Saliency-Guided Stereo Camera Control for Comfortable VR Explorations

Yeo-Jin YOON†, Jaechun NO†, Nonmembers, and Soo-Mi CHOI†a, Member

SUMMARY The quality of visual comfort and depth perception is a crucial requirement for virtual reality (VR) applications. This paper investigates major causes of visual discomfort and proposes a novel virtual camera controlling method using visual saliency to minimize visual discomfort. We extract the saliency of each scene and properly adjust the convergence plane to preserve realistic 3D effects. We also evaluate the effectiveness of our method on free-form architecture models. The results indicate that the proposed saliency-guided camera control is more comfortable than typical camera control and gives more realistic depth perception.

key words: stereo 3D, visual comfort, visual saliency, virtual reality exploration

1. Introduction

As the demand for VR contents increases, stereo 3D contents for various applications have been actively studied. In recent years, visualization techniques using stereo rendering have been further developed and used as plugins to explore virtual spaces immersively in modeling tools such as Maya, 3D Max, and Rhino. However, it is still difficult to ensure a comfortable stereo experience because visual discomfort and fatigue may occur due to different production setups.

Visual discomfort symptoms such as headaches, dizziness, eye strain and nausea are mostly associated with unnatural viewing conditions or the perceived instability of the visual world [1]. Many researchers tried to reduce these effects by proposing a capturing device [2] and optimizing post-production pipelines with image techniques [3]. However, the human visual system yields more attention to salient objects [4] and visual saliency plays an important role to determine the visual comfort in stereo 3D. Stereo disparity remapping [5] places salient objects near the display presented for the improvement of visual comfort.

In this paper, we investigate the main causes of visual discomfort and present a stereo camera control method in the virtual space for comfortable VR explorations. First, we reduce visual discomfort in stereo 3D by developing a novel virtual camera controlling method. The camera convergence and separation are controlled in dynamic scenes to keep the depth of focus as comfortable as possible. Second, visual saliency is computed and combined with the virtual camera control to preserve realistic depth perception. For temporal interpolation, the saliency detection maintains high performance with GPU based implementation.

The overall process of our method is shown in Fig. 1. When physical parameters are given, those parameters are converted from the physical space to the virtual space. Then visual discomfort is analyzed using virtual disparity and screen parallax, and visual saliency from a current scene is simultaneously computed. Finally, the virtual camera separation and the convergence plane are adjusted to reproduce visually comfortable scenes.

2. Causes of Visual Discomfort

Although previous studies have addressed biological and technological reasons [6] to cause visual discomfort, here we briefly introduce causes which are considered important in this work.

Inappropriate disparity is caused by image or screen parallax (in pixels). The term disparity is divided into vertical and horizontal ones, and the horizontal disparity mainly affects depth perception. The human visual system attempts to fuse a stereo 3D scene even with excessive disparity.

Depth-of-field refers to the depth range which appears sharp in the scene. In natural viewing, this area is where the eyes converge on an object of attention due to vergence and accommodation, known as Panum’s fusional area. When the stereo 3D scene is configured, its depth-of-field is closely related to the setting of the depth of focus or convergence plane. If an object appears far away from the convergence plane but very clear, the object causes visual discomfort.

Depth distortion is an unnatural effect from the mismatch between retinal disparities and expected scene depths. A screen is placed about one meter away from a user, but the scene is perceived smaller and narrower than reality. Moreover, the recognition of depth gradient is difficult in some stereo scenes. Such depth distortions hinder natural viewing.
Existing methods for camera control can reduce visual discomfort, but realistic depth perception might not be preserved. In contrast to previous works\cite{5,7}, our method employs a depth distortion factor and visual saliency in the camera controller to relieve the depth distortion.

### 3. Saliency-Guided Stereo Camera Control

#### 3.1 Virtual Camera Control

To reduce visual discomfort, the virtual camera separation and convergence plane are controlled by synchronizing both physical and virtual spaces, which are related to stereoscopic viewing. By synchronizing two spaces, it is possible to change the parameters in one space to control the other space\cite{7}. The physical space unit (cm) is converted to the virtual space unit by using OpenGL Normalized Device Coordinates (NDC) according to display size and resolution. Equation (1) is represented in physical space, and Eqs. (2) and (3) in virtual space.

The perceived depth $Z_p$ in Eq. (1) is obtained by the parameters in the physical space where the user is located in a certain distance $Z_D$ from the display. We assume that $t_c$ is 6.5cm as a constant value because the average eye separation of adults is approximately 6.5cm. The screen parallax $P$ is computed by using the disparity $d$ in Eq. (2), in the virtual space.

$$Z_p = Z_D t_c / (t_c - P)$$  \hspace{1cm} (1)

$$d = t_c / w_s (1 - Z_s / Z)$$  \hspace{1cm} (2)

Note that the disparity $d$ is positive when point $Z$ is placed behind the convergence plane $Z_s$, and negative in front of the convergence plane $Z_s$, where the plane width is $w_s$. In this space, the virtual camera separation $t_c$ is controlled according to the screen parallax $P$ not to exceed an interpupillary distance. When the camera separation $t_c$ is too small and objects are placed too close to the camera, the convergence plane $Z_s$ is adjusted to keep visual comfort.

When the value of $t_c$ is increased, the depth range of the scene is varied and it looks deeper and narrower than real perception. To avoid such depth distortion, shape consistency is checked by the parameter $\mu$ as follows:

$$\mu = \frac{Z_D t_c}{Z_s t_c}.$$  \hspace{1cm} (3)

During an exploration, the virtual camera separation $t_c$ and convergence plane $Z_s$ are adjusted to keep $\mu$ close to 1. Figure 2 shows the comparison of stereo scenes without and with the distortion parameter $\mu$. The graphs on the right column show the depth ranges of stereo camera frustums and the convergence planes. The minimum and maximum in the graphs indicate the bounding area of an overall 3D model.

#### 3.2 Visual Saliency Detection and Convergence Plane Adjustment

When the camera parameters for the virtual space are altered to avoid visual discomfort, the stereoscopic depth range of the whole scene tends to be placed behind of the convergence plane. In this case, the user does not feel protrusion effects and only feels depth effects because the range of screen parallaxes has mostly positive values. In natural viewing, eyes converge to the focused object and other context objects are placed in front and behind of the convergence plane. Therefore, to keep comfortable and realistic 3D effects, we extract a salient object from the scene and move the convergence plane to the salient object’s position.

The overall process of saliency detection and convergence plane adjustment is shown in Fig. 3. First, mipmap images are acquired from the frame buffer, and the images’ color space is converted from RGB to CIEL*a*b*. Second, visual saliency is extracted by conspicuity and modulation computation from a high-level feature map to a low-level feature map. Third, if the level becomes zero, the saliency candidate points are unprojected from 2D to 3D space to find a salient object from the object buffer. To place the object in focus, the position of the convergence plane $Z_s$ is properly adjusted.

For conspicuity computation, a set of six mipmap levels is obtained from the render buffer. Then conspicuity analysis\cite{8} is applied to the images, which uses center-surround differences to get three levels of the conspicuity map $C_k$. Here, the low-level represents fine features and the high-level indicates coarse features. The conspicuity map $C_k$ is computed by the following equation:

$$C_k = \sum_{n=0}^{n_m} \sum_{m=0}^{m_n} k_n - k_{n+m}. \hspace{1cm} (4)$$
The conspicuity maps $C_k$, $k = \{L^*, a^*, b^*\}$ are at different mipmap levels $n$ and $m$. We set $p = 6$, which is the level of image pyramids used in Eq. (4). In the modulation step, the conspicuity map values are divided into focus and context areas, and the focus areas are stored as modulation maps. We assume that 30% of top values of conspicuity map entries are the focus area and others are context. A threshold $t_k$ is determined by averaging the values of the focus area, and the modulation maps are calculated as follows:

\[
 m_k = \begin{cases} 
 0 & \text{abs}(c_k) < \text{abs}(t_k) \\
 \text{abs}(c_k) - \text{abs}(t_k) & \text{otherwise} 
\end{cases} \tag{5}
\]

The subscript $k$ of the conspicuity map $C_k$ indicates different color channels, $a^*$ and $b^*$. Channel $a^*$ contains red-green opponent information, and channel $b^*$ indicates blue-yellow. In the case of channel $L$, the modulation map includes only lightness information. Therefore, we calculate the modulation map $m_k$ by separating negative and positive values and taking absolute values. Once $m_k$ is calculated, the highest modulation value from the modulation map is extracted and unprojected from 2D image space to 3D space. The unprojected points of 3D space determine a salient object by detecting collisions with object’s surface. To extract salient objects in high level, an object buffer is constructed and the information of each object is stored when a 3D model is loaded. Note that the salient object can be stored as a saliency or ignored according to the scene context even if a certain object is detected as a saliency during the process of saliency extraction.

In our configuration, the wall, floor or ceiling are ignored as context by storing material information in the object buffer because some obstacles such as sculptures are more important than background objects for indoor exploration. If the number of extracted objects is more than one, the closest object to the virtual camera is selected as the salient object. On the contrary, if none of the objects are extracted, the virtual camera is automatically moved without considering saliency. For a sudden replacement of a salient object, the camera convergence plane is smoothly adjusted to the depth of a new salient area if the replaced object is in the visual comfort zone; otherwise it is ignored by the virtual camera. Figure 4 shows detected salient points and the extracted salient object from each scene. Small red squares on the left images indicate the most salient areas, respectively. There are nine candidates in total from three levels of the modulation map for each channel. Every candidate is voted for a saliency or other context, and the candidate with the highest number of votes is selected as the salient object. Then the depth information of the salient object is acquired and the convergence plane is moved to the location of the salient object to make a better stereo scene. In order to save a saliency calculation cost, the saliency extraction is performed every 5 seconds. The input mipmap images are generated using an additional frame buffer object (FBO). Real-time exploration is possible using multi-threaded shader execution.

![Fig. 4](image)

**Fig. 4** Extracted salient points (left column) and salient objects in each scene and corresponding saliency map (right column).

![Fig. 5](image)

**Fig. 5** Comparison of stereo scenes; (a) shows the scene rendered by typical camera control [9] and (b) shows the rendering result using the proposed method.

### 4. Experimental Results

We compared the proposed saliency-guided camera control with the existing camera control [9] described in Sect. 3.1. In the first experiment, camera control methods with and without saliency are compared for static scenes. The convergence plane has properly adjusted from 29.5 to 68.0 OpenGL units by the consideration of the salient object in Fig. 5. In this experiment, two anaglyph stereo scenes are rendered with the ranges of $[10.68, 151.25]$. The salient sculpture is rendered on the convergence plane and the nearest sculpture has more of a protrusion effect, as shown in Fig. 5 (b).

As a second experiment, we created dynamic stereo scenes using both camera control methods, and the perceived depth range with respect to the convergence plane traced during exploration. In Fig. 6, the left column illustrates the results of typical camera control [9], and the right column shows the results of the proposed camera control using visual saliency. In Fig. 6(a), during navigation, the perceived depth range alters only when the scene is analyzed to be uncomfortable. On the contrary, our method properly adjusts the convergence plane and the perceived depth range.
smoothly varies during the whole exploration by detecting salient objects.

In addition, the interaxial distance of the stereo camera (a camera separation) for Fig. 6 (a) becomes smaller than (b). Consequently, it slightly distorts binocular depth perception. The proposed method reduces this side effect by using the depth distortion parameter and more proper camera separation along the change of the convergence depth.

Although, the stereo quality metric based on saliency [10] was introduced, it is not suitable for the presented method because the convergence plane is particularly adjusted to the salient area where the stereo disparity is almost zero. Therefore, we simply performed a user study to show the effectiveness of our saliency-guided stereo camera control. A total of 38 subjects participated in this study with the exception of the stereo blind ones. Each subject explored two architecture models, the Guggenheim Museum and the Jubilee Church, using different camera controls and without any camera control. We randomized the order in which the test scenes were rendered by using the three different methods. In the user study, participants used a stereoscopic display (42") and wore stereo shutter glasses, and the following questions were asked:

- **Q1. Which exploration is visually comfortable?**
- **Q2. Which exploration has more 3D effects?**

More than 70 percent of participants responded that the explorations with a camera control are visually comfortable compared to without one, and the saliency-guided camera control is slightly more comfortable than the camera control without considering saliency. They also answered that the existing camera control is quite comfortable, but sometimes it has a tendency to suppress stereoscopic cues, resulting in shallow depth of field. For both architecture models, most participants answered that the saliency-guided camera control provides more realistic stereoscopic effects.

### 5. Conclusion

In this paper, we presented a camera control method by synchronizing physical and virtual spaces to minimize visual discomfort for dynamic stereo 3D explorations. In particular, the depth distortion and visual saliency factors are used to preserve a proper and realistic depth perception. The presented method can be generally applied for stereo 3D tools because it does not require additional equipment to track the viewer’s gaze position. Furthermore, it takes advantage of a large display with multiple users. In addition, the proposed method is performed with a simple calculation so that it does not affect the rendering speed significantly. In this work, we only considered a desktop VR environment, but our method can be extended to different VR platforms such as HMDs and a large VR environment.

**Acknowledgements**

This research was supported by Korean MSIP under the ITRC support program (ITRP-2017-2016-0-00312) supervised by the IITP and Korean NRF grant (2014R1A2A1A11053135).

### References

[1] M. Lambooij, M. Fortuin, I. Heynderickx, and W. IJsselsteijn, “Visual discomfort and visual fatigue of stereoscopic displays: A review,” Journal of Imaging Science and Technology, vol.53, no.3, pp.30201-1–30201-14, 2009.

[2] 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, no.4, pp.94:1–94:10, 2011.

[3] K. Carnegie and T. Rhee, “Reducing visual discomfort with HMDs using dynamic depth of field,” IEEE Comput. Graph. Appl., vol.35, no.5, pp.34–41, 2015.

[4] A. Kimura, R. Yetenari, and T. Hirayama, “Computational models of human visual attention and their implementations: A survey,” IEICE Trans. Inf. & Syst., vol.E96-D, no.3, March, pp.562–578, 2013.

[5] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, “Nonlinear disparity mapping for stereoscopic 3D,” ACM Trans. Graph., vol.29, no.4, pp.75:1–75:10, 2010.

[6] T. Oskam, A. Hornung, H. Bowles, K. Mitchell, and M. Gross, “OSCAM — Optimized stereoscopic camera control for interactive 3D,” ACM Trans. Graph., vol.30, no.6, Article No. 189, 2011.

[7] E.E. Veas, E. Mendez, S.K. Feiner, and D. Schmalstieg, “Directing attention and influencing memory with visual saliency modulation,” Proc. SIGCHI Conference on Human Factors in Computing Systems, pp.1471–1480, New York, NY, USA, 2011.

[8] Y.-J. Yoon, I.-K. Shin, Y. Jang, and S.-M. Choi, “Stereoscopic 3D exploration of freeform architecture,” Automation in Construction, vol.46, pp.1–10, 2014.

[9] Y.J. Jung, S.-I. Lee, H. Sohn, H.W. Park, and Y.M. Ro, “Visual comfort assessment metric based on salient object motion information in stereoscopic video,” Journal of Electronic Imaging, vol.21, no.1, pp.011008-1–011008-16, 2012.