TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

HAMED SARBOLANDI

SIMULTANEOUS 2D AND 3D VIDEO RENDERING

Master's thesis

Examiners:
Professor Moncef Gabbouj
M.Sc. Payman Aflaki
Professor Lauri Sydanheimo
Examiners and topic approved by the Faculty Council of the Faculty of Computing and Electrical Engineering on 5 June 2013.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Degree Programme in Electrical Engineering
Hamed Sarbolandi: SIMULTANEOUS 2D AND 3D Video RENDERING
Master of Science Thesis, 45 pages
May 2013
Major subject: Wireless communications circuits and systems
Examiners: Professor Moncef Gabbouj, M.Sc. Payman Aflaki, Professor Lauri Sydanheimo
Keywords: Stereoscopic, depth perception, subjective quality assessment, 3DV, 2DV, low pass filtering, contrast adjustment, disparity adjustment

The representation of stereoscopic video on a display is typically enabled either by using active shutter or polarizing viewing glasses in the television sets and displays available for end users. It is likely that in some usage situations some viewers do not wear viewing glasses at all times and hence it would be desirable if the stereoscopic video content could be tuned in the rendering device in such a manner that it could be simultaneously watched with and without viewing glasses with an acceptable quality. In this thesis, a novel video rendering technique is proposed and implemented in the post-processing stage which enables good quality both stereoscopic and traditional 2D video perception of the same content. This has been accomplished by manipulating of one view in the stereoscopic video by making it more similar to the other view in order to reduce the ghosting artifact perceived when the content is watched without viewing glasses while stereoscopic perception is maintained. The proposed technique includes three steps: disparity selection, contrast adjustment, and low-pass-filtering. Through an extensive series of subjective tests, the proposed approach has been evaluated to show that stereoscopic content can be viewed without glasses with an acceptable quality. The proposed methods resulted in a lower bitrate stereoscopic video stream requiring a smaller bandwidth for broadcasting.

PREFACE

This master's thesis was carried out at the Department of Signal Processing, Tampere University of Technology in collaboration with Nokia Research Centre, Tampere. I am very grateful to my head supervisor Prof. Moncef Gabbouj and co-supervisor Payman Aflaki and Distinguished Scientist Miska Hannuksela from Nokia, for much needed support for this work.

The novelty of this thesis was based on patent filed by Miska and Payman. Payman acted as my instant supervisor through all steps of implementation, tests, conducting results, and also reviewing this thesis.

The department of Signal Processing has been an excellent place for me to increase the understanding of multimedia broadcasting. I would like to thank all the staff for their intellectual and financial support.

Tampere, Finalnd.

Hamed Sarbolandi

CONTENTS

1. Introduction 3
2. Background 5
    2.1. Human Visual System 5
    2.2. Ocular information 5
        2.2.1. Stereoscopic information 6
        2.2.2. Dynamic information 6
        2.2.3. Pictorial information 6
        2.2.4. 3D video broadcast system 7
        2.2.5. Content generation 8
        2.2.6. Compression and transmission 8
        2.2.7. Asymmetric stereoscopic video 9
        2.2.8. 3D displays 10
    2.3. Quality of stereoscopic content without viewing glasses 12
    2.4. 3D video compression 13
3. Core idea and proposed technique 15
    3.1. Introduction 15
    3.2. Problem Description 15
    3.3. The Proposed Technique 16
        3.3.1. Disparity adjustment 17
        3.3.2. Contrast adjustment 18
        3.3.3. Subsampling 19
        3.3.4. View blending 19
        3.3.5. Low-pass filtering 21
    3.4. Visual illustration of the proposed technique 21
4. Software implementation 23
    4.1. Broadcasted stereoscopic video file format 23
    4.2. File format 25
    4.3. Color mapping 27
    4.4. Implementation of rendering steps 29
        4.4.1. Subsampling 29
        4.4.2. View Blending 29
        4.4.3. Low-pass filtering 31
    4.5. Optimization 32
5. Subjective test description and results 34
    5.1. Pre-test evaluations 34
        5.1.1. Visual acuity and stereoscopic vision test 34
        5.1.2. Depth perception test 34
    5.2. Test setup 35
    5.3. Preparation of Test Stimuli 36
        5.3.1. Test Procedure and Participants 37
    5.4. Results and discussion 38

6. conclusion and future work 42
References 43

1. INTRODUCTION

In the recent years, the number of three-dimentional (3D) movie titles has increased considerably both at cinemas and as Blu-ray 3D discs. Moreover, broadcast of stereoscopic video content is provided commercially on a few television channels. Hence, many user side devices are already capable of processing stereoscopic 3D content and we will consume an increasing amount of 3D video content in our daily life. Preferences of customers drive the direction of improvements and novelties in different presentation methods of the 3D content and it is therefore important to understand the habits of viewing 3D content and mechanisms of human vision. So, psycho-visual aspects have to be considered when displaying 3D content.

The human vision system (HVS) perceives color images using receptors on the retina of the eye which respond to three broad color bands in the regions of red, green and blue (RGB) in the color spectrum. The HVS is much more sensitive to overall luminance changes than to color changes. The major challenge in understanding and modelling visual perception is that what people see is not simply a translation of retinal stimuli (i.e., the image on the retina). Moreover, the HVS has a limited sensitivity; it does not react to small stimuli, is not able to discriminate between signals with an infinite precision, and also presents saturation effects. In general one could say it achieves a compression process in order to keep visual stimuli for the brain in an interpretable range.

Stereoscopic vision is the principal method by which humans extract 3D information from a scene. Left and right eyes get slightly different views due to their horizontal separation in the head. The HVS is able to fuse these two views in such a way that a 3D perception of the scene in formed in a process called stereopsis. While presenting different views for each eye (stereoscopic presentation), the subjective result is usually binocular rivalry where the two monocular patterns are perceived alternately [1]. In such a case, where dissimilar monocular stimuli are presented to corresponding retinal locations of the two eyes, rather than perceiving stable single stimuli, two stimuli compete for perceptual dominance. In particular cases, one of the two stimuli dominates the field. This effect is known as binocular suppression [2], [3]. It is assumed according to the binocular suppression theory that the HVS fuses the two images with different levels of sharpness such that the perceived quality is close to that of the sharper view [4]. In contrast, if both views show different amounts of blocking artifacts, no considerable binocular suppression is observed and the binocular quality of a stereoscopic sequence is rated close to the mean quality of both views.

In stereoscopic presentation, the brain registers slight perspective differences between left and right views to create a stable, three-dimensional representation incorporating both views. In other words, the visual cortex receives information from each eye and combines this information to form a single stereoscopic image. Left- and right-eye image differences along any one of a wide range of stimulus dimensions are sufficient

to instigate binocular rivalry. These include differences in color, luminance, contrast polarity, form, size, and velocity. Rivalry can be triggered by very simple stimulus differences or by differences between complex images. Stronger, high-contrast stimuli lead to stronger perceptual competition. Rivalry can even occur under dim viewing conditions, when light levels are so low they can only be detected by the rod photoreceptors on the retina.

Binocular suppression has been exploited in asymmetric stereoscopic video coding, for example by providing one of the views with lower spatial resolution [5] or with lower frequency bandwidth [6], fewer color quantization steps [7], or coarser transform-domain quantization [8], [9]. In this paper we exploit binocular suppression and asymmetric quality between views in another domain, namely presentation of stereoscopic 3D content simultaneously on a single display for viewers with and without viewing glasses. Such a viewing situation may occur, for example, when viewing of the television is not active, but the television is just being kept on as a habit. The television may be located in a central place of a home, where many family members are spending their free time. Consequently, there might be viewers actively watching the television with glasses and simultaneous viewers primarily doing something else (without glasses) and just momentarily peeking the television. Furthermore, the price of the glasses, particularly the active ones, might constrain the number of glasses households are willing to buy. Hence, in some occasions, households might not have a sufficient number of glasses for family members and visitors watching the television. While the glasses-based stereoscopic display systems provide a good stereoscopic viewing quality, the perceived quality of the stereo picture or picture sequence viewed without glasses is intolerable.

We tackle this problem by digital signal processing of the decoded stereoscopic video content, making the perceived quality in glasses-based stereoscopic viewing systems acceptable for viewers with and without 3D viewing glasses simultaneously. Viewers with glasses should be able to perceive stereoscopic pictures with acceptable quality and good depth perception, while viewers without glasses should be able to perceive single-view pictures i.e. one of the views of stereoscopic video. The proposed processing is intended to take place at the display and can be adapted for example based on the ratio of users with and without viewing glasses.

2. BACKGROUND

2.1. Human Visual System

One of the main functions of human visual system is to form a 3D representation of the surrounding objects. According to [7],"vision is the process of discovering from images what is present in the world, and where it is". The pictures on our retina are patterns of light intensity, reflected from our environment. In order to acquire a full representation of an object, we have to perceive all three dimensions. Although the external space is projected onto the retina of both eyes as two-dimensional images, the problem is how two-dimensional images from left and right are transformed to a three-dimensional representation? The HVS method uses to reconstruct the three-dimensional object is referred to as stereopsis.

The sources of depth information can be divided in four categories [8] [7]: ocular information (accommodation and convergence), stereoscopic information (binocular disparity), dynamic information (motion parallax) and pictorial information (occlusion, relative size, etc.). Each category will be described briefly in the following sub-sections.

2.2. Ocular information

Convergence and accommodation of the eye are means for depth perception. Convergence is rotation of the eye towards the object. When we look at an object nearby, the eyes converge more than they do when we look at an object far away. The accommodation is the process of focusing on an object by forming the lens (monocular information). Lens muscles are more relaxed when focusing on objects which are far away and contracted when focusing on the object nearby.

Accommodation and convergence do not play a main role in depth perception, but they are important at short distances for specifying the absolute distance of objects which is the perceived distance from observer to objects.

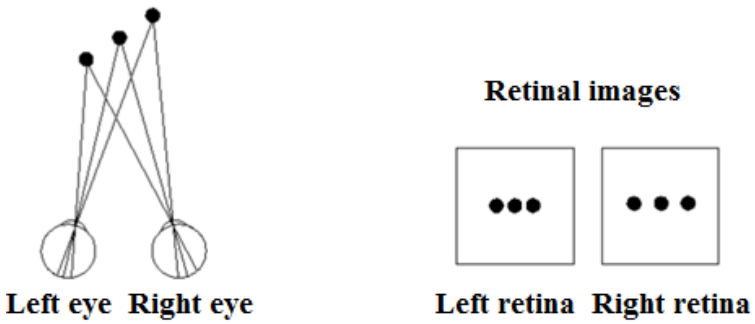


Figure 2.1. *Horizontal separation of the eyes causes an intraocular difference in the relative reflection of monocular images onto the left and right retinas.*

2.2.1. Stereoscopic information

It can be said that the most effective source of depth information is stereopsis. Since our eyes are separated by 6.3 cm on average [8], each eye observes very slightly different perspective of the same scene which is known as retinal disparity (Figure 2.1). The brain mixes these two slightly displaced images and obtains the relative and absolute depths of objects. Moreover, relative depth is perceived distance between objects, when absolute depth is perceived distance from observer to objects. As described in [8], stereopsis is the ability of the brain to perform these calculations.

2.2.2. Dynamic information

Dynamic information occurs with a change over time or changing the position. Depth information about a scene becomes more precise when the viewing point moves with respect to the scene (motion parallax). The motion parallax provides depth information because the image is seen from different distances and the velocity of the image reflected on retina is different for closer and further objects. When an observer is moving with respect to the scene, closer objects seems to be moving faster than the objects in the background.

2.2.3. Pictorial information

A flat picture can provide a good depth information of the contents and since it comes from static and monocular pictures, it is called pictorial information, although a single picture has only two dimensions. In other words, if you close one eye and keep your head still, what you see is three-dimensional and you can still discuss about depth and distance of objects. There are several monocular cues and the most powerful one is occlusion (Figure 2.2.a). Occlusion is the situation where one object is in front of another object and it is partly hidden and it tells the viewer that the hidden object is further away. Relative size refers to the fact that two objects with similar sizes make different size images on retina when they are places in different distances. Figure 2.2.b shows how further objects seem smaller than closer ones. In the visual field, the height cue refers to the fact that objects below the horizon appear closer to the viewer as they are positioned lower. Finally, the shading cue provides information about the shape of an object and it occurs because not all surfaces of an object reflect the same light (Figure 2.2.c). In fact the reflection angle and texture can greatly affect the reflected light.

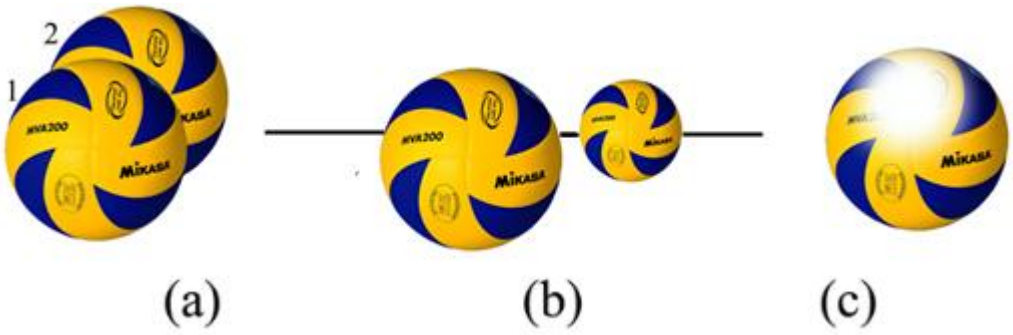
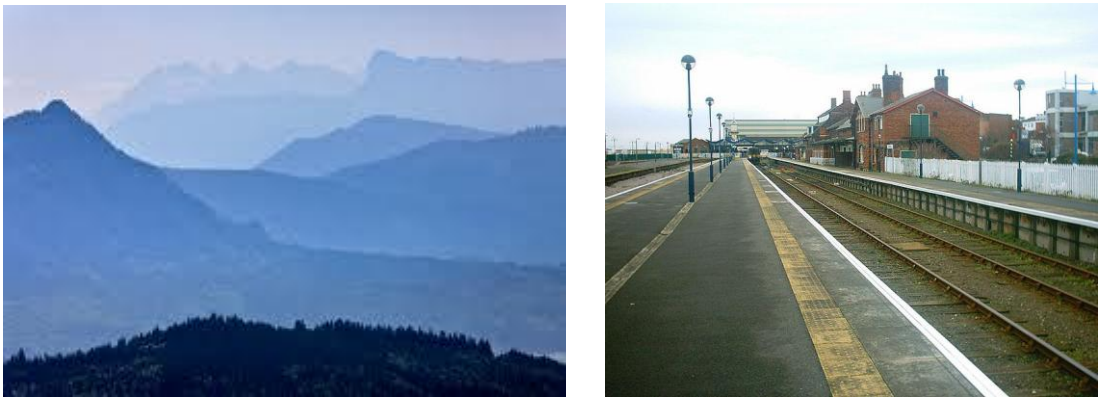


Figure 2.2. *Monocular cues providing depth information of objects in a scene: occlusion (a), relative size, height in the visual field (b), and shading (c).*

The aerial perspective is another cue caused by microscopic particles of dust and moisture in the air that makes because the air contains microscopic particles of dust and moisture that make distant objects look eliminated, less saturated and less sharp. The longer the distance and the more the atmospheric particles, the less contrast (Figure 2.3.a). Prospective is a special case for relative size, where the distance between parallel lines like railroad looks shorter and the tracks appear converged with distance. The more the lines converge, the further away they appear to be (Figure 2.3.b).



(a) (b)

Figure 2.3. *Monocular cues providing depth information of objects in a scene: (a) Aerial perspective and (b)linear perspective*

2.2.4. 3D video broadcast system

Recent technologies in image processing, display design and camera development as well as human 3D perception studies, made the introduction of 3D broadcast system increasingly feasible. A successful implementation, the 3D technology should be compatible with existing conventional broadcast system. Figure 2.4 shows a complete broadcast chain for 3D system from content generation to 3D displays.

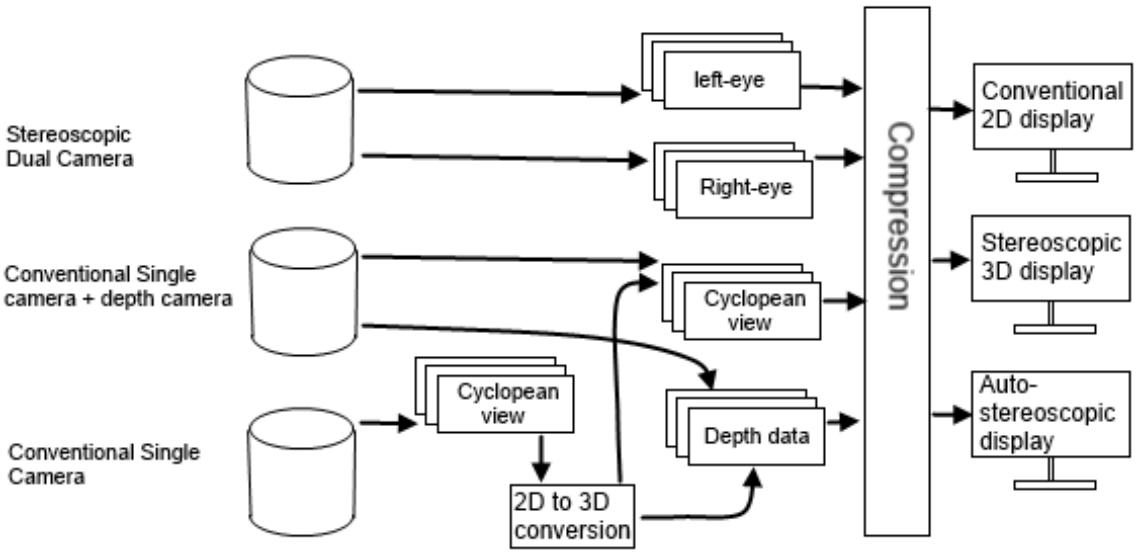


Figure 2.4. *3D-TV broadcast chain including content generation, coding, transmission, and 3D displays.*

2.2.5. Content generation

Based on the current technologies 3D videos are shot using multi-camera for multi-view or a dual-camera configuration for stereoscopic production. In general, two systems can be distinguished: 1) the parallel configuration and 2) the toed-in configuration. The important difference between these two methods is that for a parallel camera configuration, depth is conveyed by crossed disparities in which objects appear closer to the viewer in comparison with the camera fixation point because the zero-disparity point is located at infinity. Hence, binocular disparities for closer objects can be very large and cause visual discomfort. Since for a toed-in configuration the zero-disparity point is at a finite distance, depth is conveyed by both crossed and uncrossed disparities. So, objects appear closer and further away compared to the fixation point. Consequently, the same depth range is divided into crossed and uncrossed disparities for the toed-in configuration resulting in a smaller absolute disparity compared to the parallel configuration [9]. However, converging cameras have a tradeoff between reduced binocular disparities for objects at closer position to the cameras which has less discomfort on the one hand, and visual disparities on the other hand which causes more visual discomfort.

2.2.6. Compression and transmission

When it comes to storage and transmission, the data volume of multiview material involves a large amount of data due to the multiple views of the same scene. Hence, a considerable research work is conducted to reduce the redundancies and realizing the image compression such as Joint Photographic Experts Group (JPEG) or Moving Picture Experts Group (MPEG) coding to obtain saving in storage capacity and therefore smaller bandwidth in transmission.

For High-definition television (HDTV) case, a single uncompressed HDTV channel may cost up to one Gbit/s transmission bandwidth which is far beyond the capacity of low-bandwidth transmission channels such as the internet [10]. For compatibility with the existing broadcast systems the bandwidth should be twice for transmitting the left and right views.

Another approach which overcomes the problem is the use of a depth camera to transmit a single view of RGB along with the depth information per pixel which takes smaller data frame compared to a full RGB frame. Although RGB-Depth transmission is a promising technique, there are some challenges to recover the left and right view perfectly from RGB-Depth video material. Moreover, the desired video data format should be compatible with the conventional codecs H.264/AVC [25] and existing 2D TV sets as well as suited for novel 3D TV applications.

2.2.7. Asymmetric stereoscopic video

Theory of binocular suppression assumes that the binocular percept of a stereo image pair is dominated by the high quality component [11]. Thus, theoretically, when one image of the two stereo pairs is compressed with high bit-rate so that it maintains the high quality, the other view so called "Non-dominant view" in this thesis can be compressed with lower bit-rate without introducing visible artifacts in the binocular percepts. Asymmetric stereoscopic video assumes that the binocular percept is not affected when one view has higher quality and the other view has lower quality and since the quality difference makes the views asymmetric, it is called quality-asymmetric. The mixed resolution concept was introduced by [12], blur low-pass filter was applied as compression algorithm resulting in a high-resolution and low-resolution image for dominant and non-dominant views of a stereo image pair. Binocular combination of asymmetric blur and blockiness impairment images was studied in [13] and the results shows that, the success in asymmetric compression depends on the type of coding artifacts.

One method to reduce the data rate in signal processing is to reduce the sampling rate of a signal, which referred to as downsampling. Downsampling in images is reducing the spatial size of signal by an integer or rational fraction greater than unity. This factor divides the data rate twice when it is applied to both horizontal and vertical axis.

Spatial downsampling is another process to reduce the bitrate and quality in non-dominant view in asymmetric stereoscopic video before compression which is called resolution-asymmetric. The spatial resolution one view is reduced with ratio of e.g.1/2 and accordingly the decoded frames are upsampled to full resolution in a post processing stage. Obviously the quality is reduced in one view, but authors in [14] perform a series of subjective tests to prove that "in most cases, resolution-asymmetric stereo video with the downsampling ratio of 1/2 along both coordinate axes provided similar quality as symmetric and quality-asymmetric full-resolution stereo video." [14]. Figure 2.5 shows how the non-dominant view is downsampled in the transceiver structure before the encoder and upsampled with the same ratio after the decoder.

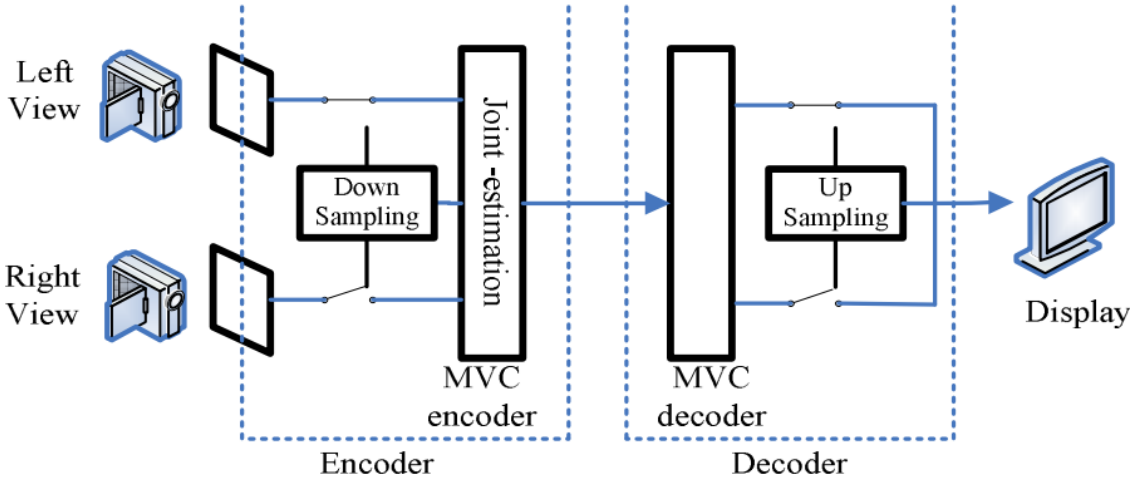


Figure 2.5. *Downsampled asymmetric stereoscopic video both in transmitter and receiver.*

2.2.8. 3D displays

Stereoscopic imaging system in principle is based on displaying two images from a single scene but from slightly different angles of view in a way that left view is seen only by left eye and right view seen only by right view. The capturing cameras resembling human eyes are aligned with the horizon and the difference in corresponding point on display is called the screen parallax. When the parallax on the screen is zero or there is no difference in left and right view, in terms of depth, this point is located at the screen plane.

The "Stereo Window" refers to the physical display surface. Viewer will be able to visualize the concept if you think of your TV screen as a real window that allows us to view the outside world. Objects in your stereoscopic scene can be behind or outside the window which is positive parallax, on the window which is the Screen Plane or zero parallax, or inside, between you and the window which is called negative parallax. Figure 2.6 shows how negative and positive parallaxes can result in objects virtually located in front or behind the display screen.

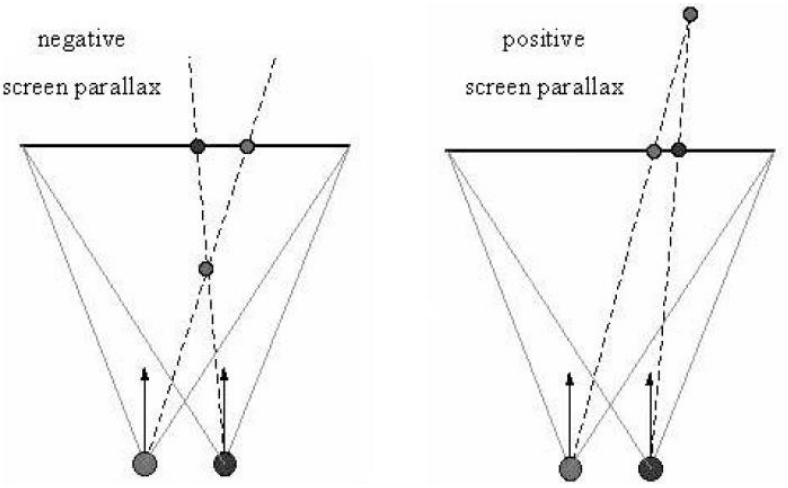


Figure 2.6. *The left image shows a negative parallax in which objects appear in front of the display screen. The right image shows a positive parallax in which objects appear behind the display screen.*

Perfect separation of left and right views is a major challenge for display designer. In general, there are three distinguishing features characterizing stereoscopic displays namely:

1. The separation technique for the left and right eye view.
2. Whether or not motion parallax (multi-view) is supported.
3. The number of observers that can watch 3D simultaneously.

Many techniques can be used to realize left/right eye separation in a stereoscopic display. Usually a distinction is made between stereoscopic and auto-stereoscopic displays

2.2.8.1 Stereoscopic displays

Stereoscopic displays require the viewer to wear polarized glasses or shutter glasses to direct the left and right view. In polarized display technology, left and right views are interleaved in rows of every frame and there are vertical and horizontal optical filters on every other pixel rows. While in shutter glasses the technology is in glasses and it can be set to almost any display. Shutter glasses have two states of transparent or shut, and the controller switches the states to let only one view at a time. As Figure 2.7 shows, the display and glasses must be perfectly synchronized. Unlike polarized glasses which don't have any connection to the display. Hence, Polarized glasses are also called passive glasses and accordingly shutter glasses are called active glasses.

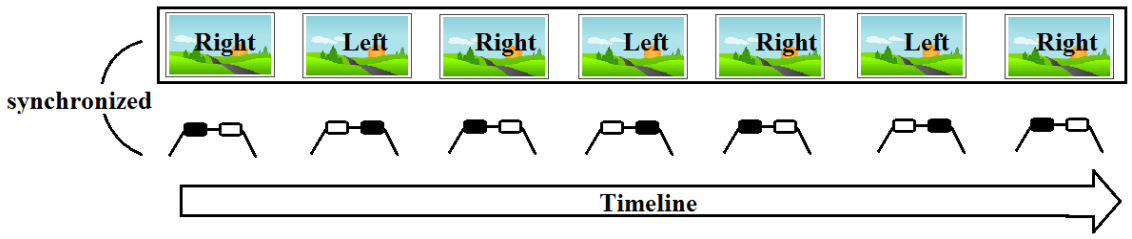


Figure 2.7 *3D display with active shutter glasses.*

2.2.8.2 Autostereoscopic displays

Specifically, Dual-view autostereoscopic displays have an especial screens that can beam two different images from different perspectives. Provided that each observer is correctly positioned, this allows several but limited number of viewers to use this type of displays (Figure 2.8b).

Figure 2.8a shows one technology that can beam two views from different perspectives. In this method a device is placed in front of an image source to allow it to show a stereoscopic image. This device is called parallax barrier.

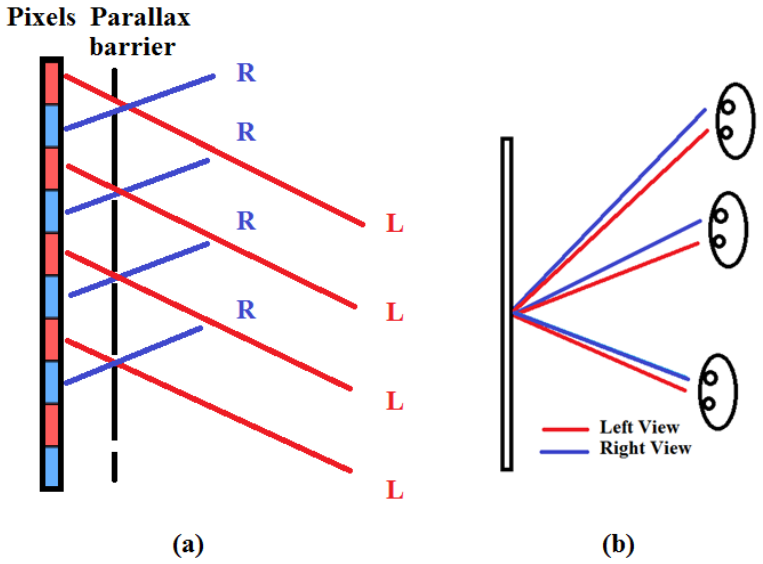


Figure 2.8. *Dual-view (a) Parallax barrier. (b)Multi-viewers principle.*

2.3. Quality of stereoscopic content without viewing glasses

In 3D video quality, we face the problem of binocular suppression [15]. This phenomenon is due to artifacts that cause contradictory depth cues to be sent to each eye. Similarly to asymmetric video encoding which results in the masking of the artifacts of the lower quality view, the risk is to suppress the stereopsis because there is no combination of both values.

Even though it has been shown that image quality is important for visual comfort, it is not the only factor for great 3D visual experience. New concepts, as widely studied in [16], have to be considered such as presence i.e. the feeling of being there and depth perception, investigated in [17], [18]. Stereoscopic vision is based on stereopsis and depth perception relies on the fusion of two slightly different viewpoints of the same scene and also on monocular cues. As described in [19], depth perception increases by increasing the disparity between left and right view. However, if the disparity of views is more than a threshold, it will cause some eye strain in the subjects. Hence, an ac-

ceptable depth perception of 3D video depends on correct selection of distance between left and right view.

One annoying artifact while observing 3D content with glasses is the ghosting effect also known as crosstalk [20]. It is perceived as ghost, shadow, or double contours due to imperfect optical separation between the left and right images by filters of each eye in passive glasses or slight imperfection in synchronization between shutters in active glasses and displayed left and right views. Crosstalk is suspected to be the main contributor to the visual discomfort and disturbing image quality for 3D viewers. This ghosting effect is mostly visible when watching a stereoscopic video on a 3D display without glasses (2D presentation), since both left and right views are visible to both eyes. Hence, the subjective quality of stereoscopic video in 2D presentation is not acceptable due to ghosting effect as depicted in Figure 2.8.

Figure 2.9. *2D perception (without glasses) of stereoscopic video.*

2.4. 3D video compression

A crude solution for coding multiview video is to encode each view separately using a standard video codec such as H.264/AVC [25]. The advantage of this approach is that it can be achieved using the existing standards and current hardware. However, it does not exploit the redundancy across views and the bit-stream would be twice the corresponding 2D video. This is potential to cause problems in existing storage and broadcast systems.

The basic idea used in all multiview compression methods is to reduce inter-view redundancies which come from the fact that all cameras are capturing the same scene. Figure 2.8 shows a sample prediction structure in which pictures are not only predicted from temporal neighbors, but also from spatial neighbors from adjacent views.

The subjective testing has indicated that the same quality could be achieved with approximately half the bit-rate for a number of test sequences. Except the coding efficiency, several other aspects of MVC standard are listed [21]:

- Scalabilities: View scalability and temporal scalability are considered in the MVC design for network bandwidth, user preferences and decoder complexity.
- Parallel Processing: Since multiple views need to be encoded simultaneously to be displayed in real time, parallel processing of different views is required.
- Random Access: Besides temporal random access, view random access is to be supported to enable accessing a frame in a given view.
- Robustness: When transmitted in a lossy channel, the MVC bit-stream will have error resiliency capabilities.

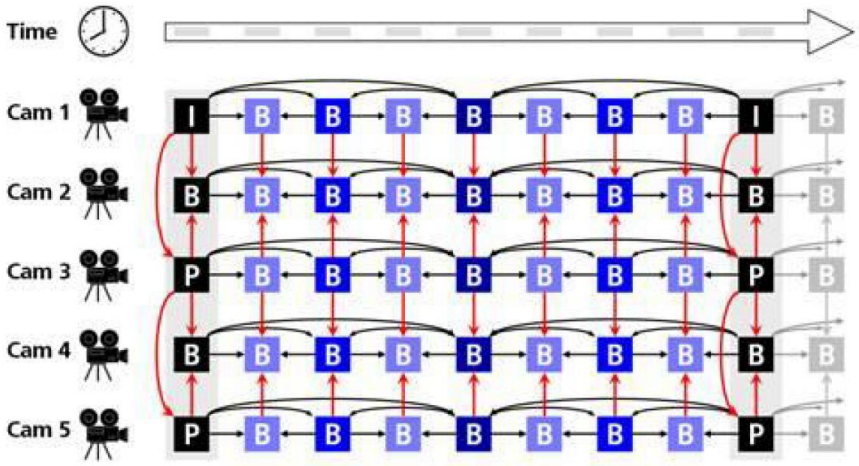


Figure 2.10. *Illustration of interview prediction in H.264/AVC.*

As it can be seen from Figure 2.10, the interview frames are predicted based on previous and later reference frames.

3. CORE IDEA AND PROPOSED TECHNIQUE

3.1. Introduction

A variety of display devices providing a three-dimensional (3D) experience have been commercialized. Among the 3D display solutions are multi-view auto-stereoscopic displays, where the views seen depend on the position of the viewer relative to the display, and stereoscopic displays requiring the use of polarizing or shutter glasses. It seems that the display solutions based on glasses are more mature for mass markets and many such products are entering the market currently or soon.

The lenses of polarizing glasses used for stereoscopic viewing have orthogonal polarity with respect to each other. The polarization of the emitted light corresponding to pixels in the display is interleaved. For example, odd pixel rows might be of a particular polarity, while even pixel rows are then of orthogonal polarity. Thus, each eye sees different pixels and hence perceives different pictures.

The shutter glasses are based on active synchronized alternate-frame sequencing. There is a synchronization signal emitted by the display and received by the glasses. The synchronization signal controls which eye gets to see the picture on the display and for which eye the active lens blocks the eye sight. The left and right view pictures are alternated in such a rapid pace that the human visual system perceives the stimulus as a continuous stereoscopic picture.

3.2. Problem Description

While the glasses-based stereoscopic display systems provide a good stereoscopic viewing quality, the perceived quality of the stereo picture or picture sequence viewed without glasses is intolerable. Figure 3.1 presents stereoscopic view perceived without glasses. An annoying shadow or ghost image can be observed.

Figure 3.1 *Original stereo pair viewed without glasses.*

However, there might be situations where there are viewers with and without glasses. For example, in many cases viewing of the television is not active, but the television is just being kept on as a habit. The television may be located in a central place of a home, where many family members are spending their free time. Consequently, there might be viewers actively watching the television with glasses and simultaneous viewers primarily doing something else (without glasses) and just momentarily peeking the television. Furthermore, the price of the glasses, particularly the active ones, might constrain the number of glasses households are willing to buy. Hence, in some occasions, households might not have a sufficient number of glasses for family members and visitors watching the television.

This thesis tackles the problem aims at making the perceived quality in glasses-based stereoscopic viewing systems acceptable for viewers with and without glasses simultaneously. Viewers with glasses should be able to perceive stereoscopic pictures, while viewers without glasses should be able to perceive single-view pictures.

3.3. The Proposed Technique

The human vision system perceives color images using receptors on the retina of the eye which respond to three broad color bands in the regions of red, green and blue (RGB) in the color spectrum. The HVS is much more sensitive to overall luminance changes than to color changes. The major challenge in understanding and modeling visual perception is that what people see is not simply a translation of retinal stimuli (i.e., the image on the retina). Moreover, the HVS has a limited sensitivity; it does not react to small stimuli, is not able to discriminate between signals with an infinite precision, and also presents saturation effects. In general one could say it achieves a compression process in order to keep visual stimuli for the brain in an interpretable range.

While presenting different views for each eye (stereoscopic presentation), the subjective result is usually binocular rivalry where the two monocular patterns are per-

ceived alternately [22]. In particular cases, one of the two stimuli dominates the field. This effect is known as binocular suppression. It is based on the binocular suppression theory that the HVS mixes the stereo images in a way that the perceived image has quality close to that of the higher quality view.

Binocular rivalry affords a unique opportunity to discover aspects of perceptual processing that transpire outside of visual awareness. In stereoscopic presentation, the brain registers slight perspective differences between left and right views to create a stable, three-dimensional representation incorporating both views. In other words the visual cortex receives information from each eye and combines this information to form a single stereoscopic image. Left- and right-eye image differences along any one of a wide range of stimulus dimensions are sufficient to instigate binocular rivalry. These include differences in color, luminance, contrast polarity, form, size, or velocity. Rivalry can be triggered by very simple stimulus differences or by differences between complex images. Stronger, high-contrast stimuli lead to stronger perceptual competition. Rivalry can even occur under dim viewing conditions, when light levels are so low they can only be detected by the retina's rod photoreceptors. Under some conditions, rivalry can be triggered by physically identical stimuli that differ in appearance owing to simultaneous luminance or color contrast.

The technique implemented in this thesis benefits from several rendering steps applied to the stereoscopic content. These steps are presents in the following sub-sections.

3.3.1. Disparity adjustment

Difference in physical positioning of human eyes makes slightly different views perceived by left and right eyes. This difference in views gives depth perception in HVS and accordingly in 3D capturing the left and right views are capture by physically separated cameras. The distance between cameras is called disparity.

The impact of camera disparity was studied in [16], in which three camera separation distances of 0, 8, and 12 cm were utilized. Results shows that the depth perception of stereo images increased by increasing the camera separation. The distance between cameras creates a pixel disparity on display for objects in the scene. Disparity of pixels can be converted to disparity distance in centimetres as shown in (3.1) and (3.2).

$$w = \mathrm{W}_{\mathrm{cm}} / \mathrm{W}_{\mathrm{pixels}} \tag{3.1}$$

Where:
W_{cm} is display width in cm.
W_{pixels} is display width in pixels.
w is pixel width in cm.

$$DD = w \times PD \tag{3.2}$$

Where:

DD is distance disparity.

PD is pixel disparity.

Assuming the viewing distance, which is the distance between the viewer and display to be (VD), the disparity in arcmin can be calculated for different objects in the scene using (3.3).

$$D_{Arcmin} = 2 \times \text{atan}\left(\frac{DD}{2 \times VD}\right) \tag{3.3}$$

Where DD and VD parameters are illustrated in Figure 3.2.

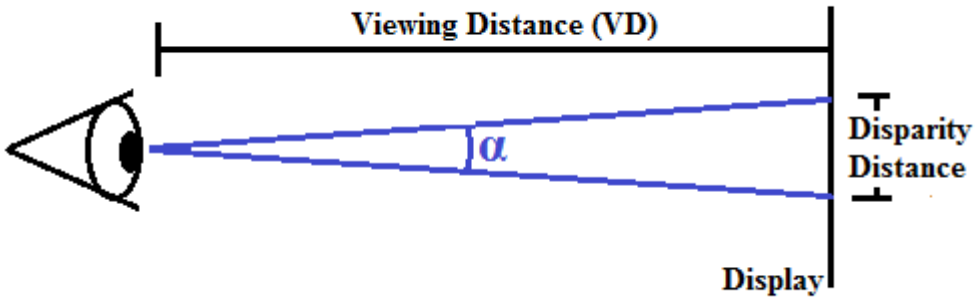


Figure 3.2 *Disparity calculation in Arcmin based on different pixel disparities on display.*

3.3.2. Contrast adjustment

Contrast is basically the dissimilarities in visual properties of objects that make it distinguishable from other objects and background. In the visual perception of scenes from different views, contrast is determined by the difference between color and brightness of each object and other objects in the same viewing field. Hence, contrast adjustment method is related to brightness and color settings e.g. the differences and changes in luminance and chrominance. On the other hand, human eye is more sensitive to the views and scenes with more contrast, or they are more interesting and have more sharpness for brain to process, rather than the views which have fewer details.

The concept described so far, is the idea of the method utilized in this experiment which is actually to decrease the contrast of the non-dominant view while keeping the contrast of dominant view unchanged. The contrast decrease of non-dominant view will help a 2D presentation of stereoscopic view that has more similarity to dominant view while stereoscopic presentation is not influenced considerably.

The contrast adjustment of an image can be done in various ways. We utilized the same formula as used for H.264/AVC weighted prediction which is presented in equation (3.1):

$$O = round\left(\frac{i \times w}{2^d}\right) = (i \times w + 2^{d-1}) \gg d \tag{3.1}$$

where:
O is the adjusted contrast value
$round$ is a function returning the closest integer
i is the input sample value
w and d are the parameters utilized to create the adjustment weight

3.3.3. Subsampling

Subsampling or half toning is a method applied to the non-dominant view in which some of the pixel positions in the non-dominant view became unused, i.e., are set to zero luma level. An additional step can be performed to adjust the non-dominant view by filling the unused pixel positions smoothly using some information from dominant view.
In this approach non-dominant view is read row by row. Along each even row, the odd pixel values will be replaced by their average value with the same pixel value in the dominant view as presented in Figure 3.3 For odd rows replacement will be applied to even pixels.

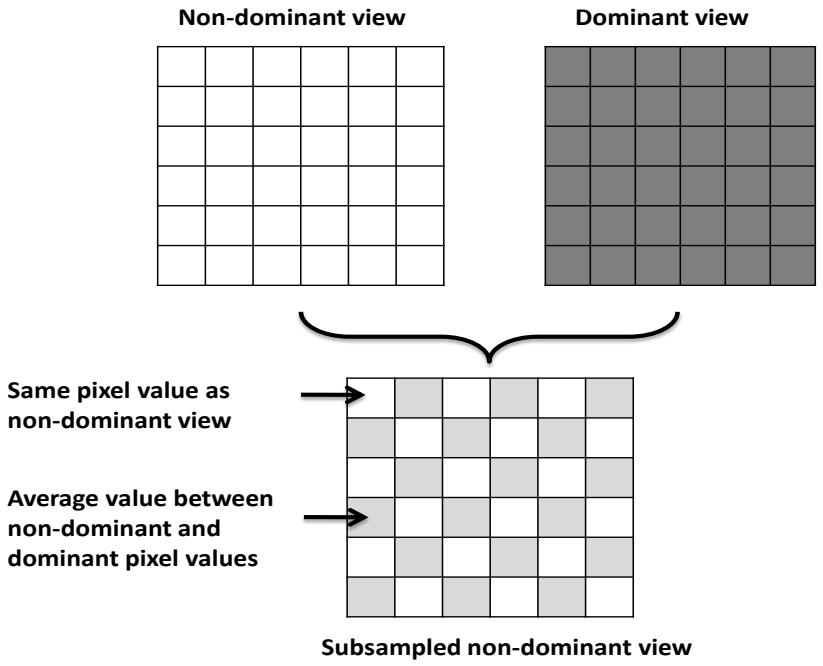


Figure 3.3 *View blending combined to sub-sampling of the non-dominant view.*

3.3.4. View blending

This approach tries to make the non-dominant view more similar to the dominant view based on a specified threshold. We denote the original dominant, original non-dominant, to be created dominant, and to be created non-dominant views as OD, OND, CD, and CND, respectively. Moreover, ω is a weighting parameter ($0< \omega <1$). By changing ω in its range, we have the possibility of adjusting the similarity extent of non-dominant

view to dominant view. Each view is scanned in blocks of 2x2 pixels. The following Error equation will be applied to each block:

$$Error = \omega * \left|\frac{CND + CD}{2} - OD\right| + (1-\omega) * \left|\frac{CND - OND}{2}\right| + (1-\omega) * \left|\frac{CD - OD}{2}\right|$$

Where OD, OND, CD, and CND are the average luma value of the respective 2x2 blocks. The term ω*abs((CND+ CD)/2–OD) represents the error observed in viewing without glasses, whereas the terms (1-ω)*abs(CND- OND)/2 + (1- ω)*abs(CD- OD)/2 jointly represent the error observed in viewing with shutter glasses. We apply a minimization algorithm on Error equation by changing the values of CND and CD in the whole range of possible values. Figure 3.4 is an illustration of view-blending method.

By solving the minimization problem for a 2x2 block, the average luma value for a 2x2 block in the output images is obtained. The ratio between OND and CND (for a 2x2 block) is then used to multiply the each luma pixel value in OND and the result is typically quantized to an integer value in the range of 0 to 255, inclusive. The potential quantization error may be randomly distributed onto the pixel values of the converted block such a way that the average luma value of the converted block becomes equal to CND.

A variety of non-dominant view presentations having different levels of similarity to dominant view can be generated with this method. By means of the parameter ω we are free to bias our final created views to satisfy more either single-view viewing without glasses (w=1) or stereoscopic viewing with glasses (w=0).

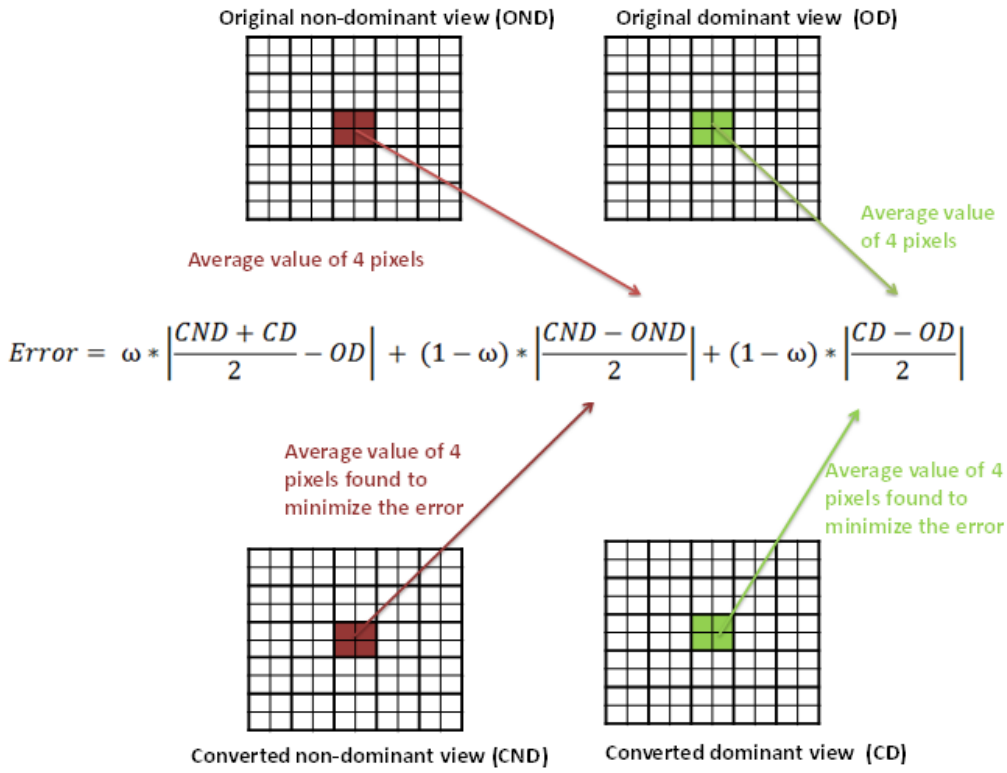


Figure 3.4. *View blending method.*

3.3.5. Low-pass filtering

This method decreases the number of high frequency components (HFCs) from non-dominant view by removing some detail. Hence, in the created asymmetric stereoscopic video, the non-dominant view will be somehow blurred compared to the dominant view. This will favor to better 2D presentation of the stereo pair while the dominant view will be sharper compared to the blurred non-dominant view and therefore it will be more perceived by HVS. Yet, as verified extensively in previous studies [23], [24] asymmetric stereoscopic video where one view has been low pass filtered provides similar subjective quality and depth perception to those of stereoscopic video where both views have the same high quality.

In our experiments, the applied LPF was a 2D circular averaging filter (pillbox) within the square matrix of side 2×radius+1, as it showed better subjective performance compared to a few other tested LPFs. In general, any LPF could be used for example on the basis of memory access and complexity constraints. The level of HFC reduction depends on the radius defined for the filter such that increasing the radius results in more reduction of HFCs. Complete 2D matrix presenting the coefficients of utilized LPF for radius equal to 6 is depicted in (3.2). This is a good approach since it could benefit coding performance by reducing the necessary bitrate for encoding the views while one view has less HFCs and hence, less details should be encoded.

$$f = 10^{-4} \times \begin{bmatrix} 0 & 0 & 0 & 0 & 13 & 36 & 44 & 36 & 13 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 61 & 88 & 88 & 88 & 88 & 88 & 61 & 8 & 0 & 0 \\ 0 & 8 & 76 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 76 & 8 & 0 \\ 0 & 61 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 61 & 0 \\ 13 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 13 \\ 36 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 36 \\ 44 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 44 \\ 36 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 36 \\ 13 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 13 \\ 0 & 61 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 61 & 0 \\ 0 & 8 & 76 & 88 & 88 & 88 & 88 & 88 & 88 & 88 & 76 & 8 & 0 \\ 0 & 0 & 8 & 61 & 88 & 88 & 88 & 88 & 88 & 61 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 13 & 36 & 44 & 36 & 13 & 0 & 0 & 0 & 0 \end{bmatrix} \quad (3.2)$$

3.4. Visual illustration of the proposed technique

It is impossible to present examples of the exact results of the thesis, because those can only be perceived on a stereoscopic display based on polarization or shutter glasses. In this section, we anyway present an example image produced by averaging the images of the left and right view, which resembles the image perceived when viewing an image from a stereoscopic display intended for shutter glasses but when no glasses are worn. We note that the perception on a stereoscopic display is different – particularly, the colors appearing in the example image in this document are more washed-out than what can be perceived on a stereoscopic display.

Figure 3.5 includes an example of an adjusted stereoscopic picture viewed without glasses as comparison to Figure 3.1 which is the original stereoscopic picture viewed without glasses. While the shadow image has become tolerable in single-view viewing without glasses, the human binocular vision still perceives three-dimensional pictures.

Figure 3.5. *An illustration of an adjusted stereo pair viewed without glasses.*

4. SOFTWARE IMPLEMENTATION

All the methods described in 3.3.1 to 3.3.5 were initially tested in Matlab and the output was confirmed. Although, the execution of the algorithms was too slow to be played in real-time, it could confirm the correctness of algorithms one by one.

In order to be able to observe the output of all methods simultaneously and in real-time, a windows application software was written to implement the whole process chain starting from raw sequence file handling to representation on the screen. This chapter describes different sections of the software.

4.1. Broadcasted stereoscopic video file format

Stereoscopic video formats are part of multiview video format but it is limited to only 2 views for stereoscopic viewing. Multiview video will be used in next generation of autostereoscopic multiview displays and these displays will provide stereoscopic perception from any arbitrary angle in front of the display. Hence, there must be several views available in the user side. Multiview Video Coding (MVC) [25] is an extension of the AVC [26] standard that provides efficient coding of multiview video. The overall structure of MVC is fed with N temporally synchronized raw video streams and after compression. The encoder receives the video streams and generates a single bit-stream. The decoder receives the bit-stream, decodes and outputs the N video signals. A high level block diagram of the MVC system is shown in Figure 4.1:

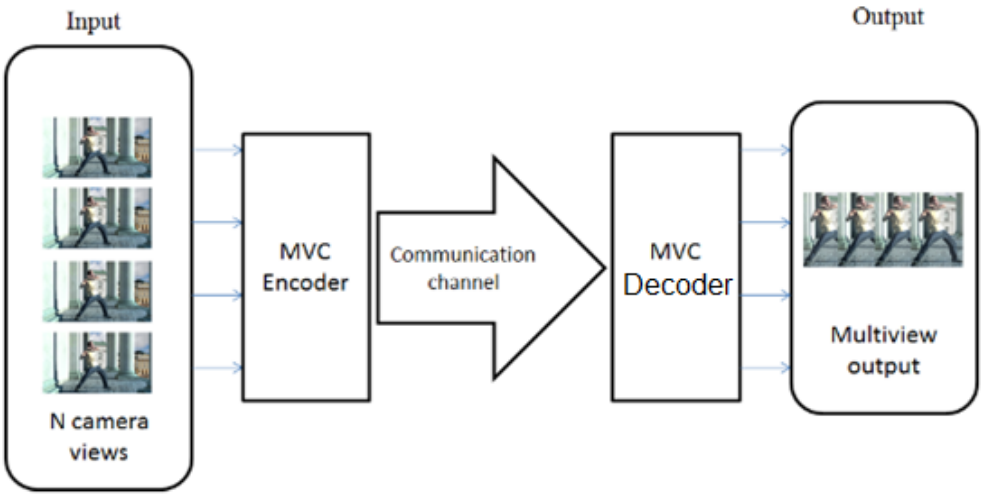


Figure. 4.1. *Multiview Video Coding.*

MVC contains a large amount of inter-view dependencies due to similar contents. Therefore, exploiting such correlation between views enables the codec to compress the content more efficiently and hence, the average bitrate per view will be considerably less that the bitrate requires to encode one view.

In current standardization, there exists one base view and few dependent views. Inter-view prediction is utilized to encode the dependent views more efficiently while the base view should be encoded in simulcast mode. Independent from multiview issues, one phase in coding is always compression phase which is highly dependent on quantization factor and image complexity. Encoding performance highly depends on the content of the sequence and amount of high frequency components presented in the scene. Therefore, Low-pass filtering as introduced in sub-section 3.4 decreases the required bitrate to encode the same content by removing the high frequency components. As a result, the same content can be transmitted occupying lower bandwidth. The major modification which is applied in section 3.4 is low-pass filtering. In low-pass filtering the complexity of one view is decreased and high frequency components in Fourier domain of the image are filtered. Therefore, in the Non-dominant view there can be higher compression factor without considerable amount of quality loss. Figure 4.2 compares the original view and the view filtered with LPF which decreases the number of high frequency components (HFCs) the applied LPF is a 2D circular averaging filter (pillbox) within the square matrix of side 2×radius+1 (Figure 4.2), The compression ratio is much higher in the filtered image. Applying such a filter to the whole sequence, decreases the bitrate considerably and consequently the stereoscopic video broadcasting with less bandwidth is feasible. Hence, there will be a considerable amount of saving in transmission power.

(a) (b)

Figure.4.2. *(a) Original and (b) low-pass filtered view.*

4.2. File format

The file format used as the input of the software is YUV 4:2:0 format. Basically, YUV is a raw data video format which has no compression no encoding and there is only a collection of raw pixel values in YUV color space. YUV color model defines a color space in terms of one luma (Y) and who chrominance (UV) components. Initially the story behind separating luminance and chrominance components comes from the analogue televisions. The time when there were a need to have a video signal transmission method to be compatible with both color-television and black-and-white infrastructure. The luma component was already available in the broadcasting technology and they added UV chroma components as a solution to keep the technology compatible with both receivers. Another advantage making YUV more useful in image processing experiments nowadays is that, the human eye is less sensitive to changes in *hue* compared to changes in brightness. Consequently, each image can be presented with less amount of information for Chroma components compared to the information for Luma component without sacrificing the visual quality (Figure 4.3).

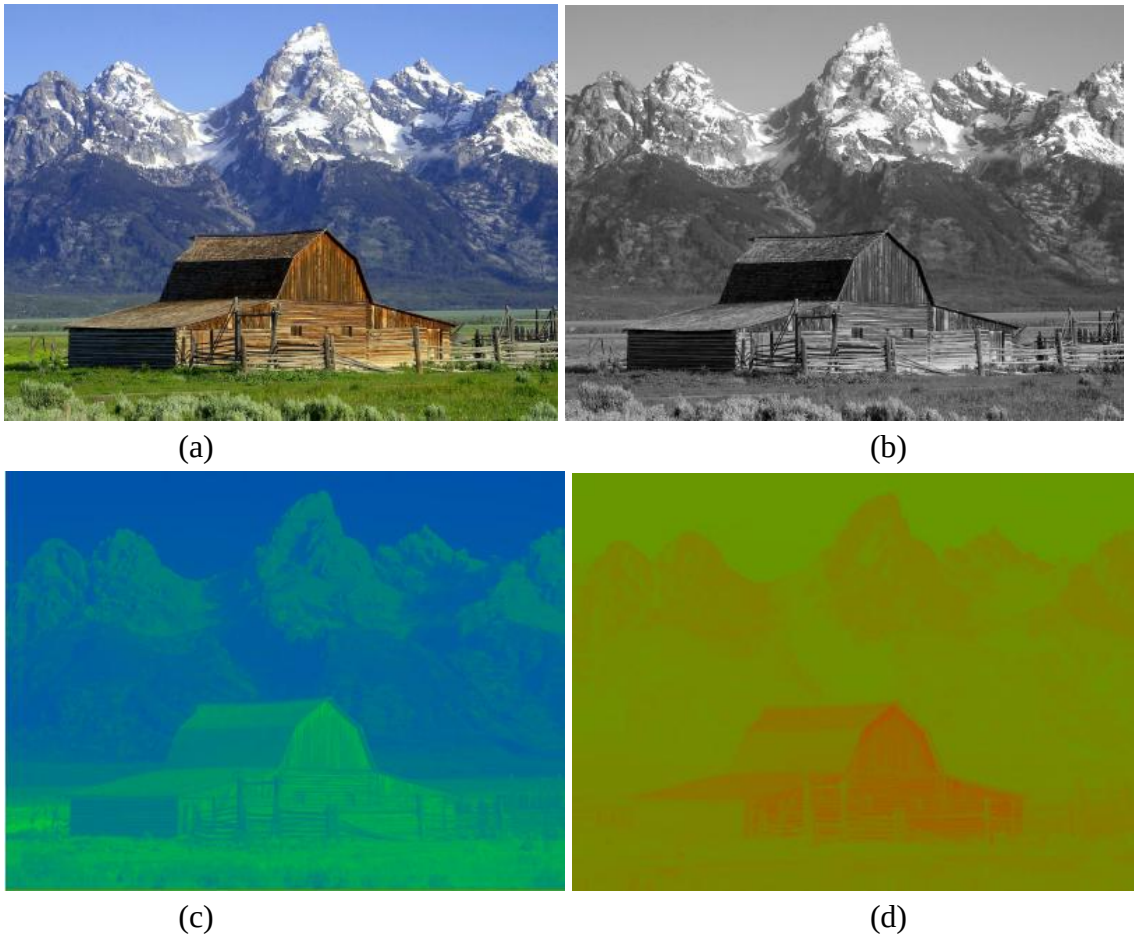


Figure 4.3 *(a) Original Image (b) Luma Y component (c) Chroma U component (d) Chroma V component*

YUV sequence format is interesting in digital video broadcasting (DVB) and we are also using this format in order to have our benchmark compatible with other tools and systems in this field. As long as YUV sequence has no header section in the file content to represent details of the sequence parameters like sampling rate, image size, number of frames etc. There has to be a standard or previously defined format when the file is ready to be played. In other words the file does not contain playing parameters, so, players cannot extract any information automatically from the file and they should manually be checked before passing to the program. Frame size determines the dimensions of every frame of the sequence and the following resolutions are widely used in this test.

There are special sampling system and ratio scheme in YUV format, which is commonly expressed as a three part ratio (e.g. 4:2:2). That describes the number of luminance and chrominance samples in a conceptual region that is J pixels wide and 2 pixels high which is described in Figure 4.3.

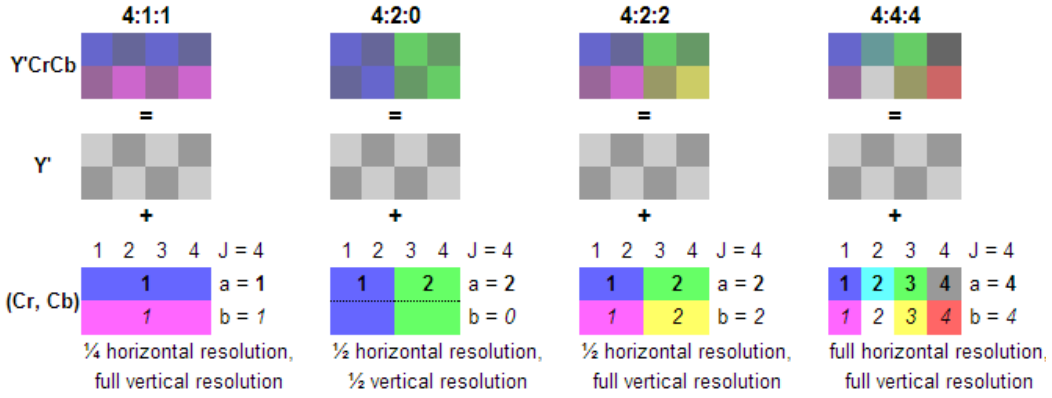


Figure 4.3 *YUV pixels formats.*

YUV 4:2:0 is mostly used in this test and all raw video sequences are based on 4:2:0 sampling format. We can simply interpret 4:2:0 sampling system so that for every standard frame size Luma frame size, there is a one-fourth standard frame for U Chroma and again a one-fourth standard frame for V Chroma as shown in Figure 4.4.

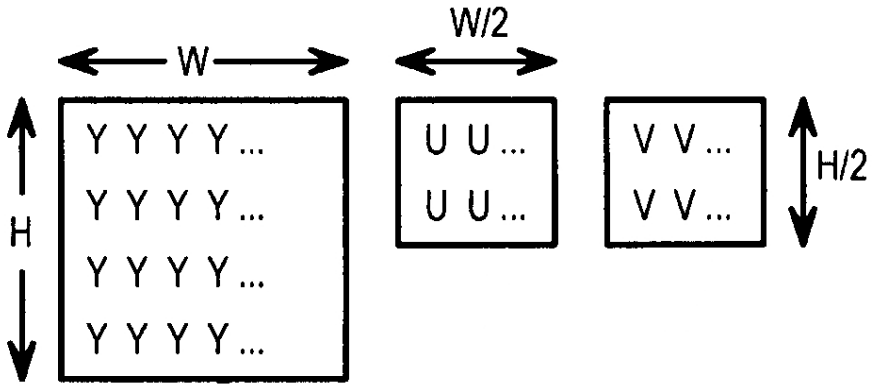


Figure 4.4 *Frame size and subsampling system in 4:2:0.*

The data volume that every frame needs in a sequence can be obtained from the equation (4.2.0).

$$Frame\ size = (H * W) + \left(\frac{H}{2} * \frac{W}{2}\right) + \left(\frac{H}{2} * \frac{W}{2}\right) = 1.5\ (H * W) \qquad (4.2.0)$$

Where:
H : is the height of a frame.
W : is the width of a frame.

Equation (4.2.0) shows how many bits are needed to save one frame in YUV 4:2:0 with respect to the subsampling in Chroma components. Frame size standards in pixels are presented in Table 4.1. As an example a frame in full HD size (1920x1080) contains 1920*1080 pixels for Luma and 960*540 pixel for each Chroma component. Then as written in equation (4.2.1) the number of pixels is multiplied by color depth to get the number of bits every frame need to be stored on memory.

$$Number\ of\ bits = \big((1920 * 1080) + (960 * 540) + (960 * 540)\big) * 2^8$$
$$= 796262400\ bits \qquad (4.2.1)$$

For instance in RGB, every frame takes 3 times of a frame size, while it has been half in YUV 4:2:0 consequently the whole file of sequences takes half space on the memory.

Table 4.1. *Flag table presenting significant differences for different test schemes.*

Standards	Frame sized in pixesl
VGA	640 x 480
HD 720	1280 x 720
Full HD	1920 x 1080

4.3. Color mapping

A large number of multimedia applications have encountered the RGB color space. A color space is in fact an association between a set of values in that color space and a color.
The RGB color space represents colors in terms of red, blue and green. The combination of these intensities by the light beams inside a display can form a wide range of color spectrum. Basically, commercial displays favor RGB color space and that is the main reason to map from YUV to RGB. Color mapping from RGB to YUV is presented in Figure 4.5.

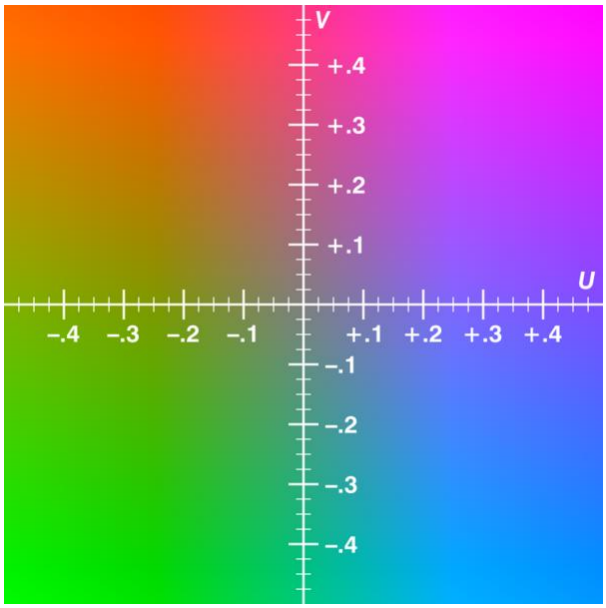


Figure 4.5. *RGB representation of U and V color space.*

There exist plenty of slightly different formulas to convert color space between YUV and RGB and the only difference is the number of decimal places. The ITU-R 601 standard [27] specifies the correct coefficients.
There is a tradeoff between precision and calculation complexity. These formulas assume Y, U, and V values are unsigned integers from 0 to 255 and presented with 8 bits. Equation (4.3.1) describes the conversion from YUV to RGB.

$$\begin{bmatrix} Y \\ U \\ V \end{bmatrix} = \begin{bmatrix} 0.257 & 0.504 & 0.098 \\ -0.148 & -0.291 & 0.439 \\ 0.439 & -0.368 & -0.071 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \tag{4.3.1}$$

Accordingly, the conversion from YUV to RGB can be obtained from the equation (4.3.2).

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix} = \begin{bmatrix} 1.164 & 0 & 1.596 \\ 1.164 & -0.391 & -0.813 \\ 1.164 & 2.018 & 0 \end{bmatrix} \left(\begin{bmatrix} Y \\ U \\ V \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right) \tag{4.3.2}$$

Both equations (4.3.1) and (4.3.2) are utilized in the first stages of software development. However, to achieve real time playback of HD content, the algorithm was optimized (as introduced in sub-section 4.6) and achieved significantly better performance without any inaccuracy in the mapping process.

4.4. Implementation of rendering steps

There are several steps to adjust the non-dominant view according to our needs. Each method is implemented in the software as a function which can be called in real time. This allows user to modify the content during the playback and select the desired combination of different methods to achieve the most comfortable viewing experience both with and without glasses. Enabling such real time modifications, requires the whole program to be relatively efficient in order to maintain the smoothness of the stream and therefore playing the final sequence in at least 25 FPS for HD (1080p) or 30 FPS for HD (720p) frame size.

4.4.1. Subsampling

A fraction of the color in in dominant view is transported to non-dominant view depending on the sampling weight. So, the main for-loop, goes through both dimensions of the image and selects every other pixel in both vertical and horizontal axis, then the selected pixel in dominant view is multiplied by weight and the same pixel in non-dominant view is multiplied by (1-w) to have the counter effect on the output value:

```
for(int ii=0;ii<imgY->height;ii+=2){
   for(int jj=0;jj<imgTempY->width;jj+=2) { //### even columns
```

Subsampling odd rows of non-dominant view:

```
for(int ii=0;ii<imgY->height;ii+=2){
   for(int jj=0;jj<imgTempY->width;jj+=2) { //### even columns
```

Subsampling even rows of non-dominant view:

```
if ((ii+1)%4 && jj%4)
   ptrTempYL [jj] = ptrTempY [jj]*Ws + ptrTempYL[jj]*(1-Ws);
```

4.4.2. View Blending

As it is described earlier, in order to implement this method, an error equation is defined as follows:

$$Error = \omega * \left|\frac{CND + CD}{2} - OD\right| + (1-\omega) * \left|\frac{CND - OND}{2}\right| + (1-\omega) * \left|\frac{CD - OD}{2}\right|$$

Error parameter must be minimized to obtain the best converted view. The algorithm

that finds this minimum value, tries all possible converted values for both dominant and non-dominant views.
First of all, all known variables like OD and OND should be calculated. There are two for-loops to repeatedly execute block averaging for 2 dimensions of the image:

```
for( int ii=0;ii<imgY->height;ii+=2 ){
   for(int jj=0;jj<imgTempY->width;jj+=2) {//### even columns

      // Original Dominant View
      OD  = (  float(ptrTempY[jj]) +
                        float(ptrTempY[jj+1]) +
                        float(ptrTempY[jj+FrameWidth])+
                        float(ptrTempY[jj+FrameWidth+1])) / 4.0 ;

      // Original Non-Dominant View
      OND = (  float(ptrTempYL[jj]) +
                        float(ptrTempYL[jj+1])+

   float(ptrTempYL[jj+FrameWidth])+float(ptrTempYL[jj+FrameWidth+1
])) / 4.0;
   }
}
```

Then the error is calculated as follows:

```
for(int CD0=0;CD0<56;CD0++){
   for(int CND0=0;CND0<56;CND0++){
      // error equation to be minimized
      Error = Wb*abs(((CND0+CD0)/2.0-OD) +
         (1.0-Wb)*abs(CND0-OND)/2.0 +
         (1.0-Wb)*abs(CD0-OD)/2.0 );
                                 }
                        }
```

Finally, the parameters which make the minimum error are replaced by the original values:

```
if (Error<Error0){
   CD=CD0;
   CND=CND0;
}
Error0=Error;
```

Where all variable are introduced in 3.3.4. It should be mentioned at this point that the weight variable varies between 0 and 1.

4.4.3. Low-pass filtering

Low-pass filtering on every full-HD frame is time consuming task and adds to the computational process of the program. Because the filter parameters are fixed and there is no need to change them during the test, this method was implemented by Matlab internal functions and the results were saved as files on hard disk. While running the program, filtered sequences were opened as well as original sequences in the memory.

Internal Matlab function of "fspecial" creates Gaussian filter using the following equations:

$$h_g(n_1, n_2) = e^{\frac{-(n_1^2+n_2^2)}{2\sigma^2}} \tag{4.4.3.1}$$

$$h_g(n_1, n_2) = \frac{h_g(n_1,n_2)}{\sum_{n1}\sum_{n2} h_g} \tag{4.4.3.2}$$

The filter matrix is made by:

h = fspecial('gaussian', hsize, sigma) (4.4.3.3)

where:
hsize is specifying number of rows and columns.
Sigma is substituted in eq(4.4.3.1)

Filter size is set to 11 and sigma to 5 in the following example which can be seen in Figure 4.2.

After opening and reading the sequences in YUV format, every frame was stored as Y, U and V component frames. The following Matlab code applies the filter to each image and later they are again saved as YUV format.

```
for frInd=1:frameCount % For each frame

    Yf = imfilter (Y,f);
    fwrite(fout,Yf','uchar');

    Uf = imfilter (U,f);
    fwrite(fout,Uf','uchar');

    Vf = imfilter (V,f);
    fwrite(fout,Vf','uchar');

end
```

4.5. Optimization

After running the program including all methods executing consecutively, a dynamic program analysis measuring the real-time output presentation shows the actual frame-per-second and compares it to the desired fps. Although the desired frame rate for full-HD sequences is 25 fps but profiling output turned out to be around 15 and it was mainly because so far all the algorithms were implemented in their simplest form.

Several optimization methods were applied to gradually improve the performance of the program such as combining the methods in a single loop to prevent multiple memory read and write operations. These methods improved the performance to around 20 fps which is insufficient for the application. The other important optimization method is to add more threads of process in parallel. Multithreading is the ability to simultaneously have multiple points of execution which are called threads. It can have several benefits such as better resource utilization, simpler program design in some situations, more responsive programs Figure 4.6.

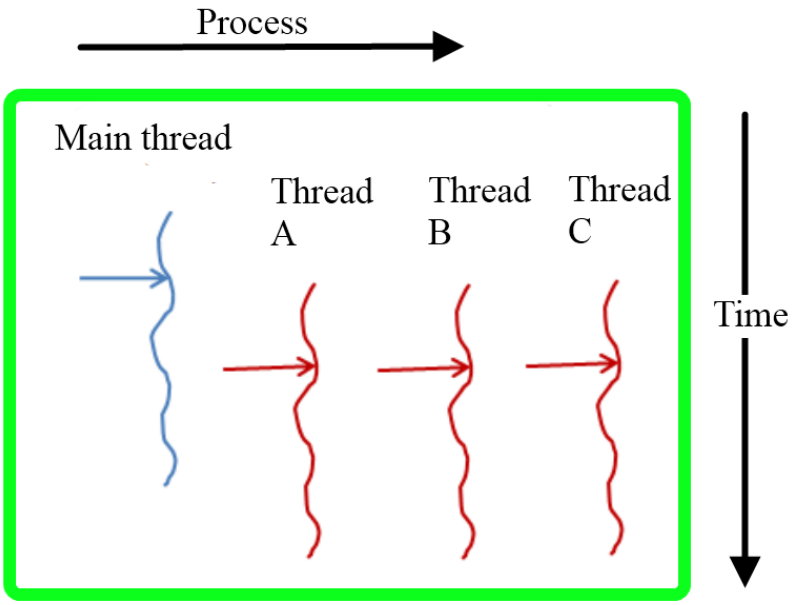


Figure 4.6. *Adding multiple threads to the main process of the program.*

Multithreading method considerably increased the performance of the program from 21 to 28 fps for full-HD sequences and from 26 to 42 fps for HD sequences. Figure 4.7 compares the computational delay for 1 to 4 processing threads in milliseconds for HD and Full HD frame sizes. The profiling is performed on Intel Core i5 CPU @ 2.4 GHz.

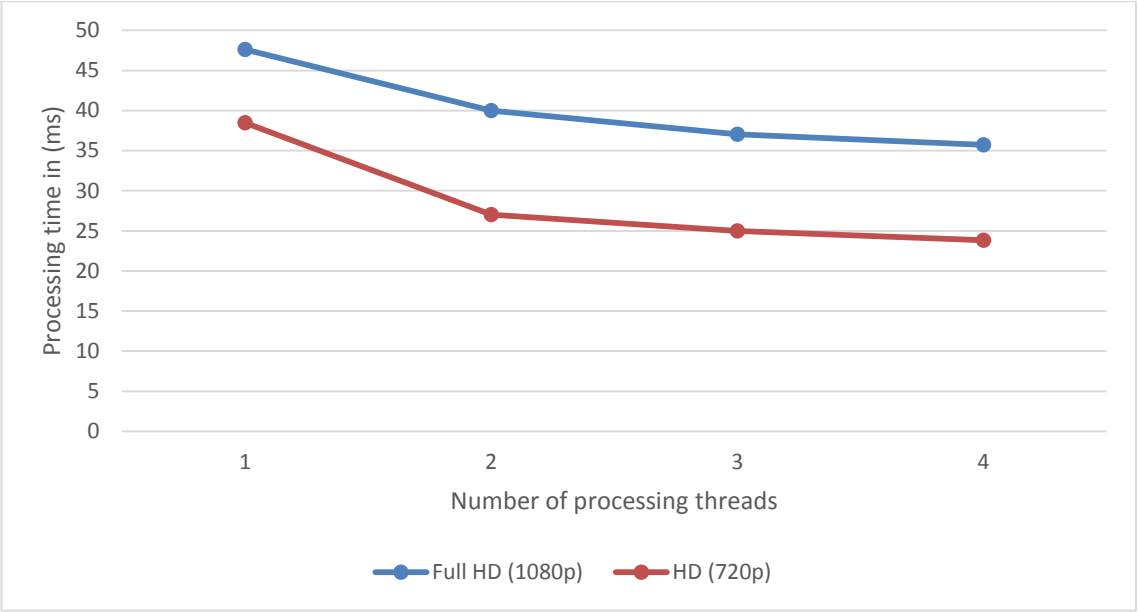


Figure 4.7. *Processing time versus number of threads for 720 and 1080 frame size.*

As it can be deduced from the chart, the number of thread has less effect on processing time after 3 or 4. So, the optimum number of threads is 4 as it adds more computational complexity in comparison to efficiency for more than 4 threads.

5. SUBJECTIVE TEST DESCRIPTION AND RESULTS

In order to evaluate the quality of the modified sequences, large scale subjective assessment was performed. The described software was used to execute all the methods during the test and a list of sequences with corresponding weight values were provided to the software as a script file. This chapter focuses on the subjective test steps, setup, and results.

5.1. Pre-test evaluations

Visual acuity test is required for the subjects to confirm that the subject has enough visual acuity and they are not suffering from stereo blindness or impaired stereo vision. Hence a pre-test evaluation is included prior to subjective tests.

5.1.1. Visual acuity and stereoscopic vision test

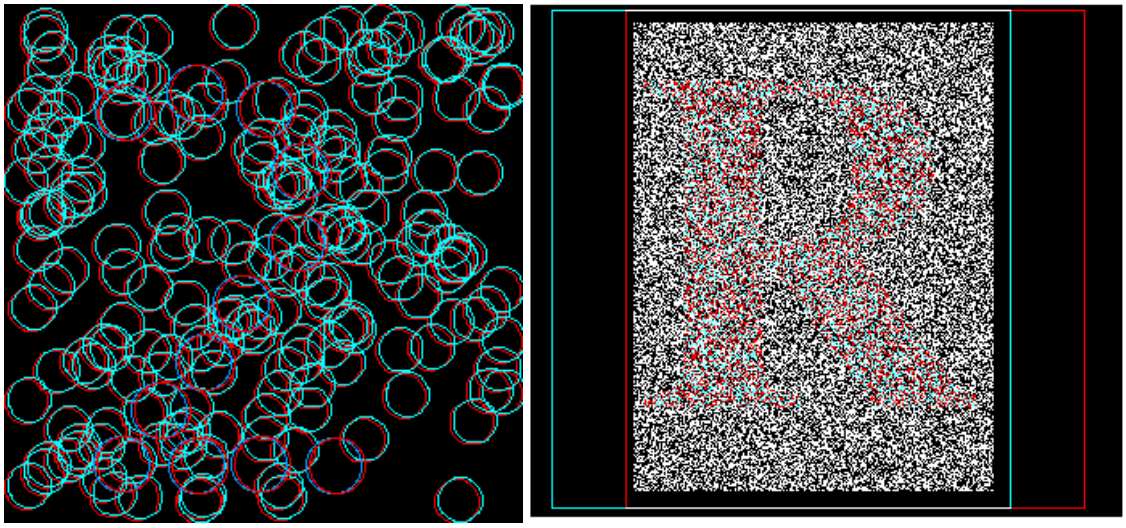
Many test methods have been performed to test the role of contrast and luminance visual acuity for medical purposes to test the acuteness or clearness of vision which depends of the sharpness of the retinal in the eye and its sensitivity. When it comes to artificial 3D vision experience, not everyone liked the 3D craze. Experts believe that 2 to 12 percent of all viewers are not happy with the video shown in 3D. There are two main reasons for that, first, they might be unable to see the 3D effect. Secondly, they might be able to see the 3D effect but it has some side effects like dizziness or headache especially after long time 3D movies observation. Based on these two categories, they are called *stereo–blind* if they cannot see the 3D effect, or having *monocular* vision or lacking *depth perception*. From ophthalmology point of view, medical disorders that prevent the eye focusing on one point properly or loss of vision in one eye can be the reason.

Unfortunately, 3D vision technology does not have a general solution for the stereo-blind. And the best approach may depend on their situation. For example if you find 3D movies uncomfortable you can watch the movie in 2D by wearing the special glasses which have the same filters on both sides. This totally filters one view and both eyes see the same view.

5.1.2. Depth perception test

The depth perception test must be performed prior to the subjective test in order to make sure that the subjects are not suffering from stereo-blindness. Depth perception test is

performed by showing a 3D image and asking the subject to try to detect the depth of objects. Figure 5.1 shows the 3D image designed to include many randomly distributed circles in same depth, while a few more circles are added to appear in a closer depth to make a recognizable figure which is the figure of number two in this example. If one look at the picture with bare eyes, nothing can be recognized but some random circles. But if a normal subject wearing 3D glasses stares at the picture, after less than one minute they must be able to figure out the number hidden in the picture. Figure 5.1a is illustrated in color anaglyph system to be more effective in printed.



(a) (b)

Figure 5.1 *Anaglyph depth perception test images.*

The second depth perception image in Figure 5.1b is using similar technique using smaller particles distributed in three dimension. If an object could recognize the outstanding particles, they could recognize the letter “R” as different depth than the rest of image.

5.2. Test setup

In order to discover good trade-offs for the three processing components, disparity selection, low-pass filtering and contrast adjustment, large scale subjective assessment was performed with four sequences: Poznan Hall2, Poznan Street [33], Ghost Town Fly (GT Fly), Undo Dancer, which are part of 3DV MPEG CfP [28]. For GT Fly and Undo Dancer sequences 500 frames were utilized while 250 and 200 frames were utilized for Street and Hall2, respectively. The frame rate was fixed to 25 Hz for all sequences. Input views and camera separation distances utilized in our experiments, for both small and big disparity stereoscopic videos, are shown in Table 5.1. Note that the camera separation of bigger disparity is the same as those introduced in MPEG 3DV CfP for C3 scenario while in the smaller disparity scheme the camera separation distance is halved.

Subjective condition was conducted according to the conditions suggested in MPEG 3DV CfP. The polarized 46'' Vuon E465SV 3D TV set by Hyundai with a total resolution of 1920x1200 pixels and a resolution of 1920x600 per view when used in the stereoscopic mode was utilized for displaying of the test material. The viewing distance was equal to 4 times of the displayed image height (2.29m).

Table 5.1. *Input views and camera distances or small and big camera separations.*

	Left view-Right view , (Camera separation in cm)	
Sequence	Small disparity	Big disparity
Poznan Hall2	7-6.5 , (6.87)	7-6 , (13.75)
Poznan Street	5-4.5, (6.87)	5-4 , (13.75)
GT Fly	3-1 , (4)	5-1 , (8)
Undo Dancer	1-3 , (4)	1-5 , (8)

5.3. Preparation of Test Stimuli

To prepare test material, we utilized three adaptation methods presented in section II and various test cases based on different combinations of adaptation methods created. In our experiments, we tested contrast reduction to 50% and 75% of the original values for different combinations. Moreover, all non-dominant views were low pass filtered utilizing the circular averaging filter with radius equal to 6 as presented in equation (3.2).

Table 5.2. *Disparities for small and big camera separation.*

	Average disparity (Maximum disparity) in arcmin	
Sequence	Small disparity	Big disparity
Poznan Hall2	18.6(22.2)	37.2(44.3)
Poznan Street	19.3(23.6)	38.6(47.2)
GT Fly	12.1(42.2)	24.3(84.3)
Undo Dancer	13.6(22.2)	27.2(47.2)

Two different disparities between the left and right view were selected for different sequences. For under test sequences the disparity was always positive for instance having the objects always behind the display level. Disparity selection was limited so that the results were in agreement with previous achievements literature to prevent eye strain due to big disparities. Considering that disparity is dependent on the location of object in the scene and even for each object it can change for different frames, we calculated the disparity of each sequence as an average over disparities of objects in foreground of each frame. Then, these values were averaged over the whole sequence to create one value presenting the disparity of respective stereoscopic sequence. This was performed by comparing the pixel values of left and right views in depth maps. Objects were found by utilizing a threshold which by crossing, we could recognize an edge and hence, pres-

ence of an object with different depth compared to background in the scene. Comparing the location of these edges in each row between left and right view, we calculated the disparity in pixels between left and right view for that specific object presented in respective row. Averaging this disparity for all rows provides the disparity in pixels for one frame. Averaging the frame disparity in pixels over the whole sequence provides the disparity in pixels for that sequence.

Table 5.2 presents the average and maximum disparities per sequence. Moreover, Table 5.1 presents the selected views and respected camera separations for different disparities of utilized sequences. For Poznan Hall2 and Poznan Street sequences, the views 6.5 and 4.5, respectively, were synthesized from original texture and depth views using the MPEG View Synthesis Reference Software (VSRS) version 3.5 [29]. The subjective quality of synthesized views was comparable to that of other original views. Moreover, since the synthesize artifacts were subjectively negligible, we assume that the synthesizing process did not affect the subjective ratings.

Combining above methods, seven following test schemes were prepared and subjectively assessed. The combinations for each scheme are presented in the format of (disparity-contrast) where for disparity the values *0, Small, Big* refer to *0* disparity (identical left and right views), *Small* disparity, and *Big* disparity, respectively. For contrast the values *X%* present the contrast reduction ratio of the non-dominant view.

1. (*0-100%*) – (O) → (Original 2D)
2. (*Small, 100%*) – (S1)
3. (*Small, 75%*) – (S2)
4. (*Small, 50%*) – (S3) → (Best 2D quality)
5. (*Big, 100%*) – (B1) → (Best 3D quality)
6. (*Big, 75%*) – (B2)
7. (*Big, 50%*) – (B3)

5.3.1. Test Procedure and Participants

Subjective quality assessment was done according to Double Stimulus Impairment Scale (DSIS) method [30] with discrete unlabeled quality scale from 1 to 10 was utilized for quality assessment. Test was divided to two sessions where in first session, subjects assessed the subjective quality of videos with glasses and in the second session, the test was performed without glasses. Two questions for each session of the test were considered and subjects wrote their ratings after each clip was played. These questions are presented in Table 5.3. Each question is associated with its short term for simplicity in reporting the results. Prior to each test, subjects were familiarized with the test task, the test sequences and the variation in quality they could expect in the actual tests. The viewers were instructed that 0 stands for the lowest quality and 10 for the highest.

Subjective viewing was conducted with 20 subjects, (16 male, 4 female), aged between 21-31 years (mean: 24.2). All subjects passed the test for stereovision prior to the actual test. Moreover, they all were considered naïve as they did not work or study in

fields related to information technology, television or video processing. To prevent subject from getting exhausted in the subjective session, the duration of the test was limited to 45 minutes.

5.4. Results and discussion

In this section we present the results of the conducted subjective test and an analysis on the statistics of the quantitative viewing experience ratings.

Figure 5.2 shows the subjective viewing experience ratings with 95% confidence interval (CI) for all sequences. The results are provided for four questions that users were asked during the test session. Subjective ratings show that scheme O achieved the highest value in 2D evaluation (session where viewing took place without glasses) and general quality of 3D presentation. However, since the smallest depth perception was rated in this scheme it cannot be considered as a competitor for best compromise for simultaneous 2D and 3D perception. Hence, it was excluded from the analysis presented in the following paragraphs. For the other tested schemes, the following general trend was observed. In both small and big disparities, while decreasing the contrast reduction ratio of non-dominant view, the ratings of the 2D evaluation session increase and at the same time the 3D evaluation ratings decrease. This was expected as reducing the contrast of non-dominant view targets ideal 2D subjective quality while compensating the 3D perception. Moreover, in all sequences, ghosting effect in 2D presentation of stereo videos without any contrast adjustment, annoyed subjects more in bigger disparity scheme. Considering large amount of viewing experience ratings, it is not possible to make many logical conclusions based on Figure 5.2. Hence, significant differences between the schemes were further analyzed using statistical analysis as presented in the paragraphs below.

Non-parametric statistical analysis methods, Friedman's and Wilcoxon's tests, were used as the data did not reach normal distribution (Kolmogorov-Smirnov: $p<0.05$). Friedman's test is applicable to measure differences between several and Wilcoxon's test between two related and ordinal data sets [31]. A significance level of $p< 0.05$ was used unless in the analysis.

The following conclusions were obtained with this statistical significance analysis presented above. In the analysis, we pairwise compared each two combinations per question rating resulting in fifteen flags presenting whether the subjective quality of different combinations have any statistically significant difference. Considering four sequences, four questions per sequence, and fifteen pairwise comparisons per question, we achieved 4×4×15 = 240 flags. Table 5.3 reports the summary of these flags. Each cell presents total number of flags from different questions where -1, 0, and 1 present significantly lower, similar, and significantly higher quality compared to other schemes. From this table it is clear that only S2 provides similar or better subjective results for all sequences while other schemes have lower performance at least in one sequence. Hence, combination utilized in S2 seems to be a well-designed potential candidate for simultaneous 2D and 3D presentation.

The conclusion that S2 provides the most acceptable trade-off for simultaneous 2D and 3D viewing is in agreement with previous findings on contrast asymmetry in [32] where contrast difference limit between left and right view was found to be equal to or less than 25% to provide equal viewing comfort. Moreover, considering camera separations presented in Table 5.2, the perceived disparity for all sequences was aligned with the results presented in [17] where the limit for maximum disparity between the left and right views was found to be 70 arcmin. Only the maximum disparity of bigger camera separation for GT_Fly is above this limit. This big disparity happens for 0.12 seconds in the 10 second sequence (3 frames in 250 frames). Moreover, as proposed by authors in [17], utilization of LPF can increase the limit of disparity. Figure 5.2 depicts a 2D

Table 5.3. *Flag table presenting significant differences for different test schemes*

		Test scheme combinations					
	Flags	S1	S2	S3	B1	B2	B3
Dancer	-1	2	**1**	0	3	2	3
	0	17	**17**	16	16	18	14
	1	1	**2**	4	1	0	3
GT Fly	-1	4	**1**	1	5	2	1
	0	16	**17**	13	13	17	16
	1	0	**2**	6	2	1	3
Street	-1	4	**2**	3	7	4	4
	0	13	**13**	9	10	15	12
	1	3	**5**	8	3	1	4
Hall2	-1	4	**3**	6	4	0	4
	0	12	**14**	12	12	14	14
	1	4	**3**	2	4	6	2

presentation of stereoscopic videos from scheme S2 and the same stereoscopic video with equal disparity and without any LPF or contrast adjustment applied.

After the test, the participants were asked whether they experienced any fatigue or eye strain during and/or after the test. Subjects seemed quite comfortable and there were no complaints regarding the 3D content and asymmetric nature of the stereoscopic videos. However, five subjects complained slightly about the variety of the clips that they observed, mentioning that sometimes it was difficult to distinguish the differences between observed clips [32].

(a) (b)

Figure 5.2. *2D presentation of stereoscopic videos from combinations (1) O and (2) S2.*

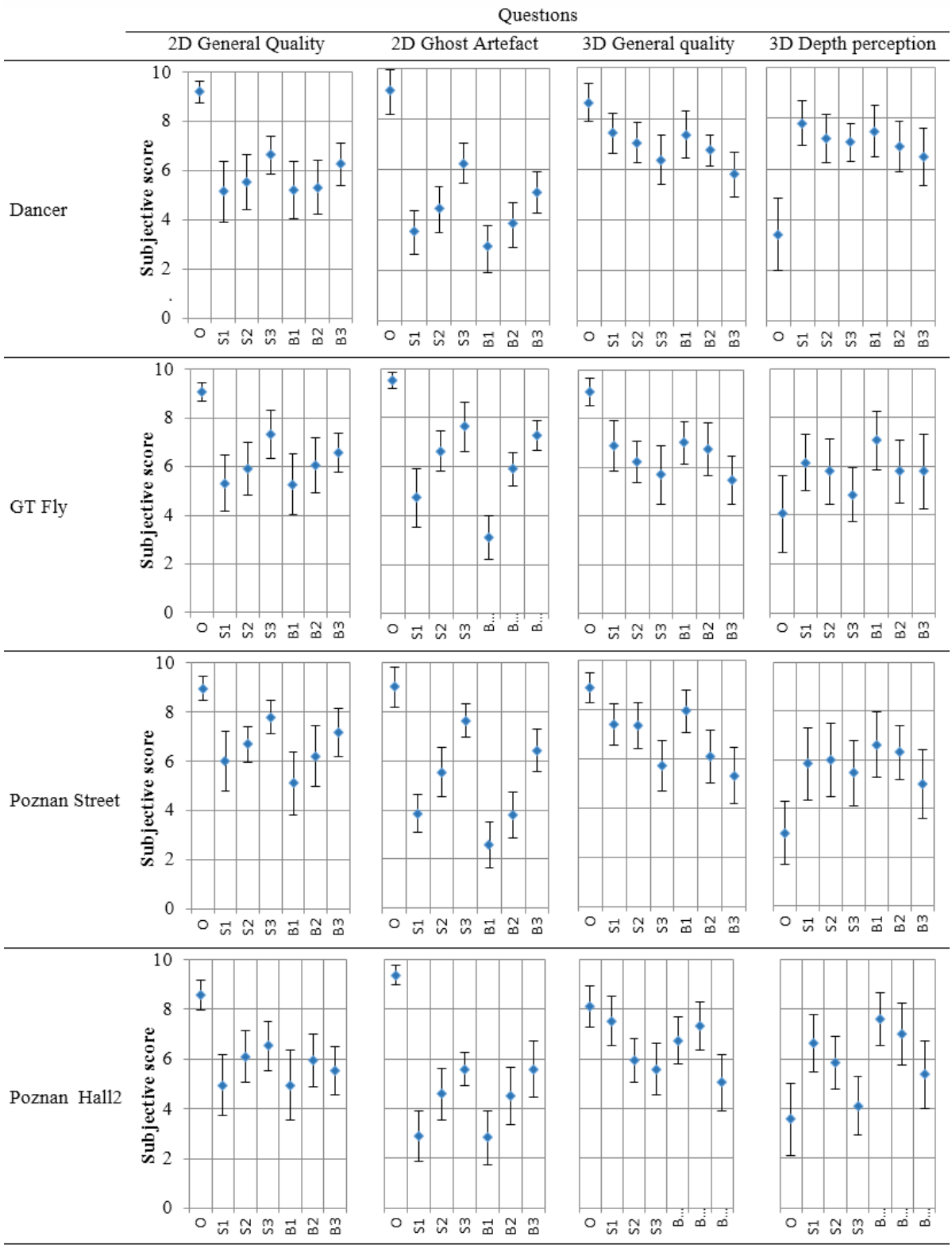


Figure 5.2. *Viewing experience ratings with 95% confidence interval [32].*

6. CONCLUSION AND FUTURE WORK

In this thesis, we tackled the problem of ghosting artefact visible for stereoscopic content while observed without glassed. Therefore, considering the possibility of observing the 3D content with and without glasses simultaneously, a new algorithm was introduced exploiting different steps to render views. These steps include:

- Disparity adjustment
- Contrast adjustment
- Subsampling
- View Blending
- Low-pass filtering

In addition to mitigation of ghosting artefact, the amount of high frequency components in non-dominant view is much lower compared to the original view. This will benefit compression and podcasting performance by reducing the necessary bitrate for encoding the respective view.

Moreover, a software was implemented to perform such rendering in real time considering polarized displays. Considering that the same content was supposed to be evaluated with and without glasses, the only way to evaluate the performance of proposed algorithm was to conduct a subjective test. Therefore, we tested the performance of proposed method and introduced software in a systematic series of subjective tests. The ratings showed that our solution better satisfies the 2D and 3D subjects compared to conventional pure 2D or 3D solutions. Considering that this is a novel topic in the research community, we expect to have more use cases for it in the near future.

As for future steps, the possibility of exploiting such system in auto stereoscopic displays should be considered. Moreover, we have plans to conduct more subjective tests including new rendering algorithms merged with current implemented methods. Furthermore, we have started testing the algorithm changing the type and strength of exploited LPF as well as adjusting the disparity to smaller values testing the limit of disparity that can be used in this rendering scenario.

REFERENCES

[1] C. Wheatstone, “On some remarkable, and hitherto unobserved, phenomena of binocular vision,” Philos. Trans. R. Soc. Lond., 1838.

[2] von Helmholtz H, “Handbuch der physiologischen Optik,” Leopold Voss 1866, (English ed. 1962, Dover New York)

[3] H. Asher, “Suppression theory of binocular vision,” The British Journal of Ophthalmology, vol. 37, no. 1, pp. 37-49, 1953

[4] B. Julesz, “Foundations of Cyclopean Perception”, University of Chicago Press, Chicago, IL, USA, 1971.

[5] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj,” Impact of downsampling ratio in mixed-resolution stereoscopic video”, Proc. of 3DTV Conference, June 2010.

[6] L. B. Stelmach, W. J. Tam, D. V. Meegan, A. Vincent, and P. Corriveau, “Human perception of mismatched stereoscopic 3D inputs,” in Proc. ICIP, Vancouver, BC, Canada, 2000.

[7] PALMER, S. E. 1999. Vision Science: Photons to Phenomenology. The MIT Press.

[8] Dodgson, N. (2004). Variation and extrema of human interpupillary distance. Proceedings of the SPIE, 5291:36–46.

[9] Meegan, D. V., Stelmach, L. B., and Tam, W. J. (2001). Unequal weighting of monocular inputs ininocular combination: implications for the compression of stereoscopic imagery. Journal of Experimental Psychology: Applied, 7:143–153.

[10] Johanson, M. (2001). Stereoscopic video transmission over the internet. Proceedings of the IEEE Workshop on Internet Applications, pages 12–19.

[11] Levelt, W. (1965). On Binocular Rivalry. Assen, The Netherlands: Royal VanGorcum.

[12] Perkins, M. G. (1992). Data compression of stereopairs. IEEE Transactionson Communications, 40:684–696.

[13] Meegan, D. V., Stelmach, L. B., and Tam, W. J. (2001). Unequal weighting of monocular inputs in binocular combination: implications or the compression of stereoscopic imagery. Journal of Experimental Psychology: Applied, 7:143–153.

[14] P. Aflaki, M. Hannuksela, J. Hakkinen, P. Lindroos, and M. Gabbouj, “Subjective study on compressed asymmetric stereoscopic video” in Proc. Int. Conf. Image Process., Sep. 2010, pp. 4021–4024.

[15] H. Asher, “Suppression theory of binocular vision,” The British Journal of Ophthalmology, vol. 37, no. 1, pp. 37-49, 1953

[16] P. Seuntiens, Visual Experience of 3D TV, Ph.D. thesis, 2006. 66, 67

[17] W.A. IJsselsteijn, H. de Ridder, J. Freeman, S.E. Avons, “Presence: concept, determinants, and measurement,” Proceedings of the SPIE, Vol. 3959, pp. 520-529, 2000.

[18] W.A. IJsselsteijn, H. de Ridder, J. Freeman, S.E. Avons., and D. Bouwhuis, “Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence,” Presence-Teleoperators and Virtual Environments, vol. 10, no. 3, PP. 298-311, 2001.

[19] P. Seuntiens, L. Meesters, and W. IJsselsteijn, “Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation,” ACM Trans. Appl. Perception (TAP), vol. 3, pp. 95–109, 2006.

[20] A. Boev, D. Hollosi, and A. Gotchev, “Classification of stereoscopic artefacts”, MOBILE3DTV Project report, 2011, available on http://mobile3dtv.eu.

[21] Y. Chen, Y.-K. Wang , K. Ugur , M. Hannuksela , J. Lainema and M. Gabbouj "The emerging MVC standard for 3D video services", EURASIP J. Adv. Signal Process., vol. 2009, no. 1, 2009

[22] Wheatstone, C. (1838). Contributions to the physiology of vision: Ii. on some remarkable, and hitherto unobserved, phenomena of binocular vision. Philosophical Transactions of the Royal Society, 128:371–394.

[23] B. Julesz, “Foundations of Cyclopean Perception”, University of Chicago Press, Chicago, IL, USA, 1971.

[24] P. Aflaki, M. M. Hannuksela, J. Häkkinen, P. Lindroos, M. Gabbouj,” Impact of downsampling ratio in mixed-resolution stereoscopic video”, Proc. of 3DTV Conference, June, 2010.

[25] A. Vetro, T. Wiegand and G. Sullivan "Overview of the stereo and multiview video coding extensions of the H.264/AVC standard" Proc. IEEE.

[26] International Telecommunications Union, ITU-R BT.601

[27] “Call for Proposals on 3D Video Coding Technology,” ISO/IEC JTC1/SC29/WG11 MPEG2011/N12036, Geneva, Switzerland, March 2011.

[28] “View synthesis software manual,” MPEG ISO/IEC JTC1/SC29/WG11, Sept. 2009.

[29] ITU-R Rec. BT.500-11, Methodology for the subjective assessment of the quality of television pictures, 2002.

[30] H. Cooligan “Research methods and statistics in psychology,” (4th Ed.). London: Arrowsmith, 2004.

[31] S. Pastoor, Human factors of 3D imaging: results of recent research at Heinrich-Hertz-Institut Berlin Proceedings of the International Display Workshop ’95 (Asia Display ’95) (1995).

[32] P. Aflaki, M. M. Hannuksela, H. Sarbolandi, and M. Gabbouj, "Simultaneous 2D and 3D perception for stereoscopic displays based on polarized or active shutter glasses," Journal of Visual Communication and Image Representation, 2013

[33] M. Domañski, T. Grajek, K. Klimaszewski, M. Kurc, O. Stankiewicz, J. Stankowski, K. Wegner, Poznan multiview video test sequences and camera parameters, MPEG 2009/M17050, October 2009.