GENERAL RATE-ALLOCATION IN FREE-VIEWPOINT TELEVISION

Camilo Dorea and Ricardo L. de Queiroz

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

ABSTRACT

We propose a framework for optimal rate-allocation in free-viewpoint television (FVTV) for a general camera arrangement based on the attention the viewers are paying to each camera. In a recent letter [1], the authors proposed a FVTV broadcast architecture and an optimal bit-allocation approach, assuming a uniformly-spaced one-dimensional arrangement of cameras. Quality (or bit-rate) at each camera was determined by viewer attention in order to minimize total observed distortion. Here, we extend the optimized bit-allocation scheme to allow for a more generic camera arrangement in FVTV. We present results on data sets from 1D and 2D array camera setups which show significant overall PSNR gains and bit-rate savings with respect to equally-balanced rate-allocation across cameras.

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

1. INTRODUCTION

In the framework proposed in [1], a free-viewpoint television (FVTV) [2] system is presented wherein there are a number of cameras used to simultaneously broadcast different point of views of a given scene. The broadcast is watched by a potentially very large audience over a cloud network and each viewer may choose its viewpoint. If the user requests a viewpoint where no camera is used, the requested view is synthesized based on the closest available camera viewpoints.

In our framework, we assume each camera's video stream is subject to compression and that depth maps [3], employed for view synthesis, are available in lossless form. Note that specialized algorithms for depth coding exist [4, 5] and, typically, the rate overhead of depth information is relatively small [6]. We further assume there are N cameras, which we generalize here each one to be at viewpoint $\vec{c_n}$, and there are M viewers each one selecting a viewpoint $\vec{v_m}$. All cameras are broadcast, according to our architecture depicted in Fig. 1, and each viewer has its viewpoint informed back to the transmitter using a slower channel of say b bits/s. Hence, the system consumes bM bits/s of transmitting band for all feedback channels spread all over the network. Viewpoint choice, or viewer attention, information is then used to allocate distinct camera bit-rates for optimal overall distortion observed by viewers.

This work was partially supported by CNPq grants 310375/2011-8, 302853/2011-1 and 470940/2010-7.

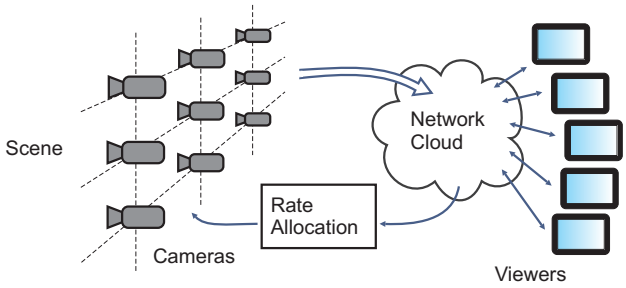

Fig. 1. General architecture for broadcast FVTV using cloud services. Viewer attention feedback is used to determine optimal rate-allocation among cameras.

Differently from [1], we do not assume a uniformly-spaced 1D camera arrangement but rather a general, possibly 2D array, arrangement. The main contribution of this paper is the generalization of the rate-allocation framework, initially proposed in [1], to accommodate general camera setups. We formulate [1] as a particular case of our framework. We present results showing significant gains in overall observed distortion and overall bit-rate savings on data from non-uniform 2D array camera setups and 1D linear setups as well.

2. BIT-RATE ALLOCATION

Each camera video, say the n-th camera, is compressed and transmitted using, for example, H.264/AVC [7] with a given quantization parameter (QP), spending an encoding bit-rate R_n and achieving a distortion D_n. Each viewer observes a synthetic image, synthesized from the video of cameras operating near the selected viewpoint. For each synthetic viewpoint we can assign a distortion δ_m which tells us the distortion incurred to the view when the camera videos are compressed.

We argue that the $\{D_n\}$ are not directly relevant and what is of real importance are the observed distortions $\{\delta_m\}$ which are experienced by the viewers, who are the actual consumers in the process. In this context, we want to minimize the overall rate

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

while also minimizing the overall observed distortion

$$D = \sum_{m=0}^{M-1} \delta_m. \tag{2}$$

For that, the mechanisms we have are the selection of the QP for each camera compressor, which controls the individual $R_n \times D_n$ trade-off.

It is natural that for synthesizing a given viewpoint m we would use a few cameras, for example, 2 in a linear arrangement: those immediately to the left and right of the requested viewpoint. Our model for the observed view distortion as a function of the distortions in the captured views is to approximate the former as a linear combination of the latter, i.e.,

$$\delta_m = \sum_{j \in \Psi_m} u_j \, D_j \tag{3}$$

where Ψ_m is the set of indexes of the cameras used to synthesize $\vec{v_m}$, and u_j are the respective weights ($0 \leq u_j \leq 1$, such that $\sum u_j = 1$). We assume the influence of a given camera to be inversely related to its distance from the desired viewpoint [8]. The closer $\vec{v_m}$ to $\vec{c_k}$, the larger u_k, until the point that the views coincide and $u_k = 1$ with all others $u_{j \neq k} = 0$. We, then, use the following auxiliary weights:

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

Let $g(\vec{v_i}, \vec{c_j}) = g(\vec{c_j}, \vec{v_i})$ be an absolute distance measure between viewpoints. And let

$$\beta_{ij} = 1 - \frac{g(\vec{v_i}, \vec{c_j})}{\sum_{k=0}^{N-1} \alpha_{ik} \, g(\vec{v_i}, \vec{c_k})}, \tag{4}$$

such that $\{\beta_{ij}\}$ enjoys the properties of the weights $\{u_j\}$ above. Then,

$$\delta_m = \sum_{n=0}^{N-1} \alpha_{mn} \, \beta_{mn} \, D_n \, , \tag{5}$$

so that

$$D = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \alpha_{mn} \, \beta_{mn} \, D_n \tag{6}$$

If we define

$$\gamma_n = \sum_{m=0}^{M-1} \alpha_{mn} \, \beta_{mn} \, , \tag{7}$$

then

$$D = \sum_{n=0}^{N-1} \gamma_n \, D_n = \sum_{n=0}^{N-1} D'_n \, . \tag{8}$$

One can view the above equation as a sum of the adjusted distortion measures $\{D'_n\}$ for each camera, which takes into account all the viewers. Here, we minimize

$$J = R + \lambda D = bM + \sum_{n=0}^{N-1} (R_n + \lambda D'_n). \tag{9}$$

At each camera compressor the optimal bit-allocation is found by adjusting the QP in order to minimize

$$J_n = R_n + \lambda \gamma_n D_n \tag{10}$$

for a given λ, responsible for the rate vs. distortion trade-off.

2.1. Special case of uniform linear arrangements

In the special case of uniformly spaced cameras in one dimension (in a line or arc), we can normalize distances to the interval between cameras and measure all distances from the first camera, such that viewpoints $\{\vec{v_i}\}$ and $\{\vec{c_i}\}$ can be described as scalars indicating the position in the path of cameras and viewpoints. Then, $0 \leq v_k, c_k \leq N-1$, $g(c_k, c_{k-1}) = 1$, and $c_k = k$. Let us separate integer and fractional parts of a viewpoint as $p_k = \lfloor v_k \rfloor$ and $q_k = v_k - p_k$. Hence, v_k is synthesized from cameras at p_k and p_k+1, such that $g(c_{p_k}, v_k) = q_k$ and $g(c_{p_k+1}, v_k) = 1 - q_k$.

The distortion associated with the k-th viewpoint is simply:

$$\delta_k = (1 - q_k) D_{p_k} + q_k D_{p_k+1}, \tag{11}$$

such that $\alpha_{ij} = 0$ except for

$$\alpha_{i,p_i} = \alpha_{i,p_i+1} = 1 \tag{12}$$

Also,

$$\beta_{i,p_i} = 1 - q_i \;\text{ and }\; \beta_{i,p_i+1} = q_i \tag{13}$$

which is the linear-arrangement case described in [1].

2.2. Extreme cases

There are two extreme cases. Firstly, the proposed bit-allocation would not improve the total observed distortion in case there is a "uniform" distribution of viewers, so that at every interval in between cameras, there is the same distribution of viewers. This would imply that all cameras (apart

from the first and last) receive exactly the same influence from viewers and, thus, they have equal weights for their distortion. Therefore, each camera would have equally important distortion and the uniform bit-allocation should be optimal. Secondly, in the other extreme, if there is only one camera being watched, no other camera is transmitted and we re-allocate all their bit budgets to the camera being watched. This would cause the largest gain in optimizing the bit-allocation for the cameras.

3. EXPERIMENTAL RESULTS

The proposed framework was tested on publicly available data sets *Akko & Kayo* [9] and *Pantomime* [10], captured through 2D array and 1D line camera arrangements, respectively. Results presented in this section employ the H.264/AVC JM Reference Software v18.0 [11] for compression and the MPEG View Synthesis Reference Software v3.5 [12].

For each data set, N cameras are selected and M viewpoints are randomly chosen. Given a viewpoint distribution, viewer-dependent camera weight γ_n is computed through (7). Distortion is measured in terms of the MSE between viewpoints synthesized from compressed and uncompressed adjacent camera views. Overall 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 camera rates and is reported in terms of bits per pixel per camera (bpc) using the first frame of each view. As noted, employed depth maps are assumed to be available in lossless form and are not considered within the rate summation. Our comparisons include a "uniform" rate-allocation in which all cameras employ the same QP and an "optimal" allocation. Uniform QPs are selected from the range $\{17, 22, 27, 32, 37\}$. Optimal allocation determines a distinct QP for each camera, by minimizing (10) subject to suitable λ.

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. An instantiation of viewer attention distribution among cameras is shown in Fig. 2. Overall rate and observed distortions for uniform and optimal allocations are presented in Fig. 3. The optimal case achieves an average PSNR gain of 1.54 dB and bit-rate savings of 24.5% over the uniform case.

Two viewer distributions were tested for the *Pantomime* (1280×960) data set with even numbered cameras in the range 20-58. A bimodal Gaussian distribution of 200 and 300 viewers centered at cameras 29 and 49 with standard deviations 5 and 4, respectively, is shown in Fig. 4. A Laplacian distribution of 400 viewpoints with mean 37 and standard deviation 3 is depicted in Fig. 5. With the bimodal Gaussian instantiation, optimal allocation achieves average PSNR gain of 0.48 dB and bit-rate savings of 9.2% relative to the uniform allocation case, as shown in Fig. 6. Of all tested viewer distributions, the sharper Laplacian instantiation produces the largest average PSNR gains and bit-rate savings, 2.41 dB and 39.3%, respectively, as pictured in Fig. 7.

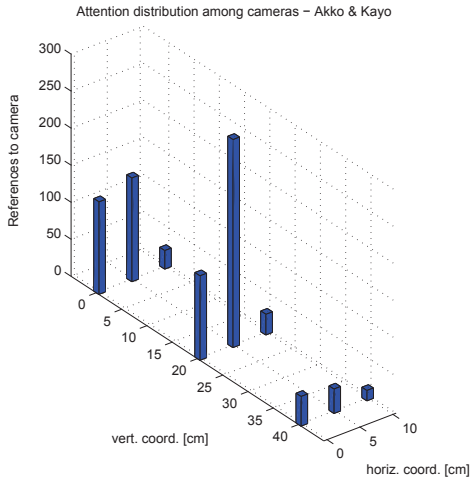

Fig. 2. Gaussian distribution of viewer attention over 2D camera array for *Akko & Kayo*.

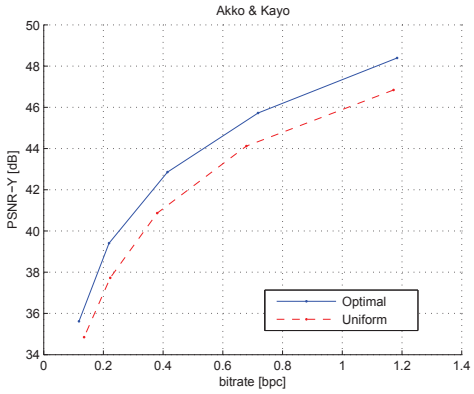

Fig. 3. Overall R-D performance comparison between optimal and uniform rate-allocation for *Akko & Kayo* with Gaussian viewer distribution.

4. CONCLUSIONS

We proposed an extension of [1] in order to accommodate a general camera arrangement within a rate-allocation framework. Based on viewer attention distribution, cameras (pos-

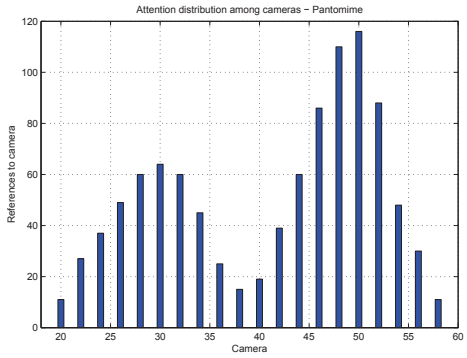

Fig. 4. Bimodal Gaussian distribution of viewer attention over 1D camera setup for *Pantomime*.

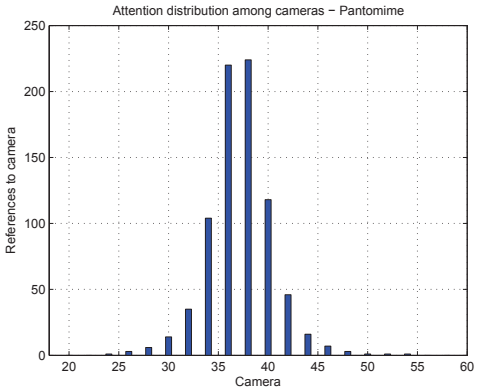

Fig. 5. Laplacian distribution of viewer attention over 1D camera setup for *Pantomime*.

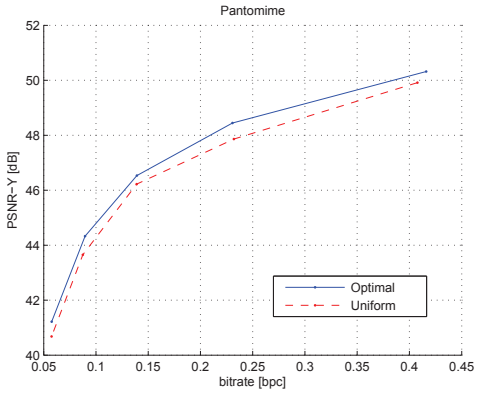

Fig. 6. Overall R-D performance comparison between optimal and uniform rate-allocation for *Pantomime* with bimodal Gaussian viewer distribution.

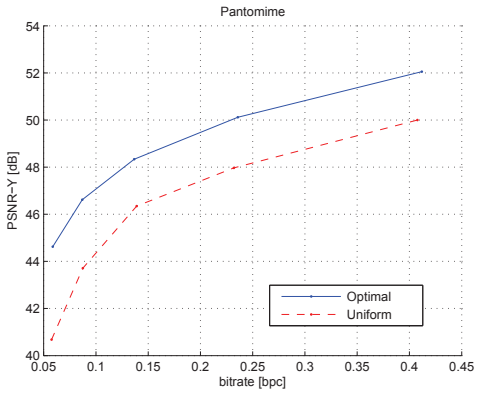

Fig. 7. Overall R-D performance comparison between optimal and uniform rate-allocation for *Pantomime* with Laplacian viewer distribution.

sibly within a 2D array setup) are weighted and optimal rate-allocation is determined for minimum overall observed distortion. In spite of the limited amount of publicly available data sets with general camera arrangements, we provide results showing overall PSNR gains and bit-rate savings for both 2D and 1D camera setups. Significant gains are achieved for sharper viewpoint distributions.

Future work includes the capture and testing of other multiview data in general camera setups. Additionally, we intend to investigate non-linear models for observed distortions as functions of distortions in captured views. The inclusion of lossy depth map compression and its impact within the rate-allocation framework is a challenging issue, which will be considered in future study. Another aspect to be examined is the application of optimal attention-weighted rate-allocation to multiview coding structures beyond simulcast.

5. REFERENCES

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

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

[3] P. Kauff, N. Atzpadin, C. Fehn, M. Muller, O. Scheer, A. Smolic, and R. Tanger, "Depth map creation and image-based rendering for advanced 3DTV services providing interoperability and scalability," *Signal Processing: Image Communication*, vol. 22, no. 2.

[4] Y. Morvan, D. Farin, and P. H. N. de With, "Depth-

image compression based on R-D optimized quadtree decomposition for the transmission of multiview images," in *IEEE Int. Conf. on Image Processing*, September 2007.

[5] K. Y. Kim, G. H. Park, and D. Y. Suh, "Bit-plane-based lossless depth-map coding," *Optical Engineering*, vol. 49, no. 6, 2010.

[6] A. Vetro, S. Yea, and A. Smolic, "Towards a 3D video format for autostereoscopic dispalys," in *Applications of Digital Image Processing XXXI, Proceedings of the SPIE*, vol. 7073.

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

[8] 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*, vol. 7443 (2009).

[9] 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.

[10] 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.

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

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