ROBUST NETWORK-ADAPTIVE OBJECT-BASED VIDEO ENCODING

Haohong Wang *Aggelos K. Katsaggelos*

Department of Electrical and Computer Engineering
Northwestern University, Evanston, IL 60208, USA
Email: {haohong, aggk}@ece.northwestern.edu

ABSTRACT

Joint source-network encoding of object-based video is a very important and challenging research topic, which has not been adequately explored. In this paper, we propose a robust network-adaptive encoding approach for object-based video. The framework jointly considers source coding, packet loss during transmission, and error concealment at the decoding. The proposed method guarantees the minimum expected distortion for the decoded video, by optimally allocating the shape and texture coding parameters at the encoder. The resulting optimization problem is solved by Lagrangian relaxation and dynamic programming. Experimental results demonstrate that the proposed method has significant gains over the non network-adaptive method.

1. INTRODUCTION

Robust video compression (see [1] for a recent review) refers to controlling the trade-off between error robustness and coding efficiency. Network-adaptive video encoding is a robust video compression approach that designs and optimizes the source encoder by considering transmission factors, such as error control, packetization, packet scheduling and retransmission, routing, and error concealment. The error resilience is gained by optimally selecting the encoding mode [2-4], such as prediction mode and quantization parameters for each packet (or macroblock), which enables the decoded video to reach the minimum expected distortion for the available resources. Various approaches have been proposed in the literature for solving such problems. In [5], the optimal encoding parameters are selected jointly considering the dynamic allocation of the transmission power, which controls the level of protection provided for each packet and directly affects the decoded video quality. In [6], the recent advances for network-adaptive video streaming are reviewed, and emphasis is placed on joint video encoding and packet scheduling and joint video encoding and control of the packet dependency.

In recent years, object-based video coding has become one of the most important research and development topics in the visual communication field, as the fast growing modern multimedia applications require content-based interactivity and scalability. Compared with the conventional frame-based video, which is represented by encoding a sequence of rectangular frames, object-based video coding is based on the concept of encoding arbitrarily shaped video objects. The video objects are described by their shape and texture, which are compressed by using totally different encoding schemes. In [7-8] we proposed an operational rate-distortion optimal bit allocation scheme for object-based video coding, which enables the rate controller to optimally allocate bits among shape, texture, and motion. In this paper, we propose a network-adaptive encoding scheme for robust object-based video transmission. To the best of our knowledge, there has been no reported work in the literature on network-adaptive encoding of object-based video.

The rest of the paper is organized as follows. In section 2, the problem is formally defined, and in section 3, the optimal solution is developed. Section 4 describes the simulations and experimental results. We draw conclusions in the last section.

2. PROBLM FORMULATION

The problem at hand is to choose the coding parameters for shape and texture data with a given rate constraint, so as to minimize the total expected distortion at the decoder. This can be represented by

$$\text{Minimize } E[D], \text{ subject to } R \le R_{budget,} \quad (1)$$

where $E[D]$ is the expected total distortion, R is the total bit rate, and R_{budget} is the bit budget for the frame. The optimization is over the source coding parameters and is restricted to the frame level.

A. System model

Let us denote by $\{m_1, m_2, \ldots, m_N\}$ the total N slices (or packets) in the frame, by $V=\{V_1, V_2, \ldots, V_N\}$ the set of admissible coding decision vectors for the slices, and v_i a coding decision vector for m_i ($v_i \in V_i$). Each packet is independently decodable. Let us denote by $R_i(v_i)$ the bit rate for m_i. Clearly,

$$R = \sum_{i=1}^{N} R_i(v_i). \quad (2)$$

In this paper, we assume a packet lossy network, that is, the packets are either received error-free or lost. We also assume that the probability with which a packet has

arrived at the receiver is available at the encoder. This can be either specified in the initial negotiations, or adaptively calculated from messages exchanged by the transmission protocol. Let us denote by ρ_i the probability of loss for the ith packet.

B. Expected distortion

For lossy networks, the distortion at the receiver is a random variable. Let $E[D_i(v_{i-a},...,v_i)]$ represents the expected distortion at the receiver for the ith packet; then

$$E[D_i(v_{i-a},...,v_i)] = (1-\rho_i)E[D_{R,i}(v_i)] + \rho_i E[D_{L,i}(v_{i-a},...,v_i)], \quad (3)$$

where $E[D_{R,i}(v_i)]$ is the expected distortion if the ith packet is received correctly at the decoder, $E[D_{L,i}(v_{i-a},...,v_i)]$ is the expected distortion if the ith packet is lost, and a is number of packets of the current packet depends on. It is clear that $E[D_{R,i}(v_i)]$ depends on the source coding parameters for the ith packet, while $E[D_{L,i}(v_{i-a},...,v_i)]$ depends on the concealment strategy used at the decoder. Clearly,

$$E[D] = \sum_{i=1}^{N} E[D_i(v_{i-a},...,v_i)]. \quad (4)$$

The calculation of $E[D]$ in Eq. (4) is dependent on a concealment strategy at the decoder. Out of a number of possible strategies, we consider the following one (an extension of ROPE in [3]), which allows for the recursive calculation of $E[D]$ at the encoder. We assume that each row of macroblocks becomes a packet. To recover the lost shape (texture) macroblocks in a packet, the decoder uses the shape (texture) motion vector of the neighboring macroblocks above as the concealment motion vector. If the concealment motion vector is not available, e.g., the above macroblock is also lost, then the decoder uses a zero motion vector for concealment.

We use the mean squared error (MSE) to measure the expected distortion. That is, the expected distortion for the ith packet is calculated by summing up the expected distortion of all pixels in the packet, $\sum_{j=1}^{M} E[d_j]$, where $E[d_j]$ is the expected distortion at the receiver for the jth pixel, and M is the total number of pixels in the packet. Let us denote by f_n^j the original value of pixel j in frame n, and $\tilde{f}_n^j$ its decoder reconstruction. Clearly,

$$f_n^j = s_n^j t_n^j, \text{ and } \tilde{f}_n^j = \tilde{s}_n^j \tilde{t}_n^j, \quad (5)$$

where s_n^j (s_n^j=0 for transparent or 1 for opaque alpha mask) and t_n^j are the corresponding shape and texture components of f_n^j, and $\tilde{s}_n^j$ and $\tilde{t}_n^j$ are the corresponding shape and texture components of $\tilde{f}_n^j$ (here we assume the background pixel intensity values at object composition are unknown for the encoder, and thus are set to 0 by default). We can then write

$$E[d_j] = E[(f_n^j - \tilde{f}_n^j)^2] = (s_n^j t_n^j)^2 - 2 s_n^j t_n^j E[\tilde{s}_n^j \tilde{t}_n^j] + E[(\tilde{s}_n^j \tilde{t}_n^j)^2]. \quad (6)$$

In calculating $E[d_j]$ in Eq. (6), the first and second moments are needed. We show next how $E[\tilde{s}_n^j \tilde{t}_n^j]$ can be computed recursively in time. $E[(\tilde{s}_n^j \tilde{t}_n^j)^2]$ is computed in a similar manner, but omitted here, due to lack of space.

Since texture and shape information can be independently encoded in the intra and inter modes, we consider the following four cases,

$$E[\tilde{s}_n^j \tilde{t}_n^j](I,I) = (1-\rho_i)\hat{s}_n^j \hat{t}_n^j + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^{k_s}\tilde{t}_{n-1}^{k_t}] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j \tilde{t}_{n-1}^j], \quad (7)$$

$$E[\tilde{s}_n^j \tilde{t}_n^j](I,P) = (1-\rho_i)\hat{s}_n^j (\hat{e}_n^j + E[\tilde{t}_{n-1}^{m_t}]) + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^{k_s}\tilde{t}_{n-1}^{k_t}] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j \tilde{t}_{n-1}^j], \quad (8)$$

$$E[\tilde{s}_n^j \tilde{t}_n^j](P,I) = (1-\rho_i)E[\tilde{s}_{n-1}^{m_s}]\hat{t}_n^j + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^{k_s}\tilde{t}_{n-1}^{k_t}] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j \tilde{t}_{n-1}^j], \quad (9)$$

$$E[\tilde{s}_n^j \tilde{t}_n^j](P,P) = (1-\rho_i)(E[\tilde{s}_{n-1}^{m_s}]\hat{e}_n^j + E[\tilde{s}_{n-1}^{m_s}\tilde{t}_{n-1}^{m_t}]) + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^{k_s}\tilde{t}_{n-1}^{k_t}] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j \tilde{t}_{n-1}^j], \quad (10)$$

where shape is intra coded in (7) and (8), and inter coded in (9) and (10); texture is intra coded in (7) and (9), and inter coded in (8) and (10); $\hat{s}_n^j$ and $\hat{t}_n^j$ are the encoder reconstructed shape and texture of the jth pixel; pixel j in frame n is predicted by pixel m in frame $n-1$ if the motion vector is available, otherwise predicted by pixel k if the concealment motion vector is available; The subscript s and t of k_s, k_t, m_s, and m_t in (7)-(10) are used to distinguish shape from texture, because the motion vector or concealment motion vector of shape could be different from that of texture. Computing $E[\tilde{s}_n^j \tilde{t}_n^j]$ in (7)-(10) depend on the computing of $E[\tilde{s}_n^j]$ and $E[\tilde{t}_n^j]$, which are calculated recursively as follows,

$$E[\tilde{s}_n^j](I) = (1-\rho_i)\hat{s}_n^j + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^k] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j], \quad (11)$$

$$E[\tilde{t}_n^j](I) = (1-\rho_i)\hat{t}_n^j + \rho_i(1-\rho_{i-1})E[\tilde{t}_{n-1}^k] + \rho_{i-1}\rho_i E[\tilde{t}_{n-1}^j], \quad (12)$$

$$E[\tilde{s}_n^j](P) = (1-\rho_i)E[\tilde{s}_{n-1}^m] + \rho_i(1-\rho_{i-1})E[\tilde{s}_{n-1}^k] + \rho_{i-1}\rho_i E[\tilde{s}_{n-1}^j], \quad (13)$$

$$E[\tilde{t}_n^j](P) = (1-\rho_i)(\hat{e}_n^j + E[\tilde{t}_{n-1}^m]) + \rho_{i-1}\rho_i E[\tilde{t}_{n-1}^j] + \rho_i(1-\rho_{i-1})E[\tilde{t}_{n-1}^k], \quad (14)$$

where shape and texture are intra coded in (11) and (12), and inter coded in (13) and (14).

It is not hard to understand that $E[\tilde{s}_n^{k_s}\tilde{t}_n^{k_t}]$ and $E[\tilde{s}_n^{m_s}\tilde{t}_n^{m_t}]$ in (7)-(10) can be computed recursively in a similar manner as $E[\tilde{s}_n^{j}\tilde{t}_n^{j}]$. However, recursively computing these inter-pixel cross-correlation terms may require computing and storing all inter-pixel cross-correlation values for all frames in the video sequence. The amount of this computation and storage is infeasible even for moderate size frame. A similar problem was reported in [9, 10] when using half-pixel motion estimation. Approximations on these cross-correlation terms were proposed to reduce the computational complexity.

We use a model-based cross-correlation approximation method, as in [10], to estimate $E[st]$ in terms of $E[s]$, $E[t]$, $E[s^2]$, $E[t^2]$ and standard deviations σ_s and σ_t. We utilize and compare the following three models:

Model I: s and t are uncorrelated. That is,
$$E[st]=E[s]E[t]. \tag{15}$$

Model II: $t=b+cs$, where b and c are unknown constants ($c \geq 0$). So,
$$E[st]=E[s]E[t]+ \sigma_s \sigma_t . \tag{16}$$

Model III: $t=N+bs$, where b is an unknown constant, and N is a zero-mean random variable independent of s. Then,
$$E[st]=\frac{E[t]}{E[s]}E[s^2]. \tag{17}$$

The value of $E[st]$ is bounded by 0 and $\sqrt{E[s^2]E[t^2]}$.

3. OPTIMAL SOLUTION

Utilizing the notation we have introduced in the previous section, the optimization problem becomes

$$\text{Minimize} \sum_{i=1}^{N} E[D_i(v_{i-a},...,v_i)],$$
$$\text{such that} \sum_{i=1}^{N} R_i(v_i) \leq R_{budget}. \tag{18}$$

We derive a solution to problem (18) using the Lagrange multiplier method to relax the constraint, so that the relaxed problem can be solved using a shortest path algorithm. We first define the Lagrangian cost function

$$J_\lambda(v) = D + \lambda R = \sum_{i=1}^{N} [D_i(v_{i-a},...,v_i) + \lambda R_i(v_i)], \tag{19}$$

where λ is the Lagrange multiplier. It has been shown in [11] that if there is a λ^* such that $v^* = \arg[\min_v J_{\lambda^*}(v)]$, and which leads to $R=R_{budget}$, then v^* is also an optimal solution to (18). It is well known that when λ sweeps from zero to infinity, the solution to (18) traces the convex hull of the operational rate distortion function, which is a non-increasing function. Hence, bisection or the fast convex search [12] can be used to find λ^*. Therefore, if we can find the optimal solution to the unconstrained problem

$$\min \sum_{i=1}^{N} [D_i(v_{i-a},...,v_i) + \lambda R_i(v_i)], \tag{19}$$

we can find the optimal λ^*, and the convex hull approximation to the constrained problem (18).

To implement the algorithm for solving the optimization problem (19), we create a cost function $C_k(v_{k-a},..., v_k)$, which represents the minimum total rate and distortion up to and including packet m_k given that $v_{k-a},...,v_k$ are decision vectors for packets $m_{k-a},..., m_k$. Clearly,

$$J_\lambda(v) = \min_{v_{N-a},...,v_N} C_N(v_{N-a},...,v_N). \tag{20}$$

The key observation for deriving an efficient algorithm is the fact that given $a+1$ decision vectors $v_{k-a-1}, ..., v_{k-1}$ for packets $m_{k-a-1},...,m_{k-1}$, and the cost function $C_{k-1}(v_{k-a-1},..., v_{k-1})$, the selection of the next decision vector v_k is independent of the selection of the previous decision vectors $v_1, v_2, ..., v_{k-a-2}$. This is true since the cost function can be expressed recursively as

$$C_k(v_{k-a},...,v_k) = \min_{v_{k-a-1},...,v_{k-1}} [C_{k-1}(v_{k-a-1},...,v_{k-1}) + D_k(v_{k-a},...v_k) + \lambda R_k(v_k)]. \tag{21}$$

The recursive representation of the cost function above makes the future step of the optimization process independent from its past step, which is the foundation of dynamic programming.

The problem can be converted into a graph theory problem of finding the shortest path in a directed acyclic graph (DAG) [12]. The computational complexity of the algorithm is $O(N \times |V|^{a+1})$ ($|V|$ is the cardinality of V), which depends directly on the concealment strategy (value of a). It is much more efficient than the exponential computational complexity of an exhaustive search algorithm.

4. EXPERIMENTAL RESULTS

Our simulations are based on MPEG-4 VM 18.0 [13]. However, the inter-mode CAE shape coding has not been considered in this work, because it violates the independently decodable video packet rule by using a special correction strategy, in which a template containing pixels from the previous VOP is used to decode the current VOP.

We first test the end-to-end distortion estimation models on both "Bream" and "Children" video sequences, and on various rate constraints. As shown in Fig. 1, the estimates are compared to the actual decoder distortion averaged over 20 different channel realizations (with different packet loss patterns). It is evident that both model I and mode II provide highly accurate estimate of the decoder distortion.

We encoded the first 30 frames of the "Bream" sequence using the proposed network-adaptive approach and using the non network-adaptive encoding method proposed in [8]. We then transmitted those coded bitstreams over lossy channels with packet loss equals 10% and 20%. Figure 2 shows the R-D curves for such experiments. Clearly, the network-adaptive method has constant gains of around *2 dB* for both channels. In addition, the expected distortion estimated using model I are considerable accurate compared to the averaged distortion over many realizations. Our experiments on "Children" sequences have shown the similar results.

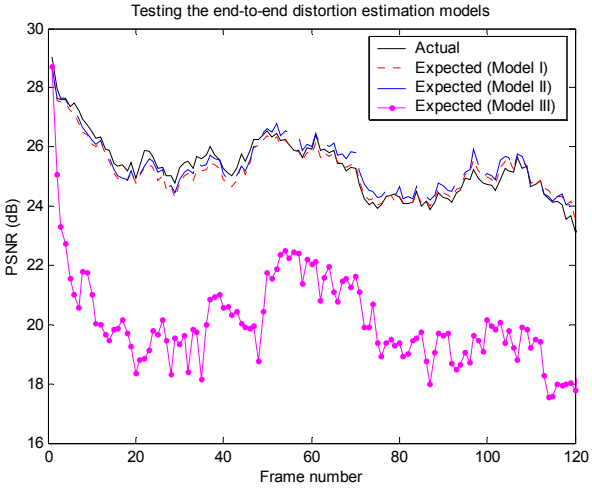

Fig. 1 Estimation performance comparison

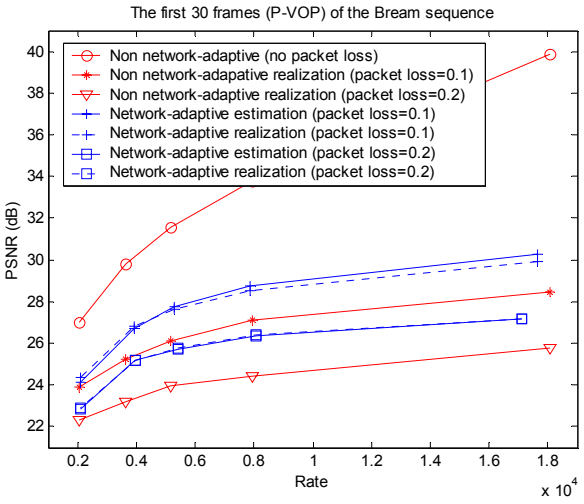

Fig. 2 Compare proposed approach with method in [8]

6. CONCLUSIONS

In this paper, we presented a network-adaptive encoding approach for object-based video. The distortion of decoded video is estimated at the encoder given the packet loss of the transmission channel, and the error concealment strategy at the decoder. The proposed framework optimally chooses the coding parameters of shape and texture to ensure the minimum expected decoder distortion. Lagrangian relaxation and dynamic programming are employed to solve the optimization problem. Experimental results demonstrated that the proposed estimation model is accurate, and that the proposed network-adaptive approach has significant gains over the method without adaptation.

ACKNOWLEDGEMENT

The authors would like to thank Fan Zhai and Yiftach Eisenberg of Northwestern University for their helpful discussion and comments on our work.

REFERENCES

[1] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, "Error resilient video coding techniques", *IEEE Signal Processing Magazine*, vol. 17, no. 4, pp.~61-82, July 2000.

[2] G. Cote, S. Shirani, and F. Kossentini, "Optimal mode selection and synchronization for robust video communications over error prone networks", *IEEE J. Select. Areas Commun.*, Vol. 18, NO. 6, pp. 952-965, June 2000.

[3] R. Zhang, S.L.Regunathan, K.Rose, "Video Coding with Optimal Inter/Intra-Mode Switching for Packet Loss Resilience", *IEEE J. on Selected Areas in Communications*, Vol. 18, NO. 6, pp..966-976, June 2000.

[4] K. Stuhlmuller, N. Farber, M. Link, B. Girod, "Analysis of video transmission over lossy channels", *IEEE J. Select. Areas Commun.*, Vol. 18, NO. 6, pp. 1012-1032, June 2000.

[5] Y. Eisenberg, C. E. Luna, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and transmission power management for energy efficient wireless video communications", *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 12, NO. 6, pp. 441-424, June 2002.

[6] B. Girod, M. Kalman, Y. J. Liang, R. Zhang, "Advanced in channel-adaptive video streaming", *Journal of Wireless Communication and Mobile Computing*, vol. 2, no. 6, pp. 573-584, Sept. 2002.

[7] H. Wang, G. M. Schuster, A. K. Katsaggelos, "Operational rate-distortion optimal bit allocation between shape and texture for MPEG-4 video coding", in *Proc. IEEE International Conference on Multimedia and Expo*, Baltimore, July 2003.

[8] H. Wang, G. M. Schuster, A. K. Katsaggelos, "Object-based video compression scheme with optimal bit allocation among shape, motion and texture", in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[9] A. Leontaris, P. C. Cosman, "Video compression with intra/inter mode switching and a dual frame buffer", in *Proc. IEEE Data Compression Conference*, Snowbird, Utah , March 2003.

[10] H. Yang, K. Rose, "Recursive end-to-end distortion estimation with model-based cross-correlation approximation", in *Proc. IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[11] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources", *Oper. Res.*, vol. 11, pp. 399-417, 1963.

[12] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion based video compression: optimal video frame compression and object boundary encoding*, Kluwer Academic Publishers, 1997.

[13] MPEG-4 video VM 18.0, ISO/IEC JSC1/SC29/WG11 N3908, Pisa, Jan 2001.