ATTENTION-WEIGHTED DEPTH MAP RATE-ALLOCATION IN FREE-VIEWPOINT TELEVISION

Camilo Dorea and Ricardo L. de Queiroz

Department of Computer Science
University of Brasilia, DF, Brazil
Email: camilodorea@unb.br, queiroz@ieee.org

ABSTRACT

We propose optimal rate-allocation, using viewer attention information among viewpoints, for depth map cameras within a free-viewpoint television broadcast system. An attention-weighted rate-allocation framework enables bit-rate, or quality, to be distributed across the multiple cameras in accordance with viewer interest, minimizing total observed distortions perceived among all viewers. Prior work has considered attention-weighted rate-allocation for texture cameras only in such systems. Depth maps, nonetheless, are a common requirement for view rendering in many systems. They may constitute a significant part of overall transmission bit-rate and they present their own distortion characteristics, different from those of texture. We model the effects of depth camera distortions upon virtual or synthesized views and propose an optimal attention-weighted rate-allocation among depth cameras. Results show significant gains in average PSNR of synthesized views and bit-rate savings of our proposal relative to the balanced rate-allocation alternative.

Index Terms— Free-viewpoint TV, rate-allocation, attention weighting, depth maps.

1. INTRODUCTION

Free-viewpoint Television (FTV) [1] enables a viewer to freely select the desired viewpoint from which to observe a scene. The scene is generally captured through a possibly large but finite number of cameras and broadcast to a potentially very large audience. Viewpoints may either coincide with an existing camera's position or must be synthesized at intermediate positions (virtual cameras) with information from existing adjacent cameras. One of the most common virtual view synthesis procedures, also adopted here, is depth-image-based rendering (DIBR) [2]. We assume depth maps are captured (or estimated) and transmitted along with texture images for each of the existing cameras in order to allow DIBR, either directly by viewers or at some point within a network cloud. Our system architecture is depicted in Fig. 1.

This work was partially supported by CNPq grant 308150/2014-7.

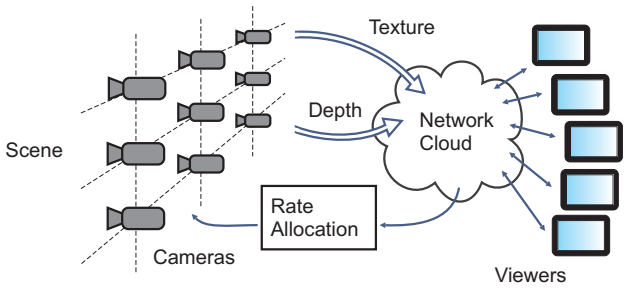


Fig. 1. General architecture for FTV broadcast using cloud services. Viewer attention feedback is used to determine optimal rate-allocation among texture and depth cameras.

All cameras are subject to standardized compression with a given quality and potentially broadcast to everyone within the viewer network. We consider simulcasting independently compressed video streams as multiview coding solutions require a single, centralized coder which may be unfeasible when dealing with a large number of dispersed cameras.

The viewpoint choices among the multiple viewers lead to an attention profile which may often reflect unequal demand across cameras. In [3] an attention-based texture rate-allocation procedure, optimal in the total observed distortion sense, was initially proposed. In it, cameras receiving greater attention also receive greater quality (or bit-rate). The proposal was extended in [4] for general camera arrangements, beyond uniformly-spaced 1D arrangements. Both proposals report significant bit-rate savings with respect to common, balanced rate-allocation. Nevertheless, neither [3] nor [4] address attention-weighted rate-allocation of depth maps. In [3] depth maps were uniformly coded (i.e., with constant quantization parameter) across all cameras whereas in [4] depth maps were assumed to be available in lossless form. Depth maps are essential in DIBR and can constitute significant portion of overall bit-rate (up to 25% in [5] and higher in our simulations). Furthermore, the effects of depth distortions, in particular, asymmetric distortions arising from non-uniform bit allocations, upon observed view distortion are distinctive from those derived for texture [3], [4]. In recent work [6],

depth map rate-allocation has been addressed, however, it is not separately treated but rather cast as a function of attention-weighted texture optimization, leading to different results.

Several other proposals [7–10] address depth, and texture, rate allocation in the context of multiview video. These consider video content distortion, not the total observed distortion of our case. Other works [11–13] present attention-driven approaches, but in response to demands of a single viewer.

This work proposes attention-weighted depth map rate-allocation. We develop a distortion model which considers effects of asymmetric depth map distortions upon view synthesis. The model is used to determine optimal, attention-weighted, bit allocation across depth cameras. We present results showing significant gains in terms of total observed distortion with respect to the uniform rate-allocation case for several viewer distributions among 1D and 2D camera arrangements.

2. DISTORTION MODEL FOR DEPTH

We label viewpoints, captured by any of N cameras, as $\vec{c}_j$. Each captured viewpoint is associated to texture and depth image components, whose pixel values are denoted $p_j(x,y)$ and $d_j(x,y)$, respectively. In DIBR, adjacent captured images are used to synthesize a virtual texture image at viewpoint $\vec{v}_m$ through the weighted blending [14] of projected pixels $p_{m|j}(x',y')$ such that

$$p_m(x',y') = \sum_{j\in\Psi_m} \beta_{mj}\; p_{m|j}(x',y') \tag{1}$$

where Ψ_m is the set of indices of cameras used in synthesis at virtual viewpoint $\vec{v}_m$, and β_{mj} are the camera blending weights which guarantee that, when available, projections from cameras closer to synthesis position are given greater relevance.

Without loss of generality, consider distances between viewpoints to be restricted to horizontal shifts described by distance measure $g_x(\vec{v}_m,\vec{c}_j)$. Thus, β_{mj} is defined as inversely proportional to $g_x(\vec{v}_m,\vec{c}_j)$ and, as a weight, $0 \le \beta_{mj} \le 1$ while $\sum_j \beta_{mj} = 1$, for a given synthesis position.

Depth information $d_j(x,y)$ is appropriately scaled into disparity values and used in determining correspondences between pixels of viewpoints $\vec{c}_j$ and $\vec{v}_m$ such that

$$p_{m|j}(x',y) = p_j\big(\; x + kg_x(\vec{v}_m,\vec{c}_j)d_j(x,y),\; y\;\big)\;. \tag{2}$$

The k component of the depth scaling factor represents intrinsic camera parameters, such as focal length and depth normalization constants [14]. Distances between viewpoints also take part in depth to disparity scaling.

Differently from texture distortions, the effects of depth distortions are only indirectly observed upon synthesized views. An error in depth pixel value of the form $\hat{d}_j(x,y) = d_j(x,y)+\varepsilon_j(x,y)$ leads to disparity error, i.e., a displacement in the position of projected pixels used towards synthesis, denoted now as $\hat{p}_m(x',y)$.

In general, larger disparity errors cause larger distortion in synthesized view as spatial correlation of projections is degraded by displacements. Effects of depth distortion on synthesis are also subject to factors such as scene content and camera parameters. Synthesis distortion has been modeled in [15] as a linear function of disparity error, subject to a constant scaling factor s. Under this model, $\hat{p}_m(x',y)$ may be expressed as a sum of an error-free projected pixel term and a scaled disparity error term

$$\begin{aligned}\hat{p}_m(x',y) = &\sum_{j\in\Psi_m} \beta_{mj} p_{m|j}(x',y) + \\ &\sum_{j\in\Psi_m} \beta_{mj} s[kg_x(\vec{v}_m,\vec{c}_j)\varepsilon_j(x,y)]\;.\end{aligned} \tag{3}$$

Note that the scaled disparity error term is also subject to weighted blending. Here, the depth error $\varepsilon_j(x,y)$ is multiplied by $g_x(\vec{v}_m,\vec{c}_j)$ and by β_{mj}, which in turn has been defined as inversely proportional to $g_x(\vec{v}_m,\vec{c}_j)$. We assume projections from multiple cameras are generally available for synthesis and that these are subject to weighted blending. Under this condition, unlike texture errors, the effects of depth error upon synthesis are in essence independent of the distance between viewpoints. This property is considered in the following optimization procedure.

3. OPTIMAL BIT-RATE ALLOCATION

Depth from each camera video, say the n-th camera, is compressed and transmitted using, for example, H.264/AVC [16] with a given quantization parameter (QP), spending encoding bit-rate R_n and achieving distortion D_n. Each viewer observes a virtual image synthesized from cameras operating near the selected viewpoint $\vec{v}_m$. For each viewpoint, we assign a distortion δ_m arising from the compression of the camera videos used in view synthesis.

We argue that $\{D_n\}$ is not directly relevant but rather the observed view distortions $\{\delta_m\}$ as these will be the ones actually experienced by the viewers. In this context, we would like to minimize overall bit-rate

$$R = \sum_{n=0}^{N-1} R_n \tag{4}$$

subject to total observed distortion

$$D = \sum_{m=0}^{M-1} \delta_m. \tag{5}$$

For such, the mechanisms we have at our disposal are the selection of QP for each depth camera compressor, which

controls rate $\times$ distortion trade-off. In order to minimize (4), we account for viewer attention in (5) while expressing observed view distortion δ_m in terms of depth camera distortions.

Assuming proportionality between observed view distortion and distortions in the captured depth views, the former is approximated as a linear combination of the latter such that

$$\delta_m = \sum_{j \in \Psi_m} u_j \, D_j \tag{6}$$

where u_j are weights ($0 \leq u_j \leq 1$, such that $\sum u_j = 1$).

Note that the initial assumption, from [15], is that depth distortion from cameras at greater distance are indeed more influential than those from smaller distances. However, the subsequent weighted blending operation, discussed in Sec. 2, assigns proportionally less weight to cameras at greater distance, canceling the initial distance dependence. Thus, differently from texture images and their distortions [4], we assume each depth camera has a proportional contribution towards δ_m, irrespective of distance to virtual viewpoint. With $\|\Psi_m\|$ being the number of cameras in Ψ_m, the depth camera distortion weights are such that $u_j = \alpha_{ij}/\|\Psi_m\|$ where

$$\alpha_{ij} = \begin{cases} 1 & \text{if viewpoint at } \vec{v_i} \text{ uses camera at } \vec{c_j} \\ & \text{for synthesis,} \\ 0 & \text{else} \end{cases}$$

and the observed view distortion of (6) may be re-written as

$$\delta_m = \sum_{n=0}^{N-1} \frac{\alpha_{mn}}{\|\Psi_m\|} \, D_n \, . \tag{7}$$

Defining viewer-dependent depth camera weights as

$$\gamma_n = \sum_{m=0}^{M-1} \frac{\alpha_{mn}}{\|\Psi_m\|} \, , \tag{8}$$

the total observed distortion of (5) becomes

$$D = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \frac{\alpha_{mn}}{\|\Psi_m\|} \, D_n = \sum_{n=0}^{N-1} \gamma_n \, D_n = \sum_{n=0}^{N-1} \hat{D}_n \, . \tag{9}$$

It is a function of adjusted distortion measures for each camera $\{\hat{D}_n\}$ which take into account all viewers as well as all depth cameras. We thus seek to minimize

$$J = R + \lambda D = \sum_{n=0}^{N-1} (R_n + \lambda \hat{D}_n) \, . \tag{10}$$

At each camera compressor, optimal bit-allocation is found by determining QP for depth in order to minimize

$$J_n = R_n + \lambda \gamma_n D_n \tag{11}$$

for a given Lagrangian multiplier λ [17], responsible for the rate $\times$ distortion trade-off.

4. EXPERIMENTAL RESULTS

The proposed attention-weighted depth map rate-allocation procedure was tested on publicly available data sets *Pantomime* [18] and *Akko & Kayo* [19]. Results employ the H.264/AVC JM Reference Software v18.0 [20] for compression and the MPEG View Synthesis Reference Software v3.5 [21].

For each data set, N cameras are selected and M viewpoints are randomly chosen. Given a viewpoint distribution, viewer-dependent depth camera weights γ_n are computed through (8). Distortion is measured in terms of the MSE between viewpoints synthesized from compressed and uncompressed adjacent camera views. Total observed distortion is taken as the MSE across all M viewpoints and reported in terms of PSNR. Overall bit-rate considers the sum of all N depth camera rates and is reported in terms of bits per pixel per camera (bpc) using the first frame of each view.

Our comparisons include a *Uniform* rate-allocation in which all texture and all depth cameras employ a uniform QP and our *Attention-weighted* rate-allocation for depth in which texture cameras employ a uniform QP. Uniform texture QPs are selected from the range $\{17, 22, 27, 32, 37\}$ and uniform depth QPs from the range $\{2, 3, 4, 5, 6\}$. The uniform depth QP range was chosen to secure proportionality between depth distortions and observed view distortion for the tested sequences. Note that for data sets with low quality depth maps, large depth QP ranges may include particular quantization levels which can violate our proportionality assumption, see Fig. 2. In specific cases, certain larger QP value (such as suggested in [22]) were observed to improve synthesis distortion with respect to lower QP values by contributing to depth noise removal.

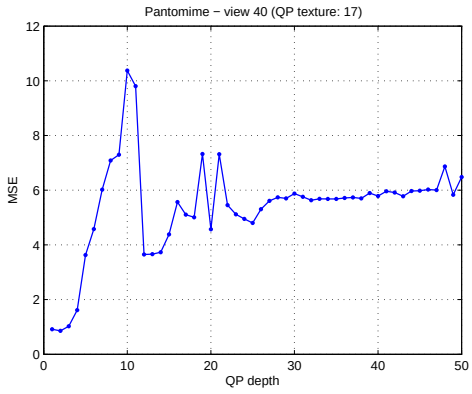


Fig. 2. MSE of synthesis of Pantomime (view 40) from coded depth images (views 39 and 41) with respect to synthesis from original reference images.

We adopt the viewer distributions of [4], briefly described next. For the *Pantomime* (1280×960) data set, with even numbered cameras in the range 20-58, two distributions were

tested: a bimodal Gaussian distribution of 200 and 300 viewers centered at viewpoints 29 and 49 with standard deviations 5 and 4, respectively (shown in Fig. 3) and a sharper Laplacian distribution of 400 viewpoints with mean 37 and standard deviation 3. For the *Akko & Kayo* (640×480) data set, we selected the cameras originally labeled 27-29, 47-49 and 67-69 with accompanying depth maps. Each group of cameras is uniformly distributed across one of three rows with 5 cm of horizontal and 20 cm of vertical spacing among them. A total of 400 viewpoints are randomly spread according to a Gaussian distribution centered at coordinates (3.75, 15) cm from the origin, set at camera 27, and standard deviation of (2, 16) cm in horizontal and vertical directions.

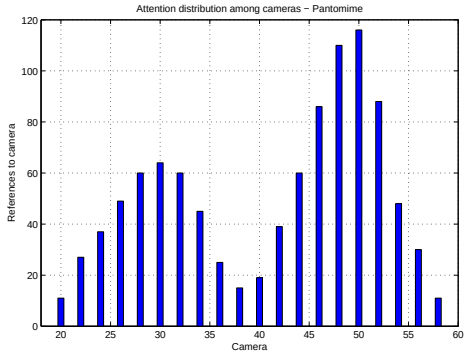
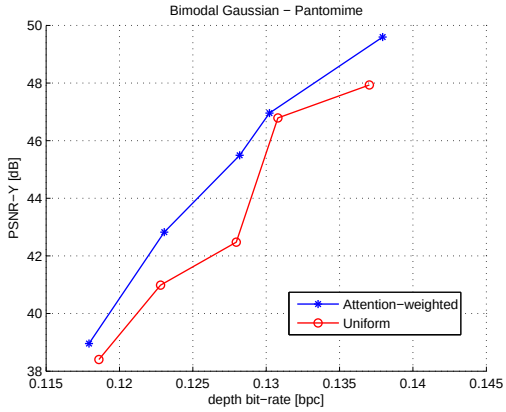


Fig. 3. Bimodal Gaussian distribution of viewer attention across cameras for *Pantomime*.

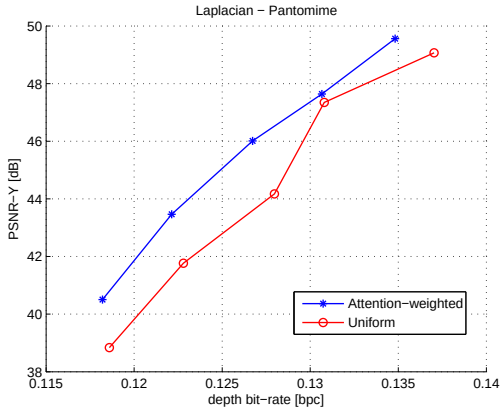
For *Pantomime*, under the bimodal Gaussian instantiation, the proposed *Attention-weighted* allocation for depth achieves average PSNR gain [23] of 1.56 dB over *Uniform*. R-D performance comparisons can be seen in Fig. 4. The sharper Laplacian instantiation produces the largest average PSNR gain, of 1.68 dB. More modest gains, of 0.64 dB, are found for *Akko & Kayo*. This may be attributable to the "flatter" 2D Gaussian viewer distribution.

5. CONCLUSIONS

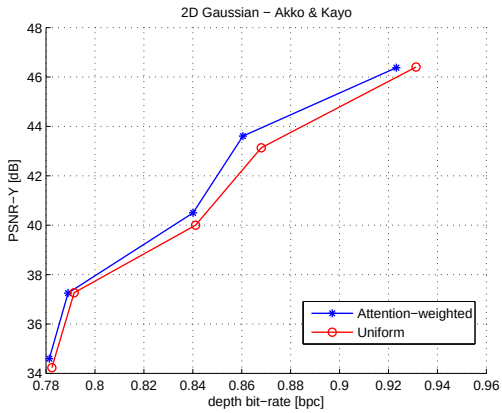
We proposed an attention-weighted depth map rate-allocation technique to minimize the total observed distortion within an FTV broadcast system. The proposal builds on and completes work for texture camera rate-allocation in attention-weighted FTV [3, 4]. We specifically model the effects of depth camera distortion upon observed view distortion. Results show significant gains in average PSNR of our proposal relative to uniform rate-allocation for depth cameras. Future work includes the extension to coding structures beyond simulcast.



(a)



(b)



(c)

Fig. 4. Overall R-D performance comparison of the *Attention-weighted* and *Uniform* rate-allocations for (a) *Pantomime* with bimodal Gaussian, (b) *Pantomime* with Laplacian and (c) *Akko & Kayo* with 2D Gaussian viewer distribution.

6. REFERENCES

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-viewpoint TV," *IEEE Signal Processing Magazine*, vol. 28, no. 1, January 2011.

[2] C. Fehn, "Depth-image-based rendering (DIBR), compression, and transmission for a new apprioach on 3D-TV," *Proc. SPIE 5291, Stereoscopic Displays and Virtual Reality Systems XI*, vol. 93, May 2004.

[3] T. Scandaroli, R. L. de Queiroz, and D. Florencio, "Attention-weighted rate allocation in free-viewpoint television," *IEEE Signal Processing Letters*, vol. 20, no. 4, April 2013.

[4] C. Dorea and R. L. de Queiroz, "General rate-allocation in free-viewpoint television," in *IEEE Int. Conf. on Image Processing*, Paris, France, October 2014.

[5] K. Muller, A. Smolic, K. Dix, P. Merkle, and T. Wiegand, "Coding and intermediate view synthesis of multiview video plus depth," in *IEEE Int. Conf. on Image Processing*, November 2009.

[6] C. Dorea and R. L. de Queiroz, "Attention-weighted texture and depth bit-allocation in general-geometry free-viewpoint television," *to appear in IEEE Transactions on Circuits and Systems for Video Technology*, 2016.

[7] Y. Liu, Q. Huang, S. Ma, D. Zhao, and W. Gao, "Joint video/depth rate allocation for 3D video coding based on view synthesis distortion model," *Signal Processing: Image Communication*, vol. 24, no. 8, 2009.

[8] Y. Liu, Q. Huang, S. Ma, D. Zhao, W. Gao, S. Ci, and H. Tang, "A novel rate control technique for multiview video plus depth based 3D video coding," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, 2011.

[9] E. Bosc, V. Jantet, M. Pressigout, L. Morin, and C. Guillemot, "Bit-rate allocation for multi-view video plus depth," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, 2011.

[10] N. Ozbek, A. M. Tekalp, and E. T. Tunali, "Rate allocation between views in scalable stereo video coding using an objective stereo video quality measure," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[11] E. Kurutepe, M. R. Civanlar, and A. M. Tekalp, "Client-driven selective streaming of multiview video for interactive 3DTV," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 11, November 2007.

[12] C. Zhang and D. Florencio, "Joint tracking and multiview video compression," in *Proceedings of the SPIE*, July 2010, vol. 7744.

[13] J. Chakareski, V. Velisavljevic, and V. Stankovic, "User-action-driven view and rate scalable multiview video coding," *IEEE Transactions on Image Processing*, vol. 22, no. 9, September 2013.

[14] D. Tian, P.-L. Lai, P. Lopez, and C. Gomila, "View synthesis techniques for 3D video," in *Applications of Digital Image Processing XXXII, Proceedings of the SPIE*, 2009, vol. 7443 (2009).

[15] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map distortion analysis for view rendering and depth coding," in *IEEE Int. Conf. on Image Processing*, Cairo, Egypt, November 2009.

[16] *ITU-T Recommendation and International Standard of Joint Video Specification*, ITU-T Rec H.264/ISO/IEC 14496-10 AVC, March 2005.

[17] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, no. 3, 1963.

[18] M. Tanimoto, M. Fujii, and K. Fukushiuma, "1D parallel test sequences for MPEG-FTV," in *ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15378*, Archamps, France, April 2008.

[19] M. Tanimoto, M. Fujii, T. Senoh, T. Aoki, and Y. Sugihara, "Test sequences with different camera arrangements for call for proposals on multiview video coding," in *ISO/IEC JTC1/SC29/WG11 MPEG 2005/M12338*, Poznan, Poland, July 2005.

[20] "JM H.264 reference software v18.0," in *http://iphome.hhi.de/suehring/tml/*.

[21] M. Tanimoto, M. Fujii, T. Suzuki, K. Fukushima, and N. Mori, "Reference softwares for depth estimation and view synthesis," in *ISO/IEC JTC1/SC29/WG11 MPEG 2008/M15377*, Archamps, France, April 2008.

[22] A. Rusanovsky, K. Muller, and A. Vetro, "Common test conditions of 3DV core experiments," in *ITU-T SG16/WP36 ISO/IEC JTC1/SC29/WG11 JCT2-A1100*, Stockholm, Sweden, July 2012.

[23] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," in *ITU-T SG16/Q6 VCEG-M33*, Austin, TX, USA, March 2001.