IEEE International Conference on Multimedia and Expo, Hannover 23-26 Juni 2008

ENCODING OPTIMIZATION OF LOW RESOLUTION SOCCER VIDEO SEQUENCES

Luca Superiori, Markus Rupp

Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology, Austria
Gusshausstrasse 25/389, A-1040 Vienna, Austria
Email: {lsuper, mrupp}@nt.tuwien.ac.at

ABSTRACT

In this article we propose an optimization method for encoding low resolution soccer video sequences. The considered video codec is the state of the art H.264/AVC, as recommended by the 3GPP specifications.

The proposed optimization aims at maximizing the user perceived quality, considering the characteristic features of soccer videos. In the typical wide angle shots, about one fourth of each frame contains the grandstands. Even if, from the observer's point of view, they can be considered as side information, the performed analysis showed that a prominent fraction of the resulting bandwidth is allocated to them. Our approach consists of a segmentation of the picture intended to recognize three components: the field, the audience and the remaining objects: the ball, the players and the field lines. The resulting map will be then utilized while encoding to optimize the bitrate assigned to each macroblock group, targeting the best resulting subjective perceived quality.

1. INTRODUCTION

The transmission of soccer video sequences, both live matches or highlights, represents one of the most common content for the emerging mobile TV paradigms, such as 3GPP Packet-Switched Streaming Service (PSS) [1] or Digital Video Broadcasting for Handheld (DVB-H). While delivering the video content to the end users, the broadcaster service providers have to face some technological challenges. To match the terminal screen size, the resolution of the video has to be reduced to QVGA (320×240 pixels), CIF (352×288) or even QCIF (176×144).

Moreover, considering wireless transmissions over the Universal Mobile Telecommunication System (UMTS) the available bandwidth for a single cell is limited by technological specification and is shared by all the users. This requires the original video, usually MPEG-2 encoded, to be transcoded using one of the codec supported by the standard [2], as H.263 and H.264/AVC [3]. The application of lossy compression leads to the degradation of the perceived quality.

The standard encoders do not distinguish the different contents of the image and, therefore, apply the same compression to all the elements of the image. This results in different degradation degrees depending on the affected elements. In soccer video sequences this can lead to annoying inconvenience, such as the blurring or even the disappearing of the ball [4]. In this work a content aware encoding mechanism specific for soccer video sequences is proposed.

This paper is organized in five sections. In Section 2 an analysis of the standard encoding mechanism applied to common soccer video sequences is presented. Three groups of image components are defined by means of a segmentation mechanism. The rate associated to each group is examined in detail. The considered video codec is the state of the art H.264/AVC in its baseline profile. In Section 3 a full standard compliant encoding optimization is proposed. Given the results of the previous analysis, each image component is encoded using an appropriate compression degree. This aims at reducing the impact of the compression degradation on the most sensitive part of the frame. Exploiting the Flexible Macroblock Ordering (FMO), a resilience feature standardized in H.264/AVC, Data Partitioning (DP) is obtained as well. The results of the proposed method, evaluated in terms of rate-distortion (both objective and subjective), are discussed in Section 4. The conclusions and potential improvements are presented in Section 5.

2. STANDARD ENCODING ANALYSIS

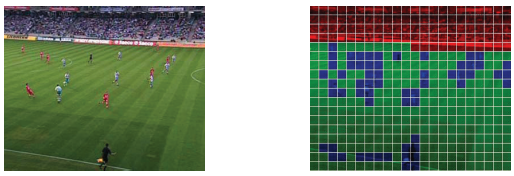
Similarly to his predecessor, the H.264/AVC is an hybrid block based codec. Each video frame is subdivided in blocks of 16×16 pixels, called Macroblocks (MBs). Depending on the frame type, such MBs are then encoded exploiting their spatial correlation with the neighboring ones (I frames) or with the ones in the previously encoded images (P frames). The best prediction (temporal or spatial, respectively) $\hat{\mathbf{k}}$ of the

The authors thank mobilkom austria AG for technical and financial support of this work. The views expressed in this paper are those of the authors and do not necessarily reflect the views within mobilkom austria AG.

original MB **k** is evaluated. The *residual block* **d** is calculated as the elementwise difference $\hat{\mathbf{k}}$ - **k**.

The difference block is then transformed in the block **t** by means of horizontal and vertical modified Discrete Cosine Transformations (DCT). The element $\mathbf{t}(0,0)$ represents the lowest frequency component of the transformed residual block (DC). Higher row and column indexes are assigned to elements associated to increasing frequency components. The block **t** is then scalarly quantized, obtaining the block **q**. The quantization steps are indexed by a Quantization Parameter (QP). Incrementing the value of QP, more components in high frequency are rounded to zero. This results in less elements required for entropy coding but, at the same time, it results in a lack of details in the reconstructed block. Such encoding scheme is applied to all the MBs of the frame.

In the following the possibility of encoding differently the characteristic elements of a soccer frame will be considered. Three different groups of scene components distinguishing their specific features and their impact on the perceived quality. The first group consists of the MBs containing the field. They are characterized by their color tone (green) and the absence of high frequency patterns. The second group comprises the player, the ball and the field lines: the elements the attention of the observer is focused to. Their movement is not consistent with the global camera movement and their shape can vary in time. The third group consists of the grandstands and the advertisements. Basically, they remain as a static background according to the camera movement.

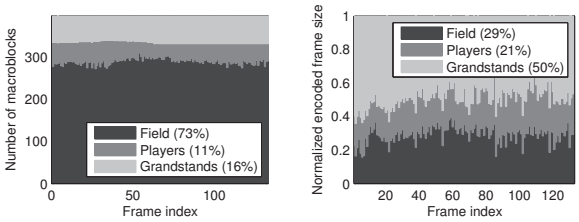


(a) Original Frame (b) Segmented Frame

Fig. 1. Segmentation of a soccer sequence

The three regions were automatically recognized by means of a segmentation mechanism, able to associate each MB to the region it belongs to. In few words, a region growing algorithm was implemented to recognize the grandstands. After placing the seeds on the corner of the frames, the neighboring MBs not containing sufficient green fraction were iteratively merged to the audience. The remaining MBs are the ones containing fields, players, ball. The field has been isolated observing the green quota of each considered MB [1]. The result of the segmentation of a soccer frame can be observed in Fig. 1(b). The MBs associated to the grandstands are highlighted in red, the one containing players, ball and lines in blue, whereas the field MBs are green.

An analysis performed over 20 different soccer sequences in CIF resolution was performed, examining the coding efficiency for the different MB groups. The analysis focuses the temporally predicted (P) frames. This is both because the spatially predicted frames (I) require much more bits than the P frames and because soccer sequences are characterized by strong temporal correlation between consecutive frames. The results for a representative sequence of 134 frames is drawn in Fig. 2. In Fig. 2(a) the distribution of the 396 MBs over



(a) MB distribution (b) Rate distribution

Fig. 2. MB and rate distribution over the three groups

the three groups is depicted. Figure 2(b) shows the resulting code size associated to each group, normalized with respect the total to size of the frame.

As expected, the bitrate associated to the MBs containing the field is in average the smallest, due to the lack of high frequency details. Surprisingly, the MBs containing the audience, representing 16% of the total number of MBs, require 50% of the total bitrate. This behaviour can be justified considering the content of the MBs belonging to the third group. The grandstands, particularly if crowded, are characterized by high frequency components. Even if imperceptibly for the human visual system, such patterns vary in time, resulting in inefficient prediction and, therefore, high frequency transformed residuals. The reduced resolution accentuates this effect.

From the observer's point of view this configuration results to be suboptimal. Most of the bitrate is, in fact, allocated for the MBs containing the least useful information concerning the match. Moreover, the information contained in the grandstands and in the advertisement remains subjectively static in time. Thus, the significant amount of bitrate is mostly associated to details not perceptible by a human viewer.

3. ENCODING OPTIMIZATION

As introduced in Sec. 2 and studied in [6], the selected QP affects strongly the size of the encoded stream as well as the quality of the decoded sequence. In an H.264/AVC encoded stream, the value of the quantization parameter is defined in the Picture Parameter Set (PPS). Usually, all the MBs utilize

[1] The current implementation of the segmentation has been written in Matlab. The algorithm can be easily optimised to be run in a real time environment

the QP specified in the PPS pointed by the frame they belong to. A deviation from that QP can be defined at *slice* level, for a whole collection of MBs, or even at MB level for each single MB, resulting in increasing signalization bits.

Our approach consists on the exploitation of the presented segmentation during the encoding. Traditionally, the MBs are encoded in a raster scan. This strategy results to be inappropriate for the proposed method. We decided therefore to exploit FMO, an error resilience tool comprised in the H.264/AVC baseline profile. FMO allows the encoder to group the MBs in slices, sorted according to some specific patterns (mode 1 to 5) or to an association map given as input (mode 6). We selected this last opportunity for two different reasons. On the one hand, a single deviation from the global QP can be defined for each slice. On the other hand, the different regions can be encoded and packetized separately, obtaining data partitions. If a priority index is associated to each packet, in case of network congestion, the least important packets can be dropped reducing the impact on the perceived quality. The proposed scheme is summarized in Fig. 3.

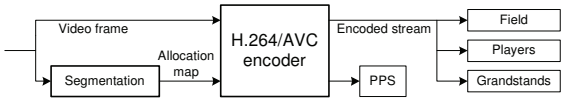


Fig. 3. General scheme of the proposed method

As a first step, the frame to be encoded is segmented. A map, containing the association between each MB and the region it belongs to, is given as input together with the frame to be encoded to the H.264/AVC encoder. The considered codec is the Joint Model (JM) H.264/AVC [5]. Accordingly to the map, the MBs belonging to each region can be separately encoded, using for each group an appropriate quantization parameter, and packetized.

The outputs of the encoder are the encoded bitstream associated to the frame, organized in three different groups of packets, and an additional PPS packet of about 1kb. This contains the allocation map that has to be transmitted to the decoder, together with the encoded stream, to allow the reconstruction of the encoded frame. In the PPS, the region each MB belongs to is sequentially exp-Golomb encoded. The state of the art allocation map encoding penalizes our method. On the one hand the exp-Golomb encoding requires 2,33 bits to encode the three possible symbols. On the other hand both the temporal correlation (within consecutive maps), as well as the spatial one (considering a single map) are not exploited. A run-length encoding of the symbols together with the observation of the slight variation within two consecutive frames, would reduce the amount of bits required for the additional PPS. This would, however, violate the current standard, delivering packets that cannot be interpreted by a standard decoder.

4. SIMULATION SETUP AND RESULTS

In the following, the considered simulation setup as well as the obtained results will be described. Different sets of QPs were assigned to the three defined regions. As a rule of thumb, we considered the information associated to the MBs containing players, ball and lines the most important one. Higher QPs are therefore assigned to the MBs containing the fields and the grandstands. For the player, ball and lines, common QP values between 26 and 30 were used. For the field and the lines, a set of QPs varying from 26 to 42 was used. A training set of sequences was encoded covering all the possible permutation of QPs.

As first analysis, the effect of the different quantization parameters were considered in terms of resulting rate, compared with the results obtained encoding the whole picture with a QP of 26. The results are shown in Fig. 4, fixing the QP of the players to 26. As expected, increasing the QP for

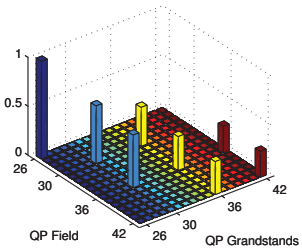


Fig. 4. Normalized rate depending on the QP settings

the field does not provide significant improvements in terms of reduction of resulting bitrate, since the number of coefficients in high frequency transformed residual is limited. On the contrary, the size of the encoded MBs associated to the grandstands can be noticeably adapted modifying the quantization parameter.

Such results were then analyzed in terms of distortion. Figure 5 shows the Peak Signal to Noise Ratio (PSNR) depending on the considered QPs. Surprisingly, the PSNR does

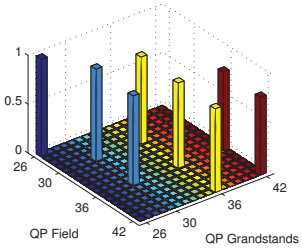


Fig. 5. Normalized PSNR depending on the QP settings

not result to be as sensitive to the QP modification as the rate was. Even for the values (42,26,42), where the rate resulted

to be about the 25% of the original, the PSNR remains about the 80% of the original. As observed for the rate, also the objective distortion metric appears to be marginally dependent on the QP applied to the field. The variations should be therefore only attributed to the effect of the quantization on the grandstands. Since the strong decrease in rate (Fig. 4), and therefore in high frequency residuals, has only a small impact in PSNR (Fig. 5), we conclude that, for an objective metric, the temporal prediction and reconstruction applied to the grandstands does not result effective, even for low QPs.

Even if the prediction at the encoder is performed minimizing an objective metric as PSNR, in this work we target the optimization of the encoding considering the subjective quality perceived by the observer. Objective metrics, as PSNR, does not correlate with the subjective optimization we are performing [7]. Exploiting the results of the previous analysis, a refined set of QPs settings for different sequences has been defined. The field was encoded with moderate QPs, varying between 26 and 30. For the grandstands higher QPs were analyzed, between 30 and 42. In average, the sequences were 135 frames long. The sequences consisted of an I frame at the beginning, encoded using QP 26 for all the MBs group in order to offer an accurate reference for the temporal prediction. All the following frames were P encoded. The frame rate was set to 30 frame per second.

The Mean Opinion Score (MOS) was selected as subjective metric. In order to make test most suitable for a large test audience, the tests were performed on a public web page. The video sequences the test subjects had to evaluate consisted of five different football sequences, encoded using nine different sets of QPs and the uncompressed ones, for a total of 50 sequences. The order of the sequences was randomized. The volunteerswere asked to evaluate the sequences without knowing which were the five uncompressed ones. Moreover, more than the 50% of the test subjects were not involved in the field of signal processing. The evaluation consisted on assigning to each displayed sequence a vote on a scale going from 1 (Bad) to 5 (Excellent). In Fig. 6 are depicted the results of

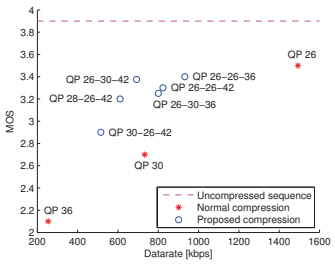


Fig. 6. MOS results

a representative sequence considering different settings of the encoder compared to the uncompressed sequence.

The results collected indicate the effectiveness of the method. The observers, indeed, resulted to be only marginally annoyed by even strong compressions of the grandstands. The viewer resulted to be extremely sensitive to even small increases in the QP used for encoding the field. This can be explained considering the different subjective response to a strong compression applied on the considered region. Even if the reconstruction of the grandstands will not be assisted by the high frequency transformed residual, their predictions still contain high frequency component. Therefore, the error occurs in the range where the human visual system results to be less sensitive [8]. Imperfections in the reconstruction of the field, contrarily, will affect blocks consisting mainly of low frequency components, causing therefore noticeable and annoying blockiness. Moreover, the field surrounds the players and the ball. Being these the objects the attention of the observer is focused on, the user experience results to be furthermore impaired.

5. CONCLUSIONS

In this paper, a novel encoding strategy increasing the perceived user quality for soccer video streaming was proposed. Three groups of scene components were defined: the grandstands, the field and one group comprising the ball, the players. All three groups provide major differences both in terms of effects of compression as well as subjective importance. Such regions were identified by means of an image segmentation mechanism. The three groups of MBs were then separately encoded using different compression degrees. Subjective tests showed that the resulting bitrate can be reduced up to a factor two compared to a standard encoded sequence, reducing the amount of bits associated to the grandstands, affecting only marginally the perceived user quality.

6. REFERENCES

[1] 3GPP TS 26.233, “Transparent End-to-End Packet-switched Streaming Services (PSS), General Description” (Release 6)

[2] 3GPP TS 26.234, “Transparent end-to-end Packet-switched Streaming Service (PSS), Protocols and codecs” (Release 6)

[3] ITU-T Rec. H.264 / ISO/IEC 11496-10, “Advanced Video Coding,” Final Committee Draft, Document JVTE022, Sept. 2002.

[4] M. Wrulich, O. Nemethova, L. Superiori, M. Rupp: “Ball Appearance Improvement in Low-Resolution Soccer Videos”; Elektro- und Informationstechnik, pp. 337 - 345, Oct. 2007.

[5] H.264/AVC Software Coordination, “Joint Model Software,” ver.12.2, available in `http://iphome.hhi.de/suehring/tml/`.

[6] Lszl Czni, Gergely Csszr, Attila Licsr, “Estimating the Optimal Quantization Parameter in H.264”, 18th Int. Conf. on Pattern Recognition, Hong Kong, Aug. 2006.

[7] O Nemethova, M Ries, E Siffel, M Rupp, “Quality Assessment for H. 264 Coded Low-Rate and low-Resolution Video Sequences” in Proc. of Conf. on Internet and Inf. Techn., 2004.

[8] ”Quantization Error and Dithering,” IEEE Computer Graphics and Applications, vol. 14, no. 4, pp. 78-82, Jul/Aug, 1994