEL). Values for RMSEL were logarithmically scaled in order to avoid biasing the result towards high intensity while PSNR already includes such a scaling.

PSNR is calculated as:

$$PSNR = 20 \log(MAX_I) - 10 \log \left(\frac{1}{n} \sum_{i=1}^n I_1(i) - I_2(i) \right) \quad (5.8)$$

where MAX_I is the maximum possible pixel value in the image, n is the total number of HDR values (including all three channels) and $I_k(i)$ is the i -th value of k -th image. The results are presented in Table 5.3 where higher measurements represent a better quality.

Table 5.3: Peak Signal-to-Noise Ratio Metric

Operator	SAD	COGC	EM	LS
Scene 1	46.21	44.73	17.76	28.49
Scene 2	43.52	48.08	14.66	10.99
Scene 3	38.50	40.98	16.51	16.42
Scene 4	43.63	43.03	19.07	11.91
Scene 5	41.95	41.41	10.86	14.48

RMSEL is calculated as follows:

$$RMSEL = \sqrt{\frac{\sum_{i=1}^n \log(I_1(i) + 1) - \log(I_2(i) + 1)}{n}} \quad (5.9)$$

where n is the total number of HDR values (including all three channels) and $I_k(i)$ is the i -th value of k -th image. The values are incremented by one to avoid negative results which occur in the range between 0 and 1. The results are presented in Table 5.4 where higher measurements represent a larger error.

Table 5.4: Root Mean Square Error of log HDR values as in Eq. 5.9

Operator	SAD	COGC	EM	LS
Scene 1	0.0021	0.0025	0.0324	0.0130
Scene 2	0.0057	0.0045	0.1457	0.3004
Scene 3	0.0183	0.0142	0.1401	0.1859
Scene 4	0.0076	0.0081	0.0713	0.1947
Scene 5	0.0024	0.0027	0.0705	0.0520

5.3.3 Image Quality Comparison

Figure 5.7 shows a comparison amongst all four methods for the reconstructed image in the stereo pair. This scene has many challenging regions including dark and bright regions, reflections and refractions. It can be noted how the stereo correspondence operators are superior to the chosen EO_s for this scene. Figure 5.8 shows further details for Scene 2 for the two correspondence methods and GT. Individual and appropriate single exposures are selected for illustrative purposes for each inset as some of them lie in very dark regions. Further analysis of this scene is provided in the next section.

5.4 Discussion

Unsurprisingly, GT was ranked in the first group for all the scenes and considered perceptually similar to itself. The SAD method was also continually found in the first group with GT which was reflected in the aggregated result. This implies that for all the scenes SAD is statistically indistinguishable from the ground truth and could, in theory, be used instead; hence reducing the cost of SHDR capture.

COGC method was in the first group for four scenes, but in *Scene 4* it was ranked third and was not in the same group as the GT. This result was reflected in PSNR and RMSEL as well. The aggregated result was affected and COGC was regarded perceptually different to GT but considered similar to the SAD method. Although this might seem counterintuitive, it means that any perceptual differences between GT and SAD were too small to be detected as were the differences between SAD and COGC. However, the difference between COGC and GT was substantial enough to be detected. Even if COGC did manage to get in the first group overall, usage of the SAD method might be preferred due to the high computational cost of COGC.

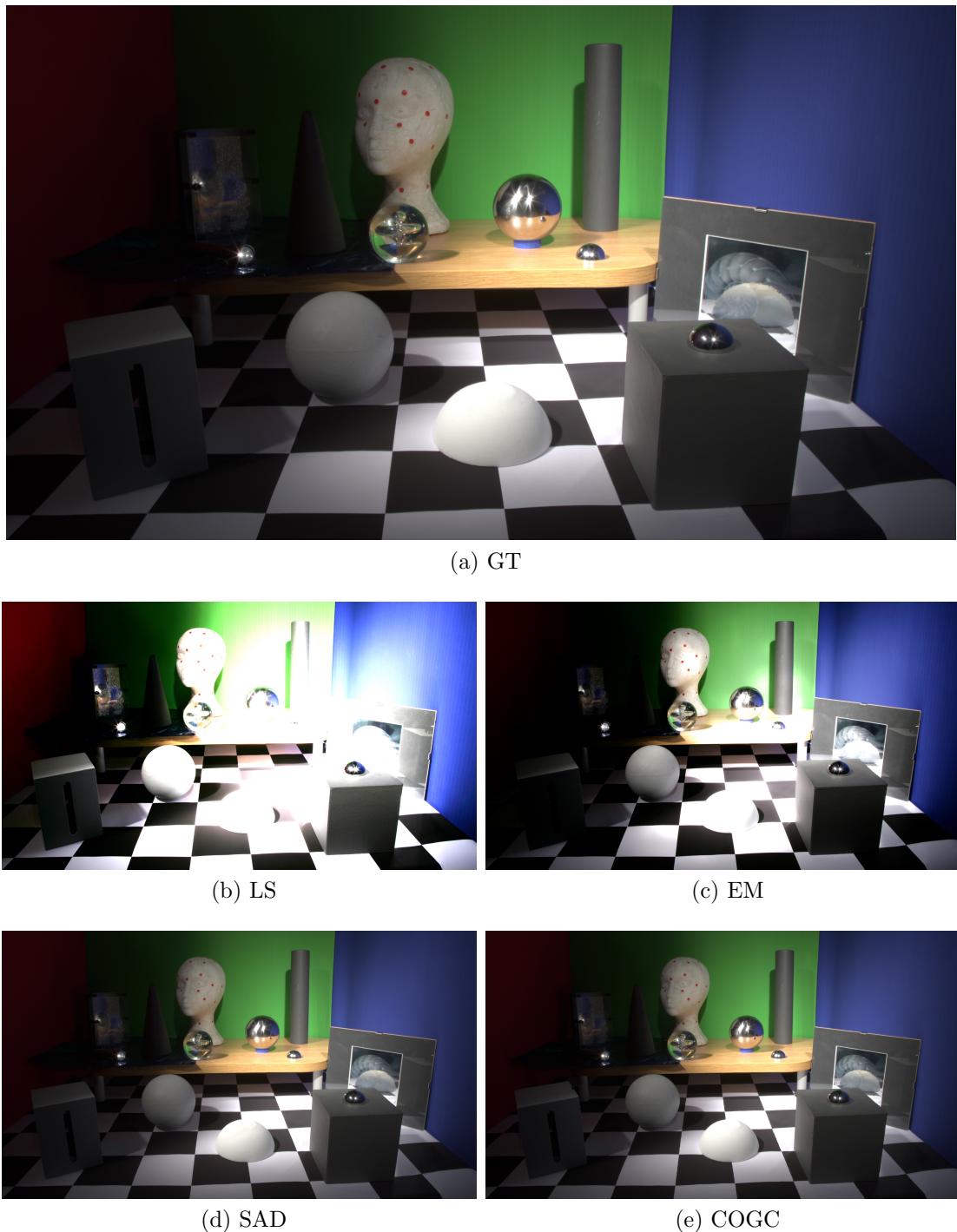
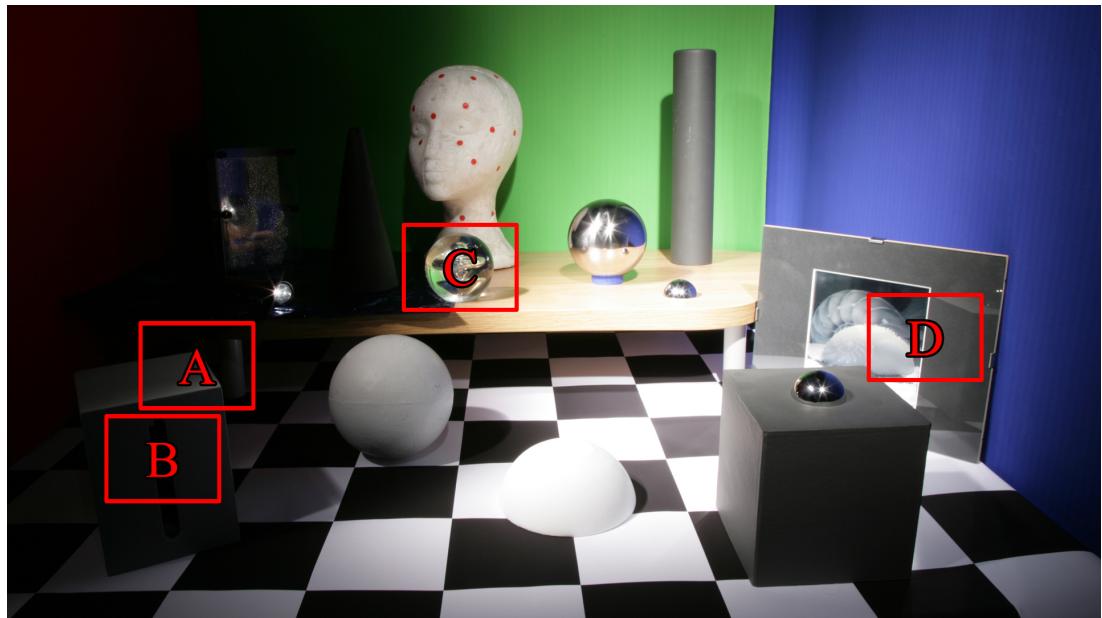


Figure 5.7: Reconstructed image from the SHDR pair for all methods for Scene 2, shown at the same single exposure level



Scene 2 - Single Exposure



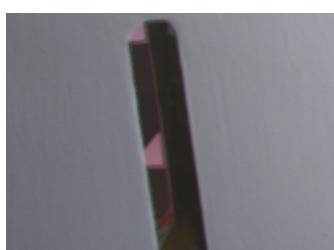
A: GT



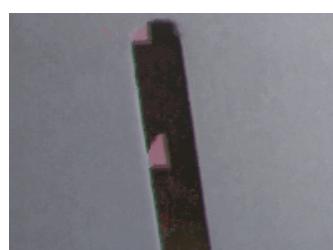
A: SAD



A: COGC



B: GT



B: SAD



B: COGC

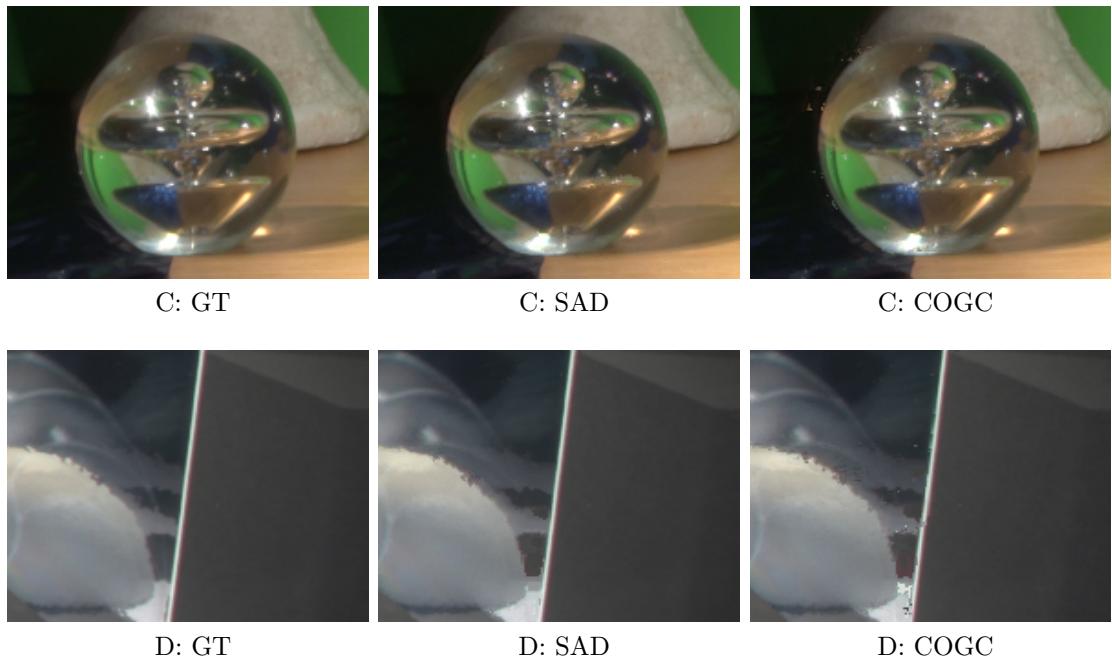


Figure 5.8: Detailed insets for the reconstructed SHDR chosen from Scene 2 showing GT, SAD and COGC. All images are shown at a single exposure.

As expected the correspondence maps generated by the SAD method contained more noise compared to COGC. While this resulted in more pixels being transferred from an incorrect position using the SAD method, it did not affect the results. It is important to note that while, COGC creates better disparity maps, SAD will find a pixel with similar RGB values. The values chosen by SAD, although less robust for traditional disparity calculations, may appear less distracting.

Both of the expansion operator based methods performed less well with considerable differences when compared to the stereo correspondence methods. The LS method was ranked last throughout the experiment making it the last choice for LDR enhancement. A possible reason for poor performance of EO_s might have been expanded brightness which did not correspond well enough to the actual brightness of the other (HDR) image. In addition there was no step which would reconstruct detail that was lost in overexposed and underexposed areas of the image.

The objective measures were consistent with the results of the perceptual experiment. For stereo correspondence methods objective scores coincided with the ranking for all scenes. In the case of EO_s, LS had better quality than EM

twice. However, in both cases (*Scene 1* and *Scene 5*) EO_s were in the same ranking group. This suggests that objective measures could be used to predict rankings for new scenes and new methods.

5.4.1 Limitations

The number of scenes tested in the study was limited by the need to maintain a reasonable amount of time for the participants' involvement with the study. To improve the general validity, future work will test more scenes. However, the scenes used in the experiment where chosen as a representative sample of those which could be encountered in everyday situations. In addition, some of the scenes had specific properties which were challenging for the proposed methods. For example Scene 2, see Figure 5.8, included transparencies, object reflections, low-frequency and high frequency regions, specular highlights, under-exposed and overexposed regions, and Scene 1 included a large overexposed region. These areas tested the limit of the stereo correspondence approach as here the matches were not guaranteed and even if present they could not be considered reliable. However, in practice the algorithms performed well.

For example, in the case of transparencies it is uncertain which depth value should be assigned (foreground or background object). The correctness of a depth map is not the primary concern for the generation of the new HDR view, as long as it provides a good correlation for the data to be transferred between the images (Figure 5.8, inset *C* and *D*). Reliable correlation is achieved using both algorithms. It is inherent for SAD which looked for the closeness in intensity value, and for COGC it is imposed using the data energy minimisation. This strong data correlations handles other, already mentioned, challenging areas, including depth of field, relatively well.

Another parameter which was limited by the experiment's length, was the number of evaluated operators. In this case, a balanced solution was sought by selecting straightforward and advanced operators from each category. It is possible that potential future operators outperform the ones suggested here. A local operator worth exploring could be a more advanced one suggested by Mei *et al.* (2011). It expands on the SAD technique by using an additional cost measure, dynamic regions for aggregation, and error correction. An interesting global operator was proposed by Lang *et al.* (2010); their hybrid approach finds initial, sparse and robust matches which are then used as a support for an optical

flow algorithm ultimately yielding dense correspondences. It is unlikely, however, that these current methods will offer a significant improvement over other global methods as the fundamental nature is not too dissimilar.

5.5 Summary

Contributions made in this chapter include four methods for generating SHDR images from an HDR-LDR image pair. Two techniques were based on expansion operators while two relied on stereo correspondence to produce a novel HDR view, avoiding the need for two rare and expensive HDR cameras. In addition, the quality of the generated images and viability of the approach was tested in a user study using the method of balanced pair comparisons. Five scenes and four operators were compared to GT using a custom built stereo rig. The stereo correspondence techniques outperformed the expansion operator methods. The SAD technique was deemed perceptually indistinguishable from the ground truth for all the scenes which confirmed the viability of the approach. This result and other insights obtained in this chapter guided development of the video algorithms which are presented in the following chapter.

CHAPTER 6

Stereoscopic High Dynamic Range Video

Chapter 5 validated the feasibility of generating static SHDR images from the HDR-LDR pairs in a user study. Four operators, two based on stereo correspondence methods and two based on dynamic range expansion, were proposed and compared to GT images where one was found statistically indistinguishable from GT. This chapter proposes generating SHDR video from an HDR-LDR video pair. Insights gained from the process of designing static operators and the obtained results guided the development of new video specific operators.

The problem is related to static images with the added challenge of temporal consistency. Consecutive frames in a video are expected to be similar with small changes in content (expect for scene changes). Algorithms that process or generate SHDR video should avoid introducing temporal artefacts which may be seen across consecutive frames, flickering for example. Therefore, methods for generating SHDR video should reinforce temporal coherence. Figure 6.1 outlines the concept.

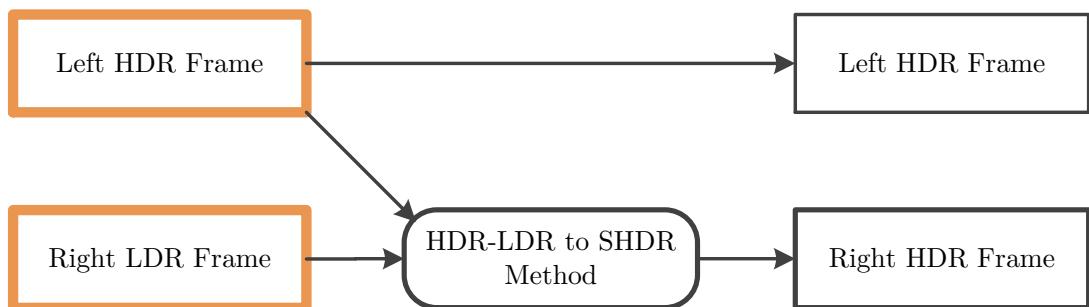


Figure 6.1: SHDR video is generated from an HDR-LDR video pair. The original HDR video is unmodified and represents one view while the second HDR view is obtained from the LDR video using the proposed methods.

The benefits of the HDR-LDR to SHDR video approach include reduced cost and complexity compared to an SHDR camera rig consisting of two native HDR video cameras. Similarly to the static image approach, having one LDR camera allows capturing one view in a standard LDR manner, to which users and movie directors are more accustomed. The LDR video stream may be used in a traditional pipeline or integrated into the SHDR one. This is important for the transition from traditional to new imaging.

The performance of the two suggested operators was compared to the technique which performed the best for static images and which was adapted for videos. Results showed that one of the two novel methods generated frames of improved spatial quality and was more temporally consistent.

Section 6.1 describes three methods used for generating SHDR video from an HDR-LDR video pair. These three methods were compared to ground truth SHDR videos in Section 6.2. Objective measurements tested temporal and spatial qualities of all the methods over five video sequences. A summary is provided in Section 6.4.

6.1 LDR to HDR Methods

Three methods to generate SHDR video from an HDR-LDR video pair are proposed. The most successful method for static images, based on using stereo correspondence, is extended to generate SHDR video, but it can suffer from flickering when extended into the temporal domain. In this chapter a further two methods are presented. One uses a novel expansion operator that extends the dynamic range of the LDR image based on the HDR data. The second approach combines the previous two methods and exploits the advantages of both. It uses stereo matching for overexposed and underexposed pixels and an EO for the remaining pixels.

6.1.1 Stereo Correspondence

The stereo correspondence approach relies on a disparity map to transfer data between the HDR and LDR view. For static images the best identified technique was *sum of absolute differences (SAD)*. It can be used in similar manner for video where it is applied on stereo frame pairs.

A detailed pipeline for generating an HDR frame using the SAD method is

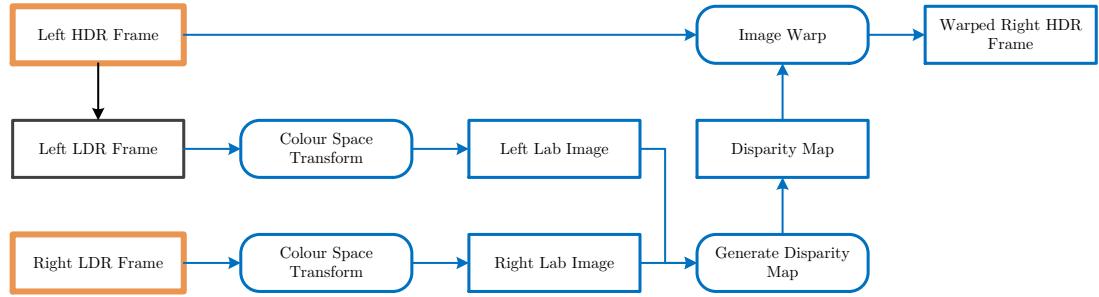


Figure 6.2: The HDR-LDR stereo correspondence pipeline finds spatial matches between HDR and LDR pixels and uses them to guide the warping of the existing HDR frame, thereby generating a novel HDR frame.

shown in Figure 6.2. Without any loss of generality the left frame is considered HDR and the right is considered LDR. First a single exposure is selected from the HDR frame to match that of the LDR one. This is achieved by minimising the difference between histograms of the existing and extracted LDR image. Values obtained in this step can be transferred to the next frames to speed up the process. Both LDR frames are then transformed to Lab colour space which approximates human vision and aspires to perceptual uniformity. This means that the differences between channels of the left and right frame are related to perceptual differences; such differences are more perceptually accurate than if RGB space was used. Next, the SAD algorithm is used to compute the disparity map between stereo frames. The disparity map then guides image warping of the available HDR image thereby generating a novel view.

The SAD stereo matching algorithm can compute the disparity map in real-time on a standard PC. The technique transfers actual HDR values to the new position in the other view and so avoids intensity quantisation. While the calculated disparity map can be noisy and incorrect offsets can be present, the algorithm always connects pixels which are close in intensity making it particularly efficient for the generation of the novel stereo view.

The overexposed and underexposed pixels are also transferred but can end up in the wrong position. As all the values in those regions have the same value (0 or 255) it is not possible to perform accurate matching. This is especially the case for larger regions where, even with increased window size, it may not be possible to find a pixel within the captured range. Disparity maps for such areas contain constant values. The first tested disparity value is selected by

the winner-takes-all (WTA) technique as all the others have the identical SAD cost. Details in these regions are present but may be out of place, and may be perceived as being at the incorrect depth (Figure 6.11, inset B). Another challenge is that of representing view dependent phenomena, such as occlusion (Figure 6.11, inset A), reflective objects, and specular highlights. Data for those might be missing from one of the views. However, SAD finds perceptually close intensities for those pixels (albeit from the spatially incorrect positions) which can alleviate the problem to an extent. Such mistakes were not perceived due to binocular fusion for static images, but they will cause temporal noise for videos as they may not be temporally consistent (Figures 6.9c and 6.9f). Thus, the main disadvantage of this approach is potential temporal incoherence due to incorrect disparity matches.

6.1.2 Expansion Operator

One of the state-of-the-art operators (Banterle *et al.*, 2006) inverse tone mapper that was evaluated as the best expansion operator in a user study (Banterle *et al.*, 2009) - did not perform well when converting HDR-LDR image pairs to SHDR images, as shown in previous chapter. When expanding the image, EOs take a small number of parameters (e.g. three in the case of the tested one Banterle *et al.* (2006)) which controls the overall brightness of the final image and its peak value.

While this is a convenient method of adjusting the output, the lack of control means that expanded images are less likely to correspond to the actual values of the imaged scene. For example, the peak luminance parameter influences the range and brightness of the expanded image, but such a value is frequently a result of noise (when the original HDR images were captured). Using existing HDR images as the means of setting this parameter is unlikely to produce appealing results so user input is required. Expansion operators are not very suitable for reconstruction of the LDR view for SHDR because discrepancies from the original can be large (especially for frames of higher dynamic range, e.g. above 10 stops), and not possible to fuse through binocular single vision.

However, for the HDR-LDR pair case, it is possible to create an EO by finding a mapping between the original HDR and LDR values by using the HDR as a reference. The problem is similar to the one faced by Mantiuk, Efremov, Myszkowski & Seidel (2006) where HDR video was compressed by using a residual

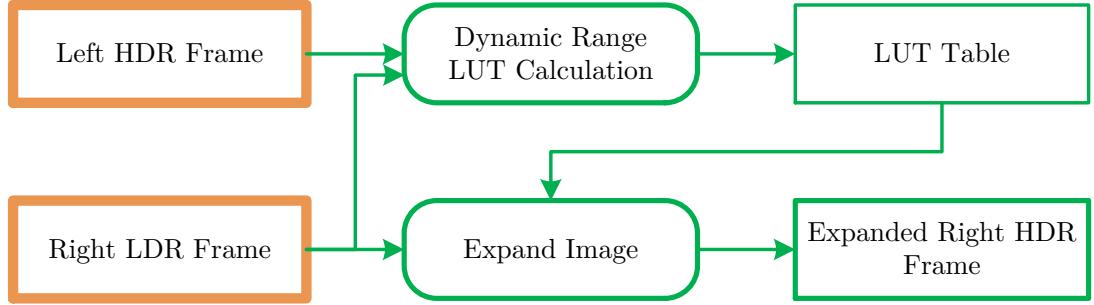


Figure 6.3: HDR-LDR expansion operator pipeline calculates intensity correspondences between LDR and HDR values and saves them in the look-up table. Expansion is performed by assigning HDR values from the table to the LDR pixels.

stream together with a tone mapped stream. It was assumed that the TMO was unknown so the correspondence between the HDR and the tone mapped values had to be calculated for the decoder. As the HDR and LDR pixels are spatially aligned, HDR values can be placed into 256 corresponding LDR bins. As this is a many-to-one mapping, multiple HDR values get assigned to a single bin. Mantiuk, Efremov, Myszkowski & Seidel (2006) used the arithmetic mean to find a single value. Their approach to generate an HDR image from an LDR image is extended and modified. The reconstruction function (RF) which maps LDR to HDR values is calculated as follows. All the HDR values are ordered. Then, an HDR histogram with 256 bins is created to emulate the LDR histogram, by putting the same number of HDR values into each bin as there are LDR values in that bin. Formally, this is expressed as:

$$RF(c) = \frac{1}{Card(\Omega_c)} \sum_{i=M(c)}^{M(c)+Card(\Omega_c)} c_{\text{hdr}}(i) \quad (6.1)$$

$$\text{where } \Omega_c = \{j = 1..N : c_{\text{ldr}}(j) = c\}$$

$c = 0..255$ is an index of a bin Ω_c , $Card(\cdot)$ is the cardinality function which returns the number of elements in the bin, N is the number of pixels in a frame, $c_{\text{ldr}}(j)$ are channel intensity values of the j -th LDR pixel, $M(c) = \sum_0^c Card(\Omega_c)$ is the number of pixels in the previous bins, and c_{hdr} are channel intensity values of all HDR pixels sorted in ascending order.

The pipeline to generate an HDR image using this approach is shown in Figure 6.3. The look-up table (LUT) is calculated using Equation 6.1. Once the LUT is

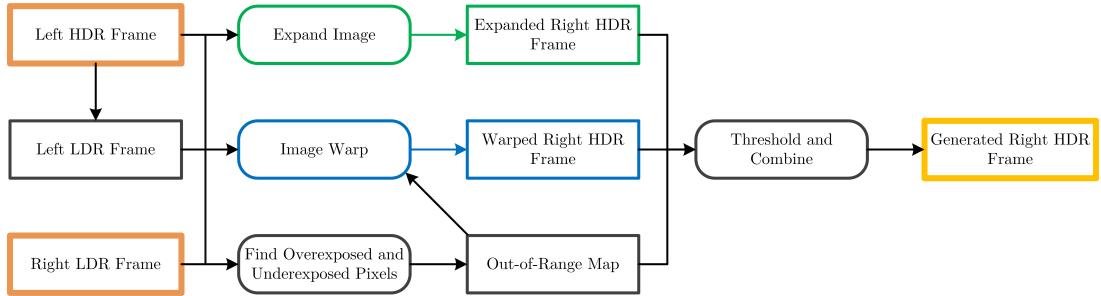


Figure 6.4: The broad pipeline showing generation of right HDR frame given a left HDR frame and a right LDR frame. Dynamic range expansion and stereo matching techniques are combined using an out-of-range map as a threshold.

obtained, expansion can be performed quickly where each LDR value is assigned a corresponding HDR value from the table. It is possible to re-use the LUT across frames and it can be used to improve temporal quality by filtering.

The proposed method of generating SHDR video from an HDR-LDR video stream using the proposed EO is quick and can be implemented in real time. Expansion is not view dependent and does not suffer from the same problems stereo matching would in occluded regions. Generated HDR values only depend on the captured HDR and LDR streams which are temporally coherent and the method is not expected to introduce flickering. The main drawback of this approach is the lack of a facility to explicitly handle overexposed regions which are of constant, maximum value without any detail. While binocular fusion can also help in those areas, differences are frequently high and noticeable (Figure 6.11, insets B, C and D).

6.1.3 Hybrid Method

The methods described above both have distinct sets of advantages and drawbacks. Hence, a novel method which combines the two, trying to obtain benefits of both while minimising disadvantages is proposed. Effectively, the hybrid method attempts to perform well in in-range regions using techniques based on the EO. For example occluded regions are handled more robustly. It also is able to handle out-of-range pixels using methods based on SAD, albeit with a further correction step.

An overview of the technique is displayed in Figure 6.4. Both an expanded HDR frame and a warped HDR frame are generated. Overexposed and underex-

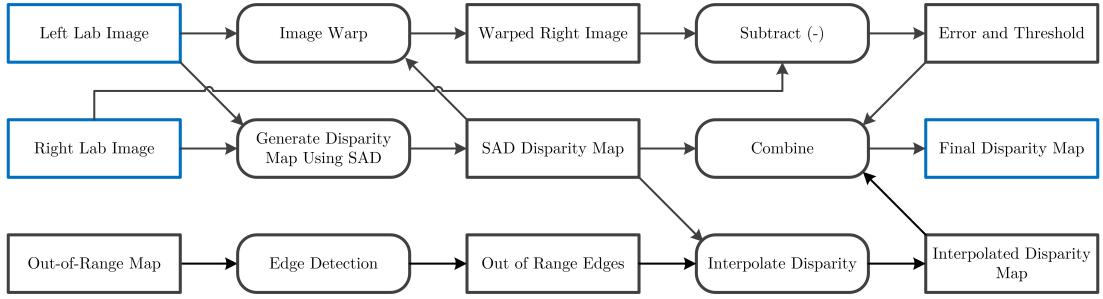


Figure 6.5: Modified disparity map generation interpolates disparities for the out-of-range region from its neighbours. Disparities which are likely to cause artefacts are replaced by the SAD generated disparities.

posed regions are identified using thresholding of the LDR frame, where a pixel is deemed out of range if the value of one channel is above or below a predefined threshold (e.g. above 250 or below 5). The out-of range pixels are assigned the warped frame data. The rest of the pixels are taken from the expanded HDR frame. The expanded frame is generated using the same approach described above while the SAD pipeline is adapted so it reconstructs out-of-range regions more accurately.

As described in the previous chapter, overexposed and underexposed regions lack any data and as such cannot be matched. The SAD method assigns the first tested disparity value to those regions, which for some cases may be correct. However, it is likely that out-of-range areas have disparities similar to the neighbouring ones. For this reason, the hybrid method interpolates disparities for overexposed and underexposed regions from well exposed edges. This modified stereo correspondence path is shown in Figure 6.5.

Images are transformed to the Lab colour space, and the stereo matching is performed using the traditional SAD approach. In addition, a map identifying out-of-range regions (generated by thresholding) is used as an input. Edges of overexposed and underexposed areas are found using an edge detection method. In the proposed implementation, image dilation is used to (spatially) expand overexposed and underexposed regions in the out-of-range map. Edges are found by subtracting the original out-or-range map form the dilated one. Once edge pixels are identified, smooth interpolation is performed inward. Figure 6.6 shows a comparison of a disparity map generated using the SAD approach and the SAD approach with interpolation.

During interpolation, foreground objects can influence disparity values of out-

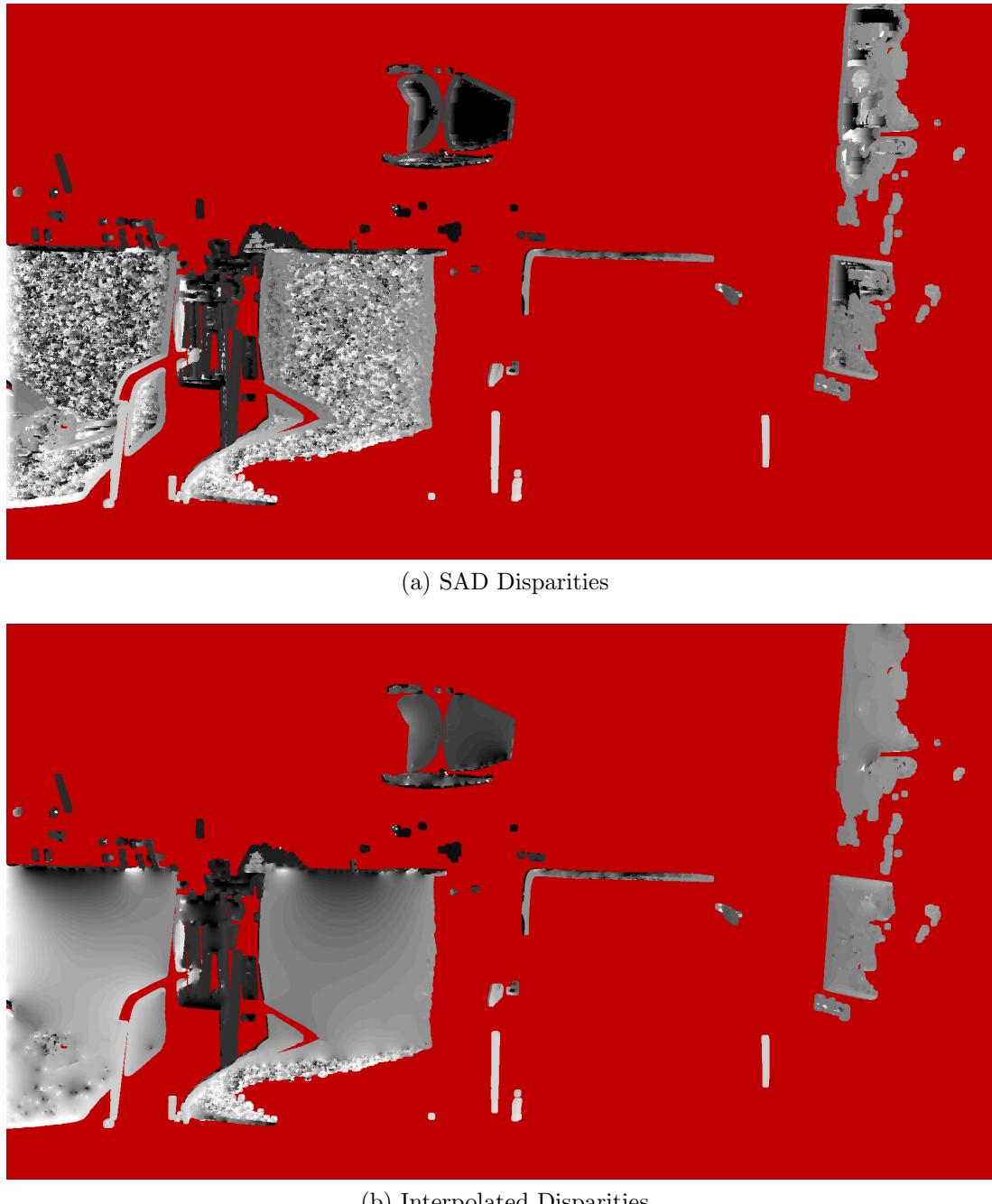
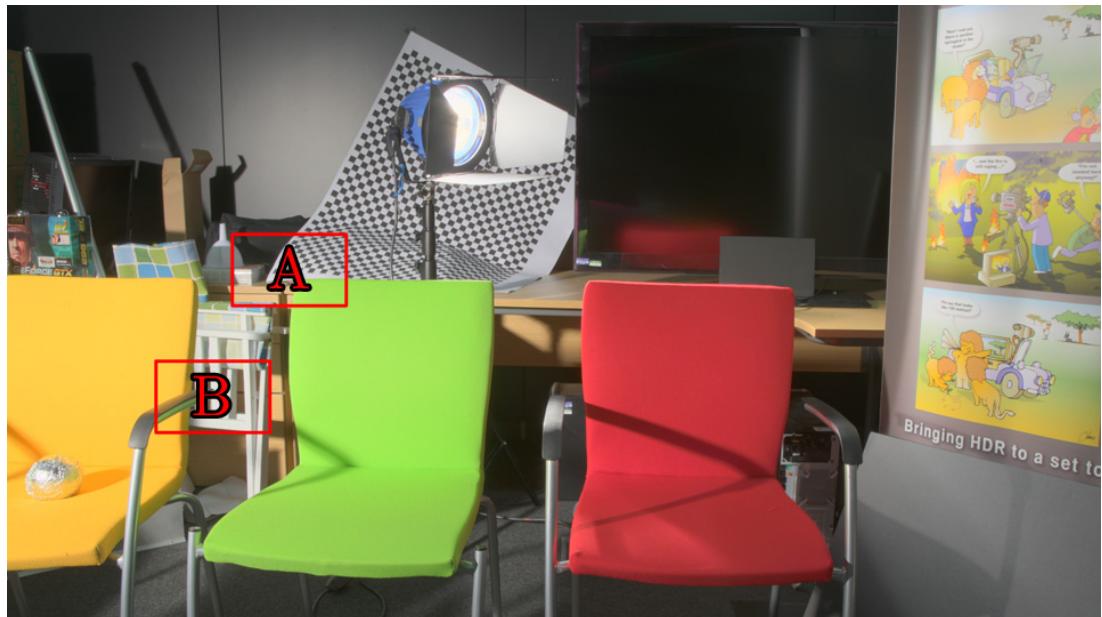


Figure 6.6: The disparities for overexposed regions generated using SAD method (a) are less smooth compared to ones obtained using interpolation (b).



(a) GT



(b) A: Artefacts



(c) B: Artefacts



(d) A: Corrected



(e) B: Corrected

Figure 6.7: Artefacts caused by interpolation are identified and corrected.

of-range background areas and vice-versa. This may result in artefacts around such objects as the values of foreground objects can be transferred to the background, as shown in Figures 6.7b and 6.7c. Such artefacts are identified by warping the extracted exposure of the HDR image using the interpolated disparity map and subtracting it from the original LDR image. Differences above the provided threshold are recognised as artefacts. In order to correct for these artefacts, pixel disparities computed by SAD are used instead. SAD matches are also potentially incorrect as they connect overexposed pixels. However the error in intensity will likely be smaller than by transferring well-exposed values from the foreground object. Results of this correcting step are shown in Figures 6.7d and 6.7e.

6.2 Results

In order to demonstrate the efficacy of the proposed methods, the methods are compared with each other and a ground truth SHDR video. The way in which the ground truth videos were obtained is explained next, after which the results of quality evaluation are provided.

6.2.1 Materials

Ground truth SHDR videos, consisting of HDR-HDR video pairs, had to be obtained in order to enable comparison with the proposed methods. As mentioned in the introduction, camera systems which record two native HDR videos simultaneously do not currently exist and are currently difficult to construct. In order to overcome that challenge, three techniques for capturing SHDR video data were employed.

Two static scenes (*Scene 1* and *Scene 2*) were recorded using stop motion by mounting a camera (Canon 1Ds Mark II) on rails and moving it laterally, in small steps (0.5 cm). At each step seven exposures separated by 2 stops were captured and later merged to produce individual SHDR video frames. As the movement was horizontal and orthogonal to the optical axis both views were obtained. The video for one eye was delayed by 13 frames compared to the other eye, which corresponded to a camera shift of 6.5 cm - an approximation of average interocular distance.

Scene 3 and *Scene 5* were dynamic and recorded using a native HDR video

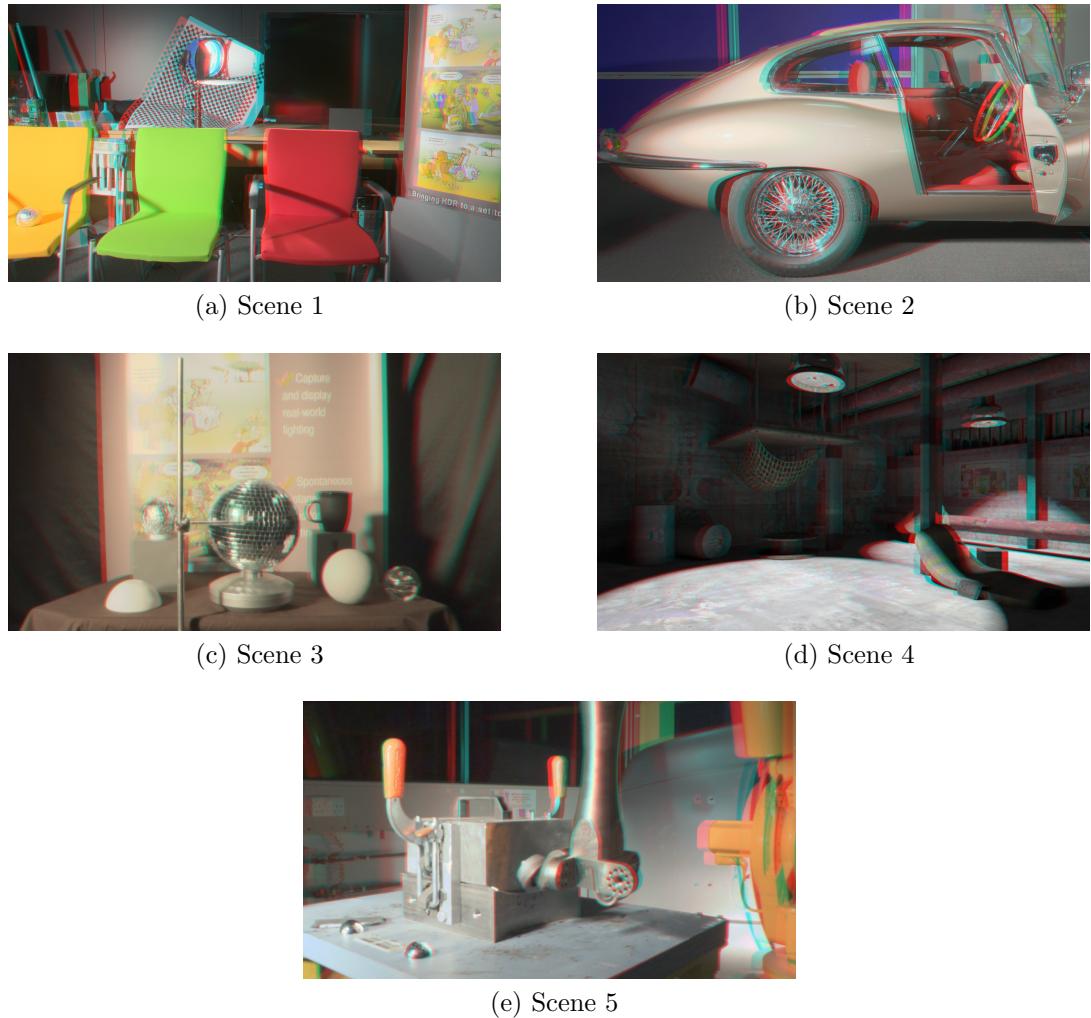


Figure 6.8: An example frame from each of the tested SHDR scenes. For illustrative purposes the frames are tone-mapped and displayed as anaglyph stereo.

camera Chalmers *et al.* (2009). Two takes, one for each eye, were required as only a single camera was available. Object movement in the scene needed to be exactly repeatable between the takes. To this end a high precision robot arm which folded aluminum sheets, and a disco ball that rotated were recorded.

The final scene was computer generated (*Scene 4*) using a virtual stereo camera rig that output HDR images. Tone mapped frames from each of the sequences are shown in Figure 6.8

All videos were captured and computed in full high definition (1920×1080 pixels). The two dynamic scenes were captured at 30 frames per second (fps), the computer generated scene was at 24 fps, while for static scenes 30 fps was chosen.

A robust measurement of the dynamic range was obtained by disregarding the top 1 % and bottom 1 % of the values in the frame; this was used to avoid extreme values caused by noise. It varied between the scenes and individual frames, peaking at 16.6 stops for *Scene 4*, and having a minimum of 11.1 stops for *Scene 5*. The length also varied between the sequences where *Scene 5* was the longest containing 720 frames, and *Scene 4* was shortest with 240 frames. Videos contained regions which would test the limits of the proposed methods including out-of-range areas, view-dependent phenomena, camera movement and object movement. Data for all sequences is summarised in Table 6.1

Table 6.1: Video data for each sequence: the average and maximum dynamic range (in stops) and the total number of frames

Data	Average DR	Maximum DR	Frame No
Scene 1	12.3	14.1	287
Scene 2	13.4	14.8	432
Scene 3	12.3	12.7	368
Scene 4	16.3	16.6	240
Scene 5	12.1	13.1	720

6.2.2 Objective Quality Measurements

Objective measurements were used to evaluate the quality of each method. To estimate the error of the individual frames, peak signal to noise ratio (PSNR) was used. It represents the ratio between the maximum possible value of an image (signal) and the power of noise which affects its quality. The measurement is logarithmically scaled making it especially suitable for images of high dynamic range, because the HVS system responds to the intensity of light approximately logarithmically (Weber, 1834; Fechner, 1838). In the previous chapter, it was also shown to correlate with subjective measures in the case of generating SHDR images from an HDR-LDR stereo pair. The averaged values for all the scenes and all the methods are shown in Table 6.2 where higher value represents better quality. Results for individual frames are presented in Figure 6.12.

As expected, the hybrid (HY) method outperformed the other two achieving the best score for all tested scenes. The SAD technique achieved better results than EO for all the scenes. The score difference was greater between HY and SAD than between SAD and EO.

Table 6.2: Peak Signal-to-Noise Ratio (higher is better)

Method	HY	SAD	EO
Scene 1	51.88	48.77	48.77
Scene 2	45.77	42.26	40.34
Scene 3	48.76	46.66	45.42
Scene 4	54.26	37.29	34.33
Scene 5	59.14	51.97	48.54
Average	51.96	45.39	43.48

In order to verify temporal quality, a temporal quality (TQ) metric is proposed, inspired largely by metrics such as RMSE and PSNR. Initially, images are converted to logarithmic space to account for perception of the HVS. Next, the metric finds differences between two consecutive frames of the ground truth and the generated streams. The differences are compared and weighted by the spatial quality of a frame pair. Finally, values are aggregated across all the pixels as shown in Equation 6.2:

$$TQ(t) = \frac{1}{Card(N)} \sum_{i \in N} \omega_i |(\Delta I_1(i, t) - \Delta I_2(i, t))| \quad (6.2)$$

where t is the frame number, N is the set consisting of all channel values for all pixels in a frame, I_1 is the ground truth image and I_2 is the generated frame, $\omega_i = |\log(I_1(i, t)) - \log(I_2(i, t))| + 1$ is the quality weight, $I(x, y, t)$ is the intensity of a pixel per colour channel at point (x, y) of the frame t and $\Delta I(x, y, t)$ is difference between logarithmically scaled consecutive frames as shown in Equation 6.3:

$$\Delta I(x, y, t) = \log(I(x, y, t) + 1) - \log(I(x, y, t + 1) + 1) \quad (6.3)$$

The summary of TQ values, averaged across the video sequence, are shown in Table 6.3 where the smaller value represents a better quality. Results for all the frames and all the videos are provided in Figure 6.13.

Overall, the HY method performed best and had the smallest error for all the scenes. SAD technique had better quality than the EO technique for four scenes while EO outperformed SAD for the last scene, which has the smallest average dynamic range. As discussed in Section 6.1.2, this is expected as the EO method should be, generally speaking, a preferable option to SAD for HDR videos with a lower dynamic range.

Table 6.3: Temporal Quality (lower is better)

Method	HY	SAD	EO
Scene 1	0.0091	0.0128	0.0156
Scene 2	0.0038	0.0063	0.0072
Scene 3	0.0043	0.0064	0.0074
Scene 4	0.0009	0.0010	0.0014
Scene 5	0.0063	0.0129	0.0110
Average	0.0049	0.0079	0.0085

6.2.3 Qualitative Results

To complement quantitative metrics, Figure 6.10 illustrates the different visual qualities of the methods. The provided example shows how EO does not manage to reconstruct any details in the out-of-range regions while SAD and the HY method appear similar, in general. To show the differences between the two, selected regions are presented in more detail in Figure 6.11. The inset C shows the lamp leg which is not reconstructed well by the SAD method, due to occlusion. The insets A and C contain occluded areas (along the edges of the chairs), where the SAD method made errors. Due to relying on the EO, the HY method is able to preserve the information available in the LDR frame. For the overexposed regions, shown in the insets A, B, and D, SAD lacks information required for accurate matches and makes mistakes. The HY method relies on interpolation to obtain disparities from the neighbouring well-exposed pixels and is able to reconstruct these regions successfully. The EO, as expected in this case, lacks the required information for reconstruction.

Disparities calculated by SAD methods are noisy in low frequency regions and in occluded areas resulting in artefacts when generating the HDR frame, as shown in Figures 6.9c and 6.9f. The HY method achieves temporal consistency by using EO for in-range pixels (Figures 6.9d and 6.9g).

6.3 Discussion

The HY method outperformed the SAD and EO methods for both single frames and video sequences resulting in the least amount of flickering. It produced the least amount of artefacts and recovered out-of-range regions well. As such, when quality of the reconstructed image is the main concern HY should be the method of choice. However, it is slowest to compute, because it performs expansion,

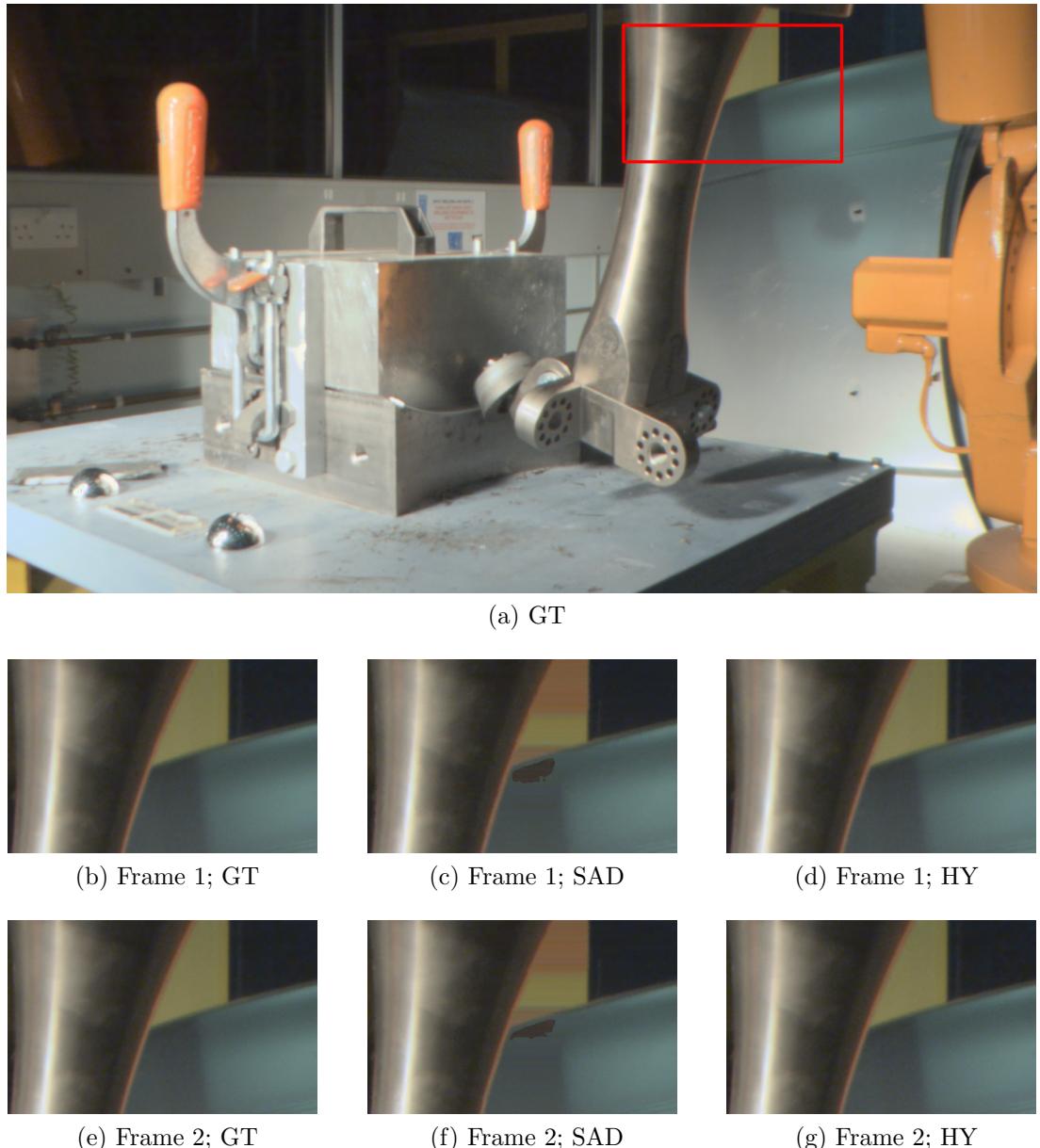


Figure 6.9: The SAD method may generate artefacts in the occluded regions which are inconsistent across frames. The HY method recovers these pixels using EO and avoids artefacts.

stereo matching, interpolation and correction steps.

The SAD method suffers from temporal flickering (especially in occluded regions) but might still be the method of choice when processing speed is crucial. The EO technique is the fastest but, due to poor performance for out-of-range regions, it should only be applied for scenes of LDR or scenes of slightly higher dynamic range (e.g. nine to ten stops).

PSNR and TQ error metrics only took into account the view which was generated. The quality of the existing natively captured HDR view was not taken into account as the influence of this is difficult to estimate objectively.

The graphs presented in Figures 6.12 and 6.13 have different characteristics but they reflect dynamics of the scenes well. *Scene 1* starts with a direct view onto the lamp which is switched on (see Figure 6.11 inset B). In the LDR frame that region is overexposed and contains objects which vary in depth (e.g. the light-bulb). All algorithms reconstruct the region with some error which is reflected in PSNR. Even HY, which manages to obtain all the correct details, shifts the position slightly resulting in objective error (which is unlikely to be perceived by the observer). TQ which is also influenced by the quality of a single frame shows similar behaviour. The error generated by EO is reflected in temporal results and it performs the worst. As the light exits the sequence, due to camera pan, the dynamic range decreases and PSNR improves for all the methods while the TQ gap reduces.

A similar trend is observed for *Scene 2* which contains two bright reflections at the beginning and at the end of the sequence. These reflections are out of view in the middle of the video which results in improved PSNR and TQ. *Scene 3* shows a rotating disco ball at which the light is pointed. This results in reflections which are constantly appearing and disappearing. As they are view dependant they may be difficult to reconstruct resulting in low PSNR and high TQ. Due to their frequency and variability both PSNR and TQ graphs oscillate.

In *Scene 4* the lamp is swinging. Two cycles of such movement can be observed on the associated graphs. As the lit (overexposed) region becomes smaller, PSNR increases and vice versa. TQ is positively correlated to the movement speed of the lamp. As the lamp reaches the apex, it slows down and TQ decreases. In *Scene 5*, the robot arm performs an operation consisting of three stages. In each stage the arm moves slowly, but quickly rotates when switching stages. These three movements are reflected in TQ graph and influence PSNR but to a much lesser extent. *Scene 5* also had the lowest average dynamic range and TQ of the EO method was better than that of SAD confirming validity of using the EO method for scenes of lower dynamic range when speed is important.

It is worth noting that the TQ is worse when there is more movement in the scene. However, it is expected that the observers will be less likely to notice flickering caused by artefacts if there is movement of objects in the scene which would attract observers' attention.

6.3.1 Limitations

The challenge of completely recovering out-of-range regions still remains. As data is missing, this method operates on a set of assumptions which may not always be true. For instance, disparities in overexposed regions may not vary smoothly as it may contain objects at multiple depths. The proposed approach is unable to identify such cases, but the generated SHDR image will still contain details of such an object (albeit at changed depth). To alleviate such challenges advanced techniques of machine learning could be used to identify objects and estimate their depth but such an approach would be the subject of future work.

The number of tested scenes was limited by the GT capture techniques which were time-consuming. To improve generalisation more videos should be tested in the future. However, the selected scenes varied in disparity, dynamic range, frequency, amount of motion, noise and contrast and as such could be considered a representative sample of those which are encountered in most situations.

6.4 Summary

Contributions of this chapter include three methods for generating SHDR videos from an HDR-LDR video pair. The first technique used the SAD method, which was developed for single images, on a per-frame basis. The second technique found intensity correspondences between LDR and HDR frames which were used to expand LDR values thereby creating an HDR frame. The final method combined advantages of both using stereo matching for out-of-range regions and EO for the rest. Results showed that the HY method generated SHDR videos of the best overall spatial and temporal quality.

This chapter concludes the consideration of the SHDR capture stage. Obtaining SHDR content produces an increased amount of data which may require compression. SHDR image compression is the topic of the next chapter.

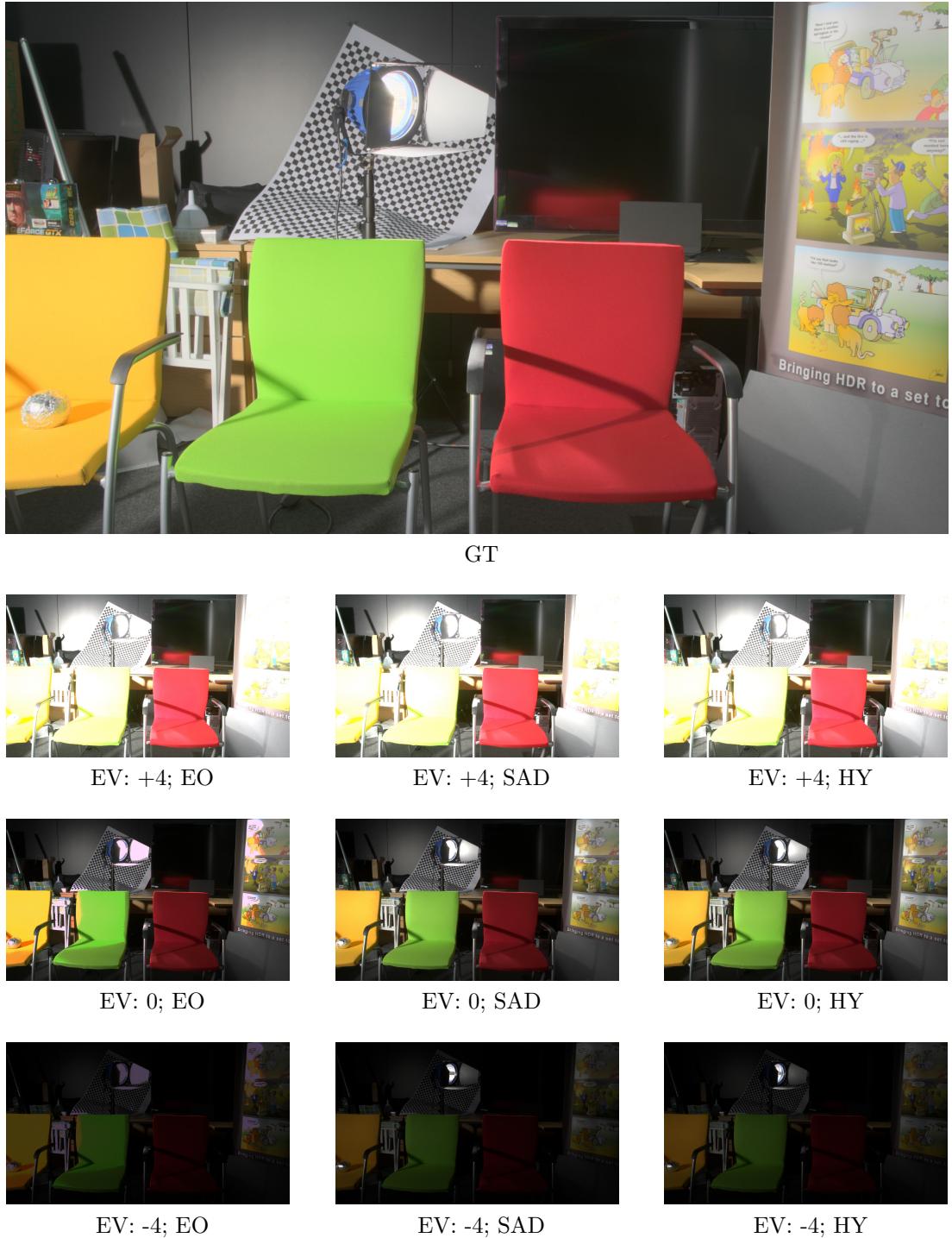


Figure 6.10: The reconstructed frame from the SHDR pair for all methods for Scene 1 are presented. GT is tone mapped to show it here. For each method three single exposures are selected and shown.

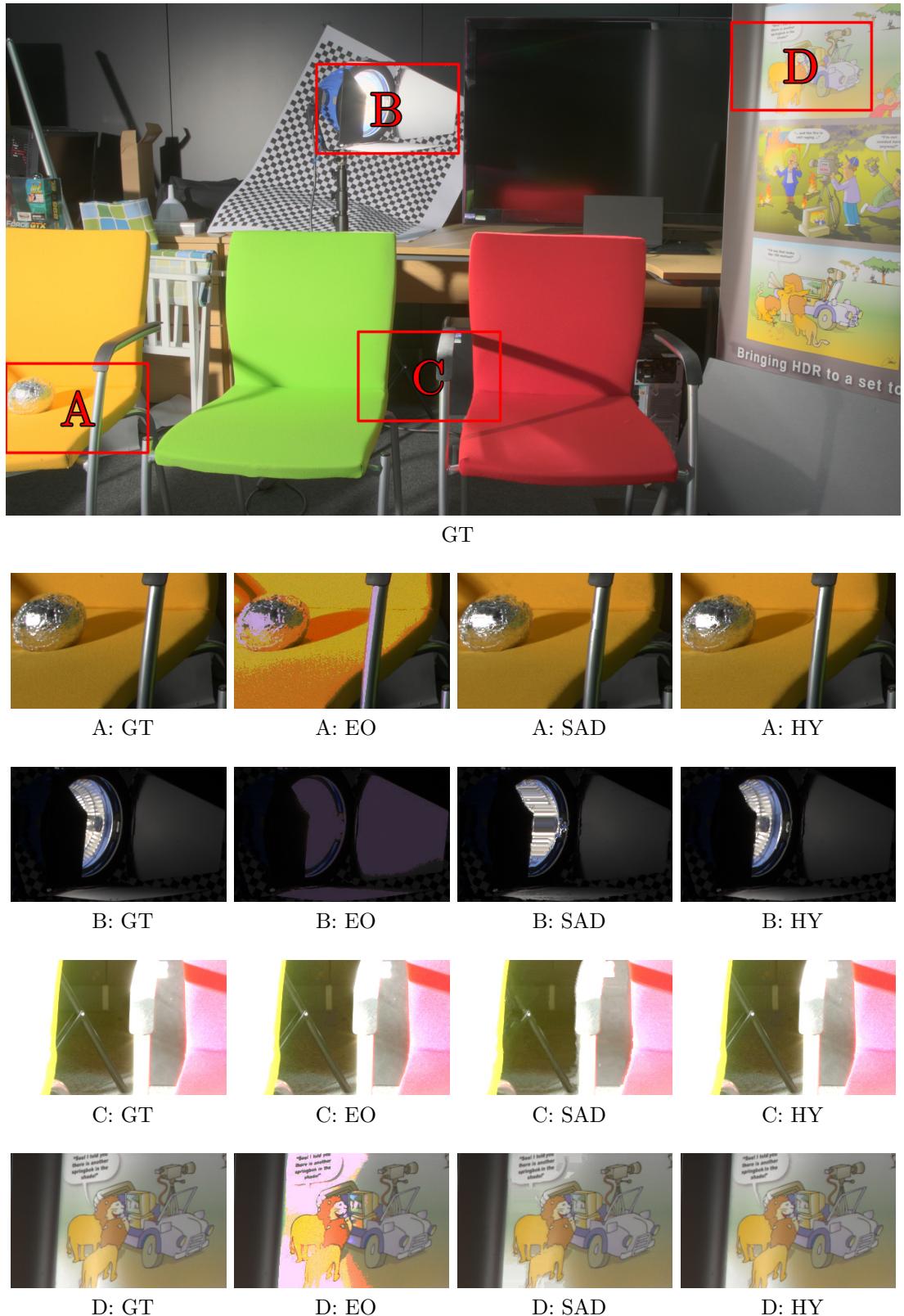


Figure 6.11: Detailed insets for the reconstructed SHDR frame chosen from Scene 1 showing GT, EO, SAD and HY. All images are shown at the appropriate single exposure.

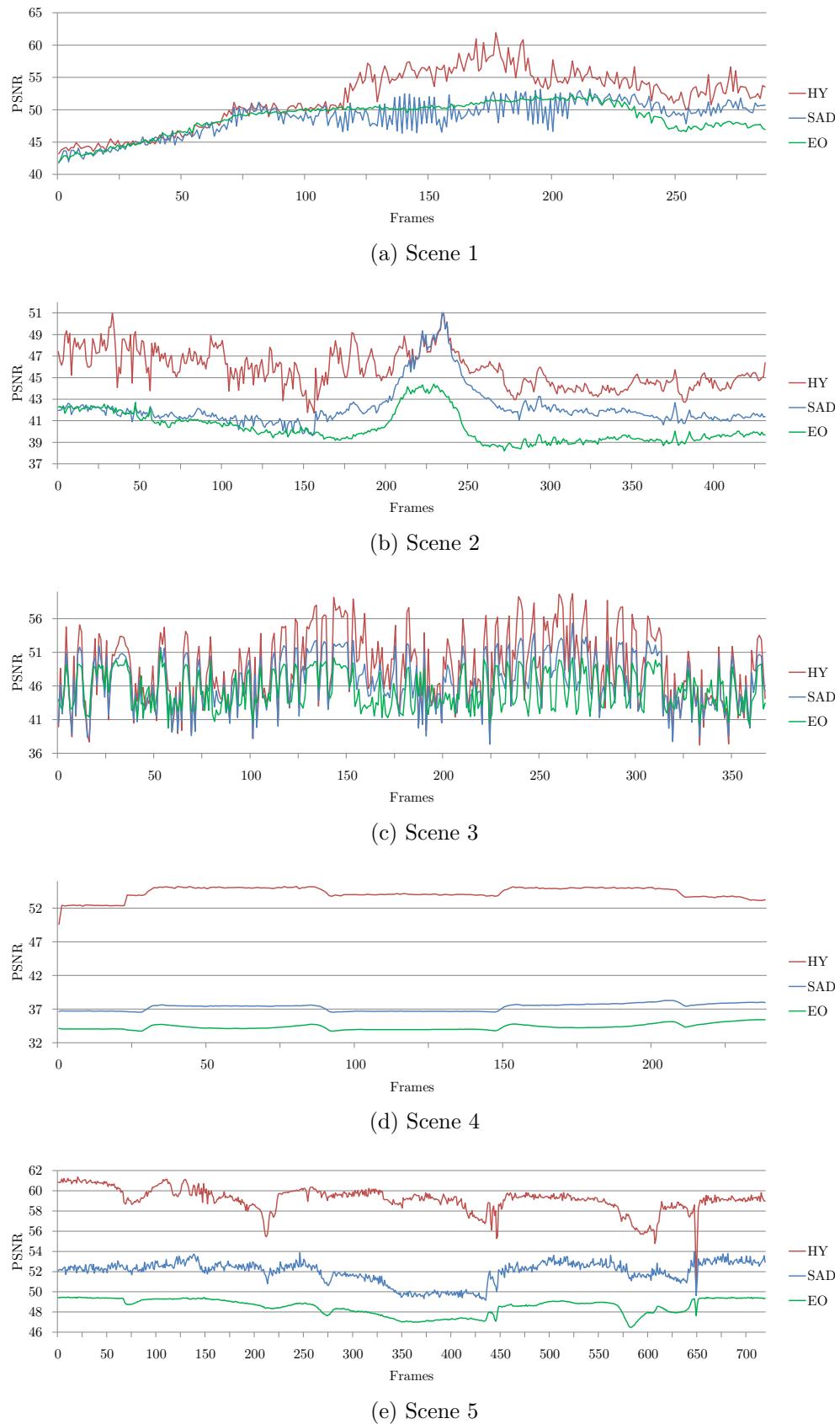


Figure 6.12: PSNR results for all the scenes an all the frames. Higher is better.

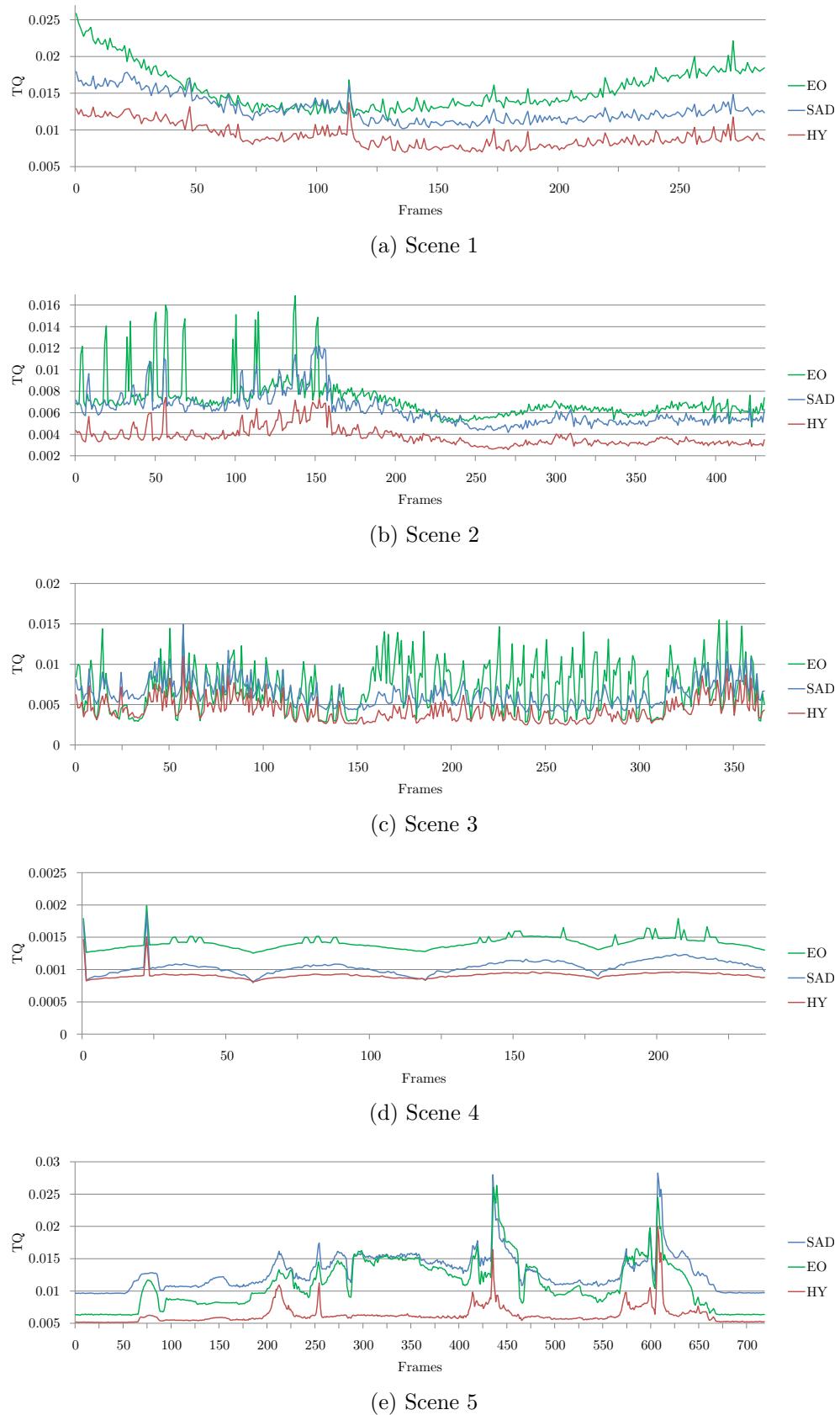


Figure 6.13: Temporal quality (TQ) results for all the scenes an all the frames. Lower is better.

CHAPTER 7

Stereoscopic High Dynamic Range Compression

Captured or generated SHDR data increases the demand for storage and bandwidth eightfold compared to standard LDR data. Uncompressed SHDR images use 32 bit floating point numbers to represent colour requiring 96 bpp per view (192 bits for both views), while uncompressed LDR images use 8 bit integers requiring just 24 bpp for a single view. This means that an SHDR image at high definition resolution of $1,920 \times 1,080$ can be just short of 48 MB. With the emergence of superHD's $7,680 \times 4,320$ resolution this would balloon to 759 MB. This amount of data poses challenges for current ICT infrastructure even when the current decrease in storage price and increase in internet speed are taken into account.

This chapter proposes five novel compression methods which enable efficient compression of SHDR image data. When using the compression methods outlined in this chapter SHDR images are only marginally larger than traditional LDR images (e.g. 30% larger for one of the methods) allowing for the usage of current ICT infrastructure and media. Furthermore, four of the five methods presented provide the ability of opening an SHDR image in a traditional LDR viewer and in a traditional HDR viewer. Two of the methods are backwards compatible with LDR stereo viewers. Such design should facilitate transition from LDR to SHDR technology.

The five proposed methods can be classified into three groups. The first group combines the standard techniques of storing stereo images with JPEG-HDR. Two such techniques are examined: *side-by-side* and *half side-by-side* methods which store tone mapped images next to each other. The only difference is that the *half*

side-by-side method initially reduces the horizontal resolution of the images. The second group stores a single view as JPEG-HDR together with a disparity map which is subsequently used to reconstruct the missing view. Two methods for generating disparity maps are considered, resulting in *image plus disparity* and *image plus disparity with corrections techniques*. The last group utilizes a motion compensation approach used in video coding. A single method which relies on MPEG compression is proposed.

The compression performance of the five proposed techniques was compared using objective measures and an SHDR image dataset. Common compression quality metrics were used: PSNR and RMSE. The *quality \times compression (QC)* provided a quick overall summary of method performance.

7.1 JPEG SHDR Methods

This section examines each of the five proposed SHDR compression methods into more detail. All methods are backwards compatible using the approach inspired by the JPEG-HDR technique of Ward (2005), which was overviewed in Chapter 3. The techniques initially separate each image of an SHDR pair into tone mapped and residual parts which are further processed. Similar to Ward (2005) the JFIF wrapper is used as a format for packing data that SHDR compression methods produce. The main entry stores a tone mapped version of one of the stereo pairs which is encoded using standard JPEG compression. The JFIF wrapper also provides storage channels for metadata. These are utilized to store the additional information such as ratio images, disparity maps and motion compensation information, depending on the technique proposed. The additional data is used to restore the full SHDR content, as shall be discussed below. While, the number (16) of metadata channels and size (64 KB) are limited, they are sufficient for the proposed methods; furthermore, if required, this limitation can be overcome by using more storage channels which have the same identifier. Compatibility with current LDR stereo standards is another goal. In an effort to achieve a balance between image size, quality and backward compatibility, five methods were proposed each of which is explained further in the following subsections.

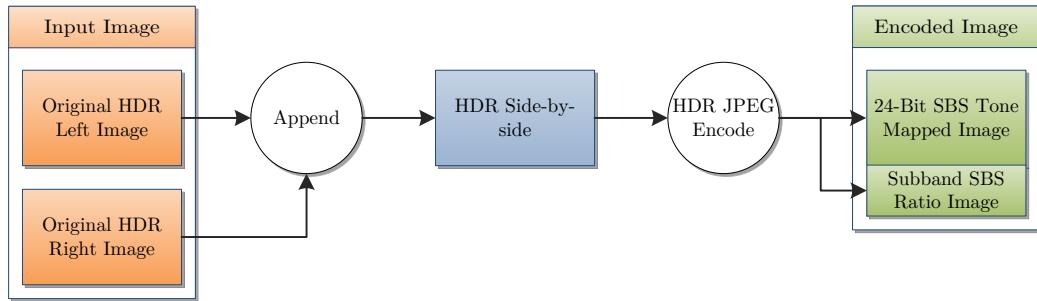


Figure 7.1: SBS encoding

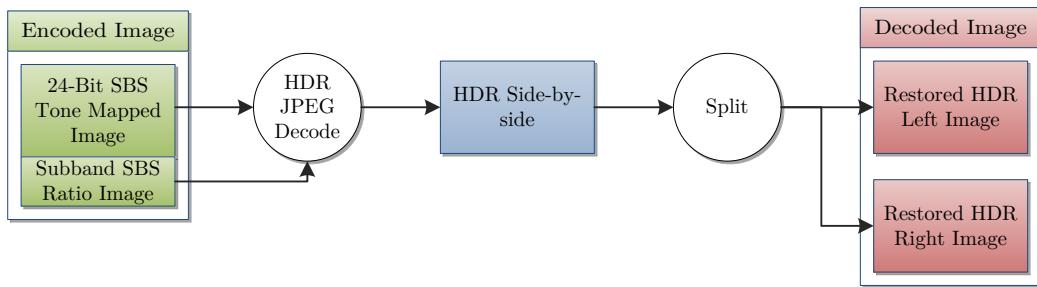


Figure 7.2: SBS decoding

7.1.1 Side-by-side (SBS) Method

The first proposed technique strives to preserve the quality of the original image and minimise data loss at the expense of larger file sizes compared to the other methods. It is also a good foundation for further examination as it provides an initial reference point in terms of quality which other approaches should try to attain. Both of the images in this case are coded using JPEG-HDR so that the quality of the restored image is kept high and further file size reduction is not considered.

This straightforward method starts by appending the right HDR image of the stereo pair to the left one (Figure 7.1). The result is a single side-by-side HDR image. This image is then compressed using JPEG-HDR. The resulting size is almost equivalent to compressing each image separately, and no advantage is taken of the large amount of stereo correspondence between the stereo pair.

There are two ways of formatting the output depending on the capabilities of the viewing software. Data from the second image can be put into JPEG subbands making it available for traditional monocular image viewers which then

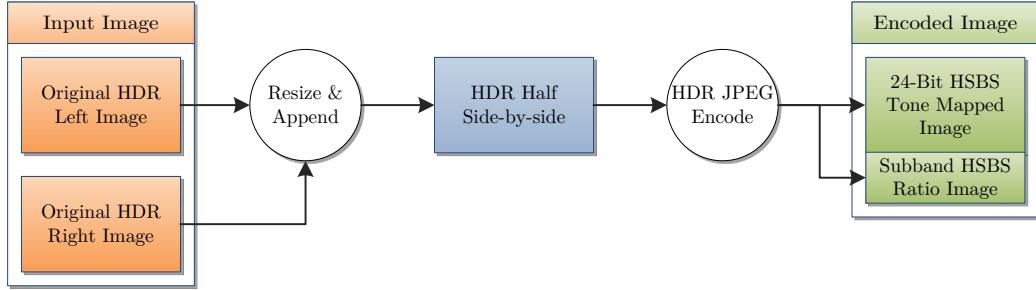


Figure 7.3: HSBS Encoding

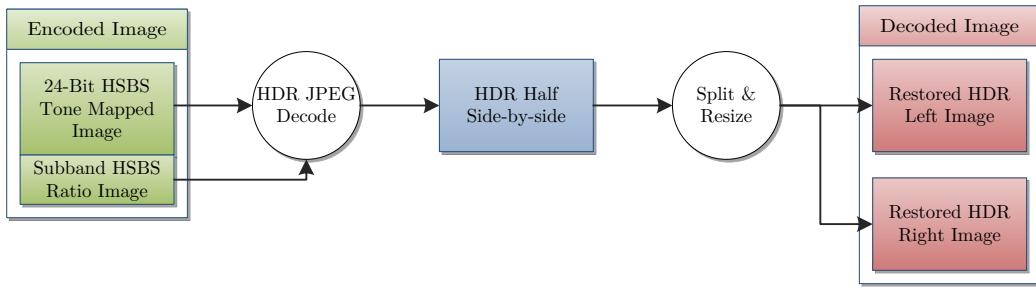


Figure 7.4: HSBS Decoding

show only one of the images of a pair. Alternatively, the tone-mapped images can be left side-by-side and saved as a stereo JPEG (JPS) which is then viewable in LDR stereo viewers (Figure 7.2). This variant can also be opened in traditional viewers but both the left and right views would be displayed. While such behaviour provides at least some insight into the content of the file, it may not be desirable for the user.

7.1.2 Half Side-by-side (HSBS) Method

Another standard way of compressing LDR stereo images puts the pair side by side but halves the horizontal resolution of each such that both can fit in the space of a single image. This is not dissimilar to interleaving the images. The encoding process is shown in Figure 7.3. It starts by resizing the images of the HDR pair as described and proceeds in the same manner as SBS. When the image is decompressed, images are resized back up again (Figure 7.4). Compared to the SBS method the image size is roughly halved but so is its resolution affecting the resulting quality. This method is included primarily because it is backward

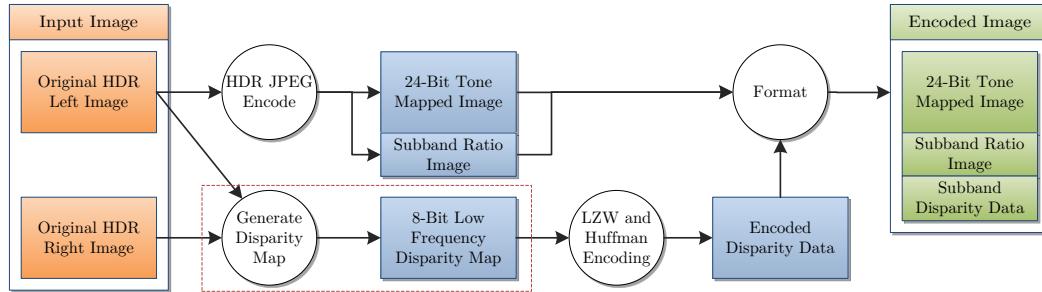


Figure 7.5: IPD Encoding

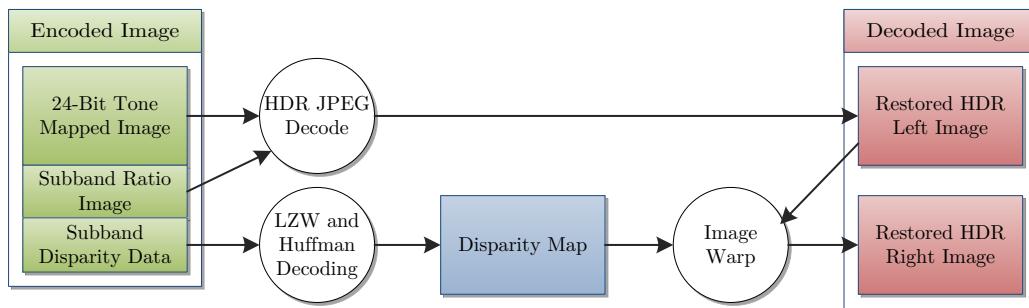


Figure 7.6: IPD Decoding

compatible with LDR stereo viewers.

7.1.3 Image Plus Disparity (IPD) Method

The IPD method exploits the correlation between the left and right views and the correspondence between most of the pixels (exceptions include occlusions, transparencies and specular highlights). The image that represents these correspondences is a disparity map (see Chapter 2). The precise maps can be obtained by using specialist equipment while taking stereoscopic photographs or they can be provided by the rendering software in the case of CG images. However, for the majority of stereo images, disparity data is not available and needs to be calculated using the stereo pair (Chapter 2).

Any disparity map can be used with this method (calculated or captured), however smooth low frequency maps are preferred because of better compression rates. For results presented here the generated disparity maps were obtained employing a technique suggested by Mei et al. Mei *et al.* (2011) which is named AD-census (ADC) and it is considered one of the best when evaluated on the

testbed (Scharstein *et al.*, 2001). This method utilises the GPU during calculations leading to fast performance, and also the technique produces the least number of errors according to the testbed (Scharstein *et al.*, 2001).

The encoding process using the IPD method is shown in Figure 7.5. It starts by generating a disparity map from the HDR pair, if it is not already present. This map is then encoded using lossless LZW and Huffman coding. The output file size is not fixed but it is a fraction of the original image and can easily fit in JPEG subbands. One of the HDR images is compressed using HDR JPEG and thus stores the ratio image in the subband and a tone-mapped JPEG image. Decoding inverses the process (see Figure 7.6) by extracting and recovering the disparity map and restoring one of the images using JPEG-HDR decoding. This HDR image is warped using disparity to obtain the missing view.

7.1.4 Image Plus Disparity with Corrections (IPDC) Method

During testing of the IPD method it was observed that some occluded regions did not restore well during the image warping stage. The problem is similar to the one faced in Chapter 5 for the COGC method. In the IPD approach edges were misplaced and a number of tested scenes had major offset issues. The cause of the problem was that disparity maps were smoothed. Also some background pixels that were required for restoring the missing view, were occluded by the foreground objects. This resulted in those foreground objects being warped to wrong positions and some of the objects being perceived at incorrect depths.

The alternative to the method used for IPD is to use a disparity map that maps only the closeness of RGB values, such as the SAD method (Chapter 2). However, the SAD method produces disparity maps with more high frequency content (see Figure 7.7), and are therefore less amenable for efficient compression. The IPDC method avoids the problems of both these methods by combining them and using SAD in regions with large differences only. This process is shown in Figure 7.8 and is used instead of the disparity generation step highlighted by the dashed square in Figure 7.5. Once the low frequency map is obtained using the ADC method, the image is warped. Artefacts are identified by dividing the warped image and the original image and finding pixels which are above an empirically obtained threshold. Those pixels on the ADC disparity map are updated with disparities from the high frequency map (which is obtained using SAD). The rest of the coding process is identical to the IPD one.

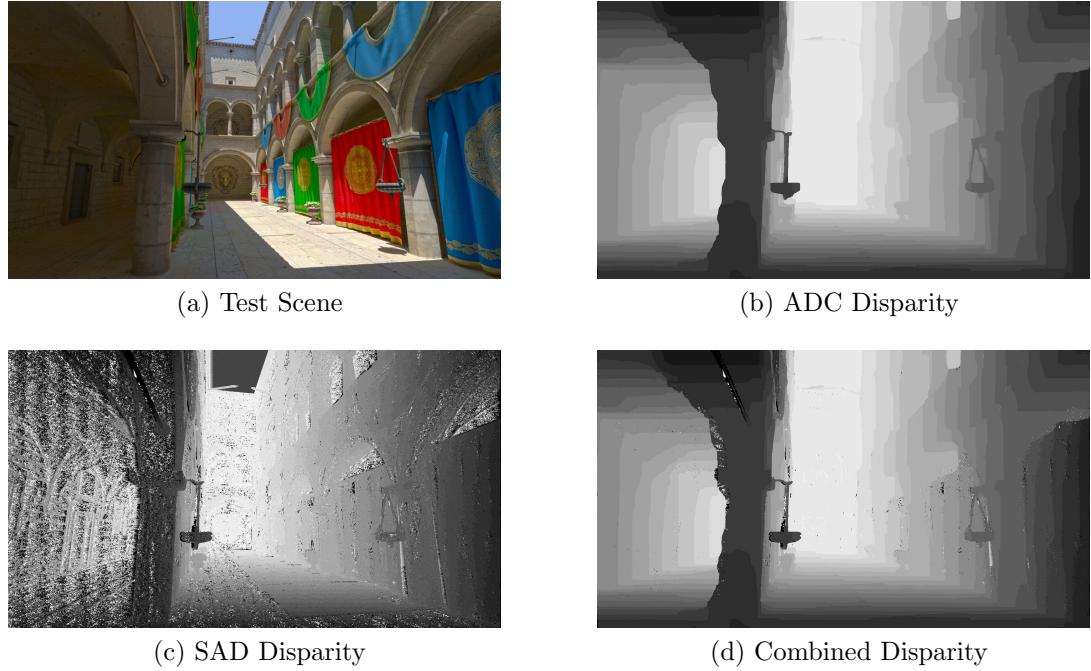


Figure 7.7: For the CG test scene (a) the ADC algorithm generates a smooth disparity map (b) which might create artefacts when generating a novel view. For example, due to smoothness constraint, the depth of the thin pole in the top left corner is lost. The SAD algorithm produces a high frequency disparity map (c) which is more difficult to compress. The two are combined (d) where SAD disparities are used for challenging regions and ADC for the rest.

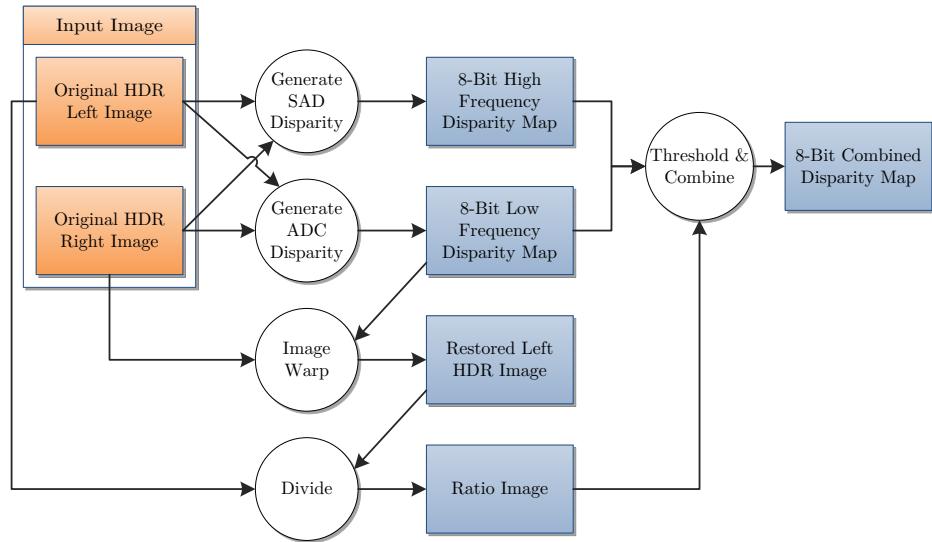


Figure 7.8: IPDC Disparity Generation

7.1.5 Motion Compensation (MC) Method

The final proposed method is based on the observation that the two views differ only by the camera position, which is similar to the temporal motion between subsequent frames in videos. This fact motivated investigation into using standard video coders to compress the SHDR image. Two views of a stereo pair are treated as two consecutive frames in a video which is then processed. Any video coder or camera motion compensation can be used here. For our tests we used the H.264 codec.

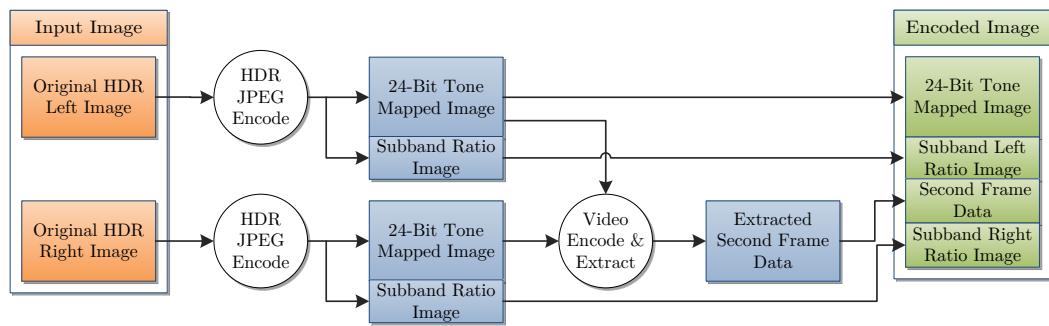


Figure 7.9: MC Encoding

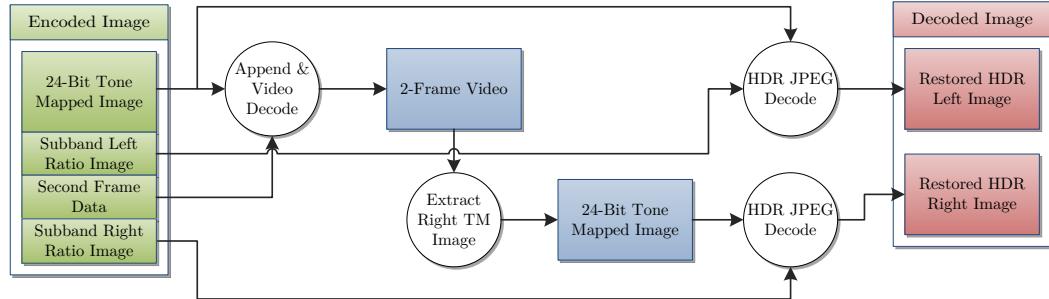


Figure 7.10: MC Decoding

The pipeline for this method is presented in Figure 7.9. Encoding starts by compressing the left and right images separately using JPEG-HDR. To support backwards compatibility the left image is stored in the JPEG-HDR format and is not MC encoded. The second view is encoded using video motion compensation. Tone-mapped images of both views are merged to create a two frame video, which is processed by the video encoder. The second frame data of the compressed video gets extracted. In the H.264 example the second frame corresponds to

Table 7.1: Compatibility. * depending on storage either Mono or Stereo LDR - not both

	SBS	HSBS	IPD	IPDC	MC
Mono LDR	✓*	✗	✓	✓	✓
Stereo LDR	✓*	✓	✗	✗	✗
Mono HDR	✓	✓	✓	✓	✓
Stereo HDR	✓	✓	✓	✓	✓

the predicted frame (p-frame) data. The size of the extracted data depends on the compression quality used but for moderate compression values it is rather small (see Table 7.2) and can fit in additional JPEG subbands together with the JPEG-HDR ratio images for both frames.

The decoding process is the inverse of the coding pipeline, (Figure 7.10). A video file consisting of two frames is generated using the tone-mapped image which is appended to the second frame data. This video is then decoded which provides the second image for reconstruction using JPEG-HDR. A standard viewer therefore opens the stored JPEG and a JPEG-HDR viewer will open only the HDR image of the stored view, while an SHDR viewer opens the SHDR image.

7.2 Results and Analysis

Nineteen SHDR images were used to evaluate and test the proposed methods. All were captured at the resolution of 1920×1080 . To test the algorithms, a variety of scenes were chosen which differ in dynamic range, depth, frequency, amount of noise and contrast. Four scenes were computer generated. Tone-mapped versions of the left view together with dynamic range and disparity range are shown in Figure 7.13. When coding JPEG-HDR images, to preserve quality, the parameter controlling it was set to 95 (out of 100). The disparity maps for the IPD method were generated using the default settings suggested by authors. The SAD method used in IPDC had a window size set to 3 pixels.

A compatibility table of all methods is presented in Table 7.1. Please note that the SBS method is either compatible with a traditional LDR viewer or an LDR stereo viewer depending on how the second frame data is formatted.

The difference between the output of each method for one of the images is shown in Figure 7.11. One image of the pair is shown. For IPD, IPDC and

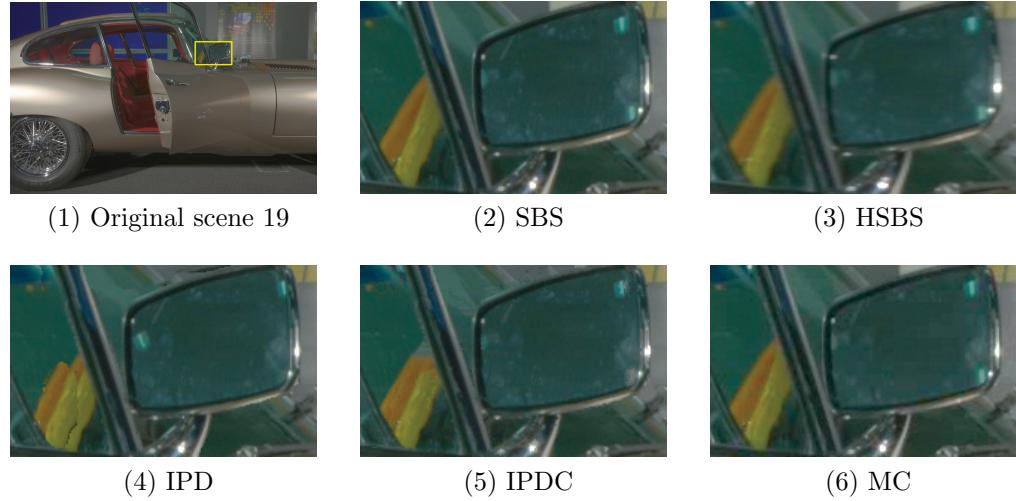


Figure 7.11: Decoded Image for Each Method

MC methods this is the disparity or motion reconstructed view, as the other is equivalent to the image obtained using the SBS method. To examine the differences in more detail a small segment (highlighted with a yellow rectangle) from Scene 19 was selected. As expected, the SBS method is the most similar to the original image as there are no prominent artefacts. The loss of resolution is visible in the HSBS image which is blurrier than the rest. Mistakes made during image warping due to the occlusion are present in the IPD method especially in the case of the steering wheel. IPDC manages to fix this problem to an extent but small mistakes are noticeable towards the top of the steering wheel which has gray values instead of the original yellowish values. The MC method appears very similar to the original but on closer inspection some “blocky” artefacts due to MPEG compression are visible.

In order to evaluate the performance of the presented methods the following quantities were measured: file size of the compressed SHDR image (in kilobytes), peak signal-to-noise ratio (PSNR) and root-mean-square error on logarithmically scaled HDR values (RMSEL) between the decoded images and the original HDR images. Values for RMSEL were logarithmically scaled in order to avoid biasing the result towards high intensity while PSNR already includes such a scaling. PSNR and RMSEL were measured for left and right views separately and then averaged. File sizes are shown in Table 7.2. The last table column contains the sizes of a single LDR image of the same scene which is JPEG encoded using the same quality settings as the proposed methods. PSNR is presented in Table 7.3

Table 7.2: Sizes of images compressed with the methods from this chapter (in KB) and a single LDR image of the same scene shown for comparison. Size increase compared to the LDR image is given as percentage.

Img	SBS	HSBS	IPD	IPDC	MC	LDR
1	1087	561	601	717	683	476
2	1088	558	613	728	671	482
3	868	459	488	616	544	370
4	897	493	494	589	557	384
5	891	506	474	514	549	383
6	1697	899	956	1392	1104	794
7	1021	512	572	704	616	449
8	784	417	436	458	484	328
9	1648	837	890	1119	977	769
10	1215	617	629	682	707	544
11	1087	549	563	577	639	487
12	436	267	245	263	285	168
13	1326	689	728	877	803	601
14	1004	530	549	589	617	439
15	888	455	549	656	543	362
16	1457	744	786	1024	913	685
17	1456	793	849	1132	883	676
18	954	491	529	603	590	412
19	1026	537	569	617	640	449
Avg	1096	574	606	729	674	487
%	125	18	24	50	38	0

and RMSEL is shown in Table 7.4. Average values for all of the scenes are shown in the last row.

Table 7.5 provides a summary of the results. The compression ratio gives the average compression ratio compared to a raw HD stereo image. The “quality \times compression” measure (QC) is a multiplication of the average image size for each method multiplied by the average NRMSE. This value is presented to give an idea of the tradeoff between quality and size, but should not be taken as a definitive measure as the different methods have distinct qualities. For “quality \times compression”, smaller values are considered better.

The MC method achieves the best overall QC results. It also has the added advantage that it is backward compatible with JPEG-HDR and JPEG. The HSBS achieves the highest compression ratio but this comes third for the QC due to the quality of images, for which it is second from the bottom. It is backward

Table 7.3: PSNR Measure

Img	SBS	HSBS	IPD	IPDC	MC
1	58	53	51	53	58
2	59	52	54	56	59
3	62	59	55	56	61
4	62	55	59	59	61
5	49	48	46	47	45
6	50	45	44	47	47
7	62	52	55	57	60
8	56	48	52	53	54
9	47	43	43	45	46
10	45	41	41	44	43
11	41	38	40	40	38
12	55	51	54	54	51
13	54	48	49	49	52
14	55	52	50	53	55
15	51	49	46	48	50
16	52	49	47	48	50
17	49	42	41	44	46
18	62	56	59	59	61
19	54	51	51	51	54
Average	54	49	49	51	52

compatible with LDR stereo JPEG but not fully compatible with JPEG-HDR and traditional JPEG; the image shown on a traditional JPEG viewer would show both images side by side. The SBS method is second overall and is backward compatible with all possible formats, however it is the largest in size, which may be too much of a prohibitive obstacle, potentially hampering the uptake of SHDR. The disparity methods IPDC and IPD are second from last and last respectively. However, the results depend on the quality of the disparity maps produced, and these may become better as this is an active research area, so the scope of such disparity based methods is likely to improve. In addition these methods are backward compatible with traditional JPEG and HDR JPEG and produce relatively small image sizes (not much larger than JPEG-HDR images).

Table 7.4: RMSEL Measure

Img	SBS	HSBS	IPD	IPDC	MC
1	0.0030	0.0062	0.0087	0.0063	0.0032
2	0.0037	0.0089	0.0085	0.0062	0.0044
3	0.0026	0.0044	0.0092	0.0075	0.0032
4	0.0029	0.0075	0.0053	0.0043	0.0033
5	0.0017	0.0020	0.0030	0.0023	0.0027
6	0.0028	0.0056	0.0071	0.0046	0.0039
7	0.0029	0.0068	0.0075	0.0059	0.0035
8	0.0019	0.0043	0.0037	0.0030	0.0025
9	0.0063	0.0126	0.0148	0.0109	0.0085
10	0.0021	0.0037	0.0040	0.0026	0.0027
11	0.0028	0.0039	0.0033	0.0032	0.0041
12	0.0017	0.0030	0.0020	0.0019	0.0027
13	0.0038	0.0084	0.0092	0.0076	0.0049
14	0.0039	0.0078	0.0069	0.0055	0.0042
15	0.0033	0.0080	0.0130	0.0078	0.0040
16	0.0043	0.0090	0.0113	0.0087	0.0062
17	0.0040	0.0089	0.0155	0.0086	0.0060
18	0.0030	0.0071	0.0058	0.0050	0.0038
19	0.0041	0.0091	0.0082	0.0070	0.0044
Average	0.0032	0.0067	0.0077	0.0057	0.0041

Table 7.5: Results Summary. Compression Ratio (CR) and Quality \times Compression (QC). NRMSE is used for quality and for $Q \times C$ lower figures are better.

	SBS	HSBS	IPD	IPDC	MC
CR	44	85	80	67	72
QC	3.51	3.85	4.67	4.16	2.76

7.3 Summary

This chapter proposed five methods for compressing SHDR images. The first technique applied the JPEG-HDR technique on an SHDR pair which was merged by positioning two images side-by-side, while the second halved the horizontal resolution before merging them in the same manner. The third proposed method encoded one of the views using JPEG-HDR compression but also utilised a pre-calculated disparity map which allowed generation of the second view during decoding. The fourth method also used a disparity map which was generated by combining two stereo matching techniques: SAD and ADC. The final technique

for compressing SHDR images relied on motion compensation by treating each of the images in a pair as consecutive frames. Objective measures showed that the last method generated the best overall QC result.

This chapter concludes the consideration of different the SHDR stages. The next chapter concludes the thesis and discusses future work.



Figure 7.12: SHDR scenes are tone-mapped for illustrative purposes. The dynamic range and the disparity range for each scene is shown below.

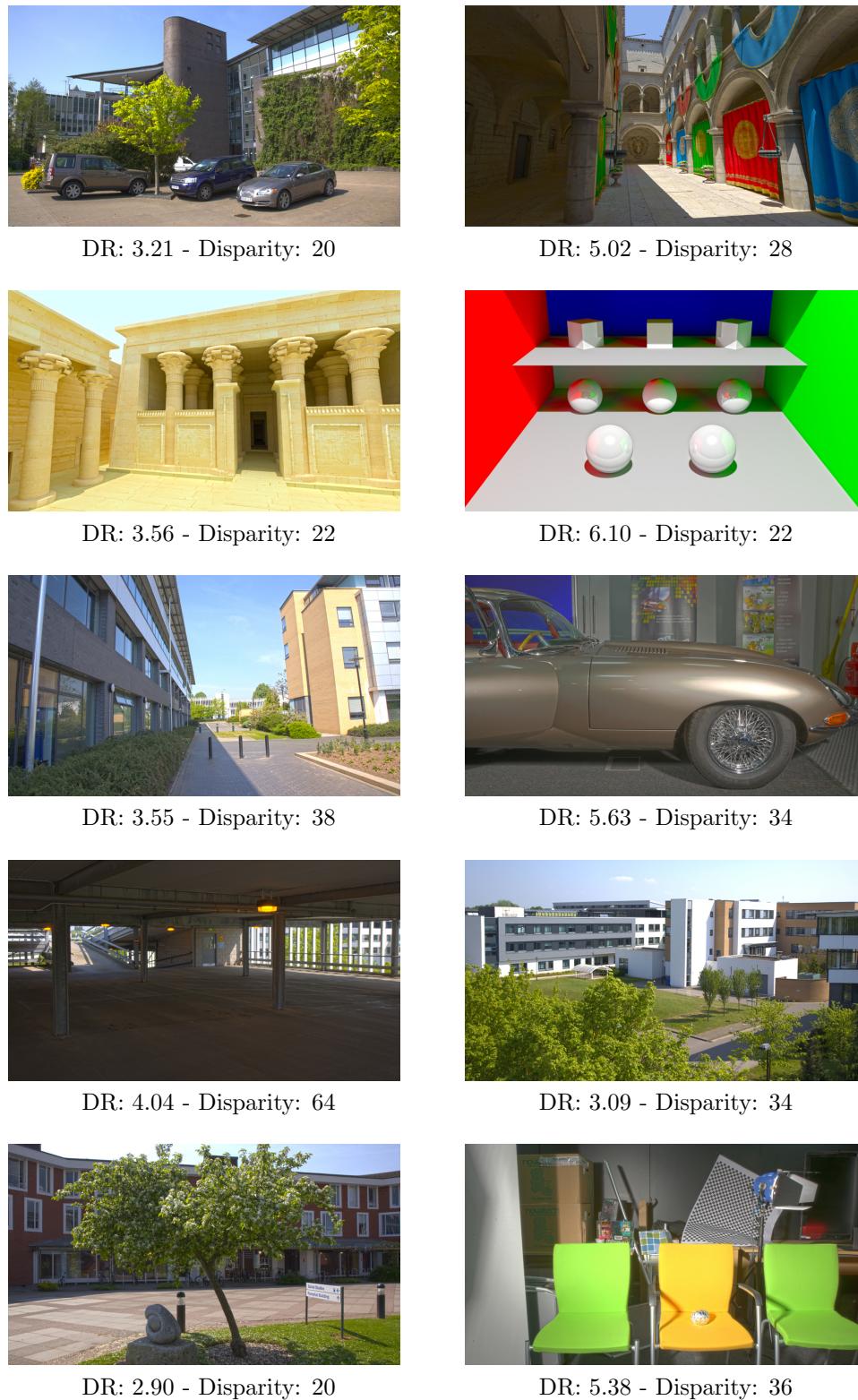


Figure 7.13: SHDR scenes continued.

CHAPTER 8

Conclusions

This thesis has introduced the novel imaging technology of stereoscopic high dynamic range imaging which aims to improve the representation of captured and generated images and videos. Chapter 4 identified and discussed possible approaches for solving challenges presented by combining stereoscopic with high dynamic range imaging. Chapter 5 suggested and validated an approach for generating SHDR images from an HDR-LDR stereo image pair. Chapter 6 generalised the approach used for images and extended it to work with SHDR videos in a temporally coherent manner. Finally, Chapter 7 detailed backwards compatible techniques for compressing SHDR images. This chapter presents a summary of the contributions and discusses possibilities for future work.

8.1 Capture of Stereoscopic High Dynamic Range Images

In Chapter 5, an approach was presented which facilitates the capture and creation of SHDR images, avoiding the need for two native HDR cameras. The approach enhanced a captured HDR-LDR image to an HDR-HDR (SHDR) image. To this end, four techniques were suggested. Two techniques used expansion operators to extend the dynamic range of the LDR image while the other two used stereo correspondence between left and right views to transfer HDR data.

Two expansion operators tested the feasibility of expanding the dynamic range of the LDR image from an HDR-LDR pair for the purpose of generating an SHDR image. The straightforward global method of LS expanded the dynamic range quickly by using the same expansion function for all image pixels. The more advanced local EM method explicitly handled overexposed image regions during expansion but took more time to compute.

Two methods for generating disparity maps were also tested. The maps guided transfer of HDR data onto the LDR view. SAD was a quick and straightforward method which produced high frequency maps but correlated pixels of similar intensities. The COGC method generated smooth disparity maps and handled occlusions explicitly but was slower than SAD and produced artefacts.

To validate the approach of generating SHDR images from an HDR-LDR image pair, a user study compared all four proposed techniques with ground truth (GT), a natively captured HDR-HDR image. Generated images were displayed using the two HDR displays and a custom built stereo rig. Five scenes were used and 26 participants took part in the experiment. The method of balanced paired comparisons allowed ranking of the tested techniques. The methods based on the EO came last for all the scenes. The COGC method was third overall. The SAD technique was continually found in the first group with GT which was reflected in the aggregated result where the SAD method was not significantly different from the reference image. This indicated that SAD or similar methods may be used for enabling SHDR content using an LDR and HDR camera setup. In addition, objective measurements of PSNR and RMSEL tested the quality of images and were found to correlate with the user study results.

8.2 Capture of Stereoscopic High Dynamic Range Video

While Chapter 5 validated the viability of SHDR from an HDR-LDR image pair concept, Chapter 6 generalised it and extended it for videos. An important aspect of the video approach was temporal coherence which had to be preserved in order to avoid noticeable artefacts such as flickering. The SAD method which performed best for static images was extended to video by treating frame pairs as a stereo images. To minimise temporal artefacts two additional methods were proposed.

The EO method for videos relied on expanding the dynamic range of the LDR image, but instead of using an existing EO designed for a generic usage, a novel technique was proposed. Intensities of HDR pixel values were sorted in bins so that the resulting histogram matched the LDR histogram. This allowed for a more precise mapping of HDR to LDR values compared to the generic methods and resulted in temporally coherent video. However, the technique was susceptible to problems in overexposed regions as one value was used for the

whole overexposed range.

The HY method, combined the previous two and took advantage of both the stereo correspondence and EO techniques. It used stereo matches to reconstruct out-of-range regions and expanded the dynamic range of the rest of the pixels. In addition, it incorporated a correction step which reduced artefacts caused by stereo matching. The technique provided temporal consistency and was able to handle overexposed and underexposed regions.

To test the quality of reconstructed videos, objective measurements were used. PSNR evaluated the quality of individual frames. Temporal quality was measured using the TQ metric which assessed differences between consecutive frames. Five video sequences of varying dynamic range, length and amount of movement were evaluated. The HY method performed best for both metrics and for all the scenes, which suggested it is a viable method for generating SHDR videos from HDR-LDR sequences. The SAD technique was second overall. The EO method had better TQ than the SAD method only for the scene of lowest dynamic range but came last elsewhere.

8.3 Compression of Stereoscopic High Dynamic Range Images

Chapter 7 was concerned with the increased size of captured SHDR images compared to the LDR images and focused on the storage aspect of the SHDR pipeline. To facilitate the storage and transmission of SHDR data, five compression techniques were suggested. Four of them were backwards compatible with monoscopic LDR and monoscopic HDR viewers while one was backwards compatible with stereoscopic LDR viewers. Including backwards compatibility was aimed at facilitating the future adoption of SHDR imaging.

The proposed compression methods were based on the JPEG standard while the concept of achieving backwards compatibility was inspired by the JPEG-HDR approach. All five methods initially separated each image of an SHDR pair into tone mapped and residual parts which were further processed and stored. The JFIF wrapper packed data, so that the main entry stored one tone mapped image of the pair which was JPEG encoded. JFIF's metadata channels stored the additional information such as ratio images, disparity maps and motion compensation information, depending on the technique proposed.

The SBS method stored tone mapped images next to each other in the main channel while the residuals were put into the auxiliary stream which enabled backwards compatibility with stereoscopic LDR. Alternatively, a single tone mapped image could have been saved in the main channel with the rest packed in an auxiliary channel thereby supporting monoscopic LDR and HDR. The HSBS technique differed from the SBS in that it halved the resolution of the HDR images horizontally before further processing. The IPD technique generated a disparity map between the left and right views. A tone mapped image of one view was saved in the main channel while the residual of the same view and the disparity map were stored in the auxiliary stream. In the decoding stage, the disparity map was used to reconstruct the missing view. The IPDC method added a correction step to the IPD disparity map generation step where more robust correspondences were used for problematic regions. The MC method used motion compensation to compress the second image. The pair of tone mapped images were treated as frames in a video and were compressed using video encoders which generated motion compensated data. The tone mapped image of one of the views was put in the main channel while residuals of both views and motion compensated data of the second view was placed in the auxiliary stream.

Objective measures compared the compression performance of the five proposed techniques. Common compression quality metrics were used: PSNR and RMSEL. In addition “quality \times compression” (QC) provided a overall summary of method performance. A data set of 19 SHDR images was used for testing. The MC method achieved the best overall QC and was also backwards compatible with both monoscopic viewing modes. The HSBS method had the highest compression ratio but had the second to last image quality which put it onto the third place for QC. The HSBS method was backwards compatible with stereoscopic LDR viewers but not with either of the monoscopic viewers. The SBS method achieved the second QC position and is backwards compatible with all the formats, but it also generated files of the largest sizes. For the QC metric, the disparity methods IPDC and IPD came second to last and last respectively. Results depended on the disparity map quality, so as the algorithms generating disparity maps improve, so might the techniques. Both techniques are backwards compatible with monoscopic LDR and HDR.

8.4 Contributions

The combination of stereoscopic and high dynamic range imaging was mostly unexplored until now. This thesis built upon the existing knowledge in both these fields and developed new methods for enabling SHDR. While multiple challenges in the whole SHDR imaging pipeline were discussed, this thesis focused on the capture and compression aspects of pipeline. The main contributions of this thesis are:

- A comprehensive literature review and critical analysis of the stereoscopic and high dynamic range imaging fields, and a discussion of how the two can be combined and the potential challenges this raises (Chapters 2, 3 and 4).
- Four methods for generating SHDR images using an HDR-LDR camera pair. Such an approach facilitates the capture of HDR content making it more accessible and feasible than using two HDR cameras (Chapter 5).
- An objective and subjective comparison of the four techniques against the ground truth - an SHDR image captured using two HDR cameras (Chapter 5).
- Temporally robust methods for capturing SHDR video using an HDR-LDR camera pair were evaluated against SHDR video. The techniques were guided with insights obtained from the static image algorithms (Chapter 6).
- Five techniques for backwards-compatible lossy compression of SHDR images which may reduce their size up to 70 times (Chapter 7).
- An objective evaluation of these compression techniques which compared the quality of encoded images against the raw image. The choice of the best technique is suggested by taking the ratio between size and quality into consideration (Chapter 7).

8.5 Impact

Techniques which enable SHDR imaging could impact areas which benefit from the improved quality of imaging. The entertainment industry is a prime example, where techniques presented in this thesis can be applied. SHDR capture methods

could allow recording of movies and shows in SHDR at a lower cost than using two HDR cameras. An added benefit includes the presence of one standard LDR camera allowing directors to use traditional filming techniques and generating a video stream compatible with current technologies. Compression techniques allow recorded data to be delivered to the end-user using existing media. As using the proposed SHDR compression techniques generates images which are only marginally larger than LDR images (e.g. 30%) current broadcasting channels and media such as DVDs and Blue-Rays could be used to deliver content. Other application areas include:

- *Virtual reality* where SHDR has the potential to improve immersion;
- *Computer vision* algorithms could benefit from SHDR data and improve their performance (e.g. feature matching, categorisation, object recognition);
- *Security* where SHDR video monitoring could provide improved depth cues and present the full range of light visible to the human eye;
- *Education and training* which relies on visual input (e.g. surgery, pilot training, driving) could benefit from the more realistic visualisation of relevant scenarios.

8.6 Future Work

This section suggests possible directions for future work, both in terms of the work presented in this thesis and for stereoscopic high dynamic range imaging in general.

8.6.1 Extended User Studies

The length of the experiment and the number of operators tested in the user study limited the number of scenes that were evaluated. A user study with more scenes, but fewer operators could improve the generality of the obtained results.

An evaluation of SHDR from HDR-LDR video operators and SHDR compression operators could also benefit from a user study. While in Chapter 5 it was shown that objective measurements correlate to the subjective results they neglect effects of binocular fusion. A user study would take binocular fusion into

the account and improve confidence in the objective results, especially for the temporal coherence of SHDR videos. For cases when running a user study is not feasible, developing a stereoscopic content quality metric which would include the binocular fusion phenomenon would be beneficial.

While it is expected that SHDR will provide improved depth perception and light representation the impact of the technology could be evaluated in a user study. The interaction of stereoscopic and HDR imaging could also be examined. For instance, HDR may also benefit depth perception increasing the impact.

8.6.2 Additional Operators

As discussed in Chapters 2 and 3 many stereo matching and expansion techniques exist, only a subset of which were evaluated for the purposes of SHDR capture and compression. Using other existing techniques may result in images of higher quality but this requires additional testing.

Chapter 6 proposed an operator which combined the advantages of both (stereo matching and EO) and had the best performance. It is likely that there are improvements to this baseline technique and that novel operators will be proposed in the future.

Compression operators were focused on both efficient encoding of images and providing backwards compatibility. By abandoning the backwards compatibility constraint, additional gains in compression performance are feasible. Once SHDR imaging enters the mainstream such techniques will become especially important.

As noted in Chapter 7 the size of a raw SHDR image is eight times larger than its LDR counterpart, and requires compression for efficient storage and transmission. SHDR videos are even larger, for instance a 100 frame raw SHDR video would consume approximately 50 GB. This highlights a need for compression of SHDR videos. Any video compression scheme would also have to take into account temporal coherence between frames. An SHDR video compression codec may follow similar concepts to those introduced in Chapter 7.

8.6.3 SHDR Display

The display aspect of SHDR imaging was not explicitly explored in this thesis as designing hardware solutions which would enable practical native visualisation of SHDR content were out of the scope of this work (the stereo rig from Chapter 5

was designed for the experiment and is not practical for everyday use). However, once created, such devices could find application in all the areas mentioned in the previous section.

Until SHDR displays become a reality, tone mapping is the approach which allows visualisation of SHDR content on currently available stereo displays. While current TMOs might be used to this end some considerations are required. For instance, the same parameters may have to be used for left and right views to ensure cross-view consistency. Local TMOs are more susceptible to inconsistencies across images as they consider local regions which might be different between left and right images due to view-dependent phenomena such as occlusions and specular highlights. The problem is similar to that of video tone mapping where two views can be treated as consecutive frames. Current TMOs do not take into account phenomena which are relevant for stereoscopic depth perception such as edge contrast. This means that designing TMOs specifically for SHDR images and videos is a challenge which requires a solution in the future.

8.6.4 Beyond SHDR Imaging

SHDR imaging has the potential to provide more visual cues than traditional LDR viewing. However, it does not cover all the cues perceived by the HVS. For instance, instead of motion parallax which occurs once the observer moves their head, stereoscopic imaging results in sheer distortion. This is because the HVS expects novel views of the scene and stereoscopy offers two predetermined views. Multiview imaging overcomes this problem by providing more views for different observer's positions. SHDR techniques based on disparity maps can be extended in a straightforward manner to enable multiview imaging, as novel HDR views can be rendered using the disparity map.

Finally, one of the ultimate goals of digital imaging is to enable a scene representation that is indistinguishable from reality. Holography allows for 3D scene representation and provides accommodation depth cues. As such, once it develops, the field is expected to provide a scene representation with unprecedented level of realism. Holography might benefit from SHDR techniques that capture the full range of light (HDR) and two views which can be used to generate partial 3D scene representation.

8.7 Final Remarks

Analog and digital imaging has been improving the way in which we represent reality for decades. Achieving a representation which is indistinguishable from reality remains out of reach for now. Stereoscopic high dynamic range imaging, introduced in this thesis, has made some headway towards this goal by adding to the body of knowledge in this domain. The proposed image and video SHDR capture methods have reduced costs and increased the practicality of recording SHDR content. The suggested backwards compatible SHDR compression techniques reduced the size of captured data, making it possible to store and transfer SHDR content using existing media while still preserving compatibility with traditional LDR technology. These approaches have tackled several important research challenges posed by the introduction of the SHDR imaging pipeline and provided a firm foundation for further research. The achieved contributions are a step towards a completely realistic scene representation.

Bibliography

- Adelson, S. J. & Hodges, L. F. (1993). Stereoscopic ray-tracing, *The Visual Computer* **10**(3): 127–144.
- Adobe Developers Association (1992). TIFF Revision 6.0, *Technical report*.
- Aggarwal, M. & Ahuja, N. (2004). Split Aperture Imaging for High Dynamic Range, *International Journal of Computer Vision* **58**(1): 7–17.
- Akyuz, A., Fleming, R., Riecke, B., Reinhard, E. & Bulthoff, H. (2007). Do HDR displays support LDR content?: a psychophysical evaluation, *ACM SIGGRAPH*, ACM, pp. 38–44.
- Akyuz, A. O. & Reinhard, E. (2006). Color appearance in high-dynamic-range imaging, *Journal of Electronic Imaging* **15**(3): 033001.
- Allen, E. & Triantaphillidou, S. (2010). *The Manual of Photography*, Focal Press.
- Andersson, M., Johnsson, B., Munkberg, J., Clarberg, P., Hasselgren, J. & Akenine-Möller, T. (2011). Efficient multi-view ray tracing using edge detection and shader reuse, *The Visual Computer* **27**(6-8): 665–676.
- Ashikhmin, M. (2002). A tone mapping algorithm for high contrast images, *Proceedings of the 13th Eurographics workshop on Rendering (EGRW '02)*, pp. 145–156.
- Banterle, F., Artusi, A., Debattista, K. & Chalmers, A. (2011). *Advanced High Dynamic Range Imaging: Theory and Practice*, A K Peters/CRC Press.
- Banterle, F., Ledda, P., Debattista, K., Bloj, M., Artusi, A. & Chalmers, A. (2009). A psychophysical evaluation of inverse tone mapping techniques, *Computer Graphics Forum* **28**(1): 13–25.

- Banterle, F., Ledda, P., Debattista, K. & Chalmers, A. (2006). Inverse tone mapping, *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia - GRAPHITE '06* p. 349.
- Barlow, H. B., Blakemore, C. & Pettigrew, J. D. (1967). The Neural Mechanism of Binocular Depth Discrimination, *The Journal of Physiology* **193**(2): 327–342.
- Battiato, S., Capra, A., Curti, S. & La Cascia, M. (2004). 3D stereoscopic image pairs by depth-map generation, *Proceedings. 2nd International Symposium on 3D Data Processing, Visualization and Transmission, 3DPVT*, IEEE, pp. 124–131.
- Battiato, S., Curti, S. & Cascia, M. L. (2004). Depth map generation by image classification, *Proceedings of SPIE Three-Dimensional Image Capture and Applications VI*, pp. 95–105.
- Benzie, P., Watson, J., Surman, P., Rakkolainen, I., Hopf, K., Urey, H., Sainov, V. & von Kopylow, C. (2007). A Survey of 3DTV Displays: Techniques and Technologies, *IEEE Transactions on Circuits and Systems for Video Technology* **17**(11): 1647–1658.
- Bergen, J., Anandan, P., Hanna, K. & Hingorani, R. (1992). Hierarchical model-based motion estimation, *Computer Vision - ECCV'92 Lecture Notes in Computer Science* **588**: 237–252.
- Bhat, P., Zitnick, C., Agarwala, N., Agrawala, M., Cohen, M., Curless, B. & Kang, S. (2007). Using Photographs to Enhance Videos of a Static Scene, *Eurographics Symposium on Rendering*, pp. 327—338.
- Bloch, C. (2007). *The HDRI Handbook*, Rocky Nook.
- Boev, A., Hollosi, D. & Gotchev, A. (2008). Classification of stereoscopic artefacts, *Technical report*, MOBILE3DTV.
- Bogaert, L., Meuret, Y., Giel, B. V., Murat, H., Smet, H. D. & Thienpont, H. (2008). Projection display for the generation of two orthogonal polarized images using liquid crystal on silicon panels and light emitting diodes, *Applied Optics* **47**(10): 1535.

- Boher, P., Leroux, T. & Collomb-Patton, V. (2010). Characterization of one time-sequential stereoscopic 3d display - Part I: Temporal analysis, *Journal of Information Display* **11**(2): 57–62.
- Boitard, R., Bouatouch, K., Cozot, R., Thoreau, D. & Gruson, A. (2012). Temporal coherency for video tone mapping, *SPIE Applications of Digital Image Processing*.
- Bovik, A. (2002). A universal image quality index, *IEEE Signal Processing Letters* **9**(3): 81–84.
- Brown, M. A. & Lowe, D. G. (2003). Recognising panoramas, *Proceedings Ninth IEEE International Conference on Computer Vision*, pp. 1218–1225 vol.2.
- Brown, M. E. & Gallimore, J. J. (1995). Visualization of three-dimensional structure during computer-aided design, *International Journal of Human-Computer Interaction* **7**(1): 37–56.
- Bulbul, A., Cipiloglu, Z. & Capin, T. (2010). A perceptual approach for stereoscopic rendering optimization, *Computers & Graphics* **34**(2): 145–157.
- Chalmers, A., Bonnet, G., Banterle, F., Dubla, P., Debattista, K., Artusi, A. & Moir, C. (2009). High-dynamic-range video solution, *ACM SIGGRAPH ASIA 2009 Art Gallery & Emerging Technologies*, ACM Press, p. 71.
- Chaudhuri, S. & Rajagopalan, A. N. (1999). *Depth from Defocus: A Real Aperture Imaging Approach*, Springer.
- Chi, K., Herf, M., Shirley, P., Swamy, S., Wang, C. & K Zimmerman (1993). Spatially Nonuniform Scaling Functions for High Contrast Images, *Proceedings of Graphics Interface*.
- CompuServe Incorporated (1990). Graphics Interchange Format.
URL: <http://www.w3.org/Graphics/GIF/spec-gif89a.txt>
- Crone, R. A. (1992). The history of stereoscopy, *Documenta ophthalmologica. Advances in ophthalmology* **81**(1): 1–16.
- Cui, Y., Pagani, A. & Stricker, D. (2011). Robust Point Matching in HDRI through Estimation of Illumination Distribution, *Lecture Notes in Computer Science* **6835**: 226–235.

- Cyganek, B. & Siebert, J. P. (2009). *An Introduction to 3D Computer Vision Techniques and Algorithms*, John Wiley & Sons, Ltd, Chichester, UK.
- Daly, S. J. & Feng, X. (2003). Bit-depth extension using spatiotemporal microdither based on models of the equivalent input noise of the visual system, *Proceedings of the SPIE, Color Imaging VIII: Processing, Hardcopy, and Applications*, Vol. 5008, pp. 455–466.
- David, H. (1988). *The method of paired comparisons*, 2 edn, Charles Griffin & Company.
- Debevec, P. (2006). A median cut algorithm for light probe sampling, *ACM SIGGRAPH 2006 Courses on - SIGGRAPH '06*, ACM Press, New York, New York, USA, p. 6.
- Debevec, P. E. & Malik, J. (1997). Recovering high dynamic range radiance maps from photographs, *ACM SIGGRAPH*, ACM Press, New York, New York, USA, pp. 369–378.
- Didyk, P., Mantiuk, R., Hein, M. & Seidel, H. (2008). Enhancement of Bright Video Features for HDR Displays, *Computer Graphics Forum* **27**(4): 1265–1274.
- Dixon, S., Fitzhugh, E. & Aleva, D. (2009). Human Factors Guidelines for Applications of 3D Perspectives: A Literature Review, *SPIE Defense, Security, and Sensing*, Vol. 7327.
- Dodgson, N. A. (2004). Variation and extrema of human interpupillary distance, *Proceedings of SPIE* **5291**(January): 36–46.
- Drago, F., Myszkowski, K., Annen, T. & Chiba, N. (2003). Adaptive Logarithmic Mapping For Displaying High Contrast Scenes, *Computer Graphics Forum* **22**(3): 419–426.
- Dubois, E. (2001). A projection method to generate anaglyph stereo images, *IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings*, Vol. 3, IEEE, pp. 1661–1664.
- Duda, R., Hart, P. & Stork, D. (2001). *Pattern Classification*, 2nd edn.

- Durand, F. & Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images, *ACM Transactions on Graphics (TOG)* **21**(3): 257–266.
- Dutre, P., Bekaert, P. & Bala, K. (2006). *Advanced Global Illumination*, 2nd edn, A K Peters/CRC Press.
- Eckhardt, M., Fasel, I. & Movellan, J. (2009). Towards practical facial feature detection, *International Journal of Pattern Recognition and Artificial Intelligence* **23**(03): 379–400.
- Eisemann, E. & Durand, F. (2004). Flash photography enhancement via intrinsic relighting, *ACM SIGGRAPH 2004 Papers on - SIGGRAPH '04*, Vol. 23, ACM Press, New York, New York, USA, p. 673.
- Elad, M. (2002). On the origin of the bilateral filter and ways to improve it., *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **11**(10): 1141–51.
- Emoto, M., Niida, T. & Okano, F. (2005). Repeated Vergence Adaptation Causes the Decline of Visual Functions in Watching Stereoscopic Television, *Journal of Display Technology* **1**(2): 328–340.
- Ens, J. & Lawrence, P. (1993). An investigation of methods for determining depth from focus, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(2): 97–108.
- Es, A. & Isler, V. (2007). GPU based real time stereoscopic ray tracing, *2007 22nd International Symposium on Computer and Information Sciences*, IEEE, pp. 1–7.
- Farid, H. (2001). Blind inverse gamma correction., *IEEE transactions on image processing: a publication of the IEEE Signal Processing Society* **10**(10): 1428–33.
- Fattal, R., Lischinski, D. & Werman, M. (2002). Gradient domain high dynamic range compression, *ACM Transactions on Graphics (TOG)* **21**(3): 249–256.
- Fechner, G. T. (1838). Ueber eine Scheibe zur Erzeugung subjectiver Farben, *Annalen der Physik und Chemie* **121**(10): 227–232.

- Feng, Y., Ren, J. & Jiang, J. (2011). Object-Based 2D-to-3D Video Conversion for Effective Stereoscopic Content Generation in 3D-TV Applications, *IEEE Transactions on Broadcasting* **57**(2): 500–509.
- Ferrari, V., Megali, G., Troia, E., Pietrabissa, A. & Mosca, F. (2009). A 3-D mixed-reality system for stereoscopic visualization of medical dataset., *IEEE transactions on bio-medical engineering* **56**(11): 2627–33.
- Ferwerda, J. (2001). Elements of early vision for computer graphics, *IEEE Computer Graphics and Applications* **21**(4): 22–33.
- Ferwerda, J. A., Pattanaik, S. N., Shirley, P. & Greenberg, D. P. (1996). A model of visual adaptation for realistic image synthesis, *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*, ACM Press, New York, New York, USA, pp. 249–258.
- Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, *Communications of the ACM* **24**(6): 381–395.
- Flack, J., Harrold, J. & Woodgate, G. J. (2007). A prototype 3D mobile phone equipped with a next-generation autostereoscopic display, *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems XIV*, Vol. 6490, pp. 64900M–64900M–12.
- Gallo, O., Gelfandz, N., Tico, M. & Pulli, K. (2009). Artifact-free High Dynamic Range imaging, *2009 IEEE International Conference on Computational Photography (ICCP)*, IEEE, pp. 1–7.
- Gautama, T. & Van Hulle, M. A. (2002). A phase-based approach to the estimation of the optical flow field using spatial filtering, *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council* **13**(5): 1127–36.
- Getty, D. J. & Green, P. J. (2007). Clinical applications for stereoscopic 3-D displays, *Journal of the Society for Information Display* **15**(6): 377.
- Ghanbari, M. (2010). *Standard Codecs: Image Compression to Advanced Video Coding*, Institution of Engineering and Technology.
- Goldstein, E. B. (2009). *Sensation and Perception*, Wadsworth Publishing Company.

- Gorley, P. W. (2012). *Metrics for Stereoscopic Image Compression*, PhD thesis, Durham University.
- Goshtasby, A. & Gruver, W. A. (1993). Design of a single-lens stereo camera system, *Pattern Recognition* **26**(6): 923–937.
- Goss, D. A. & Zhai, H. (1994). Clinical and laboratory investigations of the relationship of accommodation and convergence function with refractive error, *Documenta Ophthalmologica* **86**(4): 349–380.
- Grosch, T. (2006). Fast and Robust High Dynamic Range Image Generation with Camera and Object Movement, *Vision, Modeling, and Visualization*, IOS Press, pp. 277–284.
- Grossberg, M. D. & Nayar, S. K. (2002). What Can Be Known about the Radiometric Response from Images?, *Proceedings of the 7th European Conference on Computer Vision (ECCV '02)*, pp. 189–205.
- Hartley, R. I. & Zisserman, A. (2004a). Image rectification, *Multiple View Geometry in Computer Vision*, second edn, Cambridge University Press, pp. 302–308.
- Hartley, R. I. & Zisserman, A. (2004b). *Multiple View Geometry in Computer Vision*, second edn, Cambridge University Press.
- Hasselgren, J. & Akenine-Möller, T. (2006). An Efficient Multi-View Rasterization Architecture, *Eurographics Symposium on Rendering*, pp. 61–72.
- Heinzle, S., Greisen, P., Gallup, D., Chen, C., Saner, D., Smolic, A., Burg, A., Matusik, W. & Gross, M. (2011). Computational stereo camera system with programmable control loop, *ACM Transactions on Graphics* **30**(4): 1.
- Held, R. T. & Hui, T. T. (2011). A guide to stereoscopic 3D displays in medicine., *Academic radiology* **18**(8): 1035–48.
- Henry, P., Krainin, M., Herbst, E., Ren, X. & Fox, D. (2010). RGB-D mapping: Using Kinect-style depth cameras for dense 3D modeling of indoor environments, *The International Journal of Robotics Research*, Vol. 31, pp. 22–25.
- Hiruma, N. & Fukuda, T. (1993). Accommodation Response to Binocular Stereoscopic TV Images and Their Viewing Conditions, *SMPTE Motion Imaging Journal* **102**(12): 1137–1140.

- Hoffman, D. M., Girshick, A. R., Akeley, K. & Banks, M. S. (2008). Vergence-accommodation conflicts hinder visual performance and cause visual fatigue., *Journal of vision* **8**(3): 33.1–30.
- Holliman, N. (2006). 3D display systems, *Handbook of Optoelectronics*, Taylor & Francis, chapter C2.6, pp. 1067–1101.
- Hopf, K. (2000). An autostereoscopic display providing comfortable viewing conditions and a high degree of telepresence, *IEEE Transactions on Circuits and Systems for Video Technology* **10**(3): 359–365.
- Horn, B. K. P. & Schunck, B. G. (1981). Determining optical flow, *Artificial Intelligence* **17**(1-3): 185–203.
- Howard, I. P. & Rogers, B. J. (1995). Binocular Vision and Stereopsis.
- Howard, I. & Rogers, B. (2002). *Seeing in Depth. Volume 2: Depth Perception*, I Porteous.
- Howlett, E. M. (1990). Wide-angle orthostereo, *Proc. SPIE Vol. 1256, Stereoscopic Displays and Applications*, Vol. 1256, pp. 210–223.
- Hung, G. K. (2001). *Models of Oculomotor Control*, World Scientific Publishing.
- Iddan, G. J. & Yahav, G. (2001). Three-dimensional imaging in the studio and elsewhere, *Proc. SPIE 4298, Three-Dimensional Image Capture and Applications IV*, pp. 48–55.
- IJsselsteijn, W., de Ridder, H. & Vliegen, J. (2000). Subjective evaluation of stereoscopic images: effects of camera parameters and display duration, *IEEE Transactions on Circuits and Systems for Video Technology* **10**(2): 225–233.
- IJsselsteijn, W., Seuntiens, P. & Meesters, L. (2005). Human factors of 3D displays, *3D Videocommunication: Algorithms, concepts and real-time systems in human centred communication*, John Wiley & Sons, pp. 219–234.
- Inoue, T. & Ohzu, H. (1997). Accommodative responses to stereoscopic three-dimensional display, *Applied Optics* **36**(19): 4509.
- Jacobs, K., Ward, G. & Loscos, C. (2008). Automatic High-Dynamic Range Image Generation for Dynamic Scenes, *Computer Graphics and Applications, IEEE* **28**(2): 84 – 93.

- Jaumann, R., Neukum, G., Behnke, T., Duxbury, T., Eichentopf, K., Flohrer, J., Gasselt, S., Giese, B., Gwinner, K., Hauber, E., Hoffmann, H., Hoffmeister, A., Köhler, U., Matz, K.-D., McCord, T., Mertens, V., Oberst, J., Pischel, R., Reiss, D., Ress, E., Roatsch, T., Saiger, P., Scholten, F., Schwarz, G., Stephan, K. & Wählisch, M. (2007). The high-resolution stereo camera (HRSC) experiment on Mars Express: Instrument aspects and experiment conduct from interplanetary cruise through the nominal mission, *Planetary and Space Science* **55**(7-8): 928–952.
- Jones, G. R., Lee, D., Holliman, N. S. & Ezra, D. (2001). Controlling perceived depth in stereoscopic images., *SPIE 4297, Stereoscopic Displays and Virtual Reality Systems VIII*, SPIE, pp. 42 – 53.
- Kainz, F. & Bogart, R. (2009). Technical introduction to OpenEXR, *Technical report*.
- Kalaiah, A. & Capin, T. K. (2007). A Unified Graphics Rendering Pipeline for Autostereoscopic Rendering, *2007 3DTV Conference*, IEEE, pp. 1–4.
- Kang, S. B., Uyttendaele, M., Winder, S. & Szeliski, R. (2003). High dynamic range video, *ACM Transactions on Graphics* **22**(3): 319.
- Kawakita, M., Kurita, T., Kikuchi, H. & Inoue, S. (2002). HDTV axi-vision camera, *Proc. of International Broadcasting Conference*, number 8, pp. 397–404.
- Kendall, M. & Smith, B. (1940). On the method of paired comparisons, *Biometrika* **31**(3/4): 324–345.
- Khan, E., Akyuz, A. & Reinhard, E. (2006). Ghost Removal in High Dynamic Range Images, *2006 International Conference on Image Processing*, IEEE, pp. 2005–2008.
- Khoshelham, K. (2011). Accuracy analysis of kinect depth data, *ISPRS workshop laser scanning XXXVIII*(August): 29–31.
- Kim, D., Min, D. & Sohn, K. (2008). A Stereoscopic Video Generation Method Using Stereoscopic Display Characterization and Motion Analysis, *IEEE Transactions on Broadcasting* **54**(2): 188–197.

- Koller, D. (2011). Death knell for the lecture: Technology as a passport to personalized education.
- Kolmogorov, V. (2004). *Graph based algorithms for scene reconstruction from two or more views*, PhD thesis, Cornell University.
- Kolmogorov, V. & Zabih, R. (2001). Computing visual correspondence with occlusions using graph cuts, *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001* pp. 508–515.
- Konrad, J., Lacotte, B. & Dubois, E. (2000). Cancellation of image crosstalk in time-sequential displays of stereoscopic video., *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* **9**(5): 897–908.
- Kraus, K. (2007). *Photogrammetry: Geometry from Images and Laser Scans*, Walter de Gruyter.
- Kuang, J., Johnson, G. M. & Fairchild, M. D. (2007). iCAM06: A refined image appearance model for HDR image rendering, *Journal of Visual Communication and Image Representation* **18**(5): 406–414.
- Kubota, A., Smolic, A., Magnor, M., Tanimoto, M., Chen, T. & Zhang, C. (2007). Multiview imaging and 3DTV, *IEEE Signal Processing Magazine* **24**: 10–21.
- Kumar, R., Anandan, P. & Hanna, K. (1994). Direct recovery of shape from multiple views: a parallax based approach, *Proceedings of 12th International Conference on Pattern Recognition*, Vol. 1, pp. 685–688.
- Laframboise, S., De Guise, D. & Faubert, J. (2006). Effect of Aging on Stereoscopic Interocular, *Optometry & Vision Science* **83**(8): 589–593.
- Lai, S.-H., Fu, C.-W. & Chang, S. (1992). A generalized depth estimation algorithm with a single image, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(4): 405–411.
- Lambooij, M., IJsselsteijn, W., Fortuin, M. & Heynderickx, I. (2009). Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review, *Journal of Imaging Science and Technology* **53**(3): 30201.
- Landis, H. (2002). Production-ready global illumination, *Siggraph course notes* pp. 87–101.

- Lang, M., Hornung, A., Wang, O., Poulakos, S., Smolic, A. & Gross, M. (2010). Nonlinear disparity mapping for stereoscopic 3D, *ACM SIGGRAPH 2010 papers on - SIGGRAPH '10* p. 1.
- Larson, G., Rushmeier, H. & Piatko, C. (1997). A visibility matching tone reproduction operator for high dynamic range scenes, *IEEE Transactions on Visualization and Computer Graphics* **3**(4): 291–306.
- Ledda, P., Chalmers, A., Troscianko, T. & Seetzen, H. (2005). Evaluation of tone mapping operators using a High Dynamic Range display, *ACM SIGGRAPH 2005 Papers on - SIGGRAPH '05* pp. 640–648.
- Ledda, P., Santos, L. P. & Chalmers, A. (2004). A local model of eye adaptation for high dynamic range images, *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa - AFRIGRAPH '04*, ACM Press, New York, New York, USA, p. 151.
- Ledda, P., Ward, G. & Chalmers, A. (2003). A wide field, high dynamic range, stereographic viewer, *Proceedings of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia - GRAPHITE '03*, ACM Press, New York, New York, USA, p. 237.
- Lee, C. & Kim, C. (2008). Rate-distortion optimized compression of high dynamic range videos, *Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008)*.
- Lee, C. & Kim, C.-S. (2007). Gradient Domain Tone Mapping of High Dynamic Range Videos, *2007 IEEE International Conference on Image Processing*, IEEE, pp. III – 461–III – 464.
- Lee, C. & Kim, C.-S. (2012). Rate-distortion optimized layered coding of high dynamic range videos, *Journal of Visual Communication and Image Representation* **23**(6): 908–923.
- Lee, D. & Kweon, I. (2000). A novel stereo camera system by a biprism, *IEEE Transactions on Robotics and Automation* **16**(5): 528–541.
- Lee, J.-S., Jung, Y.-Y., Kim, B.-S. & Ko, S.-J. (2001). An advanced video camera system with robust AF, AE, and AWB control, *IEEE Transactions on Consumer Electronics* **47**(3): 694–699.

- Lillesand, T. M., Kiefer, R. W. & Chipman, J. W. (2004). *Remote sensing and image interpretation.*, number Ed.5, John Wiley & Sons Ltd.
- Lin, H. & Chang, W. (2009). High dynamic range imaging for stereoscopic scene representation, *16th IEEE International Conference on Image Processing (ICIP)*, pp. 4305–4308.
- Lin, J.-R., Zhou, H. & Lai, X.-P. (2006). Application of stereoscopy on edible birds nest identification., *Journal of Chinese medicinal materials* **29**(3): 219–21.
- Lin, S. (2005). Determining the Radiometric Response Function from a Single Grayscale Image, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2, IEEE, pp. 66–73.
- Lin, S. & Yamazaki, S. (2004). Radiometric calibration from a single image, *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004.*, Vol. 2, IEEE, pp. 938–945.
- Lindberg, D. C. (1996). *Theories of Vision from Al-kindī to Kepler*, University of Chicago Press.
- Lloyd, S. (1982). Least squares quantization in PCM, *IEEE Transactions on Information Theory* **28**(2): 129–137.
- Lo, C.-H., Chu, C.-H., Debattista, K. & Chalmers, A. (2009). Selective rendering for efficient ray traced stereoscopic images, *The Visual Computer* **26**(2): 97–107.
- Lubin, J. (1995). A Visual Discrimination Model for Imaging System Design and Evaluation, *Vision Models for Target Detection and Recognition*, World Scientific, pp. 245–283.
- MacMillan, E. S., Gray, L. S. & Heron, G. (2007). Visual adaptation to interocular brightness differences induced by neutral-density filters, *Investigative ophthalmology & visual science* **48**(2): 935–42.
- Mann, S. & Picard, R. W. (1995). On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures, *IS&Ts 48th Annual Conference on Imaging on the Information Superhighway*, pp. 422–428.

- Mantiuk, R., Daly, S. J., Myszkowski, K. & Seidel, H.-P. (2005). Predicting Visible Differences in High Dynamic Range Images - Model and its Calibration, *Human Vision and Electronic Imaging X.*, Vol. 5666, pp. 204–214.
- Mantiuk, R., Efremov, A., Myszkowski, K. & Seidel, H.-p. (2006). Backward compatible high dynamic range MPEG video compression, *ACM Transactions on Graphics (TOG)* **25**(3): 713–723.
- Mantiuk, R., Krawczyk, G., Myszkowski, K. & Seidel, H.-P. (2004). Perception-motivated high dynamic range video encoding, *ACM Transactions on Graphics* **23**(3): 733.
- Mantiuk, R., Myszkowski, K. & Seidel, H.-P. (2004). Visible difference predictor for high dynamic range images, *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)*, Vol. 3, IEEE, pp. 2763–2769.
- Mantiuk, R., Myszkowski, K. & Seidel, H.-P. (2006). Lossy compression of high dynamic range images and video, *Electronic Imaging 2006*, pp. 60570V–60570V–10.
- Mantiuk, R., Tomaszewska, A. & Heidrich, W. (2009). Color correction for tone mapping, *Computer Graphics Forum* **28**(2): 193–202.
- Masia, B., Agustin, S., Fleming, R. W., Sorkine, O. & Gutierrez, D. (2009). Evaluation of reverse tone mapping through varying exposure conditions, *ACM Transactions on Graphics* **28**(5): 1.
- Max, J. (1960). Quantizing for minimum distortion, *IEEE Transactions on Information Theory* **6**(1): 7–12.
- McIntire, J. P., Havig, P. R. & Geiselman, E. E. (2012). What is 3D good for? A review of human performance on stereoscopic 3D displays, *SPIE Defense, Security, and Sensing*.
- McMillan, L. & Bishop, G. (1995). Plenoptic Modeling: An Image-Based Rendering System, *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques - SIGGRAPH '95*, ACM Press, New York, New York, USA, pp. 39–46.

- Meesters, L., IJsselsteijn, W. & Seuntiens, P. (2004). A Survey of Perceptual Evaluations and Requirements of Three-Dimensional TV, *IEEE Transactions on Circuits and Systems for Video Technology* **14**(3): 381–391.
- Mei, X., Cui, C., Sun, X., Zhou, M., Wang, Q. & Wang, H. (2011). On Building an Accurate Stereo Matching System on Graphics Hardware, *Computer Vision Workshops (ICCV Workshops)*, pp. 467—474.
- Mendiburu, B. (2009). *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*, Focal Press.
- Mendiburu, B. (2011). *3D TV and 3D Cinema: Tools and Processes for Creative Stereoscopy*, Focal Press.
- Meylan, L., Daly, S. & Süsstrunk, S. (2007). Tone mapping for high dynamic range displays, *Proceedings of the SPIE, Human Vision and Electronic Imaging XII*, Vol. 6492, pp. 649210–649210–12.
- Meylan, L. & Süsstrunk, S. (2006). The Reproduction of Specular Highlights on High Dynamic Range Displays, *In Proc. of the 14th Color Imaging Conference*, pp. 333–338.
- Millodot, M. (2008). *Dictionary of Optometry and Visual Science*, 7 edn, Butterworth-Heinemann Ltd.
- Mitsunaga, T. & Nayar, S. (1999). Radiometric self calibration, *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, pp. 374–380.
- Motra, A. & Thoma, H. (2010). An adaptive Logluv transform for High Dynamic Range video compression, *2010 IEEE International Conference on Image Processing*, pp. 2061–2064.
- Moustakas, K., Tzovaras, D. & Strintzis, M. (2005). Stereoscopic video generation based on efficient layered structure and motion estimation from a monoscopic image sequence, *IEEE Transactions on Circuits and Systems for Video Technology* **15**(8): 1065–1073.
- Munkberg, J., Clarberg, P., Hasselgren, J. & Akenine-Möller, T. (2008). Practical HDR Texture Compression, *Computer Graphics Forum* **27**(6): 1664–1676.

- Nakayama, K. & Shimojo, S. (1990). Da Vinci stereopsis: depth and subjective occluding contours from unpaired image points., *Vision research* **30**(11): 1811–1825.
- Navarro, F., Castillo, S., Serón, F. J. & Gutierrez, D. (2011). Perceptual considerations for motion blur rendering, *ACM Transactions on Applied Perception* **8**(3): 1–15.
- Nayar, S. & Branzoi, V. (2003). Adaptive dynamic range imaging: optical control of pixel exposures over space and time, *Proceedings Ninth IEEE International Conference on Computer Vision*, IEEE, pp. 1168–1175 vol.2.
- Nayar, S. & Mitsunaga, T. (2000). High dynamic range imaging: spatially varying pixel exposures, *Proceedings IEEE Conference on Computer Vision and Pattern Recognition.*, Vol. 1, IEEE Comput. Soc, pp. 472–479.
- Neukum, G. & Jaumann, R. (2004). HRSC: the High Resolution Stereo Camera of Mars Express, *Mars Express: the Scientific Payload* **1240**: 17–35.
- Nishimoto, Y. & Shirai, Y. (1987). A feature-based stereo model using small disparities, *Proc. Computer Vision and Pattern Recognition*, pp. 192—196.
- Nojiri, Y., Yamanoue, H., Hanazato, A. & Okano, F. (2003). Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV, *SPIE 5006, Stereoscopic Displays and Virtual Reality Systems X*, pp. 195–205.
- Nojiri, Y., Yamanoue, H. & Ide, S. (2006). Parallax distribution and visual comfort on stereoscopic HDTV, *IBC*, number 3, pp. 373—380.
- Ogale, A. S., Fermüller, C. & Aloimonos, Y. (2005). Motion segmentation using occlusions, *IEEE transactions on pattern analysis and machine intelligence* **27**(6): 988–92.
- Oikonomidis, I., Kyriazis, N. & Argyros, A. (2011). Efficient model-based 3D tracking of hand articulations using Kinect, *Proceedings of the British Machine Vision Conference 2011*, British Machine Vision Association, pp. 52–73.
- Okada, Y., Ukai, K., Wolffsohn, J. S., Gilmartin, B., Iijima, A. & Bando, T. (2006). Target spatial frequency determines the response to conflicting defocus-

- and convergence-driven accommodative stimuli., *Vision research* **46**(4): 475–84.
- Okino, T., Murata, H., Taima, K., Iinuma, T. & Oketani, K. (1996). New television with 2D/3D image conversion technologies, *Proc. SPIE Stereoscopic Displays and Virtual Reality Systems III*, Vol. 2653, pp. 96–103.
- Okuda, M. & Adami, N. (2007). Two-layer coding algorithm for high dynamic range images based on luminance compensation, *Journal of Visual Communication and Image Representation* **18**(5): 377–386.
- Ostrin, L. a. & Glasser, A. (2004). Accommodation measurements in a presbyopic and presbyopic population., *Journal of Cataract & Refractive Surgery* **30**(7): 1435–44.
- Pastoor, S. (1993). Human factors of 3D displays in advanced image communications, *Displays* **14**(3): 150–157.
- Pastoor, S. (1995). Human Factors in 3D Imaging: Results of Recent Research at Heinrich-Hertz-Institut Berlin, *ASIA Display*.
- Pattanaik, S. & Hughes, C. (2005). High-Dynamic-Range Still-Image Encoding in JPEG 2000, *IEEE Computer Graphics and Applications* **25**(6): 57–64.
- Pattanaik, S. N., Ferwerda, J. A., Fairchild, M. D. & Greenberg, D. P. (1998). A multiscale model of adaptation and spatial vision for realistic image display, *Proceedings of the 25th annual conference on Computer graphics and interactive techniques - SIGGRAPH '98*, ACM Press, New York, New York, USA, pp. 287–298.
- Pattanaik, S. N., Tumblin, J., Yee, H. & Greenberg, D. P. (2000). Time-dependent visual adaptation for fast realistic image display, *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*, ACM Press, New York, New York, USA, pp. 47–54.
- Peaeson, E. & Haetlet, H. (1976). *Biometrika tables for statisticians*, 2 edn, Biometrika Trust.
- Pourazad, M., Nasiopoulos, P. & Ward, R. (2009). An H.264-based scheme for 2D to 3D video conversion, *IEEE Transactions on Consumer Electronics* **55**(2): 742–748.

- Quevedo, R. & Aguilera, J. M. (2008). Computer Vision and Stereoscopy for Estimating Firmness in the Salmon (*Salmon salar*) Fillets, *Food and Bioprocess Technology* **3**(4): 561–567.
- Reinhard, E., Heidrich, W., Pattanaik, S., Debevec, P., Ward, G. & Myszkowski, K. (2010). *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*, Morgan Kaufmann.
- Reinhard, E., Stark, M., Shirley, P. & Ferwerda, J. (2002). Photographic tone reproduction for digital images, *ACM Transactions on Graphics* **21**(3): 267–276.
- Rempel, A., Trentacoste, M. & Seetzen, H. (2007). Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs, *ACM Transactions on Graphics (TOG)* **26**(3): 39.
- Richards, W. (1970). Stereopsis and stereoblindness, *Experimental Brain Research* **10**(4): 380–388.
- Richardson, I. E. (2010). *The H.264 Advanced Video Compression Standard*, Wiley-Blackwell.
- Roelofs, G. (1999). *PNG: The definitive guide*, O'Reilly Media.
- Rollmann, W. (1853). Zwei neue stereoskopische Methoden, *Annalen der Physik und Chemie* **166**(9): 186–187.
- Rubinstein, M., Gutierrez, D., Sorkine, O. & Shamir, A. (2010). A comparative study of image retargeting, *ACM Transactions on Graphics* **29**(6): 1.
- Ruppertsberg, A., Bloj, M., Banterle, F. & Chalmers, A. (2007). Displaying colourimetrically calibrated images on a high dynamic range display, *Journal of Visual Communication and Image Representation* **18**(5): 429–438.
- Sawhney, H., Guo, Y., Hanna, K., Kumar, R., Adkins, S. & Zhou, S. (2001). Hybrid stereo camera: an IBR approach for synthesis of very high resolution stereoscopic image sequences, *Proceedings of the 28th annual conference on computer graphics and interactive techniques*, ACM, pp. 451–460.
- Saxena, A., Chung, S. H. & Ng, A. Y. (2007). 3-D Depth Reconstruction from a Single Still Image, *International Journal of Computer Vision* **76**(1): 53–69.

- Scharstein, D. & Szeliski, R. (2003). High-accuracy stereo depth maps using structured light, *Proceedings of 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. I–195 – I–202.
- Scharstein, D., Szeliski, R. & Zabih, R. (2001). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms, *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision* (1): 131–140.
- Schlick, C. (1994). Quantization Techniques for Visualization of High Dynamic Range Pictures, *Proceedings of the Fifth Eurographics Workshop on Rendering*, pp. 7–18.
- Schor, C., Wood, I. & Ogawa, J. (1984). Binocular sensory fusion is limited by spatial resolution, *Vision research* **21**(7): 661–665.
- Se, S., Lowe, D. & Little, J. (2002). Mobile robot localization and mapping with uncertainty using scale-invariant visual landmarks, *The International Journal of Robotics Research* **21**(8): 735–758.
- Seetzen, H., Heidrich, W., Stuerzlinger, W., Ward, G., Whitehead, L., Trentacoste, M., Ghosh, A. & Vorozcova, A. (2004). High dynamic range display systems, *ACM Transactions on Graphics (TOG)* **23**(3): 760—768.
- Sen, P., Kalantari, N. K., Yaesoubi, M., Darabi, S., Goldman, D. B. & Shechtman, E. (2012). Robust patch-based hdr reconstruction of dynamic scenes, *ACM Transactions on Graphics* **31**(6): 1.
- Seuntiens, P., Meesters, L. & IJsselsteijn, W. (2005). Perceptual attributes of crosstalk in 3D images, *Displays* **26**(4-5): 177–183.
- Sexton, I., Bates, R., Hopf, K. & Lee, W. (2006). Head tracked 3D displays, *Lecture Notes in Computer Science* **4105**: 769–776.
- Siragusa, J., Swift, D. C., Akka, B., Milici, D. & Spencer, A. (1997). General Purpose Stereoscopic Data Descriptor, *Technical report*, VRex, Inc.
- Smith, K., Krawczyk, G., Myszkowski, K. & Seidel, H.-P. (2006). Beyond Tone Mapping: Enhanced Depiction of Tone Mapped HDR Images, *Computer Graphics Forum* **25**(3): 427–438.

- Snavely, N., Seitz, S. & Szeliski, R. (2006). Photo tourism: exploring photo collections in 3D, *ACM Transactions on Graphics (TOG)*, Vol. 25, ACM, pp. 835–846.
- Speranza, F., Tam, W., Renaud, R. & Hur, N. (2006). Effect of disparity and motion on visual comfort of stereoscopic images, *Proc. SPIE* **6055**.
- Stelmach, L., Tam, W. J., Meegan, D. & Vincent, A. (2000). Stereo image quality: effects of mixed spatio-temporal resolution, *IEEE Transactions on Circuits and Systems for Video Technology* **10**(2): 188–193.
- Sun, J., Zheng, N.-N. & Shum, H.-Y. (2003). Stereo matching using belief propagation, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(7): 787–800.
- Tam, W. J., Speranza, F., Yano, S., Shimono, K. & Ono, H. (2011). Stereoscopic 3D-TV: Visual Comfort, *IEEE Transactions on Broadcasting* **57**(2): 335–346.
- Tam, W. J., Yee, A. S., Ferreira, J., Tariq, S. & Speranza, F. (2005). Stereoscopic image rendering based on depth maps created from blur and edge information, *Proceedings of SPIE Stereoscopic Displays and Virtual Reality Systems XII*, Vol. 5664, SPIE, pp. 104–115.
- Teoh, W. & Zhang, X. (1984). An inexpensive stereoscopic vision system for robots, *Proceedings. 1984 IEEE International Conference on Robotics and Automation*, Vol. 1, Institute of Electrical and Electronics Engineers, pp. 186–189.
- Tocci, M., Kiser, C., Tocci, N. & Sen, P. (2011). A Versatile HDR Video Production System, *ACM Transactions on Graphics (TOG)* **30**(4): 41–49.
- Tumblin, J., Hodgins, J. K. & Guenter, B. K. (1999). Two methods for display of high contrast images, *ACM Transactions on Graphics* **18**(1): 56–94.
- Ukai, K. & Howarth, P. A. (2008). Visual fatigue caused by viewing stereoscopic motion images: Background, theories, and observations, *Displays* **29**(2): 106–116.
- Ukai, K. & Kato, Y. (2002). The use of video refraction to measure the dynamic properties of the near triad in observers of a 3-D display, *Ophthalmic and Physiological Optics* **22**(5): 385–388.

- Urey, H., Chellappan, K. V., Erden, E. & Surman, P. (2011). State of the Art in Stereoscopic and Autostereoscopic Displays, *Proceedings of the IEEE* **99**(4): 540–555.
- Valencia, S. A. & Rodriguez-Dagnino, R. M. (2003). Synthesizing stereo 3D views from focus cues in monoscopic 2D images, *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems X*, pp. 377–388.
- Wade, N. (1987). On the late invention of the stereoscope, *Perception* **16**(6): 785—818.
- Wang, H., Raskar, R. & Ahuja, N. (2005). High dynamic range video using split aperture camera, *IEEE 6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, p. Cited by: 17.
- Wang, L., Jin, H., Yang, R. & Gong, M. (2008). Stereoscopic inpainting: Joint color and depth completion from stereo images, *2008 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp. 1–8.
- Wang, L., Wei, L. & Zhou, K. (2007). High Dynamic Range Image Hallucination, *Eurographics Symposium on Rendering*, p. 72.
- Ward, G. (1994a). A contrast-based scalefactor for luminance display, *Graphics gems IV*, Academic Press Professional, Inc., pp. 415–421.
- Ward, G. (1994b). The RADIANCE lighting simulation and rendering system, *Proceedings of the 21st annual conference on Computer graphics and interactive techniques - SIGGRAPH '94*, ACM Press, New York, New York, USA, pp. 459–472.
- Ward, G. (1998). Overcoming gamut and dynamic range limitations in digital images, *6th Color Imaging Conference - Color Science, Systems and Applications*, pp. 214 – 219.
- Ward, G. (2002). A wide field, high dynamic range, stereographic viewer, *Journal of Vision* **2**(10): 2–2.
- Ward, G. (2003). Fast, Robust Image Registration for Compositing High Dynamic Range Photographs from Hand-Held Exposures, *Journal of Graphics Tools* **8**(2): 17–30.

- Ward, G. (2005). JPEG-HDR: A backwards-compatible, high dynamic range extension to JPEG, *ACM SIGGRAPH 2005 Courses*, p. 8.
- Ward, G., Rushmeier, H. & Piatko, C. (1997). A visibility matching tone reproduction operator for high dynamic range scenes, *IEEE Transactions on Visualization and Computer Graphics* **3**(4): 291–306.
- Ward, G. & Simmons, M. (2004). Subband encoding of high dynamic range imagery, *Proceedings of the 1st Symposium on Applied perception in graphics and visualization - APGV '04*, ACM Press, New York, New York, USA, p. 83.
- Weber, E. H. (1834). *De Pulsu, resorptione, auditu et tactu: Annotationes anatomicae et physiologicae*.
- Wheatstone, C. (1838). Contributions to the physiology of vision.—Part the first. On some remarkable, and hitherto unobserved, phenomena of binocular vision, *Philosophical transactions of the Royal Society of London* **128**(1838): 371–394.
- Whitehill, J., Littlewort, G., Fasel, I., Bartlett, M. & Movellan, J. (2009). Toward practical smile detection, *IEEE transactions on pattern analysis and machine intelligence* **31**(11): 2106–11.
- Wilburn, B., Joshi, N., Vaish, V., Talvala, E.-V., Antunez, E., Barth, A., Adams, A., Horowitz, M. & Levoy, M. (2005). High Performance Imaging Using Large Camera Arrays, *ACM Transactions on Graphics* **24**(3): 765–776.
- Woods, A. (2010). Understanding Crosstalk in Stereoscopic Displays, (*Keynote presentation*), *3DSA*, number May, pp. 19–21.
- Woods, A., Docherty, T. & Koch, R. (1993). Image Distortions in Stereoscopic Video Systems, *SPIE's Symposium on Electronic Imaging: Science and Technology*, Vol. 1915, pp. 36—48.
- Wopking, M. (1995). Viewing comfort with stereoscopic pictures: An experimental study on the subjective effects of disparity magnitude and depth of focus, *Journal of the Society for Information Display* **3**(3): 101.
- Wu, Y.-J., Jeng, Y.-S., Yeh, P.-C., Hu, C.-J. & Huang, W.-M. (2008). 20.2: Stereoscopic 3D Display Using Patterned Retarder, *SID Symposium Digest of Technical Papers* **39**(1): 260–263.

- Wyszecki, G. & Stiles, W. S. (2000). *Color Science: Concepts and Methods, Quantitative Data and Formulae*, Wiley-Interscience.
- Xiao, Y. & Lim, K. B. (2007). A prism-based single-lens stereovision system: From trinocular to multi-ocular, *Image and Vision Computing* **25**(11): 1725–1736.
- Yamanoue, H., Okui, M. & Okano, F. (2006). Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images, *IEEE Transactions on Circuits and Systems for Video Technology* **16**(6): 744–752.
- Yang, X., Zhang, L., Wong, T.-T. & Heng, P.-A. (2012). Binocular tone mapping, *ACM Transactions on Graphics* **31**(4): 1–10.
- Yano, S., Ide, S., Mitsuhashi, T. & Thwaites, H. (2002). A study of visual fatigue and visual comfort for 3D HDTV/HDTV images, *Displays* **23**(4): 191–201.
- Yeh, Y.-Y. & Silverstein, L. D. (1990). Limits of Fusion and Depth Judgment in Stereoscopic Color Displays, *Human Factors: The Journal of the Human Factors and Ergonomics Society* **32**(1): 45–60.
- Yokota, A., Yoshida, T., Kashiyama, H. & Hamamoto, T. (2005). High-speed Sensing System for Depth Estimation Based on Depth-from-Focus by Using Smart Imager, *IEEE International Symposium on Circuits and Systems*, IEEE, pp. 564–567.
- Zhang, L., Lawrence, B., Wang, D. & Vincent, A. (2005). Comparison study on feature matching and block matching for automatic 2D to 3D Video Conversion, *Visual Media Production, 2005. CVMP, The 2nd IEE European Conference on*, pp. 122–129.
- Zhang, Y., Reinhard, E. & Bull, D. (2011). Perception-based high dynamic range video compression with optimal bit-depth transformation, *8th IEEE International Conference on Image Processing*, pp. 1321–1324.
- Zitnick, C. L. & Kang, S. B. (2007). Stereo for Image-Based Rendering using Image Over-Segmentation, *International Journal of Computer Vision* **75**(1): 49–65.