# Multi-view Depth Map Estimation Enhancing Temporal Consistency

Sang-Beom Lee and Yo-Sung Ho
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju, 500-712, Republic of Korea
E-mail: {sblee, hoyo}@gist.ac.kr

**Abstract:** In this paper, we propose a scheme for multi-view depth map estimation to enhance temporal consistency. After we divide the center image into several segments, we estimate one depth value for each segment using 3-D warping and segment-based matching techniques. In the refinement process, we apply a segment-based belief propagation algorithm. In order to enhance temporal consistency and reliability of the depth map, we define a temporally weighted matching function and apply in the initial depth estimation step. We also apply a temporal postprocessing operation to the refined depth map. Experimental results show that the final depth sequence has improved temporal consistency with reduced errors.

## 1. Introduction

A three-dimensional television (3DTV) using multi-view images is in the spotlight as one of the next-generation broadcasting systems [1]. In order to acquire multi-view images, we utilize arrays of cameras to capture a 3D scene with wide-viewing angle and we can feel the presence from the multi-view images as displaying them in a 3D display.

Generally, two major issues can be found in multi-view camera system. The first issue is the relativity between the visual quality and the distance of cameras. The other issue is flickering. When users change their views while watching contents in the 3D display, the flickering will occur on the display if the distance between cameras is large and the scene is changed suddenly. It causes a visual discomfort to viewers' eyes.

For natural reproduction of 3D contents using multi-view images, we should reconstruct intermediate views. An intermediate view is an image captured from a virtual camera between real multi-view cameras. By interpolating intermediate views, we not only provide high quality 3D contents, but also reduce the visual discomfort.

In order to reconstruct intermediate images at virtual viewpoints, we need depth information. Many works have been carried out for acquisition of the 3D depth information. As one of the passive sensing methods, stereo matching is well-known [2]. The task of stereo matching is the computation of the disparity for two stereoscopic images. Recently, segment-based stereo matching algorithms are attracting issue because of their good performance [3]. These methods assume that a scene has a set of non-overlapping planes in a disparity space and that these planes correspond to at least one homogeneous color segment obtained by color segmentation for the reference image.

However, there still exist several problems such as wide baseline, occlusion which is visible in the reference view but invisible in other views, and so on. Especially, since these methods perform each frame separately, we notice that the results have low temporal consistency. In other words, the resultant depth maps have the low consistency of depth values at the same region but in different time.

In this paper, we propose a new depth map postprocessing scheme enhancing temporal consistency. After we perform depth map estimation for each frame, we extract the background mask calculating the difference between consecutive two frames. Moreover, we calculate segment-based difference to reduce the ambiguity of the moving foreground. Then, we merge the background masks obtained by several frames. In the final step, we apply the median filter to the depth values for each pixel of the background in the temporal direction.

## 2. Multi-view Depth Map Estimation

Figure 1 shows the block diagram of multi-view depth estimation. After the color segmentation is performed for the center image, the initial depth map is determined by 3D warping technique and the segment-based matching. In the final step, the initial depth map is refined by the segment-based belief propagation.

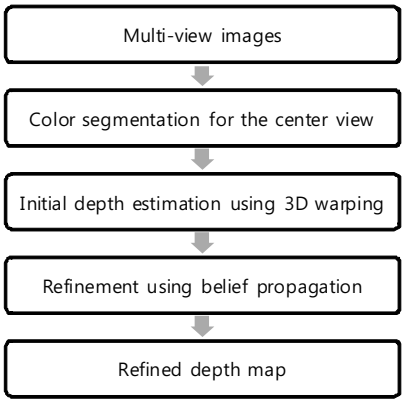
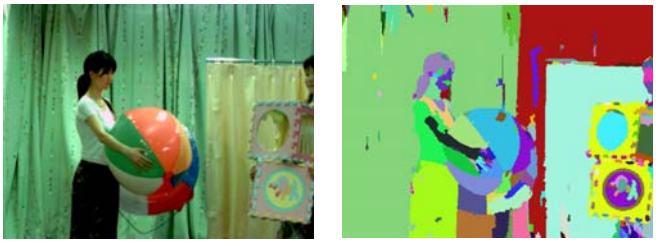



Figure 1. Block diagram of multi-view depth estimation

### 2. 1 Image segmentation

Multi-view depth map estimation scheme assumes that all the pixels in one segment have exactly the same depth value. Moreover, since most depth discontinuity occurs near the object boundary, we also assume that each segment does not contain the object boundary. We eventually have the induction that the better segmentation algorithm guarantees the higher quality of depth map. Figure 2 shows the image segmentation result for 'Akko&Kayo' (view 27) by using 'mean shift' image segmentation scheme [4]. As

shown in the Fig. 2, we confirm that the mean shift algorithm fix the accurate object boundary.



(a) Original image (b) Segmented image

Figure 2. Segmentation result for "Akko&Kayo" (View 27)

### 2. 2 Initial depth map estimation

After the center image is segmented, the initial depth map is calculated. In order to directly calculate depth values, not disparity values, 3D warping technique can be used [5]. If we conduct the 3D warping technique, we can reduce the error of the depth map since we skip the image rectification and disparity-to-depth conversion. By increasing the depth value and calculating the matching score, we determine the initial depth value when the matching score is minimized. In addition, we compare center view to both left and right view to resolve the occlusion problem.

Figure 3 shows the example of the initial depth estimation using 3D warping. The small triangle-shaped segment in the center view is warped to both left and right view and the matching score is calculated. As shown in the Fig. 3, although the triangle segment is occluded by the circle in the right view, it is not occluded in the left view. In this case, we compare the center view with the left view. Therefore, we can easily solve the occlusion problem when using multi-view images.

As a matching function, SD (squared intensity differences) and AD (absolute intensity differences) are the most popular. However, these functions are not robust to illumination changes between cameras. Recently, a self-adaptation dissimilarity measure [3] is proposed as a matching function. This function adds the conventional MAD to the mean absolute gradient difference. Since the gradient map represents the luminance change, not the absolute value, this function can be robust to absolute luminance mismatch between views.

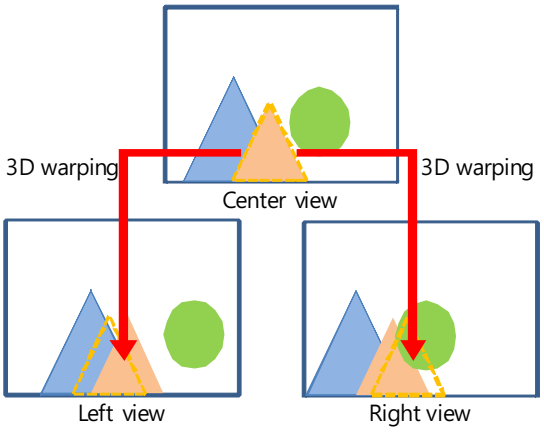



Figure 3. Initial depth estimation using 3D warping

### 2. 3 Segment-based belief propagation

Although the initial depth map reserve the object boundary, it has errorneous region, especially, the background region. The main reason is that the background area has a monotonous color and the matching score can be minimized even if we calculate the score with the wrong depth. In order to solve this problem, many algorithms are proposed such as graph cut, dynamic programming, belief propagation. Recently, many algorithms adopt belief propagation as an optimization process of disparity plane assignment. Belief propagation is an iterative inference algorithm that propagates messages in the network. The basic idea is that we refine the initial disparity map by considering the neighbors' matching score iteratively.

## 3. Depth Map Estimation Enhancing Temporal Consistency

### 3. 1 Temporally weighted matching function

Since the conventional depth estimation methods separately estimate the depth value for frame by frame, the resultant depth sequence has a low temporal consistency. Furthermore, even if the refinement is performed, there still exist errors. Figure 4 shows the depth sequence obtained by the previous method. As shown in Fig. 4, the depth values of the background are changed and there exists erroneous regions. That is, the temporal correlation of the depth sequence is low.

(a) t=2 (b) t=3

Figure 4. Depth sequence of the previous method

We propose a new matching function that refers to the depth value of the previous frame when estimating the depth of the current frame. The proposed scheme adds the self-adaptation function to the weighting function considering the depth value of the previous frame. The temporally weighted matching function is defined by

$$C(x,y,d) = C_{self-adaptation}(x,y,d) + C_{temp}(x,y,d) \quad (1)$$

$C_{temp}(x,y,d)$ can be defined by

$$C_{temp}(x,y,d) = \lambda \,|\, d - D_{prev}(x,y) \,| \quad (2)$$

where $\lambda$ represents the slope of the weighting function and $D_{prev}(x,y)$ represents the previous depth value.

Figure 5 illustrates the graph of the self-adaptation function and the proposed matching function. In Fig. 5, the dotted line represents the self-adaptation function and the chain line represents the weighting function considering the

previous depth value. Also, the solid line represents the temporally weighted matching function.

As shown in Fig. 5, in case of the previous depth value is equal to 100, the probability that the current depth value is around 100 is very high. Therefore, we apply the weighting function that increases the matching score when the distance between the current and the previous depth value are larger. Finally, we obtain the depth sequence with high temporal consistency and low error regions since the current depth value that is similar to the previous depth value is determined.

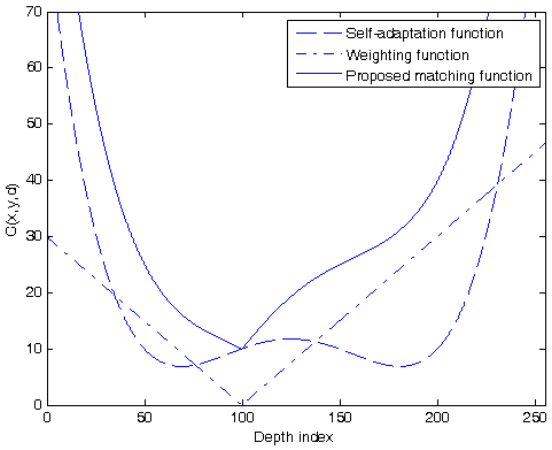



Figure 5. Graph of the proposed matching function

### 3. 2 Temporal postprocessing of the depth map

Figure 6 shows the block diagram of the temporal postprocessing. The proposed scheme consists of two parts: background mask extraction and temporal postprocessing.

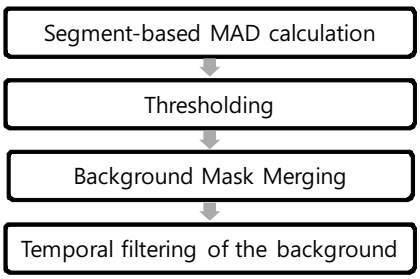



Figure 6. Block diagram of the proposed scheme

3. 2 .1 Background mask extraction

We assume that the camera arrays are fixed so that the background is not changed. In order to obtain the background mask, we first calculate the difference between consecutive two frames. Especially, we calculate segment-based MAD (mean absolute difference) twice by using the segments of those two frames bidirectionally. Figure 7 illustrates the calculation of segment-based MAD between two consecutive frames.

After calculating the differences in the different direction, we apply thresholding to them and classify the foreground and the background. Finally, we combine the forward and the backward background masks so as to obtain the background mask.

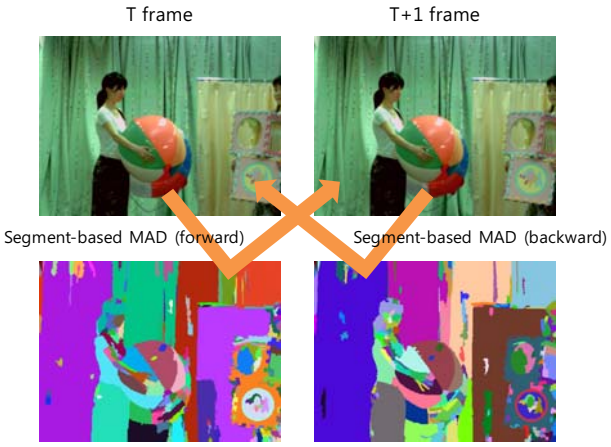



Figure 7. Calculation of segment-based MAD

3. 2. 2 Temporal postprocessing

In order to enhance the temporal coherence and the credibility, we need several frames of depth maps as inputs of postprocessing and merge those frames of masks. If we merge too many masks, the background area becomes smaller because the moving range of the foreground objects becomes larger. Because of this, we use five frames of depth maps. Figure 8 shows the background mask merging. As shown in Fig. 8, we separate the background and the foreground object by using the masks.

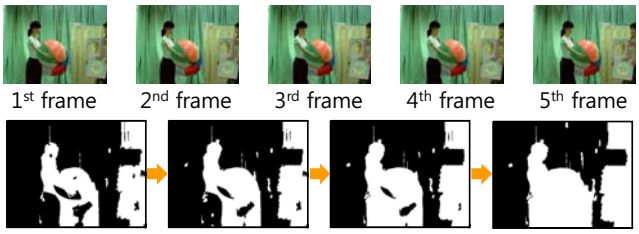



Figure 8. Background mask merging

After merging the masks, we apply a median filter to all the pixels of the background in the temporal direction. The median filter detects the most reliable depth value for several frames of the depth sequence. Figure 9 shows the temporal filtering method of depth sequence for the background area.

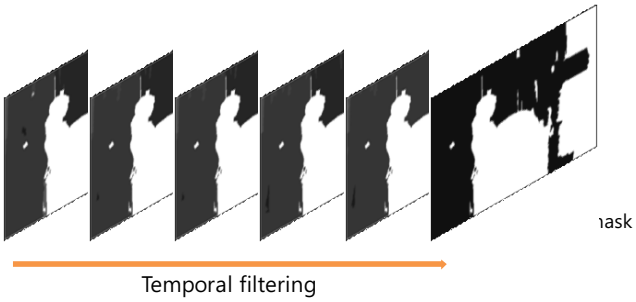
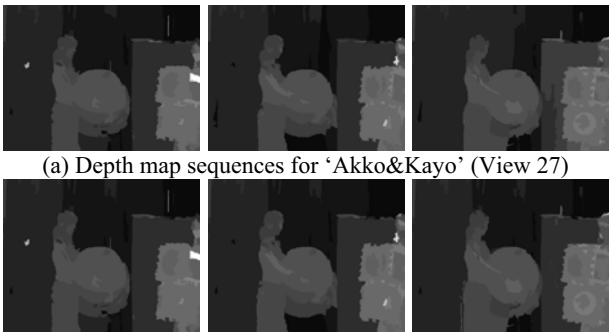



Figure 9. Temporal filtering for the background

## 4. Experimental Results

### 4. 1 Depth Sequence Estimation Results

In order to evaluate the proposed scheme, we used the rectified 'Akko&Kayo' test sequence (View 26, 27, 28) provided by Nagoya University. We used depth estimation scheme using 3D warping without refinement [5]. Figure 10 shows the experimental results of the temporal postprocessing. We notice that the background has the identical depth values and that the temporal coherence of the depth map sequences is increased.



(a) Depth map sequences for 'Akko&Kayo' (View 27)

(b) Depth map sequences after postprocessing

Figure 10. Background mask merging

Figure 11 demonstrates the depth estimation results for applying the temporally weighted matching function. As shown in Fig. 4 and Fig. 11, the proposed scheme efficiently removes the errors in the background area.

(a) t=2 (b) t=3

Figure 11. Depth sequence of the proposed scheme

### 4. 2 Comparison of the Temporal Consistency

In order to evaluate the temporal consistency, we compared the compression results. For the comparison of the previous method and the proposed scheme, we tested two depth sequences for 10 frames and compared the PSNR and bitrate. Table 1 shows the compression results. We noticed that the proposed scheme has a higher PSNR value and a lower bitrate. The proposed method reduced 30.6% of the bitrate and improved 3.02 dB of PSNR on average.

Table 1. Compression results

| QP | PSNR (dB) | | Bitrate (kbit/s) | |
|---|---|---|---|---|
| | No weighting | Proposed scheme | No weighting | Proposed scheme |
| 28 | 46.43 | 47.38 | 830.64 | 653.23 |
| 32 | 43.01 | 43.88 | 594.31 | 455.47 |
| 36 | 39.97 | 40.85 | 397.20 | 307.08 |
| 40 | 37.11 | 38.04 | 261.89 | 204.86 |

Figure 12 shows the rate-distortion curve for the Table 1. As shown in Fig. 12, the proposed scheme efficiently encodes the sequence compared to the previous method.

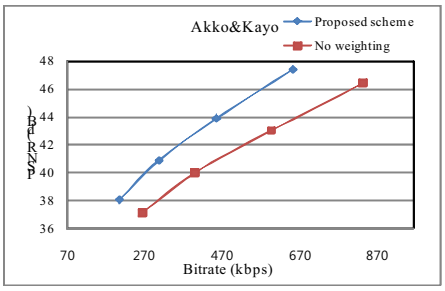



Figure 12. Rate-distortion curve

## 5. Conclusion

In this paper, we proposed a multi-view depth map estimation scheme considering the temporal consistency. We applied a temporally weighted matching function to consider the previous depth when calculating the matching score of the current frame. We added the weighting function to the existing matching function. We also applied the postprocessing scheme for the background area of a depth sequence. We extracted the background mask and we bidirectionally calculated the differences and combined two masks. After merging several frames of masks, we applied a median filter to the whole pixels of the background area. Experimental results have showed that the proposed scheme efficiently increased the temporal consistency.

## Acknowledgement

This work was supported in part by ITRC through RBRC at GIST (IITA-2008-C1090-0801-0017).

## References

[1] A. Smolic, K. Müller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D Video and Free Viewpoint Video – Technologies, Applications and MPEG Standards," IEEE International Conference on Multimedia and Expo, pp. 2161-2164, July 2006.

[2] D. Sharstein and R. Szeliski, "A Taxonomy and Evaluation of Dense Two-frame Stereo Correspondence Algorithms," Proc. of IEEE Workshop on Stereo and Multi-Baseline Vision, pp. 131-140, Dec. 2001.

[3] A. Klaus, M. Sormann, and K. Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure," Proc. of International Conference on Pattern Recognition, vol. 3, pp. 15-18, 2006.

[4] D. Comaniciu and P. Meer, "Mean shift: A Robust Approach toward Feature Space Analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, pp. 603-619, May 2002.

[5] S. Lee, K. Oh, and Y. Ho, "Segment-based Multi-view Depth Map Estimation Using Belief Propagation from Dense Multi-view Video," Proc. of 3DTV Conference, May 2008. (to be published)