ENCODING PARAMETER ESTIMATION FOR RDTC OPTIMIZED COMPRESSION AND STREAMING OF IMAGE-BASED SCENE REPRESENTATIONS

Ingo Bauermann and Eckehard Steinbach

Institute of Communication Networks, Media Technology Group,
Technische Universität München, Munich, Germany
{ingo.bauermann, eckehard.steinbach}@tum.de

ABSTRACT

Remote navigation in image-based scene representations requires random access to parts of the compressed reference image data to compose virtual views. The degree of dependencies introduced during compression has an impact on the effort that is required to access reference image data and at the same time delimits the rate-distortion (RD) trade-off that can be achieved. If a limited channel bitrate and computational power of client devices are taken into account, encoding can be performed in a RD optimal manner with respect to the expected maximum transmission data rate (T) and decoding complexity (C). In this work we present a practical framework for parameter estimation for RDTC optimal encoding of image-based scene representations.

Index Terms— RDTC optimization, IBR, compression

1. INTRODUCTION

With recent advances in image and video coding, efficient compression schemes for image-based scenes have emerged (see e.g. [1] for an overview). Additionally, techniques for streaming of such representations have been reported in the literature (e.g. [2]). For video sequences sequential play out of entire frames is dominant and therefore temporal and spatial dependency structures are known at encoding time. On the other hand, interactive navigation in image-based scenes requires random access to individual parts of the reference image data at decoding time. When only limited system resources, like computational power of the receiver device and transmission capacity, are available, traditional rate-distortion optimization is not appropriate anymore. I.e., in a remote navigation scenario with heterogeneous computational capabilities of user devices and different bitrate access there are strict requirements on the decoding time and the operational transmission data rate per virtual view. These constraints are typically not incorporated into RD optimization.

The goal of this work is to develop a practical framework for the compression and interactive streaming of image-based scene representations that allows parameter estimation and rate-distortion optimized encoding given constraints on the available computational resources at the decoder and the available channel capacity.

The remainder of this paper is structured as follows. In Section 2 we give an overview of the considered system and measures used throughout the paper. In Section 3 we introduce the encoding parameter models while Section 4 describes the optimization procedure. Section 5 discusses experimental results. Section 6 concludes the paper.

2. SYSTEM OVERVIEW AND MEASURES

Common image-based rendering (IBR) systems use image sequences that have been captured using calibrated video/still image cameras as the input to image analysis and view synthesis steps. In general, one can interpret spatially distributed camera positions within a static scene as motion trajectories of a single camera capturing a video sequence. Without loss of generality, we assume that the input to the considered system is such a calibrated video sequence. We perform offline compression on group of pictures (GOP) of size N images using motion compensated prediction on BxB pixel blocks. 2D motion compensated prediction of consecutive frames is performed using a scalar displacement Δd (in pixels) along epipolar lines which can be determined from the camera calibration.

Pixel blocks can be encoded in intra, inter or skip mode. Intensity values in intra mode undergo a transform coding step (DCT) and H.263-like quantization. For the inter block mode a residual error after motion compensated prediction is encoded using the intra encoding scheme. For skip blocks only the scalar displacement is encoded. Virtual views are rendered from pixel blocks containing relevant pixel data. The requested block, as well as reference blocks in neighboring frames, are transmitted and decoded. This recursive procedure is continued until all reference blocks can be decoded using the intra mode decoding procedure.

The smallest decodable unit in our system is a single pixel block.

The measures used to evaluate the performance of a specific compression approach with respect to scenario specific properties are summarized as follows [5]:

- The **Rate (*R*)** is the mean number of bits required to store a pixel's RGB values at the server.
- The **Distortion (*D*)** is defined as the MSE between original and reconstructed pixels.
- The **Transmission data rate (*T*)** is the mean number of bits that have to be transmitted per rendered pixel.
- The **Decoding complexity (*C*)** of a given pixel is the mean number of pixels that have to be decoded to reconstruct the current pixel.

T and *C* are strong measures for the user-perceived delay. *T* can be significantly larger than *R* as dependencies might have to be resolved. We identify four encoding parameters that have a major impact on the RDTC system measures [5]:

- The **quantization parameter *q*** (deadzone quantizer).
- The **intra-ratio *α*** which is defined as the ratio of intra encoded blocks in a GOP (except for blocks in the first frame which are all encoded in intra mode).
- The **skip ratio *β*** which is the ratio of skip blocks among all blocks not encoded in intra mode.
- The **single reference ratio *b*** which is signal dependent and is defined as the ratio of non-intra blocks that have one reference block in a neighboring frame. A single reference block refers to the case where the displacement is an integer multiple of the block size *B*. In all other cases the required prediction signal spreads across two blocks in the reference frame. *b* is defined with respect to the displacement field of a GOP as

$$b = \sum_{k=-\infty}^{\infty} p_{\Delta d}\left(k \cdot B\right) \tag{1}$$

Here, $p_{\Delta d}$ is the probability mass function of the scalar motion vector displacement.

3. MODELING RDTC MEASURES

We consider streaming of random virtual views using a sufficiently large pixel cache. To perform RDTC optimization on whole GOPs we present trained models that map the encoding parameters (α,β,*q*,*b*) to RDTC measures. In the following we assume the GOP size *N* and block size *B* to be fixed. *b* is calculated from displacement estimation on original frames and is assumed to be constant and independent from *q*. To train the models we take six sample points which we found to lie at (α,β,*q*) = (0,0,1), (0.4,0,1), (0,0.4,1), (0.6,1,1), (0,1,1), and (1,1,1).

3.1. The rate-distortion (RD) model

We choose an exponential model for the rate and interpolate $R(\alpha,\beta)$ in the α-β space:

$$R(\alpha,\beta) = R_0(0,\beta) + \left(R(1,1) - R_0(0,\beta)\right)\cdot\left(1-(1-\alpha)^{\varepsilon_1\cdot(1-\beta)+\varepsilon_2\cdot\beta}\right) \tag{2}$$

with $$R_0(0,\beta) = R(0,1) + \left(R(0,0) - R(0,1)\right)\cdot(1-\beta^{\varepsilon_3}) \tag{3}$$

Where ε_1, ε_2, and ε_3 are trained from the samples. To extend (2) to be a function of the quantization parameter *q*, we use the *ρ*-domain model introduced in [4]. First we calculate the distribution of transform coefficients at the sample positions. The relationship between the ratio of zeros $\rho(q)$ among the quantized transform coefficients and the mean rate $R(\alpha,\beta,q)$ is expressed by a trained parameter κ in the following equation:

$$R(\alpha,\beta,q) = \kappa\left(R(\alpha,\beta,q_0)\right)\cdot\left(1-\rho(q)\right) + R_0(\alpha,\beta)\,. \tag{4}$$

Here, $R_0(\alpha,\beta)$ is independent from the source and can be determined offline at the sample positions. The relationship between *ρ* and *q* is determined from the discrete probability distribution *f*(*y*) of the transform coefficients *y*:

$$\rho(q) = \sum_{|y|<2q} f(y) \tag{5}$$

Once κ is known, $R(\alpha,\beta,q)$ is determined by first extrapolating the coding samples using (4) and then applying (2).

Again, depending on α and β we choose an exponential model for the distortion $D(\alpha,\beta)$:

$$D(\alpha,\beta) = D(1,1) + \left(D_0(0,\beta) - D(1,1)\right)\cdot e^{-\alpha\cdot(\varepsilon_1\cdot(1-\beta)+\varepsilon_2\cdot\beta)} \tag{6}$$

with $$D_0(0,\beta) = D(0,0) + \left(D(0,1) - D(0,0)\right)\cdot e^{\varepsilon_3\cdot(\beta-1)}\,. \tag{7}$$

ε_1, ε_2, and ε_3 are, again, trained from the samples. Similar to the rate model, we extend (6) to be a function of the quantization parameter *q* by using [4]:

$$D(\alpha,\beta,q) = D_S\cdot e^{-\eta\cdot(1-\rho(q))} \tag{8}$$

Here, D_S is the variance of the source signal. Once η is known, $D(\alpha,\beta,q)$ is determined by extrapolating the coding samples using (8) and then applying (6).

3.2. The decoding complexity model (first view)

Before the first view is transmitted and decoded at the beginning of a streaming session the client cache is empty. For simplicity we assume that the quantization parameter *q* does not have an impact on the decoding complexity. According to the analysis in [3] we use a probabilistic model to approximate the decoding complexity *C* for a virtual view request. The decoding probabilities $a_{m,n}$ for a block at position (*m*,*n*) relative to the requested block can be expressed depending on α, β, and *b*:

$$a_{m,n} = \begin{cases} 0 & \text{if } n > m \\ 1 & \text{if } m,n = 0 \\ 1 & \text{if } m \neq 0, n = 0 \\ a_{m-1,n-1}\cdot(1-\alpha)\cdot(1-b)\cdot\left(1-a_{m-1,n-1}\right) & \\ \quad + a_{m-1,n}\cdot(1-\alpha)\cdot\left(1-a_{m-1,n-1}\right) & \\ \quad + a_{m-1,n-1}\cdot a_{m-1,n}\cdot\left(\left(1-\alpha^2\right)-\alpha\cdot b\cdot(1-\alpha)\right) & \text{else} \end{cases} \tag{9}$$

The mean decoding complexity C_b (*b* is a constant for a GOP) of a single random view access can now be written as:

$$C_b(\alpha,\beta,N)=\frac{1}{N\cdot\gamma}\cdot\sum_{f=0}^{N-1}\sum_{t=0}^{f}\left(a_{f,t}+\left(\sum_{l=0}^{t}a_{t,l}-a_{f,t}\right)\cdot\left(1-\beta\cdot(1-\alpha)\right)\right) \quad (10)$$

Here, γ is the pixel render/decode ratio which is the mean number of requested pixels divided by the number of pixels that are actually decoded per requested block for a single access. γ is a constant for a specific block size and rendering system.

3.3. The transmission data rate model (first view)

Again, assuming an empty cache prior to the view request, the weighted product of the mean rate and the mean number of pixels to be decoded per requested pixel is the mean transmission data rate and can be written as:

$$T_b(\alpha,\beta,q,N)=\frac{R(\alpha,\beta,q)\cdot C_b(\alpha,\beta,N)}{1-(1-\alpha)\cdot\beta} \quad (11)$$

Using skip encoded blocks which do not encode the residual error after motion compensated prediction leads to an imbalance between bits used for intra and inter encoded blocks compared to skip encoded blocks. The denominator in (11) considers this imbalance.

3.4. The T/C models during operation (second views)

The way a user navigates through a virtual environment is mainly characterized by smooth rotation and translation. When the first virtual view is entirely decoded at the client most of its data can be reused for nearby virtual views as many needed reference blocks reside in the client's cache.

Figure 1 illustrates this fact. Blocks marked with a small black square are requested and used for rendering while white blocks have to be transmitted and decoded. Hatched blocks are already in cache when the corresponding view is requested.

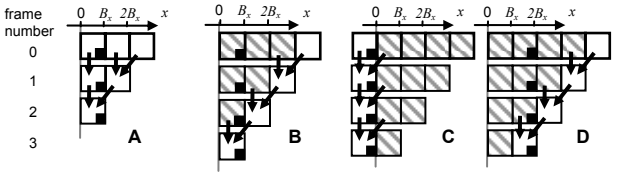

Figure 1: Access patterns for rotational and translational movement in an image-based scene representation (approximation). (A) First view and (B)-(D) second views.

In case (A) the dependency structure for a first virtual view is illustrated. Three blocks are requested and dependencies have to be resolved in exactly the same way as for the case with an empty cache prior request (arrows denote dependencies). When the user rotates (B), then the new virtual view consists of the previously requested blocks and additionally a block in frame three at position x=0 is needed which in turn can only be decoded with three further blocks to be processed. Case (C) and (D) show the same idea for translational motion. Note that, for the second and any further virtual view significantly fewer blocks have to be transmitted and decoded than for the first view.

While the models for the rate R and distortion D remain the same for the first and consecutive views, the models for C_b and T_b have to be changed for smooth navigation. To determine $C_{s,b}$ (subscript s stand for "second") we modify (9) considering only those blocks, which are not in the cache when a second view is requested. This is an approximation of the mean decoding complexity of cases B to D in Figure 1. The probability that a certain block has to be decoded is approximated as:

$$a_{m,n}=\begin{cases}1 & if\ m,n=0\\ 0.5 & if\ m\neq 0,n=0\\ 0.5\cdot a_{m-1,n-1}\cdot(1-\alpha)\cdot(1-b) & if\ m=n\\ 0 & \text{else}\end{cases} \quad (12)$$

Again, $a_{m,n}$ reflects the probability that a block in a certain relative position (m,n) to the requested block is decoded if the cache is filled with reference data of a nearby virtual view. $C_{s,b}$ is determined using (10) and (12). $T_{s,b}$ is calculated according to (11) by replacing C_b with $C_{s,b}$.

4. RDTC OPTIMIZATION

Once the mapping from encoding parameters to RDTC measures is found we perform global numerical optimization. Though the models can be used with a variety of objective functions we choose to minimize the delay for second views (represented by $T_{s,b}$). Additionally we want to guarantee a maximum distortion D_{max} and a maximum initial delay (represented by $T_b<T_{max}$ and $C_b<C_{max}$) while $C_{s,b}$ is unconstrained. This global minimization problem can now be written as:

$$\min T_{s,b} \text{ subject to}$$
$$D\le D_{max},\ R\le R_{max},\ T_b\le T_{max},\ \text{and}\ C_b\le C_{max} \quad (13)$$

The following procedure is used to find optimal encoding parameters and intra/inter/skip blocks:

1. Motion estimation on the original frames is performed and the motion vector field $\boldsymbol{M}$ and b are calculated.
2. The GOP is encoded using $\boldsymbol{M}$ at the six sample positions $\boldsymbol{z}_i=[\alpha_i\ \beta_i\ q_i]^T$ with i=1...6 producing values for R, D, T_b, C_b, $T_{s,b}$, and $C_{s,b}$. According to (α_i,β_i) intra blocks and inter/skip blocks are distributed as follows:
   a. Intra encoding is performed for blocks introducing the biggest residual error after motion compensation.
   b. Inter mode encoding is assigned to the fraction of the *remaining* blocks introducing the biggest residual error after motion compensation.
   c. All other blocks are encoded in skip mode.
3. An optimal parameter set $\boldsymbol{z}_{opt}=[\alpha_{opt}\ \beta_{opt}\ q_{opt}]^T$ is found using numerical (constrained) optimization according to the objective function (13).
4. According to (α_{opt}, β_{opt}) intra blocks and inter/skip blocks are distributed over all blocks as described in step 2.

5. RESULTS

We evaluate the proposed model-based optimization using a densely sampled image-based scene representation [5] assuming limited computational power of the client device and limited available bitrate. We set N=13, γ is determined by the rendering system [5] and is set to 3.2, R and D are given for a whole GOP. Mean values for T_b, $T_{s,b}$, C_b and $C_{s,b}$ are measured and calculated using a sufficiently large number of virtual views.

Figure 2 (left) shows operational rate distortion plots for the first virtual view. The RD optimized and the INTRA encoded rate distortion curves are shown for comparison. The solid lines denote RD curves with respect to four T and C constraints produced by our algorithm. Dots show the corresponding measurements using the optimal parameter set z_{opt}. For $T_b \leq 5bpp$ and $C_b \leq 5ppp$ pure RD optimization can achieve at most a PSNR of 28.5 dB while INTRA encoding achieves a much higher PSNR (introducing a much higher rate of course). RDTC optimization can trade-off these two extreme cases.

Figure 2 (middle) and (right) show the result of an optimization for first and second virtual views using (13). For three different maximum distortion values D_{max} the operational RT and RC plots are shown. For 35dB PSNR the minimum rate R is 0.8bpp while the transmission data rate $T_b < T_{max}$ is as high as 17.5 bpp. This point corresponds to a stream optimized using a rate distortion trade off solely (marked as "RD"). Using this configuration gives a $T_{s,b}$ of 1.7bpp. When increasing the rate R to 1.8 bpp then $T_b < T_{max}$ decreases to 6.5bpp while $T_{s,b}$ increases to 2.3bpp. This configuration corresponds to independent encoding (marked as "INTRA"). For rates between 0.8bpp and 1.8bpp a trade-off between R, T_b and $T_{s,b}$ can be achieved. A similar reasoning can be applied to the decoding complexity.

For a specific streaming scenario these results imply that depending on the available computational resources and channel capacity there is a trade-off between rate and distortion and a trade-off between the initial delay (representated by T_s and C_s) and the delay during smooth navigation (represented by $T_{s,b}$ and $C_{s,b}$).

6. CONCLUSION

In this work we present a practical framework for parameter estimation for RDTC optimal encoding of image-based scene representations. Trained models are given to estimate the rate, distortion, mean transmission data rate, and decoding complexity for the first virtual view of a streaming session as well as for succeeding views. The models allow optimizing offline compression with respect to scenario specific constraints in a remote navigation application using compressed image-based scene representations.

7. REFERENCES

[1] H.-Y. Shum, S.B. Kang, and S.-C. Chan, "Survey of Image-Based Representations and Compression Techniques," in IEEE Transactions on Circuits and Systems for Video Technology, pp. 1020–1037, Volume: 13, Issue: 11, Nov. 2003.

[2] C. Zhang and J. Li, "On the Compression and Streaming of Concentric Mosaic Data for Free Wandering in a Realistic Environment Over the Internet," IEEE Transactions on Multimedia, Vol. 7, No. 6, Dec. 2005.

[3] I. Bauermann and E. Steinbach "Analysis of the decoding complexity of compressed image-based scene representations," Technical Report, Media Technology Group, TU Munich, June 2006. Available online at: http://www.lkn.ei.tum.de\~ingob\pub\AnalysisDecComp.pdf

[4] Z. He, S. Mitra, "A Unified Rate-Distortion Analysis Framework for Transform Coding," IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 12, Dec. 2001.

[5] I. Bauermann, Y. Peng, E. Steinbach, "RDTC Optimized Streaming for Remote Browsing in Image-Based Scene Representations," In Third International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill, USA, June 2006.

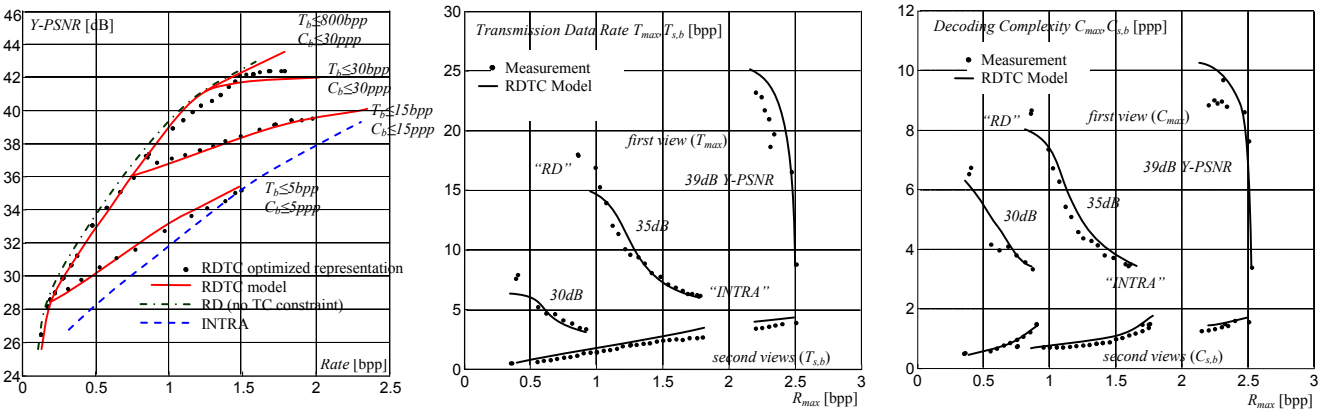

Figure 2: (left) Operational rate distortion curves for *TC* constrained optimization. Solid lines represent the trained model while dots show corresponding measurements using z_{opt}. The dotted line gives the RD performance using conventional RD optimization. The dashed line shows intra only encoding. (middle) and (right) Operational RT and RC plots for RDTC optimization using (13). For different qualities the trade-offs for RT_{max}/RT_s and RC_{max}/RC_s, respectively, are plotted as calculated from the model. Measurements using the corresponding optimized configuration are also shown as dots.