MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Intermediate View Generation for Perceived Depth Adjustment of Stereo Video

Zafer Arican, Sehoon Yea, Alan Sullivan, Anthony Vetro

TR2009-052 September 2009

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 content. This paper provides an overview of related techniques and evaluates the effectiveness of several approaches. Practical considerations are also discussed.

SPIE Conference on Applications of Digital Image Processing 2009

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 content. 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 fusion of the images received by horizontally separated two eyes. 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 and lead to depth perception. Stereoscopic displays imitate this behavior by sending a different image 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. Given a scene, the lenses in the eyes change shape to keep the object of interest in focus. This behavior, called accommodation, enables a range of distances 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 a 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).

The construction of this paper is as follows. Section 2 summarizes the theory of depth perception and reviews methods for adjusting the perceived depth. Section 3 focuses on intermediate view generation which is necessary to realize more versatile and natural adjustment of perceived depth. Various issues are discussed such as disparity estimation, temporal consistency, asymmetric view generation and non-parallel camera configurations. Section 4 provides an evaluation of performance and section 5 concludes the paper.

2. PERCEIVED DEPTH ADJUSTMENT

Binocular depth perception is formed by slight changes in two images seen by left and right eyes. 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 forming two different images for both eyes. 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 disparity, creates the depth perception. Figure 2 shows the position of perceived depth points formed by different disparities. The sign as well as the magnitude of a disparity decide on the position of the perceived 3D point. As also discussed in some previous works,[5,6] the perceived depth (Z') of a 3D point is related to the disparity (d) 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.

A stereo camera imitates the human visual system by taking images of a scene from slightly shifted positions. When the relative configuration of two cameras differs only by a shift in the horizontal direction, the configuration is called parallel and the disparity is a 1D value along the horizontal direction. The disparity is a function of the distance of a 3D point to the cameras as well as the camera parameters such as baseline distance and focal length. Figure 1 shows the relation of depth to disparity. By simple triangle equality, the well known stereo disparity formula is obtained as follows:

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

where B is the distance between two cameras called baseline distance and f is the focal length.

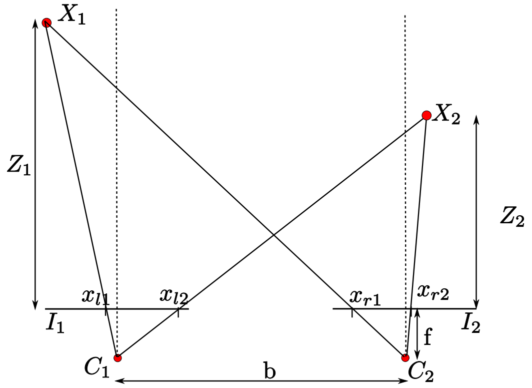

Figure 1. Formation of image points and effect of depth to disparity

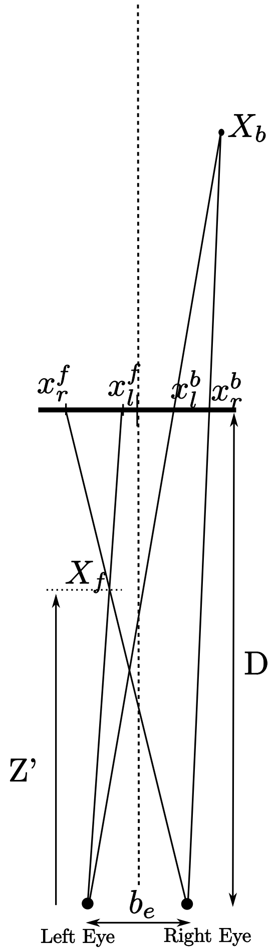

Figure 2. Perceived depths for negative and positive disparities

The baseline distance, B, and the focal length, f, are parameters which are set during the capturing stage, hence cannot be changed afterwards. Therefore, the most practical way to adjust the perceived depth would be to modify the disparity d. Following subsections will discuss two main approaches to modifying the disparity: shifting and scaling.

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 screen. For excessive disparities, this will cause eye discomfort. Thus, the 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 a configuration. This effect lets all the perceived objects shift forward or backward depending on the amount of shift in the images.

2.2 Scaling

Although shifting provides an effective method to adjust the depth, its effect is limited to moving the entire scene forward or backward as a whole. On the other hand, adjusting the range of the perceived depth - thereby

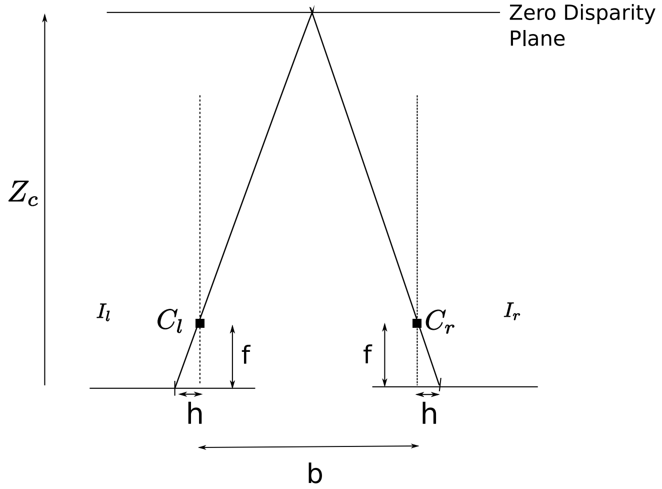

Figure 3. Image Plane shift for disparity adjustment and zero plane setting.

the degree of 3D perception - requires scaling of the disparities by a constant. Equation 2 suggests it can be achieved by changing the baseline distance, B. However, the actual baseline distance is set during capture and cannot be changed afterwards. To achieve the same effect, a virtual camera at the desired distance is assumed and an intermediate view is generated for this position. Figure 4 shows different configurations of virtual views to change the baseline for disparity scaling.

Table 1 summarizes the effect of scaling and shifting parameters.

Parameter		Screen Parallax	Perceived Depth	Object Size
Baseline Distance,b	Increase Decrease	Increase Decrease	Increase Decrease	Constant Constant

Convergence Distance, Z_c	Increase
	Decrease
		Decrease
		Increase
			Shift(Forward)
			Shift(Backwards)
				Constant
				Constant

Focal length, f	Increase Decrease	Increase Decrease	Increase Decrease	Increase Decrease

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

3. INTERMEDIATE VIEW GENERATION

Even though efforts are being made towards the standardization of data formats for the dissemination of 3D videos including the associated depth information, majority of the current 3D movie and video clips are available only as two separate left and right videos without any auxiliary disparity information. For perceived depth adjustment on the user side, 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 perceived depth range is excessive, moving the objects forward or backward will be likely to cause eye discomfort due to the accommodation-convergence breakdown and loss of stereopsis. As explained in the previous section, scaling of the perceived depth range requires disparity maps for a stereo video. This section discusses ways to estimate disparity maps and generate intermediate views based on such maps.

3.1 Disparity Estimation

When scaling of the perceived depth is performed based on the estimated disparity map, the main goal is to generate intermediate views of acceptable quality to change the depth perception. Thus, as long as a visually acceptable image is generated, the accuracy of the disparity map has a lower priority. This assumption in turn provides a relaxation on the choice of disparity estimation method. Many dense disparity estimation methods have been proposed to get an accurate map.[8,9] Among these methods, the ones based on belief propagation[10] and

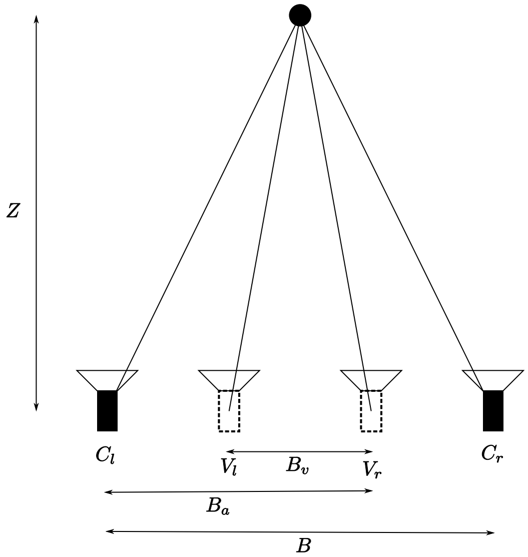

Figure 4. Virtual camera configuration for disparity scaling. Actual baseline distance is B. When both left and right images come from virtual cameras, the baseline distance is B_v. The configuration where only one of the images comes from the virtual camera is called asymmetric synthesis and the baseline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

graph-cut[11] are the most prominent methods giving the best accuracy for a given computational cost. Although there are implementations of these high-accuracy methods on GPU,[10,12] the speed is still not close to real-time particularly for high-resolution videos. Especially when the disparity estimation needs to be performed on a low-end consumer device with real-time performance constraints, these computationally demanding methods are not viable options. In this context, cost aggregation based methods[13,14] 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 local nature of optimization, cost aggregation methods might fail to provide accurate disparity values especially for low texture regions. However, the lack of accuracy in these regions often does not lead to noticeable artifacts in terms of view synthesis and various ideas such as hierarchical filtering can be incorporated to overcome such a limitation.[14] Finally, it needs to be emphasized that having an appropriate disparity search range for each frame is very important before applying any of the aforementioned disparity estimation methods. This is because it not only guarantees all the important scene features corresponding to certain disparity ranges are searched over, but also helps reducing the risk of falling into local minima in the search.

3.2 Disocclusion

After the disparity maps are estimated, the intermediate views are generated by warping either image to a desired virtual position. Some regions in the warped image at the virtual camera position do not have pixel values as they correspond to the invisible pixels from the camera the images were warped from. 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.[15] If the occluded part cannot be seen by either camera, the corresponding disoccluded part is filled by inpainting techniques.[15,16] Another approach is to smooth the disparity maps around discontinuities to avoid disocclusion. In this way, each of the pixels in the intermediate view are filled by either of the two images. Note that occluded regions can be 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

Majority of stereo disparity estimation methods consider still images as input. When each frame of a stereo video is processed independently, flickering of disparity maps will likely occur as there is no prior information inherited from previous frames that will impose temporal consistency between corresponding disparity estimates. This flickering could be particularly apparent around the object boundaries and disoccluded regions when an intermediate view is generated as inconsistent changes in these regions are easily caught by eyes. Thus, temporally consistent disparity map estimation and view synthesis will be necessary to reduce these artifacts and improve visual comfort. For example, a method using bundle adjustment[17] was recently proposed to achieve such a goal. However, when the disparity maps of the previous frames are not available and real-time performance is expected, such a method is not practical. Small time window approaches using disparity flows[18] and optical flows[19] are alternative methods to provide temporally smooth disparity maps and synthesized views. However, the problem of extending such methods for disparity maps considering both dynamic scenes and camera motion is still open.

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 sides of the center. Its symmetry will provide geometrical consistency. That is, even when the perceived depth range is changed, the position of the center will not change, hence avoid the moving camera effect. This effect is more apparent if the cameras are not perfectly parallel. However, using two synthesized images implies, in addition to its obvious computational disadvantage, that artifacts particularly around object boundaries will appear on both images and might cause depth ambiguity and discomfort.

Asymmetric view generation is based on keeping one of the reference images as fixed and generating only one intermediate view in between the two reference images. The idea is based on binocular rivalry and suppression[20] and has been used in stereo image pair compression.[15,21] 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 this 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 a perfectly aligned stereo camera pair are displayed on a stereoscopic display, all perceived objects will appear in front of the scene as all disparities are positive. This will reduce the perception of depth 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

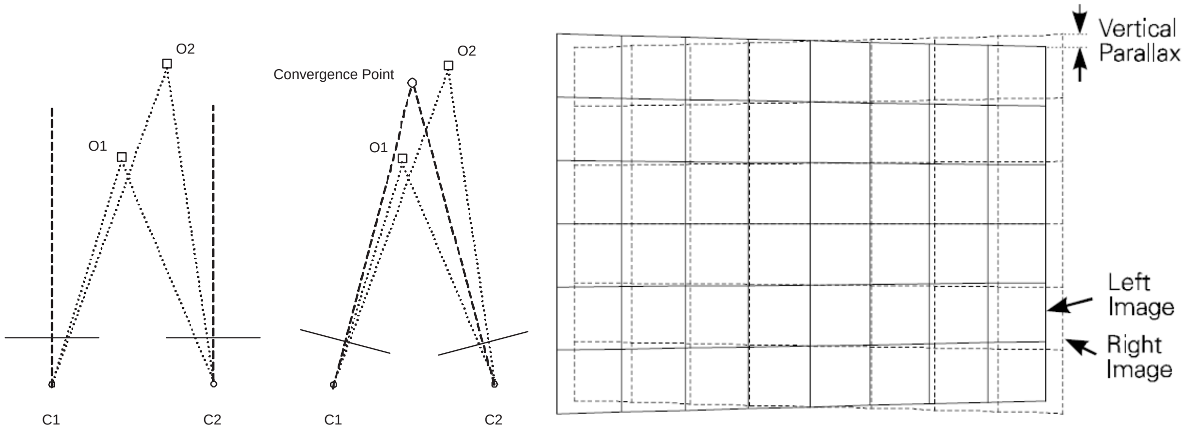

Figure 5. Parallel(left) and toed-in camera configurations(middle) with a typical keystone distortion due to toed-in camera configuration(courtesy of[22])

configuration however causes an artifact called keystone distortion. Particularly towards each side of the image vertical disparities emerge due to warping of the image. This distortion is one of reasons for eye discomfort. In addition, once set to a fixed convergence point, it cannot be changed on-the-fly. Perhaps more importantly, 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.

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.[11] We also implemented the cost aggregation algorithm using a box-filter. We tested the 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 an implementation similar to the one in.[15]

Figure 6 shows the original input left and right views (top row) together with the synthesized intermediate views with two disparity estimation methods (bottom row) for the "Sky Diving" sequence. Corresponding results for the "Lovebird" sequence are given in Figure 7. Although there are significant differences in depth estimation results as shown by the sample in Figure 8, the quality of the view synthesis is similar for both algorithms. This supports the hypothesis stating that an accurate depth estimation may not be always necessary. However, if the estimated depth map is noisy which is the case with the cost aggregation method, spurious holes could form and these holes need to be filled using an inpainting technique. These techniques are generally slow and compromise the real-time performance.

As an additional test, Figure 9 compares the view synthesis results corresponding to the cases with and without the use of correct disparity search range, respectively. Significant distortion appears in the latter case, especially around the helmet of the foremost skydiver as it corresponds to a large positive disparity area that was ignored in the disparity search.

For the tested sequences, the cost aggregation based methods provided faster operation without a big sacrifice on accuracy compared to a more sophisticated method such as graph-cut. Also note that each step of the cost aggregation has a structure convenient for parallel processing which improves the speed dramatically and allows for refinement of the accuracy. However, a more visible difference may be observed in terms of speed and accuracy between the two approaches with test materials having different scene complexity or larger image sizes.

5. CONCLUSION

Two important systems providing depth perception namely binocular perception 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 contents, the disparity maps necessary for view synthesis are not available. Thus, online disparity map estimation becomes necessary.

Due to the limited computational power and real-time processing requirements, computationally demanding disparity estimation methods cannot be implemented on typical consumer devices. Tests suggest that an accurate disparity map might not be necessary as long as visually acceptable intermediate views are generated. However, temporal consistency should be maintained in generating such a map to avoid flickering effect which will reduce the visual comfort. From a long term perspective, an accurate depth map as well as other auxiliary information

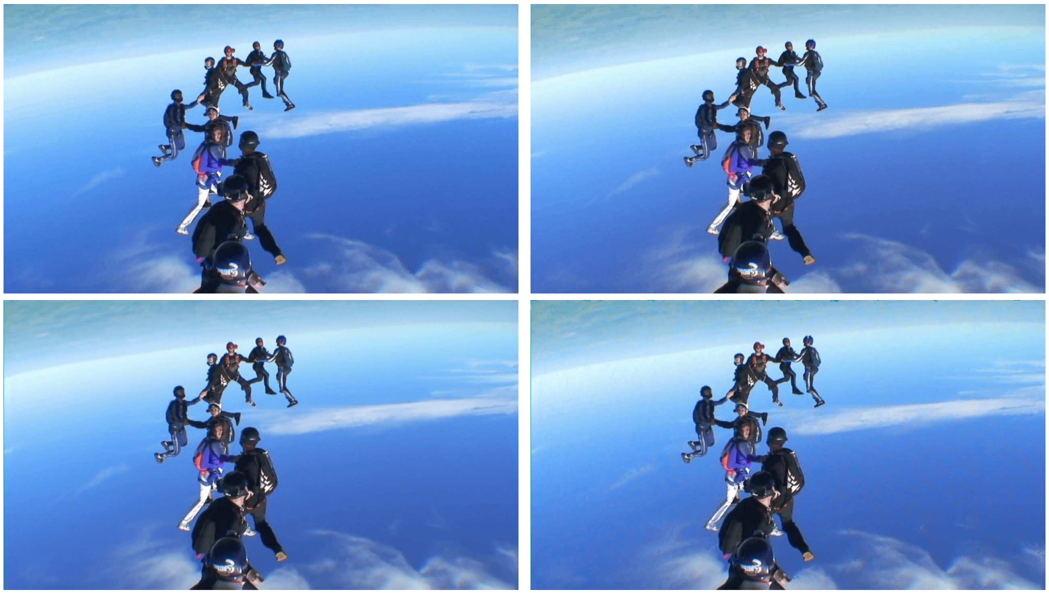

Figure 6. Left and right views from “Sky Diving” (top row) and two synthesized views using GC and cost aggregation methods, respectively(bottom row).

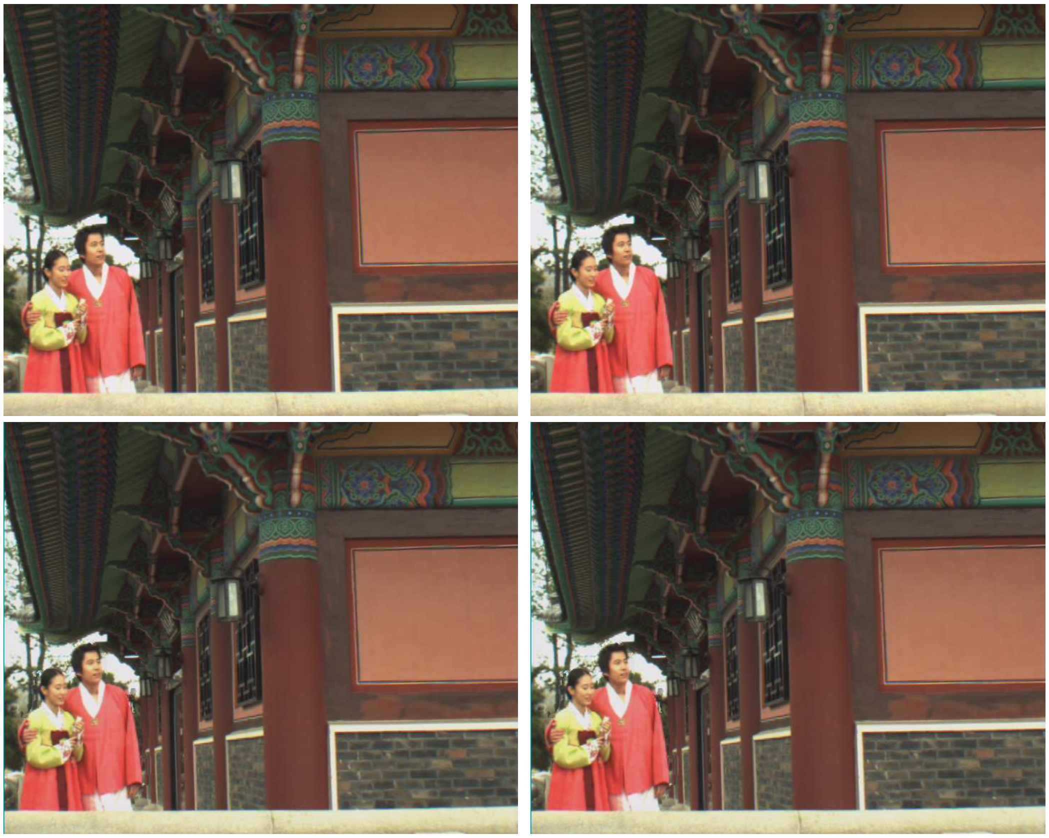

Figure 7. Left and right views from “Lovebird” (top row) and two synthesized views using GC and cost aggregation methods, respectively(bottom row).

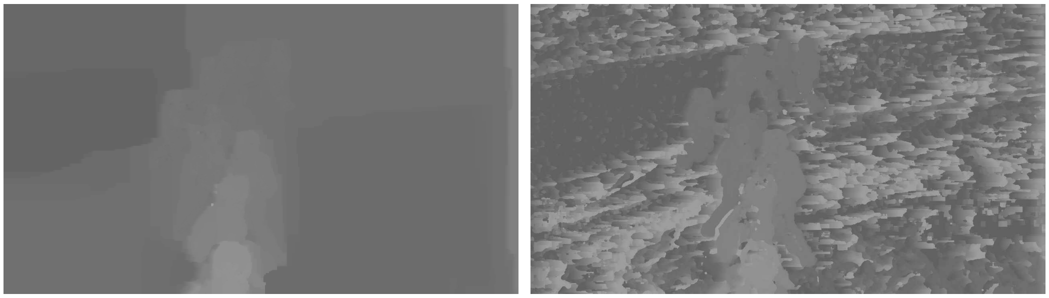

Figure 8. Estimated disparity maps by graph-cut and cost aggregation methods. Note the noisy disparity map by cost-aggregation.

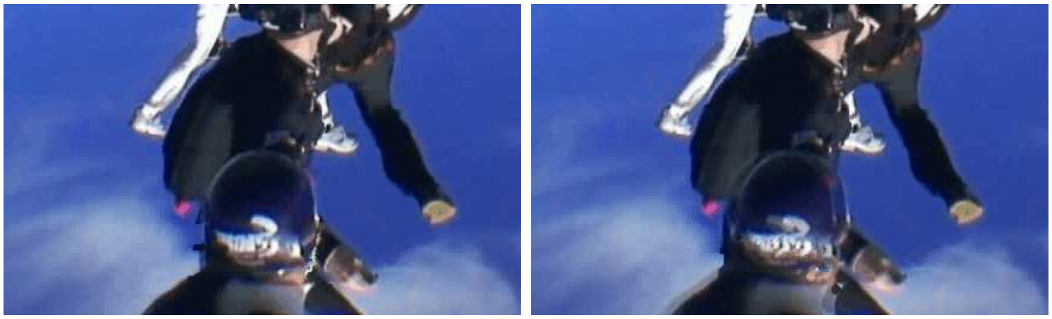

Figure 9. View synthesis based upon the disparity map with (left) and without (right) correct search range

useful for view-synthesis provided by a standardized delivery format will both reduce the load on the consumer devices and improve the image quality. 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.

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] 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).

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

[13] 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).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

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

[19] 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).

[20] 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).

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

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

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