IMMERSIVE TELECONFERENCING WITH NATURAL 3D STEREOSCOPIC EYE CONTACT USING GPU COMPUTING

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

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

ABSTRACT

This paper presents an overview of our work in creating a practical prototype for natural 3D stereoscopic eye contact in teleconferencing. The two main pillars of our system are to 'immerse' the participants and to create as much as possible a natural feel to the communication. This is mainly realized by synthesizing virtual camera viewpoints to restore the eye contact, and by creating a sense of depth perception through feeding correctly generated stereo images directly to the users' eyes. Furthermore, we present an overview of GPU computing techniques that are used to maintain the real-time processing and transmission constraints of the system.

Index Terms— immersive, teleconference, stereoscopic, eye contact, GPU computing

1. INTRODUCTION

In artificial communication or video chat, the participants are not able to simultaneously look at the screen and the camera, leading to loss of eye contact [4]. Our teleconferencing system therefore uses a practical – i.e. in the monitor frame integratable – multi-camera setup around the screen to restore the eye contact of the peers [2]. Moreover, we have enhanced the system with 3D stereoscopic rendering to produce eye contact in an immersive way with a natural sense of depth perception. To strengthen this natural way of communicating through an artificial medium, we maximize visual comfort and therefore avoid eye strain at all times, while maintaining credible 3D stereoscopic eye contact. The system harnesses GPU computing to provide real-time processing, and is designed to optimally load the network for fast transmission.

Section 2 introduces the concepts that need to be taken into account to produce natural stereoscopic eye contact. Section 3 describes the architecture of our communication system, and explains how these stereoscopic principles are absorbed into the framework. In Section 4, we discuss the performance of our system using a head mounted display to sense the stereo feed. We conclude the paper in Section 5.

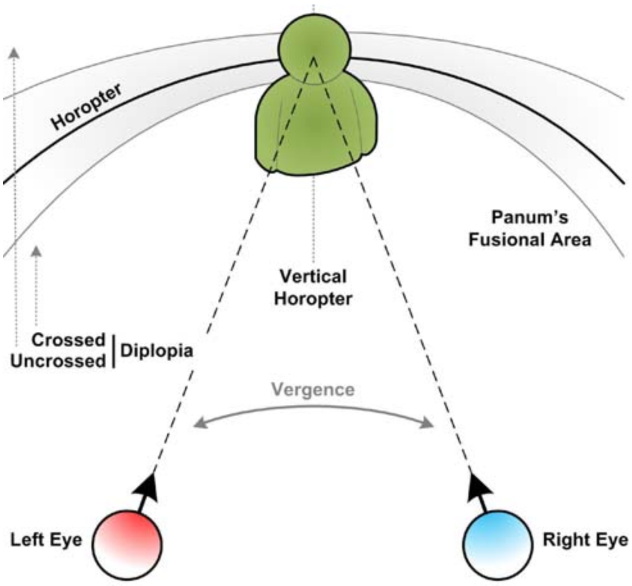

Fig. 1. Vergence, i.e. symmetrically converging the eyes to the horopter, exposing the locus of zero angular disparity. Objects' depth is only sensed within Panum's fusional area.

2. STEREOSCOPIC COMMUNICATION

In natural eye contact-driven social communication, both eyes are used to fixate on and create a sense of depth of the other party. To fully 'immerse' a participant in artificial communication, these principles of eye contact and depth perception should be preserved as much as possible.

When the eyes fixate on a given person, they focus to create a highly detailed image, resulting in a strong link between fixating and focusing. However, when the depth perception

[†] Funded by the Interdisciplinary Institute for Broadband Technology (IBBT) in the scope of the ISBO Self-Representation and Novel Interactive Experiences in Games and Applications (SEGA) project.

[*] This work is supported by the IWT within the Hasselt University under grant number SB071150.

is created artificially, this link is often destroyed – causing visual discomfort – and should therefore be avoided.

2.1. Natural Eye Contact

Whenever two persons are willing to initiate a conversation, their heads turn and their eyes immediately fixate on each other. Although the effect of genuine eye contact and its mechanisms considering the *social brain* – i.e. the cortical and subcortical regions which process social information – are until today still not fully comprehended [12], it is safe to say that in case the conversation participants are willing to make eye contact, their eyes 'seek' each other by fixating on the eyes of the other party.

Visual fixation can be best described as simultaneously moving both eyes in opposite directions, a process known as *vergence*, to converge to a specific object or point. As depicted in Fig. 1, the eyes consistently converge symmetrically, resulting in the fact that light rays from points in space will be captured by corresponding photoreceptive areas in the two retinas. Such a point is said to have zero (angular) disparity, moreover, the locus that exhibits zero disparity is called the *horopter* [1]. Objects or points that are not located on the horopter either cause crossed or uncrossed angular disparity when they are located in front or, respectively, beyond the horopter. Although disparities, combined with various monocular cues, are the fundamentals of depth perception in the human vision system, they can cause crossed or uncrossed diplopia – i.e. double vision – when they are too excessive. Within the vicinity of the horopter, *Panum's fusional area*, the eyes are still able to fuse the two input images and create a sense of depth.

Naturally, in everyday eye contact-driven social communication, participants fixate their eyes on each other, hence causing the eyes of the other party to lie on the horopter of the person seeking eye contact, while the body situates itself within Panum's fusional area. Considering immersive video communication by any artificial manner, these principles should ideally be preserved to create a natural and credible context to fully immerse the participant.

2.2. Maximizing Visual Comfort

Next to converging the eyes, the light casting from this fixated point into the iris is automatically focused by the eye lens. The eye lens changes in shape and *accommodates* for the light to fall on the fovea – i.e. the center of the macula. Since the fovea is the most susceptible photoreceptive part of the eye's retina, the object being watched is highly focused and sharply detailed. The natural process of converging (fixating) and accommodating to focus is strongly linked to each other and is performed subconsciously by the eyes and brain in a continuous manner.

Whenever an artificial way of depth perception is used – e.g. autostereoscopic displays, shutter glasses, anaglyph images or head mounted displays – the natural link between convergence and accommodation is often broken, causing eye strain and visual discomfort [6, 13]. As shown in Fig. 2, the convergence distance can differentiate from the accommodation distance, since the former is controlled directly by the disparities in the stereoscopic feed to the eyes, and the latter is determined by the distance to the (virtual) screen – or paper for that matter – that displays the stereo images.

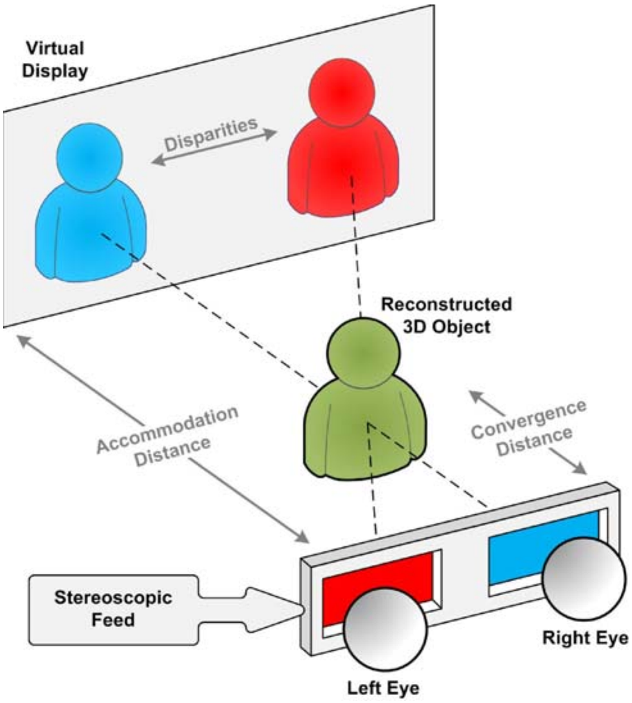

Fig. 2. Creating depth perception by an artificial manner often creates a difference between the convergence and accommodation distance, leading to visual discomfort.

To maximize visual comfort, the convergence and accommodation distance should be similar [5, 7], as the eyes and brain try to do this naturally. In concrete, this means that the stereo content should be adapted according to the artificial medium that is used to confidently perceive depth.

3. SYSTEM ARCHITECTURE

As depicted in Fig. 3, the stereoscopic processing of our video chat system consists out of four major building blocks. The preprocessing improves the data-locality and arithmetic intensity to benefit GPU computing, resulting in a drastic performance increase. The vergence control and eye tracker are closely related. Tracking the eye coordinates enables enhancing the immersive feeling, while the vergence control can also be dynamically configured to avoid breaking the link between convergence and accommodation. When the ideal camera positions are determined, a GPU-based stereoscopic plane

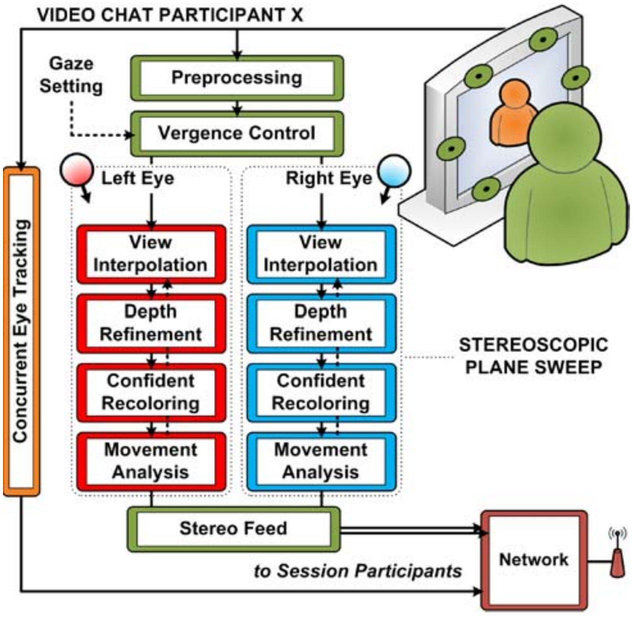

Fig. 3. Data flow of our system architecture, containing four important building blocks – i.e. the preprocessing, vergence control and eye tracking, a stereoscopic plane sweep, and the synthesized stereo feed to the network.

sweep is performed in parallel. The resulting synthesized images are packed in a stereo feed to match the used stereoscopic medium, whereas real-time transmission is maintained.

3.1. Preprocessing

In the preprocessing step, we take care of per-camera undistortion and foreground segmentation on the GPU, but mainly custom demosaicing of the input images to drastically improve the data-locality and arithmetic intensity of the algorithm [3, 9]. The undistortion is necessary to correctly access the camera images when backprojecting the plane sweep of the virtual camera, while the segmentation allows for efficient outlier detection, a significant performance increase, and the possibility to augment the participants with a virtual environment or background.

3.2. Vergence Control and Eye Tracking

As discussed in Section 2, convergence and accommodation distance should be matched to maximize the visual comfort in artificial stereoscopic communication. Although being relatively constant in most settings of video chat or teleconferencing, the accommodation is mainly determined by distance from the user to the screen being watched. Nonetheless, this distance is most definitely per-session variable. Our system allows for vergence control to dynamically adjust the disparities and the convergence of the eyes, to match the accommodation, avoid breaking their link, and ensure a natural and immersive feel when communicating.

The eye tracking is performed concurrently on the CPU and allows to adjust the virtual camera placement according to the head movements of the respective participant, further enhancing the immersive feel, often referred to as fish tank virtual reality [8]. In case of see-through anaglyph or active shutter glasses, the detected eye coordinates can be used in real-time to dynamically configure the vergence control.

3.3. Stereoscopic Plane Sweep

Our communication system allows for a stereoscopic plane sweep, which basically performs two parallel sweeps to synthesize the left and right virtual viewpoint, according to the camera positions determined in the vergence control. The first step of our GPU-based plane sweep algorithm [2] interpolates a synthetic view, as if the camera was behind the screen and is capturing through it, to directly look at the video chat participant. This approach allows for both restored eye contact and fish tank virtual reality, by coupling the virtual camera position – and displayed imagery – to the user's head.

In a second phase, the recovered depth information from the plane sweep is refined to exhibit more consistent depth information. The main purpose of this filter is to resynthesize or recolor the images with the refined depth information, and to avoid annoying artifacts or glitches when the stereo images get fused by the brain, providing a better overall quality and plausibility of fooling the brain into seeing natural images.

When the virtual camera images get recolored with the filtered depth information, it is quintessential to use synthesis techniques that lead to sharp detailed (high-frequency) images. In case of blurry images, the eye naturally starts to accommodate for this, whereas it is unable to do so. Because of this continuous and unstopple accommodation, the eyes will rapidly get tired, leading to a significant amount of eye strain. Our confident camera recoloring technique [3] copes with this problem and avoids putting too much stress on the video chat participants' eyes.

The final phase of our plane sweeping algorithm performs a movement analysis to determine the depth interval of the sweep – i.e. the depth position of the user that is being swept – which greatly reduces the complexity [10], and therefore significantly contributes to maintaining the real-time processing constraints. Furthermore, this information can simultaneously be used by the vergence control to restore the link between accommodation and convergence distance.

3.4. Stereo Feed and Network

The stereo feed packs the synthesized images together, allows for optional red/cyan anaglyph filtering, and possible compression to prepare the packet for efficient transmission.

Our system has been designed to perform cross-remote computations, to enforce data processing as close as possible to the input cameras. Although this approach induces redundant computations, the load over the network is minimized, and real-time transmission is provided.

4. RESULTS

Our enhanced teleconferencing system was tested with a head mounted display (FLCOS-type, see Fig. 4a) [11], which uses two LCD-alike screens with lenses to create a 70 inch virtual screen with an accommodation distance of 13 feet. Because the physical screens and lenses are fixed and strapped to the head, the accommodation is constant and simplifies the vergence control to a static setting that is detected in the video chat initialization. The $2 \times 800 \times 600$ stereoscopic feed can be produced from 6 input cameras in real-time at 26fps on an Intel Xeon 2.8GHz equipped with an NVIDIA GeForce 8800GTX. As shown in Fig. 4b, we have converted a stereo screenshot to anaglyph to give an idea of the system output.

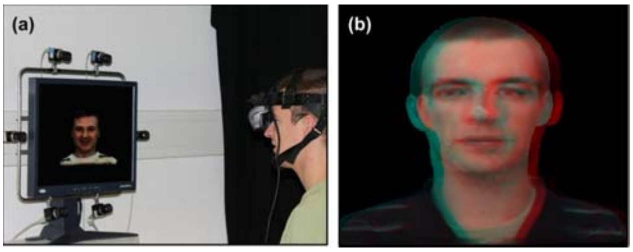

Fig. 4. The (a) head mounted display, camera-setup and (b) a screenshot of the stereo feed, converted in red/cyan anaglyph.

5. CONCLUSION

We have presented an overview of our work in creating a practical prototype for immersive teleconferencing with natural 3D stereoscopic eye contact. A significant amount of implications needs to be taken into account to preserve the natural feel of eye contact and depth perception. Mainly, the link between accommodation and convergence distance needs to be restored by dynamically adapting the stereo content to the artificial medium used for the communication, and the images need to be sharp to contain (high-frequency) details. Furthermore, we use optimized GPU computing to achieve real-time processing and transmission. More specifically, we rely on segmentation and movement analysis to drastically decrease the computational complexity, and use cross-remote computations to maximize data-locality.

6. REFERENCES

[1] Antonia Lucinelma Pessoa Albuquerque, Sergio Pinheiro, and Rubens Melo. Visualizing natural stereo images in short distance: A new approach. In *ISVC*, pages 1573–1583, Lake Tahoe, Nevada, USA, November 2006.

[2] Maarten Dumont, Steven Maesen, Sammy Rogmans, and Philippe Bekaert. A prototype for practical eye-gaze corrected video chat on graphics hardware. In *SIGMAP*, pages 236–243, Porto, Portugal, July 2008.

[3] Maarten Dumont, Sammy Rogmans, Steven Maesen, and Philippe Bekaert. Optimized two-party video chat with restored eye contact using graphics hardware. *CCIS*, 48(in press), 2009.

[4] Jim Gemmell, Kentaro Toyama, C. Lawrence Zitnick, Thomas Kang, and Steven Seitz. Gaze awareness for video-conferencing: A software approach. *IEEE Multi-Media*, 7(4):26–35, December 2000.

[5] Graham Jones, Delman Lee, Nicolas Holliman, and David Ezra. Controlling perceived depth in stereoscopic images. In *Stereoscopic Displays and Applications XII*, pages 42–53, San Jose, California, USA, January 2001.

[6] Kurtis Keller and D'nardo Colucci. Perception in hmds: What is it in head-mounted displays (hmds) that really make them all so terrible? In *Helmet- and Head-Mounted Displays III*, pages 46–53, Orlando, Florida, USA, April 1998.

[7] Marc T. M. Lambooij, Wijnand A. Ijsselsteijn, and Ingrid Heynderickx. Visual discomfort in stereoscopic displays: A review. In *Stereoscopic Displays and Virtual Reality Systems XIV*, January 2007.

[8] Jurriaan D. Mulder and Robert Van Liere. Enhancing fish tank vr. In *Virtual Reality*, pages 91–99, Washington, DC, USA, March 2000.

[9] Sammy Rogmans, Philippe Bekaert, and Gauthier Lafruit. A high-level kernel transformation rule set for efficient caching on graphics hardware. In *SIGMAP*, Milan, Italy, July 2009.

[10] Sammy Rogmans, Maarten Dumont, Tom Cuypers, Gauthier Lafruit, and Philippe Bekaert. Complexity reduction of real-time depth scanning on graphics hardware. In *VISAPP*, pages 547–550, Lisbon, Portugal, February 2009.

[11] Jannick Rolland and Hong Hua. Head mounted display. *Encyclopedia of Optical Engineering*, May 2005.

[12] Atsushi Senju and Mark H. Johnson. The eye contact effect: Mechanisms and development. *Trends in Cognitive Sciences*, 13(3):127–134, March 2009.

[13] Takashi Shibata. Head mounted display. *Displays*, 23(1–2):57–64, January 2002.