IMPROVEMENTS OF INTERPOLATION AND EXTRAPOLATION VIEW SYNTHESIS RENDERING FOR 3D AND MULTIVIEW DISPLAYS

by

ILIYA KORESHEV

B.A. (Computer Science), The University of British Columbia, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Electrical and Computer Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2013

Abstract

To display video content in 3D, traditional stereoscopic televisions require two views of the same scene filmed at a small distance from one another. Unfortunately, having the required number of views is not always possible due to the complexity of obtaining them and the required bandwidth for transmission. In cases where more advanced auto-stereoscopic televisions require more than two views, the issue of obtaining and transmitting those additional views becomes even more impractical and complex. These issues led to the idea of having a small number of real views and their corresponding depth maps, showing the distance of each object from the viewing plane, which together can be used to generate virtual intermediate views. These virtual synthesized views are generated by moving the different objects in the real views a specific amount of pixels based on how close or far they are from the viewing plane. The need for synthesizing virtual views is more pronounced with the introduction of stereoscopic and autostreoscopic (multiview) displays to the consumer market. In this case, as it is not practical to capture all of the required views for different multiview display technologies, a limited number of views are captured and the remaining views are synthesized using the available views. View synthesis is also important in converting existing 2D content to 3D, a development that is necessary in the quest for 3D content which has been deemed a vital factor for faster adoption of the 3D technology.

In this thesis a new hybrid approach for synthesizing views for stereoscopic and multiview applications is presented. This approach utilizes a unique and effective hole filling method that generates high quality 3D content. First, we present a new method for

view interpolation where the missing areas are filled with data from the other available view and a unique image warping approach that stretches out background objects to fill in the missing areas. Second, a view extrapolation method is proposed where small areas of the image are filled using nearest neighbor interpolation and larger areas are filled with the same unique image warping approach as in view interpolation. Subjective evaluations confirm that this approach outperforms the current state-of-the-art pixel interpolation-based view synthesizing method as well as the existing warping-based view synthesizing technique.

Preface

All of the work presented in this thesis was conducted in the Digital Multimedia Laboratory at the University of British Columbia, Vancouver campus. All non-original figures and tables have been used with permission from applicable sources mentioned in their descriptions.

A version of Chapter 3 has been published as I. Koreshev, M. T. Pourazad, and P. Nasiopoulos, "Hybrid View-Synthesizing Approach for Multiview Applications", 3DTV Conference, Zurich, Switzerland, Oct. 2012. I was the lead investigator responsible for all areas of research, data collection, as well as the majority of manuscript composition. M. T. Pourazad was involved in the early stages of research concept formation and aided with manuscript edits. P. Nasiopoulos was the supervisor on this project and was involved with research concept formation, and manuscript edits.

A version of Chapter 4 has been published as I. Koreshev, M. T. Pourazad, and P. Nasiopoulos, "A New Hybrid Approach for View Extrapolation and Hole Filling", IEEE International Conference on Consumer Electronics, Las Vegas, USA, Jan. 2013.. I was the lead investigator responsible for all areas of research, data collection, as well as the majority of manuscript composition. M. T. Pourazad was involved in the early stages of research concept formation and aided with manuscript edits. P. Nasiopoulos was the supervisor on this project and was involved with research concept formation, and manuscript edits.

Table of contents

List of Tables

List of Figures

1 Introduction

Three-dimensional (3D) video provides viewers with a more engaging and realistic impression of scenes than traditional two-dimensional (2D) video. Viewers can perceive depth in 3D videos the same way as if they are looking at a live scene. The first major hurdle in the proliferation of 3D display technology is the availability of 3D content. As of now, the majority of available content is still 2D, and consumers buying a 3D TV end up disappointed in the technology as they use the display less often for watching 3D content than 2D content - which they could always watch before. A further problem is that as this technology evolves and grows so do the expectations of consumers. Watching 3D content without wearing cumbersome glasses is one of the key features that 3D technology consumers demand [1][2]. In this regard, researchers and display manufacturers are working towards developing multiview displays which do not require wearing 3D glasses. This technology, however, requires several views of the scene to be captured simultaneously, and since multiview content production is expensive and highly demanding in terms of camera configuration and post processing, the problem of lack of content is even more pronounced. As multiview technology evolves, manufacturers attempt to provide viewers with a larger number of views to improve transition between sweet spots. As a result, the number of views of the preliminary multiview content will no longer be enough. In addition to the above-mentioned challenges involved with multiview content production, the transmission of multiview content which includes several number of views is extremely expensive.

To address the inherent problems with the multiview content generation, the 3D Video (3DV) ad-hoc group (part of ISO/IEC Moving Pictures Experts Group (MPEG)), recommended capturing a limited number of views (two or three views) plus their respective depth maps (3DV data format) and synthesize the rest of views for multiview applications [3]. While this type of data format reduces some of the issues regarding camera configuration and transmission bandwidth, it introduces a new challenge, which is synthesizing high quality views. This challenge also exists in converting existing 2D video content into 3D video format. For 2D-to-3D video conversion purposes, there have already been developed automated techniques for depth map generation from 2D videos [4]. Yet, after estimating the depth map, the remaining challenge is, again, to synthesize other virtual view(s). So far, the majority of research efforts on synthesizing virtual views have been focused on generating intermediate virtual views between two or three real (available) views by view interpolation as shown in Figure 1.1. However, for 2D-to-3D video conversion, synthesized virtual views are on either side of the available view and are generated via view extrapolation (see Figure 1.2). View extrapolation is more challenging than view interpolation, since in extrapolation a limited amount of information is available (only one view).

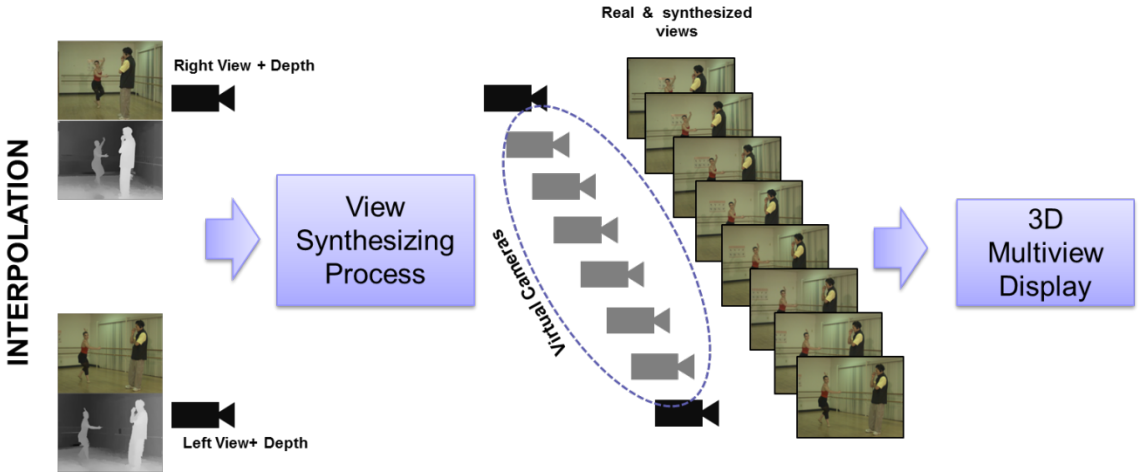


Figure 1.1: View synthesis via view interpolation

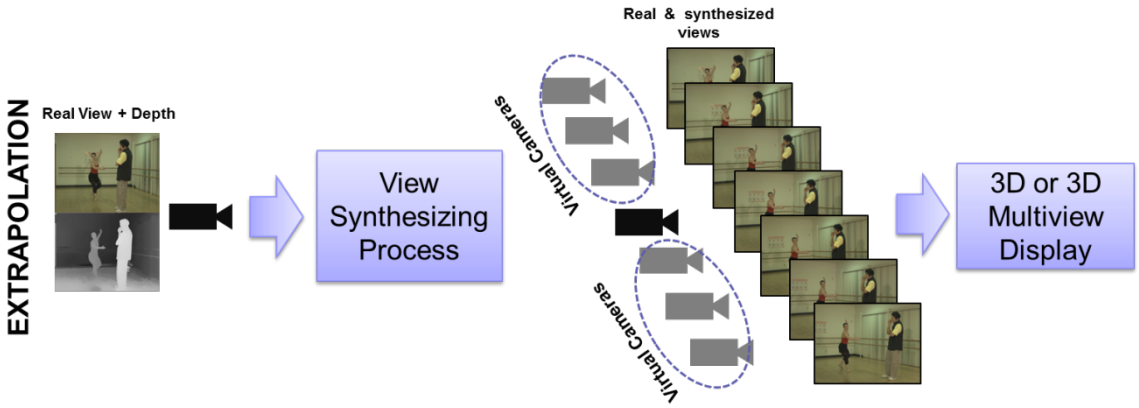


Figure 1.2: View synthesis via view extrapolation

The main issue with view synthesis is related to estimating the information of the occluded areas. During the synthesizing process, areas of the background that were occluded by foreground objects in the available views become visible in the synthesized views. These areas (holes) must be filled with realistic data to avoid noticeable artifacts. A common solution is to apply interpolation to estimate the missing texture. This approach has been utilized in the existing state-of-the-art view synthesis reference software (VSRS), which has been adopted by the MPEG-3DV group to synthesize test

sequences for 3D video compression standardization activities [5]. VSRS uses the depth and texture information of the available view(s) to generate virtual views. The foreground and background objects are segmented using the depth data and then are horizontally shifted based on their depth range to create virtual views. This shifting is what produces areas with missing texture called holes. VSRS uses the nearest neighbor interpolation approach to fill these holes. The downfall of interpolation-based hole-filling methods is that the interpolated texture does not resemble the true texture of the occluded areas, but instead looks as if a "clone tool" was applied to those areas, in a sense that small parts of the neighboring texture are simply replicated (copied) over and over. This approach usually produces a similar looking color to the true background, but fails to reproduce texture that exists in those areas, thus reducing the quality of the synthesized views and hampering the overall 3D effect.

To avoid the creation of holes in the synthesized view, a group of researchers from the Disney Research lab in Zurich have proposed to use a warping technique to generate synthesized views from the available views [6]. In this method, first a sparse saliency map is created. This saliency map helps with separating foreground and background objects. During the second stage, the saliency map information is used to stretch or compress some parts of the picture. The end-result is that this method does not produce holes. However, due to warping (stretching or shrinking), some deformation may be evident in the generated virtual views. This is more prominent around foreground objects that have large disparity (need to be stretched more) and are also close to background objects with well-defined vertical edges. In such cases, since variant amounts of warping are applied to the image, vertical edges may be deformed and become wavy.

The resulting quality of the synthesized views is very important as it is recognized that low quality 3D videos can produce eyestrain, headaches, and generally unpleasant viewing experience for the viewers [7]. Thus, in order to enable the 3D market through availability of content and ultimately use of multiview technology, there is a strong need for an effective view synthesizing approach that does not compromise the quality of the generated view with inadequate/poor hole-filling.

In this thesis, a new view synthesizing approach is proposed equipped with efficient hole-filling performance for view interpolation and view extrapolation applications. Our method takes advantage of some general ideas from the VSRS approach in [5] as well as the warping technique (Disney approach) in [6], to build a hybrid approach which results in unparalleled quality for synthesized views. In our method a view is synthesized by shifting the objects based on their depth map, similar to [5]. However, to fill the generated holes, unlike [5] which uses nearest neighbor interpolation, and unlike [6] that uses warping, our method uses a hybrid approach applying interpolation or warping for filling the holes, depending on the size and location of the holes. In addition, unlike [6], we only warp the background so that the shape and the size of the foreground objects are intact by our hole filling process. By warping the existing background texture, the holes are filled with more realistic texture that is more similar to the texture of surrounding background region. Since the texture of filled areas is similar to that of surrounding areas, the filled areas look more natural and the overall quality of the synthesized views is improved. To evaluate the performance of our algorithm, we conduct subjective tests and compare our synthesized views with those generated by the state-of-the-art view

synthesis reference software (VSRS) [8], and the Disney approach in [6] for both view extrapolation and view interpolation scenarios (see Figures 1.1 and 1.2).

The rest of this thesis is organized as follows. Chapter 2 provides background information on how 3D is displayed on conventional stereoscopic displays using glasses and glassless autostereoscopic (multiview) displays, as well as current view synthesis techniques for both view interpolation and view extrapolation. In Chapter 3, we present a new view interpolation algorithm for generating synthesized views using two existing views and their corresponding depth maps. A new algorithm for view extrapolation from only one available view and its depth map is presented in Chapter 4. Finally, conclusions and directions for further research are provided in Chapter 5.

2 Background

In this chapter, we provide background information on the fundamentals of 3D vision and how they are used in 3D and auto-stereoscopic multiview displays, as well as current synthesized view interpolation and extrapolation techniques available. Section 2.1 provides the basic ideas behind 3D vision and how humans can see 3D images on a flat 2D surface. In Section 2.2 we provide the basics behind conventional 3D displays that require glasses. Section 2.3 explains how new glassless auto-stereoscopic displays are made using multiview technology. We cover the reasoning behind accepted 3DV encoding and transmission techniques in Section 2.4. Finally a detailed summary of the current industry standard view synthesis techniques and existing research on an alternate view synthesis technique for view interpolation and extrapolation is provided in Section 2.5 and Section 2.6, respectively.

2.1 3D Vision

The first plausible explanation for why humans have two eyes was given by Charles Wheatstone who suggested that the disparity between the two eyes causes a unique sense of depth that allows us to judge how far away an object is. This has been attributed to evolution which favored those species who could judge how far or close predators or prey were. This sense of depth is also called the effect of seeing 3D; therefore, humans with an undamaged optical system can see 3D object in real life. However, due to the limitations of traditional video media we are presented with a flat screen on which all content is displayed equally for both eyes. This prevents the optical system from judging object depth and eliminates the 3D effect that we would otherwise experience in real life.

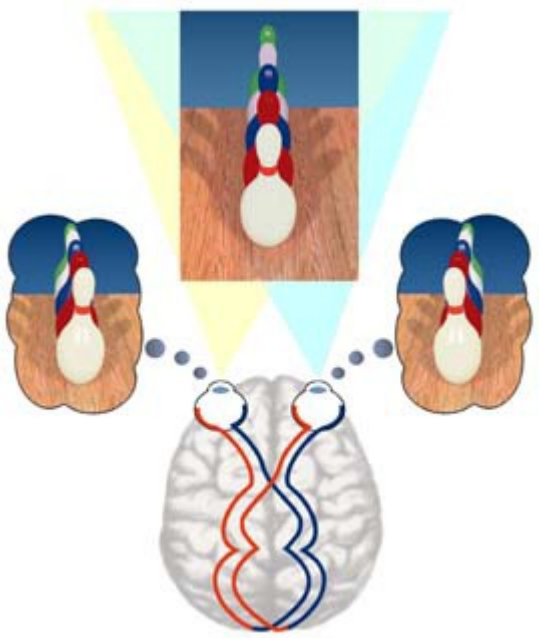


Figure 2.1: Human optical system sees depth due to disparity between the eyes

The idea that the human optical system sees 3D due to the disparity seen in the image seen by the left and right eye led to the idea of producing the same effect with traditional print media in the 1830's by Charles Wheatstone. A reflecting mirror stereoscope was used to allow the viewer to see two slightly shifted images, one with the left eye and one with the right eye, creating a 3D effect similar to what we would see in real life. Soon after that, by the 1880's, Thomas Edison's research team started investigating the possibility of using the same approach to show stereoscopic video. By 1922 the first 3D theatre, called Teleview, was opened in New York; it used two projectors synchronized to special viewers at each seat. The projectors alternated showing two slightly shifted images while the viewers used a rotating shutter effect to cover either the left or right eye so that only one image could be seen by each eye similar to what was done using the mirror stereoscope earlier.

The same approach, of showing two shifter images while blocking one image from each eye, led to the traditional anaglyph (red/blue) glasses that used the principle of blocking different color channels from reaching the left and right eye thus effectively blocking one of the two images displayed. As technology moved forward, polarized glasses were used to block the images based on their polarization, as well as shutter glasses, which worked on the same principle as the Teleview's viewers, by alternating, at a very high frequency, of making one lens opaque and the other transparent in synch with the display.

2.2 Conventional 3D Displays

There have been various approaches presented for the purpose of simulating 3D depth using a flat display so that objects look like they are in front of the user [9] [10]. The most common form of 3D displays available for the consumer market now use the approach of blocking one of two shifted views from each eye to produce the 3D effect. Using the anaglyph approach to block certain color bands from being shown to each eye greatly reduces accurate color reproduction; therefore, other methods such as polarized or rapidly alternating active shutter systems have become standard [11]. All these methods rely on using a pair of glasses which take care of blocking one view from each eye.

Active shutter systems use a pair of glasses that is synchronized with the displays and rapidly alternate each of the lenses over the eyes from being opaque to transparent [11] (see Figure 2.2). When these glasses are used in conjunction with a display that also rapidly alternated between displaying the left and right image, the user can see a 3D depth effect. The video passed to these displays is twice the regular frame rate and every

other frame is part of a single view. Due to the nature of active shutter glasses, the display has to always be synchronized to the glasses so that the alternating views are displayed properly to the viewer. This is achieved using a sensor which is attached to a controller in the display; the sensor sends a signal to the glasses and synchronizes the frame displayed on the screen with the correct lens remaining transparent. A problem with this approach is the loss of brightness in the scene. Since most shutter glasses cannot become completely transparent and retain a slight tinted feel this leads to some of the light being blocked from the screen effectively reducing the overall brightness of the video displayed.

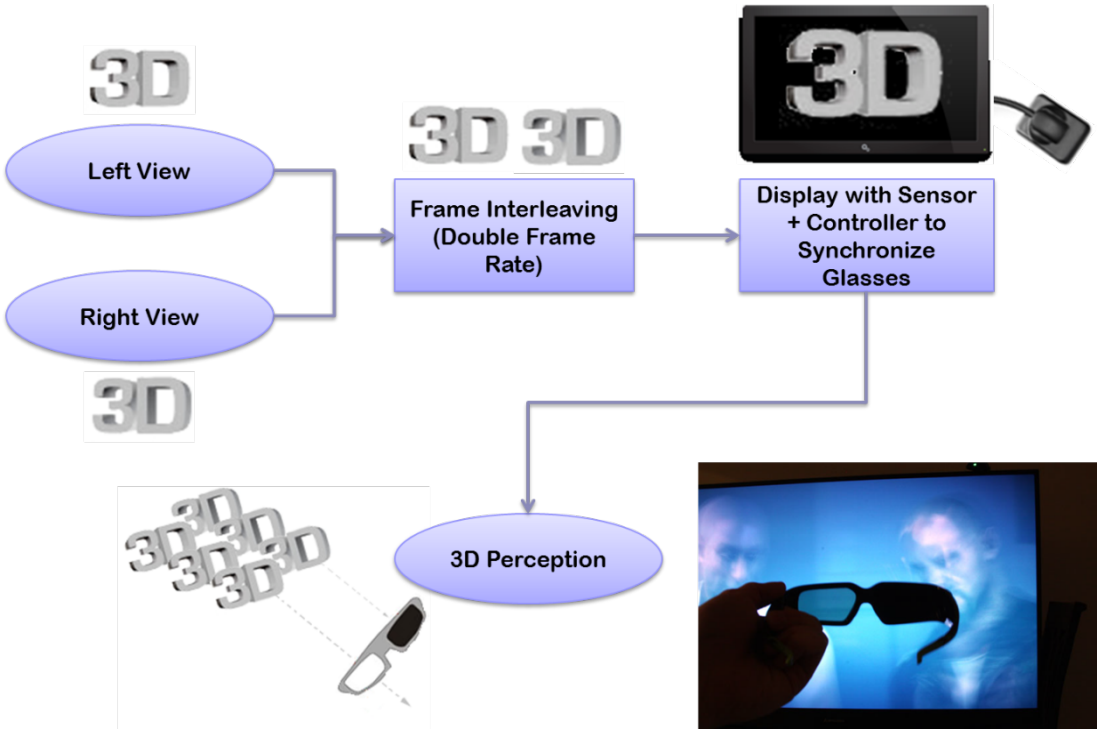


Figure 2.2: Active shutter glasses 3D technology allows the viewers to see two different frames by rapidly switching each lese from transparent to opaque

Unlike active shutter glasses, passive polarized systems do not need to synchronize the glasses to the display. Passive polarized systems use glasses that have differently

polarized lenses for each eye as well as a polarization sheet in front of the display that polarizes the two views [12]. These displays require the video to be interlaced so that every other line is polarized in the same way allowing both images to be displayed simultaneously on the display and only one image visible to each of the viewers' eyes after they have passed through the polarized glasses. Polarized systems come in two varieties: linear polarized systems and circular polarized systems. Both systems work on the principle of having each lens block part of the video frame similar to the active shutter system shown in Figure 2.2; however, the lenses do not switch between transparent and opaque as with the active shutter glasses due to the polarization effect. The main difference between the two passive polarized systems is that with linear polarization the frames are interlaces in such a way so as the light waves for one frame come at the viewer in up and down vertical waves while the light waves for the other frame come at the viewer in horizontal waves. Circular polarized systems on the other hand rotate the light waves in a circular and counter circular motion rather than just the horizontal and vertical of linear polarized systems (see Figure 2.3). The main advantage of circular polarized systems is that the viewers can tilt their heads and still maintain the 3D effect while linear systems require the viewers to maintain their eyes level to the screen. With polarized systems the video frames are interlaced rather than increasing the frame rate. During this process two frames are encoded into the size and frame rate of one 2D frame. A disadvantage of polarized systems, since the frame rate and resolution remain the same is that twice as much information needs to be shown, downgrading the overall video quality (resolution) that is presented to the viewers.. There is also a slight

loss of brightness due to the tinted nature of polarized glasses, although this is less visible than that caused by the active shutter technology.

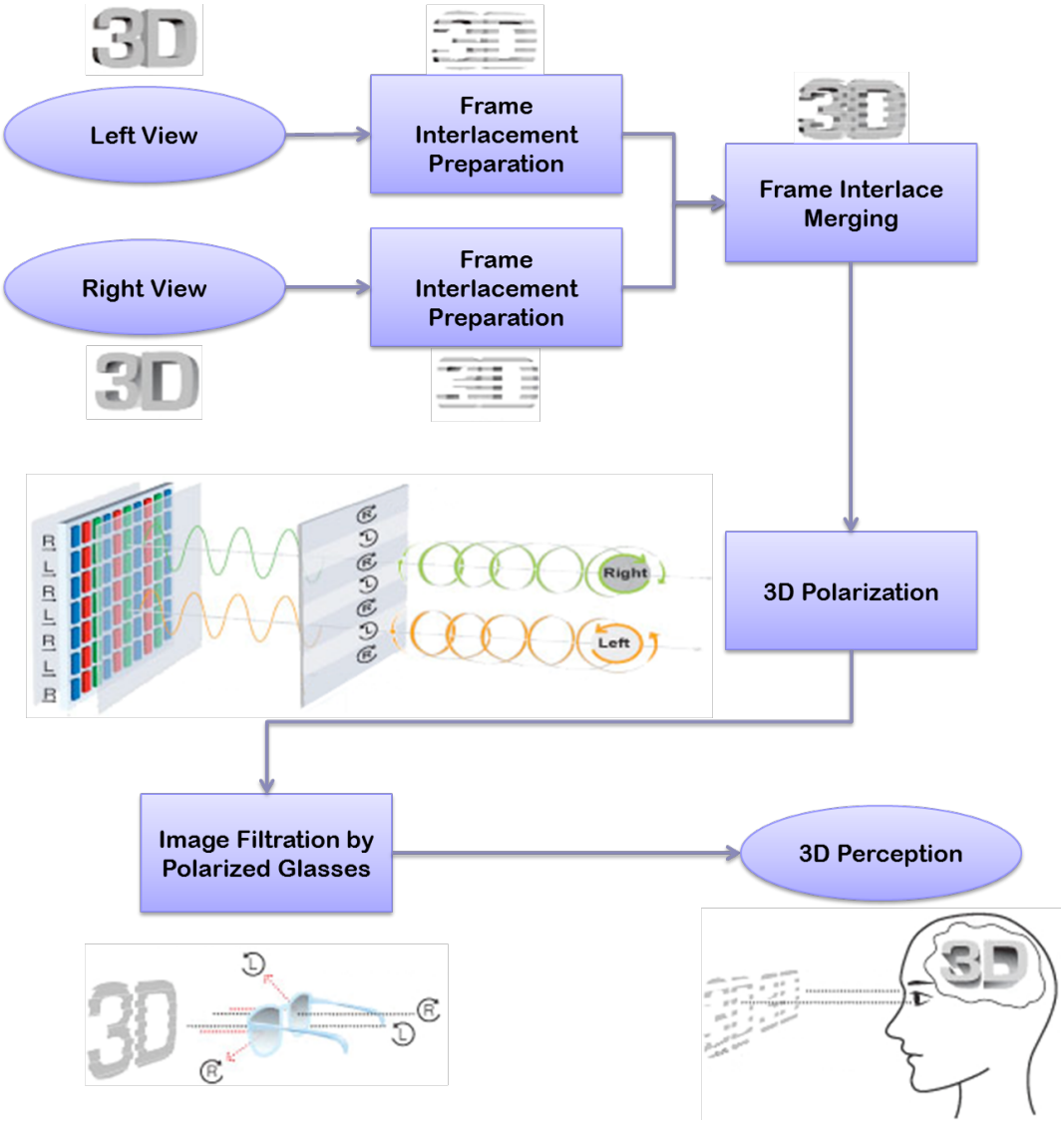


Figure 2.3: Passive polarized glasses 3D technology

2.3 Auto-Stereoscopic Multiview Displays

Conventional 3D displays that require glasses to display the 3D effect have become widely used in molecular modeling, as well as CAD fields where the requirement of

wearing special glasses is not regarded as a hindrance [13]. However, while there is currently a large market push for conventional 3D displays using glasses, consumers dislike wearing additional equipment, such as special glasses, as many state that glasses negatively affect their general ambient visual acuity [14]. Due to this, there has been much research motivation for the development of non-invasive techniques that can be used in stereoscopic display applications which do not require special glasses to be worn by the viewers.

Displays that allow the viewers to see a different view with each eye and create the 3D effect without requiring the viewers to wear special glasses are termed auto-stereoscopic [15]. Many researchers have developed displays that present a different image to each eye. These displays usually use a variation of one of two known methods to display only one view to each of the viewer's eyes. The first is the parallax barrier method, where fine vertical gratings are placed in front of the screen and block one view from each eye, as seen in Figure 2.4. The other is the lenticular lens array method which places lenticular lenses in front of the screen and instead of blocking one view it angles the displayed light in such a way so only one view is aimed at each eye, as seen in Figure 2.5. For both of these methods if the viewer remains in a fixed position, only one eye can see the even display and the other can see the odd display.

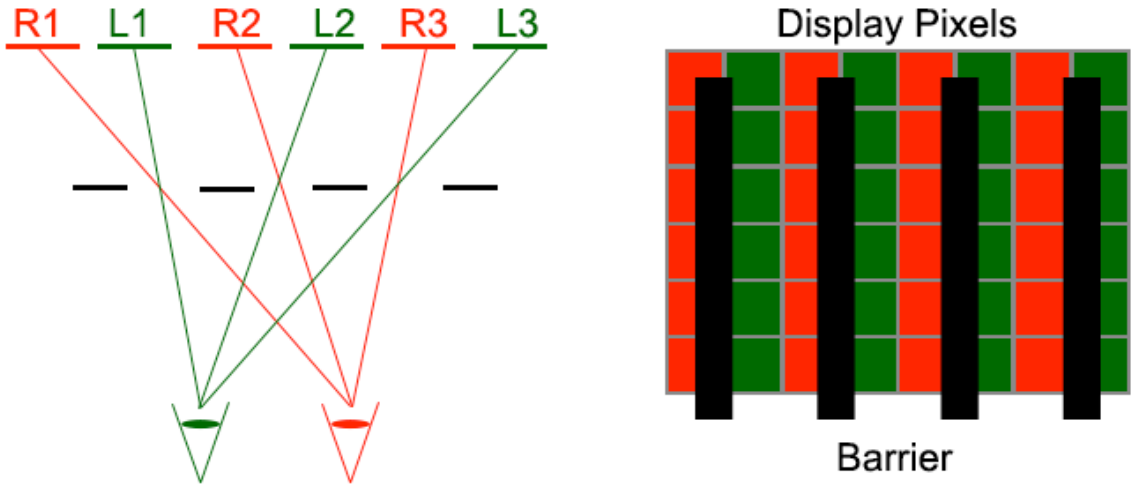


Figure 2.4: Parallax barrier auto-stereoscopic display blocks one view from each eye

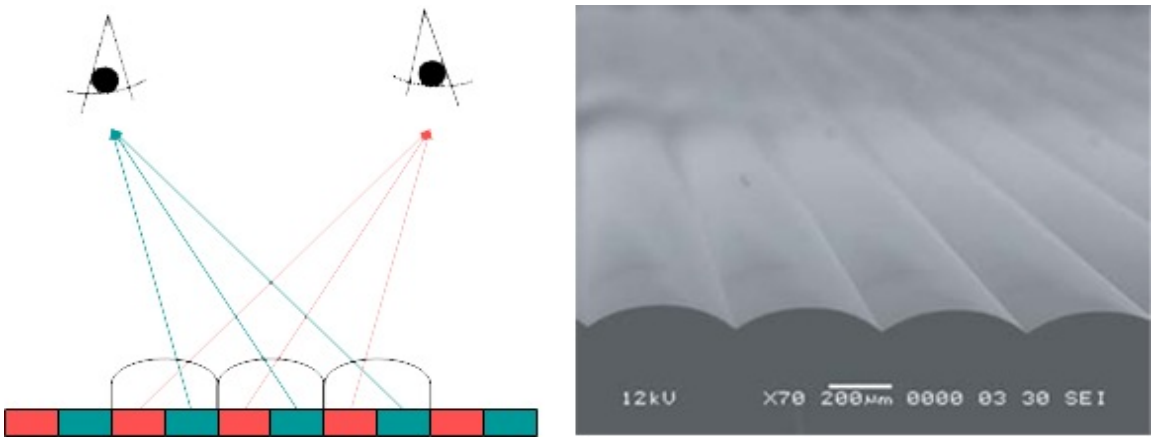


Figure 2.5: Lenticular lens auto-stereoscopic display angles one view at each eye

There are two main drawbacks to both of these techniques. One is that the viewing angle of such displays is very limited and the viewer has to remain in one specific location to be able to perceive the 3D effect [13], and the other was that the horizontal resolution of the screen would be halved since we are blocking half of the image from each eye. The first problem can be solved by reducing the size of the parallax barriers or lenticular lenses and projecting more than one view, leading to a so-called multiview display. This allows an increase of the viewing angle since there is no longer only one

"sweet-spot" for 3D perception, as seen in Figure 2.6. However, this solution only further increases the problem of reduced resolution, since more pixels are blocked from the viewer. To resolve the second problem we must observe that humans are very perceptive of vertical edges; therefore, by angling the barriers or lenses we have a trade-off between a loss of vertical and horizontal resolution, as seen in Figure 2.7, effectively increasing the overall perceived quality of the images displayed.

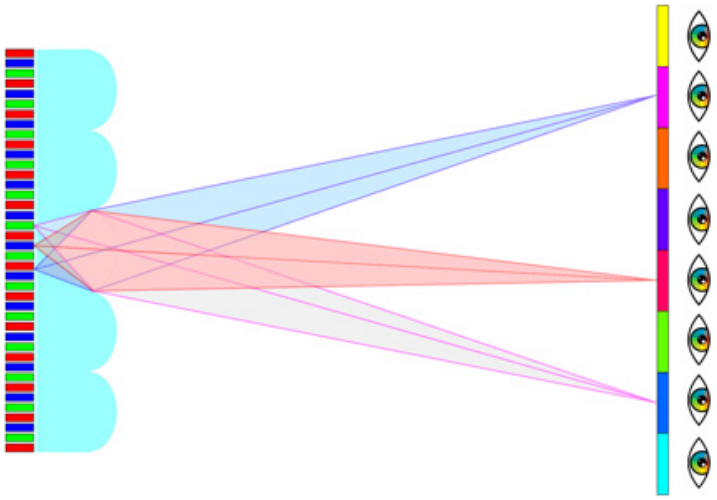


Figure 2.6: Multiview auto-stereoscopic display showing 8 views in total

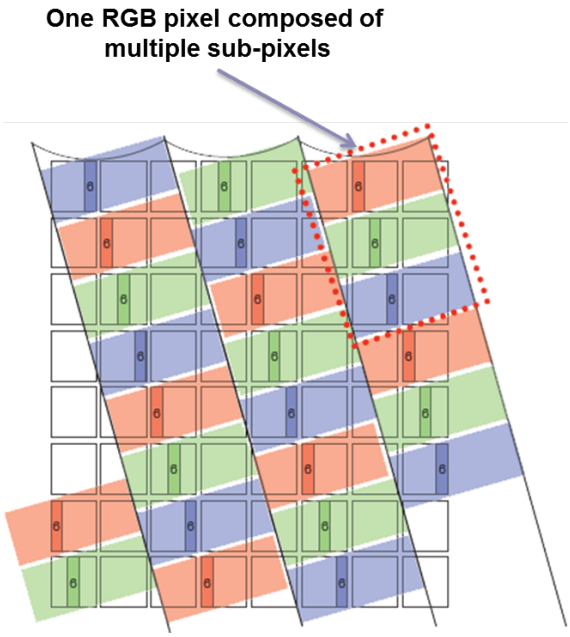


Figure 2.7: Angled lens display increases horizontal resolution at expense of vertical resolution

Today multiview auto-stereoscopic displays use anywhere from 8 to 56 views to increase the 3D perception and viewing angle. This leads to an issue with both content generation and transmission. First, it is very costly and complex to film each scene from 56 different angles and as multiview screens increase the number of views content cannot always be remade to add more views. Next, transmission of 56 views is very bandwidth intensive and thus very expensive to transmit. This means that there is a need to reduce the overall bandwidth requirements while ensuring compatibility with multiple displays.

2.4 3DV Encoding

Due to the different formats of 3D displays and the bandwidth issues of transmitting 3D video data, there is a need for a versatile encoding scheme that can reduce bandwidth and allow the same stream to be used for different types of displays. A popular format for 3DV coding, which has been around for over ten years and is still in use today, is to send

a color stream with its associated depth map [16]. This method reduces the bandwidth required to send video greatly as fewer views need to be sent and the depth map is treated as an 8-bit stream, which is enough to provide the appropriate depth data, reducing the required bandwidth further. Once this data is received, a secondary color stream can be synthesized by using a view synthesis method combined with the depth data. When dealing with traditional 3D displays that only require a left and a right stream, the content of the other stream is generated using extrapolation from the available stream and depth image-based rendering (DIBR).

This method is further extended for multiview content by sending several color streams and their associated depth maps [3]. This approach allows for higher quality views to be generated since interpolation can be used between the available views. Interpolation reduces the visible artifacts generated during the extrapolation process as we have more available data that can be used to fill in occluded areas in synthesized views as well as reduce the amount of shifting by half compared to that needed when only one view is available. This also allows for greater 3D perception as the shifting amount of objects between views can be increased without introducing as many artifacts in occluded areas as the extrapolation process (i.e., one view and depth map).

2.5 Industry Standard View Synthesis Techniques

Currently, the MPEG industry standard technique for view interpolation used in their view rendering software (VSRS) implementation uses a DIBR approach combined with a nearest neighbor interpolation approach to generate the synthesized views and fill in any

occluded areas [17][5]. VSRS uses the depth and texture information of the available view or views to generate virtual views, as seen in Figure 2.8.

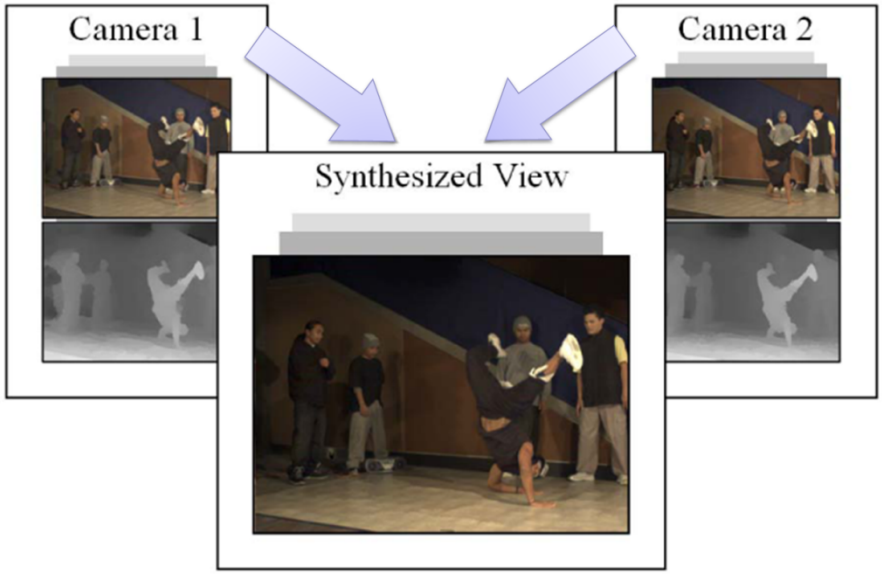
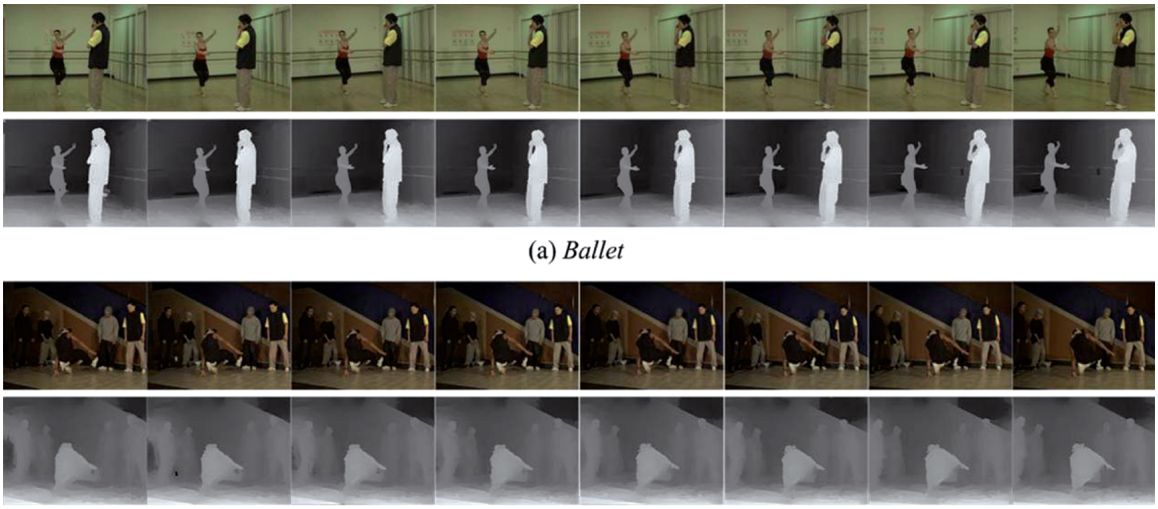


Figure 2.8: Two color streams and their depth maps generate an intermediate synthesized view

The first step of this method, for both interpolation and extrapolation applications, is to look at the provided depth map and generate the synthesized view by shifting objects according to their distance from the viewing plane. This distance is obtained from the depthmap. The depthmap itself resembles a grayscale version of the scene where the furthest objects in the background are black and the closest objects in the foreground are white, with the intermediate objects being a shade of gray representative of their distance from the viewing plane. During the shifting process, the foreground and background objects are segmented using the depthmap data and then are horizontally shifted based on their distance from the viewing plane, with foreground objects being shifter more and background objects being shifted less, to create virtual views. This process creates a synthesized view which has missing color and texture information in background areas

that were occluded by foreground objects. An example of missing texture in occluded regions of an image is shown in Figure 2.10 (green areas).



(a) *Ballet*

(b) *Breakdancers*

Figure 2.9: Two scenes showing the texture information and the corresponding depthmap information for each of the frames ©Optical Engineering

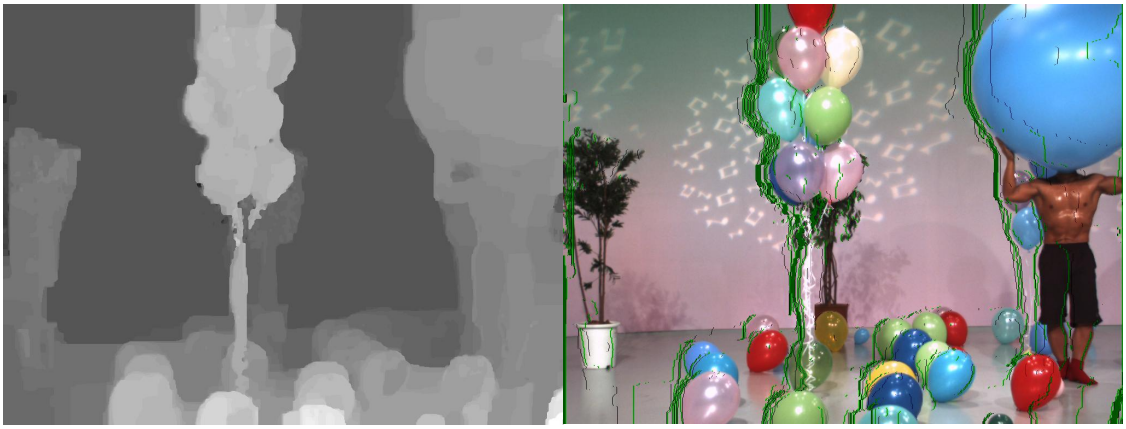


Figure 2.10: Synthesized view after depth-based shifting, before inpainting, with missing texture in occluded regions of the image visible in green

In the case of view interpolation, a secondary view exists which is used to obtain any color and texture information that might be visible from this secondary angle but is occluded in the synthesized view. This data is checked to make sure that it shows the

texture at the same depth level that is missing and, in the case that it does, is then used to fill in parts of the occluded areas during a step that is called view blending. Figure 2.11 shows the view blending step, taking two available views and looking at the available texture data in both to fill in occluded areas in the middle synthesized view. This step is possible due to the fact that the two streams show slightly different angles, thus areas occluded by the foreground objects are different (see Figure 2.12). The final step in view interpolation involves inpainting for any areas that still have missing texture after the view blending step. Inpainting is performed using depth-based, weighted, nearest neighbor interpolation to estimate the value of the missing pixels [18]. This is achieved by averaging the weighted available neighboring pixels to estimate the value of the missing pixel. The weight of each pixel is determined by its proximity to the occluded area as well as the possibility of matching depth information of both the available neighboring pixel and the missing pixel. This allows all occluded areas of the synthesized view that were missing texture data to be filled with data either from the secondary view during the view blending step or with data generated using weighted nearest neighbor pixel interpolation during the inpainting step, (see Figure 2.13).

Figure 2.11: The view blending step takes two available views and looks at the available texture data in both to fill in occluded areas in the middle synthesized view ©César Palomo

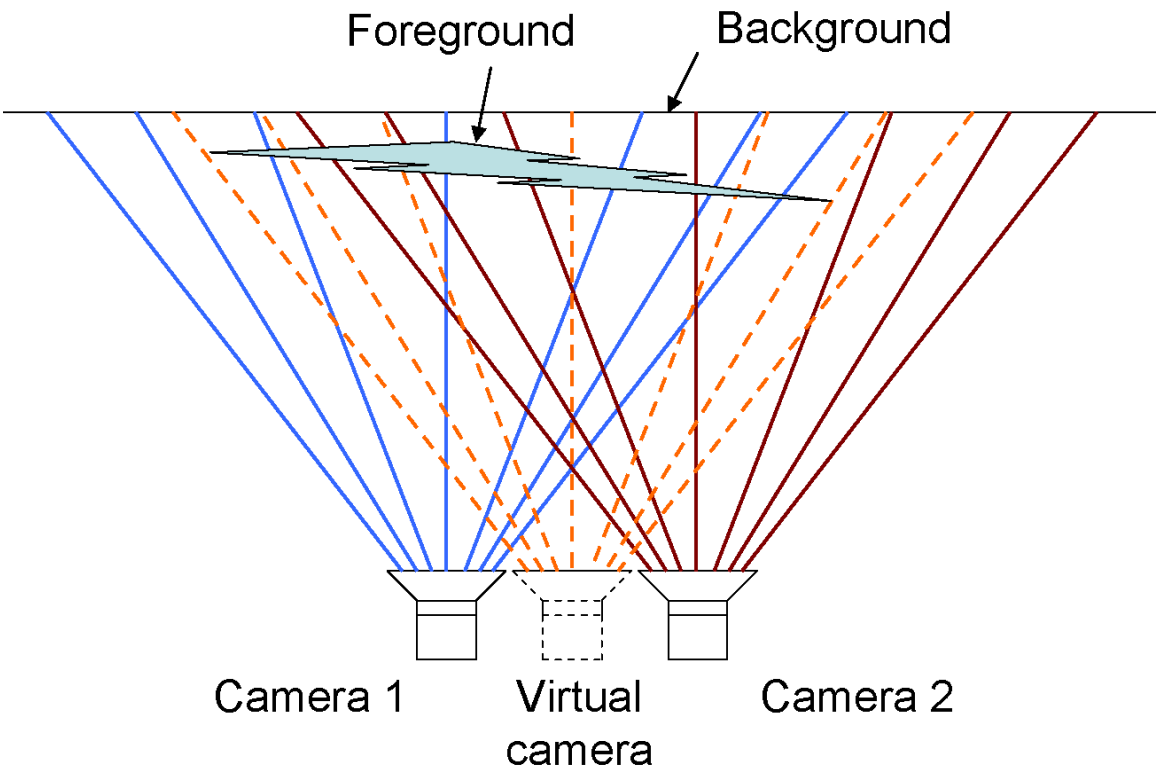


Figure 2.12: View blending is possible since the two available views are at shifted locations creating different occlusions due to foreground objects in each view (©Stanford)

In the case of view extrapolation, only one view is available. Therefore, the view blending step, that uses data from the secondary view to fill in occluded areas, is omitted and all the occluded areas in the synthesized view remain until the inpainting step, which performs depth-based, weighted, nearest neighbor pixel interpolation, as seen in Figure 2.14.

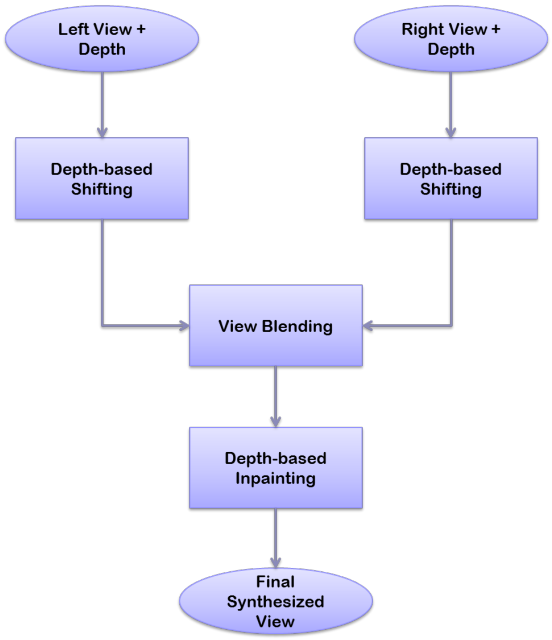


Figure 2.13: Flowchart of the DIBR and inpainting approach used for VSRS view interpolation

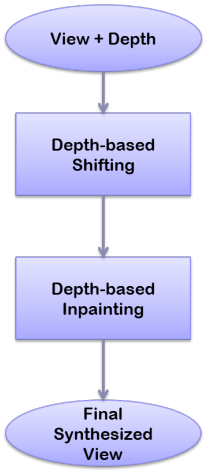


Figure 2.14: Flowchart of the DIBR and inpainting approach used for VSRS view extrapolation

The drawback of this approach is that the interpolation of pixels based on their neighbor values loses texture information and creates a repeating pattern, as can be seen in Figure 2.15. These artifacts may not be very noticeable in the case of smaller occluded areas only a few pixels wide (called cracks). However, in the case of larger occluded areas, known as holes, these artifacts are indeed visible and create a lower quality viewing experience for the end user.

Figure 2.15: Occluded areas and artifacts due to interpolation based hole filling

2.6 Alternate View Synthesis Technique

An alternate technique for view synthesis that does not produce holes or cracks in occluded areas came out of the Disney Research Group. This approach, called Image-

domain-warping-based rendered (IDWR), does not separate and shift objects based on the depth map; rather it stretches various parts of the image to produce a shifted secondary view. This approach uses a sparse saliency map (Figure 2.16) that acts in a similar way to a depth map; that is, the saliency map specifies which objects need to be stretched more or less based on their distance from the viewing plane. The saliency map is automatically generated to provide the warping data for the scene. Due to this automated generation, the saliency map is kept sparse allowing the generation process to produce fewer errors in the final result. The other reason that the saliency map is kept sparse is that, by nature of the warping algorithm, not only the marked areas are stretched but also the adjoining neighboring areas. Therefore, there is no direct need to mark every area in the image to produce proper warping results. The stretching does not produce holes, as seen in VSRS, since objects are not separated for the shifting to occur. Instead, the entire image is warped by stretching or compressing various parts of the scene, as it can be seen in Figure 2.17.

Figure 2.16: Original image and final generated saliency map for a scene ©Disney

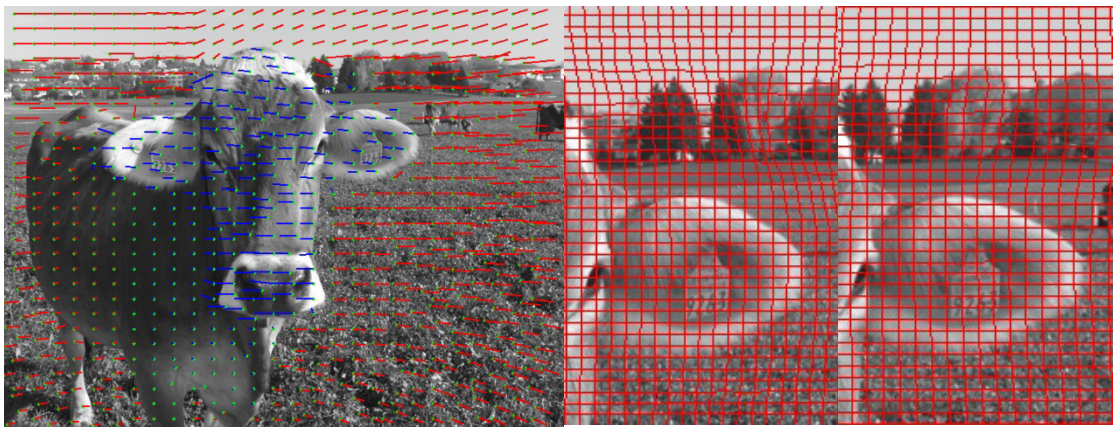


Figure 2.17: Image showing areas that are to be warped and the final result of warping on a zoomed section of the scene with foreground object distortion visible on the right side of the image ©Disney

The result of this warping procedure creates an image that appears to be shifted to the left or right of the original image. This apparent shift produces an effect similar to the shifting that is produced when multiple cameras are used to film 3D, and what is obtained by using VSRS. Since no objects are separated from the main image in this method, and smooth, saliency-driven warping functions are used [19], there are no holes that appear in the final image, so no inpainting is required. Due to this the artifacts that resemble cloned texture, visible in the VSRS method, are not present with the IDWR method.

There are, however, several tradeoffs in the IDWR method. One such tradeoff is that when creating a view far away from the available view(s), where objects have to be warped a significant amount to show their disparity, foreground objects become visibly distorted as can be seen in Figure 2.17 in the case of the ear of the cow. Another tradeoff with this method is that artifacts around straight vertical lines can be seen when there are foreground objects in front of these lines. When the image contains these elements, the foreground objects need to be shifted more than the background objects; this leads to

uneven shifting over the length of these vertical lines. Due to this uneven shifting, lines become wavy rather than straight, as can be seen in Figure 2.18. This artifact too is more prevalent when the amount of warping that has to be performed is more significant.

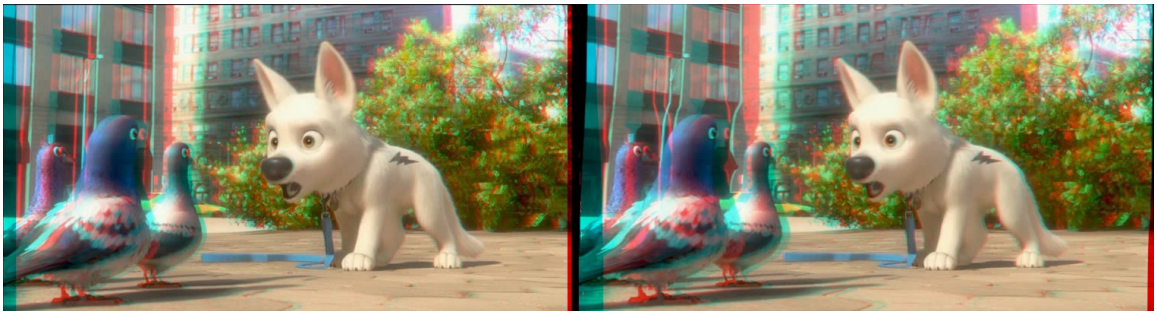


Figure 2.18: For image regions containing objects with strong change in disparity bordering each other, IDWR can create distortions visible on long vertical lines ©Disney

2.7 Conclusion

It widely accepted that the next logical step for advances in video and television is directly linked to 3D video as it provides a much more engaging experience to the end user. Over the last years much research has focused on improving 3D viewing technology to provide a more comfortable and engaging experience for the viewer. New technology no longer requires viewers to wear cumbersome glasses or sacrifice colour reproduction to be able to view 3D video. However, as the viewing technology improves, the amount of available 3D media must grow too. This requires two linked issues to be resolved, the capturing and efficient transmission of 3D media. New advances in viewing technology allow viewers to see 3D on auto-stereoscopic displays, which do not require any sort of 3D glasses, yet for these displays to function properly multiple views of the same scene are required to be available at all times. Both the generation and transmission of these views is highly complex and expensive. A solution to this problem involves generating

synthesized views using one or more real views and the depth information for the objects in those views. The depth information allows the view synthesis software packages to move foreground and background objects to reproduce the natural disparity that would be created if more views of those scenes were generated. The main challenge of this approach is caused by occlusions in the scene that are produced by foreground objects covering background objects. When the foreground objects are shifted, areas of the background for which no texture information exists become visible. These areas need to be either avoided or filled in with generated data. Two separate approaches currently exist, the DIBR approach and the IDWR approach which try to resolve this problem. DIBR uses depth data to move foreground and background objects at varying degrees to recreate the natural disparity of the objects. The occluded areas that are missing texture at the end of the shifting process are filled in using nearest neighbor pixel interpolation. This produces data in these areas that closely match the average colour of the neighboring pixels, yet most texture information is lost during this process. The IDWR approach attempts to address this problem by simply warping the entire image and not segmenting and shifting separate objects. This resolves the issue with missing texture in occluded areas but at the expense of introducing warping artifacts to the scene that make some objects appear stretched and distorts straight lines. In light of this, there is a need for further research in the area of view synthesis in order to reduce or eliminate the artifacts introduced by these techniques and provide higher quality synthesized views which will in turn improve the overall quality of the 3D video.

3 Hybrid View Interpolation

Content that is produced for multiview displays requires several views to be available for each scene so that multiple viewers enjoy watching 3D content without wearing 3D glasses. Capturing 3D content is much more complicated with multiview displays that require 8 or more views compared to traditional 3D displays that only require two views to display 3D. Capturing all required views is technically challenging, impractical and costly. This issue becomes more pronounced in the case of advanced multiview displays, which use a large number of views to allow for smooth transition between sweet spots. In this case, the existing 3D content which includes limited number of views becomes obsolete and cannot be watched on the advanced multiview displays as all the required views are not available. To address these inherent problems with the multiview content generation, the 3D Video (3DV) ad-hoc group (part of ISO/IEC Moving Pictures Experts Group (MPEG)), recommended capturing a limited number of views (two or three views) plus their respective depth maps (3DV data format) and synthesize the rest of views for multiview applications [3]. While it might result in a reduced 3D (depth) effect, this approach at least offers an cost/bandwidth effective and practical solution and guarantees that existing content will not become obsolete. With the above in mind, we propose an effective view interpolation approach (see Figure 3.1) which includes the following steps.

3.1 Creating Primary Synthesized View

In the view-synthesis problem, several different views are captured with multiple cameras (usually with a parallel setup) and additional views are synthesized from the available ones as if there were more cameras in the multiview camera setup. The closer

the real camera views are to the virtual camera view the more accurate the synthesized view is. For this reason, in our approach we create a primary synthesized view based on the closest available real camera view to the location of the synthesized view and the depth map of that real camera view. The appropriate shifting amount for different objects in the scene is calculated using the depth and texture information as follows [20]:

$$p_{pix} \approx -x_B \frac{N_{pix}}{D}\left(\frac{m}{255}\left(k_{near} + k_{far}\right) - k_{far}\right) \qquad (1)$$

where ppix is the shift parameter at depth level m, D is the viewer distance from the display, knear and kfar are the distances of the closest and farthest objects to the camera, and Npix is the user defined parameter controlling the maximum parallax based on the screen width. The maximum parallax determines the depth of the closest object in the scene when watched on the screen. This shifting process creates holes (pixels with missing color and texture information) in a way similar to a regular interpolation-based synthesizing approach [8]. Our next step is to create a binary mask of the synthesized view. This mask is used to register the coordinates of all the pixels in the synthesized view that correspond to holes, assigning the value of zero for the hole pixels and the value of one for the rest of the pixels. This mask is called "Mask I" (see Figure 3.1).

3.2 Matching-Based Hole Filling

In order to fill up the holes, the first step is to use the information of the available view that is farther from the position of the virtual camera (synthesized view). To this end, a secondary synthesized view is generated solely based on the farther view by following the same procedure as creating the primary synthesized view above. Once the secondary synthesized view is generated, the holes in the primary synthesized view

(registered in Mask I) are filled by corresponding available areas in the secondary synthesized view. At this point we have to make sure that the texture we will use from the secondary synthesized view belongs to the same object/background as in the primary synthesized view. This is achieved by keeping track of the depth value of the shifted pixels in the synthesized views and then matching them to the areas around the holes. Note that this condition is not always met, since the secondary view is generated based on the farther available view which covers different areas in the scene. At this point we create another binary mask that is used to register the coordinates of the remaining holes in the primary synthesized view; we call this mask "Mask II" (see Figure 3.1).

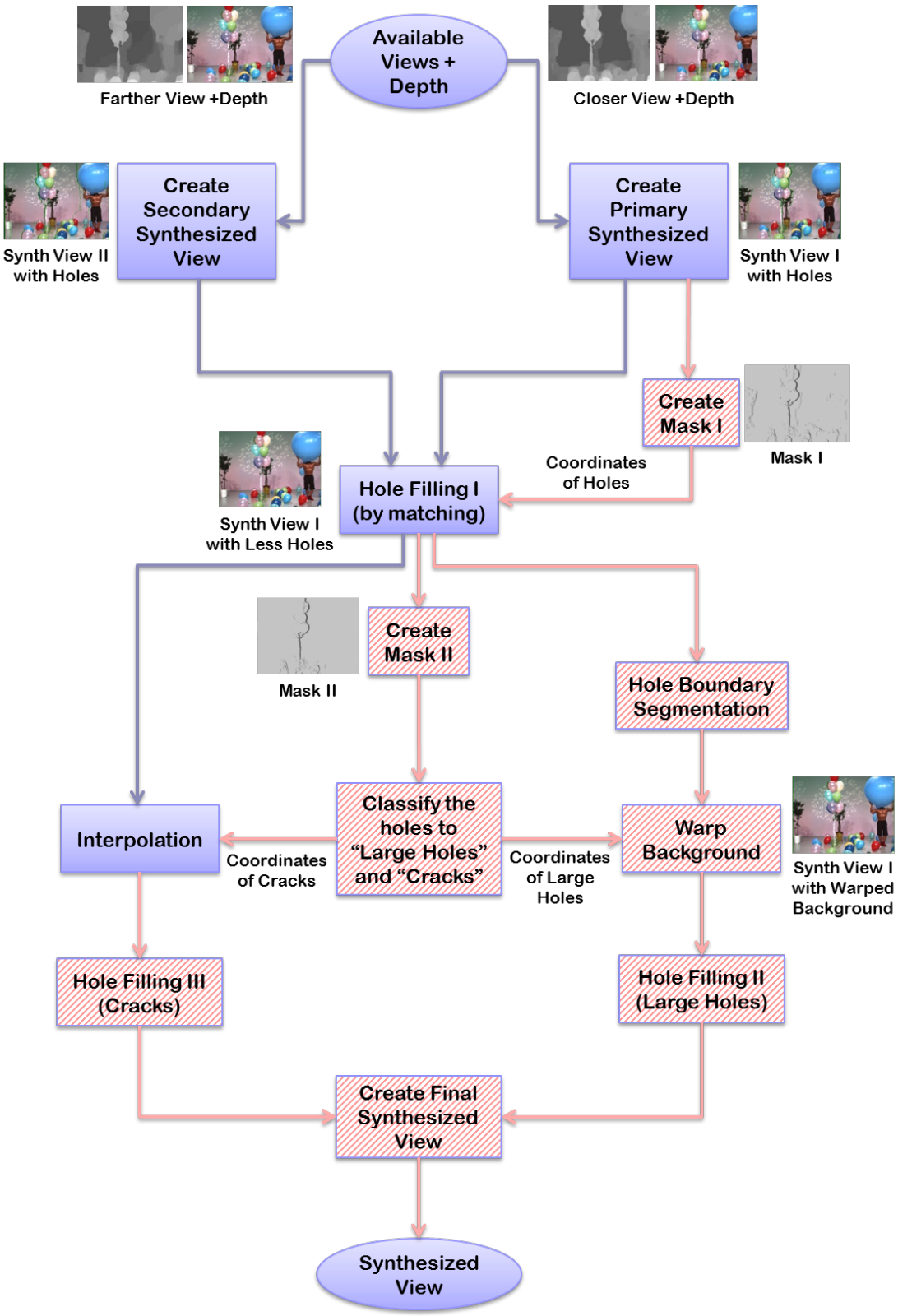


Figure 3.1: Flowchart of the hybrid view synthesis technique for interpolation; red (striped) blocks are steps unique to our method, blue blocks are hybrid steps that are based on existing methods and modified to fit our approach

3.3 Hybrid Hole Filling by Interpolation and Warping

In general, in the background areas of the synthesized view generated after the two steps described above there are smaller holes (due to small disparity), while in the foreground areas the holes are larger. In addition, some small holes are also created due to depth map imperfection in both background and foreground areas. To fill up these remaining holes we can use either warping or interpolation. Considering that warping is a time consuming and computationally expensive operation, and the size of some of the holes is too small (visually unnoticeable), we propose to use warping to fill in only large holes and for small holes we use nearest neighbour interpolation. To this end, we classify the holes in the generated synthesized view into two distinct categories based on a trade-off between visual quality and computational complexity:

a) Cracks (areas of width less than a set threshold) to be filled in by nearest neighbour interpolation

b) Large holes (areas greater than the set threshold) to be filled in by warping

The threshold for this classification was determined through subjective/empirical tests over a large number of representative videos. We found that holes smaller than 0.2% (called cracks here) of the frame-width may be filled by using nearest neighbor interpolation without any noticeable artifacts. For instance, for a high definition video of 1080x1920 any hole area with the width of less than 3.8 pixels is classified as “crack”, while the hole areas with the width of greater than 3.8 pixels are classified as “large hole”. For view interpolation of the scenes that were used in our tests, on average, 63.13% of the pixels in the occluded areas could be filled in by match based hole-filling. Of the remaining pixels, 16.23% were classified as cracks while 20.64% were classified

as holes. In summary, this classification effectively enhances the speed of the hole filling process without hampering the resulting visual quality. It can be observed that an additional classification based on the amount of texture in the background could be performed to further increase the speed of the computation as areas with no texture can be filled in with nearest neighbour pixel interpolation with no visible loss in quality. However, in practice, there are few scenes that contain completely texture-less background; in this case, the additional computational load of this further classification becomes unreasonable.

Once the holes are classified, the cracks are filled by applying nearest neighbor interpolation (which is similar to the approach used by VSRS).

To fill the large holes and preserve texture information, our hybrid approach applies warping to the background area of the synthesized view generated using the closest and farther views. This process involves two concurrent steps: 1) segmenting the boundaries of large holes and 2) generating warp points. Hole boundary segmentation is performed in horizontal direction to classify the adjacent pixels around each hole as foreground and background (using depth information), allowing us to perform warping from the background towards the foreground. Using this approach we avoid deforming foreground objects, which are usually visually important. The other step involves using the coordinates of large holes in Mask II to generate the list of warping points. The warping start-points (the points where the hole-areas start with a small overlap towards the background) and the warping end-points (the points where the hole-areas end with a small overlap towards the foreground) are identified using Mask II (coordinates of large holes only). In our warping process we need a full list of warp points to generate a

smoother warped image rather than a small subset, which can lead to deformities being generated in the warped image. To avoid vertical parallax, the warping process for filling the holes should be done in the horizontal direction, i.e., the vertical coordinate of the warping start-point and end-point are equal. We also restrict the warping process to not use the information of the corners of the synthesized image. This is because there is not enough texture data at the corners that can guarantee effective warping. Warping is performed by applying the Piecewise cubic Hermite interpolation [21] algorithm to the generated synthesized view in Section 3.2. Piecewise-Cubic Hermite interpolation constructs a interpolant of the data points $(x_1, y_1), \ldots, (x_n, y_n)$ by combining the local cubic interpolants as follows:

$$H_i(x) = a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^2(x - x_{i+1}) \qquad (2)$$

into a global interpolant:

$$H(x) = \begin{cases} H_1(x) & if\ x_1 \leq x < x_2 \\ \vdots & \vdots \\ H_{n-1}(x) & if\ x_{n-1} \leq x < x_n \end{cases} \qquad (3)$$

where H is the Hermite interpolation function and a, b, c, and d define the intervals for the interpolation. We apply this algorithm to the entire image by passing in the coordinates of the points on one side (background) of the large hole areas as well as the coordinates of the points on the other side (foreground) of the hole areas in horizontal direction. Using these coordinates the Piecewise Cubic Hermite Interpolating Polynomials $H_1, \ldots, H_{n-1}$ are calculated. Then we use the calculated Hermite polynomials as the global Hermite interpolant function H to obtain the values of every pixel in the warped image.

3.4 Creating Hybrid Synthesized View

The large hole areas in the generated virtual view in Section 3.2 (which are marked in Mask II) are filled with the data from the warped image. Note that in our approach the crack-filling and background-warping are performed in parallel and as stand-alone procedures. In other words, warping is not applied to the virtual view with filled cracks. This is because the cracks are filled with some estimated and not true information, which may cause errors in the warping process if they are used.

Once this process is complete, we obtain a virtual view where all the holes are filled either with data from the interpolation or from the warped image. As it can be observed from Figure 3.2, unlike VSRS, our hybrid approach generates more realistic texture for hole areas without hampering the quality of visually important foreground objects. At the same time, as can be observed in Figure 3.3, we avoid foreground warping artifacts, which are visible in Disney's approach, by segmenting the large hole boundaries into foreground and background and starting the warping from the background area.

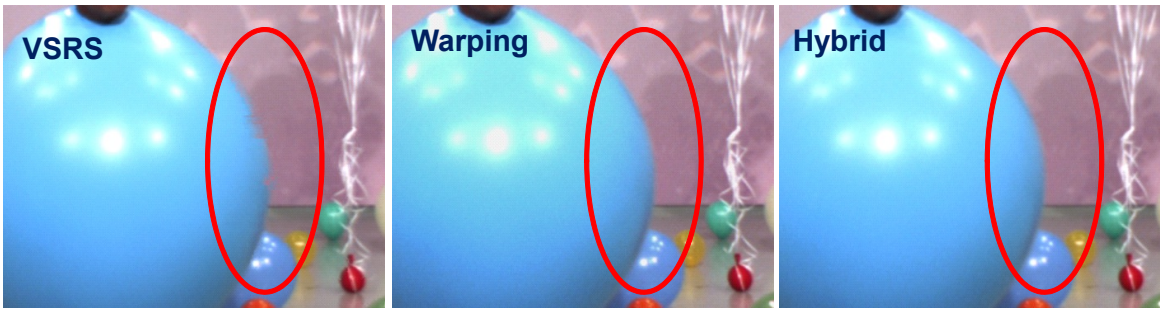


Figure 3.2: Comparison of view interpolation synthesized views generated by VSRS, Disney's Warping approach, and our Hybrid approach

Figure 3.3: Comparison of view interpolation synthesized views generated by VSRS, Disney's Warping approach, and our Hybrid approach

3.5 Experimental Setup

The performance of our method is evaluated based on subjective tests and is compared to that of the existing VSRS package (version 3.5) [8] as well as the Disney warping method [6]. For this evaluation we used three test sequences, namely "Balloons" (1024x768, 30fps, 300 frames), "Kendo" (1024x768, 30fps, 300 frames), and "GT_Fly" (1920x1088, 25fps, 250 frames). These test streams along with their depth information are selected from the database provided by MPEG for the Call for Proposals (CfP) on 3D video coding [22]. All the videos are in YUV 4:2:0 format and progressive. The synthesized views, the stereo pair used in our tests, and the real input views used for the view interpolation are presented in Table 3.1.

Seq. ID	Test Sequence	Input views	Synthesized view(s)	Stereo pair
S01	Balloons	3-5	4	3-4
S02	Kendo	3-5	4	3-4
S03	GT_Fly	3-5	4	3-4
S04	Balloons	1-3	2	2-3
S05	Kendo	1-3	2	2-3
S06	GT_Fly	1-3	2	2-3
S07	Balloons	1-3-5	2-4	2-4
S08	Kendo	1-3-5	2-4	2-4
S09	GT_Fly	1-3-5	2-4	2-4

Table 3.1: Input views, synthesized views, and stereo pair for view interpolation 2-view test scenario.

The viewing conditions were set according to the ITU-R Recommendations BT.500-13 [23]. All volunteer subjects were screened for color and visual acuity (using Ishihara and Snellen charts), and for stereo vision (Randot test – graded circle test 100 seconds of arc). All subjects had none to marginal 3D image and video viewing experience. Table III summarizes the information about the participants in our tests. Note that some of the participants have attended more than one subjective test.

The evaluation was performed using a 46" Full HD Hyundai 3D TV (Model: S465D) with passive glasses. The TV settings were as follows: brightness: 80, contrast: 80, color: 50, R: 70, G: 45, B: 30. The 3D display and the settings are based on MPEG recommendations for subjective evaluations of the proposals submitted in response to the 3DV CfP [22].

At the beginning of each evaluation session, a demo sequence ("Undo_Dancer", 1920x1088, 25fps) with different levels of synthesizing artifacts was played for the

subjects to become familiar with the artifacts and the testing process. The process of rating the sequences was explained during that time so that the subjects would know the rating scheme for the test. A five second break interval was shown at the end of each demo sequence, informing the subjects that the next sequence they see should be rated. The "Dancer" test sequence was then omitted from the actual evaluation procedure to maintain the purity of the results.

After training, viewers were shown the synthesized stereoscopic test sequences in random order. This insured that they would watch different synthesized versions of the same sequence, without knowing the video was generated by our hybrid method, the MPEG provided VSRS package, or the Disney's proposed warping approach. Between test videos, a ten-second gray interval was provided to allow the viewers to rate the perceptual quality of the content and relax their eyes. Here, the perceptual quality reflects whether the displayed scene looks pleasant in general. In particular, subjects were asked to rate a combination of "naturalness", "depth impression" and "comfort" as suggested by Hyunh-Thu et al. [24]. For ranking, there were 10 quality levels, 10 indicating the highest quality and 1 the lowest quality. In our subjective study, the performance of our Hybrid view interpolation was compared with those of VSRS and Disney's warping view interpolation. Three scenarios were examined in our tests:

1) right-view is synthesized

2) left-view is synthesized

3) both views are synthesized

Switching the synthesized view between the right and the left eye compensated for the effect of eye dominance. We repeated similar tests for view extrapolation. The last test we performed involved comparison of the performance of hybrid view extrapolation with VSRS view interpolation. The motive for this study was to examine if our method is a good candidate for situations where bandwidth drops and we can only send the information of one view. The results of our test are discussed in the following section.

3.6 Experimental Results

The first step after collecting the subjective evaluation results was to check for and remove outliers according to the ITU-R Recommendations BT.500-13 [21]. See Table 3.2 for the number of outliers per each test. We then calculated the mean opinion scores (MOS) from the viewers with a 95% confidence interval.

Tests		**Number of Subjects**	**Eye Dominance**		**Age Range**	**Number of Outliers by Eye Dominance**	
			Right	**Left**		**Right**	**Left**
View Interpolation	**Hybrid vs. VSRS**	20	7	13	18-57	2	0
	Hybrid vs. Disney	21	14	7	21-26	1	2

Table 3.2: Details about the participants in our subjective tests for view interpolation 2-view test scenario.

The evaluation results for comparing our hybrid interpolation approach to VSRS for the three video sequences and for three test scenarios: only right-view is synthesized, only left-view is synthesized, and both views are synthesized are shown in Figures 3.4a, 3.4b, and 3.4c. Figures 3.5a, 3.5b, and 3.5c show the mean opinion scores for our hybrid method against the Disney's warping approach for the same videos and same scenarios.

The black bar on each graph shows the 95% confidence interval viewers. Figures 3.4d and 3.5d show the average values for all three sequences for the two different comparisons. As it can be observed, for all the sequences the scenes generated using our hybrid approach scored consistently higher than those generated using MPEG's VSRS package or the Disney's Warping approach, confirming the superior performance of our technique. Even in the case where both views are synthesized, the MOS score for our hybrid approach is higher than that of two other approaches. In fact, the subjective tests show that the MOS scores for the case where both views are synthesized using our hybrid approach are similar or even higher than those for the case where only one view is synthesized by MPEG's VSRS or Disney's Warping.

We also observe from Figures 3.4d and 3.5d that in general the MOS scores for all evaluated methods are slightly lower for the test scenario where both views are synthesized than for the test scenario where only one view is synthesized. This is due to binocular rivalry [25]. In cases where only one view is synthesized, the information of the dominant picture (original view in our case) suppresses the information of the less-dominant view (synthesized view), thus the perceived quality of the overall picture is higher than in the case where both views are synthesized.

An interesting general observation from the results comparing VSRS to our hybrid approach (Figures 3.4a, 3.4b, 3.4c, and 3.4d) is that for the cases where the right view is synthesized the MOS is higher than the cases where left view is synthesized. We believe this can be explained by the fact that we had more left eye dominant subjects for our VSRS interpolation tests (see Table 3.2) so that when the synthesized view was shown to their left eye, the artifacts affected their 3D perception, and they rated the overall quality

much lower compared to the case that the synthesized view was shown to their right eye (non-dominant eye). It can be observed that this is not the case for the Warping evaluation results (Figures 3.5a, 3.5b, 3.5c, and 3.5d) where the number of right eye dominant subjects was greater than the number of left eye dominant subjects (see Table 3.2).

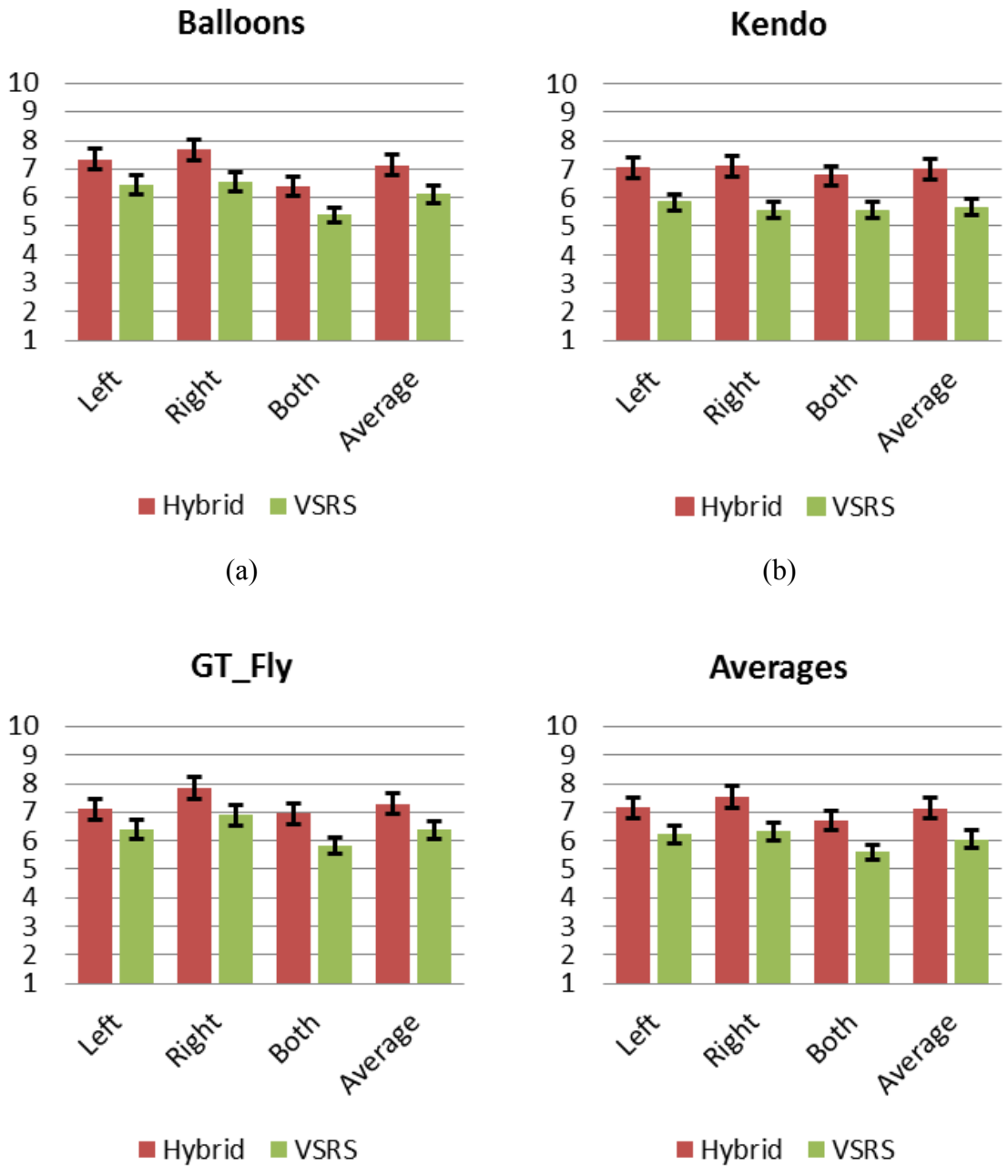


Figure 3.4: MOS for interpolation view synthesis evaluations of VSRS vs. our Hybrid method. The black bar on each graph shows the 95% confidence interval

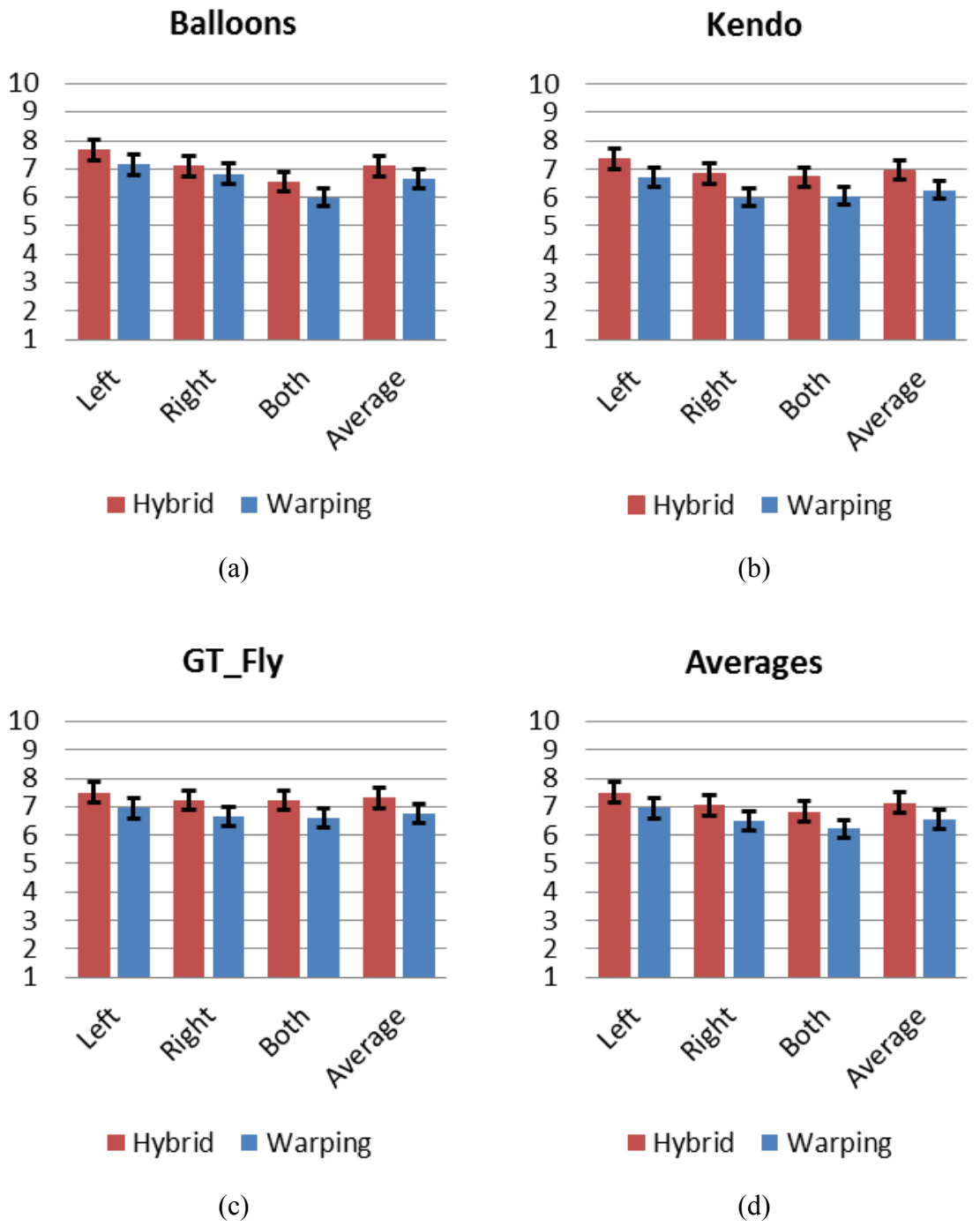


Figure 3.5: MOS for interpolation view synthesis evaluations of Warping vs. our Hybrid method. The black bar on each graph shows the 95% confidence interval

3.7 Conclusion

We proposed a new and unique hybrid view-synthesis method that addresses the limitations of the existing view-synthesis interpolation techniques. This hybrid method takes general ideas of depthmap-based shifting present in DBIR and image warping used in IDWR to create a unique approach that uses depthmap data to shift foreground objects and warping to fill in occluded areas of the final synthesized view. The initial depthmap based shifting improves on the IDWR approach by preserving the overall look of foreground objects and avoiding warping artifacts that resemble stretching of these objects. At the same time, the unique background-to-foreground warping process improves on the DBIR based VSRS approach by filling in occluded areas that contain missing texture with warped background texture. This preserves much more of the unique texture information that may exist in these areas at the expense of slightly deforming background objects. Subjective tests of our proposed hybrid method that compared it to IDWR and VSRS show that viewers find the synthesized views generated by our method to be of noticeably higher quality than the other methods.

4 Hybrid View Extrapolation

As mentioned earlier, success of the 3D technology is highly dependent on content availability. One way of resolving this problem is to convert existing 2D content to 3D format. This is similar to the old days when color television sets were introduced and the content producers manually colored in some black and white movies to resolve the issue of color content availability. In the 2D to 3D conversion process, first the depth information is estimated, and then it is used together with the existing 2D video to synthesize the second view via view extrapolation (see Figure 4.1). Depth map information can be manually extracted from the existing 2D content, or can even be automatically generated [4]. The issue with view extrapolation is that we only have one view, and we cannot use other views to extract the information of occluded areas. This in general results in holes larger than the ones in the interpolation case discussed before, which in turn present a much bigger challenge when it comes to synthesizing an accurate view. In our view extrapolation approach some of the steps are similar to our view interpolation with the exception of the matching based hole filling step (Section 3.2) which cannot be applied due to the absence of the second (farther) view. The details of our view extrapolation process are discussed in the following subsections.

4.1 Creating Synthesized View

The first part of our proposed technique uses the depth and texture information to create a virtual view in a way similar to that of the VSRS approach. Shifting of the objects based on their depth is performed in the same manner as that described in Section 3.1. For extrapolation, the shifting is done either to the left or to the right of the real view.

The direction of that shifting is decided based on the position of the virtual camera with respect to the real camera. The shifting amount is calculated based on equation (1). The main distinction from the previously described view interpolation method is that, since only one real view is available, only one intermediate synthesized view is created here.

Similar to view interpolation approach the coordinates of all the pixels in the synthesized view that correspond to holes are registered by a binary Mask, called Mask I. Also like in the view interpolation process, the holes are categorized based on their width as "Cracks" or "Large" holes. We use the same threshold as the one for view interpolation here (i.e., 0.2% of the frame-width). For view extrapolation of the scenes that were used in our tests, on average, 43.82% of the pixels in the occluded areas were classified as cracks while 56.18% were classified as holes. In addition to that, as mentioned before, the number of cracks and the size of "large" holes are much higher in the case of view extrapolation compared to view interpolation due to the absence of extra views. Thus, classifying the holes into the two categories and using hybrid interpolation in this case tremendously reduces the computation load – much more than the view interpolation case.

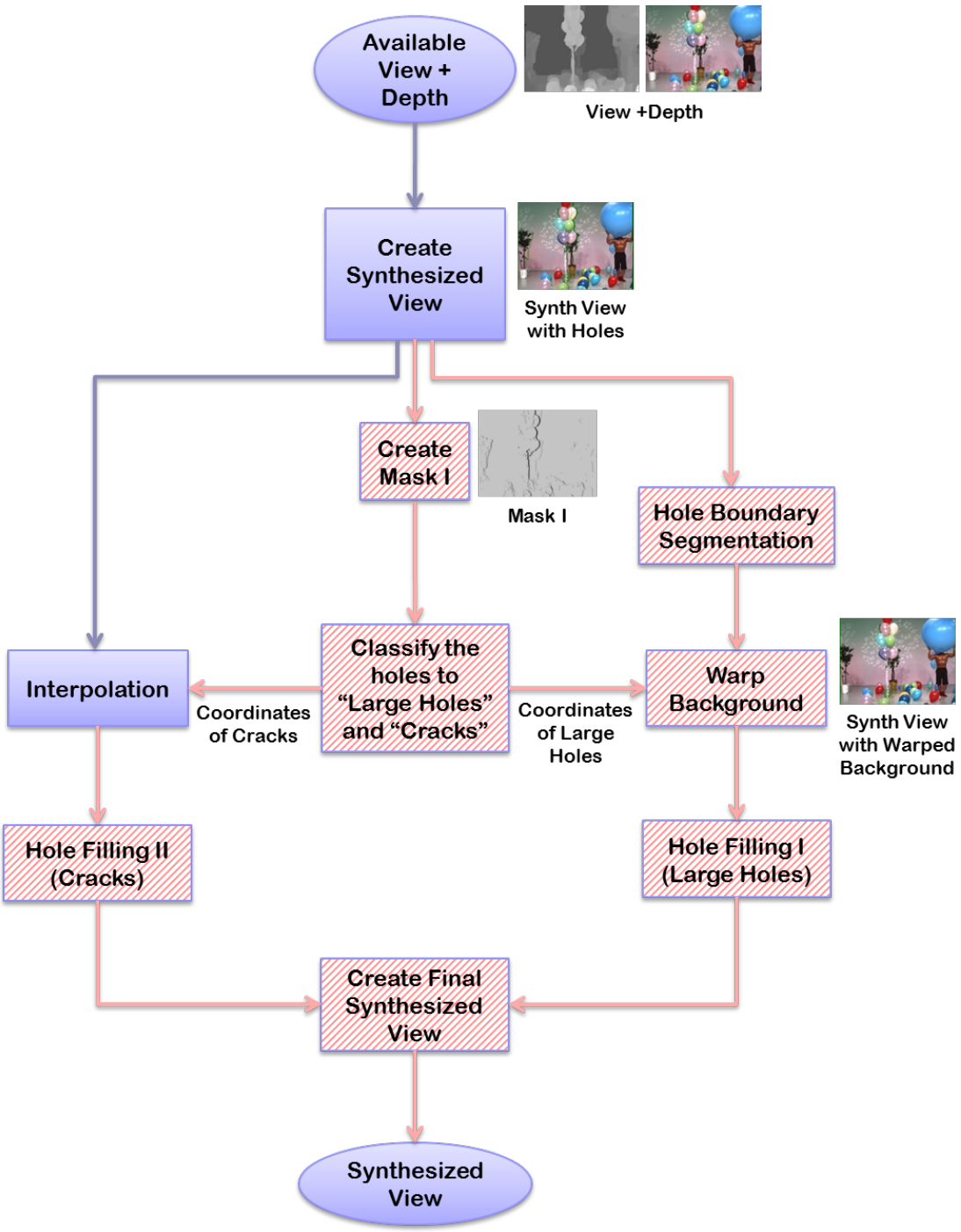


Figure 4.1: Flowchart of the hybrid view synthesis technique for extrapolation; red (striped) blocks are steps unique to our method, blue blocks are hybrid steps that are based on existing methods and modified to fit our approach

4.2 Hybrid Hole Filling by Interpolation and Warping

The holes which are categorized as cracks in the synthesized view (smaller than 0.2% of the frame width) are filled by nearest neighbor interpolation (which is similar to the approach used by VSRS). To fill large holes, similar to view interpolation we apply warping to the synthesized view generated in Section 4.1. Once more, in the warping process we ensure that warping starts from the background points towards foreground (see Section 3.3 for details), thus avoiding deforming the foreground objects, which are usually visually important, and hence improving the visual quality of the final synthesized view.

4.3 Creating the Final Synthesized View

Once the warped image is obtained, the areas of the synthesized view (with filled cracks), which are categorized as large holes in Mask I, are filled with the corresponding areas in the warped image, and the hybrid synthesized view is created. As shown in Figure 4.2, the synthesized views generated by our view extrapolation method contains more realistic texture data in the hole areas compared to VSRS extrapolation. Also, unlike Disney warping method, our approach does not deform the foreground objects since it only warps the background area (see Figure 4.3).

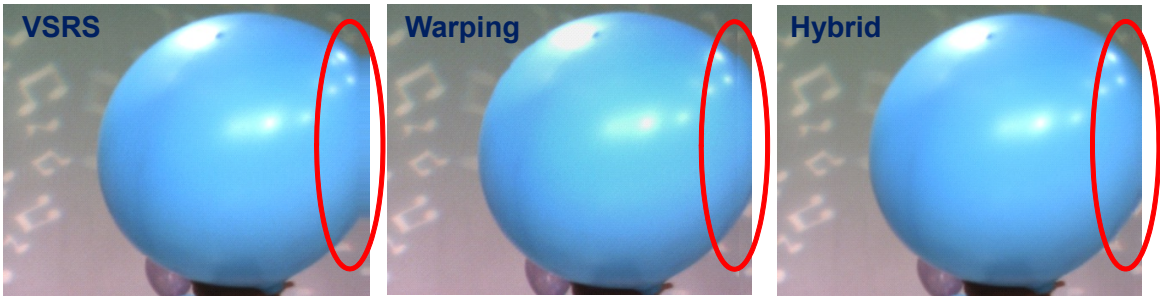


Figure 4.2: Comparison of the extrapolated synthesized views generated by VSRS, Disney's Warping approach, and our Hybrid approach

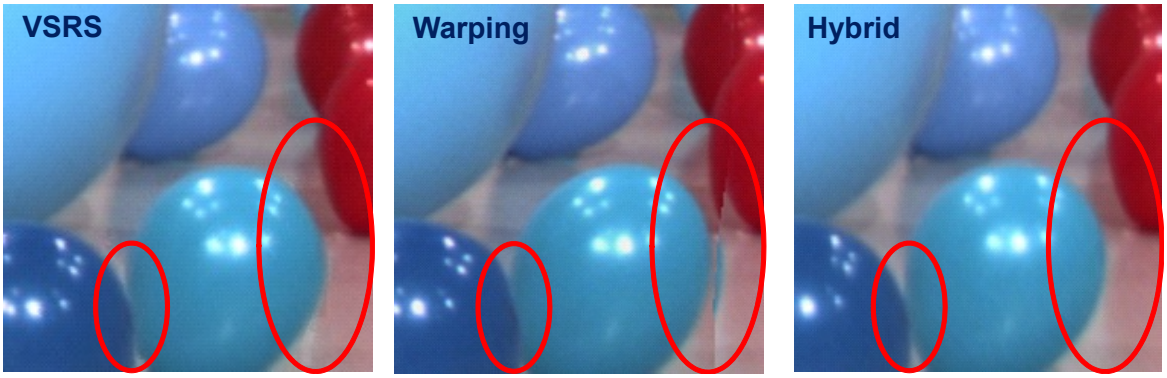


Figure 4.3: Comparison of the extrapolated synthesized views generated by VSRS, Disney's Warping approach, and our Hybrid approach

4.4 Experimental Setup

The performance of our method is evaluated based on subjective tests similar to Section 3.5 and is compared to that of the existing VSRS package (version 3.5) [8] as well as the Disney warping method [6]. The setup used the same settings for the viewing conditions, which were set according to the ITU-R Recommendations BT.500-13 [23], and the same TV settings based on MPEG recommendations for subjective evaluation of the proposals submitted in response to the 3DV CfP [22]. All the subjects were also screened for color and visual acuity, and for stereo vision similar to the previous experiments.

The subjects were trained for the evaluation session using the same approach as described in Section 3.5. The demo sequence used was again "Undo_Dancer"; however, it was specifically remade for this evaluation to show the different levels of artifacts present in views synthesized using the view extrapolation approach. The test videos and synthesized views used for this evaluation are shown in Table 4.1.

We also performed evaluations to compare our synthesized views generated using view extrapolation with the industry standard synthesized views generated using view interpolation to check the extent of the benefits of our proposed view synthesis algorithm. The test videos and synthesized views used for this evaluation are shown in Table 4.2.

Seq. ID	Test Sequence	Input view	Synthesized view(s)	Stereo pair
S01	Balloons	3	4	3-4
S02	Kendo	3	4	3-4
S03	GT_Fly	3	4	3-4
S04	Balloons	3	2	2-3
S05	Kendo	3	2	2-3
S06	GT_Fly	3	2	2-3
S07	Balloons	3	2-4	2-4
S08	Kendo	3	2-4	2-4
S09	GT_Fly	3	2-4	2-4

Table. 4.1: Input views, synthesized views, and stereo pair for view extrapolation 2-view test scenario.

Seq. ID	Test Sequence	Input view (Hybrid)	Input views (VSRS)	Synthesized view(s)	Stereo pair
S01	Balloons	3	3-5	4	3-4
S02	Kendo	3	3-5	4	3-4
S03	GT_Fly	3	3-5	4	3-4
S04	Balloons	3	1-3	2	2-3
S05	Kendo	3	1-3	2	2-3
S06	GT_Fly	3	1-3	2	2-3
S07	Balloons	3	1-3-5	2-4	2-4
S08	Kendo	3	1-3-5	2-4	2-4
S09	GT_Fly	3	1-3-5	2-4	2-4

Table 4.2: Input views, synthesized views, and stereo pair for view extrapolation vs. view interpolation 2-view test scenario.

4.5 Experimental Results

The first step after collecting the subjective evaluation results was to check for and remove outliers according to the ITU-R Recommendations BT.500-13 [21]. See Table 4.3 for the number of outliers per each test. We then calculated the mean opinion scores (MOS) from the viewers with a 95% confidence interval.

Tests		Number of Subjects	Eye Dominance		Age Range	Number of Outliers by Eye Dominance	
			Right	Left		Right	Left
View Extrapolation	Hybrid vs. VSRS	18	10	8	21-28	0	0
	Hybrid vs. Disney	20	13	7	22-31	0	2
Hybrid view extrapolation vs. VSRS view interpolation		24	13	11	21-57	1	1

Table 4.3: Details about the participants in our subjective tests for view extrapolation as well as the view extrapolation vs. view interpolation 2-view test scenario.

In our evaluations for comparing our hybrid view extrapolation to VSRS and Disney's approach, we calculate the mean opinion score with a 95% confidence interval (as in the evaluations for view interpolation). The evaluation results are shown in Figures 4.4 and 4.5 for the three video sequences and for three test scenarios:

1) only right-view is synthesized

2) only left-view is synthesized

3) both views are synthesized

Figures 4.4a, 4.4b, and 4.4c show the MOS for our hybrid method against the VSRS method. Figures 4.5a, 4.5b, and 4.5c show the MOS for our hybrid method against the Disney's Warping approach for the same videos and same scenarios. As the subjective results show, our hybrid approach, again, scored consistently higher than both MPEG's VSRS method and Disney's Warping approach. Even in the case where both views are synthesized, the MOS for our hybrid approach is higher than that for other two methods.

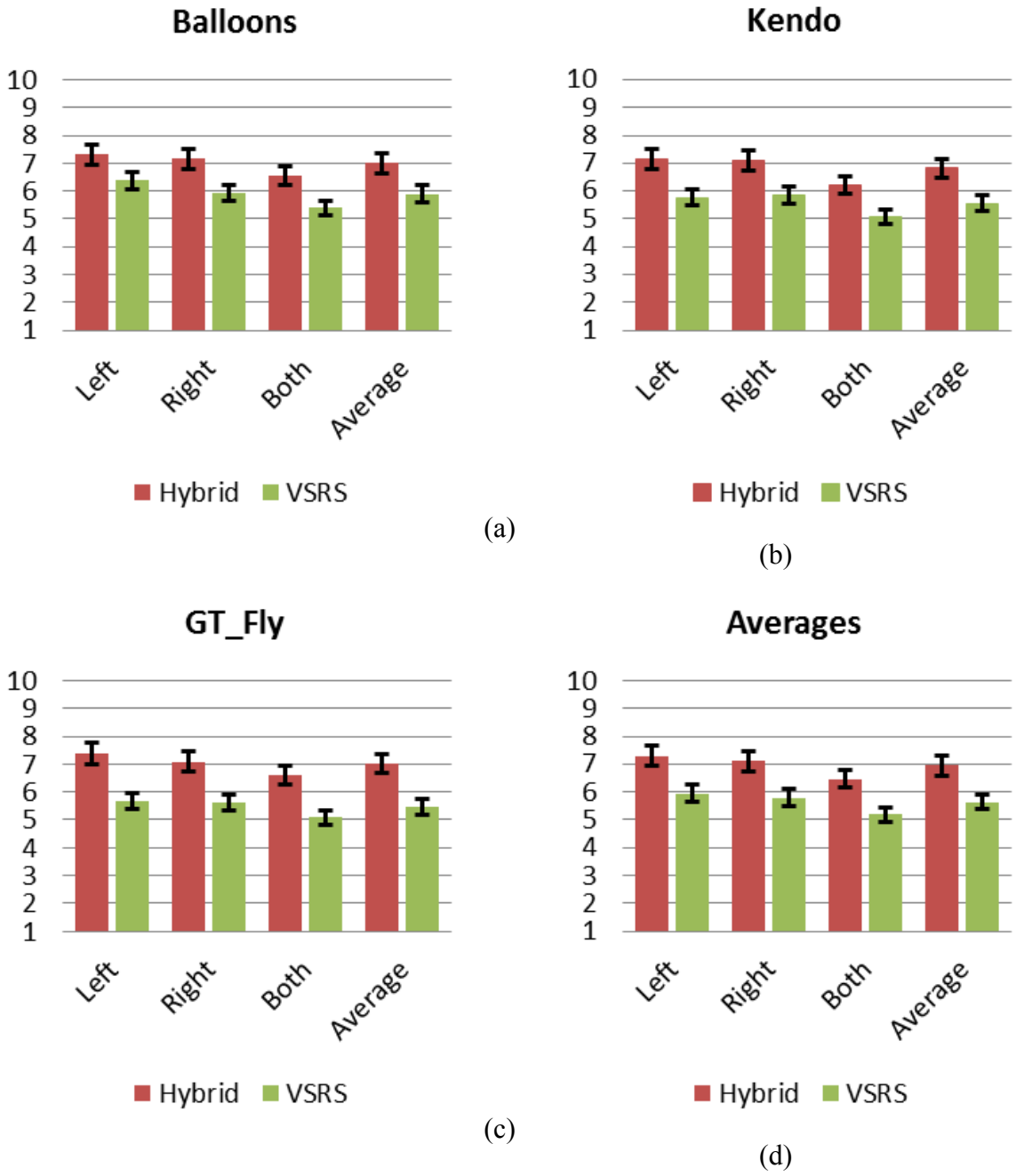


Figure 4.4: MOS for extrapolation view synthesis evaluations of VSRS vs. our Hybrid method. The black bar on each graph shows the 95% confidence interval

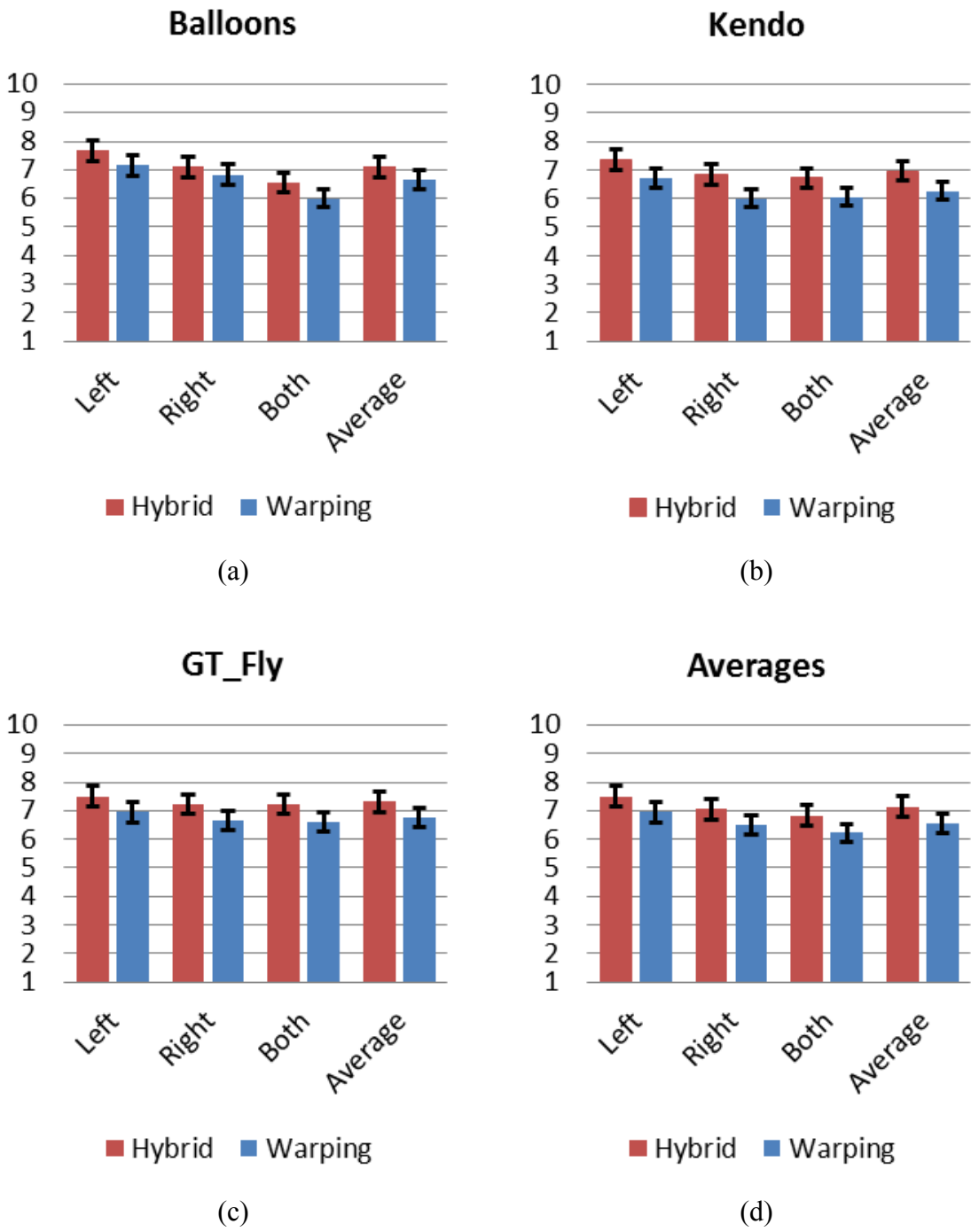


Figure 4.5: MOS for extrapolation view synthesis evaluations of Warping vs. our Hybrid method. The black bar on each graph shows the 95% confidence interval

It is known that synthesized views generated via view-interpolation have higher quality compared to the ones generated by view-extrapolation. This is due to the fact that there is more data available from the additional views in the view-interpolation case. In this case, two real views are used to generate the virtual view(s) between them. Thus, there are fewer and/or smaller hole areas in the synthesized views as objects or parts of them occluded in one real view may be visible in the other real view. In fact, the hole areas produced by occlusions in one real view can be filled with the information from the other real view which allows hole-filling with true texture and color. This effectively reduces the size and number of hole areas that need to be filled with interpolated data from neighbouring regions. In the case of view extrapolation, only one view is available, so any occluded areas need to be completely filled with generated texture. This is due to no additional views being available that would include the full or partial information of the occluded areas.

Although the perceptual quality of generated views via view extrapolation is lower than of the ones created by view interpolation, in cases where the bandwidth is limited and transmission of the additional views is impossible, using view extrapolation is inevitable. For this reason, it would be beneficial to attempt to minimize the quality reduction in these cases. To this end, we performed an additional evaluation to compare the quality of the extrapolated views generated by our Hybrid approach with that of interpolated views generated by VSRS. Since VSRS is used by MPEG for 3D video compression evaluations and the interpolated views generated by VSRS are considered to be of high enough quality for normal viewing, we only compared our method to MPEG's VSRS package.

After subjective test evaluations were performed, we calculated the mean opinion score with a 95% confidence interval in the same manner as with the other subjective evaluations. As it can be observed from Figure 4.6, the MOS for our technique performing view extrapolation is slightly lower than the MOS for the VSRS package performing view interpolation. This is expected as there is less information available for view extrapolation than for view interpolation. However, the relative closeness of the results show that for applications where bandwidth is limited our hybrid approach is a viable alternative to using VSRS view extrapolation.

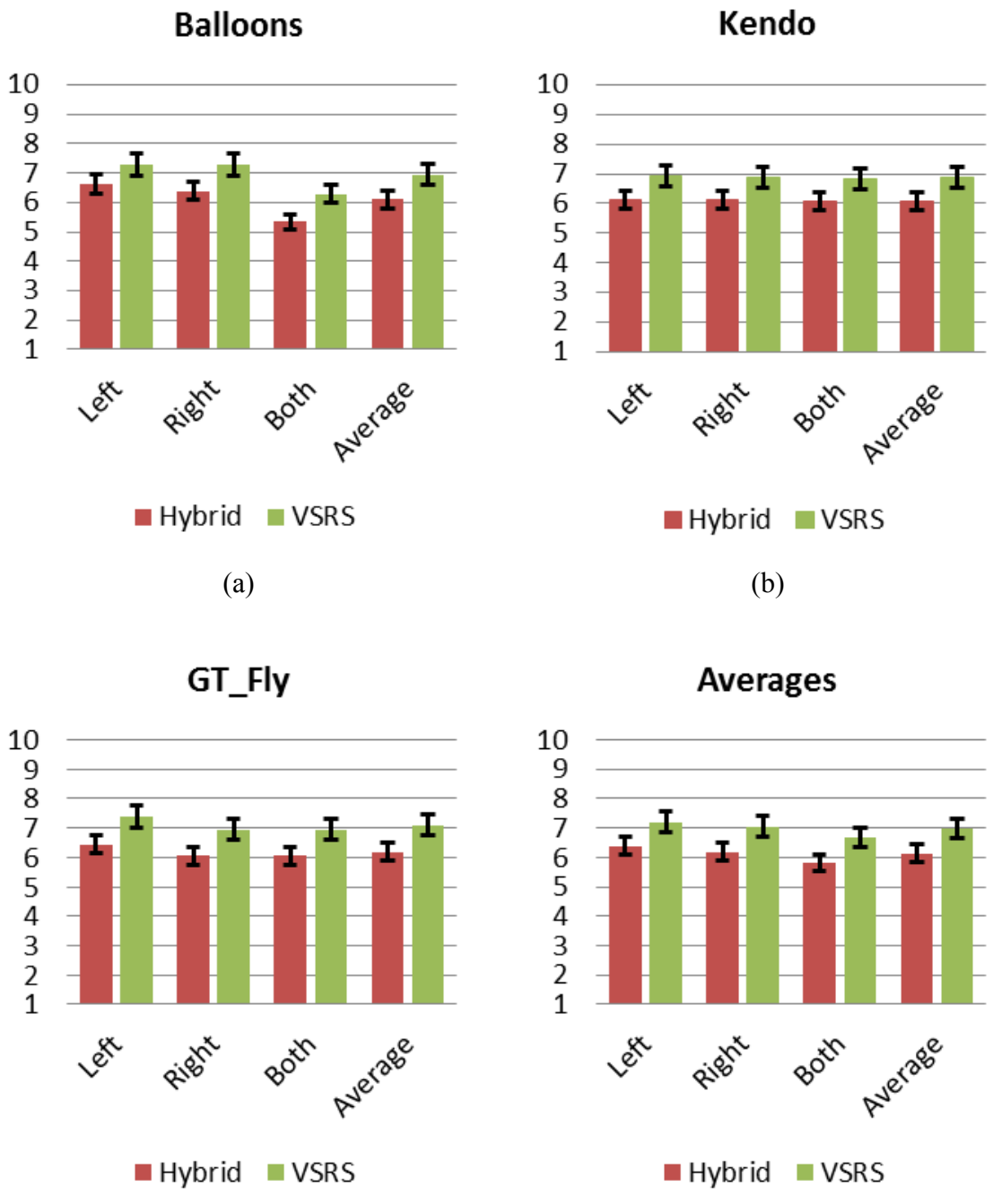


Figure 4.6: MOS for VSRS interpolation versus our Hybrid method extrapolation view synthesis evaluations. The black bar on each graph shows the 95% confidence interval

4.6 Conclusion

In the case of view extrapolation, only one real view and its depthmap are available for the creation of synthesized views. This is a much more challenging problem than view interpolation, where multiple real views are used to generate synthesized virtual views. The reason for this is that in the case of view interpolation, data that is missing in one view due to occlusions, is usually either fully or partially available in the other view(s). This data is then combined in the virtual synthesized views to reduce the number and size of missing data in occluded areas. In the case of view extrapolation, the only additional data come from the one available view, thus the missing data in the occluded areas must be filled in using either the pixel interpolation approach that is used in VSRS or the warping approach that is used in IDWR. Our proposed hybrid approach outperforms both existing methods, generating significantly better quality synthesized views. Subjective evaluations have shown that viewers find the views generated using the proposed hybrid extrapolation approach to be of noticeably higher quality with fewer artifacts.

5 Conclusion

For faster adoption of 3D display technology, there is a need for more 3D video content. This need can be alleviated by converting existing 2D videos to stereo format, through virtual view synthesis. View synthesis is also a key factor in addressing content availability for the emerging multiview display technology expected to reach the market in the next four years. Therefore, there is a need to develop efficient high quality view synthesis techniques for both stereoscopic and multiview applications. To this end, we have proposed a new hybrid view synthesizing approach, which utilizes merits of two existing techniques while overcoming their downfalls. Our proposed method synthesizes new views in a similar fashion to the interpolation-based view synthesizing techniques. However, to fill the holes, it uses an effective warping technique instead of the traditional nearest neighbor interpolation approach. Unlike the Disney's proposed IDWR approach, which warps both background and foreground objects, our approach only warps the background, thus avoiding deformation of the foreground objects. Since most of holes are present in the areas where foreground objects occlude background objects, warping the background areas keeps the more visually important foreground objects intact. Subjective evaluations confirm the superior performance of our method compared to the current interpolation-based state-of-the-art view synthesizing method available in MPEG's VSRS package as well as the new proposed warping method by Disney.

5.1 Future Work

The proposed hybrid view synthesis approach improves on existing methods of virtual view synthesis as can be seen from the subjective evaluations performed. However, further work must be done in optimizing the speed of the proposed method.

Part of this work should involve research in optimized detection of the type of missing texture present in occluded areas. As mentioned in Section 3.3, there is currently no way of distinguishing if the missing data in the occluded areas contains any texture. If no texture exists in these areas, then the nearest neighbor pixel interpolation as used in VSRS would provide results that are just as good as our hybrid approach at a fraction of the time and processing power. As also mentioned in Section 3.3, there are few cases where the background is uniform and contains no changes in texture. This is another area for future work and improvement, that is, the development of an efficient algorithm that can distinguish areas with texture from areas with little or no texture and improve the warping algorithm used in the proposed hybrid method so as to only warp the areas of the background where no texture exists. This approach would increase the overall quality of the synthesized view by further reducing background distortions. Future work could focus on real-time implementation, trying to take advantage of modern GPU processing during the warping step. This would involve offsetting more of the work onto the hardware components available in the GPU and greatly speeding up the view synthesis process and bringing it closer to on the fly virtual view generation.

Bibliography

[1] N. Holliman, “3D Display Systems,” in Press; Handbook of Opto-electronics, IOP Press. 2005.

[2] K. Perlin, S. Paxia and J. S. Kollin, “An autostereoscopic display,” Proc. of 27th Annual Conference on Computer Graphics and Interactive Technology (Siggraph), 2000.

[3] ISO/IEC JTC1/SC29/WG11 MPEG, Document N10357, “Vision on 3D Video,” N10357, 87th MPEG meeting, Geneva, February 2009.

[4] M. T. Pourazad, P. Nasiopoulos and A. Bashashati, “Random Forests-Based 2D-to-3D Video Conversion”, the 17th IEEE International Conference on Electronics, Circuits, and Systems, ICECS 2010, pp. 150-153, December 2010.

[5] ISO/IEC JTC1/SC29/WG11 MPEG Document N11631, ”Report on Experimental Framework for 3D Video Coding,” Guangzhou, China, October 2010.

[6] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross. “Nonlinear disparity mapping for stereoscopic 3D,” ACM SIGGRAPH 2010 papers (SIGGRAPH '10), Hugues Hoppe (Ed.). ACM, New York, NY, USA, Article 75, 10 pages. 2010.

[7] S. L. P. Yasakethu, W. A. C. Fernando, B. Kamolrat, and A. Kondoz, “Analyzing perceptual attributes of 3d video,” IEEE Transactions on Consumer Electronics, vol.55, no.2, pp.864-872, May 2009.

[8] ISO/IEC JTC1/SC29/WG11, MPEG, “View Synthesis Software Manual,” release 3.5, Sept. 2009.

[9] L. Lipton, and J. Halnon, “Universal Electronic Stereoscopic Display”, Stereoscopic Displays and Virtual Reality Systems III, Vol. 2653, pp. 219-223, SPIE, 1996.

[10] T. Okoshi, “Three-Dimensional Imaging Techniques”, Academic Press, New York. ISBN 0-12-525250-1, 1976.

[11] L. Lipton, et. al., U.S. Patent #4,523,226, “Stereoscopic Television System”, June 11, 1985.

[12] K. E. Jachimowicz, et. al., U.S. Patent #4,995,718, “Full Color Three-Dimensional Projection Display”, February 26, 1991.

[13] K. Perlin, S. Paxia, and J. S. Kollin, “An autostereoscopic display”, Proceedings of the 27th annual conference on Computer graphics and interactive techniques (SIGGRAPH '00). ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, 319-326, July 2000.

[14] D. Drascic, J. Grodski, “Defence Teleoperation and Stereoscopic Video”, Proc SPIE Vol. 1915, Stereoscopic Displays and Applications IV, pages 58-69, San Jose, California, February 1993.

[15] J. Eichenlaub, “Lightweight Compact 2D/3D Autostereoscopic LCD Backlight for Games, Monitor, and Notebook Applications”, Proc. SPIE Vol. 3295, p. 180-185, in Stereoscopic Displays and Virtual Reality Systems V, Mark T. Bolas; Scott S. Fisher; John O. Merritt; Eds. April 1998.

[16] C. Fehn, P. Kauff, M. Op de Beeck, F. Ernst, W. Ijsselsteijn, M. Pollefeys, L. Vangool, E. Ofek, and I. Sexton, “An Evolutionary and Optimised Approach on 3D-TV”, IBC 2002, Int. Broadcast Convention, Amsterdam, Netherlands, Sept. 2002.

[17] C. Fehn, “Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV”, Proc. SPIE 5291, Stereoscopic Displays and Virtual Reality Systems XI, 93, May 21, 2004.

[18] K. J. Oh, S. Yea; Y.S. Ho, “Hole filling method using depth based in-painting for view synthesis in free viewpoint television and 3-D video”, Picture Coding Symposium, 2009. PCS 2009 , vol., no., pp.1,4, 6-8 May 2009.

[19] Farre, M.; Wang, O.; Lang, M.; Stefanoski, N.; Hornung, A.; Smolic, A., "Automatic content creation for multiview autostereoscopic displays using image domain warping," Multimedia and Expo (ICME), 2011 IEEE International Conference on , vol., no., pp.1,6, 11-15 July 2011.

[20] ISO/IEC JTC1/SC29/WG11 MPEG, Document N8038, “Committee Draft of ISO/IEC 23002-3 Auxiliary Video Data Representations,” Montreux, Switzerland, April 2006.

[21] F. N. Fritsch and R. E. Carlson, "Monotone Piecewise Cubic Interpolation," SIAM J. Numerical Analysis, Vol. 17, pp.238-246, 1980.

[22] ISO/IEC JTC1/SC29/WG11 MPEG, Document N12036, “Call for proposals on 3D video coding technology,” 96th MPEG meeting, Geneva, March 2011.

[23] ITU-R, “Methodology for the subjective assessment of the quality of television pictures,” ITU-R, Tech. Rep. BT.500-13, 2012.

[24] Q. Hyunh-Thu, P. L. Callet, and M. Barkowsky, "Video quality assessment: from 2D to 3D challenges and future trends," Proc. of 2010 IEEE 17th International Conference on Image Processing, (ICIP), pp.4025-4028, 2010.

[25] H. Asher, “Suppression Theory of Binocular Vision,” Brit. J Ophthalmol, 37(1), pp.37–49, January 1953.

Apendix A – List of Acronyms

2D	Two dimentional
3D	Three dimentional
TV	Television
3DV	Three-dimentional Video
MPEG	Moving Picture Experts Group
VSRS	View synthesis rendering software
DIBR	Depth Images Based Rendering
IDWR	Image-domain-warping-based rendered
CAD	Computer-Aided Design
MOS	Mean opinion score