BIOLOGICAL-AWARE STEREOSCOPIC RENDERING IN FREE VIEWPOINT TECHNOLOGY USING GPU COMPUTING

Sammy Rogmans [1,2]*, Maarten Dumont* [1]*, Gauthier Lafruit* [2]*, and Philippe Bekaert* [1]

[1] Hasselt University – tUL – IBBT, Expertise centre for Digital Media
Wetenschapspark 2, 3590 Diepenbeek, Belgium
[2] Multimedia Group, IMEC
Kapeldreef 75, 3001 Leuven, Belgium

ABSTRACT

In this paper we present a biological-aware stereoscopic renderer that is used in a video communication system, to convincingly provide the participants with synthetic 3D perception. As opposed to conventional 3D systems – where pre-recorded content is presented to the viewer without taking his or her viewing location into account – we adaptively exploit both monocular and binocular cues of the human vision system, based on the viewing location. By using a GPU-based control loop, we are able to provide real-time synthetic 3D perception that is experienced as being rich and natural, without loosing any visual comfort whatsoever.

***Index Terms* —** biological, depth cues, rendering, free viewpoint, GPU computing, vergence, accommodation

1. INTRODUCTION

Presenting media content in 3D is becoming more and more popular, both in cinema theathers and at home, often relying on various hardware technologies – such as e.g. autostereoscopic displays and passive or active shutter glasses – but always presenting stereoscopic images to trigger synthetic 3D perception. However, a major drawback is that many viewers experience the 3D as unnatural, causing visual fatigue and feeling very uncomfortable [1].

The problem of unnatural 3D perception is that the presented stereo feed is pre-recorded without any knowledge of the position of the viewer [2], and that it does not properly take the biological cues of the human vision system into account. There are after all many depth cues that lead to convincing 3D perception, which are often drastically underrated. We are therefore strong proponents of decoupling the capture and render process, as free viewpoint technology – which is able to generate arbitrary virtual camera images – has the potential to solve these challenging issues.

We have developed a video communication system that uses a biological-aware stereoscopic renderer that takes the position of the participants into account. Furthermore, our system exploits the proper depth cues to trick the brain into perceiving genuine depth without loosing any visual comfort. Convincing depth is provided by enabling a rich set of biological cues next to conventional stereopsis. Moreover, as almost all algorithms are designed for use in GPU computing, our system achieves real-time speeds.

In Sect. 2, we discuss depth perception. Sect. 3 explains the problem with conventional synthetic 3D perception, while Sect. 4 discusses our novel biological-aware renderer, and Sect. 5 presents the results. Sect. 6 ultimately concludes the paper.

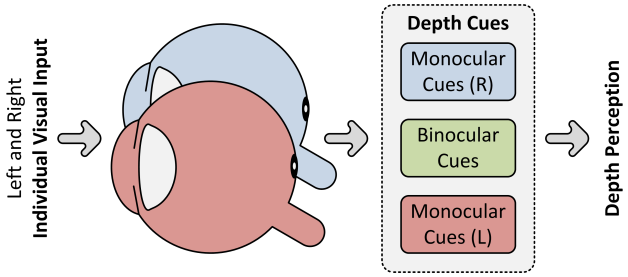

Figure 1. Schematical representation concerning the different depth cues of the human vision system that lead to natural depth perception.

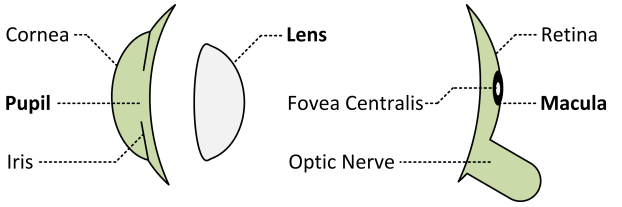

Figure 2. Simplified anatomy of the human eye, for more details the reader is kindly reffered to [3].

2. DEPTH CUES AND PERCEPTION

Perceiving depth in a natural and comfortable way is a highly complex biological process that occurs within the brain, which involves fusing and interpreting different depth cues of the human vision system. As depicted in Fig. 1, depth cues can be subdivided in two distinct groups – i.e. the monocular and binocular cues – which relate to providing additional depth information from one-eye individual and two-eye simultaneous visual input respectively.

Nonetheless individual depth cues each provide additional information to the brain, there exists a powerful link between the accommodation and vergence, i.e. the muscular reflexes to fixate upon an object, that provides with absolute depth perception. However, the majority of (monocular) depth cues provide with relative depth perception. Since the monocular cues are by far the largest group compared to the binocular ones, their significance to depth perception is often drastically underrated.

2.1. Monocular and Binocular Cues

There are over ten depth cues that provide depth information to the brain. They distinct in nine types of monocular and two types of binocular cues in total. To easily understand the monocular cue types, Fig. 2 presents the simplified anatomy of the eye. These

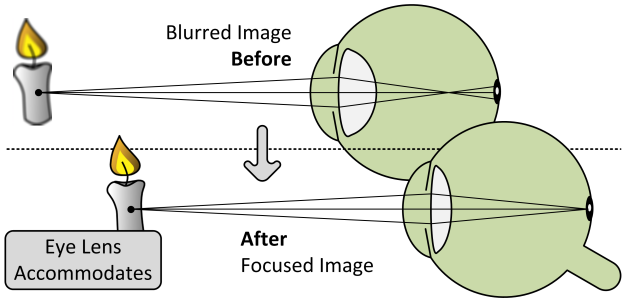

Figure 3. Accommodation of the eye lens after a change in visual fixation, placing the focal point upon the retina and macula.

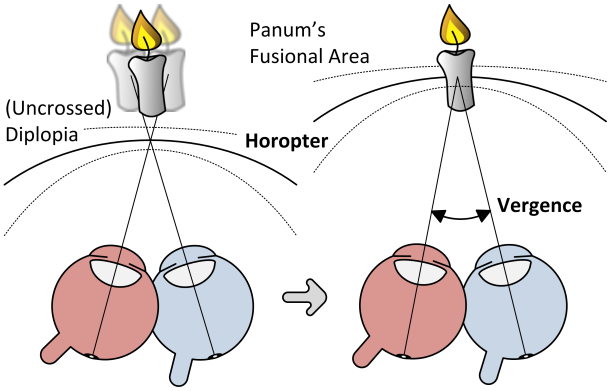

Figure 4. The process of vergence symetrically converges the two eyes, resulting in the capablity of fusing the binocular input.

cues are based on individual visual input from a single eye, and are therefore still active when one eye should be shut or disabled. The different types of monocular cues are:

1. **Accommodation**: When changing visual fixation, the image appears briefly blurred for the brain, since the focal point of the light rays entering the pupil do not coincide with the retina and macula. The brain immediately reacts to this, and the eye lens accommodates accordingly by intraocular muscles to focus the image (see Fig. 3). It therefore provides with oculomotor feedback.
2. **Occlusion**: Also called interposition or overlapping, occlusion is the most trivial and apparent depth cue. Objects that seem to occlude other ones, are interpreted as being closer.
3. **Size and Size Gradient**: As the brain aquires prior knowledge about the size of certain objects (e.g. cars, houses, etc.), estimating depth in an absolute manner becomes possible and more reliable. Furthermore, if one of multiple similar objects appears smaller, the brain will hence interprete that object as being relatively farther away.
4. **Motion Parallax**: Even a single eye can perceive depth if the head is moved, since objects that are closer will exhibit more parallax, i.e. the amount of visual shift is larger.
5. **Texture Gradient**: The more detail that can be seen, the closer an object will be interpreted. This cue should not be mistaken with accommodation, as it provides pictoral instead of oculomotor feedback to the brain.
6. **Shades and Shadows**: The brain generally tends to assume light always comes from above, due to the fact that most of the light is provided by the sun. The shades and casted shadows provide with extra relief and depth information.
7. **Linear Perspective**: The capability of recognizing planes and estimating vanishing points, e.g. parallel lines of a road that eventually meet in the horizon.
8. **Relative Height**: Objects that are smaller and closer to the horizon are observed as being farther away.
9. **Aerial Perspective**: Due to water and dust particles in the atmosphere, objects that are more in the background will appear more hazy, since more particles are in between.

In contrast with the numerous monocular cues, only two types of binocular depth cues can be noticed. They require visual input from both eyes simultaneously, and are therefore the most fragile and complex. The different binocular cues are:

1. **Vergence**: Visual fixation on an object with both eyes requires them to appropriately converge and rotate by the use of extraocular muscles. This process is guided by the brain, and is stimulated by four factors, i.e. accommodative (individual eye focus), tonic (concious use of neck muscles), proximal (awareness of proximity) and fusional (desire for single vision) convergence. Hence, this cue also provides with oculomotor feedback for absolute depth.
2. **Stereopsis**: As shown in Fig. 4, the two eyes converge symmetrically, causing light rays from points in space to be captured by corresponding photoreceptive areas in the two retinas. Such a point is said to have zero (angular) disparity. Moreover, the locus of these points is called the horopter. Points or objects in front or beyond the horopter will cause crossed or uncrossed disparities, which the brain is able to interprete as rich continious depth information.

2.2. Depth Perception

Human depth perception is a biological process – a black box if you will – that involves fusing and interpreting all individual information that is presented by each depth cue, both monocular and binocular. Inconsistent or missing cues do not always necessarily degrade the perception, because the brain is rapidly able to extract and interpolate good information. However, as accommodation induces the desire to converge, there is a significant link between them. Since the sixties, research has already shown that the process of vergence results over two thirds from accommodative convergence [4], and for only one third from tonic, proximal and fusional convergence, making the link between accommodation and vergence very powerful.

Furthermore, the brain is only capable of fusing binocular input within the vicinity of the horopter, which is often reffered to as Panum's fusional area (see Fig. 4). Objects outside this area cause crossed or uncrossed diplopia – i.e. double vision – because, identical to most monocular cues, stereopsis only provides with relative depth information, i.e. within Panum's fusional area.

Since accommodation, vergence and familiar size are the only depth cues that provide with absolute depth information, the importance of monocular (relative) depth cues is often drastically underrated in sustaining natural and rich depth perception.

3. FREE VIEWPOINT TECHNOLOGY

A major drawback in synthetic 3D perception is that very often the link between vergence and accommodation is destroyed, which quickly leads to visual fatigue and discomfort with the presented 3D content. Only small differences between these two extraretinal

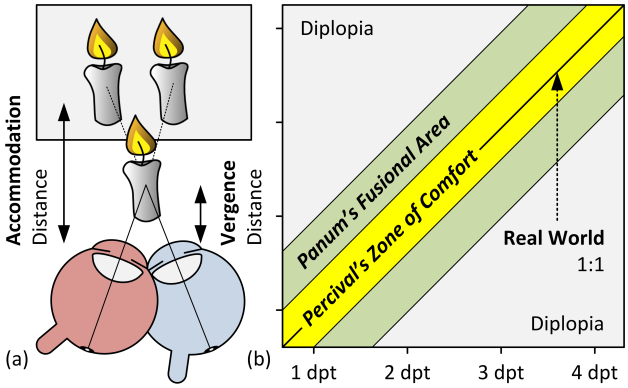

Figure 5. Overview of accommodation and vergence in (a) synthetic 3D perception, and (b) their relationship in dioptres.

cues are allowed for comfortable viewing. Pre-recorded stereoscopic footage therefore imposes drastic restrictions on the position of the viewer to correctly experience the 3D content.

However, recent free viewpoint technology – which is able to generate arbitrary virtual camera images – has the potential to solve vergence-accommodation conflicts. We have extended our previous standard free viewpoint framework [5] to enable challenging natural and comfortable synthetic 3D perception.

3.1. Synthetic 3D Perception

Recently, more and more media content is presented to allow synthetic 3D perception, using various hardware technologies such as autostereoscopic displays, passive/active shutter glasses with color-band or polarizing filters, and head-mounted displays.

The major drawback in synthetic 3D perception is that the accommodation distance – i.e. the distance to the screen – is unknown in advance and the presented stereo images cause the eyes to converge at a totally different distance (see Fig. 5a). As this destroys the link between the accommodation and vergence, recent user studies have proven the significance towards visual fatigue [6], however it already has been hypothesized for many years. As depicted in Fig. 5b, there is only a small zone where vergence-accommodation conflicts are still comfortably expierenced, which is consistently defined as Percival's zone of comfort.

Since stereopsis provides with relative depth information, the conflicts become more significant as distance between the screen and the viewer becomes smaller, making this issue even more apparent in home theathers or situations. Due to the relative behaviour of stereopsis, vergence and accommodation distance is often expressed in dioptres – i.e. a reciprocal of the distance.

When capturing stereo footage in advance, the position of the cameras automatically determines the vergence distance in function of the screen size that will be used [2]. A comfortable and natural 3D experience therefore drastically restricts the position of the viewer, if even possible (e.g. in home situations).

3.2. Virtual (Stereoscopic) Cameras

Current state-of-the-art free viewpoint technology is starting to enable real-time rendering of arbitrary camera viewpoints, based on an available set of physical cameras. Hence, it has the potential to solve current problems with synthetic 3D perception. Decoupling the capture and render process in this manner, is an almost untouched field in the domain of visual computing, and is sure to attract many researchers in the future.

In earlier research, we developed a free viewpoint framework that enables video communication with direct eye contact [5], by rendering a virtual camera image – based on a number of physical cameras placed around the screen – as if that camera would be behind and capturing through the screen, while directly looking into the participant's eyes. We further enhanced the communication by allowing synthetic 3D perception, which is quite challenging – due to the strong significance of the accommodation-vergence link – as the distance between the participants and the screen in video communication is small, i.e. about 30–70cm.

4. BIOLOGICAL-AWARE RENDERING

When rendering stereoscopic output with free viewpoint technology, the renderer needs to be aware of the biological aspects of depth perception, even more so in our experimental setup of close-range video communication. We therefore first determine the accommodation distance of the participants to the screen. The convergence angles and virtual camera positions are consequently adapted to maintain the important accommodation-vergence link, which results in stabilizing the stereopsis within the vicinity of the screen plane, and keeping it in Percival's zone of comfort.

However, stabilizing stereopsis inherently removes the frontal movements of the participants, i.e. moving towards and away from the screen. We therefore exploit the size monocular depth cue to trick the brain in seeing large depth movements, without creating any discomfort whatsoever. By additionally head-coupling the background, we efficiently create the cue of motion parallax, proving the effect as if the screen is a genuine window to the other particpant's location.

Furthermore, all remaining depth cues can either be fulfilled by a fixed setting (i.e. texture gradient, shadows and linear perspective), or are irrelevant (i.e. relative height and aerial perspective) for the entire video communication session.

4.1. Adaptive Vergence Control

As depicted on Fig. 6a, we define the two video communication participants as P_i with $i \in \{1, 2\}$. By using a well-known free viewpoint algorithm (i.e. GPU-based plane sweep [5]), we are able to extract depth information for the participants only – excluding the background – by means of silhouetting. As our framework uses a parallel histogram algorithm of the induced depth map, the accommodation distance A_i can be derived in real-time by interpreting the statistical data from the histogram, i.e. its Gaussian mean will indicate the position of the participant.

We determine the convergence $\theta_i = \arctan(2A_i/\text{IPD}_i)$, with IPD_i being the interpupillary distance, causing the vergence distance $V_i = A_i$, ergo maintaining the accommodation-vergence link (see Fig. 6b). The stereopsis therefore occurs within the vicinity of the screen plane, and in the center of Percival's zone of comfort. However, if one of the participants moves forward or backward, and therefore changes his accommodation distance, the vergence needs to be readjusted accordingly.

We implemented a (proportional) control loop, using the histogram analysis as feedback to tune the vergence to the detected accommodation distance. As the response time of the control loop is finite, sudden movements of the participants will cause the vergence and stereopsis to briefly come out or go deeper into the

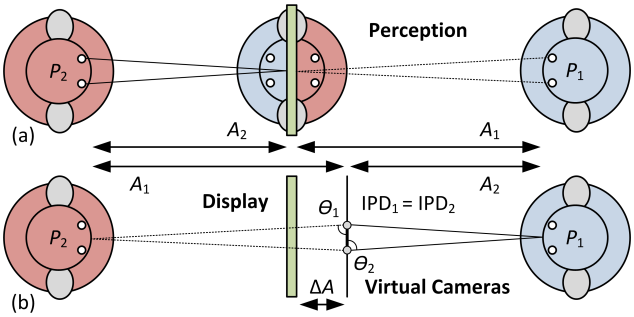

Figure 6. The (a) perceived situation, and (b) virtual camera rendering.

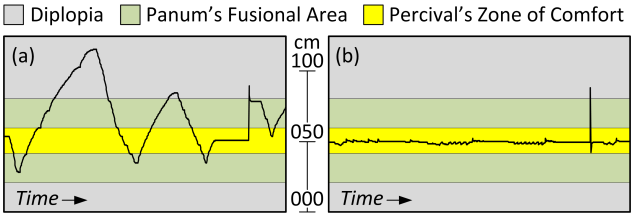

Figure 7. Depth perception (a) without and (b) with our adaptive vergence control, while sitting and remaining 50cm from the screen.

screen. However, normal speed movements are stabilized rather quickly, without allowing the stereopsis to exit the comfort zone.

4.2. Adaptive Zoom Control and Head Coupling

As shown in Fig. 6b, the distance between P_1 and the virtual cameras that capture P_1 – i.e. the virtual eyes of P_2 – is equal to A_2. Hence, these cameras almost rigidly follow P_1 and vice versa, inherently removing the frontal movements of the participants. To restore the effect of frontal movement, we exploit the familiar and gradient size depth cue by zooming in and out proportional to $\Delta A = A_1 - A_2$. This tricks the brain as if frontal movement is actually happening, while keeping the stereopsis within the vicinity of the screen plane and in the zone of comfort.

Since our video communication framework incorporates an eye-tracking module [5], we can additionally couple the head to the background, and move it accordingly to consistently provide with the motion parallax cue. Flattening the background makes it computationally lightweight, and causes no immediate cue conflicts as the stereopsis occurs in the (distant) foreground.

4.3. Fixed Depth Cue Settings

The remaining cues can be statically fulfilled, blur – i.e. texture gradient – and diplopia can be applied to the background without the brain ever noticing its static nature. Furthermore, the linear perspective cue is automatically respected thanks to the use of a perspective projection matrix in the rendering process. If extreme lighting conditions are avoided, then the same applies for the shadows cue. Because of indoor use, the relative heights and aerial perspective cues are irrelevant and cannot cause any conflicts.

5. EXPERIMENTAL RESULTS

As shown in Fig. 7a and b, we fixed the position of one of the participants, and plotted his vergence and stereopis without and with our adaptive vergence control loop. Considering normal-speed movements, the control loop is able to succesfully stabilize the depth perception in the center of Percival's zone of comfort, which is very close to real world perception.

All our implementations are designed or inherited to enable the use of GPU computing, either with traditional shaders or next-generation CUDA, whichever executes the concerning computational kernel most rapidly. Generally speaking, the performance of rendering kernels – i.e. graphics – versus computational analysis – i.e. computer vision – is faster in traditional shaders compared to CUDA and vice versa [7]. Hence, the system runs in real-time at 41 fps (800×600) on an NVIDIA GeForce 8800GTX.

6. CONCLUSION

Many cues in the human visual system contribute to natural depth perception, whereas monocular cues are often drastically underrated. Moreover, there is a powerful link between the vergence and accommodation which needs to be respected. Conventional 3D systems mostly present pre-recorded images, which imposes significant restrictions on the location of the viewer to succesfully perceive natural depth, and often leading to visual fatigue or an uncomfortable viewing experience. We therefore presented a biological-aware stereoscopic renderer which is able to use free viewpoint technology to adaptively generate images to control the vergence in function of the location of the viewer. The renderer hereby consistently maintains the accommodation-vergence link. By enabling a rich set of monocular cues, the brain can be tricked in perceiving genuine natural depth, without any visual discomfort whatsoever. As our system is driven by parallel GPU computing, it is capable of running in real-time over 40 fps.

7. REFERENCES

[1] Marc Lambooij, Wijnand IJsselsteijn, Marten Fortuin, and Ingrid Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review," *Journal of Imaging Science and Technology*, vol. 53, no. 3, pp. 030201/1–14, June 2009.

[2] Masaki Emoto, Takahiro Niida, and Fumio Okano, "Repeated vergence adaptation causes the decline of visual functions in watching stereoscopic television," *Journal of Display Technology*, vol. 1, no. 2, pp. 328–340, December 2005.

[3] Richard S. Snell and Michael A. Lemp, *Clinical Anatomy of the Eye*, Wiley-Blackwell, 1997.

[4] A. Hughes, "AC/A ratio," *British Journal of Ophthalmology*, vol. 51, no. 11, pp. 786–787, November 1967.

[5] Maarten Dumont, Sammy Rogmans, Steven Maesen, and Philippe Bekaert, "Optimized two-party video chat with restored eye-contact using graphics hardware," *Communications in Computer and Information Science*, vol. 48, pp. 358–372, December 2009.

[6] David M. Hoffman, Ahna R. Girshick, Kurt Akeley, and Martin S. Banks, "Vergenceaccommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, pp. 1–30, June 2008.

[7] Sammy Rogmans, Maarten Dumont, Gauthier Lafruit, and Philippe Bekaert, "Migrating real-time image-based rendering from traditional to next-gen GPGPU," in *Proceedings of 3DTV-CON: The True Vision Capture, Transmission and Display of 3D Video*, Potsdam, Germany, May 2009.