RATE-DISTORTION OPTIMIZED H.264/MVC VIDEO COMMUNICATIONS OVER QOS-ENABLED NETWORKS

Attilio Fiandrotti, Enrico Masala, Juan Carlos De Martin

Computer and Control Engineering Department
Politecnico di Torino – Turin, Italy
[attilio.fiandrotti|masala|demartin]@polito.it

ABSTRACT

This paper presents a rate-distortion optimized framework for the transmission of stereoscopic H.264/MVC video on QoS-enabled networks. The distortion caused by the loss of each encoding unit is computed by means of the analysis-by-synthesis approach, which is then used within a rate-distortion optimization framework to select the best transmission strategy. Simulation results, evaluated using an objective quality measure accounting for the peculiarities of the human visual system in stereoscopic video quality perception, show that the proposed rate-distortion optimized strategy significantly improves the communication quality with respect to a reference a priori optimization strategy.

1. INTRODUCTION

The interest for devices that promise to improve user experience is rapidly increasing. For instance, devices with 3D visualization capabilities are expected to augment user experience by adding depth perception. Various designs for 3D-enabled systems, with different features, have been proposed. In particular, some of them are based on capturing more than one picture of the same subject, from different point of views. Multiple views may allow to achieve a so-called free-viewpoint system, in which the user can select one of the available views or a view at an intermediate point between the cameras can be reconstructed by interpolation. Such systems usually involve a high number of views, thus a large amount of data need to be processed. However, it is clear that high correlation exists between such views that could be effectively exploited to, e.g., improve data compression efficiency.

The new H.264/Multiview Video Codec (MVC) [1], currently under standardization, builds on and substantially extends the previous H.264/Scalable Video Codec (SVC) by exploiting the redundancy between different views to improve the compression efficiency. It adopts the so called "disparity compensation" approach to reduce the amount of data using differential encoding between the views, similarly to the traditional motion compensation approach.

A particular case of multiview video is stereoscopic video. Its importance stems from the fact that binocular disparity is a sufficient, though not necessary, cue allowing to perceive depth in a scene. Therefore, many research works aimed at adding 3D perception to video mainly focused on stereoscopy instead of multiview, due to the simplicity of managing two views only.

Some aspects of stereoscopic video have begun to receive great attention recently, such as objective quality evaluation and encoding optimization. The former is, of course, the basis for all works dealing with stereoscopic video. Despite the research in this field is still in its infancy, some trends are beginning to emerge. While the best quality metric are subjective experiments, a few works proposed objective quality measures which account for some human visual system peculiarities. For instance, a PSNR-based metric has been proposed in [2], in which the asymmetric perception of the human visual system is taken into account by a weighted mean of the quality of the left and right views. Moreover, a corrective factor is inserted in case of temporal scaling of one of the views. Others, instead, showed that, to some degree, existing objective quality metrics for monoscopic video can be used, at least as a guidance, to evaluate the quality of stereoscopic video [3]. Another work [4] showed that, differently from what expected, in case one of the two views has reduced quality, the perception is mostly influenced by the characteristics of the high quality view.

Even though designing objective quality evaluation and encoding optimization methods are important research challenges for stereoscopic video to succeed, communications issues are also expected to be fundamental in future stereoscopic video systems. However, until now, very few efforts have been devoted to investigate how to efficiently transmit stereoscopic video over packet networks. In [5] a content-adaptive stereo video coding scheme is proposed, together with adaptation on the client side where each client can view the video in mono or stereo mode depending on its display capabilities.

In this work a rate-distortion optimized scheme for video communication is proposed. The scheme leverages on the analysis-by-synthesis method we developed to accurately estimate the distortion of each video unit in case of loss. Such values are used in a rate-distortion optimized framework for transmission over a generic QoS-enabled network. The proposed framework is general and can be easily adapted to different protection strategies, such as Forward Error Correction (FEC), Automatic Repeat reQuest (ARQ), and network-provided QoS levels as in the case of DiffServ or 802.11e networks. Results are evaluated using the objective distortion measure presented in [2]. Comparisons with a reference prioritization scheme relying on "a priori" considerations on dependencies between video elements are presented, showing the improvements of the rate-distortion optimized approach. Finally, a protection algorithm that randomly chooses the protection level for each video unit is evaluated to provide a lower bound of the performance, for comparison purposes.

The paper is organized as follows. Section 2 briefly reviews the emerging MVC standard, whereas Section 3 is devoted to the detailed description of the analysis-by-synthesis method which is the basis for the rate-distortion optimized stereoscopic video communication system presented in Section 4. Simulation setup and results are described in Section 5 and 6, respectively. Finally, conclusions are drawn in Section 7.

This paper describes results of authors' work carried out within the framework of a project funded by Telecom Italia S.p.A., who reserve all its intellectual property rights thereto.

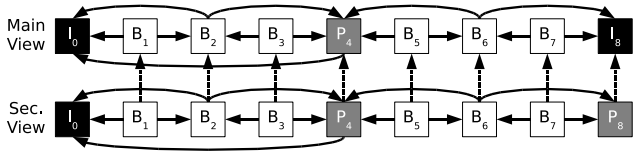

Fig. 1. A typical H.264/MVC encoding scheme for stereoscopic video coding.

2. THE H.264/MVC ENCODING STANDARD

The H.264/MVC video codec is the result of the efforts of the Joint Video Team of ISO-MPEG and ITU-VCEG to amend the H.264/AVC standard with functionalities which enable efficient encoding of multiple view video sequences. H.264/MVC achieves significant compression gains with respect to the case where each view is independently encoded, while it maintains backward compatibility with legacy H.264/AVC decoders which can discard information about the supplementary views and decode the main view only. H.264/MVC inherits from the H.264/AVC the decomposition of the encoder functionalities into a Video Coding Layer and a Network Abstraction Layer. The former layer encompasses all the encoder core functionalities (e.g., macroblock coding), while the latter is responsible for the encapsulation of each encoded data unit (pictures, auxiliary information, etc.) into independently decodable transport units known as NAL Units (NALUs). An IETF standard draft [6] defines the mechanisms to encapsulate H.264/MVC NALUs in RTP packets for real time video communication.

The H.264/MVC allows a high degree of freedom in choosing the encoding scheme which fits a given application. In more details, Figure 1 illustrates a typical Group Of Pictures (GOP) structure used in the stereoscopic video encoding case. Each box represents a NALU and each NALU contains a coded picture. The letter inside the box indicates the picture type (Intra, Predictively or Bipredictively coded) and the subscript number the picture display order (POC).

The encoding dependencies in Figure 1 suggest that the length of the dependency path could be used as a coarse measure of the importance of each picture. In fact, the loss of pictures used for the prediction of several others may cause a distortion which propagates through the prediction path and affects several other pictures. In the following, the notation $T_{V,P}$ will be used to indicated a picture of type T which belongs to the view V and whose picture order count is equal to P. Actual values of T can be I (I-type picture), P (P-type) or B (B-type), while actual values of V can be M (Main view) or S (Secondary view). For example, the loss of picture $I_{M,0}$ generates distortion which affects the whole GOP, while the loss of $P_{S,4}$ causes distortion only in the secondary view and the loss of $B_{S,1}$ does not propagate distortion to other pictures.

3. THE ANALYSIS-BY-SYNTHESIS DISTORTION ESTIMATION METHOD

The quality of multimedia communications over packet networks is affected by packet losses. The amount of resulting quality degradation strongly differs depending on the perceptual importance of the lost data. In order to design efficient loss protection mechanisms, a reliable importance estimation method for multimedia data is needed. Such importance may be defined a priori, based on the average importance of the elements as with the data partitioning approach, e.g., motion vectors are more important than residual coefficients. In order to provide a quantitative importance estimation method at a finer level of granularity, the importance of a video coding element, such as a macroblock, a slice or a frame, could be defined as a value proportional to the distortion that would be introduced at the decoder by the loss of that specific element. The analysis-by-synthesis technique [7] we developed computes the distortion caused by the loss of each element, e.g., a frame, referred to as the distortion of the frame in the following, using the following steps: 1) Decoding, including concealment, of the bitstream simulating the loss of the frame being analyzed (synthesis stage). 2) Quality evaluation, that is, computation of the distortion caused by the loss of the frame; the original and the reconstructed picture after concealment are compared using, e.g., Mean Squared Error (MSE). 3) Storage of the distortion value as an indication of the perceptual importance of the analyzed video packet.

The previous operations can be implemented by small modifications of the standard encoding process. The encoder, in fact, usually reconstructs the coded pictures simulating the decoder operations, since this is needed for motion-compensated prediction. Therefore, complexity is only due to the simulation of the concealment algorithm. In case of a simple temporal concealment technique the task is reduced to provide the data to the quality evaluation algorithm. The analysis-by-synthesis technique, as a principle, can be applied to any video coding standard. In fact, it is based on repeating the same steps that a standard decoder would perform, including error concealment. Obviously, the importance values computed with the analysis-by-synthesis algorithm are dependent on a particular encoding, i.e., if the video sequence is compressed with a different encoder, values will be different. Due to the inter-dependencies usually existing between data units, the simulation of the loss of an isolated data unit might not be completely realistic. However, the results will show that the approximation is sufficient to improve the performance of the video communication system. Note that all the considered distortion values accounts for the effect of the dependencies between macroblocks, i.e., the distortion due to error propagation. Nevertheless, experiments in [7] as well as other applications of the analysis-by-synthesis approach to MPEG coded video confirm that such an estimation technique can be successfully used to develop quality optimized video communication algorithms. Note also that such distortion values can be precomputed and stored as side information before transmission.

4. RATE DISTORTION OPTIMIZATION FOR H.264/MVC STEREOSCOPIC VIDEO

In this section, we propose a rate distortion optimization (RDO) framework to which addresses the problem of optimizing the quality of an H.264/MVC stereoscopic video communication. We consider a generic QoS-enabled channel with two QoS levels: a "premium" service with no data loss and low delay, and a "best effort" service typical of Internet-based communications. The proposed RDO approach aims at determining the transmission policy which minimizes the expected distortion at the receiver for each media segment, given a constraint on the amount of data that can be transmitted as "premium". Borrowing the notation of [8], implementing an RDO multimedia transmission systems implies solving the following minimization problem for each considered media segment (in our case, a GOP such as that in Figure 1):

$$\min_{\{\boldsymbol{\pi}\}} E[D(\boldsymbol{\pi})] \quad \text{with} \quad R(\boldsymbol{\pi}) < R_{max}. \tag{1}$$

$E[D(\boldsymbol{\pi})]$ is the expected distortion for a given transmission policy $\boldsymbol{\pi}$, taken from the set $\{\boldsymbol{\pi}\}$ which includes all the possible transmis-

Sequence	PSNR [dB]		Bitrate [kb/s]		
	Main	Sec.	Main	Sec.	Total
Breakdancer	36.29	36.05	216	146	362
Bullinger	37.95	37.75	53	36	89
Jungle	32.07	32.19	406	343	749
Leavinglaptop	34.55	34.65	120	43	163
Metu_1	36.40	36.30	100	84	184
Outdoor	36.08	36.17	120	58	178

Table 1. Characteristics of the H.264/MVC test sequences (shown separately for the main and secondary views).

sion policies for the units (pictures in our case) contained in the GOP. A specific policy $\boldsymbol{\pi}$ is an assignment to the "premium" or "best effort" service of each unit that constitutes the GOP. $R(\boldsymbol{\pi})$ is the data rate sent as "premium" given by the assignment $\boldsymbol{\pi}$, which must be lower or equal to the imposed bound R_{max}.

The cardinality of the set of transmission policies $\{\boldsymbol{\pi}\}$ can be large, since it is exponential in the number of units, however Chou and others [8] showed that it is possible to find a near-optimal solution to the problem with low complexity using the Lagrangian optimization approach. In practice, the problem is recasted into an unconstrained minimization:

$$\min_{\{\boldsymbol{\pi}\}} J = E[D(\boldsymbol{\pi})] + \lambda R(\boldsymbol{\pi}). \tag{2}$$

Under the hypothesis that distortion contributions of the various units are additive, minimizing such an expression is equivalent to perform N minimizations of the same form independently, one for each of the N units (pictures) of the GOP. The expected distortion $E[d_i]$ for each unit i is zero if the policy assigns it to the "premium" service. In case the unit is assigned to the "best effort" service, each packet may be lost with probability p_L, hence $E[d_i]$ is computed as $p_i D_i$ where

$$p_i = 1 - (1 - p_L)^{n_i}, \tag{3}$$

assuming that the unit can be successfully decoded only if all the n_i packets composing the unit are correctly received. The λ value for which the constraint is satisfied can be easily determined using a bisection algorithm, therefore the complexity to solve the expected distortion minimization problem for each couple of GOPs (one in the main and one in the secondary view) is $O(NlogN)$ where N is the total number of units involved.

The distortion D_i produced by the loss of each picture i is calculated via the analysis-by-synthesis [7] approach. The $E[D(\boldsymbol{\pi})]$ is computed as a weighted average of the estimated distortion of the main and the secondary view. Such distortion measure, suggested in [2], allows to take into account, to some degree, the asymmetric perception of the human visual system in case of stereoscopic video. The analysis-by-synthesis approach consists in simulating the loss of each picture i, decoding the resulting video sequence with an appropriate error concealment method and calculating the distortion D_i between the reconstructed and encoded sequence.

5. SIMULATION SETUP

The performance of the proposed RDO framework was evaluated by simulating the transmission of test stereoscopic sequences over a QoS enabled network. The test sequences, whose characteristics are reported in Table 1, were encoded using the H.264/MVC reference software JMVC version 3.0 [9]. Each picture was encoded as a single slice and the encoding scheme is shown in Figure 1, but the GOP size for each view was 32 pictures. Each encoded picture (NALU) was encapsulated in RTP packets as recommended by the IETF standard draft [6], fragmenting a picture into multiple RTP packets if needed. Thus, the loss of even a single RTP packet causes the loss of the whole picture to which the packet belongs. The reference decoder was extended with intra view error concealment functionalities so that each lost picture is replaced with the previous available picture (in case an I or P picture is lost) or by averaging the previous and next available pictures (in case a B picture is lost). For example, the loss of the picture $P_{M,16}$ is concealed using the picture $I_{M,0}$, while the loss of the picture $B_{S,12}$ is concealed using the pictures $B_{S,8}$ and $P_{S,16}$.

The maximum amount of premium traffic is set to 33% of the whole video stream. The service provided by the premium class is assumed to be error free while a 10% uniform packet loss rate has been imposed for the best effort service.

The performance of the proposed RDO strategy, described in Section 4, was compared with the one of an a priori transmission strategy which aims at better protecting pictures referenced by longer prediction chains. For each GOP, the pictures are considered in decoding order (e.g., $I_{M,0}$, $I_{S,0}$, $P_{M,16}$, $P_{S,16}$ etc.) and assigned to the premium class until the maximum allowed premium traffic share is reached, then the remaining pictures are assigned to the best effort class. Finally, besides the RDO and a priori strategies, a transmission strategy which randomly assign pictures to the premium service on a GOP basis while satisfying the premium traffic share constraint was also considered for reference purposes.

6. RESULTS

Table 2 shows the performance of the videocommunication for each combination of test sequence and transmission strategy, in terms of the total byte loss rate (BLR), the average number of lost pictures and the quality measured using the PSNR value.

The random strategy yields the worst visual quality performance, despite the number of lost pictures is lower than the other strategies. The reason is that the a priori and RDO strategies tend to assign pictures which have higher importance (typically, I- and P-type pictures) to the premium service. Those pictures, on average, are larger than the other ones, thus more pictures are transmitted using the best effort service, leading to a higher picture loss rate.

For each simulation, the visual quality is measured first by computing the PSNR value between the original and the received video for each view, then by computing a single quality measure using the weighted average proposed in [2]. In particular, the main view, which is usually characterized by higher quality due to the shorter prediction paths, weights 2/3 while the secondary view weights 1/3. The RDO strategy improves the overall video quality in all the simulations for all the considered video sequences with respect to the a priori strategy. Considering the quality of each of the two views, for some sequences a reduction in the video quality of the secondary view can be noticed, however the increase of the quality of the main view is more than enough to counterbalance it. In order to investigate on the different quality performance of the two views, we compared the estimated distortion for a given view with the actual distortion computed by averaging the performance over various channel realizations. Figure 6 shows the expected and the actual distortion for both views of the Jungle sequence in the case of the RDO strategy. As the figure shows, the estimated distortion on the secondary view is often overstimated with respect to the actual distortion. The same effect is present, but much more limited, on the main view. We attribute the overestimation to the longer prediction paths of the secondary view with respect to the main view, which makes the ad-

Sequence	Strategy	BLR [%]	Lost Pictures [%]	Mean PSNR [dB]	PSNR Main View [dB]	PSNR Sec. View [dB]
Breakdancer	Random	7.31	14.38	30.85	31.30	29.97
	A priori	7.35	17.01	33.21	33.77	32.10
	RDO	7.51	16.91	33.29	34.45	30.99
Leavinglaptop	Random	6.55	12.37	30.24	30.52	29.68
	A priori	7.05	18.51	33.01	33.50	32.02
	RDO	6.72	18.40	33.61	34.23	32.37
Outdoor	Random	6.67	10.88	29.55	29.97	28.71
	A priori	7.22	17.47	30.09	30.36	29.56
	RDO	7.02	15.57	30.97	31.04	30.85
Jungle	Random	6.64	18.53	22.48	23.25	20.94
	A priori	6.70	21.96	26.66	26.67	26.64
	RDO	6.82	21.60	26.95	28.42	24.01
Bullinger	Random	6.92	12.57	32.70	33.18	31.75
	A priori	6.63	17.96	34.73	35.11	33.98
	RDO	7.07	16.98	35.01	35.92	33.20
Metu1	Random	6.08	11.44	30.91	31.38	29.97
	A priori	6.47	18.56	32.95	33.34	32.18
	RDO	6.61	15.64	32.98	33.89	31.17

Table 2. Video quality performance of the different video prioritization strategies. Average of 20 channel realizations. The packet loss ratio of the best effort service (p_L) is equal to 0.1.

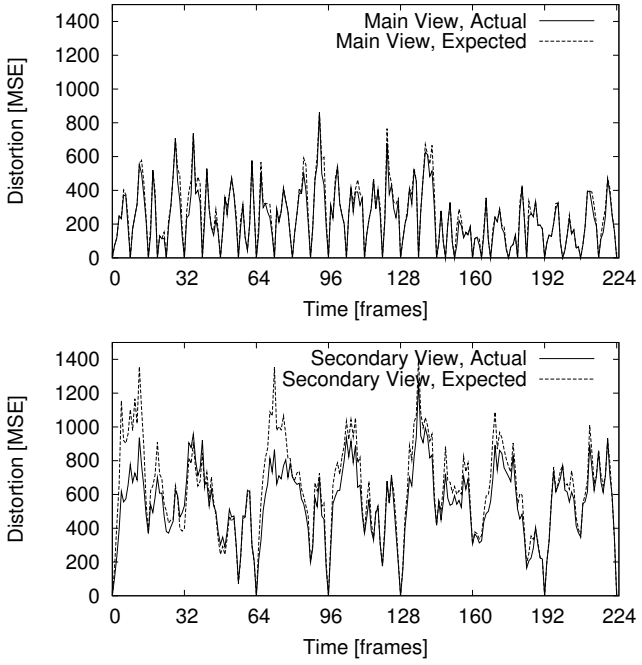

Fig. 2. Expected and average distortion for both views, for the Jungle sequence. The MSE is computed comparing the corrupted video with the error-free decoded video.

ditive distortion hypothesis used by the RDO strategy less reliable in computing the expected distortion, thus negatively affecting the performance of the RDO strategy.

7. CONCLUSIONS

In this paper we presented a rate-distortion optimized framework for the transmission of stereoscopic H.264/MVC video on networks with QoS capabilities. We also showed how the distortion values associated to each encoding unit can be computed by means of the analysis-by-synthesis approach. Simulation performance were evaluated using an objective quality measure which accounts for the peculiarities of the human visual system in stereoscopic video quality perception. Results show that the proposed rate-distortion optimized strategy significantly improves the quality of the communication with respect to a reference a priori transmission optimization strategy. Future work will be devoted to improve the accuracy of the distortion estimation for the secondary view to increase the overall perceived video quality.

8. REFERENCES

[1] Joint Video Team of MPEG and ITU-T, "Joint draft 7.0 on multiview video coding (JVT-AA209)," Apr. 2008.

[2] N. Ozbek, A. M. Tekalp, and E. T. Tunali, "Rate allocation between views in scalable stereo video coding using an objective stereo video quality measure," in *Proc. of IEEE ICASSP*, 2007, vol. 1, pp. 1045–1048.

[3] C. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *Electronics letters*, vol. 44, no. 16, 2008.

[4] L. Stelmach, W. J. Tam, D. Meegan, and A. Vincent, "Stereo image quality: effects of mixed spatio-temporal resolution," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 10, no. 2, pp. 188–193, 2000.

[5] A. Aksaya et al., "End-to-end stereoscopic video streaming with content-adaptive rate and format control," *Signal Processing: Image Communication*, vol. 22, pp. 157–168, 2007.

[6] Y. K. Wang and T. Schierl, "RTP payload format for MVC video," *draft-wang-avt-rtp-mvc-02*, Aug. 2008.

[7] E. Masala and J. C. De Martin, "Analysis-by-synthesis distortion computation for rate-distortion optimized multimedia streaming," in *Proc. of IEEE ICME*, July 2003, vol. 3, pp. 345–348.

[8] P. A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetized media," *IEEE Transactions on Multimedia*, vol. 8, no. 2, pp. 390–404, 2006.

[9] "H.264/MVC reference software JMVC version JMVC_3_0," .