# Eulerian-Lagrangian Stereo-to-Multi-view Conversion



by

Szu-Po Wang

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Masters of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2016



Signature redacted

Author .....
Department of Electrical Engineering and Computer Science
January 29, 2016

Signature redacted

Certified by .....
Wojciech Matusik
Associate Professor
Thesis Supervisor
January 29, 2016

Signature redacted

Accepted by .....
Dr. Christopher Terman
Chairman, Master of Engineering Thesis Committee

# Eulerian-Lagrangian Stereo-to-Multi-view Conversion

by

Szu-Po Wang

Submitted to the Department of Electrical Engineering and Computer Science
on January 29, 2016, in partial fulfillment of the
requirements for the degree of
Masters of Engineering in Electrical Engineering and Computer Science

## Abstract

In contrast to the popularity of stereoscopic 3D (S3D) movies in movie theaters, the adoption of S3D at home is low. It is widely believed that watching S3D with glasses is not the pratical approach for a home setting. A much more appealing approach is to use automultiscopic displays that provide a glasses-free 3D experience to multiple viewers. The main technical challenge that hampers the adaptation of this technology is the lack of multi-view content. We develop a real-time system that converts stereoscopic video to a high-quality, multi-view video, which can be directly fed to automultiscopic displays. Our algorithm uses a wavelet-based decomposition of stereoscopic images with per-wavelet disparity estimation. One key contribution lies in combining Lagrangian and Eulerian approaches. This allows us to leverage their complementary advantages. The method achieves real-time performance on current GPUs. Its design also enables an easy hardware implementation, demonstrated using a field-programmable gate array.

Thesis Supervisor: Wojciech Matusik
Title: Associate Professor

# Acknowledgments

I would like to thank Professor Wojciech Matusik, my research supervisor, for the opportunity to work on this exciting project. His patient guidance has been the greatest support throughout my time at the Computational Fabrication Group.

I would like to thank Professor Frédo Durand and Professor William T. Freeman, for their insightful advice and generous help on the direction of this research project. I would also like to thank Professor Piotr Didyk for advising me with the development of the view expansion algorithm, and Professor Pitchaya Sitthi-amorn for working with me on the hardware implementations.

Furthermore, I would like to thank the members of the Computational Fabrication Group. The fun and resourceful research environment has made my learning experience highly enjoyable.

Finally, I wish to thank my friends and my family for their support and encouragement throughout my study.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction

Stereoscopic 3D (S3D) has become much more popular during the last decade. Today, almost all movie blockbusters are also released in a stereo format. However, the popularity of S3D in the movie theaters has not translated to the equivalent popularity at homes. Despite the fact that most current TV sets support 3D glasses and the content providers offer streaming stereoscopic content, the adoption of S3D at home remains very low. It is widely believed that the use of stereoscopic glasses is not practical in a home setting [3].

We believe that the right approach to 3D viewing at home is to use automultiscopic displays that provide a glasses-free, 3D stereoscopic experience to multiple viewers in the same space. These displays are rapidly improving due to the industry drive for a higher and higher display resolution (e.g., even current 4K UHD displays can be easily converted to a 3D automultiscopic display with 8 views and an HD spatial resolution). However, using these displays presents one fundamental challenge – while there is plenty of stereoscopic content available, there is practically no multi-view content for automultiscopic displays. Therefore, there is a clear need for methods and systems that can convert streaming, high-resolution, stereoscopic video available from the standard delivery channels to a high-quality, multi-view content in real-time. Furthermore, the methods should be amenable to hardware implementations such that they can be incorporated in future streaming TV devices and smart TV-sets. Finally, the systems should support some customization of the 3D video – viewers

desire a different amount of the 3D experience.

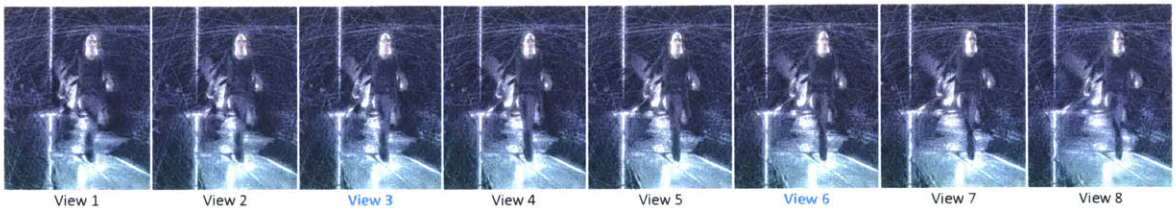


Figure 1-1: The figure presents multi-view content (8 views) generated from stereoscopic input (the views marked in blue) using our method. The view synthesis and inter-view aliasing are performed in real-time providing high-quality content. The images can be viewed in stereo using cross fusion. Scene copyright: Blender foundation (https://orange.blender.org/)

We propose a method that addresses all of the requirements we have outlined. Our method works with existing stereoscopic content expanding it to a high-resolution, high-quality, multi-view format in real-time (Figure 1-1). Furthermore, the method can be implemented efficiently in hardware and it naturally supports retargeting of the disparities. Our technique is inspired by the recent advances in phase-based approaches [9, 10, 32, 5]. Such methods use an Eulerian approach for novel view synthesis, which provides robustness, enables inter-view anti-aliasing at almost no cost, and allows for simple disparity manipulation. However, in contrast to standard depth image-based rendering methods [25], such techniques are limited to small disparities. Inspired by the work in the physics-based simulation [7], we overcome this problem by combining a phase-based approach (an Eulerian method) with standard depth image-based rendering (a Lagrangian approach). Our technique starts by decomposing the input signal using the steerable pyramid [29, 28]. The basis functions of this transform resemble Gabor-like wavelets, therefore, for the purpose of this paper we will refer to them as wavelets. Next, depth information is estimated for each wavelet separately using a combination of standard disparity estimation and phase-based measure. To synthesize new views, our method applies a wavelet re-projection which moves wavelets according to their disparities. Such an approach allows us to both handle large disparities and preserve all the advantages of the Eulerian approach [5]. Furthermore, our technique can perform nonlinear depth manipulations [14]. We demonstrate that our method can provide real-time performance both on a GPU and a field-programmable gate array (FPGA). We evaluate our method on a variety of

stereoscopic test scenes and Hollywood movies.

## 1.1 Contributions

The contributions of this thesis include:

(1) An efficient algorithm to expand existing stereoscopic content to a multi-view format in high quality, suitable for high-resolution automultiscopic displays. The core of the algorithm lies in a novel combination of Lagrangian and Eulerian approaches, which allows us to leverage their complementary advantages to handle challenging cases such as large disparity, motion blur, and complex light effects.

(2) Integration of the proposed view expansion algorithm with inter-view antialiasing and non-linear depth manipulation, which are crucial steps for delivering a good and customizable 3D viewing experience.

(3) An extensive evaluation of the proposed algorithm, including comparisons with existing Lagrangian and Eulerian methods, comparisons with ground-truth views from lightfield datasets, and a detailed evaluation on the challenging cases.

(4) Real-time hardware implementations on both a GPU and an FPGA, demonstrating the efficiency and the ease of hardware-adoption of the proposed algorithm. We believe that this would be the first step toward a glasses-free 3D experience at home.

## 1.2 Thesis Outline

This thesis is structured as follows. Chapter 2 reviews state-of-the-art approaches to generate multi-view content for automultiscopic displays. Chapter 3 describes the steps of the proposed view expansion algorithm based on a per-wavelet disparity representation. Chapter 4 discusses the two real-time hardware implementations of our algorithm, ultilizing a GPU and an FPGA. Chapter 5 compares the image quality of the synthesized views and the depth reproduction with existing Lagrangian and Eulerian approaches. Chapter 6 outlines the strengths and the limitations of our

system, and Chapter 7 concludes the thesis.

# Chapter 2

# Previous Work

In contrast to standard displays, automultiscopic displays enable reproduction of view dependent effects. Although the advantages of such systems are numerous, a big and unsolved challenge is the content production for such screen.

## 2.1 Multi-view Content Capture

A straightforward method to acquire content for automultiscopic displays is to capture it. This requires recording the scene with multiple cameras positioned at different locations, usually arranged in a camera array setup, capturing directly the views that are later displayed on the screen [21]. More advanced solutions enable acquiring over hundreds of views [35, 34]. Camera arrays, regardless of the number of cameras, are expensive and hard to manage due to their size. This makes this technology impractical from the commercial point of view. An alternative to these setups are small compact lighfield cameras [24, 17]. In order to provide multiple views, they use a lenslet array placed in front of their sensors to trade spatial for angular resolution. More recently Marwah et al. [19] proposed an improved architecture where the loss of spatial resolution is minimized. Although the image quality provided by these setups is an issue, the main limitation is the small amount of parallax that is supported. Because these devices are very compact, recorded views correspond to very close locations. This is insufficient to create true parallax and 3D effects on automultiscopic

displays.

## 2.2 Image-based Rendering

If multi-view content cannot be captured directly, it is possible to use software techniques, so-called image-based rendering, to synthesize additional views from a stereoscopic input. These techniques can be categorized into Lagrangian, which explicitly recover depth information to synthesize novel views, and Eulerian, which rely on local content changes that are amplified to obtain new views.

### 2.2.1 Lagrangian Techniques

Lagrangian techniques recover depth information first [1], and then use re-projection [18] to create novel views [30]. Using such approach, Rietchert et al. [25] built a system for real-time stereo to multi-view conversion. Although many sophisticated techniques for depth estimation have been proposed, this is still a challenging problem especially in the case of real-time applications. Lagrangian methods still suffer from low quality depth maps. This problem, to some extent, is overcome when sparse depth informtion is used together with an image warping technique [8, 31]. Such methods have an additional advantage as they do not need to deal with missing information in disocclusion regions. However, this comes at the price of poor depth quality at sharp depth discontinuities and in regions with fine depth details. The resolution of the mesh that is used for synthesizing novel views is too coarse to handle such cases. Furthermore, Lagrangian approaches rely explicitly on per-pixel depth information, which is insufficient when the depth cannot be uniquely defined. Examples include motion blur, depth-of-field effects, and transparencies, which commonly appear in the case of movie content. Hardware implementations were proposed for both depth-based rendering and image warping. Work by Rietchert et al. [25], Liao et al. [15] also presented a hardware implementation for a depth based approach, and more recently, Schaffner et al. [26] presented an implementation of warping based technique.

### 2.2.2 Eulerian Techniques

Eulerian techniques estimate local changes using local phase information, as opposed to recovering depth information explicitly. This is inspired by work on both disparity and optical-flow estimation using local phase information [9, 10]. Recently, Eulerian approaches have shown a step towards more robust view synthesis [32, 5]. There are two main advantages of these techniques. First, they do not rely on per-pixel information. Instead, the phase information being used is defined separately for each location and frequency. This turns out to be a great advantage over Lagrangian approaches in difficult cases where per-pixel depth cannot be reliably estimated [5]. Second, the phase-based disparity provides an accurate, subpixel estimation, as opposed to Lagrangian techniques which provide discretized disparity information. Phase-based techniques, however, have one significant limitation: the disparity/depth range that they can deal with is relatively low. Although there exist multiscale phase-based disparity estimation techniques which extend the supported disparity range [23, 37], their goal is to estimate per-pixel disparity estimation. Instead, we address the problem of limited disparity support by combining a phase-based technique with Lagrangian approach which pre-aligns views to reduce disparity so that the Eulerian approach can be applied. In this regard, the most similar work to ours is the technique proposed by Zhang et al. [36], which addresses the problem of reconstructing light field from a micro-baseline image pair. Similarly to our work, they also rely both on disparity and phase information. However, in contrast to their view synthesis method which relies on per-pixel disparity information, we use a concept of per-wavelet disparity which provides much richer representation. Another difference is that we propose a real-time solution which is capable of performing the stereo-to-multi-view conversion on-the-fly. To our knowledge there were no attempts of designing hardware implementation of Eulerian techniques for view expansion.

## 2.3 Post-processing Methods

Capturing or computing high-quality views is not sufficient to assure perfect viewing quality. Due to the limited angular resolution of automultiscopic screens, additional post-processing steps are usually required. First, a proper filtering needs to be applied to avoid inter-view aliasing. In this context, Zwicker et al. [38] presented a technique that accounts for the aliasing during the registration and display. They provide a filtering approach that removes both these artifacts. It has also been demonstrated that inter-view antialiasing can be easily incorporated into Eularian view synthesis [5]. Automultiscopic displays also cannot reproduce all visual cues, in particular accommodation whose lack can lead to visual discomfort [27]. To overcome this limitation, depth presented on such a display needs to be carefully managed [5, 20, 2]. Using post-processing techniques, it is also possible to reduce hotspotting, which is an intrinsic artifact of automultiscopic displays [6].

# Chapter 3

# Algorithm

In this section, we describe a new method for expanding stereoscopic content to its multi-view version which can be later displayed on an automultiscopic screen. The core of our method is a wavelet representation with per-wavelet disparity estimation. We also present a new image-based rendering approach designed for such a decomposition. Our method takes as an input the left and the right views together with a corresponding pair of disparity maps. The main advantage of our approach is that those maps can be of low quality and therefore computed very efficiently. In this work, the real-time approach proposed in [13] is used. Additionally, we assume that the input stereoscopic images are rectified [11].

Our technique consists of two main steps. First (Section 3.1), we analyze the stereoscopic content, decompose it into wavelet representation and compute per-wavelet disparity information. In this step, we compensate for the low quality of input disparity map by incorporating local phase information. Next (Section 3.2), we use the estimated per-wavelet disparity information to synthesize novel views. Additionally (Section 3.3), we discuss additional processing steps of inter-view antialiasing and nonlinear depth manipulations, which are essential to provide a good viewing experience. In our work, we consider only displays that reproduce horizontal parallax.

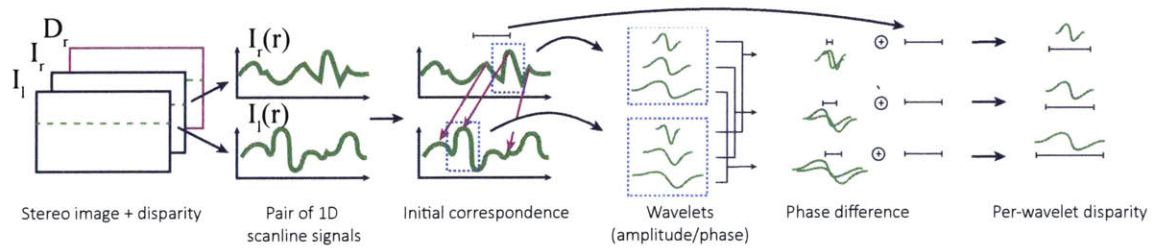


Figure 3-1: In contrast to most approaches where per-pixel disparity is estimated, in our method we consider a wavelet as a basic element of a picture, and estimate disparity for each of them separately. To this end (from left to right), we start with a stereoscopic image pair and consider each scanline of the stereoscopic image independently. We first decompose the scanlines into wavelets. Next, we find the initial correspondence between wavelets from the left view and the wavelets from the right views based on the input disparity maps. The difference in the positions of the corresponding wavelets defines the initial disparity information. To further refine it, the phase difference of the corresponding wavelets is computed and combined with the initial disparity estimation .

## 3.1 Per-wavelet Disparity Estimation

Disparity is an important cue to synthesize novel views. For stereoscopic content, disparity maps ($D_r$ and $D_r$) encode the correspondence between left and right views ($L$ and $R$). More formally, if for a given position in the world space, its projections into the left and the right views are $\mathbf{x}_l$ and $\mathbf{x}_r$, the disparity is defined as the distance between those locations in the screen space. A signed distance is considered to distinguish between locations in front and behind the zero disparity plane. For rectified views disparity maps represent a horizontal translation and can be defined as follows: $D_l(\mathbf{x}_l) = \mathbf{x}_{lx} - \mathbf{x}_{rx}$ and $D_r(\mathbf{x}_r) = \mathbf{x}_{rx} - \mathbf{x}_{lx}$, where $\mathbf{x}_{lx}$ and $\mathbf{x}_{rx}$ denote the horizontal components of $\mathbf{x}_l$ and $\mathbf{x}_r$.

In contrast to previous approaches, we consider per-wavelet, instead of per-pixel, disparity. This allows us to use phase information to improve the quality of the estimates and overcome limitations of previous Lagrangian and Eulerian approaches. To compute per-wavelet disparities, we first decompose the input images into wavelet representations. Then, for each wavelet, the initial disparity is computed from the input disparity maps. In the next step, this information is refined by additionally considering local phase information. The whole process is depicted in Figure 3-1.

Our disparity information is not a single disparity map. Instead, we obtain one disparity map for each pyramid level.

### 3.1.1 Wavelet Decomposition

As both input views are rectified, we can limit our analysis to scanlines. Here, we consider each pair of corresponding scanlines (1D signals) of the right and left views separately, and denote them by $I_r$ and $I_l$. We decompose each of these signals into wavelets. To this end, we use a steerable pyramid [29, 28], which originally decomposes a 2D signal according to spatial scale, orientation and position. In our case, as we deal with 1D signals, we do not obtain orientation differentiation, but we used this decomposition to obtain local amplitude and phase information, which we exploit in our technique.

Each of the scanlines ($I_r$ and $I_l$) is separately represented as a sum of bases functions, wavelets $b_f$, which are similar to Gabor wavelets, i. e., they are windowed sinusoids. We use a steerable filter set $b_f$, where $f \in \mathcal{F}$ specifies the central frequency of the filter. The frequency response of each filter is:

$$\hat{b}_f(\xi) = \cos(\frac{\pi}{2} \log_w(\xi/f)) \cdot \Pi(\frac{1}{2} \log_w(\xi/f)),$$

where $\Pi$ is a rectangular function and $w$ defines the width of filters – it is equal to the ratio of central frequencies of neighboring levels. In this work, we perform an octave decomposition and use $w = 2$. An additional low-pass filter $b_0$ collects the residual low-frequency components. The frequency responses correspond to cosine functions in the log-frequency domain(Figure 3-2).

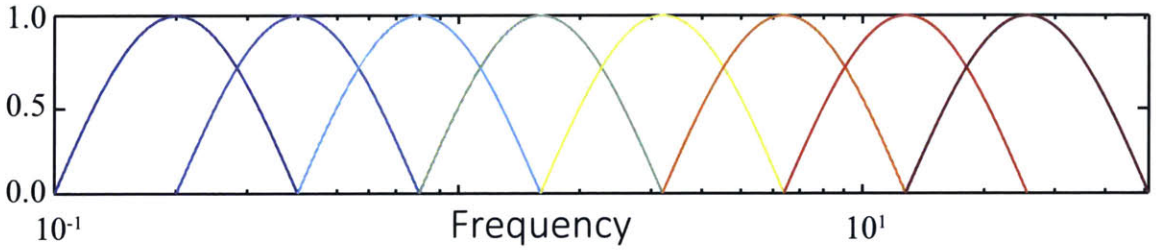


Figure 3-2: Filter bank used for wavelet decomposition

Using such a filter bank, we can compute a single wavelet coefficient for a given location $x$ and frequency $f$ as

$$A_{fx} = (b_f * I)(x),$$

where $*$ denotes a convolution. $A_{fx}$ is a complex number which contains information about local phase and amplitude. After the signal is decomposed into wavelets we can obtain the original signal as:

$$I = 2\,\mathrm{Re} \sum_{f \in \mathcal{F}} \frac{length(I)}{|X_f|} \sum_{x \in X_f} A_{fx} b_f(t - x),$$

where $\mathcal{F}$ is a discrete set of frequencies and $X_f$ is a set of locations at which wavelets are defined. The factor 2 compensates here for the fact that the complex wavelets are obtained only from positive frequency components, and factor $length(I)/|X_f|$ is necessary to compensate for the energy loss due to only $|X_f|$ wavelets representing the signal at each frequency. To prevent aliasing, $X_f$ is chosen to evenly sample the range with at least $2fw$ samples.

### 3.1.2 Initial Wavelet Disparity

After decomposing $I_r$ and $I_l$ into wavelets, we establish a correspondence between them using input disparity maps ($D_r$ and $D_l$). More precisely, for each wavelet $\psi_{rfx}$ from $I_r$ we find the corresponding wavelet $\psi_{lfx'}$ from $I_l$. To this end, for each wavelet $\psi_{rfx}$ we compute a local disparity from $D_r$. Because the signal of each wavelet covers a certain spatial extent, there is no direct correspondence between wavelets and disparity values in $D_r$. To account for this, we compute the disparity of each wavelet as an average of disparity in its local neighborhood. The size of this neighborhood is equal to the spacing between wavelet positions. Formally, the computed disparity for wavelet $\psi_{rfx}$ can be defined as:

$$d_{rfx} = \sum_{x \in S} D_r(x) \,/\, |\, S \,|, \quad S = \{x' \in \mathbb{R} \;:\; |x' - x| \leq |\, I_r \,| \,/\, |\, X_f \,|\}.$$

$\psi_{lfx'}$ is then found as the closest wavelet to the location $x - d_{rfx}$. We perform the same step for all wavelets from $I_l$.

### 3.1.3 Wavelet Disparity Refinement

The disparity information encoded in the correspondence between wavelets is often inaccurate due to insufficient quality of the input disparity maps. However, our observation is that the initial correspondence serves as a good pre-alignment of the left and the right view and therefore the residual disparity which is not captured is usually small. Recently, it has been demonstrated that such small displacements are well captured by phase information which can be used for novel view synthesis [32, 5] Inspired by this, we further improve our initial disparity using phase information of the individual wavelets. To this end, we compute a phase difference between each pair of corresponding wavelets:

$$\Delta\varphi = \arg(A_{rfx}) - \arg(A_{lfx'}).$$

Next, for the wavelets centered around frequency $f$, the phase difference can be transformed into disparity by dividing it by $f/2\pi$. As mentioned before, after our initial pre-alignment the remaining phase difference encodes residual disparity which was not captured by the initial correspondence. Hence, we can improve the disparity information for each wavelet $\psi_{rfx}$ by adding $0.5\,\Delta\varphi \cdot f/\pi$ to the previously estimated disparity $x - x'$. For color images, we compute phase differences for each channel separately, and combine them using a weighted sum to get the disparity refinement. The weights are proportional to the wavelet amplitudes to penalize phase for weak signals that can be poorly estimated.

Our per-wavelet disparity estimation is performed on individual 1D scanlines of the stereo image pair. This may lead to inconsistent disparity values in neighboring rows. Therefore, we apply an additional filtering to improve the coherence of the disparity estimation. More precisely, we filter the per-wavelet disparity using 2D median filter with a kernel size equal to twice the spacing between wavelets. Additionally, to

penalize large phase differences between wavelets, we weight the contribution of each wavelet using a Gaussian function defined on the phase differences with $\sigma = \pi/4$.

As a result of this step we obtain an accurate disparity estimation for each wavelet. In particular, in contrast to standard depth-based rendering where there is only one depth value per pixel, we obtain information for different frequencies separately. As we show later (Section 5), this allows us to handle difficult cases when we use this information for rendering novel views. In the remainder of this section we will demonstrate how our per-wavelet disparity information can be used for efficient view synthesis.

## 3.2 Novel Views Reconstruction

To reconstruct novel views, we use a technique similar to pixel re-projection [18]. However, in our case we deal with wavelet decomposition and instead of reprojecting pixels according to underlying disparities, we reproject entire wavelets.

In order to compute novel views, we modify the position of each wavelet. We perform this operation for each scanline separately. The new position for each wavelet $\psi$ at location $x$ and disparity $d$ is computed as $x + a \cdot d$, where parameter $a$ directly controls the new viewing position. After the position of each wavelet is updated, we convert the displaced wavelets back into uniform-spaced samples using non-uniform Fourier transform as described in [16]. The non-uniform Fourier transform process utilizes an oversampled grid with a oversampling factor $m = 2$. Each displaced wavelet is approximated as a weighted sum of $q = 4$ nearby samples on the oversampled grid, where the weights depend on the fractional residual in the displaced location. After the contributions from all wavelets are summed, a low-pass filter is used to downsample back into the original grid. We refer to the original paper for more details. After the wavelets are converted back to the original uniform grid, we can reconstruct the 1D signal using pyramid reconstruction. For lowest frequency wavelets corresponding to filter $b_0$, a linear interpolation of the wavelet values on the uniform grid is used. This is for preventing low-passed wavelets from accumulating and creating color bands.

### 3.2.1 Occlusions and Disocclusions

There might be two potential problems resulting from the non-uniform sampling. First, there is missing information in the undersampled regions. This lack of information does not cause significant problems, since information from the lower frequency levels will fill those disocclusions smoothly. However, as we move wavelets, some of them may overlap. This leads to ghostings with overlapping background and foreground signals. To avoid this, we detect occluded wavelets and attenuate their amplitude.

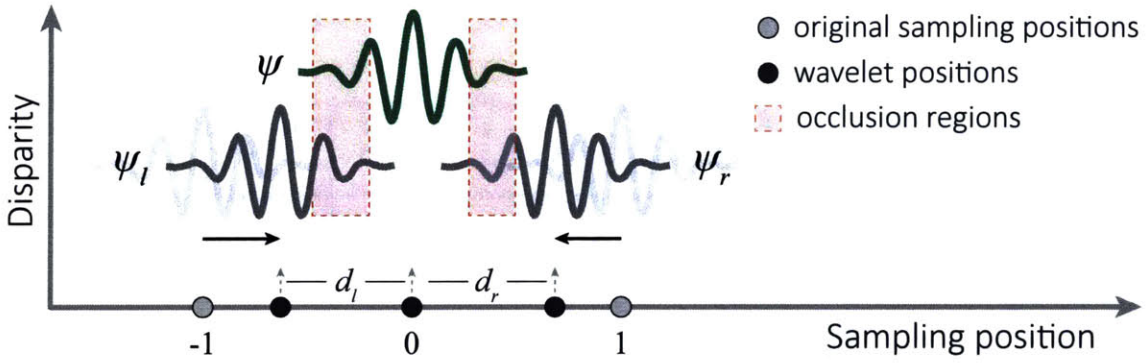


Figure 3-3: We resolve occlusion of wavelet $\psi$ by attenuating its amplitude. The attenuation is proportional to the part of the wavelet that is occluded by nearby foreground wavelets $\psi_l$ and $\psi_l$.

To this end, for a given wavelet $\psi$ we first find the closest wavelets to the left $\psi_l$ and to the right $\psi_r$ that have smaller disparities (i. e., they are in front of $\psi$). It is sufficient to consider wavelets corresponding to the same frequency. We then compute the portion of the wavelet $\psi$ that is occluded by $\psi_l$ and $\psi_r$. We assume that one wavelet completely occludes the other wavelet if the distance between them is at most half of the original sampling distance. As a result, we defined the occlusion using distances between sampling locations of $\psi$ and the other two wavelets, i. e., the occlusion caused by $\psi_l$ is defined as $\mathrm{O}_l = \max(2 - 2d_l, 0)$ and for $\psi_r$ by $\mathrm{O}_r = \max(2 - 2d_r, 0)$. Here, $d_l$ and $d_r$ are the distances, as marked in (Figure 3-3), and the original spacing between wavelets is assumed to be 1. The occlusions have constant value 1, if the neighboring wavelet moves half way to $\psi$, and 0, if the distance between them is at least the original sampling distance. To combine occlusions for both wavelets, we define the

effective occlusion of wavelet $\psi$ as $O_\psi = O_l + O_r$. $O_\psi = 0$ indicates that neither $\psi_l$ nor $\psi_r$ occlude $\psi$. $O_\psi = 1$ indicates that wavelet $\psi$ is completely occluded. Next, we attenuate wavelet $\psi$ according to a smooth function $s$ that interpolates between 0 and 1.

$$s(x) = \begin{cases} 1 & \text{if } x \geq 1 \\ 3x^2 - 2x^3 & \text{if } x \in (0,1) \\ 0 & \text{if } x \leq 0 \end{cases}$$

The amplitude of the attenuated wavelet is then defined as $\overline{A_\psi} = s(O_\psi) \cdot A_\psi$. For the real-time performance, we find $\psi_l$ and $\psi_l$ by first placing all wavelets in buckets according to their location, and then considering wavelets only from neighboring buckets within a distance of the wavelet spacing at the current level.

## 3.3 Additional Processing

A robust view synthesis is usually not enough for creating good content for a multi-view display. In order to enhance the quality and provide better experience, an inter-view antialiasing filter needs to be applied [38]. Also, in many cases, it is also necessary to perform depth manipulations to maintain viewing comfort [27] and improve perceived depth [14, 4]. Both steps can be easily incorporated into our method.

### 3.3.1 Antialiasing

Didyk et al. [5] proposed to perform the inter-view antialiasing by attenuating local amplitude according to the corresponding phase information. Because in our technique we rely on a very similar decomposition, the filtering can be performed using similar technique. To filter a view that was synthesized using our method, we attenuate every wavelet before the view is reconstructed. The amount of attenuation depends directly on the disparities between neighboring views, which can be easily obtained from our representation. For a given wavelet at frequency level $f$ with dis-

parity $d$, we attenuate its amplitude by multiplying it by $exp(-\sigma^2(\frac{2\pi d}{f})^2/2)$, where $\sigma$ is the antialiasing width as defined in [5]. An example of a synthesized view and its filtered version is shown in Figure 3-4.


Figure 3-4: The figure presents a synthesized view using our method (left) and the same view with the inter-view antialiasing applied (right).

### 3.3.2 Disparity Adjustment

Using our wavelet representation together with per-wavelet disparity information, we can easily apply non-linear disparity mapping operators. Such operators are usually defined as a disparity mapping function which maps disparity according to certain goals [14, 4]. To apply such a mapping to the disparities in our method, it is sufficient to apply the function to our estimate of the disparities between input views. Afterwards, the rest of our view synthesis technique remains unchanged. In Figure 3-5 we demonstrate one example of such manipulations.

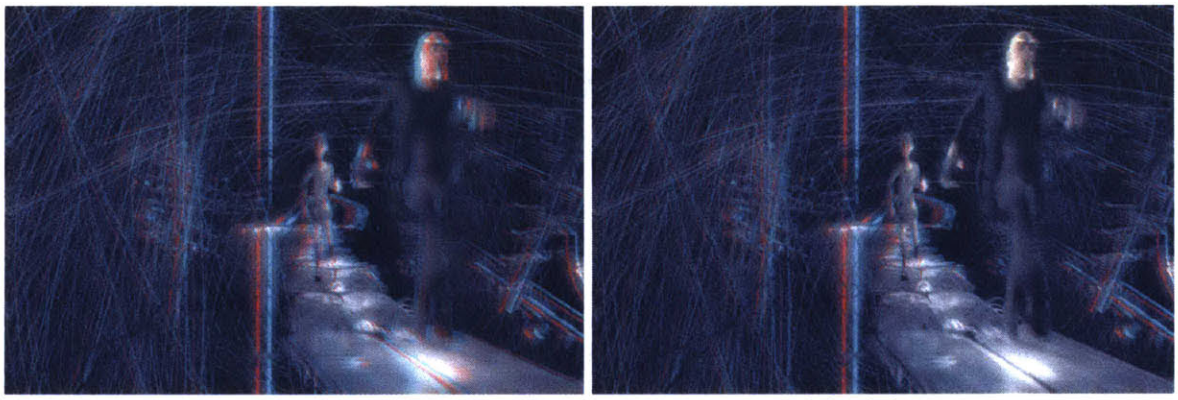


Figure 3-5: Example of nonlinear disparity remapping. The depth for the foreground objects are compressed while maintaining the same disparities in the background. As a result, the whole foreground part of the scene is pushed behind the zero disparity plane. The images can be viewed in stereo using red-cyan anaglyph glasses.

# Chapter 4

# Implementations

Our technique provides performance that is necessary to produce multi-view content for a 4K automultiscopic display in real-time. In this section, we describe two implementations: a CUDA-based GPU implementation and a hardware implementation using an FPGA with an ARM processors.

## 4.1 GPU Implementation

We produce content for an 8 view 4K ($3840 \times 2160$) automultiscopic display, where each of the output views has a resolution of $960 \times 1080$. Our method is implemented on a GPU using CUDA. To test its performance, we run it on the Nvidia GeForce GTX Titan Z graphics card. For such a setup, our technique is able to perform the conversion with the additional steps in 25 - 26 fps for all sequences presented. The breakdown of the timing and memory usages is presented in Table 4.1.

| Stage | Timing (%) | Memory (GB) |
|---|---|---|
| Pyramid decomposition | 9.9 | 1.05 |
| Initial disparity estimation | 4.9 | 0.31 |
| Per-wavelet disparity refinement | 18.5 | 0.23 |
| Wavelet re-projection | 30.5 | 0.50 |
| Pyramid reconstruction | 36.2 | 1.55 |

Table 4.1: Performance breakdown for the individual steps of our GPU implementation.

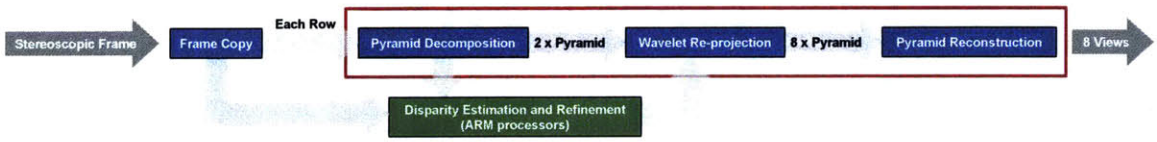


Figure 4-1: This figure presents how our method can be mapped to hardware architecture composed of an FPGA board with ARM processors.

## 4.2 FPGA Implementation

One advantage of our technique is that most stages in our algorithm can be done in a scan-line fashion. This eliminates the need for any external memory during the computation of these stages and, thus, it is suitable for a hardware implementation such as an FPGA or an ASIC. Our technique requires only low resolution disparity maps. Therefore, we leverage the ARM processors inside the SoC for this task. The ARM processor computes these disparity maps at the 240 $\times$ 180 resolution at 24 FPS.

Figure 4-1 describes the stages in our hardware implementation. The first stage decomposes the frame into two pyramids: one for the left view and the second for the right view. Both pyramids are sent to the second stage. In the second stage, each wavelet in the pyramid are re-projected according to the disparity from the ARM processor. The re-projected wavelets are filtered similarly to [16] and sent to the final stage. The final stage reconstructs views from the synthesized pyramids and outputs the result.

We test each stage of our implementation on the FPGA System-On-Chip (SoC) Xilinx ZC706 development board using Xilinx Vivado HLS 2015.4 software. The FPGA SoC has two ARM processors running up to 1GHz and programmable logic with 350K logic cells and a total of 19Mb of internal RAM. Table 4.2 shows the resource utilization of our implementation. Each stage is customized to the target generating 8 views of 512$\times$ 540 resolution at 24 FPS while running at 150MHz. The total memory utilization of our implementation is only 13Mb of the internal memory. This is a much smaller memory footprint than our current GPU implementation. Moreover, the current FPGA implementation uses only about 50% of the hardware resource on the FPGA we are using. Therefore, it is possible to double the resolution

to get a FullHD resolution in the future implementations.

| Stage | RAMs (Kb) | DSPs | LUTs | FFs |
|---|---|---|---|---|
| Pyramid decomposition | 976 | 26 | 14K | 12K |
| Wavelet re-projection | 12,960 | 427 | 74K | 85k |
| Pyramid reconstruction | 1,476 | 75 | 13k | 20K |

Table 4.2: Resource utillization on our FPGA implementation.

# Chapter 5

# Results and Comparisons

Our technique provides an up-conversion from stereoscopic content to its multi-view version. An example of applying our method to stereoscopic content is presented in Figure 1-1. In order to evaluate the performance of the technique, we consider a scenario of producing content for an 8-view 4K automultiscopic display. For the main evaluation we used four stereoscopic video sequences. Their representative frames are shown in Figure 5-1. To demonstrate both view interpolation and extrapolation, for two sequences (ELEPHANT and SKULL ROCK), we considered scenarios where two out of six additional views are located in-between the original views. For the other two sequences (BBB and BALL), we perform only extrapolation. Although we provide the evaluation only for animation sequences, we believe that their complexity justifies the extrapolation of our claims to real-world footage.

## 5.1 Comparison to State-of-the-art

Here, we compare our method to both Lagragian and Eulerian techniques. The first group consists of a depth image-based rendering (DIBR) technique [25] and an image-domain warping (IDW) technique [26]. Both of them target a real-time conversion of stereoscopic content to its multi-view version. As the source code of this first method is not publicly available, we reimplemented it. We only replaced their disparity computation with the method we use. We left, however, the pre-

processing steps which improve the disparity information. For the second technique, we provide a direct comparison to the results provided in the original paper. We compare also to the Eulerian method proposed by Didyk et al. [5] which applies a phase-based rendering approach (PBR). For all the methods, we present here only individual views.

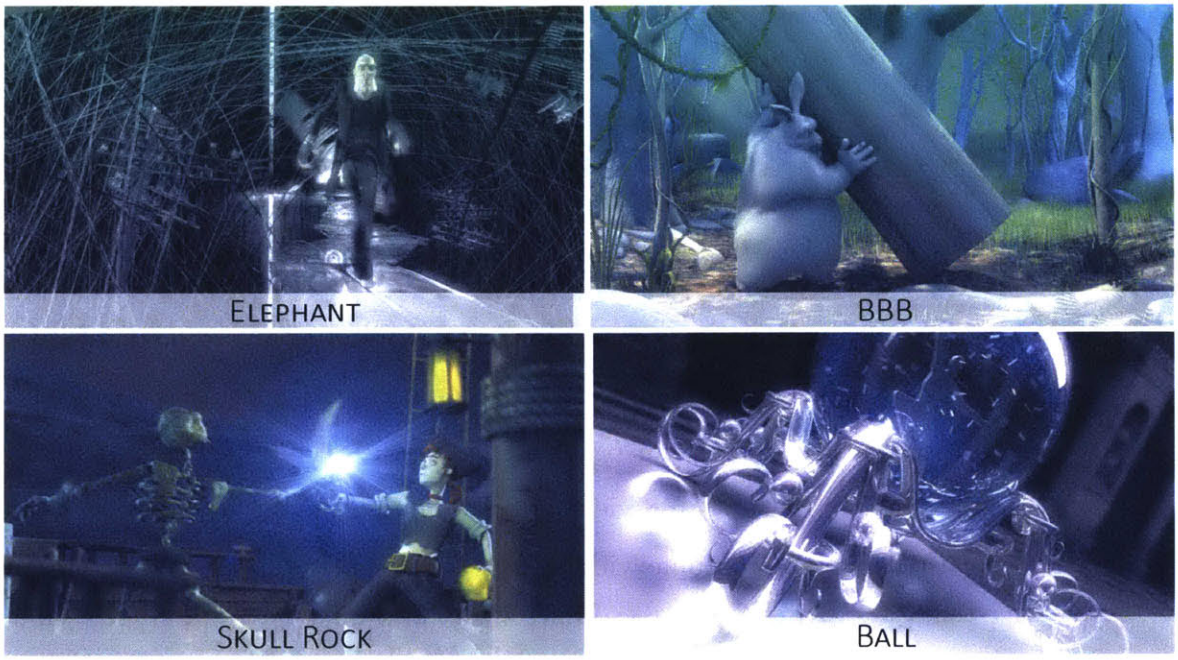


Figure 5-1: Representative frames of four stereoscopic video sequences used in our evaluation. ("Elephants Dreams" & "Big Buck Bunny" © by Blender Foundation, "The Curse of Skull Rock" © by Red Star Studio Ltd., "Ball" © by Eric Deren / Dzignlight Studios)

### 5.1.1 Lagrangian Approach

The artifacts produced by both methods can be categorized into two groups. The first group consists of errors due to poor depth quality. The second group includes artifacts coming from unsuccessful view reconstruction.

In order to assure real-time performance, depth quality is often insufficient to produce good results. In particular, both our technique and the one proposed by Riechert et al. [25] use depth map computed for downsampled images. This leads to disparity discretization artifacts. Using such disparity maps for image-based rendering results in discontinuities in perceived depth, and significant cardboard effect

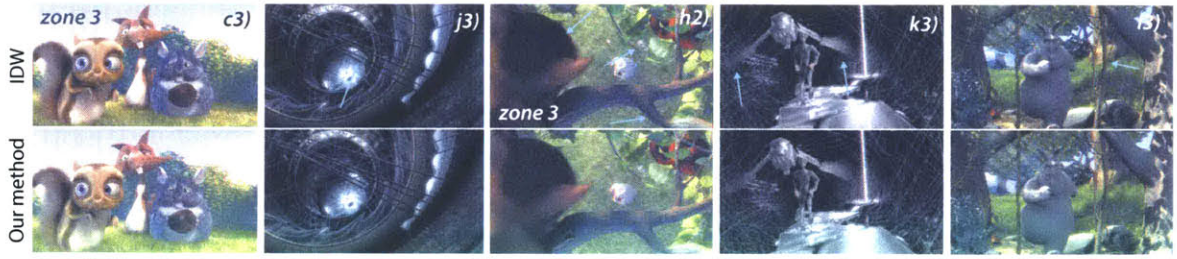


Figure 5-2: The figure presents a comparison of our technique to a recent image-domain warping technique [26]. The top row images come directly from the original publication, while the bottom row contains corresponding results of our method. The stereoscopic images are presented in anaglyph colors. Besides the first image (the most left one), the images were reported as difficult cases for IDW method. In all cases they suffer from inaccuracies with the sparse depth representation - artifacts include flattening the scene as well as losing the sharp depth discontinuities. Results provided by our technique offer more correct depth reproductions. At the same time, for the images on which IDW technique performs well (the first image on the left), our technique provides equally good results.

[22]. Whereas in the DIBR method, these artifacts can be easily observed, our step of phase-based depth correction can recover from these artifacts and provide more correct depth precept (Figure 5-3). The problem of inaccurate depth is addressed in image-domain warping techniques such as [26]. Such methods overcome the problem by warping the image according to sparse depth information. For example the above work uses mesh resolution $180\times100$. Although this avoids many visual artifacts related to depth estimation, it introduces other type of artifacts. As the depth information is represented using sparse set of features, often depth details cannot be reproduced by this technique. We compare our technique to [26] in Figure 5-2. In all cases, the IDW technique does not introduce visible 2D artifacts, however, when significant depth variations are present, it attempts to flatten the scene and smooth out sharp depth discontinuities.

Lagrangian techniques have also significant problems when depth information is not clearly defined. This can be usually observed for complex effects such as highlights, depth-of-field or motion blur. In such situations, standard depth based technique is usually unable to correctly synthesize novel views. In contrast, our technique, by leveraging the advantages of Eulerian approaches, can handle such situation more accurately (Figure 5-5). This is because we do not use per-pixel disparity information

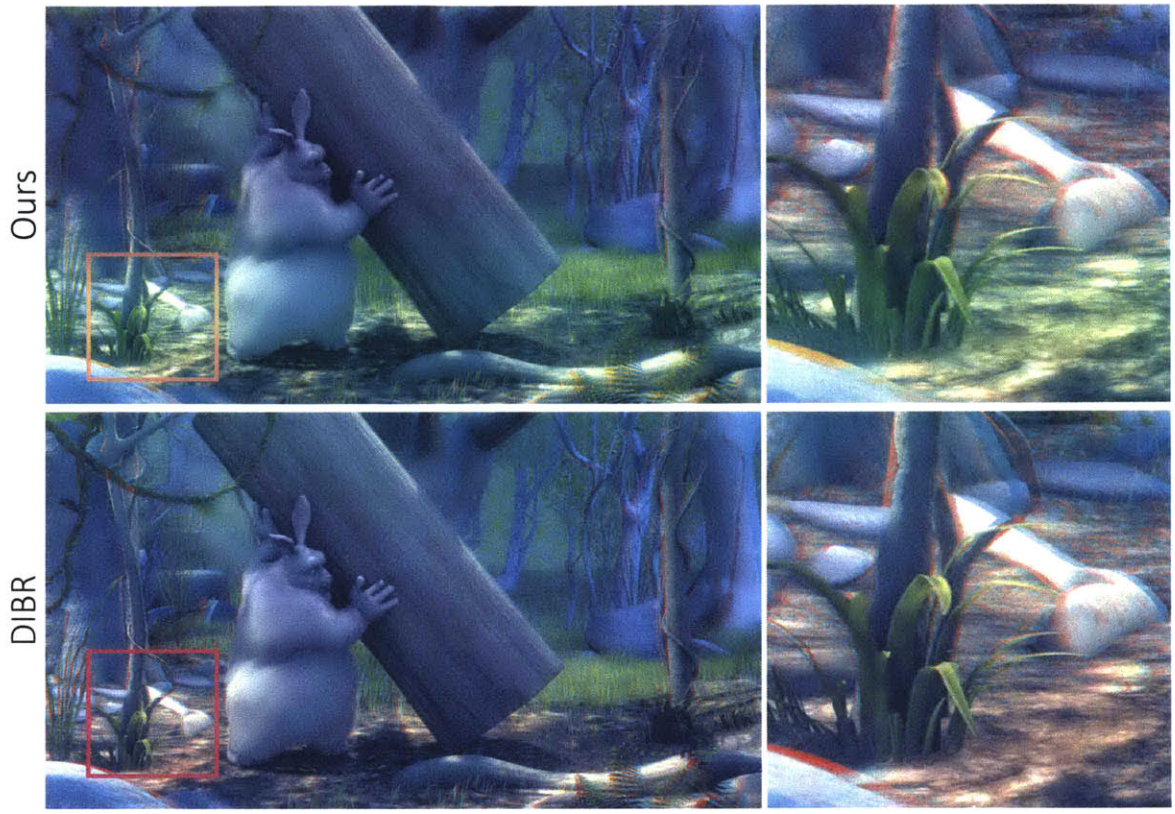


Figure 5-3: Comparison of the same stereoscopic view in red-cyan anaglyph. DIBR techniques produce large depth quantization leading to cardboard effect which can be observed on the trunk carried by the character as well as on the ground. Our technique, although it uses the same disparity map as an input can recover from these artifacts due to the additional phase information.

but rather per-wavelet disparity, which provides us with more information. In the case of motion blur and depth-of-field effects, Lagrangian techniques tend to introduce sharp edges that are not present in the original frames. This is also mentioned in [25]. Our technique does not introduce such artifacts (Figure 5-6).

Rendering novel views is challenging in places where disocclusions occur. Due to lack of information in these regions , usually a hole filling technique needs to be applied. DIBR method, due to the performance constraints, performs simple linear interpolation of neighboring pixels. Visually, this leads to an effect of stretching content over the disoccluded regions. Our technique does not need to explicitly perform hole filling. Instead, the missing information is filled during non-uniform FFT. As a result the local frequency spectrum in disoccluded regions is similar to the one in the neighborhood, which can be considered as hallucinating the unknown content. This is

different from the interpolation performed by DIBR which leaves vertical frequencies and removes horizontal ones.

### 5.1.2 Eulerian Approach

The major limitation of phase-based techniques is that they can handle only a limited range of disparities [5]. In our work, we overcome this problem by combining a phase-based approach with a standard Lagrangian approach. As a result, our technique can handle much larger input disparities and apply higher amplification factors (Figure 5-4). In contrast to PBR, which can apply only linear disparity mapping, our method enables more sophisticated depth manipulations (Figure 3-5).


Figure 5-4: PBR approach cannot handle large disparities. As a result the higher frequencies are wrongly synthesized which leads to significant ringing and blurring. On the other hand, our method is able to correctly reproduce high frequency content and the synthesized image is much clearer.

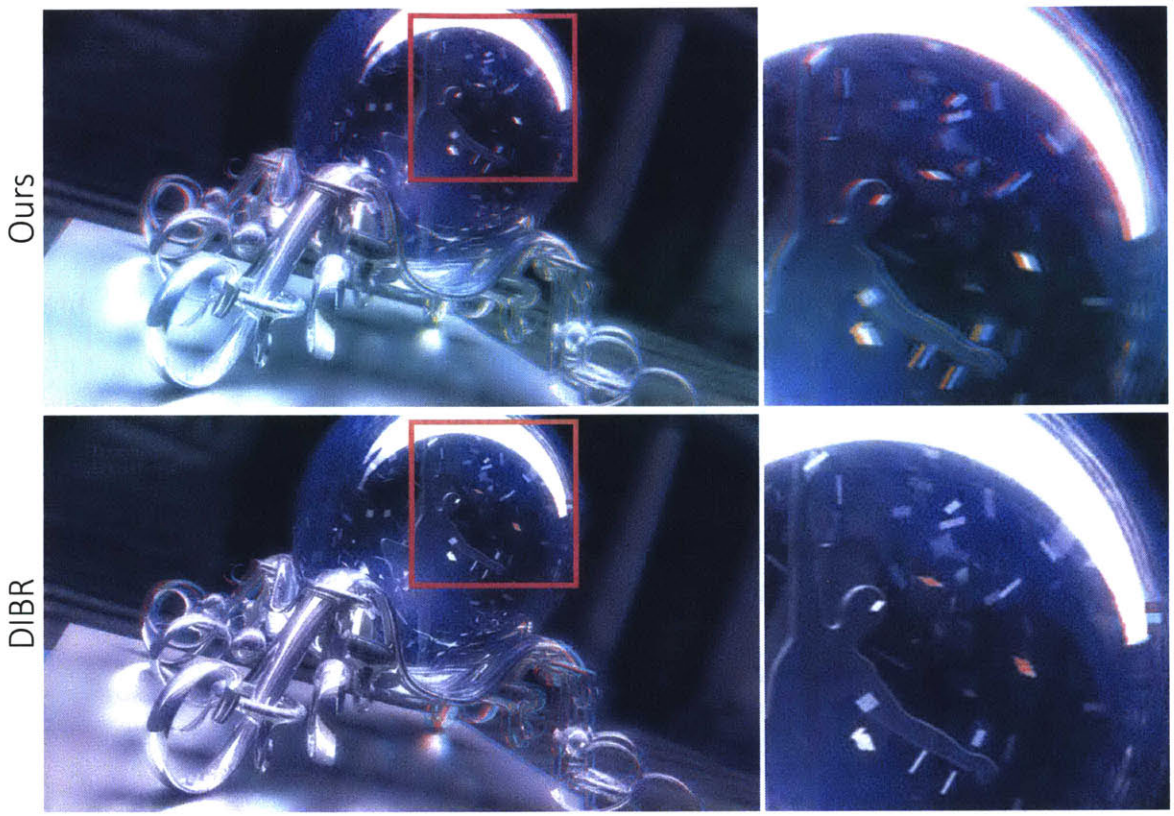


Figure 5-5: Comparison of results provided by our technique and DIBR for a challenging scene with a number of reflections and multiple depth layers. DIBR method cannot reproduce the correct depth of reflections, tiny particles and the glassy ball. In contrast our technique by leveraging advantages of Eulerian approach can reproduce the depth of these elements correctly. The images are shown in red-cyan anaglyph colors.

## 5.2 Influence of Disparity Quality

The input to our technique is a pair of disparity maps generated using [13]. The DIBR technique, that we compare against, uses the same disparity maps, but it performs a number of processing steps to improve it before it is finally used for view synthesis. To demonstrate the robustness of our technique to low quality disparity information, we checked how our technique performs if ground truth disparity information is available. Figure 5-7 presents results of our and DIBR techniques for three images from the Middlebury stereo datasets [12]. Additionally, we computed the results using the same techniques but supplied with ground truth disparity information. This is indicated by "*" next to the method names. The results are compared to original views using

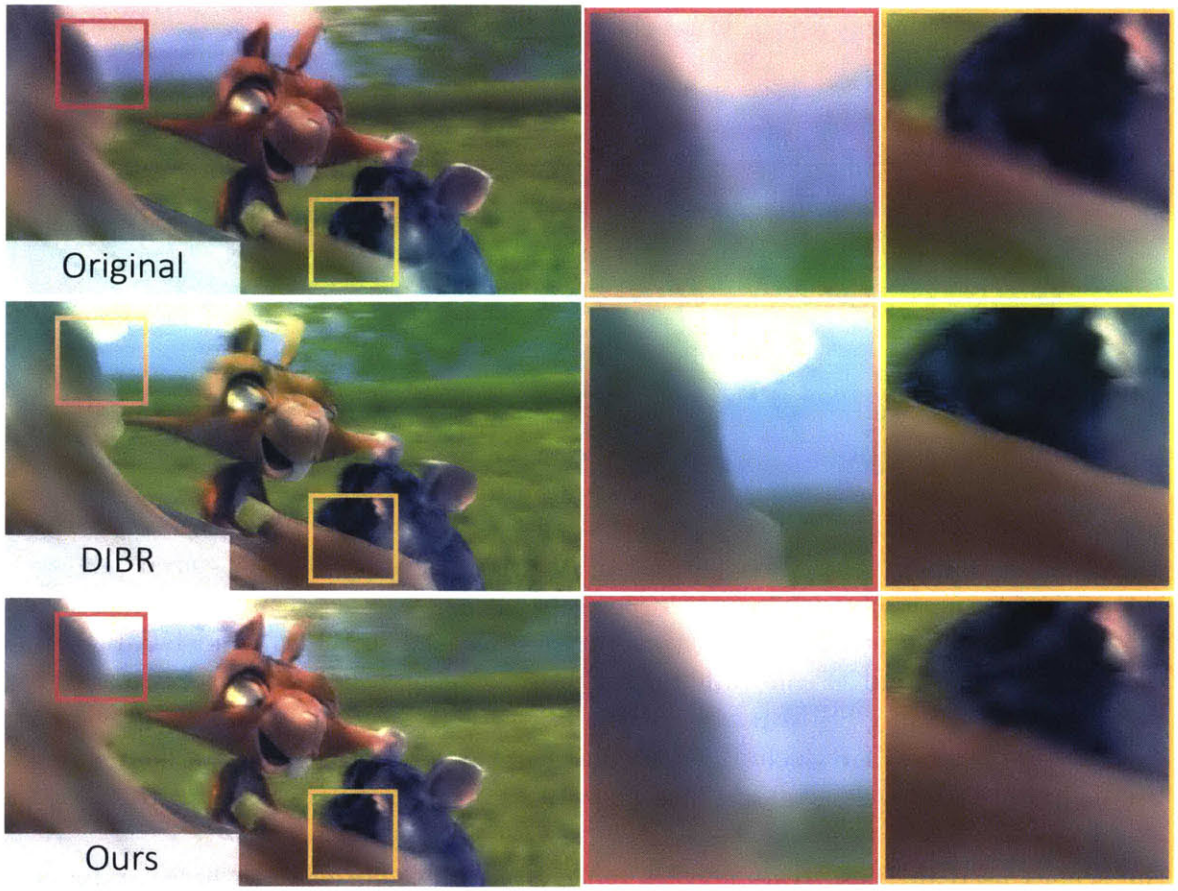


Figure 5-6: Motion blur is a difficult case for depth-based rendering. This is due to undefined depth values where the effect occurs. Here, we show one of the original views and one of the views synthesized using DIBR and our method. DIBR uses per-pixel depth information which is additionally filtered using bilateral-like filter. This introduces sharp edges in the region of blur. Our technique preserve the blur on the object boundaries.

SSIM metric [33], and the differences are reported using colormaps. It can be seen that our technique outperforms the DIBR technique, even though it uses improved disparity information. At the same time our technique provides similar results to DIBR technique which relies on ground truth disparity information. Interestingly, our technique does not significantly benefit from better disparity information. This demonstrates that our technique can use lower quality disparity information without overall quality loss and would most likely not benefit from costly depth estimations such as [36]. This is crucial for high quality view synthesis as ground truth disparity is usually unavailable.

Although our techniques performs similarly to DIBR with ground truth disparity,

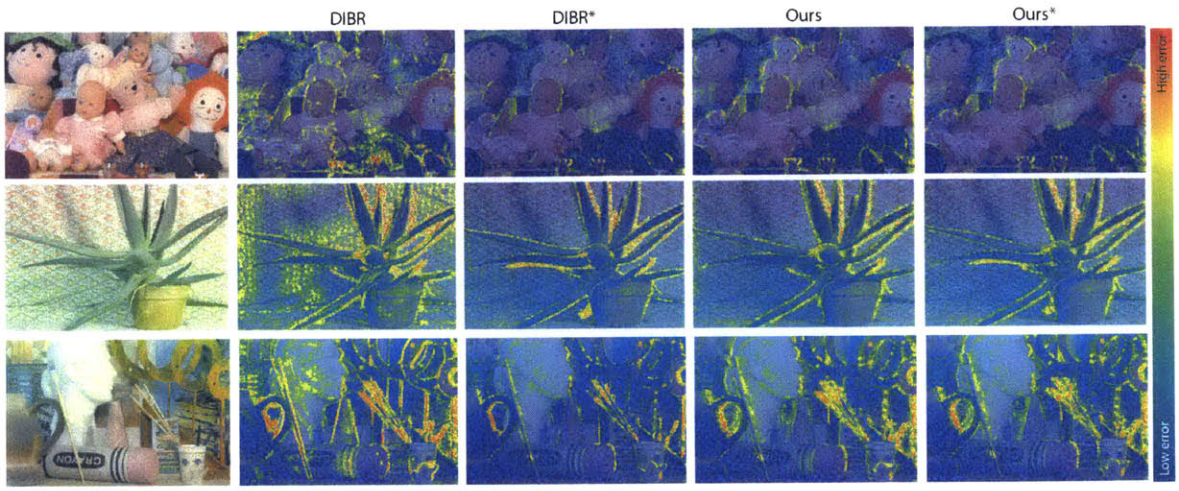


Figure 5-7: The figure presents a comparison between our technique and DIBR on three images from the Middlebury stereo datasets (http://vision.middlebury.edu/stereo/). Additionally, the results of both techniques with ground truth disparity information are shown (Ours* and DIBR*). The colormaps correspond to the differences between reconstructed images and original views measured using SSIM metric [Wang et al . 2004]. While our technique (fourth raw) outperforms DIBR technique (second raw), it produces very similar results to the techniques that use ground truth disparity information (third and fifth rows).

the scenes used in these tests consist mostly of diffuse surfaces without ambiguous depth situations like reflections, depth of field or motion blur. In more difficult cases with complex light effects, depth-of-field and motion blur effects, our method could outperform the DIBR method even if it was supplied with ground truth disparity information (Figure 5-8).

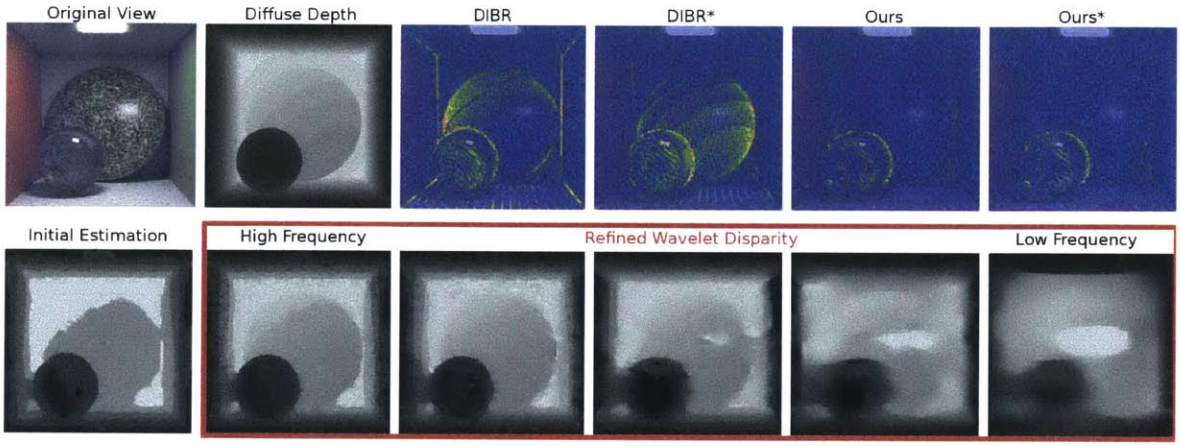


Figure 5-8: Complex light effects. As highlights or refractions have a different disparity than the diffuse component, it is challenging to expand views correctly using DIBR, even with the ground truth diffuse component depth map (DIBR*). On the other hand, our approach could handle this case by estimating disparity information separately for each wavelet frequency level. In this scene, the disparity information for diffuse components is captured by higher frequency levels, while lower frequency levels contain the disparity information for highlights and refractions.

# Chapter 6

# Discussion and Future Work

Similarly to a pure Eulerian approach, our disparity refinement step is limited to small disparity errors. However, the total wavelet disparity is not because it always includes the component from the initial disparity. This allows our method to outperform Lagrangian methods in many situations. For example, the discretization of initial disparity maps leading to cardboard effects is related to small disparity errors. These can be easily handled by our correction (Figure 5-3). In ambiguous cases, such as reflections, motion blur etc., the initial disparity is usually wrong. However, these phenomena usually correspond to lower luminance frequencies, for which the range of corrections we can perform is sufficiently large (Figures 5-5 and 5-6). Our Eulerian-based correction cannot handle large errors in the initial disparity map. These, however, usually correspond to untextured regions. Although we are not able to correct high frequencies in these areas, this does not create severe artifacts as the corresponding amplitudes are usually low.

There are two aspects that distinguish our approach from other multi-resolution techniques. First, we avoid notion of searching through a range and taking an optimal value. After the initial disparity is estimated, all refinements are expressed as closed-form expressions on wavelet phases. Second, we do not propagate disparity across frequencies. This independent estimation is the strength of our technique. This is also why our refinement is not a coarse-to-fine method.

Our method performs expansion only in the horizontal direction. Although this

is sufficient for standard automultiscopic displays, it would be interesting to consider extending it to the vertical direction. We also believe that our combination of the Lagrangian and Eulerian approaches opens up new ways of improving methods where view synthesis is necessary.

# Chapter 7

# Conclusions

We have presented a method that opens the door to practical 3D television systems at home. Our real-time method converts existing stereoscopic content to a high-resolution, high-quality, multi-view format that is suitable for automultiscopic displays. Our approach leverages advantages of both Lagrangian and Eulerian techniques by combining them into one method. This allows us to handle larger disparities than the Eulerian approach can deal with when applied alone, and to resolve difficult cases such as motion blur, depth of focus, and reflections which are challenging for Lagrangian approaches. To this end, we propose to use a steerable pyramid decomposition where disparity information is estimated for each wavelet separately. This decomposition is later used in our new wavelet-based view synthesis method which computes necessary views for autostereoscopic displays. Additional steps such as inter-view antialiasing or nonlinear disparity manipulations can be easily integrated in order to provide content customization. Our method operates locally, mostly on 1D scanlines, which allows for an efficient implementation both using a GPU and an FPGA. Our hardware implementation demonstrates that Eulerian techniques and their combination with Lagrangian approaches are good alternatives to hardware solutions that are based solely on Lagrangian approaches.

# Bibliography


[1] Myron Z Brown, Darius Burschka, and Gregory D Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993–1008, 2003.

[2] Alexandre Chapiro, Simon Heinzle, Tunç Ozan Aydın, Steven Poulakos, Matthias Zwicker, Aljosa Smolic, and Markus Gross. Optimizing stereo-to-multiview conversion for autostereoscopic displays. In *Computer Graphics Forum*, volume 33, pages 63–72. Wiley Online Library, 2014.

[3] Chris Chinnock. Trends in the 3D TV market. In *Handbook of visual display technology*, pages 2599–2606. Springer, 2012.

[4] Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, Hans-Peter Seidel, and Wojciech Matusik. A luminance-contrast-aware disparity model and applications. *ACM Transactions on Graphics (TOG)*, 31(6):184, 2012.

[5] Piotr Didyk, Pitchaya Sitthi-Amorn, William Freeman, Frédo Durand, and Wojciech Matusik. Joint view expansion and filtering for automultiscopic 3D displays. *ACM Transactions on Graphics (TOG)*, 32(6):221, 2013.

[6] Song-Pei Du, Piotr Didyk, Frédo Durand, Shi-Min Hu, and Wojciech Matusik. Improving visual quality of view transitions in automultiscopic displays. *ACM Transactions on Graphics (TOG)*, 33(6):192:1–192:9, 2014.

[7] Ye Fan, Joshua Litven, David IW Levin, and Dinesh K Pai. Eulerian-on-Lagrangian simulation. *ACM Transactions on Graphics (TOG)*, 32(3):22:1–22:9, 2013.

[8] Miquel Farre, Oliver Wang, Manuel Lang, Nikolce Stefanoski, Alexander Hornung, and Aljoscha Smolic. Automatic content creation for multiview autostereoscopic displays using image domain warping. In *IEEE International Conference on Multimedia and Expo*, 2011.

[9] David J Fleet and Allan D Jepson. Computation of component image velocity from local phase information. *International Journal of Computer Vision*, 5(1):77–104, 1990.

[10] David J Fleet, Allan D Jepson, and Michael RM Jenkin. Phase-based disparity measurement. *CVGIP: Image understanding*, 53(2):198–210, 1991.

[11] Andrea Fusiello, Emanuele Trucco, and Alessandro Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.

[12] Heiko Hirschmuller and Daniel Scharstein. Evaluation of cost functions for stereo matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[13] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(2):504–511, 2013.

[14] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross. Nonlinear disparity mapping for stereoscopic 3D. *ACM Transactions on Graphics (TOG)*, 29(4):75:1–75:10, 2010.

[15] Chao-Kang Liao, Hsiu-Chi Yeh, Ke Zhang, Vanmeerbeeck Geert, Tian-Sheuan Chang, and Gauthier Lafruit. Stereo matching and viewpoint synthesis fpga implementation. In *3D-TV System with Depth-Image-Based Rendering*, pages 69–106. Springer, 2013.

[16] QH Liu and N Nguyen. An accurate algorithm for nonuniform fast fourier transforms (nufft's). *IEEE Microwave and guided wave letters*, 8(1):18–20, 1998.

[17] Lytro Inc., January 2015. https://www.lytro.com/.

[18] William R Mark, Leonard McMillan, and Gary Bishop. Post-rendering 3d warping. In *Proceedings of the 1997 symposium on Interactive 3D graphics*, pages 7–ff. ACM, 1997.

[19] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Transactions on Graphics (TOG)*, 32(4):46, 2013.

[20] Belen Masia, Gordon Wetzstein, Carlos Aliaga, Ramesh Raskar, and Diego Gutierrez. Display Adaptive 3D Content Remapping. *Computers & Graphics, Special Issue on Advanced Displays*, 37(6):983–996, 2013.

[21] Wojciech Matusik and Hanspeter Pfister. 3d tv: a scalable system for real-time acquisition, transmission, and autostereoscopic display of dynamic scenes. *ACM Trans. Graph.*, 23(3):814–824, 2004.

[22] Lydia MJ Meesters, Wijnand A IJsselsteijn, and Piter JH Seuntiens. A survey of perceptual evaluations and requirements of three-dimensional tv. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(3):381–391, 2004.

[23] Karl Pauwels and Marc M Van Hulle. Realtime phase-based optical flow on the gpu. In *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pages 1–8. IEEE, 2008.

[24] Raytrix GmbH, January 2015. http://www.raytrix.de/.

[25] Christian Riechert, Frederik Zilly, Peter Kauff, Jens Güther, and Ralf Schäfer. Fully automatic stereo-to-multiview conversion in autostereoscopic displays. *The best of IET and IBC*, 4(8):14, 2012.

[26] Michael Schaffner, Frank Gurkaynak, Pierre Greisen, Hubert Kaeslin, Luca Benini, and Aljosa Smolic. Hybrid asic/fpga system for fully automatic stereo-to-multiview conversion using idw. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2015. to appear.

[27] T. Shibata, J. Kim, D.M. Hoffman, and M.S. Banks. The zone of comfort: Predicting visual discomfort with stereo displays. *Journal of Vision*, 11(8):11:1–11:29, 2011.

[28] Eero P Simoncelli and William T Freeman. The steerable pyramid: A flexible architecture for multi-scale derivative computation. In *Image Processing, International Conference on*, volume 3, pages 3444–3444. IEEE Computer Society, 1995.

[29] Eero P Simoncelli, William T Freeman, Edward H Adelson, and David J Heeger. Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2):587–607, 1992.

[30] Aljoscha Smolic, Karsten Muller, Kristina Dix, Philipp Merkle, Peter Kauff, and Thomas Wiegand. Intermediate view interpolation based on multiview video plus depth for advanced 3d video systems. In *IEEE International Conference on Image Processing*, pages 2448–2451, 2008.

[31] Nikolce Stefanoski, Oliver Wang, Michael Lang, Pierre Greisen, Simon Heinzle, and Aljoscha Smolic. Automatic view synthesis by image-domain-warping. *Image Processing, IEEE Transactions on*, 22(9):3329–3341, 2013.

[32] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T Freeman. Phase-based video motion processing. *ACM Transactions on Graphics (TOG)*, 32(4):80:1–80:10, 2013.

[33] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[34] Bennett Wilburn, Neel Joshi, Vaibhav Vaish, Eino-Ville Talvala, Emilio Antunez, Adam Barth, Andrew Adams, Mark Horowitz, and Marc Levoy. High performance imaging using large camera arrays. In *ACM Transactions on Graphics (TOG)*, volume 24, pages 765–776. ACM, 2005.

[35] Bennett S Wilburn, Michal Smulski, Hsiao-Heng K Lee, and Mark A Horowitz. Light field video camera. In *Electronic Imaging 2002*, pages 29–36. International Society for Optics and Photonics, 2001.

[36] Zhoutong Zhang, Yebin Liu, and Qionghai Dai. Light field from micro-baseline image pair. *displays*, 32(6):221, 2013.

[37] Jun Zhou, Yi Xu, and Xiaokang Yang. Quaternion wavelet phase based stereo matching for uncalibrated images. *Pattern Recognition Letters*, 28(12):1509–1522, 2007.

[38] Matthias Zwicker, Wojciech Matusik, Frédo Durand, and Hanspeter Pfister. Antialiasing for automultiscopic 3D displays. In *Proceedings of the 17th Eurographics conference on Rendering Techniques*, pages 73–82. Eurographics Association, 2006.