Precision Enhancement of 3D Surfaces from Multiple Compressed Depth Maps

Pengfei Wan#, Gene Cheung*, Philip A. Chou$, Dinei Florencio$, Cha Zhang$, Oscar C. Au#

Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
{leoman, eeau}@ust.hk

* *National Institute of Informatics*
2-1-2, Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430, Japan
cheung@nii.ac.jp

$ *Microsoft Research*
One Microsoft Way, Redmond, WA 98052-6399, USA
{pachou, dinei, chazhang}@microsoft.com

***Abstract*—In texture-plus-depth representation of a 3D scene, depth maps from different camera viewpoints are typically lossily compressed via the classical transform coding / coefficient quantization paradigm. In this paper we propose to reduce distortion of the decoded depth maps due to quantization. The key observation is that depth maps from different viewpoints constitute multiple descriptions (MD) of the same 3D scene. Considering the MD jointly, we perform a POCS-like iterative procedure to project a reconstructed signal from one depth map to the other and back, so that the converged depth maps have higher precision than the original quantized versions.**

I. Introduction

Texture-plus-depth is now a popular format for dynamic 3D scene representation, where texture and depth maps from multiple viewpoints are captured and compressed. Multiview depth maps are often corrupted by acquisition noise; in [1], we denoised and compressed depth maps simultaneously in a rate-distortion optimal manner. Depth maps are also constrained by sensor's limited acquisition bit-depth; in [2], we enhanced the precision of depth samples by considering multiple depth maps jointly. In contrast, in this paper we focus on reducing the distortion of depth maps caused by lossy transform coding.

The key observation in our work is that depth maps from different viewpoints can be interpreted as *multiple descriptions* (MD) of the same 3D scene. Thanks to this representation redundancy, one can enhance the quality of the reconstructed depth maps by considering multiple descriptions jointly. In particular, we propose an iterative procedure inspired by *projections onto convex sets* (POCS) [3], where a reconstructed 2D signal from the left compressed depth map is first back-projected to the 3D space, then projected to the right depth map. Then the projected signal is appropriately clipped to satisfy the quantization bin constraints of the right depth map. This transformation from left to right depth map constitutes one projection, and the second projection from right to left depth map is performed similarly, and the two projections are repeated until convergence. We show experimentally that the converged signal has higher precision than the original quantized version of individual depth map.

MMSP'13, Sept. 30 - Oct. 2, 2013, Pula (Sardinia), Italy.

II. System Overview

For simplicity, we consider a scenario where a static 3D scene is captured by depth maps of only two views (left and right). Each depth map is compressed using JPEG; 8×8 DCT transform coefficients are quantized and transmitted. At the decoder, given quantized transform coefficients, our goal is to reconstruct depth maps with enhanced precision.

Let $\mathbf{x}$ be a vectorized representation of an original 8×8 pixel block in the left depth map. The corresponding transform coefficient vector is $\mathbf{y} = \mathbf{T}\mathbf{x}$, where $\mathbf{T}$ is the linear DCT transform operator. Scalar quantization of each coefficient k, y_k, means mapping of y_k to quantization bin $\mathcal{B}_k$ of centroid $q(y_k)$ and width Δ, where

$$y_k \in [q(y_k) - \Delta/2, q(y_k) + \Delta/2] \stackrel{\text{def}}{=} \mathcal{B}_k \quad (1)$$

Only quantization bin indices are encoded, so at decoder, only quantization bins $\mathcal{B}_k$'s are known. We can hence write the quantization bin constraints for the reconstructed coefficients $\mathbf{y}^l$ for a 8×8 pixel block in the left depth map as:

$$\mathbf{y}^l \in \mathcal{S}^l \stackrel{\text{def}}{=} \{\mathbf{y} \mid y_k \in \mathcal{B}_k\} \quad (2)$$

Geometrically, $\mathcal{S}^l$ is a hyper-cube representing the range of transform coefficients for a quantized block in the left depth map. A sequence of $\mathcal{S}^l(\mathbf{b})$'s for different blocks $\mathbf{b}$ in the left map thus defines the feasible search space of the original left depth signal. Similarly, a sequence $\mathcal{S}^r(\mathbf{b})$'s for blocks in the right map can also be deduced. Collectively, $\mathcal{S}^l(\mathbf{b})$'s and $\mathcal{S}^r(\mathbf{b})$'s are interpreted as two redundant descriptions of the same 3D scene. We next describe a POCS-like iterative procedure to enhance the precision of reconstructed depth maps by considering both descriptions jointly.

III. POCS-Inspired Iteration

We perform the projection from left view to right view starting with an initial reconstructed left depth signal. The projection (composed of the following two steps) is alternated between the two views until convergence.

A. View-to-view Projection

In this step, we first back-project pixels of a reconstructed left depth map to the 3D space. In particular, for a camera with intrinsic matrix $\mathbf{K}_{3\times3}$ and extrinsic matrix $\mathbf{E}_{3\times4}$, a pixel with depth d at image coordinate (r, c) corresponds to a 3D voxel at world coordinate (x, y, d) that satisfies: $(c, r, 1)^\intercal = \alpha \cdot \mathbf{K} \cdot \mathbf{E} \cdot (x, y, d, 1)^\intercal$ for some scalar α. The obtained 3D voxel is then re-projected to the right view based on the right camera parameters. Note that in general, the projected points to the right view do not land on 2D grid points of the right depth map. Thus, to interpolate an updated right depth at (r, c), we perform a simple edge-adaptive linear interpolation using two pixels p_1 and p_2 projected from the left view (assuming two views are rectified). p_1 and p_2 are respectively the projected pixels whose depth values are: 1) close to initial right depth value at (r, c); 2) minimal among projected depth values in range $(r, (c-1, c))$ and $(r, (c, c+1))$ respectively. After interpolation, we further apply a bilateral filter [4] to remove noisy pixels.

B. Coefficient Clipping

This step is a standard projection on convex sets: we transform each depth pixel block in the right view to DCT domain, and clip a transform coefficient to its nearest boundary value if it is out of range. It is to ensure that the quantization constraints in (2) are satisfied.

After convergence, the original transform coefficient vectors $\{\mathbf{y}(\mathbf{b})\}$ becomes modified ones $\{\mathbf{y}^*(\mathbf{b})\}$. The output depth maps are then obtained using inverse transform.

IV. Preliminary Results

The test sequences are obtained from the New Tsukuba Stereo Dataset[1]. Standard decoded depth maps are denoted as $\mathbf{I}^{std}$, while the outputs of proposed method are $\mathbf{I}^{our}$. Corresponding quality metrics are calculated by:

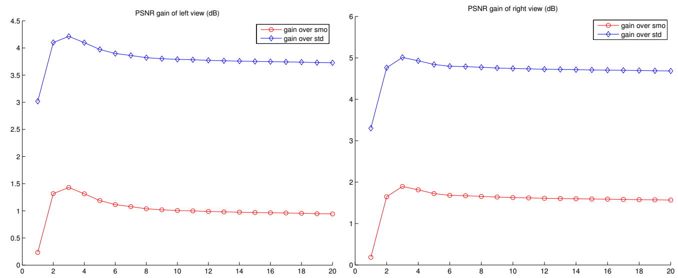
$$Q_{std} = g(\mathbf{I}_l^{std}, \mathbf{I}_r^{std}), \quad Q_{our} = g(\mathbf{I}_l^{our}, \mathbf{I}_r^{our}) \qquad (3)$$

where function $g(\mathbf{I}_1, \mathbf{I}_2) = (\text{PSNR}(\mathbf{I}_1, \mathbf{I}_l) + \text{PSNR}(\mathbf{I}_2, \mathbf{I}_r))/2$ calculates the average PSNR with regard to uncompressed 8-bit depth maps $\mathbf{I}_l$ and $\mathbf{I}_r$.

Note that bilateral filter is used in our proposed method. As a comparison, we also apply bilateral filter directly on $\mathbf{I}^{std}$, resulting in smoothed depth map $\mathbf{I}^{smo}$, whose average PSNR, Q_{smo}, is calculated similarly.

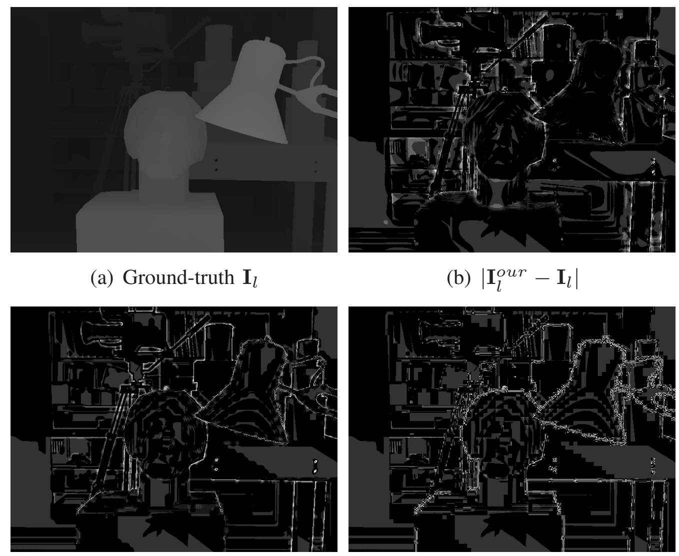
Fig. 1 shows the PSNR gain of proposed method over comparing ones: $Q_{our} - Q_{smo}$ and $Q_{our} - Q_{std}$. Error maps are also shown in Fig. 2. Overall, proposed method achieves encouraging performance: $Q_{our} > Q_{smo} > Q_{std}$.

[1] http://www.cvlab.cs.tsukuba.ac.jp/dataset/tsukubastereo.php



(a) left view (b) right view

Fig. 1. PSNR gain in dB. X-axis is the iteration index.



(a) Ground-truth $\mathbf{I}_l$ (b) $|\mathbf{I}_l^{our} - \mathbf{I}_l|$

(c) $|\mathbf{I}_l^{smo} - \mathbf{I}_l|$ (d) $|\mathbf{I}_l^{std} - \mathbf{I}_l|$

Fig. 2. Sample error maps of the left view.

Note that convergence of our POCS-like method is empirical, and that the convergence point is typically not the global optimal one, see Fig. 1. Proof of convergence and study of the optimal number of iterations are left for future work.

V. Conclusion

In this paper, we introduce a POCS-like method to enhance the bit-precision for decoded multi-view depth maps. The experimental results show that proposed method significantly outperforms comparing methods.

References

[1] W. Sun, G. Cheung, P. A. Chou, D. Florencio, C. Zhang, and O. C. Au, "Rate-distortion optimized 3D reconstruction from noise-corrupted multiview depth videos," in *IEEE International Conference on Multimedia & Expo*, San Jose, USA, 2013.

[2] P. Wan, G. Cheung, P. A. Chou, D. Forencio, C. Zhang, and O. C. Au, "Precision enhancement of 3D surfaces from multiple quantized depth maps," in *11th IEEE IVMSP Workshop: 3D Image/Video Technologies and Applications*, Seoul, Korea, 2013.

[3] P. Chou, S. Mehrotra, and A. Wang, "Multiple description decoding of overcomplete expansions using projections onto convex sets," in *In Proceedings of Data Compression Conference*, 1999, pp. 72–81.

[4] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Proceedings of the IEEE International Conference on Computer Vision*, Bombay, India, 1998.