RATE-DISTORTION-OPTIMIZED CONTENT-ADAPTIVE CODING FOR IMMERSIVE NETWORKED EXPERIENCE OF SPORTS EVENTS

Haopeng Li and Markus Flierl

School of Electrical Engineering
KTH Royal Institute of Technology, Stockholm
{haopeng, mflierl}@kth.se

ABSTRACT

This paper presents a content-adaptive coding scheme for immersive networked experience of sports events, in particular, soccer games. We assume that future sports events are captured by an array of fixed high-definition cameras which provide multiview image sequences for a free-viewpoint immersive networked experience in a home environment. We discuss a content-adaptive coding scheme for image sequences that exploits properties of such sequences and that permits efficient user interactions. In this work, we construct a rate distortion model for an image sequence to obtain the optimal bitrate allocation among static and dynamic content items. The optimal bitrate allocation results in a rate distortion performance of the coding scheme that outperforms that of conventional H.264/AVC coding significantly.

***Index Terms*—** Immersive networked experience, content-adaptive coding, rate distortion model.

1. INTRODUCTION

In recent years, contend-based coding techniques have been considered for efficient video coding [1]. These techniques provide more coding flexibility by allowing users to access video objects freely and semantically meaningfully. However, due to the high complexity of object detection and segmentation, the efficiency of most content-based coding schemes cannot be guaranteed. Additionally, shape intra coding consequently increases the coding burden. Hence, object-based coding methods usually underperform in the classical setting at high bitrates.

The rise of high-definition television (HDTV) and the desire for an immersive networked experience in a home environment have raised the interest in content-adaptive coding schemes. Popular applications include soccer games where background objects, like stadium and soccer field, are relatively static and where foreground objects, like players, are mostly dynamic. Similar to conventional sprite coding [2], the dynamic parts will be extracted from the static background and encoded separately. However, there are three important issues with this strategy. First, the computational complexity should be taken into account when extracting dynamic parts. Second, the trade-off between frame rate and rendering distortion of the static content can be optimized. Third, the optimal bitrate allocation between static and dynamic parts can maximize the overall rate distortion performance.

In this paper, we discuss a content-adaptive coding scheme for immersive networked experience of sports events. We exploit the properties of static and dynamic parts and construct a rate distortion model for optimal bitrate allocation. We consider three important aspects, namely the rendering distortion of the static part, the coding distortion of the static part, and the coding distortion of the dynamic parts. With that, we aim to maximize the overall rate distortion performance of the content-adaptive coding scheme.

The remainder of this paper is organized as follows: Section 2 presents our content-adaptive coding scheme. Section 3 discusses a rate distortion model and determines the optimal rate allocation between static and dynamic content items. The experimental evaluation of our scheme and a comparison to H.264/AVC coding are given in Section 4, followed by a short conclusion.

2. CONTENT-ADAPTIVE CODING SCHEME

To facilitate an immersive networked experience of soccer events, we discuss a content-adaptive coding scheme. To match the properties of multiview video captured by an array of static cameras in a soccer stadium, we distinguish between static and dynamic content. The static content, comprising mostly of areas capturing the field and the background, is varying slowly over time and its multiview redundancy can be exploited efficiently by inter-view prediction. On the other hand, the dynamic content, comprising mostly of regions covered by players, is changing rapidly over time and its temporal redundancy can be exploited efficiently by inter-frame prediction.

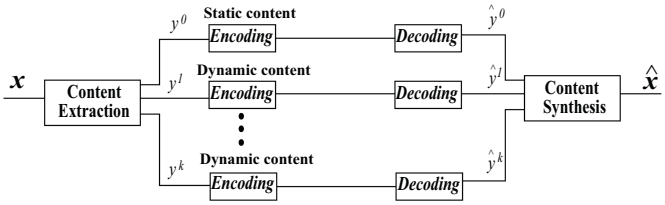

Fig. 1. Content-adaptive coding scheme.

The content-adaptive coding scheme comprises the extraction of static and dynamic content, the individual coding engines, and a synthesis unit at the decoder that facilitates the reconstruction of the output image sequence. The content-adaptive approach will also offer an advantage when considering an user-specified viewpoint in the reconstruction process. In particular, algorithms for view interpolation can be tailored to the specific content. Fig. 1 depicts the discussed content-adaptive coding scheme.

2.1. Extraction of Static and Dynamic Content

We assume that player tracking information [3] is available which allows us to extract a rectangular region for each player. Each region defines a dynamic content item and results in an individual image

sequence. All defined content items are removed from the camera images to establish the static content item. Hence, each original camera image sequence x is divided into several sub-sequences, one sequence for the static content y^0, and multiple sequences for the dynamic content items y^k, with $k = 1, 2, \ldots, K$. An example is shown in Fig. 2.

Fig. 2. Extraction of dynamic parts; the red boxes indicate dynamic content.

Foreground scenes are usually generated by applying the foreground shape macroblock approximation method [4]. However, due to the complexity of foreground object contours, the accuracy of object shape prediction is low. Additionally, the intra coding of shape information will also affect overall coding efficiency. In our work, we use rectangular boxes to capture the dynamic objects and update the position of the boxes with the help of tracking information from the dynamic objects. In other words, the image sequences of dynamic items are generated by the scene content in each box, as depicted in Fig. 2. This approach offers three advantages: First, the rectangular content can be coded efficiently with state-of-the-art standards like H.264/AVC [5]. Second, the extraction of rectangular content can be easily combined with object tracking techniques. Third, the cost of shape information coding is lower for rectangular content.

Fig. 3. Static content; occlusions are compensated.

After generating the dynamic parts, the static scene can be easily subtracted from the original image sequence. The occluded parts can be compensated by traditional temporal median methods [6]. Thus, original soccer sequences can be divided easily into static and dynamic parts with the help of object tracking information. Fig. 3 shows an example for a static part.

2.2. Content Synthesis

After decoding the individual image sequences $\hat{y}^k$ for static and dynamic parts, the reconstructed sequence $\hat{x}$ is synthesized such that dynamic parts overwrite the static part. The position of the dynamic parts on the static content is given by the synchronized tracking information.

3. RATE DISTORTION OPTIMIZATION

An efficient coding scheme allocates optimally the bitrate of each content item. To accomplish this, we introduce a rate distortion model that reflects the trade-off among static and dynamic parts. This allows us to determine the optimal trade-off and the necessary allocation of resources.

3.1. Rate Distortion Model

According to our coding scheme in Fig. 1, let the i-th pixel in the original image x be denoted by x_i, and the corresponding reconstructed pixel by $\hat{x}_i$. Let the set of pixels in the original frame be denoted by $\mathcal{A}$, the set of pixels of the static part by $\mathcal{B}$, and the set of pixels of the k-th dynamic part by $\mathcal{F}_k$. The extracted static part is denoted by y^0, the extracted k-th dynamic part by y^k. The associated decoded parts are labeled $\hat{y}^0$ and $\hat{y}^k$.

For our rate distortion model, we use the mean square error to determine the average reconstruction distortion per pixel.

$$D = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} (x_i - \hat{x}_i)^2 \tag{1}$$

The average reconstruction distortion is determined by two main contributions, namely the average reconstruction distortion of the static part $D^{(b)}$ and the individual average distortions of the dynamic parts D_k.

$$D = \frac{|\mathcal{B}|}{|\mathcal{A}|} D^{(b)} + \sum_{k=1}^{K} \frac{|\mathcal{F}_k|}{|\mathcal{A}|} D_k \tag{2}$$

As extraction and synthesis do not affect the pixel values of the dynamic parts, the individual average distortions are captured by the distortion due to coding.

$$D_k = \frac{1}{|\mathcal{F}_k|} \sum_{i \in \mathcal{F}_k} (y_i^k - \hat{y}_i^k)^2 \quad \text{for} \quad k = 1, 2, \ldots, K \tag{3}$$

However, extraction and synthesis affect the pixel values of the static part as we will reduce the frame rate for the image sequence that represents the static part. Hence, the average reconstruction distortion of the static part will have a rendering contribution $D^{(r)}$ and a coding contribution D_0. We assume that both contributions are uncorrelated such that $D^{(b)} = D^{(r)} + D_0$, where

$$D^{(r)} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (x_i - y_i^0)^2 \tag{4}$$

is the average rendering distortion of the static part, and where

$$D_0 = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} (y_i^0 - \hat{y}_i^0)^2 \tag{5}$$

is the average coding distortion of the static part.

The rate-distortion performance mainly depends on the complexity of the content items. As mentioned earlier, dynamic and

static parts have different properties. Thus, it is reasonable to allocate bitrates depending on the content. For our model, we allocate average bitrates in bits per pixel for the image sequence that represents the static part R_0 and for the image sequences that represent the dynamic parts R_k. Hence, we obtain a distortion rate function for our content-adaptive scheme as follows:

$$D(R_0, R_1, \ldots) = \frac{|\mathcal{B}|}{|\mathcal{A}|}\left[D^{(r)} + D_0(R_0)\right] + \sum_{k=1}^{K} \frac{|\mathcal{F}_k|}{|\mathcal{A}|} D_k(R_k) \tag{6}$$

Note, the average rendering distortion of the static part is determined by the algorithm that we use for extraction and synthesis.

3.2. Optimal Rate Distortion Trade-Off

With this model, we are able to find the optimal rate distortion trade-off for the individual sub-sequences. For that, we assume that the individual distortion rate functions $D_k(R_k)$ are convex. Further, we impose a bandwidth constraint. Let the constant W be the bandwidth which is allocated to the input image sequence. Let f_0 be the frame rate for the static content, and f_k the frame rate for the dynamic content items.

The optimal trade-off is obtained by minimizing the average reconstruction distortion, subject to the imposed bandwidth constraint.

$$\begin{aligned} \min \quad & D(R_0, R_1, \ldots, R_K) \\ s.t. \quad & R_0|\mathcal{B}|f_0 + \sum_{k=1}^{K} R_k|\mathcal{F}_k|f_k \leq W \end{aligned} \tag{7}$$

This constrained problem can be solved by Lagrangian relaxation and leads to the Pareto condition for our content-adaptive coding scheme:

$$\frac{dD_k}{dR_k} = \frac{f_k}{f_0}\frac{dD_0}{dR_0} \quad \text{for} \quad k = 1, 2, \ldots, K \tag{8}$$

Note, the slopes of the individual distortion rate functions need to be adjusted by the ratio between the individual frame rates with which the content items are coded.

3.3. Coding and Rendering of Static Content

For our application, we assume that the cameras are fixed. By definition, our dynamic content items expose significantly more motion activities than our static content item. Thus, it is possible to set the frame rate of the static content lower than that of the dynamic items. In other words, we update the static content less frequently than the dynamic content items.

For the extraction process of the static content, we set the i-th pixel in the extracted frame $y_i^0[p]$ at time p to be the mean value of temporally previous frame pixels $x_i[t]$, such that

$$y_i^0[p] = \frac{f_0}{f} \sum_{t=p-\frac{f}{f_0}+1}^{p} x_i[t] \quad \text{for} \quad i \in \mathcal{B}, \tag{9}$$

where f denotes the frame rate of the input image sequence $x[t]$. Note that there exists a trade-off between the frame rate f_0 of the static part and its rendering distortion $D^{(r)}$. A low frame rate may reduce the average quality of the rendered static content. However, a low frame rate is favorable for tight bandwidth constraints.

4. EXPERIMENTAL RESULTS

We evaluate our content-adaptive coding scheme with the soccer test video set *Barca-St. Andreu* which is provided by the MEDIAPRO group. The videos are captured by fixed broadcast cameras. The resolution of the videos is 1080×1920 at 25 fps. The average Y-PSNR between the reconstructed view and the original camera view will be used to evaluate the performance of the scheme. We use 175 successive frames from the test sequence. H.264/AVC encoding and decoding is accomplished by the x264 implementation [7].

4.1. Content-Dependent Rate Distortion

As we have shown in Fig. 2, an original sequence can be separated into five dynamic sub-sequences (player 1 to 5) and one static sub-sequence. In order to achieve the optimal rate allocation according to (8), we study first the rate distortion performance of the individual content items. The frame size is 1080×1920 for the static sub-sequence, and 220×140 for the dynamic sub-sequences.

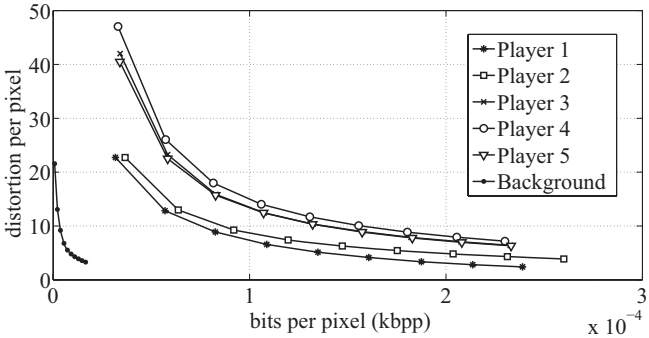

Fig. 4. Rate distortion for each content item.

Fig. 4 depicts the rate distortion performance of each content item. As the static content has significantly less motion, a much smaller bitrate is required to achieve the same average distortion level when compared to the performance of the dynamic content items.

4.2. Rendering of the Static Content

As discussed in Section 3.3, there is a trade-off between the frame rate of the static part and its rendering distortion $D^{(r)}$. The rendering distortion contributes to the overall distortion as an offset and should be chosen appropriately.

Fig. 5 shows the expected rendering distortion as luminance PSNR over the frame rate of the static sub-sequence. As expected, the quality of the rendering decreases for low frames rates.

4.3. Performance of the Content-Based Coding Scheme

To evaluate the overall rate distortion performance of our content-adaptive coding scheme, we encode each sub-sequence with x264. From (8) we know the slopes of the operation points for each sub-sequence and we may use a generalized Lagrange multiplier method [8] to obtain the optimal bitrate allocations. Note that we choose a fixed frame rate of 25 fps for the dynamic sub-sequences while we allow various frame rates for the static content.

Fig. 6 depicts the overall reconstruction quality as luminance PSNR over the total bitrate for our test data. The left plot shows

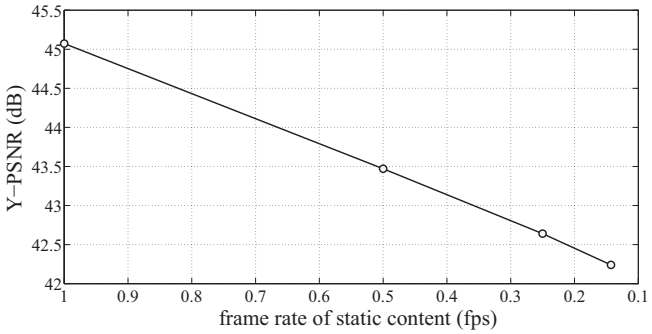

Fig. 5. Frame rate of static content and rendering distortion.

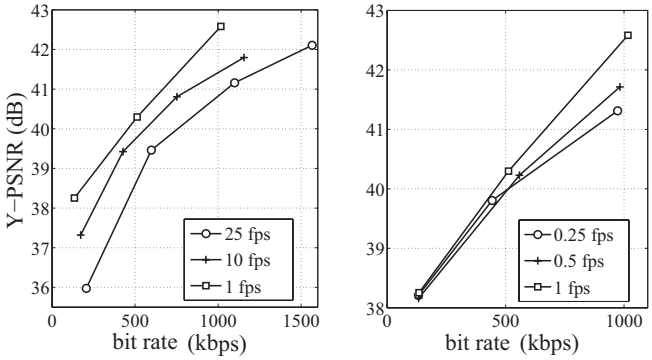

Fig. 6. Comparison of performance for various frame rates of the static content.

that the overall rate distortion efficiency improves when lowering the frame rate of the static content from 25 fps to 1 fps. However, the right plot clarifies that the overall rate distortion efficiency degrades when lowering the frame rate of the static content below 1 fps. The decreasing frame rate of the static content increases the rendering distortion. At a critical frame rate, the overall performance cannot benefit further from a decreasing frame rate.

4.4. Performance Comparison

Finally, we compare our content-adaptive coding scheme to conventional coding with H.264/AVC for our test data set. Our content-adaptive coding scheme uses a frame rate of 1 fps for the subsequence of the static content.

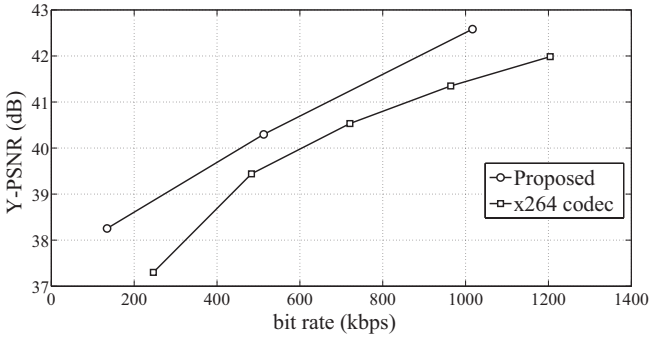

Fig. 7. Performance comparison between our scheme and x264.

As shown in Fig. 7, our content-adaptive coding scheme outperforms x264 encoding. For the same Y-PSNR values, our scheme saves up to 40% bitrate.

5. CONCLUSIONS

We discussed a content-adaptive coding scheme for immersive networked experience of sports events. The content-adaptive coding scheme extracts from an input image sequence several subsequences depending on the static and dynamic content of the input. Further, a rate distortion model is discussed to capture the rendering distortion of the static part, as well as the coding distortions of static and dynamic parts. We obtain a Pareto condition for the optimal bitrate allocation among static and dynamic content parts. The experimental results show that our content-adaptive coding scheme outperforms conventional H.264/AVC coding significantly.

6. ACKNOWLEDGMENTS

This work was supported in part by the European Commission in the context of the project ICT-FP7-248020 “FINE – Free-Viewpoint Immersive Networked Experience”. We thank the MEDIAPRO group for providing multiview video test data.

7. REFERENCES

[1] P. Ndjiki-Nya, T. Hinz, A. Smolic, and T. Wiegand, “A generic and automatic content-based approach for improved H.264/MPEG4-AVC video coding,” in *Proceedings of the IEEE International Conference on Image Processing*, Sept. 2005, pp. II 874–877.

[2] M. Kunter, P. Krey, A. Krutz, and T. Sikora, “Extending H.264/AVC with a background sprite prediction mode,” in *Proceedings of the IEEE International Conference on Image Processing*, Oct. 2008, pp. 2128–2131.

[3] P. Nillius, J. Sullivan, and S. Carlsson, “Multi-target tracking – linking identities using Bayesian network inference,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, New York, NY, June 2006, pp. 2187–2194.

[4] W. Wang, J. Yang, and W. Gao, “Modeling background and segmenting moving objects from compressed video,” vol. 18, pp. 670–681, May 2008.

[5] ITU-T and ISO/IEC Joint Video Team, *ITU-T Rec. H.264 – ISO/IEC 14496-10 AVC : Advanced Video Coding for Generic Audiovisual Services*, 2005.

[6] R. Mech and M. Wollborn, “A noise robust method for segmentation of moving objects in video sequences,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Apr. 1997, pp. 2657–2660.

[7] L. Aimar, L. Merritt, E. Petit, M. Chen, J. Clay, M. Rullgard, R. Czyz, Ch. Heine, A. Izvorski, A. Wright, and J. Garrett-Glaser, “X264 – A free H264/AVC encoder,” http:// www. videolan. org/ developers/ x264.html, Jan. 2011.

[8] H. Everett III, “Generalized Lagrange multiplier method for solving problems of optimum allocation of resources,” *Operations Research*, vol. 11, pp. 399–417, 1963.