DEPTH MAP DISTORTION ANALYSIS FOR VIEW RENDERING AND DEPTH CODING

Woo-Shik Kim[*]*, Antonio Ortega*

Signal and Image Processing Institute
Ming Hsieh Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA 90089

PoLin Lai, Dong Tian, Cristina Gomila

Thomson Corporate Research
2 Independence Way
Princeton, NJ 08540

ABSTRACT

Video representations that support view synthesis based on depth maps, such as multiview plus depth (MVD), have been recently proposed, raising interest in efficient tools for depth map coding. In this paper, we derive a new distortion metric that takes into consideration camera parameters and global video characteristics in order to quantify the effect of lossy coding of depth maps on synthesized view quality. In addition, a new skip mode selection method is proposed based on local video characteristics. Experimental results with the proposed mode selection scheme show coding gains of up to 2 dB for the synthesized views, as well as better subjective quality.

***Index Terms*—** Multiview plus depth (MVD), depth coding, view synthesis, rate-distortion optimization

1. INTRODUCTION

Efficient multiview video systems can be a significant step towards a more realistic multimedia experience, e.g., with applications such as three dimensional (3-D) video and free viewpoint video [1, 2]. Given the data volumes associated with multiview video systems, designing efficient compression techniques remains an important challenge to make these application reality. A promising compression approach is based on view synthesis [3, 4] using depth maps. This has led to recent research into multiview plus depth (MVD), efficient approaches to encode and transmit a depth map along with each view, so that at the decoder new intermediate views can be synthesized using the neighboring views and their depth maps.

A depth map can be thought of as a gray scale image, and the corresponding temporal sequence of depth maps can be treated as a standard video sequence. Thus, as a first approach to encoding a depth map sequence one could make use of standard video coding techniques. However, we note that depth map sequences have characteristics that are very different from those of standard video. For example, depth images rarely contain any texture and are predominantly flat with sharp edges marking the boundary between objects at different depths.

To exploit depth map specific characteristics for compression, various methods have been proposed. These include flat region coding with edge preservation [5], dynamic range reshaping [6], 3-D motion estimation [7], warping based inter-view prediction [8], reuse of video motion information to reduce encoding complexity [9], and sparsity-based in-loop filtering [10].

The key observation in this paper is that depth data is encoded but not displayed; it is only used to synthesize intermediate views from existing ones. Thus, the distortion that affects those synthesized views due to lossy encoding of depth is fundamentally different from the distortion affecting luminance or chrominance data in standard video. More specifically, errors in depth values at a given pixel position, affect the *position* in the intermediate view where this pixel will be used for interpolation. Thus, even small errors in depth can lead to significant errors in interpolated pixel intensity.

Note that modern video encoders, e.g., those based on H.264/AVC [11] make extensive use of rate-distortion characteristics for mode decision, rate control, etc., thus using such a coder to encode a sequence of depth maps may lead to suboptimal results if the distortion metric is simply the mean squared error (MSE) in the reconstructed depth map. Instead, we propose to *develop new distortion metrics that aim to capture the effect of depth map distortion on the final quality of the synthesized views*. Based on this distortion metric, we propose a mode selection scheme that optimizes the bitrate of depth map and the quality of the synthesized views. We achieve about 1 dB gain on average and clear subjective quality improvements, as flickering artifacts are significantly reduced in the synthesized views.

This paper is organized as follows. The problem of depth map coding is addressed in Section 2, where the effect of depth map distortion on a synthesized views is examined. The proposed solution using the new distortion metric and mode selection scheme is presented in Section 3. Experiments are performed using MVD sequences, and the results are discussed in Section 4. The conclusion is given in Section 5.

2. CHALLENGES IN DEPTH MAP CODING

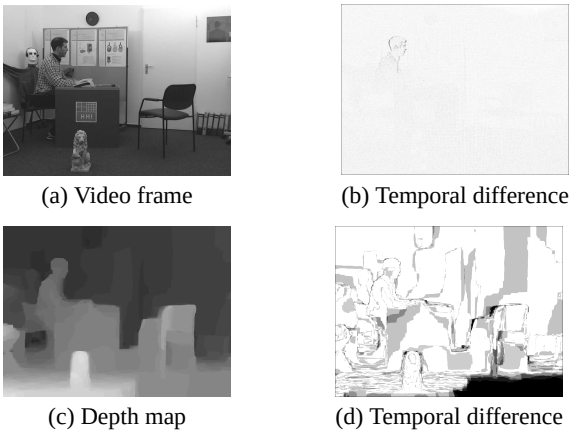
In this paper we address two main differences between depth and video data. First, while the distortion in a video directly changes the reconstructed level of luminance or chrominance, depth map distortion affects the synthesized views by causing an error in the *position* where interpolated pixels are located in the synthesized views. Moreover, the magnitude of this position error depends primarily on parameters associated to the depth map acquisition procedure. For example, in case of stereo matching this error depends on camera settings such as positions of cameras and objects, camera focal length, etc. This is important, because an error in depth reconstruction of same magnitude may have very different impact depending on the actual composition of the scene and the camera parameters. As a simple example, if we use a fixed number of bits (e.g., 8 bits) to represent depth (or in practice, disparity), the same error in reconstructed depth will have a much greater impact if the scene covers a wide range of depths (in contrast with, say, an indoor scene). In summary, for depth encoding to be optimized for view synthesis it will be necessary to introduce a new distortion metric that can take

[*]This work is supported by Thomson Corporate Research. (Further author information: Send correspondence to wooshik.kim@usc.edu)

these factors into account.

Second, it is worth noting that in case of a video, if sensor noise is negligible, the distortion at the decoder is mainly due to quantization. In contrast, in case of a depth map estimated from video data (i.e., not captured directly with special devices such as range cameras), the estimated depth itself can be very noisy. For example, using stereo matching to obtain depth will lead to more significant errors in the boundaries of near objects, as compared to the background area. This is due to large differences in projection angles between left and right cameras for near objects, which leads to large occlusion. Moreover, for areas in the scene that are predominantly flat and contain limited amounts of texture, it will be difficult to find matching points between left and right views, which will make the depth information less reliable. In addition, if the depth maps are estimated on a frame by frame basis, i.e., depth / video information from other timestamps are not considered, unreliable estimates of depth are more likely to lead to stronger temporal variations, i.e., depth estimates may vary even when the "ground truth" does not.

Fig. 1 helps illustrate these issues. From Fig. 1 (b) and (d) (where the absolute value of the temporal differences is scaled by 5 and inverted for easier visualization), it can be easily noticed that temporal variation in the depth map is very significant, even though there is practically no motion in this video. Most of these changes in the depth map can be attributed to errors in the stereo matching process. Note in particular that more errors can be observed around object boundaries and in the flat regions with less texture, where the stereo matching suffers due to occlusion and lack of matching features, respectively. This temporal variation in the depth map not only increases the coding bitrate but also deteriorates the subjective quality of the synthesized views by creating flickering artifacts in the flat region. However, as will be seen next, because these temporal variations in depth estimates do not correspond to changes in actual depth, efficient coding of depth can be achieved (e.g., by not coding many of these estimated depth changes), without significant impact on interpolated view quality. Even though it would be possible to improve depth map quality using more advanced systems such as range cameras, it will be still useful for algorithms to be robust to errors in depth map acquisition, which could be inherent to many acquisition systems.



(a) Video frame (b) Temporal difference

(c) Depth map (d) Temporal difference

Fig. 1. Example of temporal variation in depth map. (a) frame in the 'Door Flowers' video sequence, (b) difference between the first and second frame of the video, (c) corresponding depth map, and (d) difference between the depth maps of the first and second frames.

3. PROPOSED TECHNIQUES FOR DEPTH MAP CODING

To address the challenges introduced in the previous section we now propose a new distortion metric and modified mode selection technique to improve depth map coding.

3.1. New distortion metric using global parameters

As discussed earlier, distortion in a depth map results in a position error in the synthesized views. This position error can be quantified if camera parameters are known, as described in [10]. Under the assumption of a parallel camera setting, the position error ΔP can be written as a horizontal translation:

$$\Delta P = a \cdot \delta_x \cdot \frac{\Delta D_{depth}(x,y)}{255} \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right), \qquad (1)$$

where a is the focal length of the camera in the horizontal direction with the unit of pixels, δ_x is the distance between two cameras (horizontal), and Z_{near} and Z_{far} are the nearest and the farthest depth values, which correspond to the values of 255 and 0 in the depth map, respectively. This reveals that there is a linear relationship between the depth map distortion $\Delta D_{depth}(x,y)$ at pixel position (x,y) and the translation error in the synthesized view, i.e.,

$$\Delta P = k_1 \cdot \Delta D_{depth}, \qquad (2)$$

where k_1 can be calculated as

$$k_1 = a \cdot \delta_x \cdot \frac{1}{255} \cdot \left(\frac{1}{Z_{near}} - \frac{1}{Z_{far}} \right). \qquad (3)$$

Now, given the position error, we would like to estimate the resulting distortion in the synthesized view. Clearly this distortion will be content dependent. For example, position errors will have minimal effect in regions of constant intensity. Conversely, in regions with significant edge/texture information small changes in position can lead to significant changes in intensity.

We propose to capture the content-dependent nature of synthesized view distortion by estimating a simple *global* frame parameter. Define the function $d_{SSD}(t_x)$ as the sum of squared differences (SSD) between the original video frame and its horizontal translation by t_x pixels:

$$d_{SSD}(t_x) = \sum_x \sum_y \left(V_{(x,y)} - V_{(x-t_x,y)} \right)^2, \qquad (4)$$

where $V_{(x,y)}$ is the pixel value of the original video at pixel position (x,y), and $(x - t_x, y)$ is the horizontally translated pixel position. Experimentally we observe that $d_{SSD}(t_x)$ varies linearly as a function of t_x as shown in Fig. 2 (in particular for small displacements). In Fig. 2 the first frame of each sequence is used to compute $d_{SSD}(t_x)$ for t_x varying from 1 to 30 in unit increments. Hence the scale factor between d_{SSD} and translation t_x can be found using the least square fit as

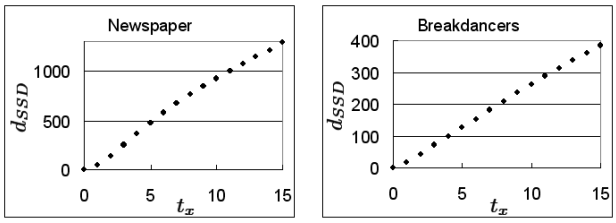


Fig. 2. Relationship between translation and distortion.

$$s = \frac{\mathbf{d}_{SSD}^{\mathrm{T}} \mathbf{t}_x}{\mathbf{t}_x^{\mathrm{T}} \mathbf{t}_x}, \quad (5)$$

where $\mathbf{d}_{SSD}$ and $\mathbf{t}_x$ are the vectors formed by aggregating multiple values of $d_{SSD}(t_x)$ and t_x, respectively, and T denotes vector transpose operand. For a given position error ΔP, this parameter s provides an estimation of the resulting distortion in the interpolated view. Note that better accuracy can be achieved if smaller area is used to reflect local video characteristics. For example, the parameter can be calculated for each blcok in order to obtain a more precise result, but computationally expensive. Therefore, we use a global parameter as a compromise between accuracy and complexity. It would be appropriate to update the parameter whenever there is a scene change.

Since the synthesis process typically uses multiple views, this factor can be scaled using the same weight the synthesis process put on the view. For example, if the synthesis process applies a weight, α, as

$$V_{synth} = \alpha \cdot V_{left} + (1 - \alpha) \cdot V_{right}, \quad (6)$$

where V_{left}, V_{right}, and V_{synth} are the pixel values in the left, right, and synthesized view, respectively, then the scale factor, k_2, to represent the global characteristic for V_{left} can be calculated as

$$k_2 = \alpha \cdot s. \quad (7)$$

Using the two parameters found above, the new distortion metric can be derived as

$$\Delta D_{synth}^2 = k_2 \cdot \Delta P = k_2 \cdot k_1 \cdot \Delta D_{depth}, \quad (8)$$

where ΔD_{synth}^2 denotes the quadratic error in the synthesized frame. This new distortion metric can be used in the rate distortion optimized mode selection process using the Lagrangian optimization [12], with Lagrangian cost J written as:

$$\begin{aligned} J &= \sum_x \sum_y \Delta D_{synth}^2(x,y) + \lambda R_{depth} \\ &= k_1 k_2 \sum_x \sum_y |\Delta D_{depth}(x,y)| + \lambda R_{depth}, \end{aligned} \quad (9)$$

where (x, y) is a pixel position in the block, λ is the Lagrange multiplier, and R is the bitrate consumed to code the depth map block. Note that the quadratic error in the synthesized view is proportional to the absolute error in the depth map.

3.2. Skip mode decision using local image characteristics

We also propose to improve the mode decision process by considering local video characteristics. As described in Section 2, distortion can occur during depth map estimation. In particular, if there is lack of features to perform stereo matching, the resulting depth map can be noisy, so that it would not be efficient to spend more bits to achieve an accurate representation of the depth map.

To solve this problem, before encoding a block of depth data we take into account how the corresponding block of video data was encoded. We note that limited motion regions are also regions where depth information is unlikely to vary over time (in particular if cameras remain fixed). Since limited motion blocks are likely to be encoded using skip mode, especially at low rates, we propose to "force" skip mode in depth coding in those blocks for which skip mode was chosen for the video data. Note that in those blocks, skip mode may not have been selected by the conventional encoding methods, because the differences in depth are non-negligible. But, since there is no motion in video, these differences in depth are very likely to be due to unreliable depth estimation, and therefore can be ignored.

In this way, better coding efficiency can be achieved by taking into consideration depth map unreliability. Flickering artifacts due to temporal variation in depth map are also reduced, leading to overall improvements in perceptual quality. In addition, with this strategy one can select temporal skip in depth automatically, whenever temporal skip in video has been chosen, so that no skip mode information needs to be inserted in the depth bitstream. This leads to reduction in not only bitrate but also encoding complexity since it is possible to skip parts of the motion estimation and mode decision processes.

4. EXPERIMENTAL RESULTS

The new distortion metric and the skip mode selection scheme are simulated using several multiview test sequences. For each sequence both video and depth map are encoded for two selected views. The decoded video and depth map are used to synthesize an intermediate view between the two views using the software developed by Nagoya University [13].

First, the scale factor k_1 is calculated using the camera setting parameters for each sequence. Then, k_2 is found as described in Section 3.1, i.e., by estimating the effect of displacements in the first frame of the sequence. The result is given in Table 1. Note that each multiview sequence is acquired in a different camera setting, which would affect the amount of geometry error differently, and this difference is well reflected in k_1. For the outdoor scene sequences Z_{far} is large, thus Z_{near} is dominant parameter to decide k_1 when the camera distance and focal length are similar. This can be seen in 'Lovebird 1' and 'Lovebird 2' cases, where the former captures nearer object, thus the position error becomes more sensitive to the depth distortion resulting in larger k_1. In case of indoor scene sequences, all parameters can affect the amount of the position error caused by the depth distortin. For example, two indoor scene sequences 'Ballet' and 'Dog' have quite different value of k_1, where the former has dense camera setting to capture near objects compared to the other. The second scale factor, k_2 depend on the image characteristics. Comparing 'Champagne Tower' and 'Ballet', k_2 is larger for the former which contains a lot of objects resulting in large distortion in the synthesized view by position error.

The video is coded using H.264/AVC (joint model reference software ver. 13.2), and the depth map is coded using H.264/AVC with and without the proposed methods. To simplify test conditions, same encoding settings are used for video and depth map including the QP values of 24, 28, 32, and 36, and the Lagrange multiplier values, and only I- and P-slices are used to code 15 frames for each view.

Fig. 3 shows the rate-distortion curves to compare the coding efficiency of the proposed methods against H.264/AVC, where 'Method 1' is the result with the new distortion metric as given in (9), 'Method 2' is the result of the skip mode selection scheme as described in Section 3.2, and 'Method 1+2' is the result of the combined method, where the skip mode selection scheme is first applied, and for non-skipped blocks mode selection using the new distortion metric is performed. In the graphs, the x-axis is the bitrate for depth map coding of two views, and the y-axis is the PSNR of the synthesized view compared to the original view[1]. Table 1 contains

[1] In our experiments we select non-adjacent views in each sequence so that after view synthesis we can measure the distortion between the synthesized view and the actual view included in the dataset.

the BD-PSNR [14] results for various test sequences with Method 1+2 compared to H.264/AVC. These results show the efficiency of the proposed methods with maximum coding gain of 2.0 dB and 0.9 dB gain on average, which corresponds to 87% and 61% bitrate reduction, respectively. Both Method 1 and Method 2 perform better than H.264/AVC, and for most of the sequences Method 2 performs better than Method 1. By combining the two methods, additional gain can be achieved as shown in Fig. 3.

In addition, subjective quality is improved because flickering artifacts are reduced. The flickering artifacts occur in the synthesized views due to the temporal variation in the depth map. By applying the skip mode selection method, erroneous depth map information is coded using the skip mode, and as a result the flickering artifact is reduced. To see the variation in the static background region, the bottom right quarter of the synthesized Ballet sequence is taken from two temporally consecutive frames, and the difference image is shown in Fig. 4. It can be easily noticed that the temporal variation has been significantly reduced by the proposed method, leading to flickering artifact reduction.

Table 1. Scale factor k_1, k_2, and BD-PSNR

Sequence	k_1	k_2	BD-PSNR (dB)
Champagne Tower	0.282	65.238	0.34
Dog	0.078	41.671	0.75
Lovebird 1	0.214	24.807	0.24
Lovebird 2	0.057	29.265	1.96
Door Flowers	0.090	15.810	1.23
Newspaper	0.275	38.653	0.75
Ballet	0.442	7.723	1.23
Breakdancers	0.383	11.430	0.29
Average	-	-	0.85

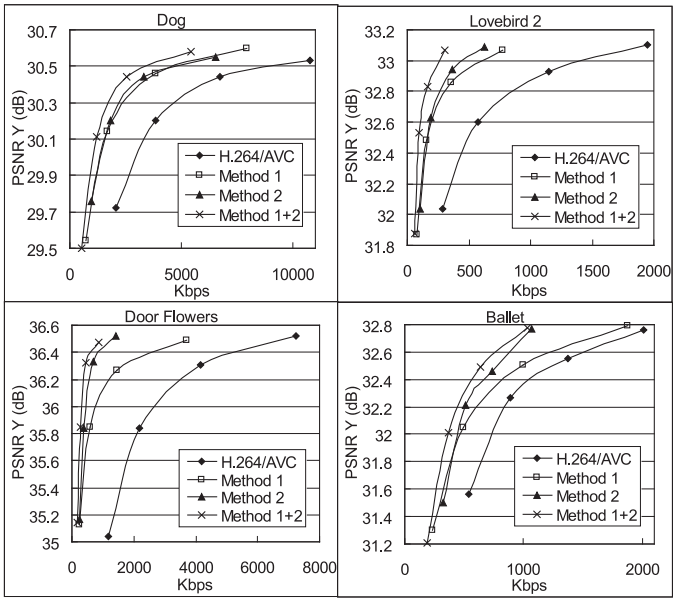
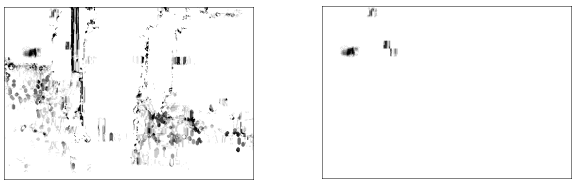


Fig. 3. Rate-distortion curves of the proposed methods.



(a) H.264/AVC (b) Proposed methods

Fig. 4. Example of flickering artifact reduction: (a) H.264/AVC and (b) proposed method.

5. CONCLUSION

Depth map distortion causes position errors in the synthesized views, which leads us to develop a new distortion metric for an optimized mode selection scheme in depth map coding. Using the proposed distortion metric and skip mode selection scheme, the experimental results show the coding gain of 0.9 dB or 61% bitrate reduction on average with better subjective quality. In future work, we plan to seek further improvements by joint optimization of video and depth map coding using the proposed distortion metric along with other coding tools.

6. REFERENCES

[1] A. Smolic, K. Mueller, N. Stefanoski, J. Ostermann, A. Gotchev, G. B. Akar, G. Triantafyllidis, and A. Koz, "Coding algorithms for 3DTV–a survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1606–1621, Nov. 2007.

[2] C. Fehn, R. de la Barré, and S. Pastoor, "Interactive 3-DTV – concepts and key technologies," *Proc. IEEE*, vol. 94, no. 3, pp. 524–538, Mar. 2006.

[3] K. Yamamoto, M. Kitahara, H. Kimata, T. Yendo, T. Fujii, M. Tanimoto, S. Shimizu, K. Kamikura, and Y. Yashima, "Multiview video coding using view interpolation and color correction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 11, pp. 1436–1449, Nov. 2007.

[4] L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2004)*, vol. 23, no. 3, pp. 600–608, Aug. 2004.

[5] Y. Morvan, D. Farin, and P. H. N. de With, "Depth-image compression based on an R-D optimized quadtree decomposition for the transmission of multiview images," in *Proc. of IEEE Int. Conf. Image Proc., ICIP 2007*. San Antonio, USA, Sep. 2007.

[6] R. Krishnamurthy, B.-B. Chai, H. Tao, and S. Sethuraman, "Compression and transmission of depth maps for image-based rendering," in *Proc. of IEEE Int. Conf. Image Proc., ICIP 2001*. Thessaloniki, Greece, Oct. 2001.

[7] D. Tzovaras, N. Grammalidis, and M.G. Strintzis, "Disparity field and depth map coding for multiview image sequence compression," in *Proc. of IEEE Int. Conf. Image Proc., ICIP 1996*. Lausanne, Switzerland, Oct. 1996.

[8] Y. Morvan, D. Farin, and P. H. N. de With, "Multiview depth-image compression using an extended H.264 encoder," *Advanced Concepts for Intelligent Vision Systems*, vol. 4678-2007, pp. 675–686, Aug. 2007.

[9] H. Oh and Y.-S. Ho, "H.264-based depth map sequence coding using motion information of corresponding texture video," *Advanced Concepts for Intelligent Vision Systems*, vol. 4319-2006, pp. 898–907, Dec. 2006.

[10] P. Lai, A. Ortega, C. Dorea, P. Yin, and C. Gomila, "Improving view rendering quality and coding efficiency by suppressing compression artifacts in depth-image coding," in *Proc. of Visual Communic. and Image Proc., VCIP '09*. San Jose, CA, USA, Jan. 2009.

[11] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[12] A. Ortega and K. Ramchandran, "Rate-distortion techniques in image and video compression," *IEEE Signal Proc. Magazine*, vol. 15, pp. 23–50, Nov. 1998.

[13] "Tanimoto Lab Nagoya University," http://www.tanimoto.nuee.nagoya-u.ac.jp/english/index.html.

[14] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T SG. 16 Q.6*, Document VCEG-M33, Apr. 2001.