Camera Placement for Handheld 3D Video Communications

Stephen Mangiat
Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106
smangiat@ece.ucsb.edu

Jerry Gibson
Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA 93106
gibson@ece.ucsb.edu

***Abstract*—We investigate the effect of stereo camera separation on 3D perception and realism for handheld 3D video communications. Using a handheld device equipped with an autostereoscopic display, a front-facing stereo camera can capture left and right views of the user's face. However, consideration must be paid to the camera separation in order to balance both viewer comfort and realism. Using display-camera geometry, we illustrate the relationship between real depths in camera space and perceived depths in display space. We then derive the optimal camera separation to capture depths within a user's face that are consistent with the size of the face on a handheld display, and contrast this result with traditional rules of thumb used by stereographers. These recommendations are evaluated by a perceptual user study with a current 3D handheld device.**

I. Introduction

Glasses-free 3D handheld devices create a new opportunity for 3D video communications [1]. Such a device would utilize a front-facing stereo camera adjacent to the autostereoscopic display. 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]. However, two fundamental questions arise in this scenario: (1) how to balance viewing comfort and 3D perception, and (2) how to optimize realism?

The realism of perceptual cues is paramount when viewing stereoscopic images of faces [4]. If there is too little depth, the face will appear as a flat layer distinct from the background, producing the "cardboard cutout effect" depicted in Fig. 1 (a) [5]. If there is too much depth, the user's face will appear elongated, producing an unrealistic "Pinocchio Effect" as illustrated by Fig. 1 (b). All perceived depths must ultimately be consistent with the size of the handheld device.

Previous work has examined the limits of viewing comfort for 3D displays. 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) [6]. If this difference is too large, the viewer will experience discomfort and fatigue.

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 face 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. Previous work in [1] therefore estimated the maximum camera separation for handheld 3D video communications to produce disparities with a "zone of comfort."

(a) (b)

Fig. 1. (a) Cardboard Cutout Effect (too little depth) (b) Pinocchio Effect (too much depth)

In Sec. II, we outline the display-camera geometry for handheld 3D video communications, and analyze its implications on viewing comfort in Sec. III. Next, stereo camera separations for realistic depths within the face are derived in Sec. IV. These results are evaluated using a perceptual user study, as described in Sec. V, followed by conclusions in Sec. VI.

II. Display-Camera Geometry

Using a pinhole stereo camera model with parallel optical axes, the image sensor disparity d is

$$d = -f\frac{b}{z}, \tag{1}$$

where b is the camera separation, f is the focal length of the cameras, and z is object depth [7]. Positive on-screen disparities create the illusion that an object is placed behind the screen, while negative on-screen disparities place objects in front of the screen. From Eq. (1), the largest disparity magnitude is associated with the object nearest to the cameras, i.e. the front of a user's face. In order to adjust convergence

depth with a parallel stereo camera, the images must be shifted in post-processing by the disparity associated with the desired convergence depth (shift-convergence). The converged object will have zero disparity and objects behind it will now have positive disparities.

The relationship between camera space (real depths) and display space (perceived depths) using a sensor-to-display magnification M can be derived as

$$z' = \begin{cases} \frac{-z_d}{-Mf\frac{b}{b'}\left(\frac{1}{z}-\frac{1}{z_o}\right)-1}, & z \geq z_o \\ \frac{z_d}{\left|-Mf\frac{b}{b'}\left(\frac{1}{z}-\frac{1}{z_o}\right)\right|+1}, & z < z_o, \end{cases} \quad (2)$$

where z' is the perceived depth of the object with respect to the viewer, z is the real depth of the object with respect to the cameras, z_o is the convergence depth, z_d is the viewing distance, and b' is the viewer interocular distance. This mapping is illustrated in Fig. 2 for several camera separations. In this example, the convergence depth of the cameras (z_o) is chosen to be 250 mm, representing the distance to the nearest point from the cameras. Similarly, the viewing distance (z_d) of the other user is 300 mm, a recommended distance for handheld autostereoscopic viewing [6]. It is important to note the significant depth compression in perceived depth space. For instance, with a camera baseline of 20 mm and $z = \infty$, the perceived depth is about 326 mm, meaning objects very far from the stereo camera will appear only 26 mm behind the display.

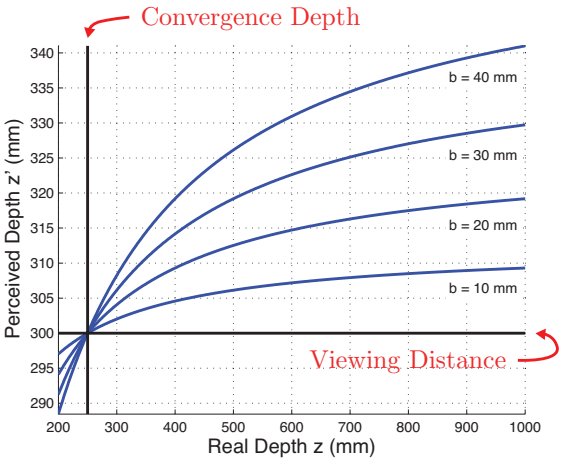

Fig. 2. Relationship between camera space and display space for different camera baselines (z_d = 300 mm, z_o = 250 mm, f = 3.94 mm, M = 16.53, b' = 65 mm)

III. Viewing Comfort

For handheld devices, Shibata et al. [6] found that objects that appear in front of the display are less comfortable to view than those that appear behind the display. Furthermore, objects that appear in front of the display at the frame edge carry harsh penalties on viewer comfort (stereoscopic window violations) [4]. Therefore, all scene depth should be placed on and behind the screen during handheld 3D video communications [1].

Using a parallel stereo camera, we can adjust only camera separation (focal length is restricted on handhelds). It seems desirable to use a separation that is equal to the human interocular distance (65 mm), yet this is not always the case. An old rule of thumb is that the separation should be at most 3% of the distance to the nearest object [2]. If the user's nose is 250 mm from the cameras, the separation would be less than 8 mm. However, the scene will look more two-dimensional as the separation is decreased. Camera separation can be *maximized* by matching the negative disparity of the nearest object to the maximum positive disparity within the zone of comfort [1].

Will maximal use of the depth budget produce the most realistic 3D effect? A second consideration is the consistency of depths with the size of the face on the handheld display [8]. If the camera separation is too large, the face will appear elongated in depth, producing the unnatural "Pinocchio effect."

IV. Camera Separation for Realistic Depths

If we consider that a user's face will occupy the maximum screen area on a handheld screen held in landscape mode, the height of the face will be approximately 50 mm (width of handheld screens). Next, if we define the height of the human face to be about 250 mm, the face-to-screen magnification (M_{face}) is roughly $1/5$.

One option is to scale depths by M_{face}, however the face should appear as though it is full-sized yet viewed from a farther distance, rather than a shrunken version held at arm's length (in the same way a toy model airplane viewed from a near distance appears different than a jumbo jet viewed from far away). Towards this, the screen can be considered a "window" into a 3D scene. When held at a distance of 300 mm to the eyes, it will appear that you are viewing a person standing about 1500 mm away. Therefore, for particular camera and display parameters, we can estimate a separation needed to create depths consistent with a real person's face at this distance.

To make an object held at 300 mm consistent with one at 1500 mm, *relative retinal disparities* must be matched, since the eyes will converge at each distance. The brain estimates the distance of objects using relative retinal disparities and a combination of vergence angle, accomodation, and head orientation [9]. It is impossible to reproduce reality entirely, since the convergence angle and accomodation between the eyes will not be the same. Still, the relative retinal disparities are a dominant depth cue that can be controlled by the autostereoscopic display. Since disparity is a nonlinear function of depth, not all of the retinal disparities relative to a convergence depth of 1500 mm can be maintained when shifted to the nearer viewing distance of 300 mm. However, it is possible to map one relative retinal disparity, and thus all disparities between zero and this value will appear close to the desired value.

Since our eyes rotate to converge at a depth of interest, we model relative retinal disparities using a toed-in camera configuration, as illustrated in Fig. 3. Retinas are spherical surfaces,

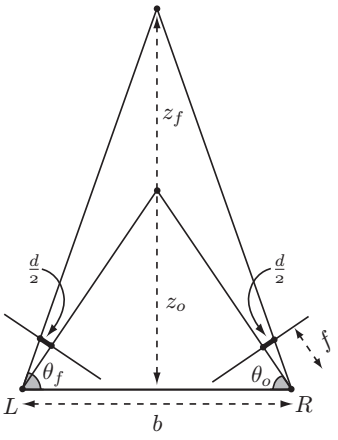

Fig. 3. Toed-in Camera Setup

however a planar model will provide a good approximation for camera placement guidelines. In Fig. 3, the eyes converge at a distance z_o, while a farther object at distance z_f produces a relative disparity d. This disparity can be estimated using the difference between the two convergence angles, θ_d, and the focal length of the eyes, f (approximately 17 mm), with

$$d = f\left(\frac{z_f - z_o}{\frac{b}{4} + \frac{z_f z_o}{b}}\right). \tag{3}$$

Using this model, we then find a relationship between two sets of depths (near and far) that have equivalent relative retinal disparities. If the viewer's eyes converge at a viewing distance of z_d, the face will appear to be viewed at a distance of $\frac{z_d}{M_{\text{face}}}$. We define the first set of depths as z_d and z_n, and the second set as $\frac{z_d}{M_{\text{face}}}$ and z_f. Setting the relative disparity between each set (d_n and d_f) equal to each other allows us to derive a relationship between z_n and z_f:

$$z_n = \frac{\frac{b}{4}\left(z_f - \frac{z_d}{M_{\text{face}}}\right) + z_d\left(\frac{b}{4} - \frac{z_f z_d}{bM_{\text{face}}}\right)}{\frac{b}{4} - \frac{z_f z_d}{bM_{\text{face}}} + \frac{z_d}{b}\left(z_f - \frac{z_d}{M_{\text{face}}}\right)}. \tag{4}$$

Using Eq. (4), we can now determine a depth z_n that has the same relative retinal disparity with respect to a convergence depth of z_d that z_f has with respect to z_{disp}/M_{face}. For example, with $z_d = 300$ mm and $z_d/M_{\text{face}} = 1500$ mm, we can choose $z_f = z_d/M_{\text{face}} + 200$ mm $= 1700$ mm. Here, 200 mm is chosen to represent the amount of depth perceived within the face that we would like to realistically map to the nearer viewing distance of 300 mm. Plugging these values into Eq. (4), we find z_n to be approximately 307.1 mm. This means that an object at 307.1 mm produces the same relative retinal disparity with respect to 300 mm that an object at 1700 mm produces with respect to 1500 mm (when interocular distance is 65 mm). In a video communications context, we would therefore like to map 200 mm behind the user's nose (which will appear on the display), to roughly 7.1 mm behind the display.

An on-screen disparity of 1.51 mm will be perceived at 307.1 mm, when viewed from a distance of 300 mm by a user with an interocular distance of 65 mm. Consequently, if we capture the user's face such that 200 mm behind his or her nose will have an on-screen disparity of 1.51 mm, we will produce retinal disparities that are consistent with a person's face viewed at 1500 mm, and thus consistent with the size of the face. To calculate the corresponding camera separation,

$$b = -\frac{d'}{Mf}\frac{1}{\left(\frac{1}{z} - \frac{1}{z_o}\right)}, \tag{5}$$

where d' is the on-screen disparity (1.51 mm), f is the camera focal length (3.94 mm), M is the screen magnification (16.53), z_o is the depth of the user's nose (250 mm), and $z = z_o + 200$ mm. For these particular camera/display parameters, the separation needed to produce retinal disparities that are consistent with the size of the face on the display is roughly 13 mm, which is smaller than the camera separation currently found on 3D handhelds (30-35 mm). This suggests that views of faces captured using the separations of current devices will appear to have exaggerated depth with respect to their size on the screen.

V. User Study

Given the same viewing conditions (viewing distance, on-screen disparities, and interocular distance), the perceived 3D effect may vary for each person. Indeed, viewers may not always prefer 3D images and video that contain the most realistic depths. As a result, we test recommendations for stereo camera placement with a user study.

A. Beam-Splitting Camera Rig

To generate data, we capture images using a custom, adjustable stereo camera rig. This beam-splitting rig using 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.

B. Questionnaire

In a blinded experiment, 22 test subjects were asked to view images on an HTC Evo 3D device and evaluate a simple questionnaire. The interocular distance of each subject was recorded, with an average of approximately 58 mm (less than the typically reported average of 65 mm). Subjects 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).

Following a sample image to help position the device in the comfortable viewing zone, users were asked to evaluate 3D images through simple multiple-choice questions. The images were presented either in sequence or simultaneously by splitting the screen in half. This questionnaire attempts to

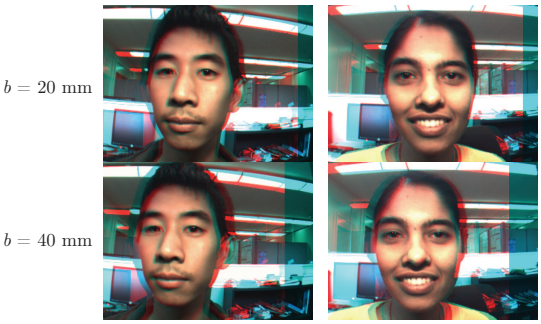

Fig. 4. Camera separation was adjusted from 10 mm to 40 mm for the same face. Each image is shifted to converge onto the person's nose.

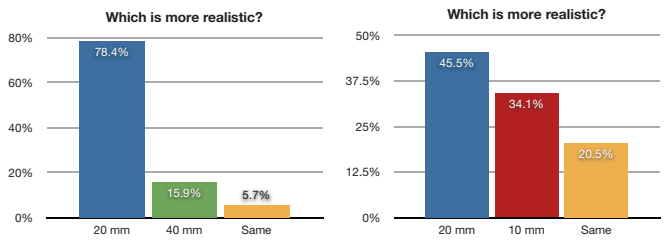

Fig. 5. Camera Separation for Realism Results: (a) 20 mm vs 40 mm (b) 20 mm vs 10 mm

determine which camera separations users prefer, and which camera separations users find most realistic.

C. Preference and Realism Results

Users evaluated four sets of images, with four images in each set corresponding to different camera separations (10 mm, 20 mm, 30 mm, and 40 mm). Between the sets, two different faces were used, processed with shift-convergence (as seen in Fig. 4) and a bi-layer disparity remapping technique that produces the same disparities in the face, while limiting uncomfortable background disparities.

Users preferred a 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% vs 55.6% for 10 mm), suggesting preference for more 3D may increase with experience.

Using the realistic depths analysis in Sec. IV, we can calculate the camera separation that will produce the most realistic depth for this dataset. In this case, the average distance of the nearest object to the cameras is 305 mm. This was calculated using the shifts found to converge each image onto the front of the nose. Furthermore, the average interocular distance of the users was 58 mm. Consequently, the estimated camera separation for realism is **16 mm**. This suggests that on average, 20 mm should appear most realistic.

To test realism, users evaluated eight images, with two faces side-by-side. In four of the images (two faces with both shift-convergence and bi-layer disparity remapping), 20 mm separation was compared to 40 mm separation. The results for these images are shown in Fig. 5 (a), with 20 mm receiving a large majority of votes (78.4%). Clearly, users found 20 mm more realistic than 40 mm, despite the 40 mm separation producing a stronger 3D effect.

Next, 20 mm separation was compared to 10 mm, with results shown in Fig. 5 (b). Here, 20 mm (45.5%) was again chosen to be more realistic than 10 mm (34.1%), supporting the result that a 16 mm separation would produce the most realistic 3D effect in this scenario.

VI. Discussion and Conclusions

It is interesting that while users chose 20 mm separation as the most realistic, they preferred 10 mm separation images. This is likely because they preferred the most comfortable images, due to inexperience with a 3D handheld. One comfort factor is the time for 3D fusion. Even if the disparities created using a 20 mm separation fall within the zone of comfort, it takes more time for the brain to fuse larger disparities. Consequently, 10 mm is still perceived to be more comfortable. However, as the user gains experience with the 3D handheld device, the time for fusion decreases, and thus the comfort advantage of the smaller camera separation decreases.

We have presented a technique to determine the optimal camera separation for realistic depths within a face during handheld 3D video communications. Ultimately, it is clear that disparities need to be minimized to maintain comfortable viewing, while maintaining realistic depths within the users face. This supports the need for disparity remapping techniques to remove unimportant depths in the background, even when the camera separation is small enough to provide depths with the zone of comfort. Since the optimal camera separation is dependent upon a number of factors (viewing distance, interocular distance, convergence depth), future work must gather viewing statistics to choose the best separation for a majority of users.

References

[1] S. Mangiat and J. 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] I. Feldmann, O. Schreer, R. Schfer, Z. Fei, H. Belt, and O. Divorra Escoda, "Immersive multi-user 3D video communication," September 2009.

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

[5] H. Yamanoue, M. Okui, and F. Okano, "Geometrical analysis of puppet-theater and cardboard effects in stereoscopic HDTV images," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp. 744–752, June 2006.

[6] T. Shibata, J. Kim, D. M. Hoffman, and M. S. Banks, "The zone of comfort: Predicting visual discomfort with stereo displays," *Journal of Vision*, vol. 11, no. 8, 2011.

[7] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Prentice-Hall, 2003.

[8] G. Sharma, L. Karam, and P. Wolfe, "Select trends in image, video, and multidimensional signal processing," *Signal Processing Magazine, IEEE*, vol. 29, no. 1, pp. 174–176, 2012.

[9] G. Blohm, A. Z. Khan, L. Ren, K. M. Schreiber, and J. D. Crawford, "Depth estimation from retinal disparity requires eye and head orientation signals." *Journal of vision*, vol. 8, no. 16, 2008.