Perceptual Enhancement for Stereoscopic Videos Based on Horopter Consistency

Zeyu Wang*[1], Xiaohan Jin[1], Fei Xue[1], Renju Li[1], Hongbin Zha[1], and Katsushi Ikeuchi[2]

[1]Key Laboratory of Machine Perception, Peking University
[2]Microsoft Research Asia

Abstract

Audience discomfort, such as eye strain and dizziness, is one of the urgent issues that virtual reality and 3D movie technologies should tackle. Except for inappropriate horizontal and vertical disparity, one major problem is that people's binocular vergence and focal length in the cinema remain inconsistent from normal visual habits. Psychologists discovered the horopter and Panum's fusional area to describe zero-disparity points projected on the retinas based on accommodation-convergence consistency. In this paper, inspired by these concepts, we propose a stereoscopic effect correction system for perceptual enhancement according to fixated region and scene information. As a preprocessing step, tracking and stereo matching algorithms are implemented to prepare cues for further transformation in 3D space. Then in order to accomplish certain visual effects, we describe a geometric framework for disparity refinement and image warping based on parameter adjustment of the virtual stereoscopic rig. For evaluation, subjective experiments have been conducted to prove the effectiveness of our method. Therefore, our work provides a possibility to improve the audience experience from a formerly underexplored perspective.

Keywords: horopter consistency; virtual rig modification; perceptual enhancement; stereoscopic videos; image warping

Concepts: •**Computing methodologies** → **Image and video acquisition; Image manipulation; Perception;**

1 Introduction

Nowadays, three-dimensional movies are broadly regarded as an indispensable branch of the film and VR industry, which take the approach of delivering video pair from two different perspectives to people's eyes. The disparity between corresponding parts in two channels will cater to the need of human visual system and hopefully generate depth perception. As a matter of fact, primitive attempts on producing 3D films took place even as early as the beginning of twentieth century. Along with the development of film cinematography, especially as we have entered the digital era, 3D films are embracing more advanced technologies as well as a wider audience. However, although there is a scramble for stereoscopic cinemas, many cases of eye strain and dizziness are reported. Hence, hot discussions emerge both in the industry and academia, regarding problems such as vertical disparity and theater layout. A somehow unsatisfactory fact is that current 3D films only bring a plausible convergence experience, which may not be consistent with the focus region of the audience, and it remains a problem urgently needing to be solved.

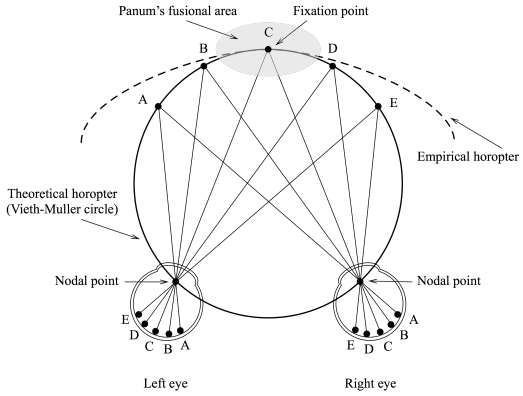

Figure 1: *The horopter in the horizontal plane. The theoretical horopter is called the Vieth-Muller circle, which passes through the nodal points of both eyes and the fixation point (C). The empirical horopter (dashed curve) is slightly behind the theoretical one. Environmental points within Panum's area are fused perceptually into a single image.*

The state of the eyes and their components can provide ocular information about the distance to a fixated surface. Of particular importance for depth perception are the focus of the lens (accommodation) and the angle between the two eyes' lines of sight (convergence). Accommodation and convergence normally covary; as the distance of the fixated object changes, both accommodation and convergence change in lockstep [Palmer 1999]. In studies of human's binocular vision, the horopter is a locus of points in space which are projected on anatomically identical positions in the two retinas, shown in Figure 1. According to the pinhole camera model, the theoretical horopter is called the Vieth-Muller circle, determined by the fixation point and the nodal points of both eyes. The empirical horopter, however, is slightly different and has become known as Panum's fusional area. Objects with disparities within an absolute limit will appear fused and single, while those with disparities outside the limit appear double [Burt and Julesz 1980]. During the long process of biological evolution, human beings have been developing a habit that focal length adjustment of the eye has to conform to the binocular vergence. The accommodation-convergence consistency based on the horopter may lower the computational cost of finding stereo correspondences, because the fovea centralis on which fixation points are projected has the highest resolution. Here comes the **principal hypothesis** in this paper:

- In our perceptually enhanced videos, the disparity of the attention area should be set to near zero in order to achieve horopter consistency, while other parts are changed correspondingly.

*e-mail: {1200012927, 1200012629, 1200018415}@pku.edu.cn, {lirenju, zha}@cis.pku.edu.cn, katsuike@microsoft.com

VRST '16, November 02-04, 2016, Garching bei München, Germany
ISBN: 978-1-4503-4491-3/16/11...$15.00
DOI: http://dx.doi.org/10.1145/2993369.2993393

This can be accomplished by multimedia techniques of disparity refinement. In our pipeline, potential fixated regions are first selected and tracked to adjust the horopter, and then virtual rig modification is implemented to produce the video with perceptually enhanced effects. As Figure 2 illustrates, our input is a video pair stream from both channels. The attention area can be found through object detection, tracking and coarse segmentation, which is described in Section 3.1. After image rectification and stereo matching, disparity and perceived depth will be reconstructed, which is described in Section 3.2. Section 4.1 introduces the geometric framework and provides the theoretical basis for image warping to improve depth perception. According to the expected effect, Section 4.2 selects the mode of parameter adjustment, including baseline, dolly, field-of-view (FOV), and translation (to simulate convergence). Finally, in Section 5, we conclude with the result of effect enhancement and evaluate our approach by subjective experiments.

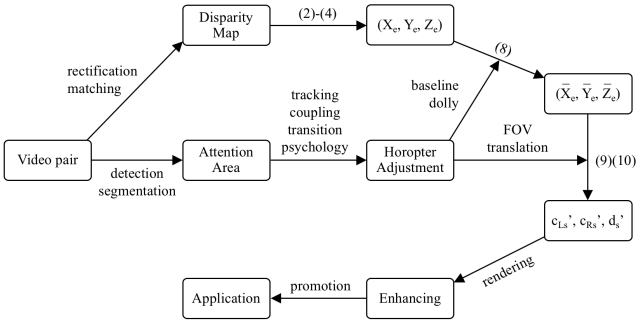

Figure 2: *Overview of our system for effect enhancement. The input is an original video pair stream and the output is the enhanced video pair stream. Several techniques of object tracking, depth reconstruction and image warping are integrated and utilized.*

2 Related Work

Along with the rapid development of the film and VR industry, related researches have emerged in large numbers, whose interests cover stereography, scene analysis, quality assessment, post-production, and artistic intent delivery. In terms of perceptual enhancement, former research focuses on warping and rendering details. For example, image warping methods are developed, so stereo editing software enables filmmakers to modify depth effects. However, they seldom consider the initial motivation and basis for manual adjustment, and rely heavily on ad hoc methods instead of a general well-defined framework. Recently, there are two remarkable pieces of work. One is adaptive 3D rendering based on region of interest [Chamaret et al. 2010], which takes advantage of the saliency map and shifts the disparity correspondingly. Another is seamless disparity manipulations named GazeStereo3D [Kellnhofer et al. 2016], which alleviates visual discomfort by gradual depth adjustments at the eye fixation stage.

Nonetheless, both of them underestimate the setup of real stereoscopic rigs, and lack a theoretical basis for disparity mapping. Therefore, we propose a horopter-based automatic system for perceptual enhancement of 3D videos. As a substitute for the previous art creation manner, the horopter constrains humans fixation and convergence to be consistent as normal habits, and provides a psychological basis for stereo effects enhancement. Thus, it is highly feasible to apply our method into real-time stereo photography to make the fixated object fall on the horopter by adjusting the camera parameters automatically. The unique preprocessing module with a geometric framework becomes able to improve our stereo comfort.

3 Preprocessing

3.1 Attention Area

Since our main purpose is to enhance depth perception according to the hypothesis of horopter consistency, it becomes the very first step to determine people's fixated region, or in other words, attention area. Considering this is a cinematographic scenario, filmmakers usually set up such a well-designed scene that can induce the audience to pay attention to a certain area such as the starring role [Andersson 2015]. Therefore, in this system, production staff are empowered to select attention area in the first frame, and tracking methods will decide attention areas in the following frames. We use tracking-learning-detection (TLD) [Kalal et al. 2012] algorithm to track target people or object in a video stream and return its position as zero disparity region for further use. For example, we can set the tracked target constantly at the center of the field-of-view and at zero disparity, so it becomes perceptually static in position and depth, while other parts in the scene form a reference system for observing dynamic flows around the target.

This state-of-the-art tracking framework TLD explicitly decomposes the object tracking task into tracking, learning and detection, which can even cope with unknown targets. The tracker follows the object from frame to frame. The detector localizes all appearances that have been observed so far and corrects the tracker if necessary. The online learning module estimates errors of the detector and updates it to avoid these errors in the future. P-N learning is adopted to estimate errors by a pair of "experts": P-expert estimates missed detections; N-expert estimates false alarms.

3.2 Disparity Map

Disparity between stereo images creates the illusion of depth, and manipulation of depth scales may create a visually stunning experience for 3D movie. In order to apply pixel-by-pixel transformations, we need a dense disparity map, which leads to the fundamental subject of stereo correspondence in computer vision. Stereo algorithms based on local correspondences are typically fast, but require a proper choice of window size. Meanwhile, poorly-textured and ambiguous surfaces cannot be matched consistently. Global algorithms impose smoothness constraints on disparities in the form of regularized energy functions, such as method based on Markov random field (MRF), graph cuts, belief propagation and mean-shift segmentation. However, they generally require large computational efforts and high memory capacities. Some algorithms involve time-consuming sub-pixel refinement dealing with discontinuities.

In the field of 3D reconstruction, it is widely accepted that recovering accurate and dense range data only from an image pair is very difficult. In this system, a synthesis of methods is designed to generate usable disparity map as depth cues at a low computational cost. First, we use a generative Bayesian probabilistic approach for stereo matching, called efficient large-scale stereo (ELAS) [Geiger et al. 2011], which computes accurate disparity maps of high resolution images at frame rates close to real time. By computing a piecewise linear function induced by disparities and triangulated mesh of a set of robustly matched support points from Sobel filter responses, it builds a prior to disambiguate the matching problem and automatically searches the disparity range. Since this method causes edge blurring and performs poorly on textureless regions, we enhance the disparity map by a noise-aware edge-preserving bilateral filter, termed noise-aware filter for depth up-sampling (NAFDU) [Chan et al. 2008], which adaptively weighs the RGB intensity map $\tilde{I}_p$, $\tilde{I}_q$ and disparity map I_p, I_q at positions p, q, and consequently dampens the influence of edge blurring and texture copying. The

mathematical description of this image joint filter is

$$\tilde{S_p} = \frac{1}{k_p} \sum_{q_\downarrow \in Q} I_{q_\downarrow} f(||p-q||)[\alpha(\Delta_\Omega)\, g(||\tilde{I_p} - \tilde{I_q}||) + (1-\alpha(\Delta_\Omega))h(||I_{p_\downarrow} - I_{q_\downarrow}||)], \quad (1)$$

where $\tilde{S_p}$ is the updated disparity value, k_p is a normalization factor, and $\alpha(\Delta_\Omega)$ is the weight estimated in a local window. f, g, h are all Gaussian functions.

ELAS with NAFDU for stereo matching can achieve good performance with significant speedup. Ambiguities on the correspondences are reduced, although textureless surfaces such as sky and ground still need to be filled. We assume these areas usually have a stable location over time, so we can also adopt a less efficient but more precise algorithm for several frames to make up for the holes, named cross-scale cost aggregation (CSCA) [Zhang et al. 2014]. Inspired by human visual experience when processing stereoscopic correspondence across multiple scales, this stereo matching framework generally takes four steps: matching cost computation, cost aggregation (patch matching), disparity computation and disparity refinement. Cost volume can be computed from any stereo algorithm and then be aggregated through different scales.

So far we have introduced our real-time approaches in the preprocessing step. How attention area and disparity map are used to enhance stereoscopic effects is the key geometric problem of our framework, and will be explained in detail in the next section.

4 Image Warping

4.1 Geometric Model

In stereo display system, camera intrinsics, viewing location, projector-screen configuration, psychological factors, and their combination all play a critical role in creating visual experience, especially for depth perception. Inappropriate settings may lead to dizziness and eye strain. Over the years, the communities of 3D filmmakers and photographers have learned various heuristics for enhancing well-known stereo effects such as pinching and gigantism. As shown in Figure 3, the geometric assumption is a rectified stereo setup with the eyes represented as pinhole cameras with parallel optical axes, which is suggested as a conservative predictor of what humans can actually fuse [Held and Banks 2008]. Under this geometric framework, it becomes possible to describe horopter consistency using mathematical equations.

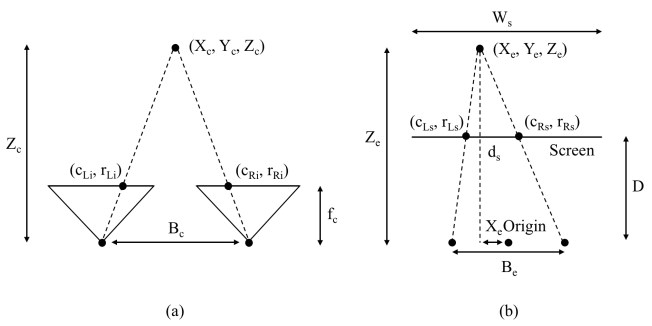

Figure 3: *Geometric framework of our approach. (a) A rectified stereo setup. (b) Eyes represented as pinhole cameras with parallel optical axes.*

All the necessary constants and variables are assumed to share the same units, except that image width, coordinates, and disparity are specified in pixels. Denote human's binocular distance by B_e, then the world coordinates are centered between the viewer's eyes, so the left and right eyes have a position of $\left(-\frac{B_e}{2}, 0, 0\right)$ and $\left(\frac{B_e}{2}, 0, 0\right)$ respectively. Let (X_c, Y_c, Z_c) and (X_e, Y_e, Z_e) be real world and perceived coordinates. First we want to investigate the relation between screen disparity d_s and the perceived 3D locations. Denote the viewer screen distance by D. According to similar triangles, the base ratios $\frac{d_s}{B_e}$ should equal to the height ratios $\frac{Z_e - D}{Z_e}$, thus we get perceived depth

$$Z_e = \frac{DB_e}{B_e - d_s}. \quad (2)$$

Similarly, from the viewer's perspective, X_e can be calculated from the left triangle created by dropping the normal, which is an analogous manner to compute Y_e with screen height H_s given. We equate the base ratios $\frac{\frac{W_s}{2} - c_{Ls} - X_e}{\frac{B_e}{2} - X_e}$ and the height ratios $\frac{Z_e - D}{Z_e}$, or the base ratios $\frac{-\frac{W_s}{2} + c_{Rs} + X_e}{\frac{B_e}{2} + X_e}$ and the height ratios $\frac{Z_e - D}{Z_e}$ for the right triangle, and get

$$X_e = \frac{B_e}{2} - \frac{Z_e}{D}\left(\frac{B_e}{2} - \frac{W_s}{2} + c_{Ls}\right), \quad (3)$$

$$Y_e = \frac{Z_e}{D}\left(\frac{H_s}{2} - r_{Ls}\right). \quad (4)$$

Reversely, if we want to modify the perceived 3D coordinates to $\left(\bar{X}_e, \bar{Y}_e, \bar{Z}_e\right)$, which are the adjusted coordinates according to horopter consistency, the corresponding screen coordinates $\bar{c}_{Ls}$, $\bar{c}_{Rs}$, and screen disparity $\bar{d}_s$ should be computed as follows (note that $\bar{r}_{Rs} = \bar{r}_{Ls}$ because there should be no vertical disparity)

$$\bar{c}_{Ls} = \frac{W_s}{2} + \frac{D}{\bar{Z}_e}\left(\frac{B_e}{2} - \bar{X}_e\right) - \frac{B_e}{2}, \quad (5)$$

$$\bar{c}_{Rs} = \frac{W_s}{2} - \frac{D}{\bar{Z}_e}\left(\frac{B_e}{2} + \bar{X}_e\right) + \frac{B_e}{2}, \quad (6)$$

$$\bar{r}_{Ls} = \frac{H_s}{2} - \frac{D}{\bar{Z}_e}\bar{Y}_e. \quad (7)$$

4.2 Virtual Rig Modification

In the next step, we discuss definitions of four stereoscopic parameters that users are allowed to modify: camera baseline B_c, dolly Z_c, camera FOV θ_c, and horizontal image translation V_c. Perceived depth is basically decided by these four parameters, which is a similar idea of ratio modification from Koppal et al. [Koppal et al. 2010] with several mathematical mistakes corrected. In a typical stereographic system, we give interaxial distance and convergence another name, baseline and translation (to simulate changes of the inward angle). As proportional representations, $B_c = \alpha_B B_{c0}$, $Z_c = \alpha_Z Z_{c0}$, $\tan\left(\frac{\theta_c}{2}\right) = \alpha_\theta \tan\left(\frac{\theta_{c0}}{2}\right)$, $Z_{c0} = \frac{B_{c0} W_s}{2B_e \tan\frac{\theta_{c0}}{2}}$. α_B is the change ratio of original camera baseline B_{c0}. α_Z is the "normalized" dolly using the unit distance Z_{c0}. Z_{c0} is computed as a function of the viewer to screen depth as reprojected in camera space, by simply scaling screen width W_s in the eye diagram by the ratio $\frac{B_c}{B_e}$. Applying changes of camera baseline and dolly, we can find a new set of perceived coordinates

$$\left(\bar{X}_e, \bar{Y}_e, \bar{Z}_e\right) = \left(\frac{X_e}{\alpha_B}, \frac{Y_e}{\alpha_B}, \frac{Z_e + \alpha_Z D - D}{\alpha_B}\right). \quad (8)$$

α_θ scales the image about its center. Unlike changes in baseline or dolly which need the scene to be re-rendered, changes in FOV

and horizontal translation just need resizing and shifting the images respectively without their disparity maps. The new screen coordinates c'_{Ls}, c'_{Rs} ($= c'_{Ls} + d'_s$), and warped screen disparity d'_s are

$$c'_{Ls} = \alpha_\theta \left(\bar{c}_{Ls} - \frac{W_s}{2} \right) + \frac{W_s}{2} - \frac{V_c}{2}, \quad (9)$$

$$d'_s = \alpha_\theta \bar{d}_s + V_c. \quad (10)$$

Based on our horopter model, the attention area should have a disparity of nearly zero, thus can make human eye's focal length stay consistent with binocular vergence. Once the original video stream goes through the preprocessing modules, a region of interest can be determined with average disparity offset calculated. Given $d'_s \approx 0$ for pixels in this region, we can generate parameters and apply disparity re-rendering as long as the scene remains in the comfort zone.

5 Experiment and Conclusion

To evaluate our method, we shot four scenes using the TenYoun ORRO stereoscopic rig. Original and enhanced videos of the first scene are shown in Figure 4, where we can see the disparity of the salient target is adjusted to zero in enhanced videos. That by manipulating baseline is even more perceptually friendly because other parts in the video fall within a reasonable disparity range. We also conducted a subjective experiment of visual perception by analyzing evaluations from 40 participants (17F, 23M, 18 to 40 years old). For each scene, they were asked to score three videos in random order based on stereo comfort with standard red-cyan glasses. To emphasize the difference, video scores adopt the three-point style and those of the same scene have to be different. Figure 5 illustrates the result and proves the effectiveness of our approach.

In this paper, we propose a system to enhance the perceptual stereoscopic effects based on psychological concepts and multimedia techniques. According to accommodation-convergence consistency, we refine such a disparity map that the attention area is fused into a single image and other parts fall in the comfort zone. Our horopter-based system is of great importance as the 3D film and VR industry is ushering in resurgence. In the future, our framework will be promoted to cinema-specific situations. The scale of user study will also be expanded to verify the psychological view. Moreover, eye tracking can be performed on each individual, so a more accurate effect will be generated with the gaze information.

Figure 4: *An example of original and perceptually enhanced videos, where a motorbike driver is set as the attention area. From left to right, we use three keyframes to represent each video. The first row is the original video; the last two rows are enhanced by manipulating translation and baseline in line with our hypothesis.*

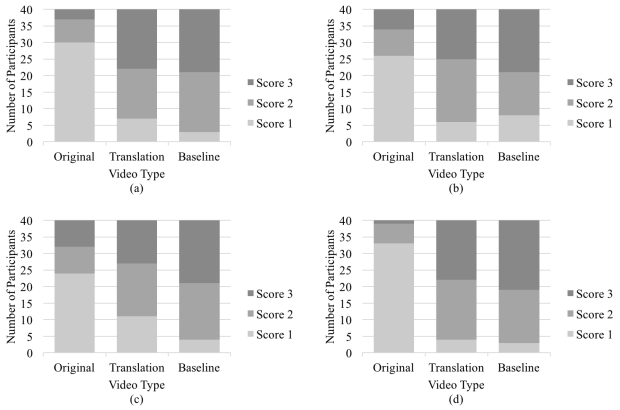

Figure 5: *The subjective experiment result of visual perception. The vertical axis stands for the number of participants, while the horizontal axis stands for different video types. The darker the columns are, the higher the scores. More people consider enhanced videos to have better stereo effects. (a) for the above example. Corresponding videos can be found in the supplementary material.*

References

ANDERSSON, B. 2015. *The DSLR filmmaker's handbook: real-world production techniques*. John Wiley & Sons.

BURT, P., AND JULESZ, B. 1980. Modifications of the classical notion of panum's fusional area. *Perception 9*, 671–682.

CHAMARET, C., GODEFFROY, S., LOPEZ, P., AND LE MEUR, O. 2010. Adaptive 3D rendering based on region-of-interest. In *IS&T/SPIE Electronic Imaging*, International Society for Optics and Photonics.

CHAN, D., BUISMAN, H., THEOBALT, C., AND THRUN, S. 2008. A noise-aware filter for real-time depth upsampling. In *Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*.

GEIGER, A., ROSER, M., AND URTASUN, R. 2011. Efficient large-scale stereo matching. In *Asian Conference on Computer Vision*. Springer, 25–38.

HELD, R. T., AND BANKS, M. S. 2008. Misperceptions in stereoscopic displays: a vision science perspective. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization*, ACM, 23–32.

KALAL, Z., MIKOLAJCZYK, K., AND MATAS, J. 2012. Tracking-learning-detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 34*, 7, 1409–1422.

KELLNHOFER, P., DIDYK, P., MYSZKOWSKI, K., HEFEEDA, M. M., SEIDEL, H.-P., AND MATUSIK, W. 2016. GazeStereo3D: seamless disparity manipulations. *ACM Transactions on Graphics 35*, 4, 68.

KOPPAL, S. J., ZITNICK, C. L., COHEN, M. F., KANG, S. B., RESSLER, B., AND COLBURN, A. 2010. A viewer-centric editor for 3D movies. *IEEE Computer Graphics and Applications 1*, 20–35.

PALMER, S. E. 1999. *Vision science: Photons to phenomenology*. MIT Press.

ZHANG, K., FANG, Y., MIN, D., SUN, L., YANG, S., YAN, S., AND TIAN, Q. 2014. Cross-scale cost aggregation for stereo matching. In *Computer Vision and Pattern Recognition, IEEE Conference on*, IEEE, 1590–1597.