PERCEPTUALLY DRIVEN STEREOSCOPIC CAMERA CONTROL IN 3D VIRTUAL ENVIRONMENTS

A THESIS

SUBMITTED TO THE DEPARTMENT OF COMPUTER ENGINEERING

AND THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE

OF BILKENT UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE

By

Elif Bengü Kevinç

August, 2013

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Asst. Prof. Dr. Tolga Çapın(Advisor)

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Prof. Dr. Bülent Özgüç

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

_____

Prof. Dr. Veysi İşler

Approved for the Graduate School of Engineering and Science:

_____

Prof. Dr. Levent Onural
Director of the Graduate School

ii

ABSTRACT

PERCEPTUALLY DRIVEN STEREOSCOPIC CAMERA CONTROL IN 3D VIRTUAL ENVIRONMENTS

Elif Bengü Kevinç
M.S. in Computer Engineering
Supervisor: Asst. Prof. Dr. Tolga Çapın
August, 2013

Depth notion and how to perceive depth have long been studied in the field of psychology, physiology, and even art. Human visual perception enables to perceive spatial layout of the outside world by using visual depth cues. Binocular disparity among these depth cues, is based on the separation between two different views that are observed by two eyes. Disparity concept constitutes the base of the construction of the stereoscopic vision.

Emerging technologies try to replicate binocular disparity principles in order to provide 3D illusion and stereoscopic vision. However, the complexity of applying the underlying principles of 3D perception, confronted researchers the problem of wrongly produced stereoscopic contents. It is still a great challenge to give realistic but also comfortable 3D experience.

In this work, we present a camera control mechanism: a novel approach for disparity control and a model for path generation. We try to address the challenges of stereoscopic 3D production by presenting comfortable viewing experience to users. Therefore, our disparity system approaches the accommodation/convergence conflict problem, which is the most known issue that causes visual fatigue in stereo systems, by taking objects' importance into consideration. Stereo camera parameters are calculated automatically with an optimization process. In the second part of our control mechanism, the camera path is constructed for a given 3D environment and scene elements. Moving around important regions of objects is a desired scene exploration task. In this respect, object saliencies are used for viewpoint selection around scene elements. Path structure is generated by using linked Bézier curves which assures to pass through pre-determined viewpoints.

Though there is considerable amount of research found in the field of stereo creation, we believe that approaching this problem from scene content aspect

provides a uniquely promising experience. We validate our assumption with user studies in which our method and existing two other disparity control models are compared. The study results show that our method shows superior results in quality, depth, and comfort.

ÖZET

3B SANAL ORTAMLARDA ALGIYA DAYALI STEREOSKOPİK KAMERA KONTROLÜ

Elif Bengü Kevinç
Bilgisayar Mühendisliği, Yüksek Lisans
Tez Yöneticisi: Asst. Prof. Dr. Tolga Çapın
Ağustos, 2013

Derinlik kavramı ve derinliğin nasıl algılandığı psikolojide, fizyolojide, hatta sanatsal çalışmalarda uzun süredir incelenmektedir. İnsanlardaki görsel algı sistemi, dış dünyanın yerleşimini görsel derinlik ipuçlarını kullanarak anlamaktadır. Bu derinlik ipuçlarından biri olan binoküler disparite iki göz tarafından yakalanan iki farklı görüntü arasındaki ayrılığa dayalı olarak oluşmaktadır.

Gelişen teknolojiler 3B yanılsamasını sağlamak ve stereoskopik görüntüleri oluşturabilmek amacıyla binoküler disparite prensiplerini kopyalamayı denemektedirler. 3B algısının oluşturulabilmesi için gereken prensiplerin uygulanabilirliğinin karmaşıklığı, araştırmacıları yanlış şekilde üretilen stereoskopik içerikler oluşturmaları problemiyle karşı karşıya getirmiştir. Gerçekçi ve konforlu 3B deneyimi sunabilmek hala zor bir çalışma konusudur.

Çalışmamızda disparite kontrolünü sağlayan yeni bir yaklaşım ile yol oluşturmayı sağlayan bir modelden oluşan kamera kontrol mekanizması sunulmuştur. Kullanıcılara konforlu bir seyir deneyimi sunmak adına stereoskopik 3B üretimi esnasında karşılaşılan sorunların üzerine eğilmeye çalışılmıştır. Akomodasyon ve yakınsama uyuşmazlığı 3B sistemlerde karşılaşılan göz yorgunluğuna neden olan en büyük problemdir. Bu nedenle sunduğumuz disparite sistemi akomodasyon ve yakınsama kavramlarının uyuşmamasından doğan problemi, sahne elemanlarının önem derecelerini dikkate alarak ele almaktadır. Stereo kamera parametreleri bu evrede optimizasyon işleminden geçirilerek otomatik olarak hesaplanmaktadır. Kontrol mekanizmamızın ikinci kısmında ise verilen bir 3B ortam için kameranın izleyeceği yol oluşturulmaktadır. Önemi yüksek olan objelerin dikkat çeker kısımlarına bakarak o sahneyi incelemek, tercih edilen bir sahne analiz yöntemidir. Sahne elemanları etrafındaki bakış noktalarının seçilebilmesi için objelerin dikkat çekerlilikleri kullanılmıştır. Yol yapısı belirlenen

bakış noktalarından geçmekte olan, birbirlerine bağlı Bézier eğrileri kullanılarak oluşturulmuştur.

Stereo oluşturulması için çok çeşitli çalışmalar bulunmakla birlikte bu konuya sahne içerikleri açısından yaklaşmak ümit verici bir deneyim sağlamıştır. Sunmuş olduğumuz yaklaşımın geçerliliği, kendi methodumuzu var olan diğer iki disparite kontrol modelleriyle karşılaştırdığımız deneyler ile gösterilmiştir. Deneyler methodumuzun görsel kalite, derinlik ve rahatlık üzerine üstün sonuçlar gösterdiğini doğrulamaktadır.

Anahtar sözcükler: Stereoskopik 3D, Kamera Kontrolü, Disparite Kontrolü.

Acknowledgement

I would like to express my thanks to my advisor Asst. Prof. Dr. Tolga Çapın for giving me the opportunity to do my master in a leading university. I also appreciate him for his courtesy, guidance, and support.

I would like to thank to my thesis committee members Prof. Dr. Bülent Özgüç and Prof. Dr. Veysi İşler for accepting my invitation without hesitation, spending their time to evaluate my thesis, and their valuable comments.

The biggest part of my gratitude belongs to my lovely family. My mother Nuray Kevinç, my father Kahraman Kevinç, and my brother Bilinç Kevinç always endeavoured to provide me the best of all, also encouraged me to do my best. Without their endless love, guidance, help, and support through all my life, not only this thesis would be completed, but also I would not be the person who I am. They taught the importance of being a good person, righteous, and kind.

Also, special thanks go to my great friends, Sinan Arıyürek, Gökçen Çimen, Gizem Mısırlı, Elif Eser, Seher Acer, Zeynep Korkmaz, Can Telkenaroğlu, Sami Arpa, Bertan Gündoğdu, and Shatlyk Ashyralyyev. They coloured my life in many ways, and always with me during tough times. Thanks to them, my graduate education is filled with unforgettable memories. I am so lucky to have these people in my life.

Finally, I would like to acknowledge the Scientific and Technical Research Council of Turkey (TÜBİTAK) for financially supporting my graduate education.

To my family...

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Understanding the layout of the outside world is important to perceive shapes and to estimate distances of objects which are main capabilities of our visual system. Therefore, the question of how our surrounding is understood has always been an important issue in a variety of fields. Physiologists try to solve how the brain shows the world as a result of a visual construction process and psychologists analyse how this shaping process occurs by approaching from perception angle. Even artists ponder this issue in order to replicate this feature in their work of art in order to create more realistic products. All these researches focus at one point that main principles exists in order to perceive surrounding.

Visual cortex is responsible of constructing visual representation of the world we are living in, which is also called as depth perception. Spatial layout between the objects is processed in the cortex by using depth cues which are responsible of constructing the outside world. These depth cues can be categorized as pictorial, oculomotor, binocular, and motion-related cues [2]. Among depth cues, binocular cues come to the fore with its feature of providing depth information and distance, while other cues help to understand spatial relationships between objects located in the three-dimensional (3D) space of our surrounding. Figure 1.1 shows a wireframe cube known as Necker cube. Visual depth cues except binocular cues are not sufficient to understand locations of sides of the cube with respect to each other. The most basic working principle underlying human visual
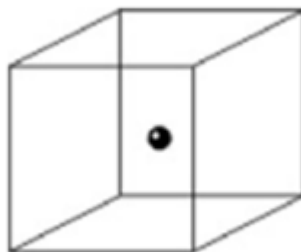
Figure 1.1: Which part is front? Which part is back? Where the dot is standing on? A wireframe structure so-called Necker Cube contains no depth cues.

perception mechanism in order to comprehend the world in 3D is based on separation between two different views that are observed by two eyes, which is binocular disparity, and replicating this feature enables to convey depth realistically. Therefore stereoscopic displays use the same principle and produce binocular disparity by providing two different perspective images, captured from two cameras, for two eyes. Binocular disparity is the underlying principle of stereoscopic 3D.

3D analogy is an intriguing concept and earlier studies on depth illusion date back to 17th century. In the late 17th century, it was discovered that presenting two separate images instead of one image enhance the depth feeling in the paintings. The desire of feeling immersion led to rising of stereoscopic products in the 19th century. After the rise of the film industry, 3D notion attracted producers and first 3D movie was released in 1952. However, image quality issues restrained to create high quality 3D production which is a process far beyond of that analog age. 3D became a breakthrough in the mainstream cinema in the beginning of the 21st century with the help of technological developments; thrived in many other entertainment areas. 3DTV sets are sold at remarkable numbers, more tv channels begin 3D broadcasting, 3D games attracted people day by day. Information display industry also resort to utilities that 3D presents, since complex data can be comprehended easier by using 3D technology rather than flat 2D images. In spite of all these rapid developments, stereoscopic content production and visualization is still a great challenge in order to provide realistic and comfortable viewing experience. The fundamental problem lies in the complexity of applying

the underlying principles of 3D perception of the human visual system (HVS) and its capabilities/limitations for displaying content in stereoscopic displays.

In this study, we address the challenges of presenting a comfortable viewing experience while displaying stereoscopic contents. The horizontal separation of two eyes, as being the basis to create depth feeling, is applied as the main principle in stereoscopic displays. The horizontal separation of two eyes is the basis to create depth perception as it is explained in the above. We present a novel method to calculate screen disparity, which creates a perceived depth around the display screen. The perceived depth in stereoscopic scenes is achieved by adjusting stereoscopic camera parameters automatically. Interaxial separation, one of the stereoscopic camera parameters, is the distance between two cameras and corresponds to interocular distance or eye separation in HVS. This camera parameter is responsible of generating two slightly different images of the scene like two captured vision from left and right eyes. Convergence distance is the other camera parameter and refers to the distance between the center of two cameras and a point or a plane focused. Convergence distance arise from the need to replicate the effect generated when eyes are rotated. The difference in the views, or screen disparities, are designated by using these stereoscopic camera parameters by taking "stereoscopic comfort zone", which is a notion used for comfortable range of the perceived depth, into consideration.

Our stereoscopic camera system starts with a user-based disparity calibration phase. Perceived depth range varies from person to person, since stereoscopic comfort zone limits change for each user. The maximum and minimum disparity limits that the user is able to perceive is found via this phase. After disparity calibration, our system starts to show given scene content in 3D with screen disparity values that are calculated through our approach.

Our stereo rendering approach composes of three consecutive steps. Depth range is calculated in the first part which simply calculates interaxial separation and convergence distance by geometrically modelling the stereoscopic vision with respect to the user's personal disparity extrema. Then we map scene depth to the obtained depth range. However, we believe that this geometric approach is not

sufficient to handle accommodation/convergence conflict that is the main reason of uncomfortable 3D experience. We enhance this methodology by incorporating scene elements' importances into our algorithm in the second part. With this aim, our system analyzes of the scene environment and finds attention-grabbing objects. Then, the location of the convergence plane is modified according to significance scores of these objects. Our aim is to specify the location of the convergence plane, on which scene elements are captured with exactly zero disparity. This is achieved by locating convergence plane nearer to objects with higher significance rather than other scene elements. This motivation comes from that the user focuses on attention-grabbed scene elements longer and little disparity value of these elements ensures comfort viewing experience. Finally, optimization of stereo camera parameters is performed in the third part. The distance between the convergence plane and scene elements which have relatively higher significance score and lower radial distance from the user's center of attention that is center of the display in our case, is minimized. At the same time, our system aims to maximize the total screen disparity. Our system repeats these process steps for every frame and automatically adapts interaxial separation and converge distance for any scene content. With the user tests we validate that our approach among existing stereo rendering methods presents a more comfortable 3D experience remarkably without losing image quality, or perceived depth aspects.

Researches in the stereo field focus on disparity computation and miss out the other main part of the camera control systems: path finding. Though our proposed system can be used interactively with user input where the user freely navigates in the dynamic environment, we extend the system with saliency based path generation in order to visualize interactive scenes in 3D. Therefore, we combine our disparity control mechanism with camera path finding approach in order to produce an entire 3D camera system. Path generation is done by calculating object saliency, which is used to obtain viewpoints around objects. Passing directions of the camera is also based on these viewpoint selection. Then, the directions of the camera are converted to control points, which refers to key locations that camera passes from. Camera path is generated based on Bézier curves between

control points. The overall process is performed in a semi-automatic manner.

Chapter 2 presents main principles in order to comprehend underlying concepts of our system. Then, existing approaches in disparity control and stereo content production is represented comprehensively in Chapter 3. Proposed system is explained in detail in Chapter 4 and camera control mechanism is explained in Chapter 5. User study to validate our methodology and experimental results are presented in Chapter 6. Chapter 7 concludes the thesis with a summary of the overall system and future work is discussed finally.

Chapter 2

Background

How do we perceive our surrounding world is a question with several answers, and also a complicated procedure processed by HVS. To replicate this process in 3D generation by providing realistic depth feeling illusion is a complicated process as well. Depth perception, stereo geometry, and accommodation/converge conflict are the three key concepts that underlie stereo content production pipeline and our system makes use of characteristics of these concepts. In order to comprehend stereoscopic systems, a summary of basic principles behind them is given in the following sections.

2.1 Depth Perception

Depth cues, which help the human visual system to perceive the spatial relationships between the objects, construct the core part of our depth perception. These depth cues are investigated under two main titles; which are oculomotor and visual depth cues.

Oculomotor Depth Cues

Oculomotor system is responsible of movements of eye muscles as well as pupillary control like constriction or dilation. Therefore, oculomotor depth cues

include the data obtained by muscular activities of the eye lens. In order to fixate on an object, eyes show a muscular response like focusing on object which is known as accommodation, rotating to object which is known as vergence, also pupil size is increased or decreased. These are three depth cues processed by oculomotor system in physiology.

Visual Depth Cues

Visual depth cues are divided into two groups: monocular and binocular depth cues.

Monocular: These depth cues give visual feedback that comes from one eye to HVS. Pictorial and motion based cues constitutes monocular depth cues. Pictorial ones provide to extract depth information from a single and flat 2D view and include occlusion, cast shadow, shading, linear perspective, relative height, relative size, texture gradient, aerial perspective etc. Pictorial cues are also used by artists in 2D paintings for centuries. Motion based cues allow us to understand a depth information during a motion, by using movements of objects or viewers. The difference of their motion in a short time period creates difference between their images relative position on retina. The difference between images on each view gives an approximate movement information. These cues include motion parallax, motion perspective, and dynamic occlusion. Although, all these monocular cues give information about outside world and positions of objects from one single view, they are not enough to give illusion of depth and absolute distance. Binocular cues come into play at this point.

Binocular: Binocular visual depth cues make a comparison between point of views of two eyes by using discrepancies between two retinal images on two eyes. Stereoscopic production researches focus on binocular visual depth cues in order to take advantage of this concept in stereoscopic applications. Binocular disparity, also known as stereopsis, constitutes the base of the stereo geometry in the construction of the stereoscopic vision, which is covered extensively in the following subsection.

2.2 Stereo Geometry

In stereoscopic image creation, the main difficulty arises while controlling the stereoscopic camera parameters. There are two principal parameters to control disparity: *interaxial separation* (t_c) and *convergence distance* (Z_c). Disparity is used to gather absolute depth information of the observed scene. Therefore, proper interplay of interaxial separation with convergence distance is an important process in order to create realistic 3D percept.

When the viewer is looking to an object or a surrounding field, left and right eyes do not see exactly the same view due to the fact that left and right eyes view the world from slightly different angles. The difference between two eyes is called interocular distance or eye separation. This separation generates different left and right retinal images which hold views captured by two eyes. Binocular disparity is the difference between these two retinal images, forming binocular vision. In stereoscopic systems two cameras are placed at slightly different positions from each other horizontally. These cameras are used to represent left and right eyes. The distance between two cameras is called as interaxial separation which corresponds to interocular distance in the HVS.

Convergence and divergence constitute the vergence notion. This notion is the synchronical movement of two eyes in physiology. Convergence represents movement of two eyes rotating towards each other when eyes are focused on a close object; whereas, divergence represents movement of two eyes rotating away from each other when eyes are focused on a farther object. Since both convergence and divergence define the rotating movement of the eyes, convergence is used solely in the literature in order to reduce terms. Similarly, convergence distance corresponds to the distance between the plane or object in focus and the middle point between two cameras in stereoscopic applications. Convergence distance in stereoscopic applications replicate the vergence effect in HVS.

In HVS, interocular distance and vergence movements generate retinal images. Similarly in stereoscopic systems, interaxial separation and convergence distance generate disparities, or screen parallaxes. A virtual environment that is captured
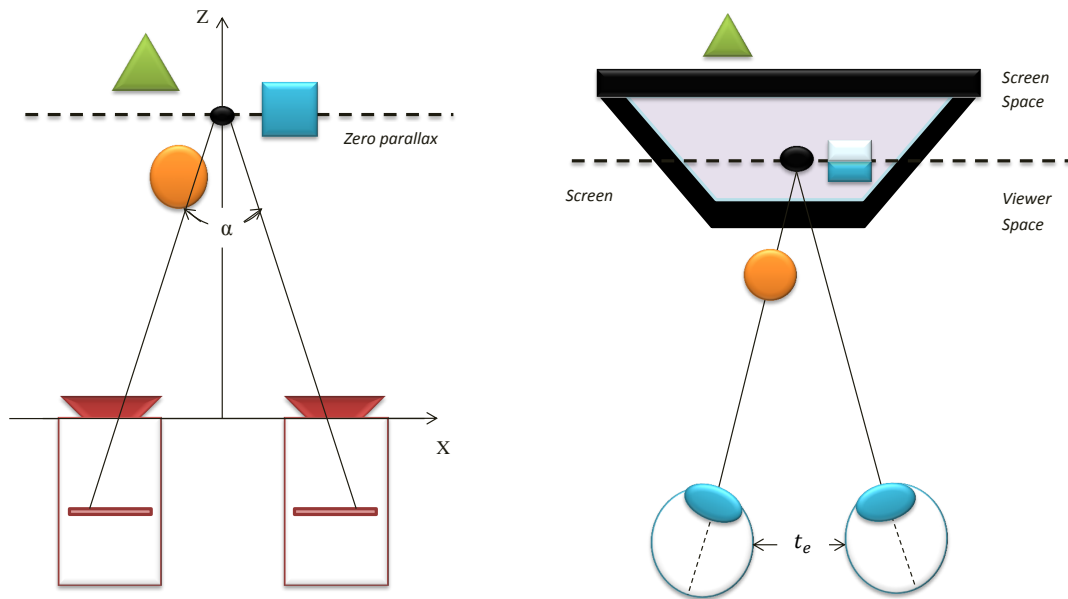
Figure 2.1: Employment of two cameras in a virtual space at the top, corresponding screen space at the bottom

with two cameras and corresponding 3D view on a display screen is given in Figure 2.1.

There are two setup types for converging cameras explained as follows:

- **_Toed-in setup:_** Two cameras are rotated inward towards a plane or object in focus. This approach adapts the convergence mechanism of HVS literally.

- **_Parallel sensor-shifted setup:_** The rotation of two cameras remain still and cameras stand in parallel as their view directions are in parallel too. Image shift is used in the camera sensors to replicate the resulting disparity if cameras were actually rotated.

Parallel sensor-shifted setup is preferable instead of toed-in setup in stereoscopic

systems especially for virtual environments. Although toed-in seems to be a more natural way since convergence mechanism works alike HVS, parallel approach produces stereoscopic images with higher qualities and less artifacts. The underlying reason that toed-in is an approach with stereoscopic impairments is *Keystone distortion.* The positioning of left and right cameras at an angle toward each other causes to capture slightly different image planes. This condition brings about the problem of capturing a trapezoid-like image in opposite directions by two cameras. Scene part closer to the left camera looks larger on the right part of the screen surface; whereas, scene part closer to the right camera looks larger on the left part of the screen surface. This situation induce to have incorrect vertical parallax, which is one of the dominant factors of visual discomforts like eye-strain. Since both left and right cameras are directed toward the same image plane, parallel camera configuration does not suffer from Keystone distortion and only generates the desired horizontal parallax.

Figure 2.2 is an illustration of the relation among interaxial separation and convergence distance geometrically. Given this parallel sensor-shifted camera setup geometry, two equations are extracted by using similar triangles:

$$\frac{t_c}{2(h - \frac{d}{2})} = \frac{Z_v}{f} \tag{2.1}$$

$$\frac{t_c}{2h} = \frac{Z_c}{f} \tag{2.2}$$

These two equations are employed to obtain the *disparity* of an object, located at a distance Z_v away from two cameras, depends on interaxial separation (t_c) and convergence distance (Z_c), and is given as:

$$d = f t_c (\frac{1}{Z_c} - \frac{1}{Z_v}) \tag{2.3}$$

The distance between the projection of a 3D point on the one camera's image plane and the projection of the intersection point of the two camera viewing
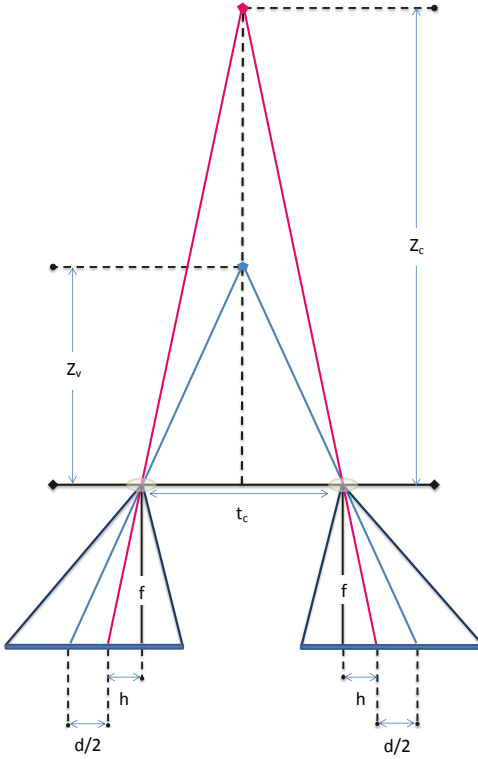
Figure 2.2: Parallel sensor-shifted camera setup configuration in a virtual environment

directions on the same image plane constructs one pair of the disparity, while the other pair comes from the other camera's image plane. In this equation, d represents this disparity. Focal length of the cameras is denoted by f. The vision of a 3D point in a real world or a virtual environment on the camera sensor's image plane is projected on f in toed-in configuration; however, this case is not same for parallel setups. The projection shifts by h on the image plane, that is why the parallel camera setup is entitled as sensor-shifted.

There is a correlation between disparity and parallax notions. Since disparity represents a distance on image plane of the cameras, it is called as image disparity either. Parallax represents the difference in the produced left and right views on the screen plane. The conversion from image disparity d to screen parallax p

simply requires scaling the image disparity from image sensor metric to display size metric, by multiplying it with a scale factor W_s/W_i, where W_i and W_s denote the image sensor width and screen width respectively.

$$p = d(W_s/W_i) \tag{2.4}$$

While maintaining stereoscopic depth, the viewer reconstructs a 3D environment around the display screen. This constructed 3D environment involves objects that actually appear on the display screen but perceived as they stand in front or behind the screen. The distance, how much further away each object appears than the display screen, is determined by each object's corresponding parallax values. The distance of this perceived point between the viewer is Z, while the distance between the viewer and physical display screen is the viewing distance Z_d. The correlation between Z_d and Z is given as:

$$Z = \frac{Z_d t_e}{t_e - p} = \frac{Z_d t_e}{t_e - d(W_s/W_i)} \tag{2.5}$$

where p is parallax and t_e is the human interocular distance, and the physiologically average of interocular distance is approximately 65 mm.

The perceived depth, generated around the display screen, is affected by the *type* of the parallax as well as the *amount* of the parallax. Amount determines the distance between the appeared position of the reconstructed object and the display screen; while, type determines the region of the appeared position. Regions are divided into three, in the light of following cases: viewer space includes positions in front of the screen, screen space includes positions behind the screen, and positions are located on the screen as it is illustrated in Figure 2.3.

- ***Zero parallax:*** On the plane at convergence distance the retinal positions of objects appear at the same point which results, in turn, they appear at the physical screen surface ($Z = Z_c$). This condition is called *zero parallax setting*. Two conditions occur when object distances Z are different from Z_c.
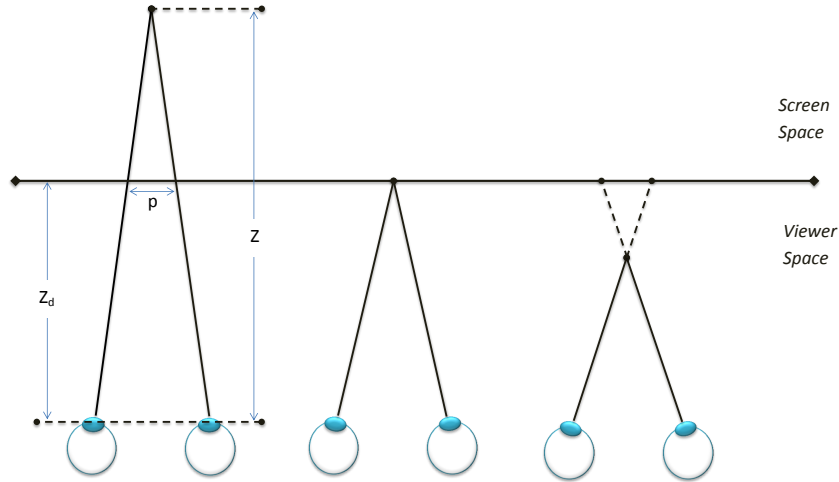
Figure 2.3: Positive, zero, and negative parallaxes for screen space respectively.

- **Positive parallax:** In this case, $(Z > Z_c)$, the object appears inside the screen space, which is the condition that objects appears behind the display screen. When this condition occurs, the object has a positive disparity, or screen parallax.

- **Negative parallax:** On the other hand, in the case $(Z < Z_c)$, the object has a negative disparity, or parallax. These objects appear as if they are physically located in front of the screen.

Physiological experiments have proven that the human visual system has more tolerance to positive parallax than negative parallax [3]. However, the human visual system is still limited to comfortably perceive all objects which appear in positive or negative parallax regions. It has been shown that locating the scene in a limited area around the screen surface gives more reasonable results for avoiding accommodation-convergence conflicts.

The perceptual effects of the stereoscopic camera parameters are summarised in the Table 2.1. Interaxial separation (t_c) directly affects the disparity and eventually the amount of depth perceived in the final image. The convergence

Table 2.1: The review of the perceptual effects of stereo parameters (adapted from Milgram and Kruger [1])

Parameter	Increase/Decrease	Parallax	Perceived Depth	Object Size
t_c	Increase	Increase	Increase	Constant
	Decrease	Decrease	Decrease	Constant
z_c	Increase	Decrease	Shift forward	Constant
	Increase	Increase	Shift backward	Constant

distance, on the other hand, does not affect the overall perceived depth, but effects objects' individual perceived depths.

2.3 Accommodation and Convergence Conflict

Accommodation and convergence are two important oculomotor cues which have a big role on binocular viewing after binocular disparity. Accommodation refers to the eye lens activity when eyes are fixated at a point or region and driven by a monocular cue that is retinal blur. The object or area is observed sharper; whereas, remaining regions look smoother as if blur effect is applied. This occasion enables HVS not to process details and insignificant parts of the scene. Convergence denotes the rotation of two eyes towards each other when eyes are focused at a point or region. Both cues are used in conjunction with each other. They are triggered by looking to same specific location, HVS operates such that eyes converge to and accommodate at the same point. Nevertheless, replicated stereoscopic vision is in contrast to vision in real world. The working principle of stereoscopic displays is based on providing an amount of perceived depth around the display screen. This means, the scene is located on the display screen physically; however, scene elements are visualized around the display screen. As a result, the conflict is caused by the fact that when looking at the stereoscopic 3D display, viewer's eyes are accommodated on the display plane, while they are

forced to converge towards scene elements on the perceived depth Z_c at a distance of the display.
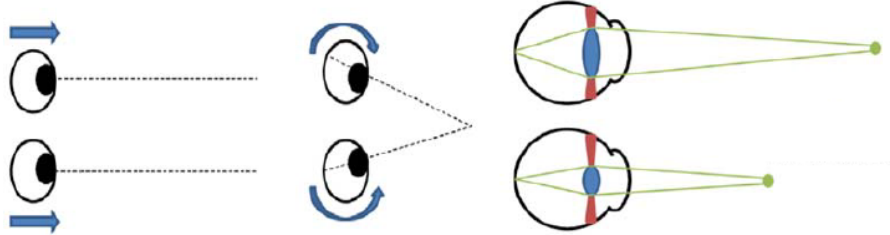
Figure 2.4: Convergence and accommodation

The discrepancy between focused positions causes an undesirable phenomena, so-called accommodation and convergence conflict, happens for all planostereo-scopic displays, i.e. displays where the views are presented on a planar screen. There is a threshold for a relaxing configuration for HVS to bear this discrepancy between accommodation and convergence. If threshold is exceeded, the viewer gradually suffers from eye-strain, visual fatigue, and diplopia. This threshold varies for everyone and investigated under stereoscopic comfort zone. There are several earlier studies on the issue of stereoscopic comfort zone. The conclusion pointed out by these studies is that the amount of perceived depth in stereo-scopic displays should be limited; and the conflicts related to accommodation and convergence should be controlled.

Chapter 3

Related Work

3D notion has recently gained importance and a number of techniques have been proposed for 3D camera systems for real environments, stereoscopic post-production pipeline and editing of stereoscopic images, also stereoscopy adjustment in virtual environments, which are presented in this section respectively, while the second section summarises a large body of studies on camera control for virtual environments.

3.1   Stereoscopy Production Studies

Rapid development in technology and industry revived 3D production which became popular again after almost fifty years. This current renaissance, as called in 3D literature, aroused 3D production based research. There is considerable amount of researches found in the field of stereo creation and these are represented under three main subsections.

3.1.1 3D Camera Systems and Stereo Acquisition

The conventional way for capturing real scenes in 3D is done by using two physical camera equipments. Relative positions of two cameras and their lens settings are important to produce good stereo content. One of the recent approaches which focus on production of high quality stereoscopic content capture is presented by Zilly et al.[4] as a software system. Their system, called as Stereoscopic Analyzer, is a 3D production tool for stereo shooting by assisting stereographers and camera teams in real environments. Video streams are used to compute disparities by correcting deficiencies such as camera misalignments and keystone distortions. Their system analyses depth structure of the captured scene and proposes proper suggestions for stereo camera parameters, also provides to adjust camera calibration manually.

Heinzle et al. [5] develop a computational stereo camera system for controlling physical camera and rig properties automatically with a control loop that comprise capture and analysis of 3D stereoscopic parameters. They propose their system as being a combinable design for existing stereo camera rigs. The system architecture includes configurable unit which performs scene analysis in real time and programmable unit to utilize different algorithms for different scene and shot properties.

3.1.2 Stereoscopic editing on still images

Recent work on stereoscopic image editing focuses on correction of imperfect stereoscopic images and videos. Koppal et al. [6] present an editor for live stereoscopic shots. They concentrate on the viewer's experience and transform desired visual experience settings into camera parameters. As a previewing step, an estimation of the viewers' 3D perception is predicted from robustly obtained scene videos or still images. Replanning of the shot is done by using new camera parameters that are procured from editing tool if the predicted perceived effect is found as incorrect or insufficient by the user.

Lang et al. [7] focus on the problem of remapping the disparity range of stereoscopic images and video. Perceptual aspects of stereo vision are formalized into disparity mapping operators which control and retarget depth range in the produced stereoscopic images and videos to different displays and viewing conditions in a nonlinear way. These operators are implemented based on steroscopic warping strategy. A sparse set of stereo correspondences, presented algorithm computes disparity and image-based saliency estimates, and uses them to compute a deformation of the input views so as to meet the target disparities.

Didyk et al. [8] have recently proposed a disparity model that estimates the perceived disparity change in processed stereoscopic images to control distortions and make enhancements. They perform psychophysical experiments in order to derive a metric for modelling disparity. Their study also presents a backward compatible stereo application that produces images which looks ordinary; moreover, if required equipments are used depth illusion occurs. Didyk et al. [9] also extend their disparity model by considering luminance effect on the perception of disparity. In their work, they presented disparity retargeting as one of its applications.

3.1.3 Stereo parameter adjustment in virtual environments

Post processing and image shifting methods are used for retargeting disparity in offline applications such as digital cinema and 3D content retargeting. On the other hand, interactive applications require real-time techniques. Among recent works, Jones et al. [10] propose a geometrical framework for real-time stereoscopic camera parameters calculation by providing a transformation between camera space and screen space in order to map specified depth range of the scene to perceived one. They also ensure that no depth distortion occurs with viewer movements while using head tracked displays. Their model is employed for generating still images, digital photography, and real time computer graphics.

Oskam et al. [11] present a controller for real-time applications which produces

a final disparity value for the viewed frame by calculating camera convergence and interaxial separation, while scene depth is assigned to a desirable depth by using control points. Stereoscopic camera parameters change automatically by taking minimum and maximum scene depth values into account in order to handle excessive binocular disparities. Since unpredictable object or viewer motion changes the depth of the scene instantly, a temporal constraint interpolation phase is performed to avoid sudden depth jumps which result in uncomfortable stereoscopic perception.

3.2 Camera Control Studies

The viewer's experience of a 3D environment is highly correlated with the success of the presentation of the scene. The camera motion, position, and orientation and their conjunction with scene elements are used to present a scene. There are several studies which address the camera control issue in different fields such as data visualization, 3D games, and virtual walk-throughs. In addition to virtual environments, camera control techniques are employed for real world camera systems especially in robotics.

3.2.1 Path Planning and Scene Exploration

Knowledge of the environment is used to assist users in order to make them explore the environment or navigate in the environment, classified under two parts based on local or global awareness. The aim is observing scene objects by determining important viewpoints around them while maintaining occlusion free camera paths in object-based assistance systems. Navigation and exploration in the environment establish the framework of environment based assistance. Robotics based approaches and path planning techniques are used for navigation and exploration tasks. These techniques are analysed under potential fields, cell decomposition, and roadmaps. Potential fields, a sub topic of theoretical physics, use the same principle of charged particle interactions in electrostatic

fields. Similarly obstacles and the camera is put in charged particles positions. Khatib [12] proposes a solution is based on steepest descent algorithm. Low cost is an advantage of potential fields technique that provides usability in real time aplications; however, management of local minima causes problems for highly dynamic environments. Cell decomposition is a technique that divides environment into smaller regions as cells and builds a network between these regions. Roadmaps specify candidate configurations and connect consecutive ones with a graph search algorithm.

Salomon et al. [13] describe an approach for navigating avatars in complex environment based on a variant of the probabilistic roadmap planning algorithm. Their algorithm searches roadmap graph for a path between two points by performing path smoothing and collision detection via bounding volumes. Nieuwenhuisen et al. [14] exploit probabilistic roadmap method in the pre-process step in order to compute a path through the environment. Resulting path is improved by using circular blends between edges, parabolic blends, Beziers, or clothoids may be used as alternatives.

3.2.2 Cinematographic Practice in Camera Control

Cinematography provides guidelines for how the camera should be moved and positioned. Scene descriptions, camera angles, shot types, and camera movement types compose principles of cinematography. In order to implement a camera system by using cinematographic principles, the system must know the layout of the scene, principal characters, important objects, while principles must be encoded in the system as well.

Kneafsey and McCabe [15] summarize existing studies on camera control through cinematographic principles. They classify techniques by approaching simply positioning and orienting the camera within virtual world for still images, for shots with a moving camera e.g. for museum walkthroughs, for following moving subjects.

3D computer graphics applications observe the scene from a particular characters point-of-view or from a small set of prespecified viewpoints. Camera placement by cinematic rules is generally ignored. The approach in the study of Christianson et al. [16] extends camera placement approach by applying cinematic principles; therefore, it benefits from storytelling capabilities. They describe several cinematography principles and then formalize them into declarative language which is used to adapt cinematography principles in computer graphics applications.

Chapter 4

Automatic Adjustment of Stereoscopic Parameters

In this part of our camera control mechanism, we propose a novel method for adjustment of stereoscopic camera parameters, interaxial separation and convergence distance, in order to improve viewer comfort during 3D experience. We have tested our system in order to gauge the effectiveness of our approach by comparing with existing methodologies.

4.1 General Architecture

Our method exploits parallel sensor-shifted setup instead of toed-in setup for disparity calculation due to stereoscopic impairments explained in Chapter 2. We enhance this geometrical framework by utilizing stereoscopic comfort zone principles and incorporating importance of scene elements. A number of researches address disparity control problem by correcting disparity on captured images in the post production pipeline. However, we approach this issue for interactive environments where the position of the camera is dynamically changing. We render the environment by employing two virtual cameras for real-time disparity range adoption. Figure 4.1 shows an overview of the proposed method.
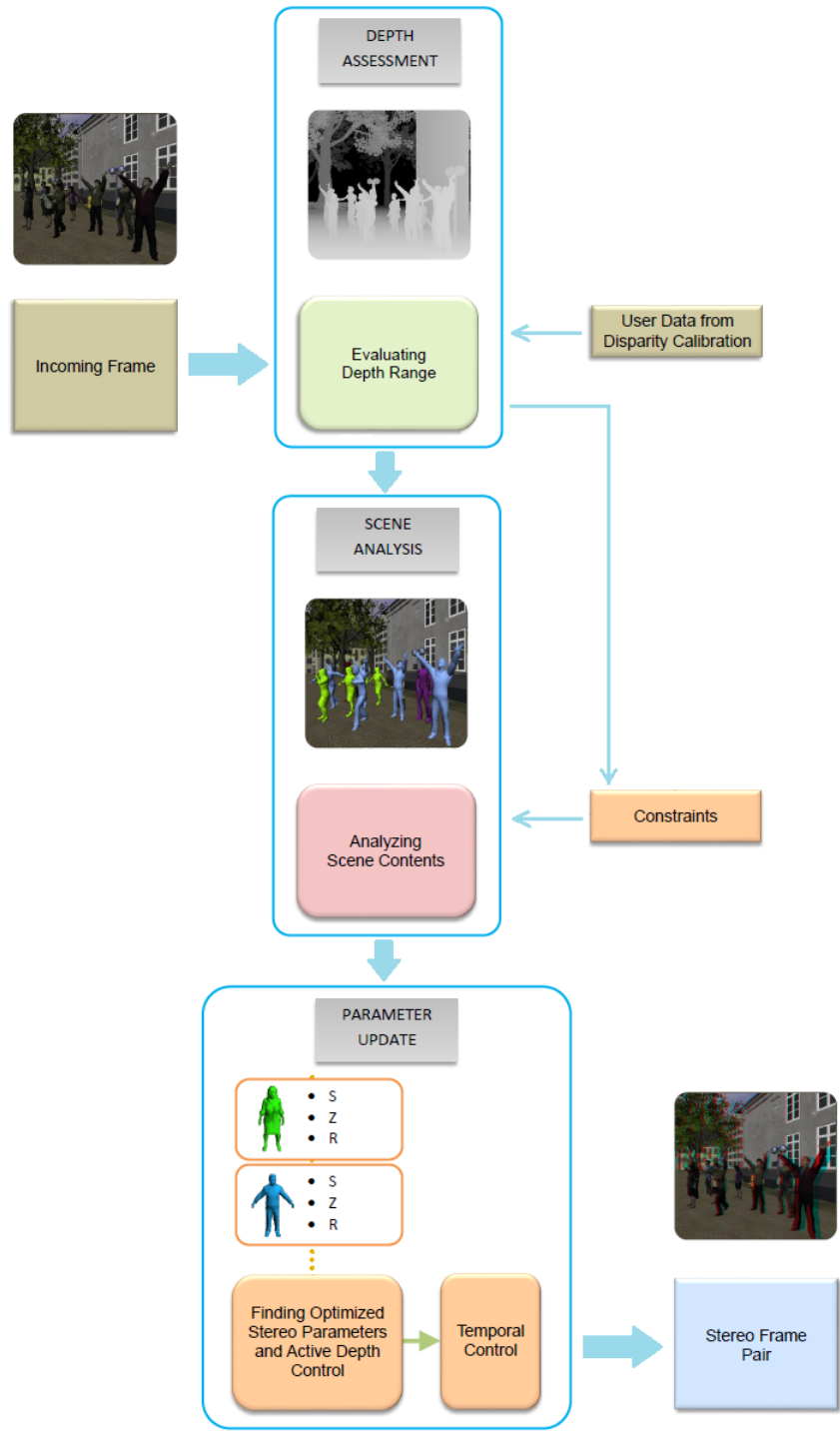
Figure 4.1: An overview of our methodology

Our proposed stereo rendering method consists of four main stages. The first stage applies a disparity calibration phase, where the depth range extrema that the viewer is able to perceive is found. A depth assessment process is applied in the second stage in order to calculate scene depth. Scene elements are analysed in the third stage in order to extract attention-grabbing objects and these objects' corresponding significance scores, locations in the virtual environment, also positions on the display surface. Finally, stereo parameters are calculated through an optimization phase that is performed according to our two assumptions for comfortable and effective 3D experience. Total screen disparity is aimed to be maximized and convergence distance is aimed to be located to nearer to the most attention-grabbing objects.

4.2 Depth Range Control

The most naive approach for stereoscopic rendering is based on assignment of fixed values for interaxial separation and convergence distance. This is an expedient solution, since it may provoke excessive disparities, also deprives updating parameters continuously. A control facility for perceived depth range around the screen display is required in order to make scene elements appear within the stereoscopic comfort zone. This control mechanism enables to map a specific range of scene distances to a perceived depth range by updating parameters for changing scene contents.

Several studies, like the model of Jones et al. [10] and the model of Guttmann et al. [17], make use of depth range control approach. Depth range control employs geometric formulation of stereo vision, which are presented comprehensively in Chapter 3. Oskam et al. [11] propose that a series of points in the scene can be mapped onto a series of points in the target space by using Equation 4.1 which is obtained by the conjunction of similar triangles in 3D Display and Camera Geometry.

$$fbc_i - fbc_{cvg} - c_i d_i c_{cvg} = 0 \, for \, i = 0, 1, ..., n \tag{4.1}$$

where f is focal length, d is image disparity, b and c_{cvg} stands for interaxial separation and convergence distance. If we utilize this equation for two constraints, which stand for minimum and maximum distances of the scene, then we obtain the following equations 4.2 and 4.3 to calculate convergence distance and interaxial separation.

$$Z_c = \frac{Z_{max} Z_{min}(d_{max} - d_{min})}{(Z_{max} d_{max} - Z_{min} d_{min})} \tag{4.2}$$

$$t_c = \frac{Z_{max} Z_{min}(d_{max} - d_{min})}{f(Z_{max} - Z_{min})} \tag{4.3}$$

where Z_{max} is the distance between the camera and the farthest visible scene element, Z_{min} is the distance between the camera and the nearest visible scene element, d_{max} is maximum disparity value of the farthest object, and d_{min} is minimum disparity of the nearest object. We obtain Z_c, the distance between zero parallax plane and viewpoint plane, and t_c, the separation between two virtual cameras.

4.3   Attention-Aware Disparity Control

In order to improve viewer comfort in 3D experience, significant scene elements should appear within the stereoscopic comfort zone of viewers. In other words, scene elements should be located nearer to the convergence plane; consequently, they appear in regions closer to the display screen. Stereoscopic comfort zone is illustrated in Figure 4.2. However, scene contents cannot be rearranged and objects cannot be relocated in pre-produced scenes. Consequently, convergence distance should be adjusted by maintaining the total disparity as high as possible.
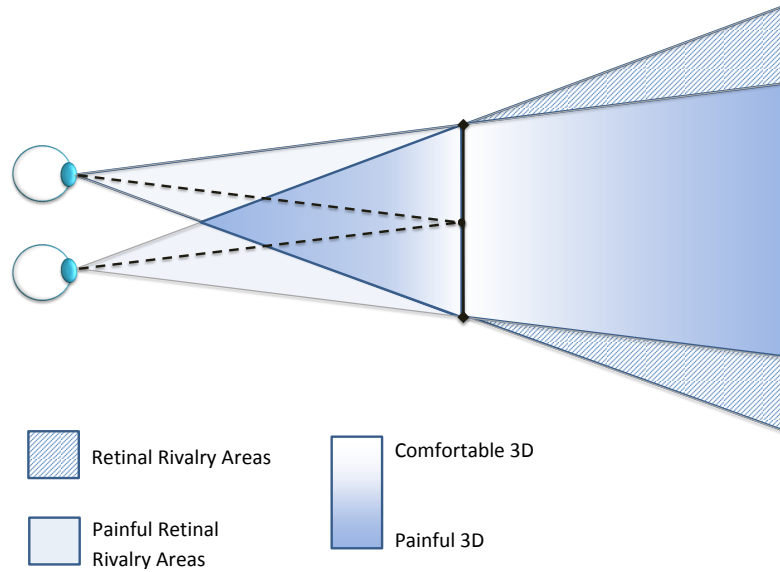
Figure 4.2: The stereoscopic comfort zone

4.3.1 Viewer-Based Disparity Calibration

Perceptual experiments indicate that in stereoscopic systems, the same disparity range creates different visual feedback for different users, due to the fact that stereoscopic comfort zone limits change for each person. There is a significant variation in the physiological capabilities of each people. A content may present a comfortable 3D experience to a viewer, while the same content with the same disparity range may cause eyestrain to another viewer. This fact brings about the need for a user-adaptive control in stereoscopic systems.

Some stereoscopic products, especially 3D games, make use of individual control over depth and let the viewer adjust disparity while displaying 3D contents. It is not an ideal solution to provide proper amount of disparities for that viewer. The viewer may adjust depth range so high in order to generate depth-rich contents, which results in excessive disparities and uncomfortable experience. Conversely, the viewer may keep disparity range lower than it is expected in order to avoid visual fatigue, which decreases depth illusion. We perform this disparity
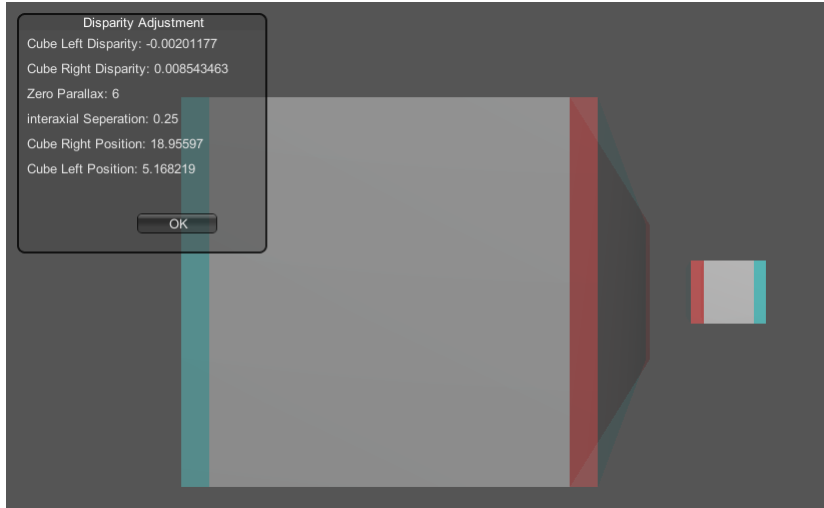
Figure 4.3: A screenshot from disparity calibration stage

calibration stage in order to detect the viewer's perceived depth limits and provide a 3D experience where scene elements appear within the stereoscopic comfort zone of that viewer.

Disparity calibration stage of our system is shown in Figure 4.3. The scene content composes of only two elements, two side-by-side cubes with zero parallax setting. The viewer moves one of the objects in the forward direction, where the object appears in front of the display surface in order to find the maximum disparity limit for positive parallax. If the viewer is not able to fuse two distinct images on the screen for two eyes, then corresponding disparity to this position is assigned as the positive parallax limit for this viewer. Similarly, the same procedure is repeated by moving the other object in the backward direction. If the viewer loses 3D effect and observes the object like a 2D still image, then this corresponds to the maximum disparity limit for negative parallax.

We believe that a simple scene structure rather than a complex environment is more suitable for finding limits of perceived depth range. When the viewer is looking for the maximum disparity limits, his/her focus is at one object and corresponding disparity amount. Remaining scene objects in a complex environment confuses the viewer, since they appear in front of or behind the focus object and have disparities over limits.

4.3.2 Scene Depth Calculation

When mapping of scene elements into a target depth space is the case, also that is the motivation of our research, the virtual world distances have a direct effect on generated disparity value as it is explained in Section 2. Therefore, correct extraction of the minimum and maximum distances of the furthest and closest point of the scene is an important process that should be done rigorously. Depth buffer is used for calculation of these distances.

Using Depth Buffer

The scene content gives us the location information about closest and furthest scene elements; however, using depth buffer provides a better solution in order to gather visible scene depth extrema. All objects may not seen by the camera, they may be occluded by other objects if the scene depth range is too high. In this case, the distance between the furthest element and the camera is assigned for maximum depth distance of the scene; however, depth range of visible scene is lower. This case leads to low disparity range for the visible scene content. In order to avoid this kind of an issue, depth buffer is used to gather depth information for each frame.

Depth buffer transforms the z distance of each pixel's corresponding 3D point between *zNear*, near clipping plane and *zFar*, far clipping plane of the camera in a non linear way, then stores this transformed value in the buffer. This value is in the range of [0, 1], where 0 corresponds to the *zNear*, 1 corresponds to the *zFar*, and remaining values' corresponding 3D positions are distributed in a non linear way between *zNear* and *zFar*. Depth buffer representation of a scene is given in Figure 4.4.

Eq. 4.4 gives the relation between depth buffer value and corresponding z distance of a point.

$$Z = \frac{\frac{Z_{far}}{Z_{far}-Z_{near}} + \frac{Z_{far}Z_{near}}{Z_{near}-Z_{far}}}{Z_{buffer}} \tag{4.4}$$

Figure 4.4: A grey-scale output of a sample view rendered by corresponding depth buffer values in pixels

Min Max Reduction

Depth buffer provides to gather maximum and minimum distances of the visible scene; however, extraction of this information is a costly process. It requires a search operation, in which comparison of each pixel value with minimum and maximum values is performed for every frame. Therefore, we take the advantage of parallel processing feature of the GPU, in order to efficiently obtain minimum and maximum depths in the scene in real-time applications.

Reduction operation on GPU provides to adjust sizes of input and output textures. In our case, we search for minimum and maximum values among all pixel values from the captured still image of the visible scene. This captured image is given as an input texture to reduction process, then parallel mechanism of GPU comes into play. Texture is divided into 2x2 sample blocks and local maximum and minimum of each 2x2 group of pixel values is designated in the parallel manner. After values are determined, input texture of size M by M is reduced to a texture of size M/2 by M/2. This procedure is repeated until the size of output texture becomes 1 by 1. This output texture stores minimum and maximum values. A simple illustration of min max reduction is shown in Figure 4.5. Greßet al.[18] present a GPU-based collision detection method which is an example research that utilizes GPU for reduction process.
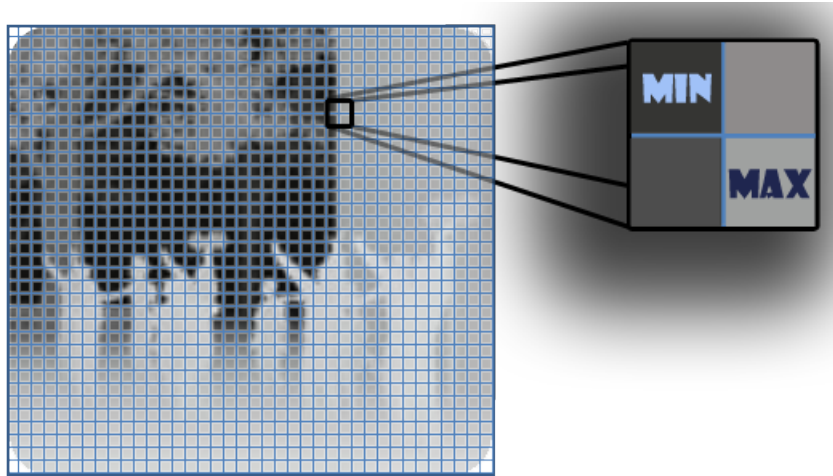
Figure 4.5: Min max reduction

4.3.3 Analysis of Scene Elements

Our motivation for generating a camera control system relies on presentation of attention-grabbing scene elements comfortably and realistically. Therefore, it is an important task to characterise significances of scene elements. There are three features of a scene element we need to gather.

Significance score is the most prominent feature of a scene element. This score indicates the importance degree of scene elements. In our system, application developer or scene author assign these scores after scene content is prepared.

Forward distance is the distance between the scene element and camera. We need forward distance values since we modify convergence distance in accordance with this distance.

Radial distance is the distance between the scene element and forward camera axis. If a scene element draws attention, the viewer prefers to watch this element closer and tries to position it onto the center of the display.

We need to perform an analysis of scene elements in order to detect important scene attributes and gather these three features. A sample pseudocode for analysis phase Algorithm 1 is given below, where S stands for significance score,
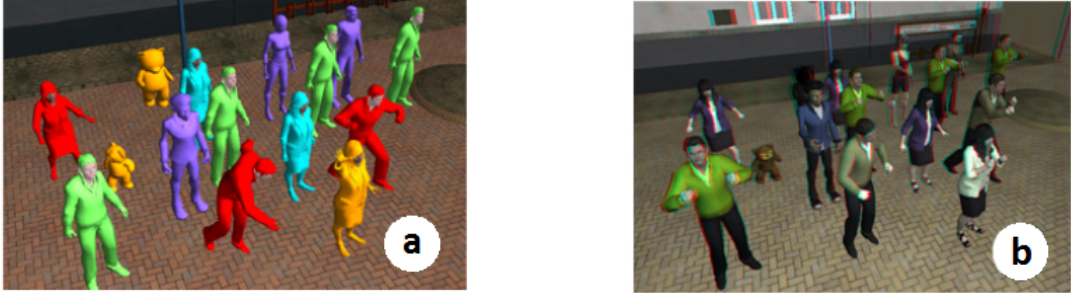
30

Figure 4.6: (a) Analysis of scene elements based on their significance scores. (b) Corresponding view of the scene.

Z for forward distance, and R for radial distance.

Algorithm 1 Scene content analysis algorithm

1: $e[\,] \leftarrow getSignificantElements()$
2: $\quad\quad$ ▷ Acquiring all significance score assigned elements in the current scene
3: $j \leftarrow 0$
4: **for** $\forall e[i]$ **do**
5: $\quad$ **if** $e[i]$ is visible in the current frame **then**
6: $\quad\quad$ $e[i].Z \leftarrow ForwardDistanceFromCamera()$
7: $\quad\quad$ **if** $e[i].Z \leq D_{max}$ **then**
8: $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ ▷ D_{max}: maximum forward distance allowed
9: $\quad\quad\quad$ $o[j] \leftarrow e[i]$
10: $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ ▷ implies $o[j].S \leftarrow e[i].S$ and $o[j].Z \leftarrow e[i].Z$
11: $\quad\quad\quad$ $o[j].R \leftarrow RadialDistanceFromCameraAxis()$
12: $\quad\quad\quad$ $j \leftarrow j + 1$
13: $\quad\quad$ **end if**
14: $\quad$ **end if**
15: **end for**
16: **return** $o[\,]$

4.3.4 Disparity Production

Required geometric formulations, that are employed in our system, are explained in the previous sections. However, we believe that disparity production issue should not approached from geometric aspect only. For a more perceptual approach, there is a need for a control mechanism that optimizes calculated camera parameters in accordance with two assumptions.

- Convergence plane should tend to be nearer to both important scene elements and elements with lower radial distances.

- Total scene disparity should be maximized.

The center of attention represents scene parts, where the viewer focuses on longer than remaining scene elements. Attention-grabbing objects or environments are positions that viewers focus on, also viewers tend to look toward the center of the display device. A comfortable presentation is required for the center of attention. The first assumption stands for locating scene elements in the center of attention nearer to zero parallax state, which results in minimization of visual artifacts i.e. ghosting effect for these objects. For a realistic one, the second assumption enables to compensate disparity which are decreased for our first assumption.

We first formulate an energy term $E_o(Z_c)$ in order to move convergence plane towards scene elements with higher significance scores and with relatively less radial distances from the forward axis of virtual camera.

$$E_o(Z_c, t_c) = \sum_{i=1}^{n} \frac{S_i}{R_i^2} \left(Z_i - Z_c \right)^2, \tag{4.5}$$

where n is the number of significant scene elements found in the scene analysis stage.

Eq. 2.3 is employed in order to define a second energy term $E_d(Z_c, t_c)$ which maximizes total scene disparity, consequently total perceived depth is maximized as well.

$$E_d(Z_c, t_c) = \sum_{i=1}^{n} S_i f t_c \left(\frac{1}{Z_c} - \frac{1}{Z_i} \right), \tag{4.6}$$

Our objective function $E(Z_c, t_c)$ is a combination of these two energy terms and presented in the following Eq. 4.7. In our case, optimization problem consists of the minimization of $E_o(Z_c)$ which aims to compute a value as close as to the center of attention and the maximization of $E_d(Z_c, t_c)$ which pursues to obtain

a value as much as for a larger perceived depth range. Therefore, the system searches for the optimal parameter set by minimizing $E(Z_c, t_c)$.

$$E(Z_c, t_c) = \hat{E}_o(Z_c) - \hat{E}_d(Z_c, t_c), \tag{4.7}$$

where $\hat{E}_o(Z_c)$ and $\hat{E}_d(Z_c, t_c)$ are the normalized energies of $E_o(Z_c)$ and $E_d(Z_c, t_c)$ s.t.

$$\hat{E}_o(Z_c) = E_o(Z_c) / (Z_{max} - Z_{min})^2, \tag{4.8}$$

$$\hat{E}_d(Z_c, t_c) = E_d(Z_c, t_c) / (d_{max} - d_{min}), \tag{4.9}$$

Normalization process for $E_o(Z_c)$ and $E_d(Z_c, t_c)$ is required in order to make our methodology applicable to any given scene content with different depth ranges and viewers with different stereoscopic comfort zone limits.

There are two constraints, which are d_{max} and d_{min}, employed during the minimization of $E(Z_c, t_c)$ in order to ensure that resulting optimized parameters will not produce a disparity value that exceeds upper or lower bounds of the viewer's comfort zone which are specified in the disparity calibration phase.

$$d_{max} \geq ft_c \left(\frac{1}{Z_c} - \frac{1}{Z_i} \right) \geq d_{min}, \ \forall i \,|\, 1 \leq i \leq n, \tag{4.10}$$

This nonlinear system is solved by utilizing improved stochastic ranking evolution strategy (ISRES) algorithm [19] in NLOpt library [20]. ISRES algorithm is based on a simple evolution strategy augmented with a stochastic ranking that decides by carrying out a comparison, which utilizes either the function value or the constraint violation. The optimization process results in interactive speed, that enables to update stereo camera parameters dynamically by employing this process for each frame. There are two cases that our system switches from optimization phase to depth range control (DRC) method and these situations are

indicated below.

- If only a single element is within the center of attention in a frame, then the system detects only one element that has a significance score. In this case, the system locates convergence plane on this scene element i.e. $Z = Z_c$ and computes the other parameter interaxial separation by using DRC method.

- If no scene element, which has an assigned significant score, is visible in a frame, then importance notion cannot be employed in this case. The system computes stereo camera parameters by DRC method in these frames.

Temporal Control: Our system considers snapshots in time for calculation of stereo camera parameters; therefore, resulting disparities are found for each frame. Since the scene depth changes from time $t - 1$ to t, a discontinuity may cause a large variance between corresponding disparity values d_{t-1} and d_t if there is an instant scene depth change is observed. This situation results in undesired visual artifacts and excessive disparities. Therefore, the system controls optimized parameters over time and produces final ones through a threshold function $f(\cdot)$ that is presented in Eq. 4.11 in order to provide temporal coherence and avoid instant depth jumps.

$$f\left(x(t)\right) = \begin{cases} x(t-1) + x_1, & \text{if } x(t) - x(t-1) \leq x_1; \\ x(t-1) + x_2, & \text{if } x(t) - x(t-1) \geq x_2; \\ x(t-1) + k\left(x(t) - x(t-1)\right), & \text{otherwise.} \end{cases} \quad (4.11)$$

where k is chosen to be $0 < k < 1$.

Chapter 5

Scene Exploration

We define stereoscopic camera control issue as a two-part process. Since our camera control mechanism addresses producing a comfortable and realistic 3D experience, our main motivation is automatic calculation of stereoscopic camera parameters which is explained comprehensively in the previous section. However, we believe that camera control is not only a parameter calculation process and a camera control system should also include a mechanism for scene exploration task. With this motivation, we extend our system by presenting a model for path generation.

The entire system is proposed in order to generate a perceptually driven camera control mechanism which makes use of HVS and perception principles. In the first phase, this feature is derived from important scene elements and their assigned significance scores which are employed in the optimization phase of automatic adjustment of parameters. We aim to utilize significance characteristics of scene elements in the second phase while exploring scene environment. In order to offer this kind of a model for scene exploration task, we utilize *saliency* concept that is used for finding attention-grabbing regions of 3D models. Three main parts constitute the skeleton of scene exploration mechanism: viewpoint selection, path generation and camera transformation. The flow of the system is given in the Figure 5.1.
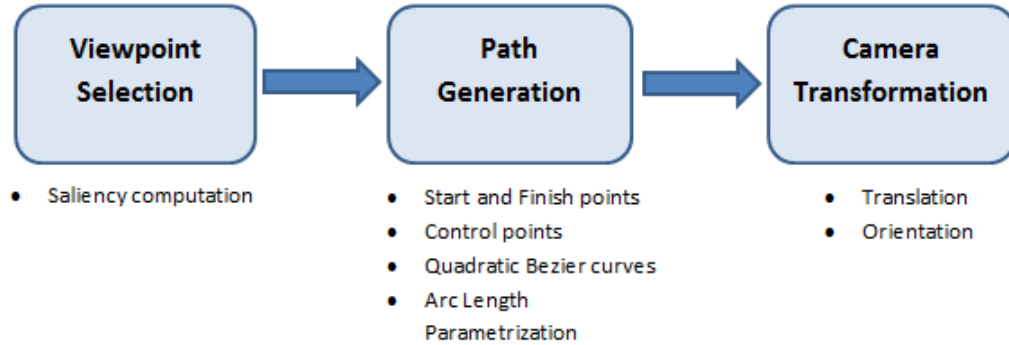
Figure 5.1: A step-by-step working principle of scene exploration mechanism.

Viewpoint selection part deals with importance regions of scene elements, while path generation part focuses on modelling of the path. The last phase executes camera motion in the light of path characteristics obtained from previous phases. A step-by-step working principle of our path generation mechanism is as follows:

- Saliency values of each important scene elements are calculated,

- Start and finish positions are determined,

- Control points are specified,

- Quadratic Bézier curves are fitted between specified control points and linked with each other,

- Positions on Bézier curves are parametrized,

- Each point on Bézier curves corresponds to the camera position for each frame.

- Camera orientation towards the important objects is executed.

5.1 Viewpoint Selection Using Saliency

The scene content has a significant role for exploring task in our approach, since our main motivation is based on producing a path which presents important scene elements rather than other regions of the scene content. Our assumption relies on the fact that important scene elements' corresponding significant scores are assigned directly proportional with their attention-grabbing degree. We want the viewer to observe the scene by moving around important objects, in this way we can form our path model into an attention-aware structure. The need of a viewpoint selection technique comes into play at this point, to determine a position around each important object. Underlying idea is constructing a path between these positions which ensures to view important scene elements.

Viewpoint selection is not only used in the field of path planning, but also it is a key issue in computational geometry, robot motion, graph drawing etc. based applications. The most accepted judgement about the quality of a viewpoint is in highly correlated with how much information this viewpoint gives about the environment or scene element. Vazquez et al. [21] propose a viewpoint selection algorithm which selects a set of good views to understand the scene. Their algorithm is based on viewpoint entropy that is derived from Shannon Entropy of Information Theory. Viewpoint entropy stands for the amount of information that one of the selected point of views provides. The amount of information is obtained by the projected areas and number of faces of scene elements.

The work of Vazquez et al. [21] is a satisfying solution for determination of a viewpoint around objects; however, it is not an approach which considers attention grabbing regions of objects. The problem evolves from the fact that presentation of detailed regions of objects is prior than crude geometry. Surface visibility is an example for the latter that ignores details but highlights total amount of projected areas. Thus, this approach may not be adequate for choosing most attractive viewpoint. Our viewpoint selection procedure is based on a more perceptual approach, so-called *saliency*, than visible scene elements in the capture. We employ the work of Lee et al. [22] who proposed mesh saliency concept in order to formalize searching process for the most significant parts of an
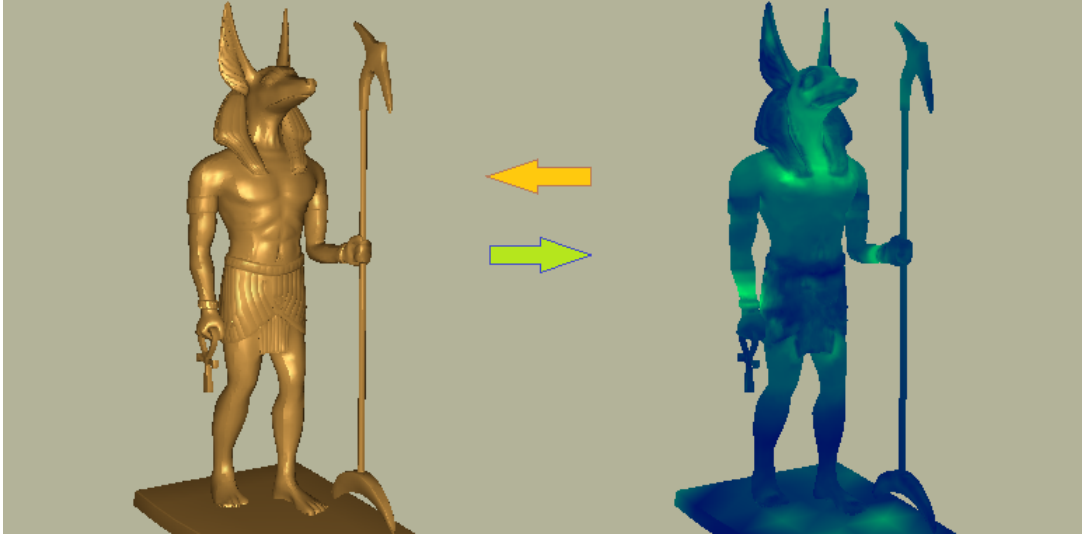
Figure 5.2: An important scene object on the left and corresponding salient regions are given on the right.

object that is investigated in cognitive science. Their work is based on calculating mean curvatures of meshes and finding regions which show considerably different mean curvatures than their neighbors. Salient parts of a 3D object are detected at the end of this process.

The computation of saliency is a costly procedure which cannot be processed in real-time, since the method deals with calculation of mean-curvature properties of each mesh and examination of their differences. Also, the processing time depends on the object size that yields different output-time for different objects. On the other hand, we achieve to calculate stereo camera parameters for each frame in real-time in the disparity adjustment stage. Therefore, saliency computation for each important object is applied in the pre-production stage in order to ensure that our system runs is in real-time. An important scene object that is one of the components of our scene and saliency output of the same object are given in Figure 5.2.

In addition to computation of saliency values for important objects in the scene, our system also detects the most salient part of the object. This most salient part represents the viewpoint of each important object. These determined viewpoints are used in the path generation stage. Therefore, the viewer

is guaranteed to be able to observe important scene contents by passing through positions that are located onto the directions which are objects' the most salient regions.

5.2   Path Generation

We confront the need for a curvature-based path structure in order to provide an exploration experience by strolling around important scene elements in a smooth manner. Different curve types are analysed in order to select the proper one for our need. We decide to employ quadratic Bézier curves among them; however, the disadvantageous side of Bézier curve is not providing same speed between curves, also within a curve too. In order to handle this problem, arc length parametrization is used which yields to provide a smooth camera movement that makes same distance in each frame. In the light of this work flow, our path generation mechanism is investigated under two main categories.

Bézier Curves

Several curve types are convenient to model a path structure. B-spline is one of the basic functions to generate curve shapes. The curve-fitting feature of B-spline provides to produce a smooth curve structure. On the other hand, this feature causes the difficulty for estimating the exact positions on the curve, since the curve is fitted to the control points that generates a structure positioned not around but between these control points. In a crowded scene, this situation may cause overlapping of curve positions and scene elements, which raises the problem of occlusion. In addition to this problem, computational complexity of B-spline function is more than tolerable limit for a real time application. Therefore, B-spline is not a suitable solution for our path structure.

A special case of a cardinal spline that is Catmull-Rom spline presents a reasonable solution for this issue. Curves are generated between two control points in this technique; however, slope of a curve is controlled by two other control points. In order to adjust the slope to a desired level, locating two other

control points requires a careful attention.

Instead of examined curve types, we make use of Bézier curve, a basic parametric curve that is easier to compute. Bézier curve is employed extensively in computer graphics for modelling smooth curves. Similarly, it is constructed using control points, where the number of control points represent the order of the curve.

In our case, control points are designated around important scene elements in order to move around them. If we divide whole path into smaller segments, then each segment corresponds to one curve around each important scene element. One control point stands for the start position of the movement around the object, the other one is obtained from the viewpoint selection phase. The last one represents the finish position of the movement for that object; whereas, it stands for the start position of another movement of another important object as well. The resulting curve is generated by interpolating endpoints while the remaining control point influence the curvature. In our case, three control points are sufficient in order to generate one Bézier curve around each important object in the scene content. Therefore, we employ quadratic Bézier curve in our system. The combination of these quadratic curves around important scene elements generates the overall path structure in the scene. The mathematical basis for quadratic Bézier curve is given in Eq. 5.1.

$$B(t) = (1 - t)^2 P_0 + 2(1 - t)tP_1 + t^2 P_2, t \in [0, 1] \tag{5.1}$$

where P_1, P_2, and P_3 are three control points respectively, each parametric entry produces a position along the curve and denoted by t. In our case, t is incremented by 0.005 for each frame.

Arc Length Parametrization

Bézier curve is in the shape of an arc, in which each calculated $B(t)$ value corresponds to a point on the arc length, that is also the position of the camera in each frame. However, Bézier equation is not a linear function, distances between a

point and successive ones are not equal. Therefore, resulting camera motion does not change at a constant speed which is an undesirable outcome of employing Bézier for a smooth camera motion. In order to address this issue, arc length parametrization is utilized to find camera positions which provides to move at a constant speed [23]. In the process, the curve length is estimated by calculating the linear distance between consecutive positions. Then, the curve is sampled and divided into equal distances. Each parametric entry that corresponds to the each sampled position is calculated by using ratios and proportions between sampled position and its closest two points, also previous value of the parametric entry. At the end of this procedure, equally incremented new positions along the curve are found which provides a smooth camera motion with constant speed.

Another case that should be considered for a correct camera motion is based on equalising camera speed among each curve. Each Bézier curve around an important element is generated by using different three control points. This situation leads to the problem of generating curves with different lengths. Small variances between each curve yields observable change of speed, which is handled by our system as well.

5.3 Camera Transformation

Points along the Bézier curve is finalized after arc length parametrization. Since these points correspond to the positions on the camera path, camera moves from one to another at each frame. Orientation of the camera has an important role like translation of the camera in order to maintain continuity during a camera motion. When camera is rotating, smooth transitions should be satisfied, otherwise instant changes of rotation causes an unnatural viewing experience. In our system, camera rotates towards important scene elements, since displaying attention-grabbing scene elements is an important exploration task. Camera orientation is performed differently in the two regions of the Bézier curve and disjunctive is the control point obtained from viewpoint selection. In the first half of

the Bézier curve, camera looks towards the important scene element. The direction of the third control point which represents finish position of that curve looks towards to the next important object. Therefore, camera viewpoint direction is linearly interpolated from that direction to the third control points' direction in the second half of the curve. This linear transition generates a smooth rotation between each quadratic curve.

Chapter 6

User Study and Evaluations

Our assumption is proposing a novel method for stereo rendering that presents both comfortable and realistic 3D illusion without losing perceived depth feeling. In order to evaluate the effectiveness of our proposed method, we have conducted several user study cases in two different scenes. Three aspects of the displayed contents, image quality, perceived depth, and visual (dis)comfort, are individually graded by subjects. Also, we compared our method with two other existing methodologies which are Naive approach where fixed stereo parameters are used and Depth Range Control where scene range is mapped to the desired perceived depth range. Subjects are asked to select one of the methods after displaying the content in pair-wise manner in order to assess the relative preference of the user. Experiment procedure is detailed in the following sections. The resulting output in comparison with Naive approach is given in 6.1.

6.1   Testing of Disparity Control

Subjects

We recruited 15 subjects, aged between 20 to 28, with a mean average 25. The subjects were among voluntary undergraduate and graduate students with
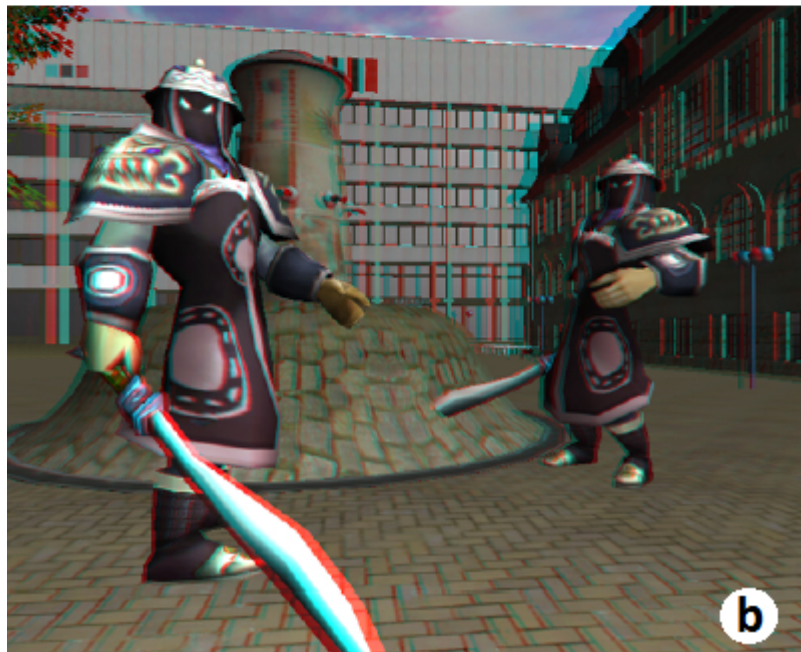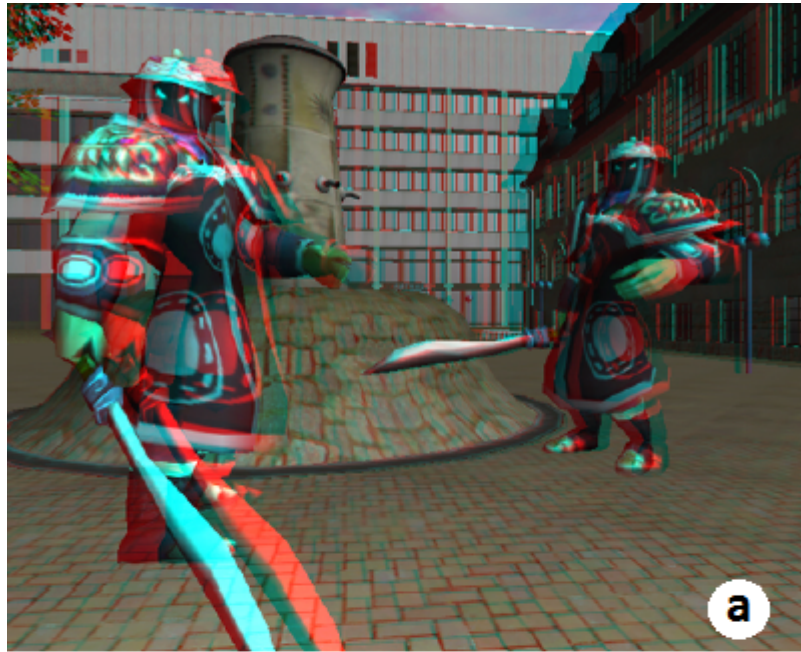
Figure 6.1: (a) An example capture of the scene with parameters calculated by Naive Method (b) The same capture with parameters calculated by our method
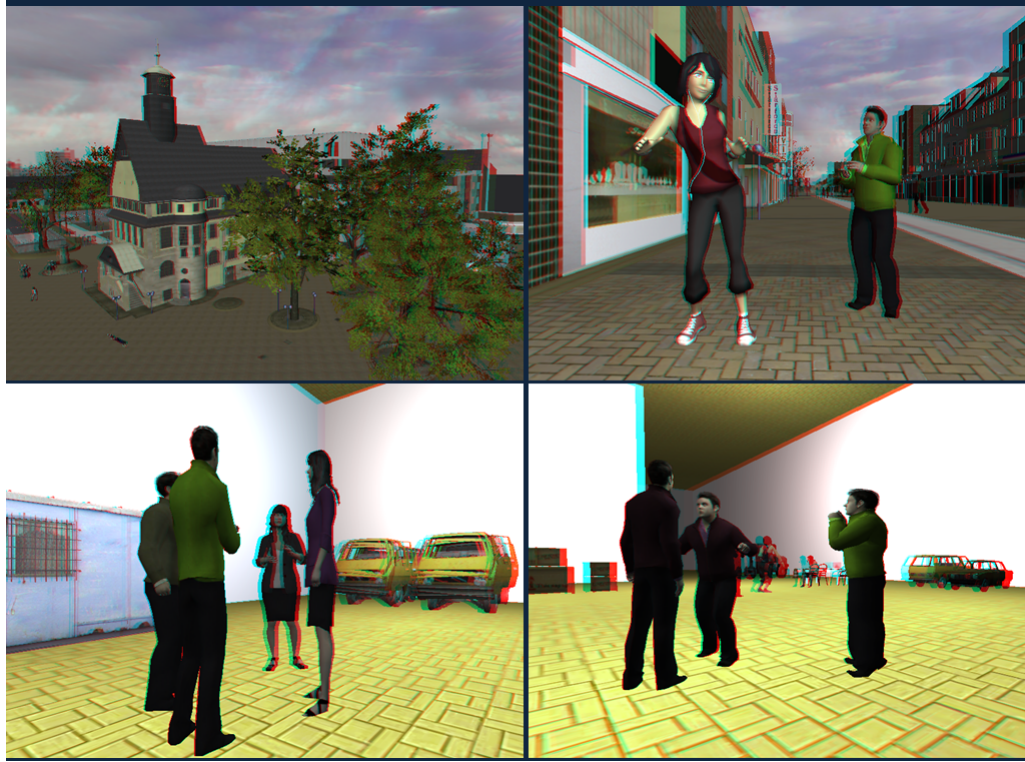
Figure 6.2: Sample snapshots of outdoor and indoor scene contents.

computer science background; and most of them did not have previous detailed experience on rendering on stereoscopic displays. The subjects were not informed about the purpose of the experiment. They were tested for proper stereoscopic visual acuity using random dot stereogram tests. The subjects who failed the random dot stereogram test did not participate in the user study.

Equipment

We used Nvidia GeForce GT 540M as graphics card and a 2.20 GHz Quad-Core laptop with 6 GB RAM for rendering. The stereoscopic pairs are displayed on a 40 inch 3D display with active shutter glasses, with a resolution of 1920 x 1080, in a dimly lit environment. The subjects were seated at a viewing distance of 2m.

Scenes

We built two interactive scenes (Figure 6.2) for the tests. The first scene

contains an indoor setting, where several groups of human characters, each of which performing various gestural movements, randomly distributed in a room. The second one is a city scene, which presents more dynamic environment in terms of variety of characters and their movements. Important scene elements are virtual human characters in both scenes. Significance scores are assigned in compatible with attention-grabbing degree of the character. As an example, a dancing character is more significant than a standing one. In each test, the user was asked to navigate freely in the environment.

Procedure

In the beginning of the experiment, each subject is being informed about the 3D stereo vision and possible encountered issues by giving related text document. Also, written instructions that describe tasks needed to be performed during experiment are presented to the subjects. The three attributes, which are used to grade displayed contents, are explained to the subjects in order to avoid potential incomprehension about the topic.
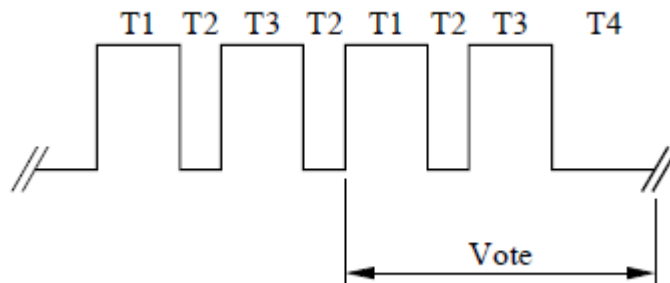

Figure 6.3: Presentation of test materials

We have followed the double stimulus continuous quality scale (DSCQS) method for the experiment design, given in Figure 6.3. DSCQS is one of the mostly used method for the subjective assessment of the quality of television pictures [24]. According to this procedure, subjects were shown a content, either test or reference; after a break, the other content was presented to subjects. Then, both contents were shown for the second time, to obtain the subjective evaluations. This process is illustrated in Figure 6.3. ITU-R BT.2021 [25], which is

a commonly employed recommendation that includes subjective methods for the assessment of stereoscopic 3DTV systems, states that DSCQS can be successfully used for the assessment of stereoscopic imaging technologies. Therefore, we decided to employ DSCQS method for the experiment procedure and order of test material, to make our user study more reliable.

Interactive tasks are performed in two different scene settings under two main phases. Each phase includes two sessions, corresponding two evaluation process and four test cases. As presentation of test materials is illustrated in Figure 6.3, two of test cases correspond to T1 and T3 are shown consecutively. These test cases are rendered with two different methods which stand for reference and test content in DSCQS. T2 is a mini break, where a mid-grey level screen is displayed for approximately 3 seconds. After that, T1 and T3 are shown again, while the subjects are asked to evaluate the presented materials at the same time. The same procedure is repeated in the second session. This time, one of the test cases is switch to the remaining stereo rendering method, while the second test case is shown with our method again. Moreover, both in the two sessions, subjects do not know which stereo rendering is shown first, The order of the reference and the test contents was determined in a randomized manner.

Assessment of Contents

Displayed stereoscopic contents, which are rendered with three different disparity adjustment methods including our approach, are evaluated by three primary perceptual dimensions: quality, depth, and comfort. These three criteria affect the quality of immersion feeling in stereoscopic systems according to [25]. Subjects evaluated both test and reference contents of all the cases separately. The meaning of each criterion, which were explained to the viewers before the experiment begins, is given as follows:

- *Image Quality:* Image quality denotes the perceived overall visual quality of the shown content. Ghosting, defined as the incomplete fusion of the left and right image so that the image looks like a double exposure, is a critical factor determining the image quality of stereoscopic content. A

good quality 3D stereo image should eliminate the ghosting effect.

- *Perceived Depth:* This criterion measures the overall perceived depth range of the scene content as reported by the user, so that the effect of the methods on apparent depth should be taken into account.

- *Visual (Dis)comfort:* This assessment item measures the subjective sensation of discomfort that can be associated with the improperly produced stereoscopic contents. A good quality 3D stereo image should provide a comfortable viewing experience. Otherwise, long-term exposure causes visual discomfort issues such as eye strain, fatigue, headache.

The quality of the displayed content is strongly related with sense of *presence*, a psychological state that describes the involvement and immersion feeling of an individual in a virtual environment. Two questionnaire types are used to measure effectiveness of virtual environments [26]. Presence Questionnaire (PQ) stand for measuring presence in a tested virtual environment. Immersive Tendencies Questionnaire (ITQ) is utilized to find out the capabilities and tendencies of individuals to experience presence. Underlying idea of PQ and ITQ is employed in our experiments. Random-dot stereogram test is used to adopt ITQ principles. For assessment of the stereoscopic content, an evaluation process is conducted at the end of each session aiming to employ PQ features simply. We first asked the subjects to rate the quality, depth, and comfort of both the reference and test methods separately, by filling out a 5-point Likert scale for each method. For grading of quality, depth, and comfort, we used the discrete scale with the labels "bad", "poor", "fair", "good", and "excellent". An example for this grading part is shown in Figure 6.4. We also asked the subjects to assess the relative comparison of the reference and our methods. For this purpose, at the end of each pair, we asked the subjects the following questions in the response form:

- Which session provided better image quality?

- Which session offered more depth?

- Which session was more comfortable to watch?

Which session was more comfortable to watch?	Session 1 No Difference Session 2
Grade comfort of Session 1	Bad Poor Fair Good Excellent
Grade comfort of Session 2	Bad Poor Fair Good Excellent

Figure 6.4: An example snippet from our questionnaire

6.2 Discussion

In order to analyze the results of the conducted user studies, firstly we computed the average scores of user ratings, as well as user preferences. These ratings and preferences are obtained by evaluation forms which are filled by each user.

Figure 6.5 illustrates the rating results for three employed perceptual dimensions, image quality, depth, and comfort. In each chart, the average grade is indicated in a circle. The results show that our method yields better average scores than other approaches in all three dimensions. Specifically, our method has achieved a considerable improvement in the stereoscopic image quality, due to the fact that our method ensures elimination of ghosting effect of the important objects which have higher significant scores. Since, convergence plane is located closer to these objects which results in lower screen parallax around them on the display image. Regarding to the Figure 6.5 the average rating of our method in the perceived depth is slightly better than the other two methods, also less number of subjects have evaluated our method as "bad" or "poor", compared to the other methods. The comfort ratings also reveal that our method is generally rated better than the other methods.

Figure 6.6 shows results of the preferences by comparing three perceptual dimensions of our method with other methods. These preferences are collected from the questions described in the Assessment of Contents. Different from the rating analysis of the methods, this chart shows the preferences in percentages for our method directly in comparison with other two methods. The study showed
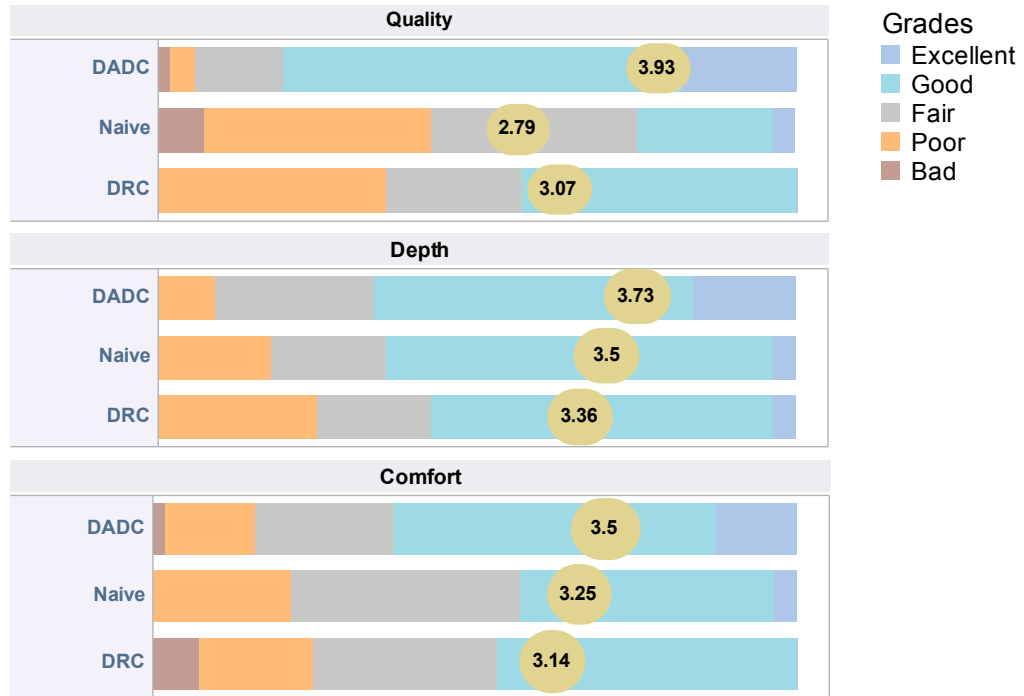
Figure 6.5: Comparison between three methodologies

that our approach was preferred over the other two methods, with a 64,28 % preference ratio; whereas 21,48 % of the results preferred the Naive over ours and 25 % of the cases showed preferences of DRC. The high performance of the Naive method is due to the fact that the static disparity levels were chosen to be compatible with the scenes, for a fair comparison.

To evaluate the cinematographic quality of each method, we have plotted a depth chart [27][28], which shows the distribution of the depth budget over time. The charts in Figure 6.7 shows the minimum and maximum depth values of the scene, with respect to the physical display surface. Figure 6.7 also shows the perceived depth of the most attention-grabbing object. The highest significant score assigned scene element is selected to be the most attention-grabbing object in the environment (orange curve). The results show that our method achieves the goal of keeping the most significant object appears closer to the display screen as much as possible. Based on these results, we can claim that our method prevents the accommodation-convergence conflict in a large extent.
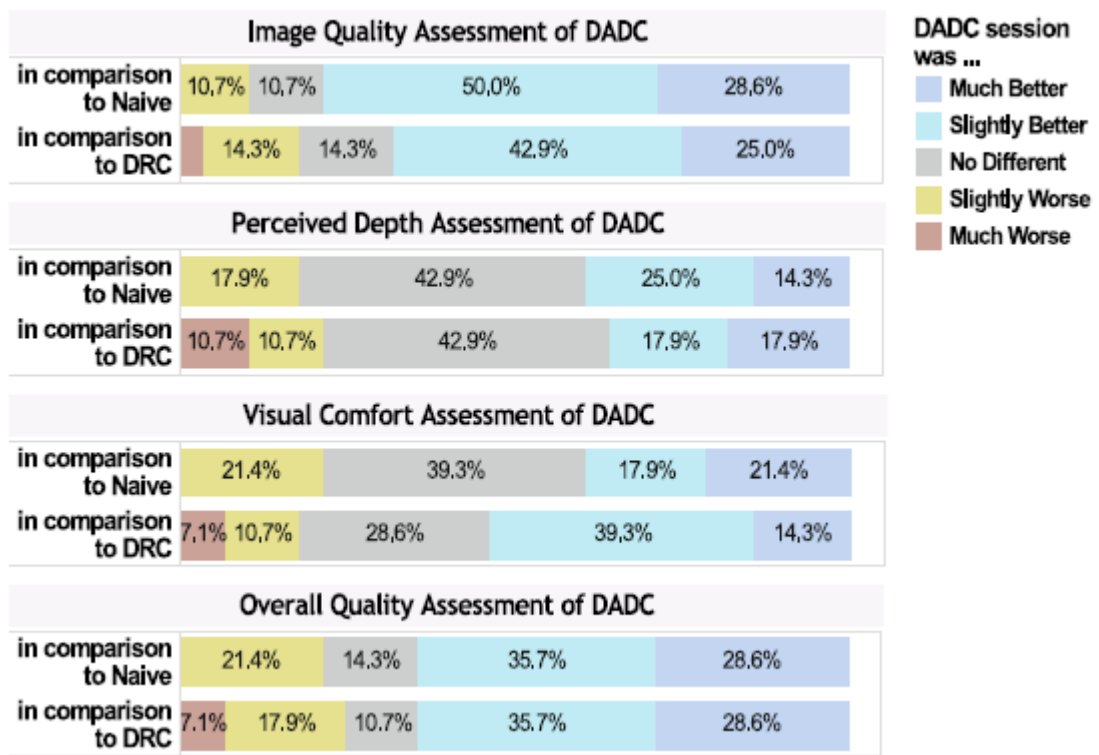
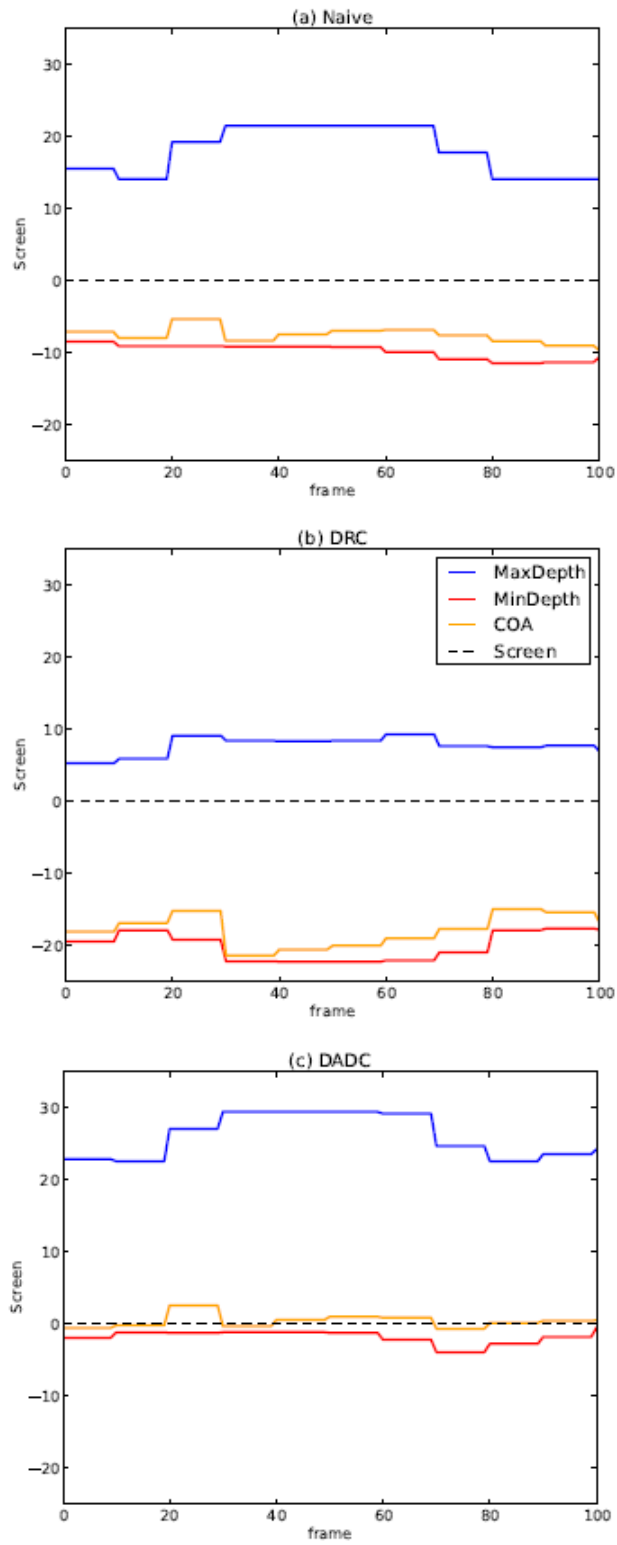Figure 6.6: Comparison results of our methodology with Naive and DRC approaches

Figure 6.7: Depth charts obtained by three different stereo rendering methods, Naive method (a), DRC (b), and our proposed method (c)

Figure 6.8: A sample scene prepared for the scene exploration task

In order to test the presented scene exploration approach, we have prepared an environment consisting of different 3D objects in the shape of sculptures and figurines 6.8. Significance scores assigned to these 3D objects; therefore, our approach asserts to generate a path, in which camera direction looks towards salient parts of these objects. Another sample scene environment composes of two scene elements with corresponding camera locations is given in Figure 6.9.
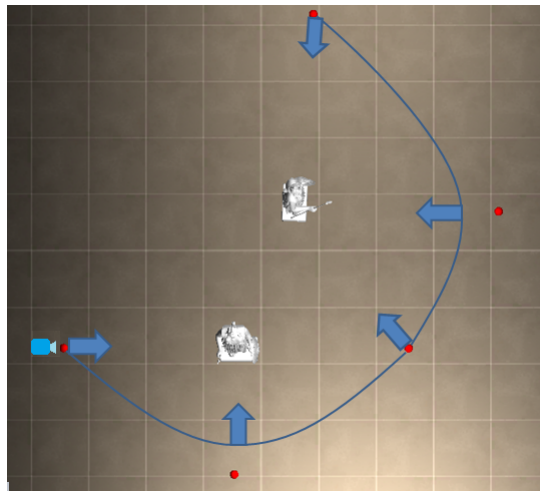

Figure 6.9: Orientation of the camera

Chapter 7

Conclusion

We have presented a camera control system by utilizing HVS perception principles for 3D contents. Our camera control mechanism composes of two main parts. We have introduced a novel approach for stereoscopic rendering which addresses calculation of stereoscopic camera parameters automatically and dynamically in the first part of the system. Our approach conveys scene depth in any arbitrary interactive 3D scene content by automatically calculating the stereoscopic camera parameters that are convergence and camera separation. Our method specifies a depth configuration according to the distribution and importance degree of attention-grabbing elements and depth range of the scene. It also automatically finds the camera parameters for mapping total scene depth to this specified depth range.

This new method for stereoscopic camera parameter adjustment allows 3D scene content creators to adjust available perceived depth in a way that the perceived depth is controlled and limited to the stereoscopic comfort zone of the users. This process is ensured by employing a disparity calibration phase, where the viewer's perceived depth limits are found. Also, accommodation/convergence conflict is handled by keeping the focus or the convergence of the camera closer to the elements of interest.

The other part of our system addresses scene exploration issues in a virtual environment. We employ linked and quadratic Bézier curves as the base for smooth curves around important scene elements to generate corresponding path in any arbitrary environment. Mesh saliency is used in order to find the most attention-grabbing regions of important objects that is used as the viewpoint selection phase. Therefore, output path passes through these saliency-based viewpoints. However, the process is done in a semi-automatic manner since control points around important objects are selected by the user. In order to convert the system into a fully automatic system, positions of control points should be found as a future work. With regard to test the efficiency of our proposed path framework, several user studies can be conducted by comparing our approach with different path finding algorithms as a future work. Also, different approaches can be compared with saliency for the viewpoint selection phase.

We have presented the results of user studies in order to comprehend the effectiveness of our disparity adjustment approach. Also, we compared our approach with two other existing disparity control methodologies, Naive and Depth Range Control. Results prove that our method shows superior results in quality, depth, and comfort.

Bibliography

[1] P. Milgram and M. Krueger, "Adaptation effects in stereo due to on-line changes in camera configuration," in *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, vol. 1669 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, pp. 122–134, 1992.

[2] I. P. Howard and B. J. Rogers, "Seeing in depth, volume 2: Depth perception," *Toronto: I Porteous*, 2002.

[3] B. Mendiburu, *3D movie making: stereoscopic digital cinema from script to screen.* Focal Press, 2009.

[4] F. Zilly, M. Muller, P. Kauff, and R. Schafer, "Stan - an assistance system for 3d productions: From bad stereo to good stereo," in *Electronic Media Technology (CEMT), 2011 14th ITG Conference on*, pp. 1–6, IEEE, 2011.

[5] S. Heinzle, P. Greisen, D. Gallup, C. Chen, D. Saner, A. Smolic, A. Burg, W. Matusik, and M. Gross, "Computational stereo camera system with programmable control loop," *ACM Trans. Graph.*, vol. 30, pp. 94:1–94:10, August 2011.

[6] S. Koppal, C. Zitnick, M. Cohen, S. B. Kang, B. Ressler, and A. Colburn, "A viewer-centric editor for 3d movies," *Computer Graphics and Applications, IEEE*, vol. 31, pp. 20 –35, jan.-feb. 2011.

[7] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3d," *ACM Trans. Graph.*, vol. 29, pp. 75:1–75:10, July 2010.

[8] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, and H.-P. Seidel, "A perceptual model for disparity," *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011, Vancouver)*, vol. 30, no. 4, 2011.

[9] P. Didyk, T. Ritschel, E. Eisemann, K. Myszkowski, H.-P. Seidel, and W. Matusik, "A luminance-contrast-aware disparity model and applications," *ACM Trans. Graph.*, vol. 31, pp. 184:1–184:10, Nov. 2012.

[10] G. Jones, D. Lee, N. Holliman, and D. Ezra, "Controlling perceived depth in stereoscopic images," in *in Stereoscopic Displays and Virtual Reality Systems VIII, Proceedings of SPIE 4297A*, vol. 4297, pp. 42–53, 2001.

[11] T. Oskam, A. Hornung, H. Bowles, K. Mitchell, and M. Gross, "Oscam-optimized stereoscopic camera control for interactive 3d," in *SA'11 Proceedings of the 2011 SIGGRAPH Asia Conference*, vol. 30, p. 189, ACM New York, 2011.

[12] O. Khatib, "Real-time obstacle avoidance for manipulators and mobile robots," *Int. J. Rob. Res.*, vol. 5, pp. 90–98, Apr. 1986.

[13] B. Salomon, M. Garber, M. C. Lin, and D. Manocha, "Interactive navigation in complex environments using path planning," in *Proceedings of the 2003 symposium on Interactive 3D graphics*, I3D '03, (New York, NY, USA), pp. 41–50, ACM, 2003.

[14] D. Nieuwenhuisen and M. Overmars, "Motion planning for camera movements," in *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 4, pp. 3870–3876 Vol.4, 2004.

[15] J. Kneafsey and H. Mccabe, "Camera control through cinematography for virtual environments: A state of the art report abstract," in *Proceedings of Eurographics Ireland Chapter Workshop*.

[16] D. B. Christianson, S. E. Anderson, L.-w. He, D. H. Salesin, D. S. Weld, and M. F. Cohen, "Declarative camera control for automatic cinematography," in *Proceedings of the thirteenth national conference on Artificial intelligence - Volume 1*, AAAI'96, pp. 148–155, AAAI Press, 1996.

[17] M. Guttmann, L. Wolf, and D. Cohen-Or, "Semi-automatic stereo extraction from video footage," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 136–142, IEEE, 2009.

[18] A. Greß, M. Guthe, and R. Klein, "Gpu-based collision detection for deformable parameterized surfaces," in *Computer Graphics Forum*, vol. 25, pp. 497–506, Wiley Online Library, 2006.

[19] T. P. Runarsson and X. Yao, "Stochastic ranking for constrained evolutionary optimization," *Evolutionary Computation, IEEE Transactions on*, vol. 4, no. 3, pp. 284–294, 2000.

[20] S. Johnson, "The nlopt nonlinear-optimization package."

[21] P.-P. Vázquez, M. Feixas, M. Sbert, and W. Heidrich, "Viewpoint selection using viewpoint entropy," in *Proceedings of the Vision Modeling and Visualization Conference 2001*, VMV '01, pp. 273–280, Aka GmbH, 2001.

[22] C. H. Lee, A. Varshney, and D. W. Jacobs, "Mesh saliency," *ACM Trans. Graph.*, vol. 24, pp. 659–666, July 2005.

[23] R. Parent, *Computer Animation: Algorithms and Techniques*. Morgan Kaufmann, 2002.

[24] "Recommendation itu-r bt.500-13, methodology for the subjective assessment of the quality of television pictures," 2012.

[25] "Recommendation itu-r bt.2021, subjective methods for the assessment of stereoscopic 3dtv systems," 2012.

[26] B. G. Witmer and M. J. Singer, "Measuring presence in virtual environments: A presence questionnaire," *Presence: Teleoper. Virtual Environ.*, vol. 7, pp. 225–240, June 1998.

[27] C.-W. Liu, T.-H. Huang, M.-H. Chang, K.-Y. Lee, C.-K. Liang, and Y.-Y. Chuang, "3d cinematography principles and their applications to stereoscopic media processing," in *Proceedings of the 19th ACM international conference on Multimedia*, MM '11, (New York, NY, USA), pp. 253–262, ACM, 2011.

[28] F. Zilly, J. Kluger, and P. Kauff, "Production rules for stereo acquisition,"
Proceedings of the IEEE, vol. 99, no. 4, pp. 590–606, 2011.