DISPARITY REMAPPING FOR HANDHELD 3D VIDEO COMMUNICATIONS

Stephen Mangiat and Jerry Gibson

Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106
{smangiat, gibson}@ece.ucsb.edu

ABSTRACT

With the advent of glasses-free autostereoscopic handheld displays, a number of applications may be enhanced by 3D perception, particularly mobile video communications. On handheld devices, front-facing stereo cameras can capture the two views necessary for 3D display. However, the short distance between the user and cameras introduces large disparities and other stereoscopic challenges that have traditionally plagued close-up stereo photography. Maintaining both viewer comfort and 3D fusion under varying conditions is therefore a technological priority. In this paper, we discuss the main stereoscopic concerns of handheld 3D video communications and outline a new post-processing technique to remap on-screen disparities for viewer comfort.

Index Terms— 3D Video, Disparity Remapping, Stereoscopy, Mobile Videoconferencing

1. INTRODUCTION

Stereoscopy enhances the realism of images and video by presenting different views to each eye, creating an illusion of depth. The brain can reconstruct 3D volumes from 2D imagery using a variety of monoscopic depth cues (often shaped by prior experience). However, stereoscopy eases this process by directly presenting a fundamental optical depth cue [1]. Views of faces can be dramatically enhanced by 3D, as the mind naturally expects faces to exhibit particular structure and depth features.

Traditional 3D displays are unsuitable for video communications because they require glasses, yet handheld autostereoscopic displays eliminate this hurdle. There now exists a new opportunity for realistic communications, using handheld systems that can comfortably display 3D video of a user's face. The key difficulty is finding a balance between viewing comfort and 3D perception.

Stereoscopy stimulates eye *convergence* (the brain processes retinal disparities and converges the eyes to fuse a particular depth). When viewing real-world objects, convergence is directly linked to *accommodation* (pupils adjust to focus light from the desired depth). However, when viewing imagery on a stereoscopic display, this link is broken because the eyes must always focus light at the distance of the display, regardless of their convergence angle. This disconnect is a main source of discomfort and eye fatigue [2], and it is inherent to all stereoscopic systems that use planar screens.

The discomfort caused by this "vergence-accomodation conflict" can be mitigated by shrinking the stereo baseline to limit disparities, which unfortunately means reducing the 3D effect. When the stereo baseline is too small, the resultant video will exhibit the "cardboard cutout" effect, and will not capture 3D structure *within* the user's face [1]. In this case, stereo cameras would fail to enhance immersion.

In Sec. 2, we discuss stereoscopic concerns that help determine the placement of two front-facing cameras for 3D video communications. Once camera positioning is fixed, post-processing methods may be used to remap disparities for comfortable viewing. Prior methods are not immediately applicable to video communications due to artifacts such as depth discontinuity and warping, in addition to complexity constraints. In Sec. 4, we introduce a new method to shift disparities using a rough depth estimate and knowledge of depths within a face. Sample results are shown in Sec. 5, followed by conclusions and future work in Sec. 6.

2. STEREOSCOPIC VIEWING COMFORT

Problems contributing to stereoscopic viewing discomfort include ghosting/crosstalk, misalignment, vertical disparities, temporal discontinuities, and the vergence-accomodation conflict [1]. Video "quality" is also heavily dependent upon the unique perception and anatomy of individual users. Stereographers typically rely on rules of thumb, resulting in ambiguity throughout the literature. However, recent advances in the science of stereoscopic viewing provide a foundation for new stereo video guidelines [2], [3].

2.1. The Zone of Comfort

The vergence-accomodation conflict arises from the difference between vergence distance (in front of or behind the screen) and viewing distance (always on the screen). Figure 1 shows the relationship between these distances, both measured in diopters D (the inverse of distance in meters) [3].

This work has been supported by Huawei Technologies Co. Ltd. and HTC Corporation.

The center diagonal represents standard viewing (no conflict). The surrounding lines marked "near" and "far" represent the largest conflict that can be comfortably viewed, either behind the screen (far) or in front of the screen (near). The resultant region between these lines is referred to as the "zone of comfort" (Percival and Sheard [3]).

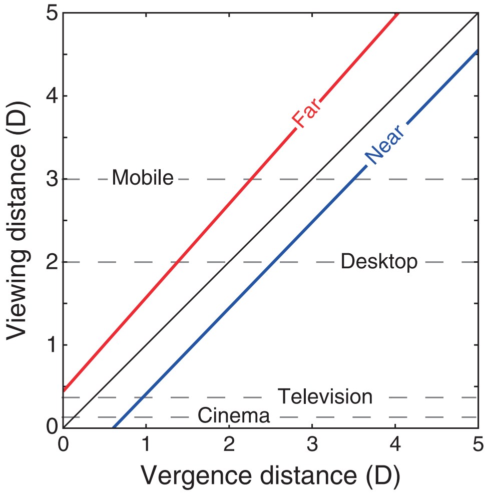


Fig. 1. Zone of Comfort [3]

Typical viewing distances for different sized displays are also marked in Fig. 1. For mobile devices, the average viewing distance shown here is 3 diopters (33 cm), however for our analysis, we use a nearer viewing distance of 30 cm. Figure 1 illustrates that a typical viewer can comfortably fuse disparities depicting 3D objects between 26 cm and 39 cm away. Given an interocular distance of 65 mm, these depths correspond to on-screen disparities of 10 mm (crossed) and 15 mm (uncrossed). As depicted in Fig. 2, crossed disparities appear in front of the display and uncrossed disparities appear behind the display. For mobile devices, Shibata et al. [3] found that objects that appear in front of the display are less comfortable to view than those that appear behind the display (the opposite is true for larger displays).

2.2. The Stereoscopic Window

In addition to disparity range, the stereoscopic window (aka proscenium rule [4]) has a significant effect on viewing comfort. A 3D scene reproduced by stereoscopic images is viewed through a window defined by the edges of the display. Information behind the window edges is missing, producing conflicts that the brain cannot resolve. As a rule, objects in front of the screen should never cross the left or right edges of the window (except in brief bursts). It may be possible to "bend" the stereoscopic window at the top and bottom edges, however when viewing closeups of a face, stereographers advise that the top of the head should never appear cut off while in front of the display [1]. It is difficult to completely avoid a scenario where the user's head is cropped, and the neck/shoulders will always reach the bottom and sides of the screen. It is therefore evident that scene depth should only be placed on and **behind the screen** during a 3D video call.

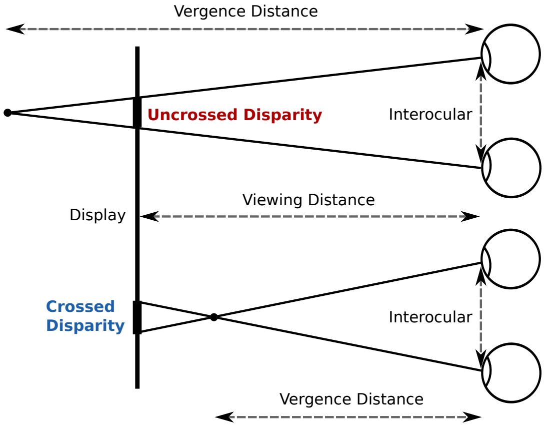


Fig. 2. Uncrossed and Crossed On-Screen Disparities

3. CONTROLLING DISPARITY

The factors described in Sec. 2 provide guidelines for the range of disparities that can be comfortably viewed on a handheld stereo display. Methods to control disparity can then be divided into two categories: (1) stereo camera/display setup and (2) post-processing.

3.1. Camera Convergence

One option to remove disparity at a particular depth is to "toe-in" the stereo camera, yet this is impractical since the optimal depth varies. Alternatively, when the optical axes are parallel, the convergence depth is placed at infinity (objects at infinite depth will appear in the same location in each image). All objects will appear in front of the display, so images must be shifted in post-processing. Yet, since disparity goes to zero as depth increases, the largest disparity is now determined by the depth of the closest object, i.e. the user's face.

3.2. Camera Baseline

For parallel cameras, the disparity (d) of an object is proportional to baseline (distance between the cameras), with

$$d = f\frac{b}{Z}, \quad (1)$$

where f is camera focal length, Z is object depth, and b is the baseline (pinhole camera model). It is often desirable to use a baseline equal to the human interocular distance (65 mm), yet this is not always the case. A rule of thumb used by stereographers is that the baseline should be at most 3% of the distance to the nearest object [1]. In a handheld scenario, with the user's face about 300 mm from the cameras, the baseline would be only 9 mm! However, since the amount of depth is also proportional to baseline, the scene will appear more two-dimensional as the baseline is decreased.

In fact, the difference between recording and viewing geometries introduces depth distortion. In order to accurately preserve 3D shape, a camera/display system must adhere to the depth consistency rule,

$$\frac{b}{Z} = \frac{b'}{Z'}, \tag{2}$$

where b and Z are the baseline and depth in camera recording space, and b' and Z' are the baseline and depth in display space [4]. If two users communicate using the same device, then Z and Z' are approximately equal (depending on arm length). However, b must be significantly smaller than b', the interocular distance. This means that the roundness factor (ratio between $\frac{b}{Z}$ and $\frac{b'}{Z'}$) is much less than one, leading to the "cardboard effect", where objects appear flat. As such, if the stereo baseline is too small, the only depth that may be perceived is the difference between the user's face and the background, and there will be no depth across facial features.

The camera baseline should therefore be *maximized*, so that the maximum on-screen disparity is at least equal to the maximum "far" disparity defined by the zone of comfort. The disparity given by Eq. (1) is converted to on-screen disparity using the ratio of sensor pixel size to display pixel size. For example, a Nintendo 3DS using a standard VGA sensor (f = 1.34 mm) has a sensor pixel size of .0022 mm and a display pixel size of .1918 mm [5]. For a viewing distance of 300 m, we can estimate that the point nearest to the camera (the nose) is about 225 mm away. Using Eq. (1), the maximum baseline that produces a comfortable range of disparities for a face at this depth is 29 mm, which is smaller than the outward facing stereo camera baseline of the device (35 mm).

3.3. Post-Processing

As discussed in Sec. 3.1, a parallel camera configuration limits the maximum disparity, yet all objects appear in front of the display. Consequently, disparities must be *remapped* to move the scene depth behind the display. Most commonly, "shift convergence" shifts the images to line up features at the desired zero disparity depth. Often done manually, current 3D handhelds can automatically shift by minimizing the difference between the entire images. This method typically converges near the largest on-screen object, and is prone to jitter as the scene changes.

Nonlinear disparity remapping adjusts different disparities by varying amounts. Kim et al. [6] present a method to reconstruct stereoscopic imagery for visual comfort using a dense disparity map. The most common disparity is reassigned to zero, while the maximum disparity can also be constrained [6]. This method relies heavily on the accuracy of the depth map. It also requires an inpainting method to fill missing regions introduced by occlusions.

A second class of nonlinear disparity remapping algorithms bypasses the need for a dense disparity map. In [7], a sparse set of correspondences and pixel importance metrics are used to compute a deformation of the input views in order to meet target disparities. Although it attempts to warp the images only in smooth and unimportant regions, this method may introduce distortions that would counteract any perceptual advantages of 3D. These methods are also ill suited for real-time applications. As such, we investigate a new shift convergence algorithm informed by the unique stereoscopic constraints of handheld 3D video communications.

4. DISPARITY REMAPPING

In order to maximize both depth perception and viewing comfort during a 3D video call, **the object nearest to the cameras should always be placed on the screen**. This will normally correspond to the tip of the user's nose, yet this may vary. The nearest depth could be found by calculating a dense disparity map and returning the largest disparity. A sparse set of feature correspondences may miss the nearest object. Complexity is a concern, yet more importantly, these methods are susceptible to noise and temporal discontinuity.

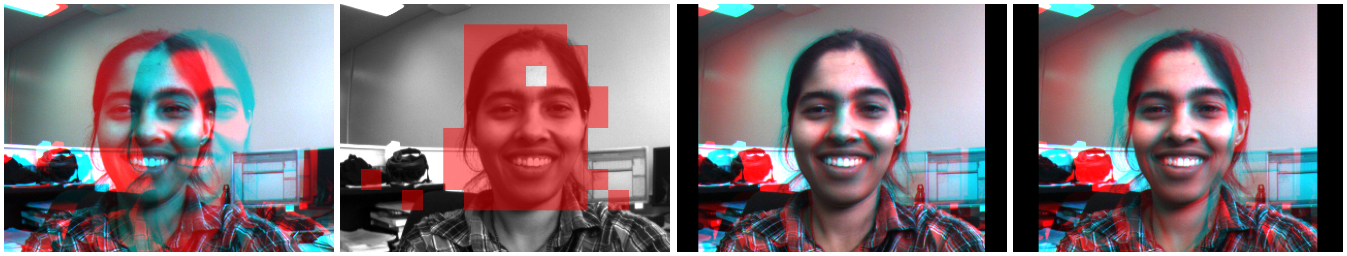
The first step in our disparity remapping method is a rough segmentation of foreground and background. This can potentially be done using face detection, yet we use template-based stereo matching since face detection can fail when the user turns his/her head. The input images are first downsampled by a factor of 8. The sum of absolute differences (SAD) between blocks in the left and right downsampled images is then minimized for a range of horizontal disparities (the images may be rectified, yet this is not required). Blocks that are too smooth for reliable matching (standard deviation less than a threshold), are skipped and assigned a disparity of zero.

The segmentation step does not need to be exact, as it only attempts to locate a majority of foreground pixels. A foreground mask is returned by selecting blocks with disparity greater than a threshold and upsampling the result. The next step is to calculate the mean absolute difference (MAD) between foregrounds in the *full-sized* images, as the amount of shift is varied. Robustness is gained by calculating the MAD over a larger number of pixels. The segmentation step is necessary since high contrast features in the background may dominate, even if the user's face fills much of the screen. The shift that minimizes the foreground MAD minimizes the average disparity across the entire foreground.

At this point we have a reliable estimate of average foreground disparity, but not the disparity of the nearest point (i.e. the nose). To account for this we use knowledge of the scene, and add a shift adjustment that accounts for the depth of a face. Using Eq. (1), we can calculate the change in disparity Δd for a change in depth ΔZ as

$$\Delta d = \frac{1}{\frac{1}{d} - \frac{\Delta Z}{fb}} - d, \tag{3}$$

where f is focal length, b is baseline, and d is the original disparity (all in mm). All disparities must be converted from



(a) Input Images (b) Foreground Mask (c) Shifted (d) Shifted with Adjustment

Fig. 3. Disparity Remapping Results (images best viewed in color with red-blue anaglyph glasses)

millimeters to pixel count using the pixel size of the camera sensor (and vice versa). In our tests, we set ΔZ to be 120 mm. The final adjusted shift is then $\lceil d + \Delta d \rceil$.

This procedure calculates the best shift for the first frame. For subsequent frames, we utilize previously calculated shifts in order to maintain temporal smoothness. Studies show that humans can tolerate changes in convergence, as long as the speed of change is limited [1]. In our tests with the Nintendo 3DS and HTC Evo 3D devices, sharp jitter in shift convergence is perceived as flickering and leads to eye fatigue. In order to enforce smoothness, the MAD search within the foreground of the full-sized images is done only for the previous shift (without adjustment) plus or minus one pixel. Furthermore, to reduce jitter we add a slight bias to the previous shift by subtracting a small constant from its MAD (.25 in tests). The algorithm is re-initilialized if no foreground is detected.

5. RESULTS

Results of disparity remapping for a sample stereo frame are shown in Fig. 3 using red-blue anaglyphs, though these images are best viewed on handheld 3D devices[1]. Figure 3 (a) shows the input left/right images without any shift. Here, the entire scene lies in front of the display and is impossible to fuse. The segmented foreground (blocks highlighted in red) is illustrated in Fig. 3 (b). Note that segmentation is imperfect (several background blocks are marked in red), yet a majority of foreground pixels are correctly labeled, significantly reducing the impact of background pixels on MAD calculations.

Figure 3 (c) shows the shifted output, without adjusting for the depth within the foreground. Even though average disparity across the foreground is minimized, this output is unsuitable due to crossed disparities that remain across the front of the face (disparities are represented by visible shades of blue or red). These disparities will appear in front of the display and are difficult to view. Finally, Fig. 3 (d) shows the final shifted output including the adjustment. Now the images converge onto the nose and the rest of the scene recedes behind the display, providing the most comfortable viewing experience on a handheld device. Even if the user's face is cropped by the frame edge, the stereoscopic window will not be violated since the entire frame lies on or behind the screen.

6. CONCLUSIONS

We have described an exciting new application for 3D handheld devices, and outlined some important stereoscopic challenges posed by 3D video communications. The proposed disparity remapping algorithm automatically eliminates uncomfortable crossed disparities, and avoids breaking the stereoscopic window by shifting the entire scene onto and behind the display. With respect to the tradeoff between depth perception and viewing comfort, tests need to be performed to choose the best stereo baseline to provide sufficient depth within the face. Since a user's gaze is likely fixed to the other person's face, it may in fact be possible to introduce disparities in the background beyond the zone of comfort.

7. REFERENCES

[1] B. Mendiburu, *3D Movie Making: Stereoscopic Digital Cinema from Script to Screen*, Elsevier Science, 2009.

[2] D.M. Hoffman, A.R. Girshick, K. Akeley, and M.S. Banks, "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *Journal of Vision*, vol. 8, no. 3, pp. 33, 2008.

[3] Takashi Shibata, Joohwan Kim, David M. Hoffman, and Martin S. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *Journal of Vision*, vol. 11, no. 8, 2011.

[4] R. Ronfard and G. Taubin, *Image and Geometry Processing for 3-D Cinematography*, Geometry and Computing. Springer, 2010.

[5] "Nintendo 3DS hardware specs," http://www.nintendo.com/3ds/hardware/specs, 2011.

[6] Hye Jin Kim, Jae Wan Choi, An-Jin Chaing, and Ki Yun Yu, "Reconstruction of stereoscopic imagery for visual comfort," 2008.

[7] Manuel Lang, Alexander Hornung, Oliver Wang, Steven Poulakos, Aljoscha Smolic, and Markus Gross, "Nonlinear disparity mapping for stereoscopic 3d," *ACM Trans. Graph.*, vol. 29, no. 3, pp. 10, 2010.

[1] For videos, please visit http://vivonets.ece.ucsb.edu/handheld3d.html