DEPTH PIXEL CLUSTERING FOR CONSISTENCY TESTING OF MULTIVIEW DEPTH

Pravin Kumar Rana and Markus Flierl

ACCESS Linnaeus Center, School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden
{prara, mflierl}@kth.se

ABSTRACT

This paper proposes a clustering algorithm of depth pixels for consistency testing of multiview depth imagery. The testing addresses the inconsistencies among estimated depth maps of real world scenes by validating depth pixel connection evidence based on a hard connection threshold. With the proposed algorithm, we test the consistency among depth values generated from multiple depth observations using cluster adaptive connection thresholds. The connection threshold is based on statistical properties of depth pixels in a cluster or sub-cluster. This approach can improve the depth information of real world scenes at a given viewpoint. This allows us to enhance the quality of synthesized virtual views when compared to depth maps obtained by using fixed thresholding. Depth-image-based virtual view synthesis is widely used for upcoming multimedia services like three-dimensional television and free-viewpoint television.

Index Terms— Depth map enhancement, depth pixel clustering, hypothesis testing, inter-view connection information.

1. INTRODUCTION

Three-dimensional television (3D TV) and free-viewpoint television (FTV) are emerging visual media applications that use multiview imagery [1]. 3D TV aims to provide a natural 3D-depth impression of dynamic 3D scenes, while FTV enables viewers to freely choose their viewpoint of real world scenes. In conventional multiview systems, view synthesis is required for smooth transitions among captured views. Usually, view synthesis uses multiple views and depth maps acquired from different viewpoints. Each depth map gives information about the distance between the corresponding camera and the objects in the real world scene. Depth maps for chosen viewpoints are estimated by establishing stereo correspondences only between nearby views [2]. However, the estimated depth maps of different viewpoints usually demonstrate only a weak inter-view consistency [3]. Furthermore, depth estimation does usually not consider inherent temporal similarities among the multiview imagery, which results in temporal depth inconsistency. Several methods have been proposed to enhance the temporal consistency of depth, such as [4] and [5]. Nonetheless, any inconsistency of depth affects the quality of view synthesis negatively and, hence, FTV users experience visual discomfort.

In [3], we addressed the problem of inter-view inconsistency by testing connection evidence among multiple depth values from various viewpoints and using a single hard connection threshold per frame. In contrast to [3], this paper defines clusters of depth pixels and chooses adaptive connection thresholds based on the statistics of each cluster under consideration. First, the proposed clustering algorithm assigns each principal depth pixel to a particular cluster. If the current cluster satisfies the sub-clustering requirements, the algorithm will split this cluster into two sub-clusters. Second, based on the cluster or sub-cluster statistics, a connection threshold is defined and the corresponding connection evidence is tested for depth pixels that belong to this cluster. Finally, by utilizing the resulting connection information, inter-view consistent depth maps are generated. Further, these depth maps improve the visual quality of the synthesized virtual views.

The remainder of the paper is organized as follows: In Section 2, we summarize the depth consistency testing algorithm (DCTA) to obtain the inter-view consistent depth map. In Section 3, we describes the proposed clustering algorithm of depth pixels. We use the depth consistency information for view synthesis in section 5. The experimental results for virtual view synthesis are given in Section 6.

2. DEPTH CONSISTENCY TESTING

As estimated depth maps usually show weak inter-view consistency, we proposed a method in [3] to achieve inter-view depth consistency at a given viewpoint. As summarized in Fig. 1, the algorithm warps more than two depth maps from multiple reference viewpoints to a principal viewpoint. The principal viewpoint may or may not coincide with any reference viewpoint. However, the principal viewpoint should be close to the reference viewpoints. At this stage, the reference depth maps are used for 3D warping [6]. Holes that occur during warping are masked for testing.

In the next stage, the consistency among all warped depth values at the principal viewpoint are examined. To assess consistency, the absolute differences between all possible

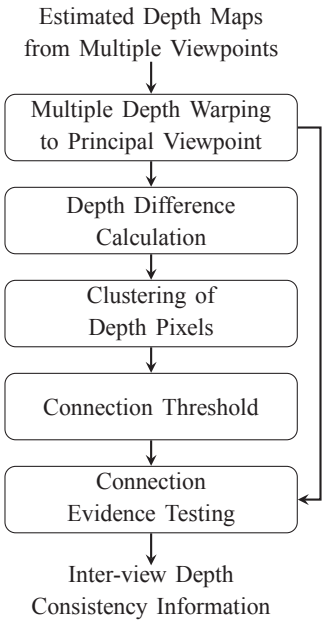


Fig. 1. Block diagram for depth consistency testing.

pairs of depth values for each given principal pixel are determined. For example, with n reference views, there are up to $N = \frac{n!}{(n-2)!2!}$ possible pairs of depth values for a given principal pixel p. This can be represented by the following skew-symmetric matrix

$$\mathbf{D}_p = \begin{pmatrix} 0 & \Delta_{1,2} & \dots & \Delta_{1,n} \\ \Delta_{2,1} & 0 & \dots & \Delta_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ \Delta_{n,1} & \Delta_{n,2} & \dots & 0 \end{pmatrix}, \quad (1)$$

where $\mathbf{D}_p$ is the difference matrix of all possible pairs of the depth values at the principal pixel p for a given frame, and $\Delta_{j,k} = d_j - d_k$ is the difference of depth values between warped depth map j and warped depth map k at p. Indexes $j, k = \{1, \dots, n\}$ represent the warped views from different viewpoints. Since $\mathbf{D}_p$ is skew-symmetric with diagonal elements being zero, its upper triangular part is sufficient for testing. Each $\Delta_{j,k}$ is an *inter-view connection evidence*, which is a measure of depth consistency between the corresponding depth pairs (d_j, d_k) at the principal pixel. If an inter-view connection evidence is below a given connection threshold, the testing algorithm marks the corresponding depth pair at the principal pixel as connected, and assumes that the depth pair relates to the same 3D point in the world. This connectivity information is then used to estimate the final depth value at the principal pixel.

3. DEPTH PIXEL CLUSTERING

Depth pixel clustering will allow us to exploit the local statistical properties for depth consistency testing. Given the depth difference matrix for all principal depth pixels, we classify these pixels into three clusters as follows:

Cluster $\mathcal{A}$: Pixel $p \in \mathcal{A}$, if

$$[\mathbf{D}_p]_{j,k} = 0 \ \forall\, j, k;\ j < k. \quad (2)$$

For pixels $p \in \mathcal{A}$, the corresponding warped depth values from all reference depth maps are consistent and describe the same 3D point in world coordinates.

Cluster $\mathcal{B}$: Pixel $p \in \mathcal{B}$, if

$$\exists (j, k) : [\mathbf{D}_p]_{j,k} = 0;\ j < k. \quad (3)$$

At least one guaranteed consistent depth pixel pair exists in cluster $\mathcal{B}$ for pixels $p \in \mathcal{B}$.

Cluster $\mathcal{C}$: Pixel $p \in \mathcal{C}$, if

$$[\mathbf{D}_p]_{j,k} \neq 0 \ \forall\, j, k;\ j < k. \quad (4)$$

The corresponding warped depth values from all reference depth maps are inconsistent for pixels $p \in \mathcal{C}$.

3.1. Sub-clustering of Cluster $\mathcal{C}$

All depth pixels belonging to $\mathcal{C}$ are inconsistent. Furthermore, if the number of pixels in the cluster $\mathcal{C}$ is large and exceeds the maximum number of pixels P_{max}, then cluster $\mathcal{C}$ is split into two sub-clusters $\mathcal{C}_1$ and $\mathcal{C}_2$. For our experiments $P_{max} = 45\%$ is chosen. We assign each pixel $p \in \mathcal{C}$ to one of these sub-clusters. The assignment is based on relative evidence.

First, we sort the off-diagonal elements of $\mathbf{D}_p$, $p \in \mathcal{C}$, in ascending order.

$$\theta_{p,i} = \min\{|\Delta_{j,k}| \mid |\Delta_{j,k}| > \theta_{p,i-1}, j < k\} \quad (5)$$

for $i = 1, \dots, N$, where $\theta_{p,0} = 0$. This sorting gives $\theta_{p,max} = \theta_{p,N}$. Second, we define the i^{th} relative evidence, $\tilde{\theta}_{p,i} = \frac{\theta_{p,i}}{\theta_{p,\max}}$ for the pixel p and the vector

$$\rho_i = [\tilde{\theta}_{1,i}\ \tilde{\theta}_{2,i}\ \dots\ \tilde{\theta}_{M,i}], \quad (6)$$

where M is the total number of pixels in $\mathcal{C}$. Finally, we define the sub-clustering criteria for each $p \in \mathcal{C}$ according to

$$\forall p \in \mathcal{C} : \begin{cases} p \in \mathcal{C}_1 & \text{if } \exists i : \tilde{\theta}_i \leq min\{\eta\}, \\ p \in \mathcal{C}_2 & \text{otherwise,} \end{cases} \quad (7)$$

where $\eta = \{\eta_1, \eta_2, \dots, \eta_N\}$ and $\eta_i = median\{\rho_i\}$. Fig. 2 shows these clusters and sub-clusters for a test depth map.

4. CONNECTION THRESHOLD

For a given principal pixel p in a cluster, say $\mathcal{C}$, we test each inter-view connection evidence by checking the corresponding value of $\Delta_{j,k} \in \mathbf{D}_p$. As a result of this testing, we get the inter-view connection information across multiple reference views and, hence, consistent information about depth pixels for the given principal pixel. If an inter-view connection evidence $\Delta_{j,k}$ is less than a given connection threshold T_c, the evidence is accepted and it is assumed that the corresponding two depth values in the two warped depth maps are consistent for the given principal pixel. Hence, these depth pixels have

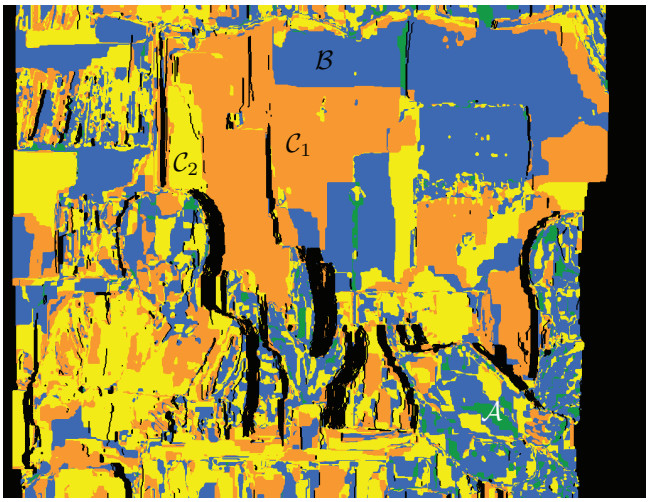


Fig. 2. Depth pixel clusters for the Newspaper sequence, where clusters $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}_1$, and $\mathcal{C}_2$ are depicted by green, blue, orange, and yellow colors, respectively. Black color depicts masked hole areas for the consistency testing.

a consistent depth representation of the corresponding 3D object point in world coordinates. Otherwise, the connection evidence is rejected and it is assumed that the corresponding two depth pixels in the two warped depth maps do not have a consistent depth representation. The connection threshold relates to the quality of the connectivity and defines a criterion for depth consistency testing for all pixels in a given cluster.

Fig. 3 shows the distribution of all depth differences depending on the cluster. Due to varying variances of $\Delta_{j,k}$ in each cluster, it is efficient to define individual connection thresholds for each cluster. For example, we define the connection threshold for cluster $\mathcal{C}$ as

$$T_c = \frac{\sigma_{\mathcal{C}}}{2}, \tag{8}$$

where $\sigma_{\mathcal{C}}$ is the standard deviation of $\mathbf{D}_{\mathcal{C}} = \{\mathbf{D}_p | p \in \mathcal{C}\}$.

As each principal pixel falls into a specific cluster or a sub-cluster, we use the corresponding connection threshold to test the inter-view connection evidence in the corresponding depth difference matrix. Based on the inter-view connection evidence, various cases of inter-view connectivity can arise, as depicted in Fig. 4. The different cases of inter-view connectivity and the corresponding pixel selection from multiple reference viewpoints are tested. For any accepted connection, the resulting inter-view connectivity information is used to choose a consistent depth value for the corresponding principal pixel.

5. VIRTUAL VIEW SYNTHESIS

Now, we consider the principal viewpoint as a virtual viewpoint and use the resulting inter-view connectivity information for virtual view synthesis. If the viewpoint to be tested coincides with any reference viewpoint, we may use the virtual view synthesis approach as proposed in [7]. From the connectivity information, we have information about reference viewpoints and corresponding pixel positions which have a consistent depth description of the same 3D object point for each virtual pixel. Consequently, we warp the pixel intensity values of the reference views to the virtual viewpoint using the corresponding connectivity information.

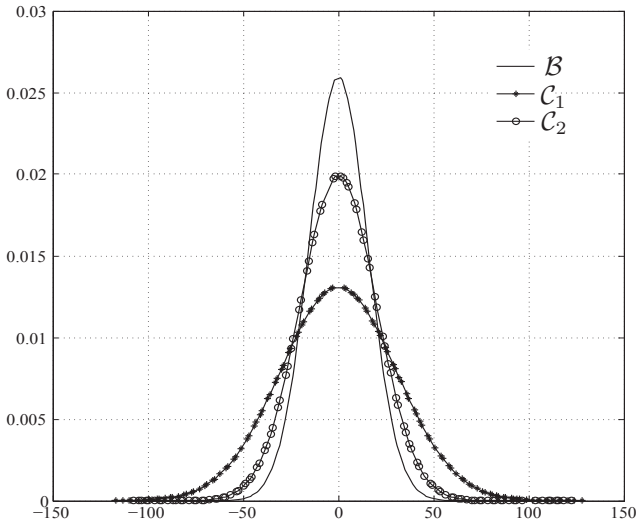


Fig. 3. Distribution of $\mathbf{D}_{\mathcal{B}}$, $\mathbf{D}_{\mathcal{C}_1}$, and $\mathbf{D}_{\mathcal{C}_2}$.

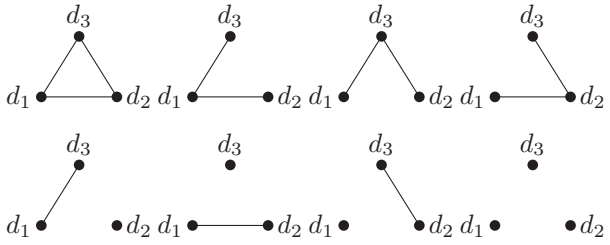


Fig. 4. Different cases of inter-view connectivity with $n = 3$.

If there is no inter-view connection information available for a given virtual pixel, the proposed rendering method is not able to determine a virtual pixel intensity from the reference views. In this case, we set a mask for inpainting [8] to determine the pixel intensities for such unconnected inter-view pixels in the virtual view. However, if there is inter-view connection information available for a given virtual pixel, we use the connectivity information to warp the specified pixels in the reference views to the virtual viewpoint.

To determine the final pixel intensity in the virtual view, we use various approaches depending on the baseline scenario and the varying illumination condition among reference views. If the pixel intensities of inter-view connected reference pixels are similar, averaging of the warped pixel intensities is feasible. However, if the pixel intensities among the connected and warped texture views vary significantly due to

varying illumination, we assume that the virtual pixel value is best described by the warped texture pixel of the nearest reference view. The reference view which has minimum baseline distance from the virtual viewpoint is defined as the nearest view. In this case, we simply set the pixel intensity in the virtual view by copying the pixel intensity from the warped texture pixel of the nearest reference view that is connected. If the reference views are captured from multiple viewpoints using irregular camera baseline distances between viewpoints, we estimate the virtual pixel intensity by weighted-baseline averaging of the connected and warped pixel intensities. Further, to determine the virtual pixel values, the advantage of color consistency testing may also be used.

When reference views are warped to a virtual viewpoint, some areas may be affected by occlusion. Increasing the number of reference views is likely to decrease occlusion areas. However, occlusions cannot be ruled out completely. Therefore, occlusions are detected by checking the generated depth map at the virtual view. If some holes remain due to unconnected pixels, their intensity values are filled by inpainting.

6. EXPERIMENTAL RESULTS

To evaluate the proposed algorithm, we use the video test material and the corresponding depth maps as provided by MPEG [9]. In the experiments, we assess the quality of the synthesized virtual view by using the proposed clustering algorithm for a view configuration with three reference views. First, we test the depth consistency to obtain the inter-view connection information at the virtual viewpoint by utilizing the proposed clustering approach. Second, a view at the virtual viewpoint is synthesized by using the obtained inter-view connectivity information. We measure the objective video quality of the synthesized virtual view in terms of the PSNR with respect to the captured view of a real camera at the same viewpoint. We also study the effect of depth consistency testing and clustering for quantized depth maps.

Furthermore, the proposed algorithm is also compared to virtual views as synthesized by the MPEG 3DV/FTV View Synthesis Reference Software 3.5 (VSRS 3.5) [10], [11]. To synthesize a virtual view, VSRS 3.5 uses two reference views, left and right, and utilizes the two corresponding reference depth maps. The reference software employs mainly pixel-by-pixel mapping of depth maps and texture views, hole filling, view blending, and inpainting of remaining holes. We synthesize virtual views by using the general synthesis mode with half-pel precision.

Table 1 shows a comparison of the average PSNR (in dB) of the synthesized virtual views over 50 frames. The quality of the reference depth maps is indicated by the QP value. The value zero indicates that the provided MPEG depth maps are used. VSRS 3.5 is used in column (a). In column (b), depth consistency testing uses a fixed connection threshold per frame. Column (c) gives the results for the clustering approach and column (d) reflects the additional sub-clustering. The proposed methods do not offer gains for the original synthetic test material, i.e., the Dancer sequence. This is because the synthetic depth information is consistent across all viewpoints. However, downsampling, quantization, and upsampling introduce inconsistencies in the synthetic depth information. Here, our clustering approach can provide gains up to 0.6 dB for highly quantized content . For unquantized content, we gain up to 0.1 dB. However, the proposed algorithm offers improvements of up to 1.8 dB when compared to VSRS 3.5. The improvement of the quality depends on the input reference depth maps and the level of quantization.

Table 1. Objective quality of synthesized virtual views

Test Material (Virtual View)	QP	VSRS 3.5 (a)	DCTA Supported View Synthesis [dB] No Cluster (b)	Cluster (c)	Sub-cluster (d)
Dancer (3)	0	37.00	38.72	38.72	38.72
	28	35.50	36.85	37.31	37.31
	42	33.02	33.52	34.15	34.15
Kendo (4)	0	37.10	38.05	38.12	38.12
	34	36.60	37.56	37.60	37.60
	40	36.25	37.17	37.21	37.20
Balloons (4)	0	35.46	35.83	35.85	35.91
	34	35.00	35.03	35.12	35.10
	40	34.73	34.76	34.87	34.83
Newspaper (5)	0	32.54	33.32	33.32	33.31
	35	32.37	33.30	33.30	33.31
	40	32.16	33.17	33.17	33.20

Fig. 5 shows synthesized virtual views using quantized depth maps for subjective evaluation. The proposed clustering reduces the artifacts around the hand for the Dancer sequence and for the Kendo sequence, as shown in Fig. 5 (a) and Fig. 5(b), respectively. The balloons and corresponding surrounding areas in the Balloons sequence are improved, when compared to the view synthesized without depth pixel clustering; see Fig. 5(c). Furthermore, in Fig 5(d), the blue sweater and the background in the Newspaper sequence are well synthesized by our clustering method.

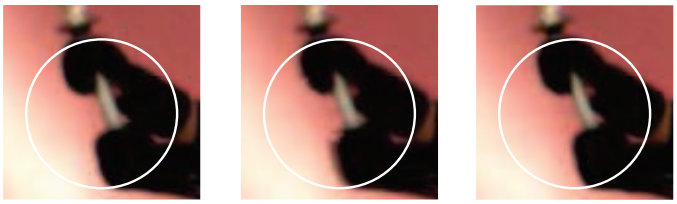
Regarding the increase in complexity when compared to [3], the proposed algorithm adds only the mapping of depth pixels to clusters and sub-clusters as well as the usage of different thresholds. On an implementation level, this can be handled by memory lookup, which increases complexity only marginally.

7. CONCLUSIONS

In this paper, we proposed a clustering algorithm for depth pixels based on the inter-view consistency of depth pixels. Depending on the statistical properties of the inter-view consistency, we determine improved depth information for virtual viewpoints. This approach is useful for depth maps that have been deteriorated by coding. Moreover, the resulting connection information can help to improve the visual quality of synthesized virtual views.

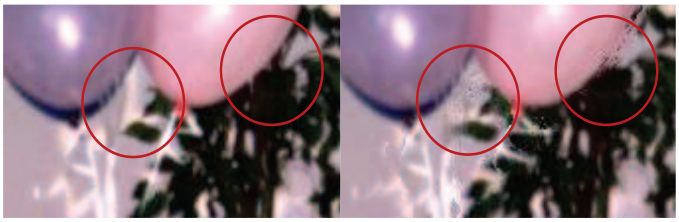
Original. Without clustering. With clustering.

(a) Dancer with QP = 42.



Original. Without clustering. With clustering.

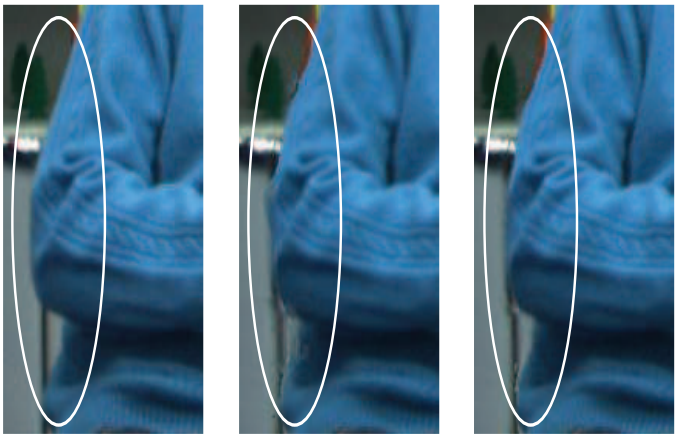
(b) Kendo with QP = 40.



Original. Without clustering.

With clustering.

(c) Balloons with QP = 40.



Original. Without clustering. With clustering.

(d) Newspaper with QP = 40.

Fig. 5. Effect of depth pixel clustering on synthesized virtual views using quantized depth maps.

8. ACKNOWLEDGMENT

This work was supported in part by Ericsson AB and the ACCESS Linnaeus Center at KTH Royal Institute of Technology, Stockholm, Sweden.

9. REFERENCES

[1] M. Tanimoto, "Overview of free viewpoint television," in *Signal Processing: Image Communication*, vol. 21, no. 6, pp. 454–461, Jul. 2006.

[2] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *International Journal of Computer Vision*, vol. 47, pp. 7–42, Apr. 2002.

[3] P. K. Rana, and M. Flierl, "Depth consistency testing for improved view interpolation," in *Proc. of the IEEE International Workshop on Multimedia Signal Processing*, Saint Malo, France, pp. 384–389, Oct. 2010.

[4] C. Cigla and A. A. Alatan, "Temporally consistent dense depth map estimation via Belief Propagation, " in *Proc. of the 3DTV-CON*, Potsdam, Germany, pp 1–4, May 2009.

[5] S. Lee and Y. Ho, "Temporally consistent depth map estimation using motion estimation for 3DTV, " in *International Workshop on Advanced Image Technology*, Kuala Lumpur, Malaysia, pp. 149(1–6), Jan. 2010.

[6] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D TV," in *Proc. of the SPIE Stereoscopic Displays and Virtual Reality Systems XI*, San Jose, USA, pp. 93-104, Jan. 2004.

[7] P. K. Rana, and M. Flierl, "View interpolation with structured depth from multiview video," in *Proc. of the European Signal Processing Conference*, Barcelona, Spain, pp. 383–387, Aug. 2011.

[8] M. Bertalmio, A. L. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proc. of the IEEE Conference on CVPR*, Kauai, HI, USA, vol. 1, issue 1063-6919, pp. 355–362, Dec. 2001.

[9] MPEG, "Call for proposals on 3d video coding technology," Tech. Rep. N12036, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Mar. 2011.

[10] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference Softwares for Depth Estimation and View Synthesis, " Tech. Rep. M15377, ISO/IEC JTC1/SC29/WG11, Archamps, France, Apr. 2008.

[11] MPEG, "View synthesis software manual," ISO/IEC JTC1/SC29/WG11, Sept. 2009, release 3.5.