Depth Consistency Testing for Improved View Interpolation

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*—Multiview video will play a pivotal role in the next generation visual communication media services like three-dimensional (3D) television and free-viewpoint television. These advanced media services provide natural 3D impressions and enable viewers to move freely in a dynamic real world scene by changing the viewpoint. High quality virtual view interpolation is required to support free viewpoint viewing. Usually, depth maps of different viewpoints are used to reconstruct a novel view. As these depth maps are usually estimated individually by stereo-matching algorithms, they have very weak spatial consistency. The inconsistency of depth maps affects the quality of view interpolation. In this paper, we propose a method for depth consistency testing to improve view interpolation. The method addresses the problem by warping more than two depth maps from multiple reference viewpoints to the virtual viewpoint. We test the consistency among warped depth values and improve the depth value information of the virtual view. With that, we enhance the quality of the interpolated virtual view.**

I. Introduction

Due to advances in camera and display technology, three-dimensional television (3D-TV) and free-viewpoint television (FTV) are emerging visual media applications [1]–[3]. Such advanced applications offer highly interactive 3D experiences to viewers. Rapidly dropping costs of digital cameras enable us to acquire multiview imagery by utilizing arrays of cameras to capture dynamic 3D scenes from various viewpoints simultaneously. 3D-TV aims to provide natural 3D-depth impressions of dynamic 3D scenes. FTV enables viewers to freely choose their viewpoint of real world scenes. In both cases, view interpolation is used to facilitate smooth transitions among captured views.

View interpolation algorithms may use multiple views acquired from different viewpoints and their corresponding depth maps. Each depth map gives information about the distance between the corresponding camera and the objects in the 3D scene. The quality of interpolated views depends on the accuracy of the estimated depth maps. Algorithms for depth estimation are widely used in computer vision [4]. These vision algorithms estimate depth information mainly by establishing stereo correspondences using belief propagation [5], graph-cut methods [6] and segmentation [7]. These methods usually estimate depth information for different viewpoints independently by using nearby views only. Given such estimates, the resulting depth informations at different viewpoints has weak spatial consistency, as shown in the Figure 1. These inconsistent depth maps affect the quality of view interpolation negatively.

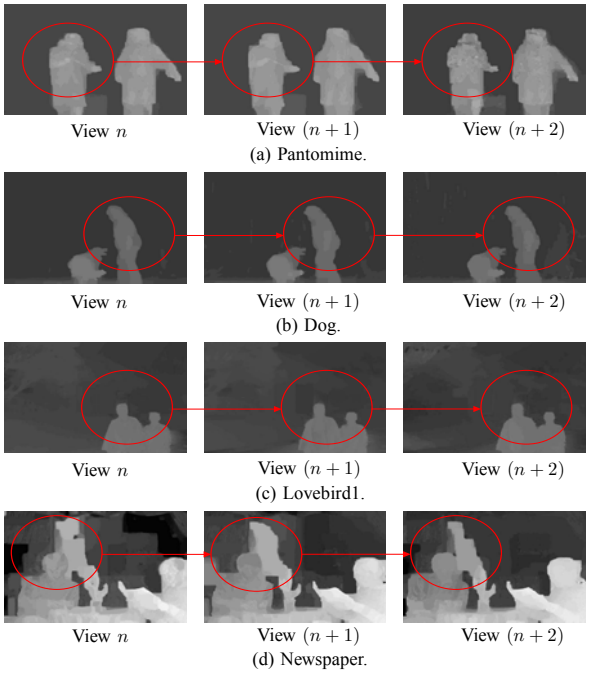


Fig. 1. Inconsistency among the estimated depth maps at different viewpoints for four multiview video sequences. The red circles mark the most prevailing inconsistent areas in the depth maps.

In this paper, we propose a novel method to address the problem of weak depth map consistency. First, the method warps more than two depth maps from multiple reference viewpoints to the virtual viewpoint by 3D warping [8]. Second, it tests the consistency among all warped depth values at the given virtual viewpoint. Third, an enhanced depth map is generated that is used to improve the quality of the interpolated view at the virtual viewpoint.

The remainder of this paper is organized as follows: In Section II, we introduce the algorithm for depth consistency testing. In Section III, we use the enhanced depth information for view interpolation. In Section IV, we present experimental results on view interpolation. Finally, Section V gives concluding remarks.

MMSP'10, October 4-6, 2010, Saint-Malo, France.
978-1-4244-8112-5/10/$26.00

II. Depth Consistency Testing

As conventional depth estimation algorithms estimate depth maps for different viewpoints separately, the resulting depth maps may have weak spatial consistency. Inconsistent depth maps affect the quality of view interpolation at a given viewpoint. We addresses the problem by first warping more than two depth maps from multiple reference viewpoints to the virtual viewpoint by using 3D warping techniques. We examine the consistency among all warped depth values by taking the absolute difference between all possible pairs of depth values at the virtual viewpoint. We use the resulting consistency information among various depth values at the virtual viewpoint for warping corresponding pixels of the multiple reference views to the virtual viewpoint. These warped pixels from the multiple reference views are used for view interpolation at the virtual viewpoint. First, we will explain the warping of the depth maps from multiple viewpoints. Then, we will examine the consistency among all warped depth values at the virtual viewpoint. Figure 2 shows the block diagram of the depth consistency testing algorithm and its application to virtual view interpolation.

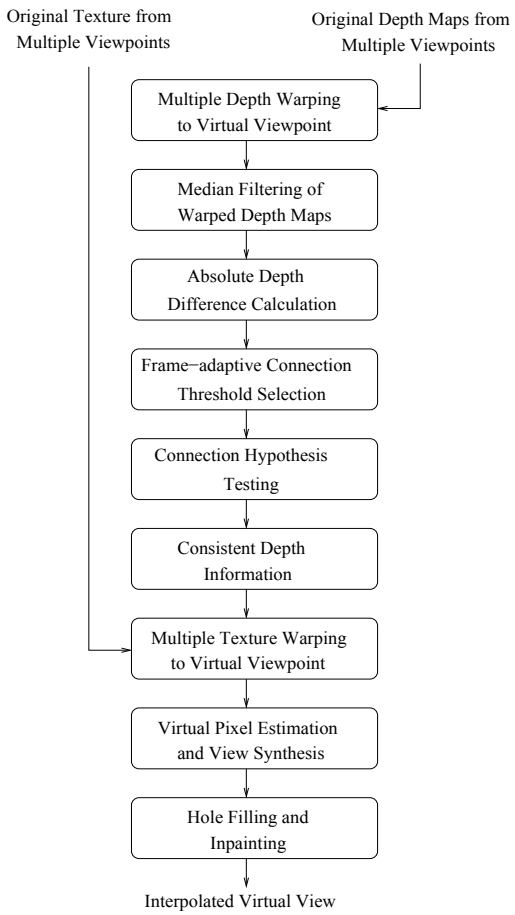


Fig. 2. Block diagram of the depth consistency testing algorithm with application to view interpolation.

A. Multiple Depth Warping and Connection Hypotheses

The 3DV/FTV ad-hoc group of MPEG is currently working on the Depth Estimation Reference Software (DERS) [9], which is provided by Nagoya University. It uses left, intermediate and right viewpoint views as a reference to estimate the depth map of the intermediate viewpoint. It incorporates many matching features, like pixel-matching, block-matching, and soft-segmentation matching for estimation of depth maps. It uses graph-cut algorithms for disparity enhancement, sub-pel precision, and temporal enhancement. We estimate depth maps of reference viewpoints by using DERS 5.0 [10]. We warp the estimated depth maps from multiple reference viewpoints to the virtual viewpoint by 3D warping [11] with the reference depth map itself.

3D warping is a depth-image-based rendering technique for rendering novel views from arbitrary viewpoints. The novel view is projected from reference views by using depth information and camera calibration parameters. The latter describe the relationship between camera coordinates (CCs) and the world coordinates (WCs). Each camera is described by one intrinsic parameter matrix A and two extrinsic parameters, the rotation matrix R and the translation vector t. When a point I in WCs is projected to reference CCs, a pixel i_r in the reference view is obtained by the projective relation

$$i_r = A_r[R_r|t_r]I. \tag{1}$$

If we know the virtual camera calibration parameters (A_v, R_v, t_v), the relation between the WC of the 3D point I and the corresponding virtual view pixel i_v is obtained by

$$i_v = A_v[R_v|t_v]I, \tag{2}$$

where $I = [X, Y, Z, 1]^T$ is the point in homogeneous WCs and $i = [x, y, 1]^T$ the projected point in homogeneous pixel coordinates. Expressing point coordinates in homogeneous form has many advantages [12]. For example, multiple matrices, including intrinsic parameter matrix A, the rotation matrix R, and the translation vector t, can be combined efficiently into a single matrix. We can project the reference view pixel i_r back into a reference 3D point I_r in homogeneous WCs by

$$I_r = R_r^{-1}A_r^{-1}i_r d_r - R_r^{-1}t_r, \tag{3}$$

where d_r is a depth value for the position of i_r in the reference view. The relationship between pixels in the reference view and in the virtual view of the same 3D object point is obtained by projecting I_r into the virtual CCs by using the relation [12]

$$i_v = A_v[R_v|t_v]I_r. \tag{4}$$

This relation describes the 3D warping of the reference view to the virtual viewpoint by using depth information from the reference view.

Usually, depth maps are 8-bit single-channel gray value images [8]. The gray value zero represents the farthest value and 255 the closest. We warp the depth maps from various reference viewpoints to the virtual viewpoint by 3D warping. During 3D warping of depth maps, small holes occur due

to the change in viewpoint. These small holes are filled by using a 3x3 median filter. This reduces also effects due to rounding operations during warping. We have now multiple warped depth maps from various reference views for a given virtual viewpoint.

Next, we examine the consistency among all warped depth values in the virtual view. We consider the absolute differences between all possible pairs of depth values for each given pixel in the virtual view position. For example, with n reference views, there are up to $\mathrm{N} = \frac{n!}{(n-2)!2!}$ different possible pairs of depth values for a given pixel. This can be represented by the following symmetric matrix

$$ADM = \begin{pmatrix} 0 & h_{1,2} & h_{1,3} & \dots & h_{1,n} \\ h_{1,2} & 0 & h_{2,3} & \dots & h_{2,n} \\ h_{1,3} & h_{2,3} & 0 & \dots & h_{3,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ h_{1,n} & h_{2,n} & h_{3,n} & \dots & 0 \end{pmatrix}, \quad (5)$$

where ADM is the absolute difference matrix of all possible pairs of depth values at the virtual viewpoint pixel i_v in the frame f, and $h_{j,k} = |d_j - d_k|$ is the absolute difference of depth values between warped depth map j and depth map k at the virtual pixel i_v. $j, k = \{1, 2, \dots, n\}$ represent the warped views from different viewpoints. Since ADM is symmetric with diagonal elements being zero, its upper triangular part is sufficient for our considerations.

Now, each $h_{j,k}$ relates to the depth pixel connectivity between warped depth map j and k at pixel i_v, and is a measure of spatial consistency among multiple depth maps of different viewpoints. We call each a *connection hypothesis*.

B. Connection Hypothesis Testing

We assume that each individual connection hypothesis $h_{j,k}$ is a measure of consistency between the corresponding two depth values d_j and d_k for a given virtual pixel i_v. We test each connection hypothesis by checking the corresponding value of $h_{j,k}$ in the ADM for a given pixel i_v. If an absolute depth difference is less than a given connection threshold, we accept the connection hypothesis and assume that the corresponding two depth values in the two warped depth maps are consistent for the given pixel. Hence, these consistent depth pixels at a given virtual viewpoint have a consistent depth representation of the corresponding 3D object point in the world coordinates. Otherwise, we reject the connection hypothesis and assume 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 pixel consistency. The connection threshold is estimated by considering statistics of all depth difference values in the ADM per frame. We calculate statistical mean μ_f and standard deviation σ_f of the ADM for frame f. We choose a connection threshold for frame f according to

$$T_f = \mu_f + \lambda\sigma_f, \quad (6)$$

where λ is a parameter which takes values in the interval $[0, 1]$. We manually select the appropriate λ for each test sequence.

Based on the number of connection hypotheses in the ADM per pixel, our method allows for various cases of connectivity. The different cases of connectivity and the corresponding reference pixel selection from multiple reference viewpoints are described below for four reference views n = 4. We distinguish among several cases.

Null-Hypothesis Case: No connection hypothesis is below the threshold. In this case, the absolute differences between all possible pairs of warped depth values for a given virtual pixel are above the connection threshold. Hence, we reject any connection hypothesis. We may use inpainting [13] to set intensity values for such pixels in the virtual view because we are not able to find a consistent depth representation of the corresponding 3D object point.

One-Hypothesis Case: One out of six connection hypotheses is below the threshold.

$$h_{j,k} < T_f \quad (7)$$

We accept this connection hypothesis because it satisfies our consistency criteria at the given virtual pixel i_v. The corresponding depth pixel pair (d_j, d_k) in the warped depth maps j and k are consistent at the virtual pixel i_v and describe the same 3D object point. Using this depth consistency information at the virtual pixel i_v and the 3D warping relations as described in subsection (II-A), the corresponding depth values $d_{j_r}(w_r)$ and $d_{k_r}(u_r)$ in the reference depth maps j_r and k_r are obtained at reference pixels w_r and u_r, respectively. Here, we assume that the depth information at the specified pixel pair (w_r, u_r) relates to the same 3D point in world coordinates. Hence, two reference depth maps from reference viewpoint j_r and k_r have consistent depth representation for a given virtual view pixel i_v.

Two-Hypothesis Case: Two out of six connection hypotheses are below the threshold. This is a special case for the four reference view configuration because two distinct pairs of warped depth values are individually consistent for a given virtual pixel. For example, reference maps j and k as well as l and m may be consistent according to $h_{j,k} < T_f$ and $h_{l,m} < T_f$, where j, k, l, and m represent four reference depth maps of four reference views. Figure 3 shows all possible cases of connectivity among four warped depth values in the two-hypothesis case at a given virtual pixel.

The two-hypothesis case for n=4 may lead to two constellations. In the first constellation, one connection hypothesis has a significantly lower value than the other. Here, the best connection hypothesis may be chosen to define connectivity. In the second constellation, both connection hypotheses have very similar or identical values. Here, we assume that the depth information from the two nearby warped depth maps is more accurate and relates to the same 3D point in world coordinates.

By using consistent depth information at a given virtual pixel and 3D warping, the corresponding consistent depth information in the reference depth maps can be obtained for

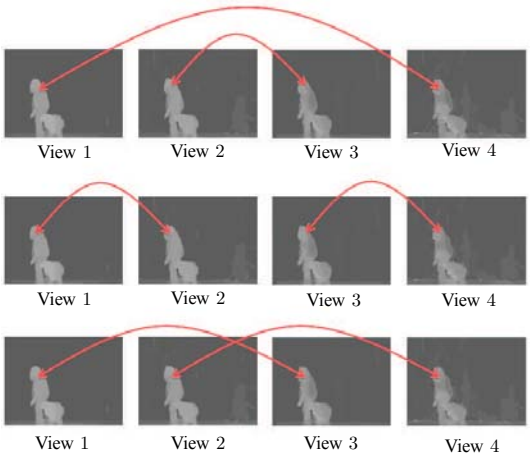


Fig. 3. All possible two-hypothesis cases for a four reference view configuration.

all reference pixels. This depth information of a specified pixel pair relates to the same 3D point in world coordinates.

Three-Hypothesis Case: Three out of six connection hypotheses are valid. In this case, the absolute differences between any three warped depth values are below the threshold. We now assume that the connection hypotheses are correct for three depth values at the given virtual pixel. The corresponding pixels in the three reference depth maps obtained by 3D backward-warping are consistent and describe the same 3D point in world coordinates. Hence, three reference depth maps at the reference viewpoints have a consistent depth representation for a given virtual view pixel.

Similarly, with any specified number of connection hypotheses below the threshold, we use the resulting depth pixel connectivity information for selecting consistent depth values from reference viewpoints. In the six-hypothesis case, all four pixels in the four reference depth maps are consistent and describe the same 3D point in world coordinates.

III. Application to View Interpolation

View interpolation has been extensively investigated in computer vision [14], [15]. It is an image-based rendering technique to generate a novel view by using a small set of captured views and estimated depth maps of a real 3D scene. In this paper, we use the resulting consistent depth information from our algorithm for view interpolation.

A. Multiple View Warping

For view interpolation, we warp pixel values of the reference views to a given virtual viewpoint by using our consistent depth map and additional information from hypothesis testing. Based on the number of accepted connection hypotheses, 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.

If there is no connection hypothesis for a pixel i_v in the virtual view, i.e., the null-hypothesis case, the proposed algorithm is not able to determine a consistent pixel intensity from the reference views. In this case, the algorithm sets a mask for inpainting to determine the intensity values for such pixels in the virtual view. If the one-hypothesis case applies to the pixel i_v, we use the determined connectivity between two reference depth pixels w_r and u_r to warp the two corresponding pixels w_r and u_r in the reference views j_r and k_r to the given virtual pixel i_v. Similarly, if there is a larger number of connection hypotheses satisfying the connection criteria, we use the determined connectivity 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 illumination-adaptive estimation techniques per virtual pixel. If the pixel intensities of 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, we assume that the virtual pixel value is best described by the warped texture pixel of the nearest reference view. In this case, we simply set the pixel intensity in the virtual view by copying the pixel intensity information 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.

B. Hole Filling and Inpainting

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 of the virtual view. If some holes remain due to unconnected pixels, they are filled by inpainting using the mask as defined by the null-hypothesis case.

IV. Experimental Results

In order to evaluate the performance of the proposed algorithm, we assess the quality of the synthesized virtual view by using the proposed algorithm. We measure the objective video quality of the synthesized view at a given viewpoint by means of the Peak Signal to Noise Ratio (PSNR) with respect to the view of a real camera at the same viewpoint. We use four standard multiview video test sequences. Two test sequences, Pantomime and Dog, are provided by Nagoya University [16]. The Newspaper test sequence is provided by GIST [17] and the Lovebird1 test sequence by ETRI [18]. The Nagoya University test sequences include 80 camera positions at pixel resolution 1280x960. Newspaper and Lovebird1 are captured from 9 and 12 camera positions, respectively, at a resolution of 1024x768 pixels.

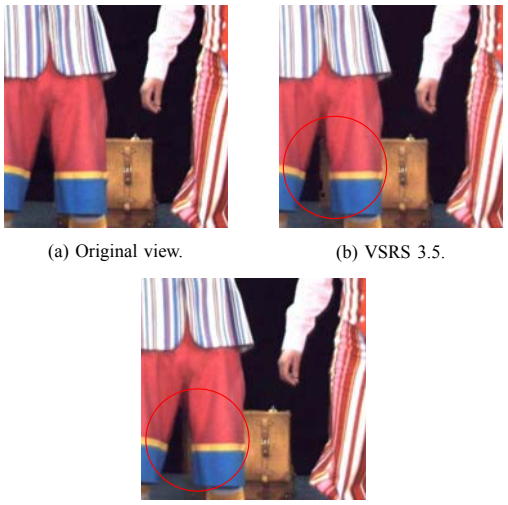
We evaluate the proposed algorithm for the four reference view configuration. For example, we synthesize the virtual view 40 of the Pantomime test sequence by using reference

views and corresponding depth maps at viewpoints 38, 39, 41, and 42. We estimate the required depth maps with DERS 5.0.

The proposed algorithm is compared to the synthesized view as obtained by the View Synthesis Reference Software 3.5 (VSRS 3.5) [9], [19]. VSRS 3.5 is provided by Nagoya University for MPEG 3DV/FTV exploration experiments. It uses two reference views, left and right, to synthesize the intermediate virtual view by using the two corresponding reference depth maps. The reference software mainly employs pixel-by-pixel mapping for depth maps and texture views, hole filling, view blending, and inpainting for the remaining holes. We synthesize virtual views by using the general synthesis mode with quarter-pel precision and with both narrow-baseline and wide-baseline configurations. In the narrow-baseline configuration, near left and near right viewpoints are used for interpolation. In the wide-baseline configuration, far left and far right viewpoints are used. The synthesized virtual view with the best average luminance signal Y-PSNR generated by VSRS 3.5 is preserved for the comparison.

TABLE I
COMPARISON OF AVERAGE LUMINANCE SIGNAL Y-PSNR FOR MULTIVIEW TEST SEQUENCES

Test Sequence	Virtual View	Proposed Algorithm (a)	VSRS 3.5 (b)	△ Y-PSNR [dB] (c)=(a)-(b)
Pantomime	View 40	39.3	38.7	0.6
	View 43	38.5	37.2	1.3
Dog	View 40	35.0	31.4	3.6
	View 43	32.4	31.2	1.2
Lovebird1	View 5	33.2	32.6	0.6
	View 8	32.6	32.0	0.6
Newspaper	View 4	30.0	29.5	0.5
	View 5	32.5	29.3	3.2



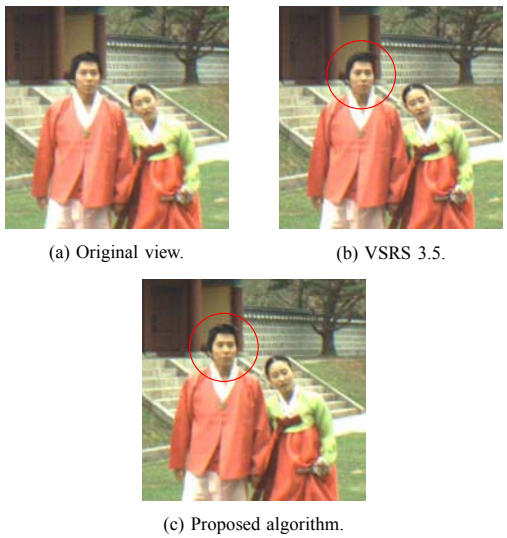
(a) Original view. (b) VSRS 3.5.

(c) Proposed algorithm.

Fig. 4. Synthesized view of the Pantomime sequence.

Table I shows a comparison of the average luminance signal Y-PSNR (in dB) of the synthesized virtual view, averaged over 50 frames, for two different viewpoints using the proposed algorithm and VSRS. In the case of view 40 of the Dog sequence and the view 5 of the Newspaper sequence, the proposed algorithm offers improvements of up to 3 dB. The improvement of the quality depends on the input reference depth maps from various viewpoints.

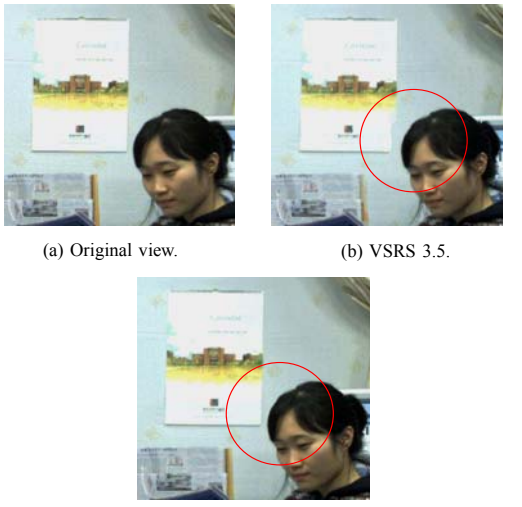
For the Pantomime sequence, the proposed method reduces the noise around the suitcase, as shown in Figure 4. In Figure 5, we compare the visual quality of the Lovebird1 sequence. The quality of the synthesized man's hair in the Lovebird1 sequence is better when compared to the view synthesized by VSRS 3.5. The corner area of the wall calender in the Newspaper sequence is very well synthesized by using the proposed algorithm, as shown in Figure 6.



(a) Original view. (b) VSRS 3.5.

(c) Proposed algorithm.

Fig. 5. Synthesized view of the Lovebird1 sequence.



(a) Original view. (b) VSRS 3.5.

(c) Proposed algorithm.

Fig. 6. Synthesized view of the Newspaper sequence.

V. Conclusion

In this paper, we propose a novel depth consistency testing algorithm. We improve the weak spatial connectivity among depth maps for efficient view interpolation at given virtual viewpoints. We compare average Y-PSNR values of synthesized virtual views calculated with respect to real camera views at the same viewpoints. The proposed scheme improves the quality of the synthesized views up to 3 dB when compared to VSRS for the four reference view configuration. In future work, we will investigate backward warping of consistent depth maps to improve the depth maps of reference views.

Acknowledgment

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

References

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

[2] ——, "FTV (free viewpoint television) creating ray-based image engineering," in *Proceedings of the IEEE International Conference on Image Processing*, vol. 2, Genova, Italy, Sept. 2005, pp. II–25–II–28.

[3] ——, "Overview of FTV (free-viewpoint television)," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, vol. 3, New York, NY, USA, Jul. 2009, pp. 1552–1553.

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

[5] Q. Yang, L. Wang, R. Yang, H. Stewenius, and D. Nister, "Stereo matching with color-weighted correlation, hierarchical belief propagation, and occlusion handling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 3, pp. 492–504, Mar. 2009.

[6] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 9, pp. 1124–1137, Sept. 2004.

[7] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proceedings of the 18th International Conference on Pattern Recognition*, vol. 3, Hong Kong, China, Aug. 2006, pp. 15–18.

[8] C. Fehn, "A 3D-TV approach using depth-image-based rendering (DIBR)," in *Proceedings of the 3rd IASTED Conference on Visualization, Imaging, and Image Processing*, Benalmdena, Spain, Sept. 2003, pp. 482–487.

[9] M. Tanimoto, T. Fujii, K. Suzuki, N. Fukushima, and Y. Mori, "Reference softwares for depth estimation and view synthesis," ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15377, Apr. 2008.

[10] M. Tanimoto, T. Fujii, M. Panahpour, and M. Wildeboer, "Depth estimation reference software DERS 5.0," ISO/IEC JTC1/SC29/WG11, Xian, China, Tech. Rep. M16923, Oct. 2009.

[11] M. Tanimoto, T. Fujii, and K. Suzuki, "Experiment of view synthesis using multi-view depth," ISO/IEC JTC1/SC29/WG11, Shenzhen, China, Tech. Rep. M14889, Oct. 2007.

[12] R. I. Hartley and A. Zisserman, *Multiple View Geomtery in Computer Vision*, 2nd ed. Cambridge, United Kingdom: Cambridge University Press, 2004.

[13] M. Bertalmio, A. Bertozzi, and G. Sapiro, "Navier-stokes, fluid dynamics, and image and video inpainting," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, Kauai, HI, USA, Dec. 2001, pp. I–355–I–362.

[14] S. Chan, H.-Y. Shum, and K.-T. Ng, "Image-based rendering and synthesis," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 22–33, Nov. 2007.

[15] H.-Y. Shum, S. B. Kang, and S.-C. Chan, "Survey of image-based representations and compression techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1020–1037, Nov. 2003.

[16] M. Tanimoto, T. Fujii, and N. Fukushima, "1-d parallel test sequences for MPEG-FTV," ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15378, Apr. 2008.

[17] Y. S. Ho, E. K. Lee, and C. Lee, "Multiview video test sequence and camera parameters," ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15419, Apr. 2008.

[18] G. M. Um, G. Bang, N. Hur, J. Kim, and Y. S. Ho, "3d video test material of outdoor scene," ISO/IEC JTC1/SC29/WG11, Archamps, France, Tech. Rep. M15371, Apr. 2008.

[19] *View Synthesis Software Manual*, ISO/IEC JTC1/SC29/WG11, MPEG, Sept. 2009, release 3.5.