DENOISING OF VOLUMETRIC DEPTH CONFIDENCE FOR VIEW RENDERING

Srinivas Parthasarathy, Akul Chopra, Emilie Baudin, Pravin Kumar Rana, and Markus Flierl

School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm, Sweden

ABSTRACT

In this paper, we define volumetric depth confidence and propose a method to denoise this data by performing adaptive wavelet thresholding using three dimensional (3D) wavelet transforms. The depth information is relevant for emerging interactive multimedia applications such as 3D TV and free-viewpoint television (FTV). These emerging applications require high quality virtual view rendering to enable viewers to move freely in a dynamic real world scene. Depth information of a real world scene from different viewpoints is used to render an arbitrary number of novel views. Usually, depth estimates of 3D object points from different viewpoints are inconsistent. This inconsistency of depth estimates affects the quality of view rendering negatively. Based on the superposition principle, we define a volumetric depth confidence description of the underlying geometry of natural 3D scenes by using these inconsistent depth estimates from different viewpoints. Our method denoises this noisy volumetric description, and with this, we enhance the quality of view rendering by up to 0.45 dB when compared to rendering with conventional MPEG depth maps.

Index Terms — Volumetric depth confidence, denoising, superposition, discrete wavelet transforms, adaptive thresholding, view rendering

1. INTRODUCTION

Advances in visual media technology have led to applications such as three-dimensional television (3D-TV) and free-viewpoint television (FTV) [1]. 3D TV aims to provide a natural 3D-depth impression of dynamic 3D scenes, while FTV enables viewers to dynamically choose their viewpoint of real world scenes. This is realized by using an array of cameras which enable us to acquire multiview imagery by capturing dynamic natural world scene from multiple viewpoints simultaneously. In conventional multiview systems, view rendering is required for smooth transitions among captured views. Usually, view rendering uses multiple views and depth maps acquired from different viewpoints. Each depth map gives information about the shortest distance between the corresponding camera plane and object points in the real world scene. Usually, depth maps for chosen viewpoints are estimated by establishing stereo correspondences only between nearby views [2], [3], [4], [5]. However, the estimated depth maps of different viewpoints usually lack inter-view consistency [6]. This inconsistency of depth estimates affects the quality of view rendering negatively and, hence, FTV users experience visual discomfort.

The availability of highly consistent and accurate depth maps of natural 3D scenes is relevant for enabling these emerging visual media applications. Therefore, in order to have an unique and a consistent description of the underlying geometry of the natural 3D scene, we first generate a noisy volumetric description of the scene geometry by using inconsistent depth estimates from multiple viewpoints. In this description of the scene geometry, depth estimates from multiple viewpoints are fused into a single volume. In [7] and [8], the fusion of depth estimates from multiple viewpoints has been investigated for 3D model reconstruction from video. However, our proposed fusion of depth estimates is based on the superposition principle, where each voxel indicates a 3D position of an object point and holds an additive confidence value. The confidence value of each voxel is updated by depth evidence from multiple viewpoints. Second, we perform a 3D wavelet transform on the volumetric depth confidence. Third, we denoise the volumetric confidence in the wavelet domain by using adaptive thresholding. Finally, we improve the quality of virtual view rendering by utilizing the denoised volumetric depth confidence.

The remainder of this paper is organized as follows: Section 2 describes the concept of volumetric depth confidence and discusses briefly the effect of quantization. Section 3 presents our 3D wavelet thresholding method to denoise the volumetric depth confidence. Section 4 discusses the obtained results.

2. SUPERPOSITION OF CONFIDENCE EVIDENCE

For a given natural scene, inconsistency is inherent among the individually estimated depth maps from multiple viewpoints due to limitations of conventional stereo-matching algorithms. In order to have a consistent description of scene geometry, i.e., depth information, of natural 3D scenes for emerging visual media applications, we uniquely describe the scene geometry by using inconsistent depth estimates from multiple viewpoints. We generate an unique volumetric description of the scene geometry by using depth estimates from multiple viewpoints, where each depth estimate from a viewpoint contributes to a voxel in the volume with a confidence value. Moreover, multiple depth estimates of a 3D object point from multiple viewpoints are fused in the volume by superposition. In the following, we discuss the concept of volumetric depth confidence and consider briefly the effect of quantization.

2.1. Volumetric Depth Confidence

Consider that we capture a natural dynamic scene at multiple viewpoints and that we have conventionally estimated depth maps at viewpoints $1, \ldots, N$. Further, we assume that we know the camera parameters at these N viewpoints. Next, in order to generate

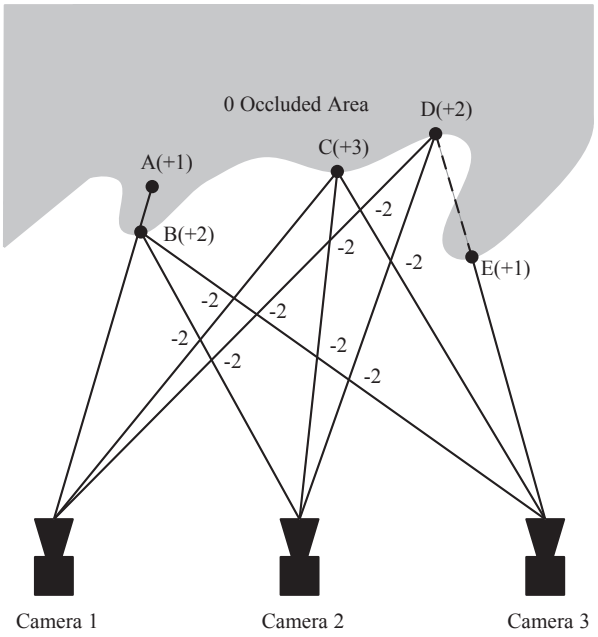


Figure 1. Confidence assignment to the voxel.

the volumetric depth confidence of the scene, we map each estimated depth pixel $d^i(x,y), i \in \{1,\ldots,N\}$, at image coordinates (x,y) to a voxel $v(X,Y,Z)$ in world coordinates by using perspective projection [9].

For each visible object point in the camera plane, we have a corresponding depth pixel in the depth map. This depth value allows us to determine the corresponding voxel in the volumetric description. If a 3D object point is visible from N viewpoints and the corresponding depth information is consistent, we assign the number N to the confidence value of the corresponding voxel, i.e., $v(X,Y,Z) = +N$. Here, each mapped consistent depth pixel from a viewpoint increments the confidence value of the corresponding voxel by one. However, due to inconsistent depth map estimates, the confidence value of a visible object point may vary between $+1$ to $+N$. Specially, object points that are not visible from all viewpoints will have lower confidence values. Object points which are not visible in any of the given camera viewpoints will be assigned a confidence value of zero. With that, depth confidence information from any added camera viewpoint may be superimposed with the volumetric depth confidence.

The transparent voxels between the camera planes and object points are handled similarly. We trace rays from pixels in the camera planes to the respective object points and decrement the confidence value of transparent voxels that lie on the path of each ray. With N available viewpoints, we can decrement the confidence value of transparent voxels to $-N$, at most. Hence, the confidence value of a transparent voxel can vary from $-N$ to -1. Fig. 1 shows the assignment of volumetric confidence values for three camera views.

2.2. Impact of Depth Error

The noise in the volumetric confidence data of the scene geometry is mainly due to the inconsistency in the estimated depth maps. However, the error in estimating depth is bounded by quantization in stereo imaging as explained in [10], [11]. Due to discretization in the imaging system, each image point can suffer a quantization error of up to $\pm 1/2$ pixel. That is, the left and right depth maps can be in error by up to $\pm\delta/2$, where δ is the image sampling interval. The disparity will therefore be in error by up to $\pm\delta$. Further, the error due to quantization depends on the baseline b of the stereo system, the common focal length of the cameras f, and the minimum and maximum depth values in the range of view z_{near} and z_{far}. We assume for the depth z that $0 < z_{near} \leq |z| \leq z_{far} < bf/\delta$. Now, we study the variance of the quantization error as a function of these parameters. We use the assumption in [10] and have for the depth error

$$\Delta z = \frac{-z^2 \Delta d}{bf + z\Delta d} \tag{1}$$

and for the probability density function of the disparity error

$$\begin{array}{rcll} f_{\Delta d}(\Delta d) & = & \frac{\delta+\Delta d}{\delta^2}, & -\delta \leq \Delta d \leq 0 \\ & = & \frac{\delta-\Delta d}{\delta^2}, & 0 \leq \Delta d \leq \delta. \end{array} \tag{2}$$

With that, we have for the conditional expected absolute depth error

$$\begin{array}{l} E[|\Delta z||z] = \frac{1}{2\delta^2}[\frac{2bf}{z}(bf - z\delta)\log(bf - z\delta) - \\ \frac{2bf}{z}(bf + z\delta)\log(bf + z\delta) + 4bf\delta\log(bf) + 4bf\delta], \end{array} \tag{3}$$

and following similar lines, we obtain for the conditional expected squared depth error

$$\begin{array}{l} E[(\Delta z)^2|z] \approx \frac{1}{\delta^2}[\delta^2 z^2 + 3b^2 f^2 \log\left(\frac{b^2 f^2}{b^2 f^2 - \delta^2 z^2}\right) \\ -2bfz\delta\log\left(\frac{bf+\delta z}{bf-\delta z}\right)]. \end{array} \tag{4}$$

A numerical study will follow in the experimental section.

3. 3D WAVELET DENOISING

Wavelet denoising attempts to remove the noise present in the signal while preserving the signal characteristics, regardless of its frequency content. However, denoising is heavily dependent on the thresholding parameter. Using a small threshold may result in ineffective denoising, while a large threshold may yield a smooth output which lacks necessary detail and has blurs and artifacts. Adaptive thresholding improves upon the wavelet thresholding performance by allowing additional local information of the signal to be incorporated into the algorithm [12]. Wavelet denoising is mainly used for three reasons: (1) noise is spread out equally among all the coefficients; (2) it creates a sparse signal, and (3) the signal coefficients are clearly distinguished from the noisy coefficients. Denoising involves taking the discrete wavelet transform (DWT) of the signal, setting an optimum threshold for the coefficients and taking the inverse to get back the denoised signal.

The concept of the wavelet denoising can be explained by assuming that the input data $\underline{g}$ is given by

$$\underline{g} = \underline{s} + \underline{n}, \tag{5}$$

where $\underline{s}$ is the uncorrupted signal and $\underline{n}$ the additive noise. Let $W(\cdot)$ and $W^{-1}(\cdot)$ denote the forward and inverse wavelet transform operators. Let $D(\cdot,\theta)$ denote the denoising operator with threshold θ. We intend to denoise $\underline{g}$ to recover $\underline{\hat{s}}$ as an estimate of $\underline{s}$. The procedure can be summed up in three steps:

1. $\underline{c} = W(\underline{g})$

2. $\hat{\underline{c}} = D(\underline{c}, \theta)$
3. $\hat{\underline{s}} = W^{-1}(\hat{\underline{c}})$

In this paper, we use both manual and adaptive thresholds for denoising the volumetric depth confidence. Manual thresholds are set by trial and error and are the same for all the sub-bands. However, adaptive thresholds are set separately based on the respective sub-band.

3.1. Adaptive Thresholding

To find the optimum adaptive threshold θ for denoising, we use a method known as SURE Shrink [13]. Let $\underline{\mu} = (\mu_1, \mu_2, \ldots, \mu_l)$ be a length l vector and $\underline{c}$ be multivariate normal observations with mean vector $\underline{\mu}$. Using the results from [14] and [15], we have

$$SURE(\theta; \underline{c}) = l - 2|\{i : |c_i| < \theta\}| + \sum_j \min(|c_j|, \theta)^2. \quad (6)$$

For every observed noisy coefficient vector $\underline{c}$ in a sub-band, we obtain the $SURE$ threshold θ^S by minimizing $SURE(\theta; \underline{c})$.

$$\theta^S = argmin_\theta SURE(\theta; \underline{c}) \quad (7)$$

3.2. View Rendering

To evaluate the performance of the proposed methods, we render a virtual view by utilizing the denoised volumetric depth confidence. For rendering, we first obtain three depth maps at left, center, and right viewpoints from the denoised volumetric depth description by using perspective projection [9] [16]. If multiple object points with volumetric depth confidence values correspond to a same depth pixel, we use the one which is closer to the camera in order to obtain the depth map. This is because the depth pixel values describe the shortest distance between the camera plane and the object points. Second, we warp the left view and the right view to the center viewpoint by using the depth map of the center viewpoint [17]. Thus, we obtain two warped version of the central view, one from the right view and the other from the left view, respectively.

We use both of these warped versions of the central view to render the final central view. These warped versions of the central view are used to handle occlusion. For example, if a pixel is occluded in one of the warped views, we use the pixel information from the other warped view. Furthermore, we consider depth maps from various viewpoints, as obtained from the denoised volumetric depth confidence, to tackle occlusion more efficiently. For each object point, we compare the depth pixel value of the left, center, and right depth maps. If the difference between them is below a certain threshold, these depth values are consider to be consistent and averaging of warped pixel intensities is feasible. If only one of the left and the right depth values is close to the central depth value, we keep only the corresponding depth value and pixel intensity. Increasing the number of reference views is likely to decrease occlusion areas. However, occlusions cannot be ruled out completely. If some holes remain, we use a 3×3 median filter to fill isolated intensity pixels.

4. RESULTS

To evaluate the proposed methods, we assess the quality of the rendered virtual views. We measure the objective image quality of the rendered view at a given viewpoint by means of the Peak Signal-to-Noise Ratio (PSNR) with respect to the captured view of a real camera at the same viewpoint. We use two standard Motion Picture Expert Group (MPEG) multiview video test sets, Newspaper, and Dancer [18]. For each of test set, we generate a volumetric depth confidence by utilizing estimated depth information from three different viewpoints, where the voxel confidence can have values between -3 and $+3$.

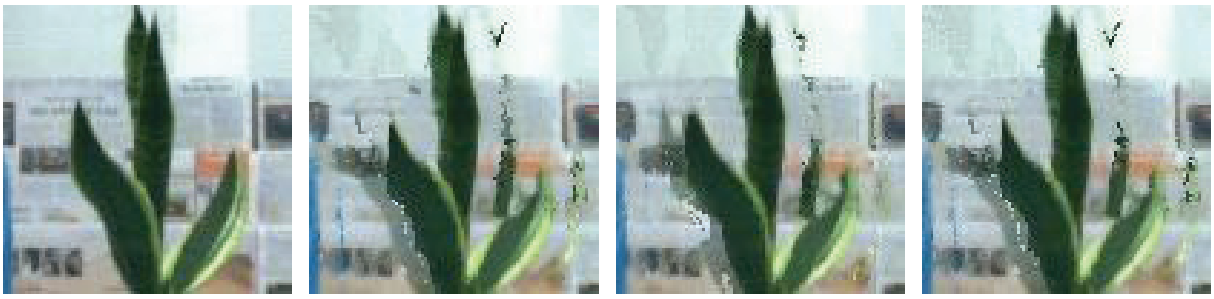
From the study of the depth error for the Newspaper test data with $\delta = 1$, $b = 92.68$, and $f = 2929.49$, as described in Section 2.2, we find the conditional expected values of the absolute error for the depth range from $z_{min} = -2715.18$ to $z_{max} = -9050.60$ to be $E[|\Delta z||z] = 9.04$ and $E[|\Delta z||z] = 100.59$, respectively. The resulting conditional standard deviations for the depth range $z_{min} = -2715.18$ to $z_{max} = -9050.60$ are 6.40 and 71.20, respectively. This analysis confirms that the depth error due to quantization has less effect for points that are close to the camera plane when compared to points that are farther away. Hence, to represent the depth in the volumetric data, we have chosen a uniform quantization between zero and 255 in a ν domain with the relation

$$z = \frac{1}{\frac{\nu}{255}\left[\frac{1}{z_{min}} - \frac{1}{z_{max}}\right] + \frac{1}{z_{max}}}. \quad (8)$$

We denoise the noisy volumetric depth confidence of the scene geometry by using both manual and adaptive thresholds. As described in the Section 3.2, we render the virtual view for the central viewpoint using left and right views with projected central depth maps. The depth maps at the central viewpoint are calculated by projecting voxel points with highest confidence value from the denoised volumetric description onto the central camera plane. The results in terms of PSNR are presented in Table 1. We see that for manual thresholding, the best results are obtained with the db3 filter. Therefore, we choose the db3 filter for adaptive thresholding. Adaptive thresholding improves marginally the objective quality for Dancer. However, for Newspaper, adaptive thresholding improves the objective quality by up to 0.45 dB when compared to rendering with MPEG depth maps only, and 0.24 dB when compared to using the noisy volumetric data. The insignificant improvement for Dancer is due to the fact that its synthetic geometry description is very consistent and that any quantization error can be neglected.

Furthermore, we study the effect of increasing the voxel confidence range on denoising and, hence, on virtual view rendering. We generate a volumetric depth confidence for Newspaper by utilizing estimated depth maps from 7 different viewpoints, where voxel confidence values range between -7 and $+7$. Table 2 shows that an increase in voxel confidence range will improve the objective quality of the rendered views. We notice an improvement even without denoising the volumetric depth confidence. This is because the depth map at the central viewpoint is calculated by projecting voxel points with highest confidence value onto the central camera plane.

Fig. 2 shows rendered central views for subjective evaluation. The visual quality of the rendered virtual view is improved by using the noisy volumetric depth confidence in the range [-7,+7] and by using the denoised volumetric depth confidence. Furthermore, visual quality improves with increasing voxel confidence range.



(a) Original. (b) With MPEG depth maps. (c) With noisy depth volume. (d) With denoised volume.

Figure 2. Rendered views of Newspaper.

Table 1. Objective quality of rendered views with confidence [-3, 3].

Filter	Best Threshold	Rendered Virtual View [dB]	
		Dancer	Newspaper
No denoising	-	38.87	31.84
db1	0.01	38.70	31.33
db2	1.4	38.71	31.42
db3	1.2	38.72	31.43
db3	Adaptive	38.87	32.08
db4	1	38.64	31.38

Table 2. Voxel confidence and objective quality of rendered views.

Used Depth Data	Voxel Confidence Range	
	[-3, +3] [dB]	[-7, +7] [dB]
MPEG Depth Maps	31.63	32.04
Noisy Volume	31.84	32.33
Denoised Vol. (db3, adpt.)	32.08	32.32

5. CONCLUSIONS

This paper discusses volumetric depth confidence, outlines methods to denoise this type of volumetric data and, hence, improves the quality of virtual view rendering. Confidence information from additional camera views can be incorporated by superposition. Further, we use 3D wavelets and adaptive thresholding to denoise our volumetric depth confidence. Finally, experimental results show the advantage of volumetric depth confidence as well as the improvement in rendering quality by increasing the confidence range and by 3D wavelet denoising.

6. REFERENCES

[1] M. Tanimoto, "Overview of free viewpoint television," *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," *Int. J. Computer Vision*, vol. 47, pp. 7 –42, Apr. 2002.

[3] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1124 –1137, Sept. 2004.

[4] A. Klaus, M. Sormann, and K. Karner, "Segment-based stereo matching using belief propagation and a self-adapting dissimilarity measure," in *Proc. Int. Conf. Pattern Recognition*, Hong Kong, China, Aug. 2006, vol. 3, pp. 15 –18.

[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 Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 492 –504, Mar. 2009.

[6] P. K. Rana and M. Flierl, "Depth consistency testing for improved view interpolation," in *Proc. IEEE MMSP*, St. Malo, France, Oct. 2010, pp. 384 –389.

[7] M. Goesele, B. Curless, and S.M. Seitz, "Multi-view stereo revisited," in *Proc. IEEE CVPR*, New York, NY, USA, Jun. 2006, vol. 2, pp. 2402–2409.

[8] P. Merrell, A. Akbarzadeh, L. Wang, P. Mordohai, J.-M. Frahm, R. Yang, D. Nister, and M. Pollefeys, "Real-time visibility-based fusion of depth maps," in *Proc. IEEE ICCV*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1 –8.

[9] R. I. Hartley and A. Zisserman, *Multiple View Geomtery in Computer Vision*, Cambridge University Press, Cambridge, United Kingdom, second edition, 2004.

[10] J. J. Rodriguez and J. K. Aggarwal, "Quantization error in stereo imaging," in *Proc. IEEE CVPR*, Ann Arbor, MI, USA, Jun. 1988, pp. 153 –158.

[11] R. Balasubramanian, S. Das, S. Udayabaskaran, and K. Swaminathan, "Quantization error in stereo imaging systems," *Intern. J. Computer Math.*, vol. 79, no. 6, pp. 671–691, 2002.

[12] S.G. Chang, Bin Yu, and M. Vetterli, "Adaptive wavelet thresholding for image denoising and compression," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1532 –1546, Sep. 2000.

[13] R. Rangarajan, R. Venkataramanan, and S. Shah, "Image denoising using wavelets," Dec. 2002, [Online] Available: http://www.eecs.umich.edu/techreports/systems/cspl/cspl-391.pdf.

[14] C. M. Stein, "Estimation of the mean of a multivariate normal distribution," *The Annals of Statistics*, vol. 9, no. 6, pp. 1135–1151, 1981.

[15] I. M. Johnstone and D. L. Donoho, "Adapting to unknown smoothness via wavelet shrinkage," *J. Statist. Assoc.*, vol. 90, no. 432, pp. 1200–1224, Dec. 1995.

[16] D. Tian, P. Lai, P. Lopez, and C. Gomila, "View synthesis techniques for 3d video," 2009, vol. 7443, p. 74430T, SPIE.

[17] C. Fehn, "Depth-image-based rendering DIBR, compression, and transmission for a new approach on 3D-TV," 2004, vol. 5291, pp. 93–104, SPIE.

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