ROBUST TRANSMISSION OF MULTI-VIEW VIDEO STREAMS USING FLEXIBLE MACROBLOCK ORDERING AND SYSTEMATIC LT CODES

S. Argyropoulos[1,4]*, A. S. Tan*[2]*, N. Thomos*[3]*, E. Arikan*[2]*, and M. G. Strintzis*[1,4]

[1]Electrical and Computer Engineering Dept., University of Thessaloniki, Hellas
[2]Electrical and Electronics Engineering Dept., Bilkent University, Ankara, Turkey
[3]Signal Processing Institute - ITS, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, Switzerland
[4]Informatics and Telematics Institute, Thessaloniki, Hellas

ABSTRACT

The transmission of fully compatible H.264/AVC multi-view video coded streams over packet erasure networks is examined. Macroblock classification into unequally important slice groups is considered using the Flexible Macroblock Ordering (FMO) tool of H.264/AVC. Systematic LT codes are used for error protection due to their low complexity and advanced performance. The optimal slice grouping and channel rate allocation are jointly determined by an iterative optimization algorithm based on dynamic programming. The experimental evaluation clearly demonstrates the validity of the proposed method.

Index Terms— Multi-view coding, FMO, Fountain codes, LT codes, unequal error protection.

1. INTRODUCTION

Streaming of multi-view video content has received much attention as it enhances user experience. The multi-view coding (MVC) systems capture video streams using multiple cameras located in various angles and positions. Unlike single-view video coders which perform spatial and temporal decorrelation, MVC schemes also exploit the redundancy among adjacent views. The MVC schemes, as the one presented in [1], can use synthesis prediction and multi-view reference picture management. However, the generated MVC coded streams are sensitive to channel errors and few errors may cause significant degradation in video quality. Thus, the design of reliable multi-view transmission systems involves sophisticated selection of the employed error resilient tools and channel coding techniques.

A variety of error resilient video transmission schemes have been recently presented in the literature. Most of them employ Reed-Solomon (RS) codes to cope with erased packets. In [2], a rate-distortion algorithm was proposed for the determination of source and channel rate for each quality layer. Other schemes, such as the Data Partitioning system of [3], use Turbo codes for channel protection. However, RS and Turbo codes are impractical for real-time applications because of their high computational cost. Instead of the usually employed channel coding tecnhiques, LT codes can be used in packet loss networks [4]. LT codes are low-complexity rateless codes able to generate, on-the-fly, unlimited number of protection symbols from given source symbols. This property makes LT codes appropriate for multi-view video transmission systems where large files are transmitted. Moreover, feedback channels are not necessary with LT codes since every received packet is decodable regardless its transmission order.

In this paper, a novel slice-group multi-view video coding scheme (SG-MVC) is presented. The proposed scheme employs the MVC scheme presented in [5]. This codec is an extension of the H.264/AVC standard and performs disparity estimation using five different reference modes for improved performance and reduced complexity. Unequally important slice groups are formed by adapting FMO to MVC, as suggested in [6]. The macroblock (MB) classification is dynamic and depends on the multi-view video content and the estimated end-to-end distortion. The resulting stream is unequally protected using systematic LT codes. The optimal MB classification and channel code rate allocation are jointly determined by an interdependent procedure of two successive steps. The proposed scheme is experimentally evaluated and validated.

2. SG-MVC CODEC

The proposed SG-MVC scheme is based on the multi-view extension of the H.264/AVC standard [5] developed under the 3DTV project. In this coder, the buffering structure of H.264/AVC is modified to implement multiple referencing modes and exploit the redundancy among views. The MBs are processed in raster scan order to form slices as in conventional hybrid video coding systems. Error concealment is used to moderate the impact of erased packets occurring during transmission. However, it can not provide much assistance when neighbouring blocks are also erased.

In this work, the dynamic formation of the Macroblock Allocation Map (MBAMap) using the explicit mode of FMO is proposed. Specifically, MBs are classified into slice groups by examining their relative significance. For the optimal MB classification, the contribution of each MB to the overall video quality should be determined. Next, a rigorous expression for the end-to-end MB distortion estimation is defined which will assist in formulating an efficient MB classification algorithm.

2.1. End-to-end distortion estimation

In hybrid video coding systems, MBs not only affect the current frame but also the following frames. The expected end-to-end distortion of a MB could be used as a significance metric [7]. In an error-prone environment, the end-to-end MB distortion should be estimated both for errorless reception at the receiver and for MB

This work was supported by the EC under contract FP6-511568 3DTV and in part by TUBITAK under contract BTT-Turkiye 105E065. A. S. Tan is supported in part by graduate scholarship of TUBITAK.

erasure. When a MB is received intact, the end-to-end distortion is equal to the source distortion. For inter frames, the end-to-end distortion should also take into account the distortion of the referenced MBs. For erased MBs, the distortion due to error concealment should be considered. Below, the calculation of each of the above distortion terms is explained briefly.

Let $s_i, i = 1, 2, 3$ and $MB_{n,m}$ denote respectively the i_{th} slice group and the m_{th} MB in frame n. The overall end-to-end distortion $D(n, m, s_i)$ takes into account the distortion due to both packet erasures and error propagation. In particular, the MBs that minimize end-to-end distortion $D(n, m, s_i)$ are classified to the i_{th} slice group. The distortion $D(n, m, s_i)$ is given by:

$$D(n,m,s_i) = (1-p)\cdot\Big(D_q(n,m,s_i) + D_{prop}(n,m)\Big) + p\cdot D_{ec}(n,m) \quad (1)$$

where p denotes the packet error rate, and $D_q(n, m, s_i)$, $D_{prop}(n, m)$, and $D_{ec}(n, m)$ represent the distortions due to quantization, error propagation, and concealment respectively. The terms $D_q(n, m, s_i)$ and $D_{ec}(n, m)$ are derived directly during the encoding process while the error propagated distortion $D_{prop}(n, m)$ is estimated as explained in the following.

In order to calculate $D_{prop}(n, m)$, an error propagated distortion map D_{ep} is built on a block basis after each frame is coded, as suggested in [7]. $D_{prop}(n, m)$ is zero for intra frames since MBs are coded without reference to previous frames. For the reference frames, the motion compensated blocks are estimated.

The error concealed distortion $D_{ec}(n, m)$ is computed using non-normative techniques for spatial and temporal concealment. Alternatively, more advanced error concealment techniques, like the stereoscopic error concealment method introduced in [8], which perform concealment among the multiple views could be used. However, they are not employed in the proposed scheme for complexity reasons.

Since the transmission scenario is over packet erasure networks, channel codes should be used for efficient protection. In the following section, the utilized systematic LT codes are presented in detail.

3. FOUNTAIN CODES

Most of the transmission systems in the literature employ RS codes for channel protection. Nevertheless, the RS codes can also be considered as a fountain code. Although RS codes are perfect codes, which means that they can recover encoded information from any set of packets larger or equal to the original set of packets, they are practical for small codewords. Moreover, RS codes can generate limited number of packets, which is restrictive for transmission applications. The decoding requires quadratic time which is too slow even for small sets of packets. Conversely, LT codes have lower computational cost with small performance loss compared to RS codes due to overhead.

LT codes are the first practical realization of fountain codes operating on lossy packet networks. LT coders partition the original data into packets called *input symbols* $\mathrm{I}_i, i = 1, ..., k$, where I_i is a row vector of length L. Using the input symbols the encoder generates an arbitrary number of packets called *output symbols* $\mathrm{S}_i = \sum_{j\in\psi_i} \oplus \mathrm{I}_j$ where S_i is a row vector of length L, ψ_i is a set of input symbol indexes and $\sum \oplus$ denotes XOR-sum. The encoding process of original LT coding is performed in linear time. The set of all input symbols and output symbols is represented in matrix form as $\mathbf{I} = [\mathrm{I}_1^T, \mathrm{I}_2^T, ..., \mathrm{I}_k^T]^T$ and $\mathbf{S} = [\mathrm{S}_1^T, \mathrm{S}_2^T, ..., \mathrm{S}_n^T]^T$ respectively. During the lossy transmission the receiver does not request retransmission of lost symbols. The receiver waits for reception of at least $k(1 + \delta)$ output symbols to complete the decoding, where δ is the overhead of LT coding and tends to zero as k increases. Thus, the LT coding scheme is adaptive to the erasure rate of the channel and scalable in the number of users. Decoding of LT codes is a sequential process based on the belief propagation method. The proposed system employs systematic LT codes which are presented in the following section.

3.1. Systematic LT Codes

The direct access to original source data is beneficial in video transmission applications since error concealment techniques are applied for lost segments of original data when the channel decoder fails to recover all the input symbols. Systematic codes provide direct access to the original data. The original structure of LT coding is non-systematic as proposed in [4]. Systematization of the Raptor codes which use LT codes as inner codes is described in [9]. Based on a similar notation, the systematization of LT codes is described in the following lines.

Fountain codes can be represented as linear block codes when the number of output symbols is fixed. Let the *k* input symbols be denoted as $\mathbf{I} = [\mathrm{I}_1^{\mathrm{T}}, \mathrm{I}_2^{\mathrm{T}}, ..., \mathrm{I}_{\mathrm{k}}^{\mathrm{T}}]^{\mathrm{T}}$. Define Γ_i as a row vector with ones at positions corresponding to the index of XOR-summed input symbols for the formation of an output symbol. A generator matrix for the first *n* output symbols is defined as $\mathbf{G}_{\mathrm{LT}} = [\Gamma_1^T, \Gamma_2^T, ..., \Gamma_n^T]^T$. Then the generated output symbols $\mathbf{S} = [\mathrm{S}_1^T, \mathrm{S}_2^T, ..., \mathrm{S}_n^T]^T$ are represented as $\mathbf{S} = \mathbf{G}_{\mathrm{LT}}\mathbf{I}$. Denote the first *k* row of $\mathbf{G}_{\mathrm{LT}}$ as $\mathbf{G}_{1:k}$ such that $\mathbf{G}_{1:k} = [\Gamma_1^T, \Gamma_2^T, ..., \Gamma_k^T]^T$. Similarly, denote the $(k+1)_{th}$ to n_{th} row of $\mathbf{G}_{\mathrm{LT}}$ as $\mathbf{G}_{k+1:n}$. If the rows of $\mathbf{G}_{1:k}$ are independent then the inverse of $\mathbf{G}_{1:k}$ exists. Just before the input symbols are given to the encoder, $\mathbf{I}$ is multiplied with $\mathbf{G}_{1:k}^{-1}$. As a result of this process the following result is obtained at the output of encoder:

$$\begin{bmatrix} \mathbf{G}_{1:k} \\ \mathbf{G}_{k+1:n} \end{bmatrix} \mathbf{G}_{1:k}^{-1}\mathbf{I} = \begin{bmatrix} \mathbf{I} \\ \mathbf{S}'_{k+1:n} \end{bmatrix} = \mathbf{S}'$$

where $\mathbf{S}'$ represents the new output symbols, $\mathbf{S}'_{1:k} = \mathbf{I}$ and $\mathbf{S}'_{k+1:n} = [\mathrm{S}'^T_{k+1}, ..., \mathrm{S}'^T_n]^T$. Thus, the LT encoder is transformed into a systematic encoder and still preserves the fountain property of generating potentially limitless output symbols. The decoding of systematic LT codes is the same as in the non-systematic case where belief propagation method is used. However, due to the increased degree of nodes the belief propagation decoder does not perform well and fails due to lack of degree-1 nodes. ML decoding is used when the belief propagation decoder fails. The ML decoder finds the reduced row echelon form of the generator matrix and it solves the resulting equation system for recovering the input symbols.

4. CHANNEL RATE ALLOCATION

In the preceding analysis for an optimal classification, it was assumed that the distortion between the original and reconstructed coefficients is known. In practice, the actual distortion depends on the reconstructed coefficients *after* channel decoding. This means that the processes of slice grouping and channel allocation are actually interdependent. For this reason, the formation of slice groups and their unequal error protection (UEP) are optimized in the proposed system by iterating two interdependent steps.

The employment of the FMO enables the formation of slice groups of unequal importance. In the proposed scheme, the unequally important slice groups consist of equally-sized slices (pack-

ets). LT codes were chosen for UEP protection. Since, in general, different frames have different classification maps, channel rate allocation is performed at the frame level. The optimization seeks for:

- the optimal classification of MBs into slice groups
- the optimal channel channel protection of slice groups.

The optimization algorithm intends to minimize the average expected distortion $\overline{D}$ subject to a channel rate constraint determined by experimentation. This is necessary to avoid overprotection of the first frames. Without this constraint, the first frames in the sequence would be protected by the maximum allowed channel protection and drift would occur.

The average expected distortion $\overline{D}$ in case of s slice groups is equal to[1]:

$$\overline{D} = \sum_{l=1}^{s} \left\{ \sum_{i=1}^{N_l} D_{f,l} \cdot P_l(i) \cdot P_{LT}(i) + \sum_{i=N_l+1}^{N_l+K_l-1} D_{f,i,l} \cdot P_l(i) + D_{f,PC,l} \cdot P_l(N_l+K_l) \right\} \quad (2)$$

where K_l and N_l are the number of source and LT packets of the l_{th} slice group respectively. $P_{LT}(i)$ is the probability of decoding all source symbols given that i symbols were lost and can be determined via simulations. $P_l(i)$ is the probability of losing i packets out of the $N_l + K_l$ packets of the l_{th} slice group and is given by:

$$P_l(i) = \binom{N_l + K_l}{i} \cdot p^i \cdot (1-p)^{N_l+K_l-i} \quad (3)$$

where p is the packet erasure probability.

The distortion $D_{f,PC,l}$ in the last term of (2) expresses the distortion when all packets of the l_{th} slice group are erased and concealed by slice group replication. Finally, $D_{f,i,l}$ represents the distortion introduced when the current frame slice group is concealed by slices received intact and $D_{f,l}$ the distortion when the channel protection is sufficient to recover all erased packets.

4.1. Packet classification optimization

In this section, a solution to the previously formulated optimization problem is proposed. The optimization objective is actually twofold. Specifically, it includes the determination of both the number of slices classified into each slice group and their channel protection. In general, reaching an optimal solution of the joint problem is a difficult task. In this work, a two-step optimization procedure, which iteratively determines the packet classification and the channel protection is proposed. Although, this approach to the solution of the optimization problem does not guarantee global optimization, in practice it yields very satisfactory results. The optimization procedure is summarized as follows:

1. Determine the channel protection of each frame.
2. Classify MBs into slice groups according to the algorithm presented in Section 2.
3. Find the optimal channel protection for the above classification.
4. Calculate the expected distortion of allowable neighboring MB classifications with the restriction that a single packet can be exchanged between successive classes.
5. Compare the expected distortion of the ancestor classification with the lowest average distortion of all descendant classifications of step 3. If a classification with lower expected distortion exists, it is considered as optimal and steps 2 to 6 are repeated, else the algorithm is terminated. When the same packet is exchanged, in two successive iterations, between two slice groups the algorithm is again terminated.

The objective is to optimize the channel rate allocation by minimizing the expected distortion given by (2). The dynamic programming algorithm presented in [10] is used to reduce the computational cost.

5. EXPERIMENTAL RESULTS

The proposed scheme for the transmission of multi-view video streams over IP/UDP/RTP was evaluated for the "Race1" sequence coded at $30fps$. The size of each frame is 640×480. 60 frames from each of the eight views were coded using the reference mode 3 [5]. Group of Pictures (GOPs) of $IPPP\ldots$ structure were considered. Intra update of 10% of the total number of frame MBs was used to prevent error propagation. The MBAMap is contained into the Picture Parameter Set (PPS) which was transmitted to the decoder.

The NS-2 network simulator [11] was used to simulate the transmission scenarios. The systematic LT codes were used as the channel protection scheme. The LT codec was implemented in the application layer in NS-2 and the encoded packets were generated according to the channel rate allocation algorithm given in Section 4. The packet size was set to 200 bytes, which is a good compromise between error resiliency and compression efficiency. Since the decoder needs to know the indexes of the XOR-summed input symbols the random seed number generators of the encoder and the decoder are synchronized. This approach does not increase the overhead and it is suitable for unicast type applications where synchronization is possible.

Two transmission scenarios were considered for the simulation tests. In the first one, the multi-view video is transmitted through a packet erasure channel, while in the second case, there are three nodes and four external TCP connections as shown in the topology in Fig. 1. The external TCP sources are connected to an exponential traffic generator to approximate the behavior of the Internet, i.e. the packet inter-arrival time is random variable following an exponential distribution. Packet losses in the TCP scenario occur due to the packet drops caused by network congestion. The external TCP traffic rates R_{TCP} are modified to obtain different loss patterns with different loss rates. All reported PSNR results are the average of 50 simulations.

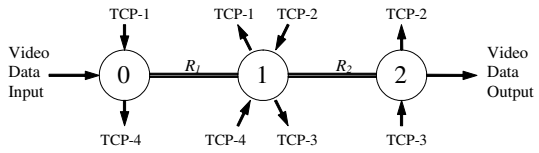
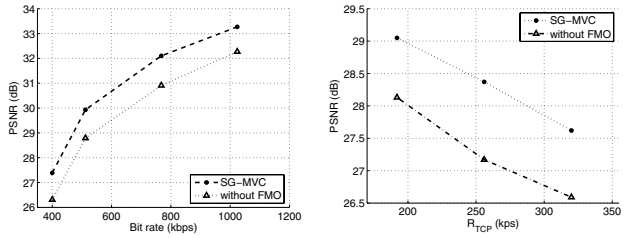


Fig. 1. Network topology for congestion simulations.

The proposed SG-MVC codec was compared with a variant of the scheme which does not use FMO (non-FMO coding). Equal strength channel codes were used in both variants. The results for transmission over a packet erasure network with 10% packet error rate (PER) are illustrated in Fig. 2 (a). Optimization was performed

[1]The probability of LT decoding failure is assumed equal to 1, without loss of generality, when the number of erased symbols exceeds the total number of parity symbols.



(a) (b)

Fig. 2. PSNR results for the transmission: (a) over a packet erasure network with 10% PER, (b) using the topology of Fig. 1.

for 10% PER. The performance gain is over $1dB$ for a variety of transmission rates.

Moreover, the results for the transmission scenario using the topology of Fig. 1 are depicted in Fig. 2 (b). The channel link capacities were set to $R_1 = R_2 = 1Mbps$ and simulations for three different values of R_{TCP} as $192, 256$ and 320 *kbps* were performed. These results can be considered as a mismatch case since burst of erasures occur. The proposed scheme yields approximately $1dB$ improvement in $PSNR$ for the simulated parameters compared to the case without FMO.

The proposed scheme was also evaluated for channel mismatch conditions. In Fig. 3 results are presented for the transmission of the "Race1" sequence coded at 512 *kbps*. All schemes are optimized for 10% packet loss rate and tested for a large variety of channel conditions. The results demonstrate that the proposed scheme provides better error resilience since the quality degrades more gracefully. The worse performance of the full scheme in error free case is attributed to the inferior compression efficiency when FMO is used. The gain of the full scheme over the scheme without FMO coding widens when the channel conditions deteriorate.

The enhanced performance of the proposed SG-MVC scheme is attributed to the more effective MB classification, which boosts the performance of the error concealment algorithm, and to the efficient channel protection. More specifically, the UEP algorithm enables the application of channel codes of higher code rate. Considering the above, the performance gain should not be attributed solely to the adaptive slice grouping itself or the UEP algorithm, but rather to their synergistic cooperation.

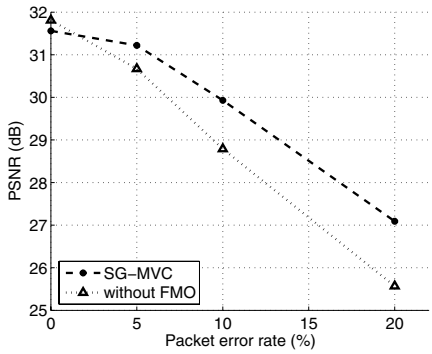


Fig. 3. PSNR results for the mismatch transmission scenario.

6. CONCLUSIONS

A novel method for the transmission of multi-view video sequences over packet erasure channels was proposed in this paper. The error resilient features of the H.264/AVC standard were exploited and systematic LT codes were employed in order to protect effectively the resulting streams against transmission errors. A framework for optimal classification of MBs into slice groups using FMO and optimal UEP was also proposed. The performance of the algorithm was evaluated experimentally and its performance was found to be significantly better than that of MVC schemes using one slice group.

7. ACKNOWLEDGEMENTS

The authors would like to thank C. Bilen, A. Aksay, and G. Bozdagi Akar for providing the multi-view codec of [5].

8. REFERENCES

[1] E. Martinian, A. Behrens, J. Xin, A. Vetro, and H. Sun, "Extensions of H.264/AVC for multiview video compression," in *IEEE Int. Conf. on Image Processing*, Atlanta, USA, Oct. 2006.

[2] C. M. Fu, W. L. Huang, and C. L Huang, "Efficient post-compression error-resilient 3D-scalable video transmission for packet erasure channels," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Philadelphia, PA, USA, 2005.

[3] P. Y. Yip, J. A. Malcolm, W. A. C. Fernando, K. K. Loo, and H. K. Arachchi, "Joint source and channel coding for H.264 compliant stereoscopic video transmission," in *Canadian Conf. on Electrical and Computer Engineering*, Saskatoon, Canada, May 2005.

[4] M. Luby, "LT Codes," in *Proc. of the 43rd Annual IEEE Symposium on Foundations of Computer Science (FOCS)*, 2002.

[5] C. Bilen, A. Aksay, and G. Bozdagi Akar, "A multi-view video codec based on H.264," in *IEEE Int. Conf. on Image Processing*, Atlanta, USA, Oct. 2006.

[6] N. Thomos, S. Argyropoulos, N. V. Boulgouris, and M. G. Strintzis, "Robust transmission of H.264/AVC streams using adaptive group slicing and unequal error protection," *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006.

[7] Y. Zhang, W. Gao, H. Sun, Q. Huang, and Y. Lu, "Error resilience cideo coding in H.264 encoder with potential distortion tracking," in *IEEE Int. Conf. on Image Processing*, Singapore, Oct. 2004.

[8] S. Knorr, C. Clemens, M. Kunter, and T. Sikora, "Robust concealment for erroneous block bursts in stereoscopic images," in *2nd Int. Symposium on 3D Data Processing, Visualization, and Transmission (3DPVT'04)*, Thessaloniki, Greece, Sept. 2004.

[9] M. Luby, M. Watson, T. Gasiba, T. Stockhammer, and W. Xu, "Raptor codes for reliable download delivery in wireless broadcast systems," in *Proc. of the IEEE CCNC*, 2006.

[10] N. Thomos, N. V. Boulgouris, and M. G. Strintzis, "Wireless image transmission using turbo codes and optimal unequal error protection," *IEEE Trans. on Image Processing*, vol. 14, no. 11, pp. 1890–1901, 2005.

[11] "The network simulator - ns2," http://www.isi.edu/nsnam /ns/index.html.