Optimization of Audience Encoding in Low-Resolution Soccer Video Sequences

Luca Superiori, Alfredo Font Perez and Markus Rupp
Institute of Communications and Radio-Frequency Engineering
Vienna University of Technology
Gusshausstrasse 25/389, A-1040 Vienna, Austria
Email: {lsuper, afont, mrupp}@nt.tuwien.ac.at

***Abstract*—Soccer is one of the most streamed video contents over the European mobile networks. However, technical limitations, such as stringent bandwidth constraints and reduced size of the mobile terminal displays, make the delivery of soccer videos with satisfactory quality a challenging task. In this work we propose the optimization of the encoding process by minimizing the bandwidth associated to the audience while preserving acceptable quality for the viewer. The macroblocks belonging to the audience are identified by means of a segmentation algorithm. For these macroblocks, the camera movements will be compensated using a single motion vector. The optimized sequence can be decoded by a baseline profile H.264/AVC decoder.**

I. Introduction

Mobile video streaming is one of the emerging application offered by the latest wireless transmission standards such as DVB-H (Digital Video Broadcast for Handheld), UMTS (Universal Mobile Transmission System) as well as its successor LTE (Long Term Evolution). Addressing a nomadic usage, the content of mobile video streaming mainly covers entertainment, such as sport, in particular football, news and music video clips.

The delivery of video sequences with satisfactory quality represents a challenging task. Mobile video streaming copes with two main technical limitations: the reduced screen resolution of the mobile terminals and the limited data rate available in the wireless link. This calls for spatial (resolution) and temporal (frame rate) downsampling of the original sequence. Before transmission, the available data rate is matched by encoding the video sequence using standard video codecs such as H.263 or H.264/AVC. Small objects in the original video are, at the user side, barely visible due to the low pass filtering performed by the mentioned processing. In soccer videos, the ball is one of the key elements necessary to follow the game. At the user side, however, the visibility of the ball in wide angle shots is unsatisfactory, as discussed in [1].

In this work, we focus on the optimization of the video encoding at the server side, using the H.264/AVC video codec [2] in its baseline profile, as recommended by the 3GPP specifications [3]. The standard video encoder applies the same grade of compression to each element of the frame, not being designed to distinguish between their subjective importance. However, in a soccer frame it is possible to distinguish elements of greater importance, such as the players of the ball, and other elements of lower significance, such as the audience and the field.

In [4] an H.264/AVC video encoder mixed with object-based encoding was proposed. However, this requires modifications at the decoder that would violate the standard. In [5] the application of segmentation for identifying a Region of Interest (ROI) was proposed. Such an ROI is then cropped and transmitted to the users. However, in this work we address the optimization of the frame without modifying its content. In [6] it has been proposed to detect the ball (as in [7]) and enhance its visibility of the ball replacing it with an artificial template. This method implies a content modification that, for commercial purposes, might be forbidden.

The macroblocks belonging to the audience are encoded with high quality in the reference pictures, usually Inter (I) encoded. In the following frames, the temporal correlation between consecutive pictures is exploited by means of Inter (P) encoding. Besides motion compensation, the differences between the current picture and the previous one, the *residuals*, are signalised. In soccer sequences, the macroblocks containing the audience consists mostly of high resolution patterns. Although differences between consecutive frames are barely detectable by the human viewer, they require a considerable amount of bits, compared to the macroblocks containing the ball or the players. In a previous work [8], we proposed the application of different compression rates to the different regions of the frame, identified as (1) field, (2) players and ball and (3) audience. An analysis performed over typical soccer sequences showed that almost half of the bitrate was used to transmit the encoded macroblock belonging to the audience. A higher Quantization Parameter (QP) was therefore used when encoding the audience. This approach exploits the property of the audience to be an element of the sequence where the attention of the viewer is not focused on.

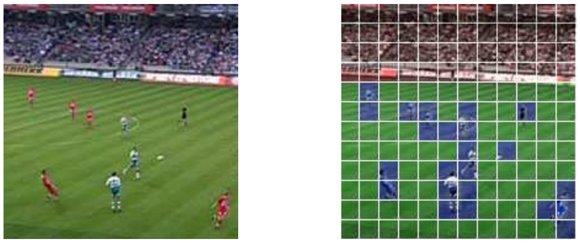
In the present article, we also consider the audience as a static background element. In low resolution videos, besides special instances such as the celebration of a goal, this can be considered true. After segmenting each frame in the three regions before mentioned, we compensate the camera movement applying a single *global motion vector* to the audience. Specific issues, such as new elements appearing at the border of the image as well as the presence of zoom, are handled by means of trace analyses.

This article is structured as follows: Section II briefly recalls the segmentation mechanism proposed in [8]. Section III describes the proposed encoding mechanism. In Section IV objective results in terms of bit rate saving and subjective results in terms of perceived quality are discussed. In Section V the conclusions will be drawn.

II. Segmentation of the Soccer Frames

In our proposal, each frame of a soccer video has been subdivided in three main regions. The field is labeled as Region 1 (R1), the ball, the players and the field lines as Region 2 (R2) and the audience as Region 3 (R3). In [8], an unsupervised segmentation mechanism was designed in order to associate each MacroBlock (MB) to the region it belongs to.

The segmentation has been performed considering the different color characteristics of the macroblocks. A region growing algorithms, with seeds placed on the corner of the picture, has been implemented in order to recognize the macroblocks belonging to the audience. New macroblocks are included in the R3 if they border R3 and their green quota do not exceed a given threshold. Once determined R3, the green quota of the remaining elements is analyzed in order to separate elements belonging to R1 from elements belonging to R2. The result is shown in Figure 1. Exploiting the property of the field pixels to have a green color tone, each picture was transformed in its HSV color components.



(a) Original frame

(b) Segmented frame

Fig. 1. Segmentation algorithm

The segmentation algorithm returns a map containing the association between MBs and the three regions. H.264/AVC defines an error resilience feature called Flexible Macroblock Ordering (FMO). It allows for customizing the order in which the macroblocks are encoded (alternative to the classical raster scan) as well to which *picture slice* each MB belongs to. Since MBs belonging to the different slices are stored in different packets, Data Partitioning is enabled.

A different QP has been defined for each of the three slices. The lower the QP, the higher is the resulting quality as well as the amount data associated to the encoded picture. The MBs belonging to R3 are the more expensive in terms of associated data. The differences within two frames are located in the high frequency components, reducing the compression capabilities of the encoder. By increasing the QP of R3, most of the encoded high frequency residuals are cut. Although the majority of the MBs belong to the field, it has been measured that the encoding of the whole R1 requires a marginal amount of bits, if compared with R2 and R3. The application of an higher QP would therefore not be significantly beneficial in terms of rate, but would result in annoying blockiness.

Mean Opinion Score (MOS) tests confirmed the effectiveness of the method.

III. Encoding Mechanism

The present work has to be considered as an improvement of the method summarized in Sec II. In this Section a proposal for further reducing the data rate associated to the audience is discussed. The strong correlation of the audience region within two consecutive frames will be exploited. Whereas the movement of the player is hardly predictable, the elements containing the audience move coherently with the camera movement. The latter will be compensated by means of a single *Global Motion Vector* (GMV).

The proposed method consists of a two pass encoding of the video sequence. As a first step, the sequence is encoded using the standard development codec Joint Model (JM) [9] H.264/AVC in its baseline profile. A single low QP is used for the whole frame. Information regarding the horizontal and vertical motion vectors associated to each MB is collected and stored during the encoding. The movement data associated to the audience is extracted and processed to find, possibly, a characteristic component, in the following the *global motion vector*, as indicated in Fig. 3.

The GMV is used in the second pass encoding, performed by a modified JM encoder. While the macroblocks belonging to R1 and R2 are encoded individually as described in Sec II, we apply a single motion compensation to the whole R3. The residuals result from differences between the original block and the predicted one. The H.264/AVC calculates the difference in a discrete cosine transformed plane. The macroblocks belonging to the audience mainly consist of high frequency components, their residuals are inefficiently encoded, resulting in low coding rate. We propose to apply motion compensation without transmitting residuals.

The encoding mechanism is modified in such a way, that during the encoding of a macroblock belonging to R3, its associated MVs are forced to be equal to the GMV. Although they might be present, we also force the encoder not to transmit any residual. In H.264/AVC only the difference between the original MV and the predicted value is transmitted. The macroblocks having MVs equal to zero and no residuals are *skipped*, i.e. no data at all is sent.

The scheme in Fig. 2 describes the proposed approach. Besides the two mentioned encoders, a processing block consisting of three algorithms has been developed. In the following subsection, their functionalities are discussed in details.

A. Global Motion Vector Estimation

As previously discussed, the distribution of the MVs associated to the audience is analyzed, in order to extract a characteristic component, the *global motion vector*. Fig. 3

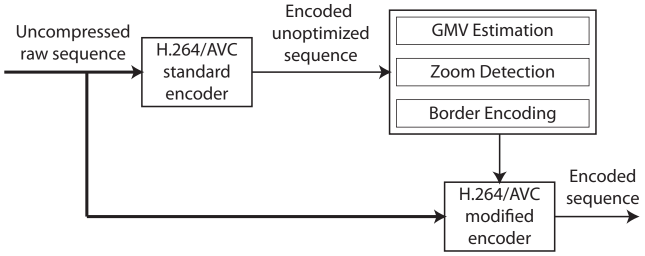

Uncompressed raw sequence
H.264/AVC standard encoder
Encoded unoptimized sequence
GMV Estimation
Zoom Detection
Border Encoding
H.264/AVC modified encoder
Encoded sequence

Fig. 2. Proposed encoding scheme

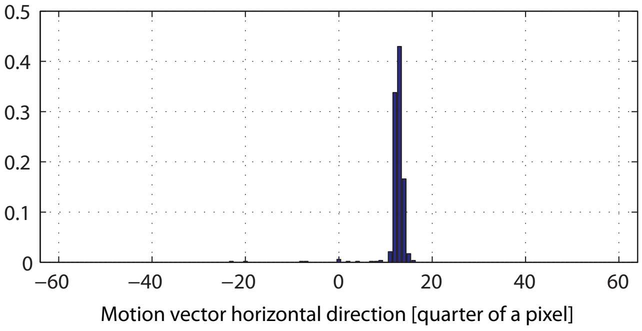


Fig. 3. Distribution of the horizontal MVs

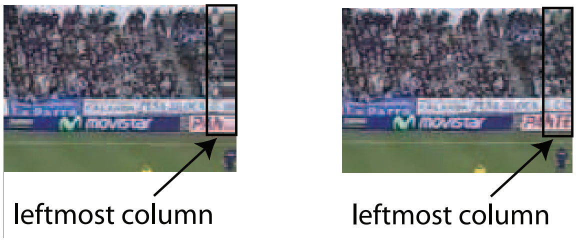
shows the histogram of the horizontal component of the MVs of a single frame. The magnitude of the motion vector has been limited by 16 pixels. Having the MVs in H.264/AVC the resolution of one quarter of a pixel, they are spread in the range [-64,64].

Not integer values of the motion vector cause the reconstruction of pixels by means of interpolation of the neighbouring ones. In contrast to the standard encoding mechanism where the residuals are then summed up at the receiver, our approach does not rely on residuals. Fractional values of the MVs cause, therefore, the reconstructed picture to become blurry due to the performed interpolation.

A motion vector buffer has been therefore considered. The motion compensation, by means of the global motion vector, has been applied for integer multiples of a pixel.

B. Macroblocks at the border of the frame

The proposed approach exploits the visual information already contained in the previously encoded frames. However, vertical as well as horizontal camera movements make new elements appear in the border of the picture. The global motion vector applied in the MBs at the borders of the picture, may point to pixels outside the frames. H.264/AVC support this, reconstructing such MBs using a weighted interpolation of the border pixels. However, this results in blurriness, as shown in the most right-hand MB column in Fig. 4(a). It has been therefore decided to handle differently the MBs located at the border pointed by the global motion vector. Without valid reference in the previous pictures, their reconstruction necessitates the use of residuals. Whenever the movement applied to the audience overcomes half of a macroblock (eight pixels), the macroblocks on the proper border are encoded as standard P macroblocks. In Figure 4(b), the macroblocks on the right hand column are encoded as regular P macroblocks.



(a) Frame 101

(b) Frame 102

Fig. 4. Recovery of the right border

C. Zoom detector

Being the macroblocks belonging to the audience translated according to the direction pointed by the global motion vector, the performance of the proposed method strongly depends on the quality of the reference pictures.

In case the sequence includes zoom, the previous frames do not offer an appropriate reference, if no residual correction is considered. The motion compensation relies on the assumption that the elements of the picture only translate within consecutive frames. In case of zoom, such elements also vary in size, making the motion compensation alone inefficient.

It has been therefore decided to develop a zoom detector. Whenever the zoom is detected, the picture is encoded as a normal P frame. In order to limit the computational overhead, the zoom detector is based on the already available data extracted from the first pass encoding. Besides other factors, such as the number of residuals sent, the presence of zoom influences the distribution of the motion vectors.

Figure 5, upper row, shows the distribution of the motion vectors in a frame where zoom was not present. More than 95% of the motion vectors were bounded within two quarter of a pixel. The remaining MVs are pointing incoherent directions and are associated to the macroblocks at the border of the picture. Figure 5, lower row, refers to a picture with zoom. In order to collect 85% of the MVs, a window of 11 quarter of a pixel has to be considered. In other words, in case of zoom, the MVs are spread over a wider range of values. It has been therefore decided to discriminate whether a zoom has been detected or not, basing the decision on the observation of the size of the range where 90% of the MVs are located.

IV. Results

The proposed method aims at the reduction of the required data rate while keeping the subjective quality unaltered. The performance will be, therefore, discussed with respect to the associated code size as well as with respect to the collected Mean Opinion Score (MOS) results.

The sequence used as a reference are shots taken from a match of the Spanish first division. The original sequence, in MPEG2 DVD format, was deinterlaced and downsampled in

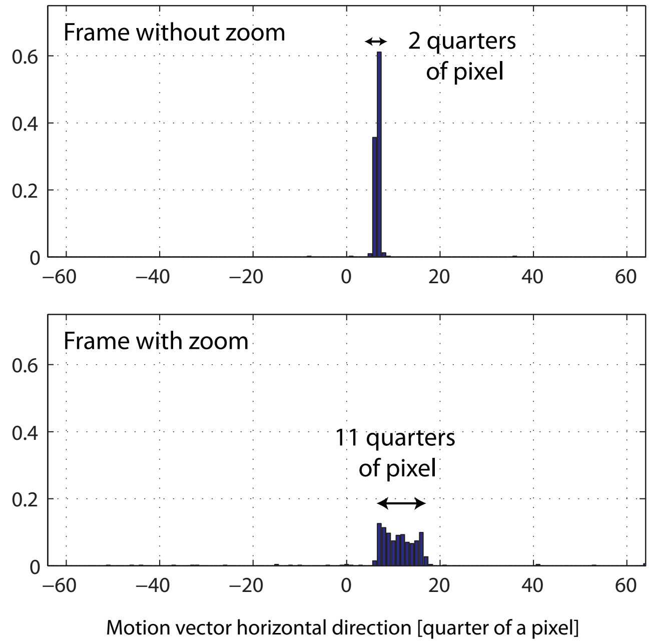


Fig. 5. Distribution of the MVs

Common Intermediate Format (CIF) format, equal to 352×288 pixels. Each sequence comprises a single inter encoded picture at the beginning. The remaining frames have been encoded as intra pictures.

The chart in Fig. 6 shows the code size associated to a test sequence encoded using different coding strategies. The label used in the following are explained in Table I. The leftmost bar represents the resulting size when using the standard H.264/AVC encoder. The whole frame has been encoded using a single QP equal to 26. It has been considered as reference and therefore normalized to one. Although flexible macroblock ordering was not used when encoding the sequence, the size associated to each region has been calculated. As already discussed in [8], more than 60% of the data is used for encoding the macroblocks belonging to the audience.

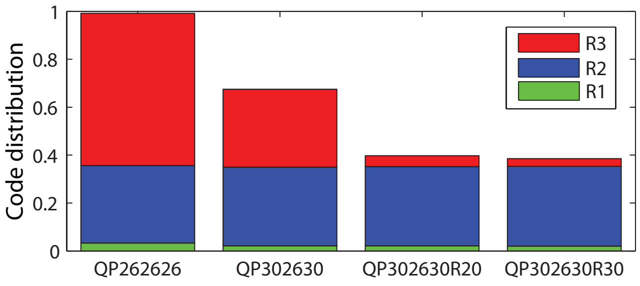


Fig. 6. Code distribution for different settings

Label	QP	QP	QP	FMO	GMV
	R1	R2	R3	used	used
QP262626	26	26	26	NO	NO
QP302630	30	26	30	YES	NO
QP302630R20	30	26	30	YES	YES (20)
QP302630R30	30	26	30	YES	YES (30)

TABLE I
SIMULATION SETTINGS

The bar QP302630 represents the results obtained using the encoding mechanism described in [8]. A higher quantization parameter has been used for encoding the field and the audience compared to the one used for the player and the ball. This reflects on a noteworthy reduction of the data associated to the R3, being 51% of the original. Since the field consists of low frequency components, the reduction for R1 is much lower. Using the same QP settings, the bars QP302630R20 and QP302630R30 describe the results when using the proposed method. Due to perspective distortion and reference picture degradation, the refreshment of the sequence has been applied each 20 and 30 frames, respectively. Considering the latter case, the proposed method requires for the encoding of the audience 6.17% of the data required by the original encoder.

A detailed description of the behaviour of the proposed method QP302630R30 compared to QP302630 is drawn in Fig. 7. The two graphs show, for each frame of the sequence, the size associated to each region. Since the proposed method affects only the encoding of the audience, R1 and R2 require the same amount of data. The size of R3 differs considerably. The behaviour of the graph associated to the audience encoded with the proposed method deserve further discussion. Most of the frames are encoded storing solely the value of the global motion vector. Almost all the macroblocks are skipped. Small peaks can be recognized frequently. They are associated to frames whose border macroblocks are encoded as normal P macroblocks. Large peaks are caused by refresh pictures inserted at regular intervals. The size of the refreshment pictures is slightly bigger than the size of the same frame encoded without the proposed method. This is caused by the

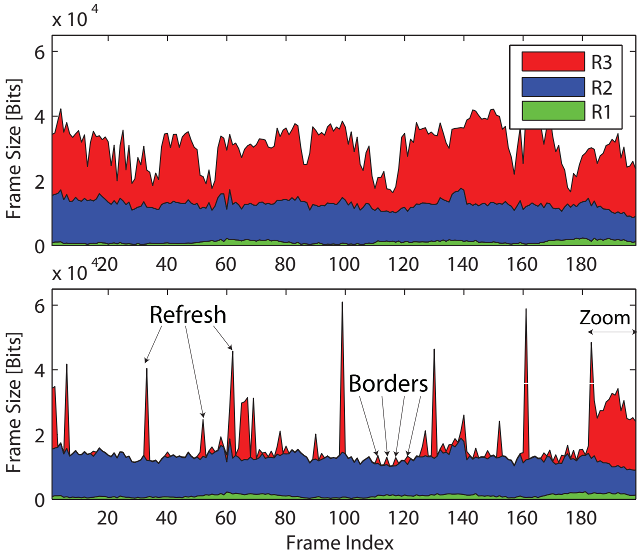
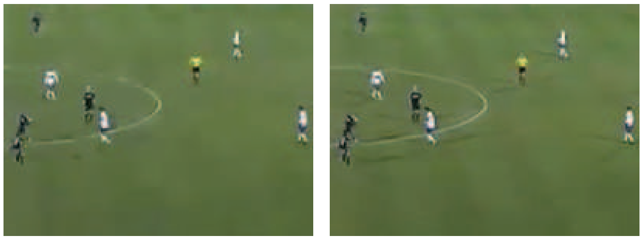


Fig. 7. Frame variant code size distribution

considered reference picture that, being the result of successive picture translations, represents a worst source of prediction.

In this work we address the transmission of soccer video sequences over UMTS networks. The available net datarate, excluding the IP (Internet Protocol), RTP (Real Time Protocol) and UDP (User Datagram Protocol), for the encoded video is usually bounded to 220 kbps. Preliminary test with a limited number were performed to determine the optimal QP sets for low bitrate. Surprisingly, it has been noticed that with such a bitrate the quantization parameter of the field has to be kept smaller than the QPs of the R2 and R3. As mentioned before, the macroblocks containing the fields consist mainly of low frequency components, since they usually represent a flat green area. Possible high frequency components are responsible for the transition between different tones of green, due to, for instance, shadows or artificial lights. Cutting these high frequency components cause the reconstructed picture to be affected by blockiness, as shown in Fig. 8. Even though the QP of the field is significantly smaller than other two, for the above mentioned reasons, this does not result in considerable increase of the associated rate.



(a) Standard encoding

(b) Proposed mechanism

Fig. 8. Quality comparison

In order to subjectively evaluate the perceived quality of the video, a web page was realized. 22 volunteers were asked to rate a set of sequences with grades varying from one (bad) to five (excellent). Since this process is time consuming, the number of sequences has been reduced to 12. Two shots from the original football sequences were encoded using three different QP sets. For each resulting bitrate, the sequences were encoded using the standard encoder with Constant Bit Rate (CBR) enabled.

The results of the investigation is plotted in Fig. 9. The score assigned to the optimized encoded (OE) sequences clearly overcomes the standard encoded (SE) ones. In terms of average MOS, we measured an increase of over 1.24 points. We also discriminated two main groups of test persons. The first one assigned almost the same mark (high or low) to all the sequences. Quite all the outliers were produced by such users. For the second one, a high variance in the marks has been noticed. We assume that the users belonging to this group are the one more accustomed with low resolution football streaming and, therefore, already knew the quality to be expected.

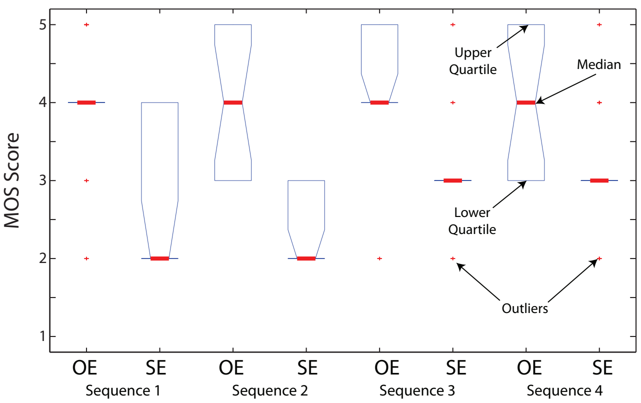


Fig. 9. MOS tests: Standard Encoding (SE) and Optimized Encoding (OE)

V. Conclusion

This paper proposes an encoding mechanism optimized for low resolution soccer video streaming over wireless channel with limited capacity. The method relies on a segmentation algorithm designed to group macroblocks in regions: (1) field, (2) player/ball and (3) audience. An appropriate quantization parameter is associated to each region. Instead of applying the original intra prediction to the audience macroblocks, they are reconstructed using a single motion vector for the whole region without transmitting residuals. Subjective tests confirmed the effectiveness of the method.

Acknowledgment

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. The authors also thank the volunteers for spending their valuable time evaluating the sequences.

References

[1] O. Nemethova, M. Zahumensky, M. Rupp: “Preprocessing of a Ball Game Video Sequences for Robust Transmission over Mobile Network,” Proc. of the 9th CDMA Int. Conf. (CIC), Seoul, Korea, Oct. 2004.

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

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

[4] A. Krutz, M. Kunter, M. Drose, M. Frater, and T. Sikora, “Content-Adaptive Video Coding Combining Object-based Coding and H.264/AVC,” Picture Coding Symposium 2007, Lisbon, Nov. 2007.

[5] K. Seo, J. Ko, I. Ahn, C. Kim, “An Intelligent Display Scheme of Soccer Video on Mobile Devices,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 17-10, Oct. 2007.

[6] M. Wrulich, L. Superiori, O. Nemethova, M. Rupp: “A Robust Preprocessing Algorithm for Low-Resolution Soccer Videos,” ACM Multimedia 2007, Augsburg, Sep 2007.

[7] L. Superiori, O. Nemethova, M. Rupp: “Clustering-based Object Detection for Low-resolution Video Streaming,”; IEEE Symp. on Broadband Multimedia Systems and Broadcasting, Orlando, USA, Mar. 2007.

[8] L. Superiori and M. Rupp, “Encoding Optimization of Low Resolution Soccer Video Sequences,” International Conference on Multimedia and Expo 2008, Hannover, Jun. 2008.

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