

# COMPRESSION AND TRANSMISSION OF DEPTH MAPS FOR IMAGE-BASED RENDERING

*Ravi Krishnamurthy, Bing-Bing Chai, Hai Tao, and Sriram Sethuraman*

Sarnoff Corporation, 201 Washington Rd., Princeton, NJ 08540, USA  
{rkrishnamurthy,bchai,htao,ssethuraman}@sarnoff.com

## ABSTRACT

We consider applications using depth-based Image-based rendering (IBR), where the synthesis of arbitrary views occur at a remote location, necessitating the compression and transmission of depth-maps. Traditional image compression has been designed to provide maximum perceived visual quality, and a direct application is sub-optimal for depth-map compression, since depth-maps are not directly viewed. In other words, the sensitivity of the rendering error depends on image content as well as on the depth map; we propose two improvements to take this into account. Firstly, we consider region-of-interest (ROI) coding, where we identify those regions of the image where accurate depth is most crucial. Secondly, we reshape the dynamic range of the depth map. Our experiments show a significant improvement in coding gain (1.1 dB) and rendering quality when we integrated these two improvements into a standard JPEG-2000 coder.

## 1. INTRODUCTION

Over the past several years, image-based rendering (IBR) techniques have received great attention in the research communities of image processing, computer vision, and computer graphics. Unlike the traditional rendering methods where the geometry and the texture of the scenes are modeled, IBR methods synthesize arbitrary views of a scene from a collection of images observed from known viewpoints. The main advantage of the IBR approach is that it can render complicated scenes that are otherwise difficult to model geometrically. Depending on the amount of 3D information being employed, a continuum of IBR methods have been developed, ranging from pure image based methods such as light field rendering [1] and lumigraph [2] to depth based image warping method [3].

In this paper, we consider a particular application where the 3-D dynamic scene is represented by

- a) Multiple views from multiple cameras distributed in such a way as to “cover” the scene. This information is captured as multiple 2-D video streams.
- b) Dense accurate depth maps estimated for each camera view and each time instant, which are then used for image-based rendering (IBR).

We are looking at applications running over the Internet, where the compression and transmission of this 3-D information is very important. In this paper, we address the problem of coding the depth images and their transmission over the Internet.

## 2. DEPTH MAP ESTIMATION AND IBR

Depth information of a scene can be recovered from multiple close-by images using generalized stereo algorithms. These algorithms compute depth using the triangulation method based on image correspondences across multiple views. With the collection of images and their associated depth maps, new views are generated from individual given views through the depth-based warping.

If all cameras are calibrated, for a pixel in the reference with image coordinates  $p=[x, y, 1]$ , the depth-based image warping function is

$$p' = p^w + k(T - T_z p^w) \quad (1)$$

where  $p'=[x', y', 1]$  is the homogeneous coordinate of the

pixel in the new view,  $p^w = \frac{Rp}{[Rp]_3}$  is the result of an

intermediate warping that is affected by the camera rotation  $R$  but not by the camera translation and the depth,  $k = 1/Z$  is the depth information,

$T = [T_x, T_y, T_z]^t$  is the camera translation [4].

For rendering purpose, it is crucial to obtain accurate depth information. We have developed a color segmentation based stereo algorithm for recovering accurate depth maps with sharp depth discontinuity boundaries and fine details of thin structures. More details of the algorithm can be found in [5].

## 3. CODING OF DEPTH MAPS

In this paper, we will consider the coding of single depth images, an example of which is shown in Figure 1 (c). Extensions to temporal sequences will be a subject for further study.

A large variety of techniques exist for coding of still images, from the ubiquitous DCT-based JPEG standard to the emerging, wavelet-based JPEG-2000 standard [6]. However, most of these standards have been developed

for “real-world” imagery and may not be optimal for encoding depth information.

The main difference is that these existing coding methods use an error criterion similar to mean-squared error (MSE) in order to compress the images, which may be close to optimal from a visual standpoint. However, MSE in depth map may not be very meaningful, because the depth maps are never directly viewed—rather they are used for rendering new images whose quality is of importance in the application. Thus, the sensitivity of the final rendered image due to errors in the depth map cannot be described by the MSE-like criterion. For example, errors in the depth map close to an intensity edge can result in ugly rendering artifacts, while errors on smooth surface will be unnoticeable. Thus, a degree of perceptual weighting needs to be applied and an appropriate cost function needs to be used for depth map compression. This is similar to the problem of encoding dense optical flow fields, where the use of an appropriate cost function for encoding is crucial in order to get a compact representation [7].

A most general cost function for depth-map coding can take on the form

$$E(\mathbf{dk}) = \|I(p'(k+\mathbf{dk})) - I(p'(k))\|^2 \quad (2)$$

Where  $I$  is the intensity of the rendered pixel,  $k$  is the original high-quality depth map and  $\mathbf{dk}$  is the coding error in the depth map. This is similar to the cost function that is used for depth-map estimation.

In this paper, we do not consider this complicated cost function, which will be the focus of further study. Instead, we consider the disparity function, which is the shift of a pixel  $p$  from the reference view to the corresponding pixel  $p'$  in the rendered view,

$$e = \|p' - p\|^2 = (p' - p)^t (p' - p) \quad (3)$$

The sensitivity of this function to depth can be described by  $\partial e / \partial k$  and can be used for compression. However, this depends not only on  $k$ , but also on the pixel position in the image and on the final rendered view, which is unknown at the time of compression.

One possibility is to optimize the compression for a range of views, which will be rendered using this particular depth map. The problem of an optimal tessellation of 3-D space with cameras, and thus determining the range of views that will be rendered from this depth map, is an interesting and extremely challenging problem in itself.

We begin our experiments on the *Quantico* sequence, an outdoor sequence of 14 images captured from a helicopter. This sequence contains a number of sharp boundaries with high contrast. The estimated depth map ( $k$ ) for one of the views is shown in Figure 1(c). Depth maps are compressed using VM8 of the JPEG-2000

standard [8]. For this purpose, the floating point number is scaled to a 16-bit number and then input to the JPEG-2000 algorithm. And the procedure can be reversed at the decoder. We observe that at least  $0.3 \text{ bpp}$  is required in order to create a rendered image that is indistinguishable from the original rendered image. In Figure 1(e), we show the rendering result (a magnified region in the high-contrast foreground is shown) when using  $0.2 \text{ bpp}$ , and we can see significant errors when compared to the rendering using the uncompressed depth (Figure 1 (d)). From these experiments, a few observations can be made:

- a) Sharp discontinuities in depth and intensity require more accurate depth-maps.
- b) The sensitivity of the disparity with respect to the depth varies with depth for any given view.

Based on this, we come up with two coding improvements that give us significant gains in depth map coding.

#### 4. CODING IMPROVEMENTS

The two major improvements that we propose are

- a) Region of Interest (ROI)-based coding
- b) Reshaping dynamic range of depth in order to reflect the different importance of different depths

##### 4.1 ROI coding

In previous discussions, we have seen how different parts of the depth-map are more important for IBR. This ROI can be generated from the depth estimation algorithm and could reflect the confidence and accuracy of the depth-map in that particular region. However, for our initial experiments, we used an ROI, which was generated by running an edge detector on the depth map (Figure 1(e)). For image discontinuities, which do not occur at depth discontinuities, we expect the smooth property of the wavelet transform to provide us with a fairly uniform depth, which will result in reasonable rendering.

JPEG-2000 provides us with the means of selectively encoding ROIs using the *maxshift* method [9]. In this technique, the coefficients belonging to the ROI are shifted such that their bit-planes do not overlap with the non-ROI coefficients. This ensures that the ROI is first completely encoded, and then if bits are left over, the non-ROI regions are encoded. At the decoder the ROI coefficients can be implicitly identified from their shifted values, and thus no shape information needs to be explicitly transmitted to the decoder [9].

##### 4.2 Reshaping the dynamic range of depth maps

In our experiments, we observed that the errors were much more significant in the areas of smaller depth (larger  $k$ ). For a range of views where rotation is small, this can be theoretically derived as shown in Section 5.

This observation gave us the idea of reshaping the dynamic range of depth, an idea similar to *companding* [10], which is used in encoding speech signals.

First the depth map was scaled to a value [0,1]. Then we passed the depth-map through a function that expands the dynamic range for higher  $k$ 's and compresses the dynamic range for lower  $k$ 's (Note that this is opposite to the action of a traditional compander used in speech). A simple such function is the quadratic function

$$k' = \alpha k^2 + (1-\alpha)k \quad (3)$$

The theoretical justification for using such a scaling function is given in Section 5. The choice of  $\alpha$  depends on the range of views that are being rendered.

We then encode  $k'$  using the JPEG-2000 algorithm after scaling this floating point number to a 16-bit number.

#### 4.3 Experimental Results

The results obtained by using ROI and reshaping of the dynamic range are shown in Figure 1(f). For these results, we used a value of  $\alpha=1$ , which corresponds to  $k'=k^2$ . This assumption is valid when the rotation is small and when the translation in the  $z$ -direction is small.

The rendering results are significantly improved at 0.2 *bpp* as can be seen from Figure 1(f), and are very similar to the rendering using original depth map.

To measure the relative performance, we measured the PSNR of the rendered views using compressed depth maps with reference to the view rendered by the uncompressed depth map. Direct application of JPEG-2000 algorithm gave us a PSNR of 33.17 *dB*, and the coder incorporating our proposed improvements gave a PSNR of 34.27*dB*, an improvement of 1.1 *dB*.

#### 5. THEORETICAL ARGUMENTS

In this section, we provide theoretical arguments to support our choice of Equation (3).

Errors in the depth map are transformed into displacement errors in the warping process. If for a given new view, the displacement vector of a pixel  $p$  is measured as

$$e = \|p' - p\|^2 = (p' - p)^t (p' - p) \quad (4)$$

Then the sensitivity of this measurement with respect to depth is described by the derivative

$$\begin{aligned} \frac{\partial e}{\partial k} &= 2(p' - p)^t \frac{\partial p'}{\partial k} \\ &= 2(p^w - p + k(T - T_Z p^w))^t (T - T_Z p^w) \\ &= 2(p^w - p)^t (T - T_Z p^w) + k(T - T_Z p^w)^t (T - T_Z p^w) \\ &= a + bk \end{aligned}$$

An optimal dynamic range transformation should take the form  $k' = k \frac{\partial e}{\partial k}$ . Thus, the transformation becomes  $k' = ak + bk^2$ , which can be normalized and written as  $\alpha k + (1-\alpha)k^2$ .

It can also be seen that when there is no rotation,  $p^w = p$  [4], which implies that  $k' = k^2$ .

#### 6. SUMMARY

Our experimental results have demonstrated considerable improvement in depth-based IBR, when depths are encoded using the proposed improvements to a standard JPEG-2000 coder.

In future study, we intend to study the optimal reshaping for a given range of arbitrary views that will be rendered using this depth-map. Further, the ROI can be refined using inputs from the depth-estimator like confidence and rendering error. And finally, we will study motion-compensated encoding of depth sequences.

#### 7. ACKNOWLEDGMENTS

We would like to thank Harpreet Sawhney and Rakesh Kumar for ideas and discussions that greatly contributed to this paper, and Supun Samrasekhara and Rakesh Kumar for providing us with the *Quantico* sequence.

#### 8. REFERENCES

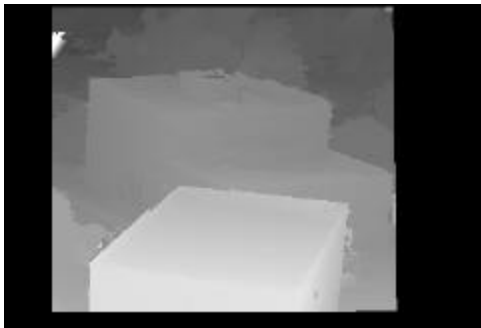
- [1] M. Levoy and P. Hanrahan. Light field rendering. In *Computer Graphics Proceedings, Annual Conference Series*, pages 31–42, Proc. SIGGRAPH'96 (New Orleans), August 1996, ACM SIGGRAPH.
- [2] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The Lumigraph. In *Computer Graphics Proceedings, Annual Conference Series*, pages 43–54, Proc. SIGGRAPH'96 (New Orleans), August 1996. ACM SIGGRAPH.
- [3] J. Shade, S. Gortler, L.-W. He, and R. Szeliski, Layered depth images. In *Computer Graphics (SIGGRAPH'98) Proceedings*, pages 231–242, Orlando, July 1998. ACM SIGGRAPH.
- [4] Irani and Anandan, Parallax Geometry of Pairs of Points for 3-D Analysis, *Proceedings of ECCV*, 1996
- [5] Hai Tao and Harpreet S. Sawhney, Global matching criterion and color segmentation based stereo, in *Proc. Workshop on the Application of Computer Vision (WACV2000)*, pp. 246-253, December 2000.
- [6] M. J. Gormish, D. Lee, M. W. Marcellin, JPEG-2000: Overview, Architecture and Applications, *Proceedings of ICIP-2000*, September 2000.
- [7] P. Moulin, R. Krishnamurthy and J.W. Woods, "Multiscale modeling and estimation of motion fields for video coding", *IEEE Trans. Image Processing*, vol 6 no. 12 pp 1606-1620, 1997.
- [8] JPEG-2000 Verification Model Software 8.5, ISO/IEC JTC1/SC29/WG1 N1894, 2000.
- [9] C. Christopoulos, J. Askelof, M. Iarsson, "Efficient regions of Interest coding Techniques in upcoming JPEG2000 still image coding standard, *Proceedings of ICIP*, September 2000.
- [10] J. Proakis, Digital Communications, McGraw Hill, 1989, pp 91-92.



(a)



(b)



(c)



(d)



(e)



(f)



(g)

**Figure 1:** (a), (b) Two neighboring views (c) Depth map for View 1 (d) Rendering of intermediate view using Original Depth Map – (high-contrast foreground shown) (e) Rendering using JPEG-2000 compressed depth ( $0.2 \text{ bpp}$ ,  $PSNR=33.17 \text{ dB}$ ) (f) Rendering from depth map compressed using ROI and reshaping ( $0.2 \text{ bpp}$ ,  $PSNR=34.27 \text{ dB}$ ) (g) Edge mask for ROI