BI-LAYER DISPARITY REMAPPING FOR HANDHELD 3D VIDEO COMMUNICATIONS

Stephen Mangiat, Kuo-Chin Lien, and Jerry Gibson

Department of Electrical & Computer Engineering, University of California, Santa Barbara, USA
{smangiat, kuochin, gibson}@ece.ucsb.edu

ABSTRACT

Handheld devices with "glasses-free" autostereoscopic displays present a new opportunity for 3D video communications. 3D can enhance realism and enrich the user experience, yet it must be employed without visual discomfort. A simple shift-convergence disparity remapping technique can align a user's face throughout a 3D video call, eliminating uncomfortable crossed disparities. However, this can produce large disparities in the background that the viewer is unable to fuse. Furthermore, reducing camera separation and thus all disparities may lead to a flat appearance that does not aid realism. Using foreground/background segmentation, we propose a novel bi-layer disparity remapping algorithm to limit uncomfortable background disparities during handheld 3D video communications. A user study with the HTC Evo 3D handheld device shows that this method improves visual comfort while preserving the critical depths within the face.

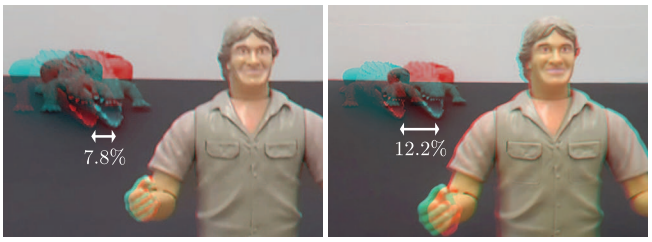
Index Terms— 3D video communications, 3D viewing comfort, disparity remapping

1. INTRODUCTION

Glasses-free 3D handheld devices can enable 3D video communications, using a front-facing stereo camera adjacent to the autostereoscopic display [1]. Stereoscopy produces a 3D illusion by displaying separate images to the viewer's left and right eyes [2]. The glasses often cited as the main dissatisfaction with 3D cinema are impractical for video communications because they change the appearance of participants. Using a glasses-free device, depths within facial features can enhance realism and the quality of experience [3].

Viewing comfort is critical using a 3D handheld device. The "vergence-accomodation conflict" arises from the difference between vergence distance (possibly in front of or behind the screen) and viewing distance (always on the screen) [4]. If this difference is too large, the viewer will experience discomfort and fatigue. Since viewing distance is constrained to within an arm's length, this conflict is controlled by limiting on-screen disparities.

Close-up stereo photography is particularly difficult due to the *range* of disparities that result from depths within the scene. Changes in depth correspond to larger disparity differences for nearer objects. Shift-convergence can successfully converge onto the close-up object and minimize disparities in the foreground. However, background objects will now have very large uncrossed disparities, with a maximum magnitude equivalent to the size of the shift [1].



(a) (b)

Fig. 1. (a) Placing the camera farther from the nearest object ($2'$) produces a smaller disparity range. (b) Placing the camera nearer to the nearest object ($1'$) produces a larger disparity range. Disparities are measured with respect to image width.

This effect is illustrated by Fig. 1. Two objects are placed one foot apart and imaged using a stereo camera. In Fig. 1 (a), the cameras are approximately two feet from the near object. In Fig. 1 (b), the cameras are one foot from the near object. Each image pair is shifted to converge onto the near object, and scaled so the apparent size of the near object is constant. Moving the cameras closer to the objects increases the size of the foreground compared to the background, however the disparity of the background object increases from 7.8% to 12.2% (measured with respect to image width).

Nonlinear disparity remapping applies varying horizontal shifts to each pixel, typically based on a depth map estimated beforehand [5, 6]. By carefully shifting pixels, the disparity range of the video can be remapped into a "zone of comfort" to tailor different viewing conditions. Although this operation is sound in principle, the quality of the resulting stereoscopic images heavily depends on the accuracy of the depth map. After nonlinear shifting, stereoscopic inpainting [7] is needed to fill holes that were occluded by foreground objects, a process that also demands an accurate depth map.

Other nonlinear disparity remapping techniques attempt to sidestep the hole-filling process and the estimation of a dense and perhaps noisy depth map. Inspired by work on

content retargeting, Lang et al. [8] proposed a warping-based method to manipulate disparity. Instead of estimating the disparity value for every pixel, the authors proposed to detect a sparse yet robust set of correspondences between the two views. Shifting is then only applied to the corresponding points. With these points as anchors, the remaining pixels in the two views are filled by image warping. A drawback of this approach is that it requires flexibility to warp images in non-salient regions. The resulting images may be inaccurate in geometry, and spatial distortion is noticeable when there are vertical structures present in the input image pairs.

A perceptual user study on a current 3D handheld device shows that users generally prefer 3D images with smaller disparities [9]. At the same time, reducing camera separation and thus all disparities will flatten the face, negating any sense of realism gained by 3D display [9]. Therefore, we describe a novel bi-layer disparity remapping method in Sec. 2 that preserves all depths within the foreground and compresses the background depths that are unimportant to video communications. Section 3 outlines our approach toward foreground/background segmentation, followed by sample results in Sec. 4. Evaluation through a user study is discussed in Sec. 5, with conclusions and future work in Sec. 6.

2. BI-LAYER DISPARITY REMAPPING

In a typical video communications scenario, the foreground face is undoubtedly the most meaningful part of the scene. Therefore, preserving depths within the face without distortion is the highest priority. In order to maximize viewing comfort during a 3D video call on handheld devices, the object nearest to the cameras should also be placed on the screen [1]. The left and right images must therefore adaptively converge onto the front of the user's face.

A bi-layer shifting scheme assumes that the user's face itself can only produce a limited disparity range and never exhaust the depth budget of the zone of comfort. This observation suggests that we can always display the original 3D face without any risk of visual discomfort. In contrast, the background layer is less important so we can safely (1) apply a large compression on its disparities and (2) put it at an arbitrary position behind the foreground. As objects get farther away from the stereo camera, changes in depth are less noticeable, particularly because a 3D handheld can display only a limited depth range (a few centimeters). As such, depth compression is mainly gained by reducing the gap between the foreground and the background. Based on these observations, we design the following shifting scheme to produce a perceptually satisfying and comfortable stereo pair.

To guarantee that all disparities are mapped into the zone of comfort [4], we shift the foreground so that the nearest point lies on the screen. This prevents stereoscopic window violations, which are caused by objects that appear both in front of the screen and cut off by the frame edge [10]. We then compress the entire background to a flat layer, and shift it to just behind the foreground, as illustrated in Fig. 2. In order to remap disparities, one of the views or both views can be shifted (the results shown here shift only the left image).

Three shifting parameters are first determined using a depth map and foreground mask, which are extracted using methods described in Sec. 3. We find the nearest point and the farthest points within the foreground and denote their disparity values as a^- and a^+ respectively. For video communications, a^+ is determined as the median disparity of pixels on the border of the foreground. The disparity interval $c = |a^- - a^+|$ thus corresponds to the depth range of the foreground face.

To shift the background, we initialize a new left frame by duplicating the right frame. We then horizontally shift this new left frame L' by c pixels, the foreground disparity range. This produces a comfortable background with c disparity, but the critical depths in the foreground must now be recovered.

We restore foreground disparities by shifting the isolated left foreground region FG by a^- pixels and pasting the result FG' onto L'. During this operation the borders of the two foregrounds, FG' and the foreground region in L', are roughly aligned. In fact, this shift is necessary to minimize any holes along the foreground border. This is a natural outcome for head/shoulders scenes when the camera separation is relatively small (< 30 mm), since much of the foreground border will have a disparity equal to the farthest depth in the foreground, a^+. In other words, the new foreground in the left frame will cover up the foreground from the shifted right frame when the front of the face is aligned. Therefore, the resulting holes (if any) must be small. Additionally, any holes are filled with true content copied from the right view, so they can be smoothly blended with the shifted foreground FG'. This method is summarized in Algorithm 1.

Algorithm 1 Bi-layer Disparity Remapping

Require: L and R: the left and the right images, FG: the extracted foreground of the left image, a^+: the disparity value of the farthest point within the foreground, a^-: the disparity value of the nearest point within the foreground.

Process background:

1: Assign the new left frame $L' = R$. ▷ Now both the FG and BG have zero disparity

2: Shift L' by $c = |a^+ - a^-|$ pixels. ▷ Now both the FG and BG have c disparity

Process foreground:

3: Shift FG by a^- pixels; The result is FG$'$.

4: Paste FG$'$ on L'. The result is L''. ▷ Now the FG pixels have disparities ranging from 0 to c.

return The new stereo pair L'' and R.

During a video, we also utilize previously calculated shifts in order to maintain temporal smoothness. Studies show

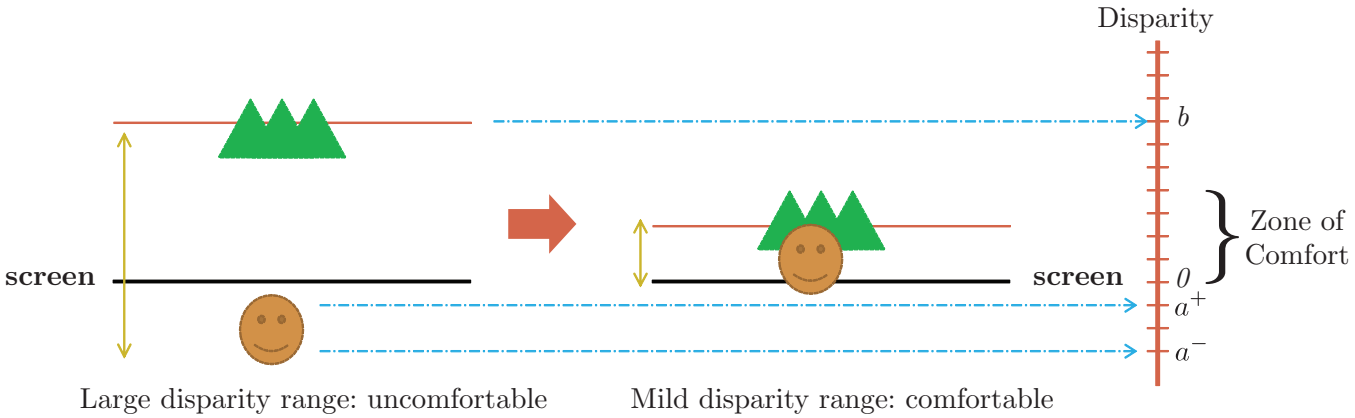


Fig. 2. A 3D video can be modeled by two independent layers: the foreground face and the background scene. A comfortable and realistic disparity configuration maintains the nearest point on the screen and the most distant point at $|a^- - a^+|$.

that humans can tolerate changes in convergence, as long as the rate of change is limited [10]. In our tests with the HTC Evo 3D device, sharp jitter in shift-convergence is perceived as flickering and leads to eye fatigue. In order to enforce temporal smoothness, we limit the change in shift between frames to two pixels. This limit applies to both the foreground shift (a^-) and background shift (c).

3. FOREGROUND/BACKGROUND SEGMENTATION

Foreground/background segmentation is a technique widely studied for video surveillance and background substitution [11]. Segmentation presumes that an image consists of two layers that can be independently processed. More precisely, the video frame can be split into the foreground face and the background scene, with some distance between them. This distance produces the large disparity range that is the root of 3D visual discomfort.

The simplest segmentation method is to threshold a dense disparity map, though this is susceptible to noise. Kolmogorov et al. [11] investigated bi-layer segmentation using Layered Dynamic Programming and a Layered Graph Cut method, which we adapt here. The foreground/background labeling problem considers multiple features for robust prediction [11]. Here, we integrate color and depth information into a conditional random field (CRF) [12] framework that predicts the best possible label for each pixel in the image by maximizing a joint probability of the multiple cues.

For the depth cue, an open-source 3D library[1] is used to generate a dense depth map. With a two-layer assumption, we can determine a threshold to separate the foreground peak from the background peak using the median of the disparity histogram. Since this assumption may not always hold, it is also possible to use a priori information of the user's viewing distance to determine this threshold. For instance, during handheld 3D video communications, the foreground must be within an arm's length of the stereo camera.

A sample stereo frame is shown in Fig. 3 (a). This image is first labeled with 1 representing a foreground pixel and 0 otherwise, using depth map thresholding. Because of the inherent deficiency of stereo matching, this rough segmentation is often noisy across textureless regions or around the foreground boundary, as seen in Fig. 3 (b). A foreground color model is then utilized for more accurate segmentation. In our current implementation, a color histogram in the YCbCr domain is trained offline. Real-time implementations may adaptively train this color model as a video call progresses. To introduce spatial coherence, we also incorporate gradient as suggested in [13]. The segmentation considering color, depth, and gradient is more reliable than thresholding the imprecise depth map, as seen in Fig. 3 (c).

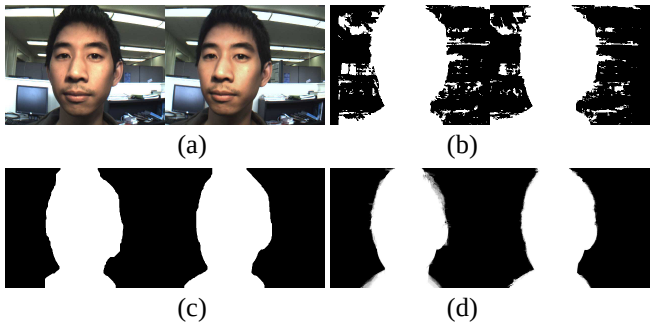


Fig. 3. Foreground/Background Segmentation: (a) Input Stereo Pair (b) Depth Thresholding (c) Graph-cut Color plus Depth Model (d) Image Matting

The foreground of the left image will eventually be blended with the background of the right image during our shifting scheme, so care must be taken to assure a seamless and accurate blend between the two layers. Image matting can be used to improve labels at the boundary and estimate

[1]Openvis3d: http://code.google.com/p/openvis3d/

the opacity α in the following equation for fuzzy or hairy objects:

$$I = \alpha F + (1 - \alpha)B, \quad (1)$$

where I is the input image composed of the foreground F and the background B.

The first step here is to relabel the image as foreground, background, or unsure. The unsure region is found by extending a 20 pixel border around the foreground/background boundary seen in Fig. 3 (c). These labels are then used to extend and erode the foreground label using Closed-Form Matting [14]. The resultant matte will contain values between 0 and 1 in the unsure region. Figure 3 (d) illustrates the improved foreground segmentation using image matting.

4. RESULTS

Bi-layer disparity remapping results[2] are shown in Fig. 4, with the input stereo pair followed by the output stereo pair. Notice here that the foregrounds in the left image have been shifted more than the background. This converges the images onto the front of the face, while limiting uncrossed background disparities. Furthermore, there are no holes in the output images, with smooth transitions between the new foreground and background regions of the shifted left images.

This is more clearly illustrated using red-blue anaglyphs comparing the output of shift-convergence and bi-layer disparity remapping, as shown in Fig. 5. The foregrounds have the exact same disparity using both methods. However, using shift-convergence, where all pixels are shifted the same amount, there are extremely large background disparities (visible in the ceiling lights and in the desktop monitor in the background). In the bi-layer output, the depths between the foreground and background are removed, while the crucial depths within the face are maintained without distortion.

5. USER STUDY

To test our algorithm, we captured images using a custom stereo camera rig. This beam-splitting rig uses two Point Grey Research Firefly MV cameras and a half-silvered mirror. The cameras are arranged such that 50% of incoming light is reflected upward towards a camera placed above the mirror, while the other 50% of the light transmits through the mirror to the second camera placed behind it. In this way, the cameras can capture stereo images with very small separations (less than 50 mm) that are not possible if the cameras are placed side-by-side.

In a single-blind viewing experiment, 22 users were asked to view images on an HTC Evo 3D device and evaluate a questionnaire. The average interocular distance was 58 mm, less than the typically reported average of 65 mm. Users were first asked whether or not they had any prior experience with a 3D handheld device, with 9 out of 22 (41%) reporting that they did have some prior experience, though this prior experience was not substantial (limited to several brief viewings).

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

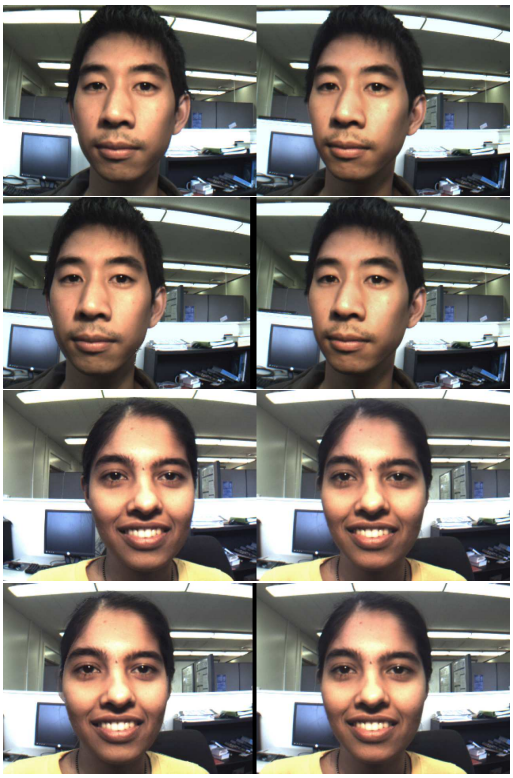
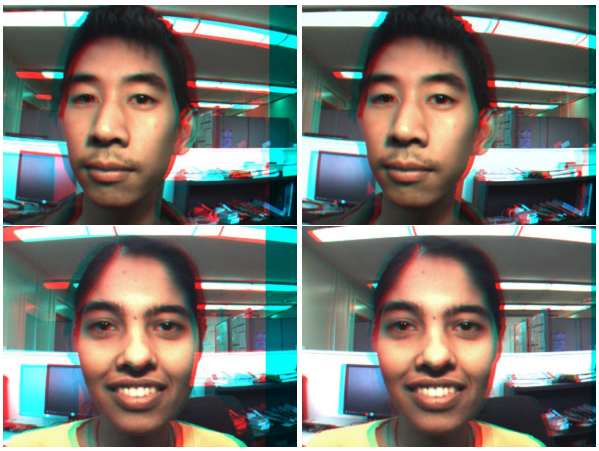


Fig. 4. Bi-layer Disparity Remapping Results (Input stereo pair on top with output stereo pair below)



(a) (b)

Fig. 5. Disparity Remapping Results: (a) Shift-Convergence (b) Bi-layer Disparity Remapping — note the foreground disparities are the same but the background disparities are significantly reduced. (Images best viewed in color with red-blue anaglyph glasses)

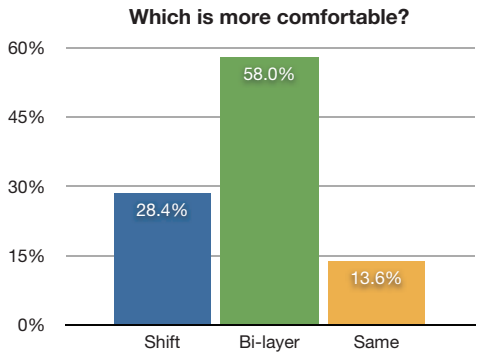


Fig. 6. Shift-convergence vs Bi-layer Disparity Remapping

Next, users were presented random images with four camera separations (10 mm, 20 mm, 30 mm, and 40 mm) and asked to indicate their preference. Users preferred 10 mm separation (59.1%) most and 20 mm (25%) second, thus preferring the least 3D depth. When isolating the results for users that had prior experience, a larger percentage chose the 20 mm separation (36.1% for 20 mm vs 55.6% for 10 mm), suggesting preference for more 3D may increase with experience [9]. For this image dataset, the optimal camera separation for realistic depths within the face was found to be about 16 mm, and users confirmed 20 mm as being the most realistic [9].

To evaluate the bi-layer method, users were asked to compare comfort for shift-converged images and images that had been processed with bi-layer disparity remapping. Here, ground truth foreground/background segmentation was used, to evaluate the perceptual effect of bi-layer remapping without potential segmentation artifacts. For this test, four image pairs were evaluated with two different faces captured at two camera separations. As seen in Fig. 6, bi-layer disparity remapping was chosen as more comfortable than shift-convergence by a significant majority (58%). While the depth within the faces was held constant, reducing disparities in the background provided a clear improvement in comfort.

6. CONCLUSIONS AND FUTURE WORK

In general, disparities need to be minimized to maintain comfortable viewing on a 3D handheld device, while maintaining realistic depths within the users face. This supports the need for a bi-layer disparity remapping technique to reduce unimportant depths in the background, even if the camera separation is small enough to ensure all depths fall within the zone of comfort. Future work must create a real-time implementation and examine the perceptual effects of errors in foreground/background segmentation. Image matting adds to the computational complexity, so future work must also evaluate the trade-offs of complexity reduction schemes such as reduced-resolution foreground segmentation.

7. REFERENCES

[1] Stephen Mangiat and Jerry Gibson, "Disparity remapping for handheld 3D video communications," in *IEEE Emerging Signal Processing Applications (ESPA)*, Las Vegas, NV, 2012.

[2] L. Lipton, *Foundations of the Stereoscopic Cinema: A Study in Depth*, Van Nostrand Reinhold, 1982.

[3] Ingo Feldmann, Oliver Schreer, Ralf Schfer, Zuo Fei, Harm Belt, and Oscar Divorra Escoda, "Immersive multi-user 3D video communication," September 2009.

[4] T. Shibata, J. Kim, Hoffman D., and M. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *Journal of Vision*, Jul 2011.

[5] C. Wang and A. A. Sawchuk, "Disparity manipulation for stereo images and video," *Proc. SPIE*, vol. 6803, 2008.

[6] H. J. Kim, J. W. Choi, A.-J. Chaing, and K. Y. Yu, "Reconstruction of stereoscopic imagery for visual comfort.," *Proc. SPIE*, vol. 6803, 2008.

[7] L. Wang, H. Jin, R. Yang, and M. Gong, "Stereoscopic inapinting: Joint color and depth completion from stereo images," *IEEE CVPR*, June 2008.

[8] M. Lang, A. Hornung, O. Wang, S. Poulakos, A. Smolic, and M. Gross, "Nonlinear disparity mapping for stereoscopic 3d," *ACM Transaction on Graphics*, 2010.

[9] S. Mangiat and Jerry Gibson, "Camera placement for handheld 3d video communications," in *Proc. Conf Signals, Systems and Computers (ASILOMAR) Record of the Forty Sixth Asilomar Conf*, 2012.

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

[11] V. Kolmogorov, A. Criminisi, A. Blake, G. Cross, and C. Rother, "Bi-layer segmentation of binocular stereo video," June 2005, vol. 2, pp. 1186 vol. 2–.

[12] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data.," *International Conf. on Machine Learning*, pp. 282–289, 2001.

[13] A. Blake, C. Rother, M. Brown, P. Perez, and P. Torr, "Interactive image segmentation using an adaptive gmmrf model," *The European Conference on Computer Vision (ECCV)*, 2004.

[14] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Transactions On Pattern Analysis And Machine Intelligence*, vol. 30, no. 2, Feb. 2008.