INTERACTIVE MULTIVIEW IMAGE CODING

Andriy Gelman[1], *Pier Luigi Dragotti*[1], *Vladan Velisavljević*[2]

[1]Communications and Signal Processing Group, Imperial College, London, UK
[2]Deutsche Telekom Laboratories, Technische Universität Berlin, Germany

ABSTRACT

We propose a novel multiview compression method for multiview images. The algorithm supports random access for interactive applications and has low storage requirements. The fundamental component of the method is the layer-based representation, which partitions the data set into redundant layers characterized by a constant depth value. We exploit the redundant property of each layer and remove the side information uncertainty using Distributed Source Coding (DSC) principles. In comparison to independent coding, our method achieves a PSNR improvement of 3dB. Furthermore, we present a rate-distortion (RD) analysis which demonstrates that the proposed algorithm can achieve a better performance in comparison to independent coding.

1. INTRODUCTION

In recent years, Image Based Rendering (IBR) has been proposed as an alternative to the traditional rendering algorithms. The approach has a lower computational complexity and achieves photorealistic results by interpolating the novel viewpoints from existing data. To obtain artifact-free results, however, the scene must be sampled with a large number of cameras. These images are either transmitted or stored, which means efficient compression is an essential part of IBR systems [1].

The majority of the compression literature has focused on hierarchical prediction [2] or subband coding [3, 4]. Although these algorithms achieve high compression, they have limited random access. These techniques are, therefore, not suitable in an interactive setting, where the images are stored at a server and transmitted to the remote users on request. The key point is that the viewing trajectory is unknown prior to encoding.

A number of techniques have been proposed which achieve high compression and still maintain random access. For example, in [5] the authors propose storing multiple representations of an image for a set of possible predictions to reduce the transmission rate and eliminate drift. This method, however, requires high storage requirements at the server. A different approach [6, 7] has been to use DSC principles to reduce the storage size and eliminate the side information uncertainty.

In this paper we propose a novel multiview image coding method with random access and low storage requirements at the server. The fundamental component of the algorithm is the layer-based representation [8], which partitions the data set into layers each modeled by a constant depth plane. The redundancy of each layer is exploited using a spatial Discrete Wavelet Transform (DWT) and DSC principles, which we also use to remove side information mismatch at the user. The obtained transform coefficients are efficiently entropy coded and transmitted to the remote user on request. Additionally, the algorithm is complemented with a model which demonstrates that the approach can achieve a better RD performance than independent coding of images.

The outline of this paper is as follows. Multiview data structure and the layer-based representation are reviewed next. In Section 3 we present the proposed algorithm and in Section 4 discuss its RD performance. The results are presented in Section 5 and the paper is concluded in Section 6.

2. REVIEW OF MULTIVIEW IMAGE REPRESENTATION

In this section, we analyze the redundancy of multiview images and review the layer-based representation. For clarity, we simplify the setup to a 1D array of evenly spaced cameras perpendicular to the baseline, also known as the EPI [1]. This type of data set is parameterized as:

$$I = P_3\left(x, y, V_x\right), \tag{1}$$

where *I* is the pixel intensity, (x, y) are the spatial coordinates of the image and V_x is the camera location.

2.1. Multiview image data structure and redundancy

Assuming the scene is Lambertian and has no occlusions an object appears at different pixel locations x and x' seen from different viewpoint coordinates (frames) V_x and V'_x. This shift in pixel locations (*disparity*) $\Delta x = x - x'$ can be represented as a function of the corresponding viewpoint coordinates, depth Z of the object and focal length f, that is,

$$\Delta x = \frac{f\left(V'_x - V_x\right)}{Z}. \tag{2}$$

The obtained relation between the viewpoint V'_x and the spatial coordinates x' is also known as EPI line, along which the pixel intensity is constant. Observe that the occlusion ordering can also be predicted. Since the disparity Δx is inversely proportional to the depth, when two lines intersect, the line corresponding to the larger disparity will occlude the other.

2.2. Layer-Based Representation

The layer-based representation is an extension of the EPI line concept. The representation partitions the data into redundant regions where each layer is a collection of EPI lines modeled by a constant depth. Fig. 1(b) illustrates the representation of the dataset in Fig. 1(a). It can be observed that each layer preserves the linear structure corresponding to an object location in a 3D space.

Extraction of layers from a general 3D scene is a non-trivial task. Here, we use a variation of the level-set segmentation algorithm which was proposed in [8]. An advantage of this unsupervised method is that it can be extended to an arbitrary number of

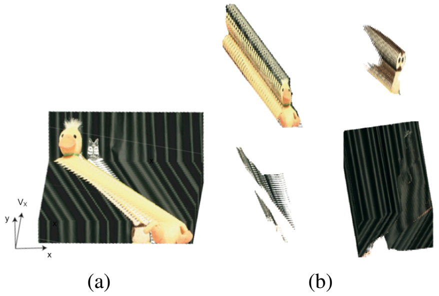

Fig. 1. (a) Multiview image cross-section. The data can be analyzed as a set of EPI lines with varying gradients. (b) Layer-based representation.

dimensions. Additionally, the algorithm efficiently handles occlusions, which is an important property for the subsequent compression algorithm.

3. PROPOSED ALGORITHM

In this section we propose our novel approach to encoding multiview images. First, we describe the general overview of the algorithm and, then, we outline an approach to obtain bit rate scalability.

3.1. Algorithm Overview

The concept of the proposed algorithm is to substitute the inter-view transform with DSC coding. This property allows the decoder to correctly reconstruct the transmitted data given any side information available in the cache of the user. The DSC ideas are applied to each layer independently in the spatial transform domain. We note that the redundant properties of the layers reduce the number of data bits which must be transmitted, thus providing a bit-rate saving in comparison to independent encoding of the images. Next, we outline the encoding process for one layer, which can be generalized to the complete dataset.

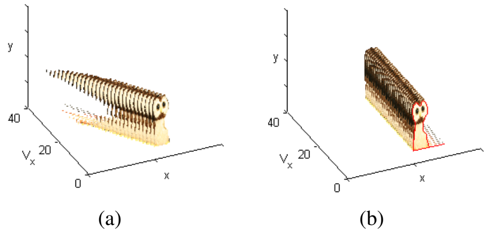

Fig. 2. (a) Due to occlusions, extracted layers might have discontinuities in the EPI lines. (b) Occluded pixels are interpolated using the mean of the non-occluded pixels and each image is disparity compensated onto a common view. The layer contour outlined with the red curve is efficiently encoded using a modified version of the Freeman algorithm [9] and transmitted.

Consider a layer from the Animal Farm dataset shown in Fig. 2(a). Initially, a preprocessing step is applied where the occluded regions are interpolated using the mean along the EPI lines and each image is disparity compensated onto a common view. The obtained layer is shown in Fig. 2(b). Note that the layer contour in each image is constant. This boundary is losslessly encoded using modified version of the Freeman algorithm [9] and transmitted along with the disparity.

In the following step, we reduce the intra-view redundancy by applying a 9/7 DWT to each image. We use a shape-adaptive implementation as proposed in [10] to remove the boundary effects associated with the irregular contour. Then, the resulting DWT subbands are quantized in a similar approach to [4], where the step-size is chosen using a Lagrangian multiplier λ. The obtained low-pass transform coefficients from the three images are illustrated in Fig. 3. Observe that the subbands are correlated across the views, which is exploited by the DSC algorithm in the following stage.

Recall that the cache of the remote user may contain DWT blocks from any image as side information. Our approach is to use DSC principles to remove the side information uncertainty. Consider the following model:

$$y = x + n, \tag{3}$$

where y is the transform coefficient requested by the user, x is the side information available in the cache and n is the residual signal. Recall that y can be correctly reconstructed transmitting at least $\lceil \log_2(2n) + 1 \rceil$ least significant bits (LSB) from y. To encode a sequence of blocks shown in Fig. 3, we take the worst case scenario, where any image can be used as side information. For example, the transform sequence $\{55, 51, 53\}$ requires $\lceil \log_2 8 + 1 \rceil = 4$ LSB to correctly reconstruct the data.

Observe that the outlined approach is inefficient when the transform coefficients across the views are the same except for one frame. For example in a sequence $\{55, 55, 57\}$ we have to transmit 3 LSB from each coefficient. Intuitively, a better solution would be to set 57 to 55, so that zero LSB are encoded. We solve this problem in an RD sense as follows: For each set of transform coefficients, we find the value which corresponds to the largest error and set it to the median of the set. Then, we evaluate the change in distortion ΔD and estimate change in rate ΔR. Using a greedy approach, if

$$\Delta D + \lambda \Delta R < 0 \tag{4}$$

we make the substitution and iterate the process until the cost is positive. The trade-off between rate and distortion is set using λ, which is determined when choosing the quantization step-size.

The server subsequently encodes the data using a bit-plane context adaptive arithmetic coder to attain rates close to the entropy of the source. The number of retained LSB is also encoded and transmitted with the data. This information is stored by the user for future reference. Note that the number of retained LSB provides a bit-plane significance map, which is further exploited by the entropy coder to reduce the encoding rate.

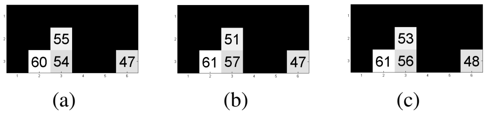

Fig. 3. Quantized low-pass subbands from three images. Observe that the blocks are correlated across the views.

3.2. Bit rate scalability

Wavelet based coding inherently supports scalable compression. This means that the data can be encoded once at the server and decoded multiple times for different target bit rates. To support bit rate scalability we operate as follows: for each residual block, we choose a transmission mode which minimizes the RD cost

$$D_i + \lambda_T R_i, \tag{5}$$

where D_i is the distortion and R_i is the transmission cost of the i-th mode and λ_T is chosen to meet a target bit rate. Note that to support interactive viewing, we evaluate the RD values off-line and store them at the server.

The blocks have three different modes. These are 'skip', 'skip-2' and 'lsb'. The first skip mode sets the data to zero and the second uses the side information in the cache as a prediction. The 'lsb' mode, however, transmits the residual bits to correctly reconstruct the data.

4. THEORETICAL MODELING

In this section we present a theoretical model of our encoding scheme. The aim of the model is to show that the proposed algorithm can achieve improved compression performance over independent coding. First, we present a synthetic data model which well approximates real multiview images, then we evaluate the RD relation of the independent and the proposed DSC algorithm for this class of signals.

4.1. Data Modeling

4.1.1. 2D $\alpha-$Lipschitz model

We model the layer images using a globally smooth 2D α-Lipschitz function $f_\alpha(x, y)$, which satisfies the following condition:

$$\mid f_\alpha(x_1, y_1) - f_\alpha(x_2, y_2) \mid \leq K \left(\mid x_1 - x_2 \mid^2 + \mid y_1 - y_2 \mid^2\right)^{\alpha/2}, \tag{6}$$

where $K > 0$. Transforming the signal using a 2D wavelet having at least $\lfloor \alpha + 1 \rfloor$ vanishing moments yields wavelet coefficients with the following decay [11]:

$$|d_{j,n}| \leq A 2^{-j(\alpha+1)}, \tag{7}$$

where j is the wavelet scale and constant $A > 0$. A linear compression scheme based on (7) can be designed by appropriately choosing constant quantization step size across all the subbands [11]. It can be shown that a high bit-rate assumption of the compression algorithm yields the following RD function:

$$D(R) \leq cR^{-\alpha}, \tag{8}$$

where R is the total number of bits allocated to encoding the signal and constant $c > 0$.

4.1.2. Contour Model

In practice, the layers are outlined by a segmentation and are therefore not globally α-Lipschitz smooth. To obtain the decay in (8), we transmit the contours and encode the texture using a shape adaptive scheme. We model the contour of the texture as a piecewise linear curve having V vertices. The RD function due to quantizing the location of the vertices can be upper bounded as [4]:

$$D(R) \leq A^2 T^2 V 2^{-R/2V}, \tag{9}$$

where A is the maximal magnitude of the texture and T is the maximal length of a side of the bounding box. The RD function is obtained by upper bounding the number of pixels affected due quantizing the vertex locations and then scaling this number by the amplitude of the texture to obtain the distortion.

4.1.3. Multiview Image Model

Using the analysis in Section 2.1, the layer images are modeled as a shifted version of the first view and a 2D α-Lipschitz error term. The error term corresponds to either lighting changes, layer extraction errors or non-Lambertian surfaces.

$$f_i(x, y) = f_1(x + (i-1)\Delta x, y) + \epsilon_\alpha^i(x, y), \tag{10}$$

where Δx is the layer disparity defined in (2) and i is the image location.

4.2. Independent Encoding

In the case of independent encoding, the 2D α-Lipschitz signal and the layer contour are separately encoded from each view. Using (8) and (9), the total distortion due to encoding the texture and the contour is bounded as:

$$\mathbf{D_{ind}}(\mathbf{R_t}) \leq \sum_{i=1}^{N} c_i \left(R_x^i\right)^{-\alpha} + NA^2T^2V2^{-R_v/2V}, \tag{11}$$

where R_x^i and R_v is the rate allocated to the α-Lipschitz texture in the i-th view and the contour encoding rate in each image, respectively and N is the total number of views. The total bit-rate can be shown to be:

$$\mathbf{R_t} = \sum_{i=1}^{N} R_x^i + NR_v. \tag{12}$$

The correct rate allocation which minimizes the distortion for a total bit budget can be solved using Lagrangian multipliers. A high rate analysis yields:

$$R_x^i \approx \mathbf{R_t} \left(\sum_{l=1}^{N} \left(\frac{c_l}{c_i}\right)^{\frac{1}{\alpha+1}}\right)^{-1}, \tag{13}$$

and

$$R_v = 2V \log_2 \left(\frac{A^2T^2 \ln(2)}{2\alpha c_1}\right) + 2V(\alpha+1)\log_2 R_x^1. \tag{14}$$

The minimized RD function in terms of the total rate can be obtained by substituting (13) and (14) into (11).

4.3. Proposed Algorithm

Using (10) we note that the wavelet coefficients in the residual frames can described using:

$$d_i^j = \widehat{d_1^j} + d_\epsilon^j, \tag{15}$$

where $\widehat{d_1^j}$ and d_ϵ^j denote the disparity compensated wavelet coefficients in the first frame and the wavelet coefficients of the α-Lipschitz error, respectively. By definition, the wavelet coefficients of the α-Lipschitz error can be upper bounded as:

$$|d_\epsilon^j| \leq A_\epsilon 2^{-j(\alpha+1)}. \tag{16}$$

Referring to (3), this analysis can be used to determine the number of LSB which must be transmitted to correctly reconstruct the texture. A similar analysis can be applied when the subbands of the signal are quantized.

The RD function of a globally smooth 2D α-Lipschitz signal, which is encoded using a DSC scheme can be shown to have the same RD behaviour as in (8). Therefore, the total RD due to encoding the dataset at the server using the DSC scheme is identical to (11) with different scaling constants. This behaviour will be validated in the following section.

Note that the RD bound in (8) can only be attained when the texture is globally α-Lipschitz smooth. This supports our layer-based coding scheme which partitions the data into redundant regions.

5. SIMULATION RESULTS AND ALGORITHM ANALYSIS

To evaluate the proposed algorithm, we compare the RD performance of our method to JPEG2000. To attain a fair comparison we have modified JPEG2000 to include the same entropy coding as our algorithm. In addition we compare the results with H.264/AVC High Profile when the viewing trajectory of the user is known by the server prior to encoding.

We first show that the proposed method achieves the RD performance presented in Section 4.3 for the signal model of Section 4.1. We then present numerical results on real data.

To this end, we encode an $\alpha = 1.5$ Lipschitz multiview image array consisting of four images. The data is encoded using a linear compression strategy where the optimal rate allocation for each image is obtained using (13). The model parameters are estimated by separately encoding each image in DSC or independent mode. Fig. 4 shows the theoretical and practical results when encoding the α-Lipschitz signal. Observe that the performance of the proposed algorithm is better in both the theoretical and practical results. As conjectured in Section 4.3, the independent and proposed algorithms have the same rate of decay but with different scaling constants.

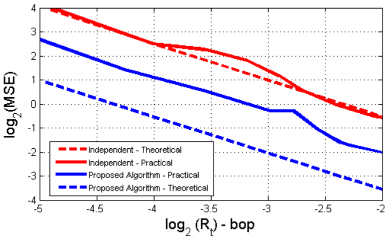

Fig. 4. Practical and theoretical RD performance when encoding an $\alpha = 1.5$ Lipschitz signal. The proposed algorithm achieves an improved performance in both practical and theoretical cases. Observe that the rate of decay in both the theoretical and practical cases is the same.

To analyse the performance when encoding real data, we use a multiview image sequence called Tsukuba $[282 \times 382 \times 4]$ from [12]. The first image of each layer is transmitted using the intra modality. Then, to mimic random access, the other images are randomly chosen and DSC encoded using the proposed method. Fig. 5 shows a quantitative comparison of the proposed scheme with H.264/AVC and independent coding. Observe that the predictive structure of H.264/AVC would require a significantly larger storage size to allow interactive viewing and is therefore not a fair comparison. However, we demonstrate that the proposed algorithm is competitive at low rates and incurs a 15% rate loss at 0.26bpp. In comparison to independent coding our approach achieves a gain of 3dB at 0.3bpp. Due to a lack of space, we only show experimental results for one real dataset. However, similar results have been obtained when encoding Teddy [12].

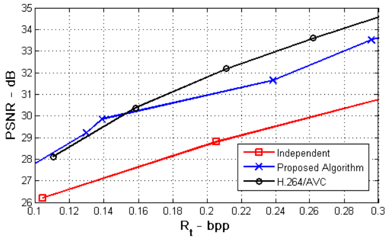

Fig. 5. Comparison of the proposed algorithm with H.264/AVC High Profile and independent compression when encoding Tsukuba.

6. CONCLUSION

We presented a novel multiview image compression algorithm with random access at image level. The fundamental component of the algorithm is the layer-based representation, which partitions the data into redundant layers each modeled by a constant depth value. Each layer is encoded independently and we use robust DSC coding principles to remove the side information ambiguity at the decoder. The algorithm achieves an improved RD performance with gains of up to 3dB over independent encoding. Furthermore, we have presented a RD analysis of our algorithm which demonstrates that the proposed approach can achieve a better performance in comparison to independent coding. Future work includes making the algorithm robust to segmentation errors which affects the RD performance of the proposed scheme.

7. REFERENCES

[1] C. Zhang and T. Chen, "A survey on image-based rendering–representation, sampling and compression," *Signal Processing: Image Communication*, vol. 19, no. 1, pp. 1–28, 2004.

[2] M. Magnor and B. Girod, "Data compression for light-field rendering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 3, pp. 338–343, Apr 2000.

[3] B. Girod, C.L. Chang, P. Ramanathan, and X. Zhu, "Light field compression using disparity-compensated lifting," *IEEE Transactions on Image Processing*, pp. 761–764, 2003.

[4] A. Gelman, P.L. Dragotti, and V. Velisavljevic, "Multiview image coding using depth layers and an optimized bit allocation," *submitted to IEEE Transactions on Image Processing*.

[5] P. Ramanathan and B. Girod, "Random access for compressed light fields using multiple representations," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*, 2004.

[6] N.-M. Cheung, A. Ortega, and G. Cheung, "Distributed source coding techniques for interactive multiview video streaming," in *Proceedings of the 27th conference on Picture Coding Symposium*, Piscataway, NJ, USA, 2009, PCS'09, pp. 269–272, IEEE Press.

[7] A. Aaron, P. Ramanathan, and B. Girod, "Wyner-ziv coding of light fields for random access," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*, 2004.

[8] J. Berent and P.L. Dragotti, "Plenoptic manifolds: Exploiting structure and coherence in multiview images," *IEEE Signal Processing Magazine*, vol. 24, no. 6, pp. 34–44, November 2007.

[9] Y.K. Liu and B. Zalik, "An efficient chain code with huffman coding," *in Pattern Recognition*, vol. 38, no. 4, pp. 553 – 557, 2005.

[10] S. Li and W. Li, "Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 5, pp. 725–743, Aug 2000.

[11] S. Mallat, *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*, Academic Press, 3rd edition, 2008.

[12] D. Scharstein and R. Szeliski, "Middlebury data sets," vision.middlebury.edu/stereo/.