Intermediate view generation for perceived depth adjustment of stereo video.

Zafer Arican[a] and Sehoon Yea [b] and Alan Sullivan[b] and Anthony Vetro [b]

[a]Ecole Polytechnique Federale de Lausanne, Lausanne, Switzerland;
[b]Mitsubishi Electric Research Labs, Cambridge,MA, USA

ABSTRACT

There is significant industry activity on delivery of 3D video to the home. It is expected that 3D capable devices will be able to provide consumers with the ability to adjust the depth perceived for stereo contents. This paper provides an overview of related techniques and evaluates the effectiveness of several approaches. Practical considerations are also discussed

Keywords: Intermediate View Generation, Stereoscopic Display, Perceived depth adjustment, Stereo video

1. INTRODUCTION

Commercialization of 3d ready stereoscopic televisions, new storage technologies and current trend on producing 3D media content have made 3D viewing possible and reachable to home cinema users. Various viewing technologies from anaglyph glasses to lcd shutter glasses and current research on glasses-free autostereoscopic displays aim to provide 3d perception by sending different images to each eye of the viewer. Variety in the 3D content and different sensitivity levels of human visual system, however, require a customization mechanism for comfortable viewing of the 3d contents. Range of the perceived depth in the 3D content is the main factor which determines the limits of comfortable viewing.

Binocular depth perception in human visual system is achieved by horizontally separated two eyes and fusion of the images received by each eye. Because there is a distance between two eyes, each eye receives slightly different view of the scene. The distance between the corresponding points called disparities in the images are different depending on the distance of the object to the viewer. These differences are analyzed in the brain to have the depth perception. Stereoscopic displays imitate this behavior by sending different images of the same scene to each eye. These two images are captured by two cameras slightly separated horizontally. Depending on the distance between these cameras the formed images and the disparities will differ. Various techniques to send different images to each eye have been proposed.[1,2]

Another dynamic for depth perception is the accommodation(focusing)-vergence relation. When looked at an object, the lenses in the eyes change shape to keep the object of interest in focus. This behavior,called accommodation enables a range of distance to be sharper. This range is called depth of field and changes with distance. Another motor behavior which controls the eyeballs is called vergence. The eyeballs rotate so that the optical axis of the eyes converge around the point of interest. Both convergence and accommodation points are close to each other and two motor systems controlling these points are linked during normal viewing. However, during viewing of 3D media on a stereoscopic display, these two systems will not operate together. The eyes will focus on the screen to keep the images sharp. However, the vergence system will fixate on a perceived 3D point which is in front or back of the screen creating an inconsistency in the two motor systems. If the distance between the convergence point of the optical axis and focusing point is large, it causes eye discomfort.[3,4] In addition, if the depth range changes rapidly due to camera movement or scene changes, the speed of accommodation will not be sufficient to follow. Thus, the perceived depth range should be arranged to avoid such breakdowns.

Further author information: (Send correspondence to Z.A.)
Z.A..: E-mail: zafer.arican@epfl.ch, Telephone: 41 76 4375271
S.Y.: E-mail: yea@merl.com, Telephone: +1 617 6217500

Depending on the gender, race and age, the distance between two eyes (Interpupillary distance) differs. This will cause different depth perception as the observed images on each eye are different. For example, if the interpupillary distance (IPD) is small, the close objects will look closer and far objects will be perceived as farther away. Considering, the accommodation and vergence system behavior, people with smaller IPD are more likely to suffer from a possible breakdown. Hence, adjusting the depth range has a paramount importance to avoid eye discomfort. (See Lambooij *et al*[3] for an extended study on eye discomfort for stereoscopic displays.)

This paper reviews the methods to adjust the perceived depth range of stereo videos. Section 2 summarizes the theory for perceived depth formation and methodology to change the perceived depth. Section 3 reviews different view generation methods used to scale the depth range and analyzes different aspects considered during view generation, namely, disparity estimation, temporal consistency, asymmetric generation and non-parallel camera configurations. Section 5 concludes the paper with a discussion and summary.

2. PERCEIVED DEPTH ADJUSTMENT

Binocular depth perception is formed by slight changes in two images seen by left and right images. Depending on the depth of the object in the scene, a slight displacement occurs in the position of the corresponding pixels in both images. Stereoscopic displays together with the filtering glasses help formation of two different images for each eye. Overlapped images on the display are filtered by special glasses (anaglyph, lcd shutter etc.) to discriminate received images by each eye. The displacement of image pixels, called disparities, creates the depth perception. Figure 1 shows the position of perceived depth points formed by different disparities. Sign, as well as quantity of disparity decide on the position of the perceived 3D point. As also discussed in some previous works[5,6] position of the perceived depth point is related to the disparity by the following formula.

$$Z' = \frac{b_e D}{b_e + d} \tag{1}$$

where b_e is the distance between two eyes, D is the viewing distance.

Stereo camera imitates the human visual system by taking images of a scene from slightly shifted positions. When two cameras differ only by position in horizontal direction, the configuration is called parallel camera configuration and the disparities are in horizontal direction. The disparities are a function of distance of the 3D point to the cameras, the distance between the cameras and focal length. Figure 2 shows the relation of depth to disparity. By simple triangle equality, the well known stereo disparity formula

$$Z = \frac{Bf}{d} \tag{2}$$

is obtained. B is the distance between two cameras called baseline distance and f is the focal length. The baseline distance, B and the focal length, f are parameters which are set during capturing. Thus, once captured these parameters cannot be changed leaving the disparity, d. There are two main approaches to modify the disparity, d to adjust the perceived depth. Following two subsections explain these two approaches.

2.1 Shifting

When images are captured by two cameras with displacement in one dimension and no rotation, all formed disparities are positive putting all perceived 3D objects in front of the scene. For excessive disparities, this will cause eye discomfort. Thus, either cameras are turned slightly to each other to form a convergence point around which the disparity is zero. For objects behind this point, the disparities are negative and for objects in front of the convergence point, the disparities are positive. Besides being a capture time solution, this camera configuration introduces some artifacts such as keystone distortion and depth plane curvature. To overcome this problem, a method called "zero plane setting" has been proposed. This method is based on shifting the image planes to create an artificial convergence point. In terms of image processing, this method shifts both images in opposite directions to add or subtract a constant from all disparities. Figure 3 shows such configuration. This effect let all the perceived objects shift forward or backward depending on the amount of shift in the images.

2.2 Scaling

Although scaling provides an effective method to adjust the depth, its effect is limited to moving the objects forward or backward as a whole. Changing the perceived depth range, however, requires scaling of the disparities by a constant. By equation 2, scaling is achieved by changing the baseline distance, B. However, the actual baseline distance is set during capturing and cannot be changed afterwards. To achieve the same effect, a virtual camera at the desired distance is placed and an intermediate view is generated for this position. Figure 4 shows

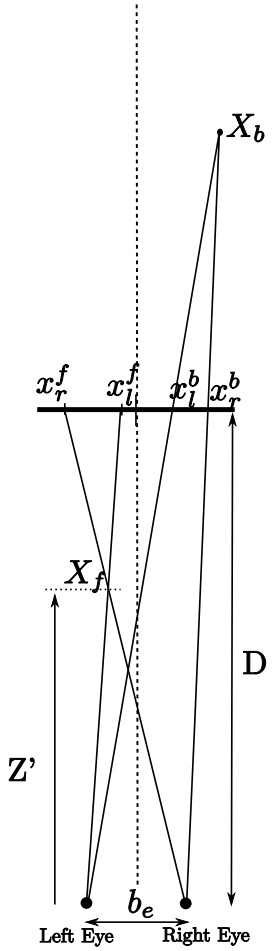

Figure 1. Perceived depths for negative and positive disparities

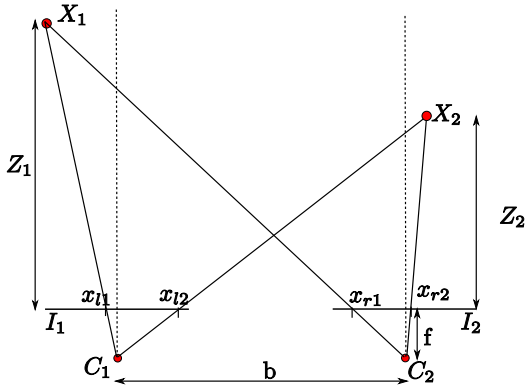

Figure 2. Formation of image points and effect of depth to disparity

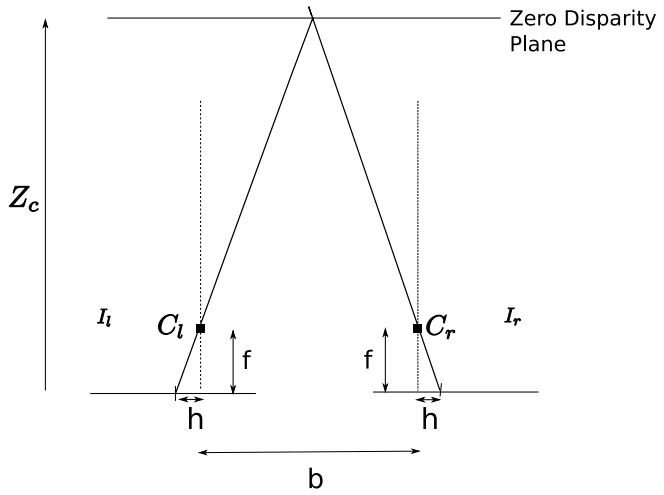

Figure 3. Image Plane shift for disparity adjustment and zero plane setting.

different configuration of virtual views to change the baseline, and scale the disparities.

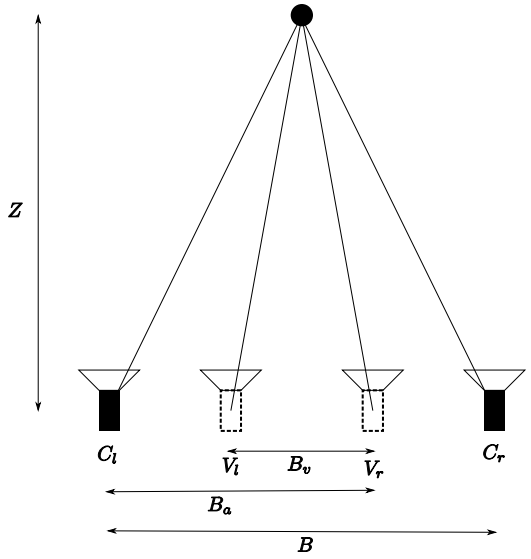

Figure 4. Virtual Camera Configuration for disparity scaling. Actual baseline distance is B. When both left and right images come from virtual cameras, the baseline distance is B_v. The configuration where only one of the images comes from the virtual camera is called asymmetric synthesis and the baseline distance is B_a

Table 1 summarizes the effect of scaling and shifting parameters.

Parameter		Screen Parallax	Perceived Depth	Object Size
Baseline Distance,b	Increase	Increase	Increase	Constant
	Decrease	Decrease	Decrease	Constant
Convergence Distance, Z_c	Increase	Decrease	Shift(Forward)	Constant
	Decrease	Increase	Shift(Backwards)	Constant
Focal length, f	Increase	Increase	Increase	Increase
	Decrease	Decrease	Decrease	Decrease

Table 1. Effect of different parameters for perceived depth adjustment[7]

3. INTERMEDIATE VIEW GENERATION

Many of the stereo content consist of left and right video files. In spite of the standardization efforts for 3DTV to include disparity maps, there is not any consensus yet on how to add the disparity information. Hence, many of the current movies and stereo clips are broadcasted as two seperate videos without any prior disparity information. For perceived depth adjustment on the user side, the shifting is an efficient method to change the perceived depth as it only requires shifting of the images and do not need any disparity maps. However, if the disparity range, thus the perceived depth range is large, moving the objects forward or backward will cause eye discomfort due to accommodation-convergence breakdown and loss of stereopsis. As explained in the previous section, scaling of the perceived depth range requires disparity maps for stereo video. This section discusses different view generation methods based on disparity maps and estimation of disparity maps.

3.1 Disparity Estimation

Although the scaling of the perceived depth is based on scaling of the disparity, the main goal is to generate acceptable intermediate views to change the depth perception. Thus, as long as visually acceptable image is generated, the accuracy of the disparity map has a lower priority. This assumption provides a relaxation on the choice of disparity estimation method too. Many stereo dense disparity estimation methods are proposed to get an accurate map.[8,9] Among these methods, the ones based on belief propagation and graph-cut are the most prominent methods giving the best accuracy with the cost of computational burden. Because the disparity estimation is performed on the low-end user side with real-time performance constraint, computationally demanding methods cannot be applied. Although, there are implementations of high-accuracy methods on GPU,[10,11] particularly for high-resolution videos, the speed is still not close to real-time. Cost aggregation based methods[12,13] offer a balanced solution between accuracy and speed. Each of the steps are parallelizable and filtering steps provide an implicit smoothness on the disparity. Due to their limited support, cost aggregation methods will not provide accurate disparity values for low texture regions, however, for view synthesis, the lack of accuracy will not make an impact around these regions.

3.2 Disocclusion

After the disparity maps are estimated, the intermediate views are generated by warping either image at the desired position. At the desired virtual camera position, there are regions not seen by the camera of which the images were warped. This is called disocclusion and these regions must be filled by some relevant pixel values. Unseen part is either by one camera or both cameras. If it is unseen by only one camera, that region is filled by patches from the other image together with matting.[14] If the occluded part cannot be seen by both cameras, the corresponding disoccluded part is filled by inpainting techniques.[14,15] Another approach is to smooth the disparity maps around discontinuities to avoid disocclusion. By this way, each of the pixels in intermediate view are filled by either of the images. Note that occluded regions are mostly determined by the disparity maps. However, discontinuities in disparity maps do not necessarily follow the object boundaries. This will create cracks and artifacts around object boundaries.

3.3 Temporal Consistency

Many of the stereo disparity estimation methods consider still images as input. When stereo videos are processed and each frame is processed independently, the flickering of disparity maps will occur as there is no prior information inherited from previous frames. This flickering will be apparent particularly at the object boundaries and disoccluded regions when an intermediate view is generated. These regions are unstable regions as changes around these high frequency regions are detected by the eyes and variations in disparity maps affects these regions. Thus, temporally consistent disparity map estimation and view synthesis reduces these artifacts and improve visual comfort. Because it is assumed that a disparity estimation is not available a priori and real-time performance is expected, the methods using bundle adjustment[16] are not applicable. Small time window approaches using disparity flows[17] and optical flows[18] are some of the methods to provide temporally smooth disparity maps and synthesized views. However, disparity maps considering both dynamic scenes and camera motion are still to be addressed.

3.4 Symmetric/Asymmetric View Generation

Changing the baseline distance via intermediate view generation can be performed in two ways. Symmetric view generation keeps the center of the baseline fixed and creates two synthetic view on both side of the center. Its symmetry will provide geometrical consistency. That is, if the perceived depth range is changed, the position of the center does not change avoiding moving camera effect. This effect is more apparent if the cameras are not perfectly parallel. However, because both images are to be synthesized, artifacts particularly around object boundaries exist on both images causing depth ambiguity and discomfort. In addition, it will double the computation. Asymmetric view generation is based on keeping one of the reference images as fixed and generating only one intermediate view in between two reference images. The idea is based on binocular rivalry and suppression[19] and has been used in stereo image pair compression.[14,20] When there is an inconsistency on the depth cues of both eyes, the one with low-frequency components is suppressed and replaced by the other cue. This behavior is observed on asymmetric view generation and artifacts due to disocclusion around object boundaries are smoothed out by the reference image which has higher frequency. However, note that as the artifacts occur always on one side, eye fatigue may occur on that eye. For that reason, alternating the reference views will distribute the fatigue on both eyes enabling longer viewing sessions. Figure 4 shows the typical symmetric and asymmetric virtual camera configurations.

3.5 Toed-in Camera Configuration

If two stereo images captured by perfectly stereo camera pair are displayed on a stereoscopic display, all perceived depth objects will appear in front of the scene as all disparities are positive. This will reduce the perception and cause eye discomfort. To eliminate this problem, often the cameras are slightly rotated towards each other so that their optical axis intersect at a convergence point. This configuration also imitates the convergence motor system of the human visual system. Figure 5 illustrates two camera configurations. This configuration however cause an artifact called keystone distortion. Particularly towards each side of the image vertical disparities emerges due to warping of the image. This distortion is one of reasons for eye discomfort.

Due to non-parallel disparity lines, disparity estimation when performed directly on the two images will give erroneous results reducing the quality of the synthesized view. Although this degradation is low and concealable for small degrees of rotation, it is not possible to process the images as-is for disparity estimation for larger rotations. In such cases, rectification of images to align the disparities is necessary. However, this process requires estimation of the camera parameters and an additional warping is performed which reduces the image quality and size. In addition, the same effect of the convergence is achieved by Zero-plane setting explained in section 2 on parallel camera configuration without any keystone distortion.

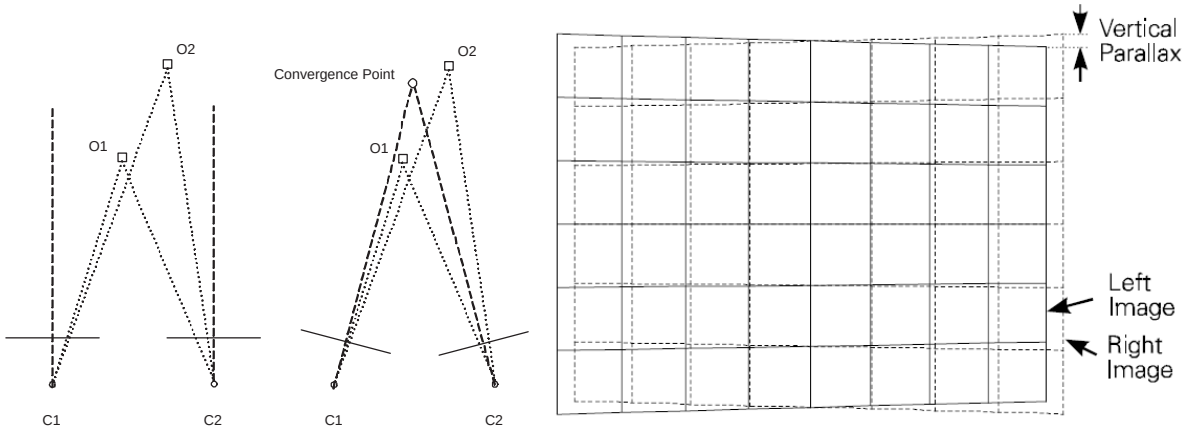

Figure 5. Parallel(left) and toed-in camera configurations(middle) with a typical keystone distortion due to toed-in camera configuration(courtesy of[21])

4. PERFORMANCE EVALUATION

We tested two key disparity estimation methods mentioned in section 3.1, namely graph-cut and cost aggregation. For graph-cut we used the α-expansion implementation of Kolmogorov.[22] We implemented a non-parallel version of cost aggregation using box-filter as the smoothing filter. We tested two methods on "Skydiving"[23] and "Lovebird" sequences. We generated a synthesized view in the middle of two cameras which cuts the baseline distance to half using the estimated depth maps calculated by these two methods. For view synthesis we used the implementation of.[14] Figure 6 shows the left and right views with synthesized views with two views for "Sky Diving" sequence. Results for the "Lovebird" sequence are given in Figure 8. The view synthesis performances are similar for both algorithm although there are significant differences in depth estimation results which is given in Fig. 7. This supports the hypothesis stating that an accurate depth estimation is not necessary. However, if the estimated depth map is noisy which is the case for cost aggregation method, spurious holes will form and these holes need to be filled using some digital inpainting techniques. These techniques are generally slow and affects the real-time operation.

The cost aggregation based methods provide faster operation without a big sacrifice on accuracy compared to more sophisticated methods such as graph-cut. The difference is more apparent as the image size gets bigger. In addition, each step of the cost aggregation has a structure which is convenient for parallel processing which improves the speed dramatically and allows some refinement on the accuracy. The view synthesis times are similar, however as mentioned before, if the holes are filled by using some digital inpainting techniques, a smooth disparity map is favored in terms of speed.

Figure 6. Left and right views from "Sky Diving" (top row) and two synthesized views using GC and cost aggregation methods respectively(bottom row).

5. CONCLUSION

Two important systems providing depth perception namely binocular reception and accommodation-convergence system work collaboratively in normal viewing. However, stereoscopic displays interrupt these systems by causing a gap between the accommodation and convergence points. When the gap between these two points increases, eye discomfort will occur. Thus, 3D content viewed on these displays should be arranged for increased eye comfort considering both the content and the physiology of the viewer such as IPD. This adjustment requires generation of synthesized intermediate views for scaling of the perceived depth range. For many existing stereo video content, the disparity maps necessary for view synthesis is not available. Thus, an online disparity map estimation is required. Due to limited computation power and real-time processing requirements, computationally demanding

disparity estimations cannot be implemented on these devices. An accurate disparity map is redundant as long as visually acceptable intermediate views are generated and desired depth range is achieved. However, temporal consistency should be maintained to avoid flickering effect which will reduce the comfort.

If moved to the capture side, an accurate depth map has a positive effect on the image quality. Standardization efforts including the disparity maps and extra information necessary to fill occluded regions will both reduce the load on the consumer-side devices and improve the image quality. In addition, zero-plane setting is shown to give the same convergence effect without introducing artifacts. Thus, a capture time rotation of the cameras can be avioided. This will facilitate the processing on both capture and display side. For further improvement on the processing speed, asymmetric view generation can be adopted as the human visual system will suppress the artifacts and low resolution components.

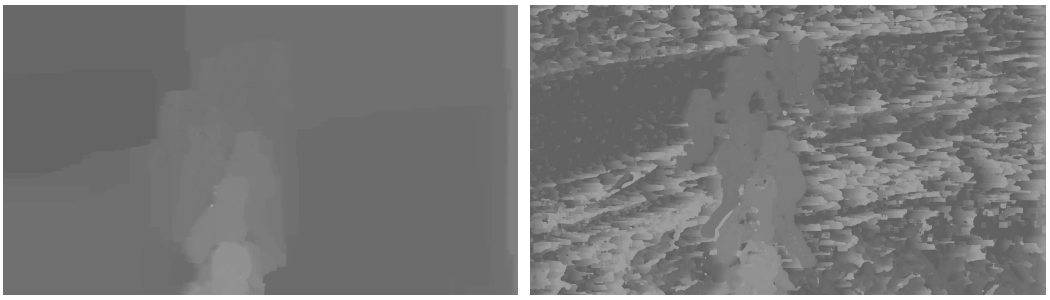

Figure 7. Estimated disparity maps by graph-cut and cost aggregation methods. Note the noisy disparity map by cost-aggregation.

Figure 8. Left and right views from “Lovebird” (top row) and two synthesized views using GC and cost aggregation methods respectively(bottom row).

REFERENCES

[1] Dodgson, N., "Autostereoscopic 3D displays," *Computer* **38**(8), 31–36 (2005).

[2] Benzie, P., Watson, J., Surman, P., Rakkolainen, I., Hopf, K., Urey, H., Sainov, V., and Von Kopylow, C., "A survey of 3DTV displays: techniques and technologies," *IEEE Transactions on Circuits and Systems for Video Technology* **17**(11), 1647–1658 (2007).

[3] Lambooij, M., IJsselsteijn, W., Fortuin, M., and Heynderickx, I., "Visual discomfort and visual fatigue of stereoscopic displays: a review," *Journal of Imaging Science and Technology* **53**, 030201 (2009).

[4] Hoffman, D., Girshick, A., Akeley, K., and Banks, M., "Vergence-accommodation conflicts hinder visual performance and cause visual fatigue," *J Vis* **8**(3), 33 (2008).

[5] Konrad, J., "Enhancement of viewer comfort in stereoscopic viewing: parallax adjustment," in [*Proc. SPIE Stereoscopic Displays and Virtual Reality Systems*], **3639**, 179–190, Citeseer (1999).

[6] Holliman, N., "Mapping perceived depth to regions of interest in stereoscopic images," *Stereoscopic Displays and Virtual Reality Systems XI, Proceedings of SPIE* **5291** (2004).

[7] Fehn, C., Hopf, K., and Quante, B., "Key technologies for an advanced 3D TV system," in [*Proceedings of SPIE*], **5599**, 66 (2004).

[8] Scharstein, D. and Szeliski, R., "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision* **47**(1), 7–42 (2002).

[9] "Middlebury Stereo Vision Page," (2009).

[10] Yang, Q., Wang, L., Yang, R., Wang, S., Liao, M., and Nister, D., "Real-time global stereo matching using hierarchical belief propagation," in [*The British Machine Vision Conference*], 989–998 (2006).

[11] Brunton, A., Shu, C., and Roth, G., "Belief propagation on the GPU for stereo vision," in [*Canad. Conf. on Comput. and Robot Vision*], (2006).

[12] Gong, M., Yang, R., Wang, L., and Gong, M., "A performance study on different cost aggregation approaches used in real-time stereo matching," *International Journal of Computer Vision* **75**(2), 283–296 (2007).

[13] Min, D. and Sohn, K., "Cost aggregation and occlusion handling with WLS in stereo matching," *IEEE Transactions on Image Processing* **17**(8), 1431–1442 (2008).

[14] Mori, Y., Fukushima, N., Yendo, T., Fujii, T., and Tanimoto, M., "View generation with 3D warping using depth information for FTV," *Signal Processing: Image Communication* (2008).

[15] Tauber, Z., Li, Z., and Drew, M., "Review and preview: Disocclusion by inpainting for image-based rendering," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **37**(4), 527–540 (2007).

[16] Zhang, G., Jia, J., Wong, T., and Bao, H., "Recovering consistent video depth maps via bundle optimization," in [*IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*], 1–8 (2008).

[17] Gong, M., "Enforcing temporal consistency in real-time stereo estimation," *LECTURE NOTES IN COMPUTER SCIENCE* **3953**, 564 (2006).

[18] Bartczak, B., Jung, D., and Koch, R., "Real-Time Neighborhood Based Disparity Estimation Incorporating Temporal Evidence," *Lecture Notes in Computer Science* **5096**, 153–162 (2008).

[19] Hayashi, R., Maeda, T., Shimojo, S., and Tachi, S., "An integrative model of binocular vision: a stereo model utilizing interocularly unpaired points produces both depth and binocular rivalry," *Vision Research* **44**(20), 2367–2380 (2004).

[20] Perkins, M., "Data compression of stereopairs," *IEEE Transactions on communications* **40**(4), 684–696 (1992).

[21] Woods, A., Docherty, T., and Koch, R., "Image distortions in stereoscopic video systems," in [*Proceedings of SPIE*], **36**, SPIE (1993).

[22] Boykov, Y. and Kolmogorov, V., "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , 1124–1137 (2004).

[23] "3dtv.at - 3D movies," (2009).