OPTIMAL OBJECT-BASED VIDEO COMMUNICATIONS OVER DIFFERENTIATED SERVICES NETWORKS

Haohong Wang, Fan Zhai, Yiftach Eisenberg, and Aggelos. K. Katsaggelos

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

ABSTRACT

In this paper, we propose an optimal unequal error protection scheme for object-based video communications over differentiated services networks. Our goal is to achieve the best video quality (minimum total expected distortion) with constraints on transmission cost and delay. An end-to-end distortion estimation approach for object-based video is proposed, which can be used for different packetization schemes. The problem is solved using Lagarangian relaxation and dynamic programming. Experimental results indicate that the proposed unequal error protection schemes can significantly outperform equal error protection methods.

1. INTRODUCTION

The Differentiated Services (DiffServ) network architecture has recently been proposed by the IETF (Internet Engineering Task Force) for supporting Internet quality of service (QoS) [1]. It allocates resources discriminatorily for aggregated traffic flows according to various QoS service classes. Typically, in using a pricing model, higher QoS classes cost more, but have smaller probabilities of packet loss than lower QoS classes. To better utilize network resources, a sender must classify the priority of each packet based on its relative importance. In this paper we consider packet classification jointly with error resilient object-based video coding.

Recent research, such as [2-4], has shown the potential benefits of video communications over DiffServ networks. In [2] an adaptive packet forwarding mechanism with QoS accommodation was achieved by mapping video packets onto different service levels. In [3] a general cost-distortion optimized multimedia streaming framework over DiffServ network was proposed, which is based on pre-encoded media. In [4] a real-time video transmission problem was studied by jointly considering source coding and packet classification.

Object-based video is one of the most important topics for interactive multimedia applications. In object-based coding, the video is composed of arbitrarily shaped video objects, which are described by their shape and texture. A rate-distortion optimal video encoding scheme was proposed in [5] for object-based video, which enables the optimal bit allocation among shape, texture and motion. In [6] a robust network-adaptive encoding scheme was proposed by jointly considering the source coding, packet loss during transmission, and error concealment at the decoder. However, there is very few reported literature results on object based video communications over DiffServ networks. In this paper, we propose a general cost-distortion optimal unequal error protection scheme for object-based video communications over DiffServ networks.

2. END-TO-END DISTORTION ESTIMATION

Several methods have been proposed to estimate the end-to-end distortion. In [7] a model-based method is proposed to estimate the distortion, and in [8] a recursive distortion calculation approach with pixel level precision is proposed. However, those results only deal with rectangular video frames. In this section, we propose an end-to-end distortion estimation approach for object-based video. We consider two packetization schemes; 1) combined packetization scheme, where shape and texture data are packed in the same packet; 2) separated packetization scheme, where shape and texture are packed in separate packets [9]. In our framework, we assume that each packet is independently decodable.

In object-based video communications, video objects are compressed and transmitted separately. The decoder has the flexibility to decide how to combine the video objects in order to compose the video object plane (VOP). To evaluate the distortion caused by a video object, we assume that the transmitter knows the background on which the transmitted video object will be composed at the receiver. Otherwise, a default background will be adopted. We assume that the concealment strategy is also known at the encoder. We use the expected mean squared error (MSE) as our objective distortion metric. The expected distortion for the *i*th slice can be calculated by summing up the expected distortion of all pixels in the slice,

$$E[D_i] = \sum_{j=iN}^{iN+N-1} E[d_j], \tag{1}$$

where $E[d_j]$ is the expected distortion at the receiver for the *j*th pixel in the VOP, and N is the total number of pixels in the slice. Let us denote by f_n^j the original value

of pixel *j* in VOP *n*, and $\tilde{f}_n^j$ its reconstructed value at the decoder. By definition,

$$f_n^j = s_n^j t_n^j + (1-s_n^j) g_n^j,\ \tilde{f}_n^j = \tilde{s}_n^j \tilde{t}_n^j + (1-\tilde{s}_n^j) g_n^j, \quad (2)$$

and

$$E[d_j] = E[(f_n^j - \tilde{f}_n^j)^2] = (f_n^j)^2 - 2 f_n^j E[\tilde{s}_n^j \tilde{t}_n^j] - 2 f_n^j g_n^j + 2 g_n^j E[\tilde{s}_n^j] + E[(\tilde{s}_n^j \tilde{t}_n^j)^2] + (g_n^j)^2 - 2(g_n^j)^2 E[\tilde{s}_n^j] + (g_n^j)^2 E[(\tilde{s}_n^j)^2] + 2 g_n^j E[\tilde{s}_n^j \tilde{t}_n^j] - 2 g_n^j E[(\tilde{s}_n^j)^2 \tilde{t}_n^j], \quad (3)$$

where s_n^j ($s_n^j = 0$ for transparent or *1* for opaque block; only binary shape is considered in this work) and t_n^j are the corresponding shape and texture component of f_n^j, $\tilde{s}_n^j$ and $\tilde{t}_n^j$ are the corresponding shape and texture component of $\tilde{f}_n^j$, and g_n^j is the background pixel value at the same position. In calculating $E[d_j]$ in Eq. (3), the first and second moments of the reconstructed shape and texture intensity value for the *j*th pixel are needed. In the following paragraph, we show how the first moment can be recursively calculated in time. The second moment is computed in a similar fashion, but omitted here, due to lack of space.

For the separated packetization scheme, since the shape data and texture data are independently transmitted and decoded, we have $E[s_n^j t_n^j] = E[s_n^j] E[t_n^j]$ and $E[\tilde{s}_n^j \tilde{t}_n^j] = E[\tilde{s}_n^j] E[\tilde{t}_n^j]$. This observation enables us to efficiently calculate the expected distortion (3) by recursively calculating the necessary moments for shape and texture independently. That is,

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

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

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

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

where shape and texture are intra coded in the first 2 equations, $\hat{s}_n^j$ and $\hat{t}_n^j$ are respectively the reconstructed shape and texture of the *j*th pixel at the encoder, and ρ_{S_i} and ρ_{T_i} are respectively the probability of loss for the *i*th shape and texture packet. If the *j*th pixel is lost, but its concealment motion vector is available, then it is concealed using pixel *k* in frame *n-1*. In the last 2 equations, shape and texture are inter-coded, the prediction error is $\hat{e}_n^j = \hat{t}_n^j - \tilde{t}_{n-1}^m$, and pixel *j* in frame *n* is predicted from pixel *m* in frame *n-1* (given the motion vector).

For the combined packetization scheme, the task of calculating $E[\tilde{s}_n^j \tilde{t}_n^j]$ in (3) is complicated because of the involvement of some inter-pixel cross-correlation terms [6], which require computing and storing all inter-pixel cross-correlation values for all frames in the video sequence. The amount of computation and storage is infeasible even for a moderate sized frame. In order to reduce computational complexity, a model-based cross-correlation approximation method was proposed in [6] to estimate $E[\tilde{s}_n^j \tilde{t}_n^j]$ in terms of $E[\tilde{s}_n^j]$, $E[\tilde{t}_n^j]$, $E[(\tilde{s}_n^j)^2]$, $E[(\tilde{t}_n^j)^2]$ and standard deviations σ_s and σ_t (see [6] for the detail). Experiments indicate that the proposed distortion estimation approach is accurate for both packetization schemes.

3. PROBLEM FORMULATION

The problem at hand is to choose coding parameters for the shape and texture of a VOP (or frame), so as to minimize the total expected distortion, given a cost constraint and a transmission delay constraint in a DiffServ network environment, that is,

$$\text{Minimize } E[D_{tot}], \text{ s.t. } C_{tot} \le C_{\max} \text{ and } T_{tot} \le T_{\max}, \quad (4)$$

where $E[D_{tot}]$ is the expected total distortion for the frame, C_{tot} is the total cost, T_{tot} is the total transmission delay, C_{max} is the maximum allowable cost, and T_{max} is the maximum amount of time that can be used to transmit the entire frame.

We consider an MPEG-4 compliant object-based video application, where the video is encoded using different algorithms for shape and texture. We use the separated packetization scheme as the default packetization scheme. The coded video frame is divided into 16x16 macro blocks, which are numbered in scan line order and divided into groups called slices. For each slice, there is a corresponding shape packet and a corresponding texture packet. Let *I* be the number of slices in the given frame and *i* the slice index. For each macro block, the coding parameters for both shape and texture are specified. We use μ_{S_i} and μ_{T_i}, respectively, to denote the coding parameters for all macro blocks in the *i*th shape and texture packets, and use $B_{S_i}(\mu_{S_i})$ and $B_{T_i}(\mu_{T_i})$, respectively, to denote the corresponding encoding bit rates of these packets. Let us denote by π_{S_i} and π_{T_i}, the selected service classes for the *i*th shape and texture packet, respectively. Similarly, let $\rho(\pi_{S_i})$ and $\rho(\pi_{T_i})$ denote the corresponding probability of packet loss, and $R(\pi_{S_i})$ and $R(\pi_{T_i})$ the corresponding transmission rate. In our work, we assume that the processing and propagation delays are constant and can therefore be ignored in this formulation. The only delay we are concerned with is the transmission

delay. The total transmission delay per fame is represented by

$$T_{tot} = \sum_{i=1}^{I} [\frac{B_{S_i}(\mu_{S_i})}{R(\pi_{S_i})} + \frac{B_{T_i}(\mu_{T_i})}{R(\pi_{T_i})}]. \quad (5)$$

Let $C(\pi_{S_i})$ and $C(\pi_{T_i})$ denote respectively the transmission cost per bit for the ith shape and texture packets. The total cost used to transmit all the packets in a frame is therefore

$$C_{tot} = \sum_{i=1}^{I} [B_{S_i}(\mu_{S_i})C(\pi_{S_i}) + B_{T_i}(\mu_{T_i})C(\pi_{T_i})]. \quad (6)$$

We assume that the service level can be pre-specified in the service level agreement (SLA) between the Internet service provider (ISP) and the users [1]. Typically a set of parameters is used to describe the state of each service class, including the transmission rate bound and probability of packet loss. In this setting, a cost is associated with each service class as specified in the SLA. By adjusting the prices for each service class, the network can influence the class a user selects. The sender classifies each packet according to its importance in order to better utilize the available network resources. Also, we assume that the transmitter knows the probability with which a packet has arrived at the receiver. Thus, from the point of view of the transmitter, the distortion at the receiver is a random variable. We use the approaches described in section 2 to estimate the expected distortion $E[D_{tot}]$.

4. OPTIMAL SOLUTION

In this section, we present an optimal solution for problem (4). We use the Lagrange multiplier method to relax the cost and delay constraints. The Lagrangian relaxation method leads to a convex hull approximation to the constrained problem (4). Let U be the set of all possible decision vectors u_i for the ith slice ($i=1,2, \ldots, I$), where $u_i = (\mu_{S_i}, \mu_{T_i}, \pi_{S_i}, \pi_{T_i})$. We first define a Lagrangian cost function, which is minimized $J_{\lambda_1,\lambda_2}(u) = E[D_{tot}] + \lambda_1 C_{tot} + \lambda_2 T_{tot}$

$$= \sum_{i=1}^{I} \left\{ E[D_i] + \lambda_1 [B_{S_i}(\mu_{S_i})C(\pi_{S_i}) + B_{T_i}(\mu_{T_i})C(\pi_{T_i})] + \lambda_2 [\frac{B_{S_i}(\mu_{S_i})}{R(\pi_{S_i})} + \frac{B_{T_i}(\mu_{T_i})}{R(\pi_{T_i})}] \right\},$$

where λ_1 and λ_2 are the Lagrange multipliers. It can easily be derived from [10] that if there exists a pair λ_1^* and λ_2^* such that $u^* = \arg[\min_u J_{\lambda_1^*,\lambda_2^*}(u)]$, which leads to $C_{tot}=C_{max}$ and $T_{tot}=T_{max}$, then u^* is also an optimal solution to (4).

Most decoder concealment strategies introduce dependencies between slices. Without loss of generality, we assume that the concealment strategy will cause the current slice to depend on its previous a slices ($a \geq 0$). We define a cost function $G_k(u_{k-a}, \ldots, u_k)$, which represents the minimum total cost, delay and distortion up to and including the kth slice, given that $u_{k-a}, \ldots, u_k$ are decision vectors for the (k-a)th to kth slices. Therefore, $G_I(u_{I-a}, \ldots, u_I)$ represents the minimum total cost, delay and distortion for all the slices of the frame, and thus

$$\min_u J_{\lambda_1,\lambda_2}(u) = \min_{u_{I-a},\ldots,u_I} G_I(u_{I-a},\ldots,u_I). \quad (7)$$

The key observation for deriving an efficient algorithm is the fact that given $a+1$ decision vectors $u_{k-a-1}, \ldots, u_{k-1}$ for the (k-a-1) to (k-1) slices, and the cost function $G_{k-1}(u_{k-a-1}, \ldots, u_{k-1})$, the selection of the next decision vector u_k is independent of the selection of the previous decision vectors $u_1, u_2, \ldots, u_{k-a-2}$. This is true since the cost function can be expressed recursively as

$$G_k(u_{k-a},\ldots,u_k) = \min_{u_{k-a-1},\ldots,u_{k-1}} \{ G_{k-1}(u_{k-a-1},\ldots,u_{k-1}) + \lambda_1 \cdot [B_{S_i}(\mu_{S_i})C(\pi_{S_i}) + B_{T_i}(\mu_{T_i})C(\pi_{T_i})] + \lambda_2 [\frac{B_{S_k}(\mu_{S_k})}{R(\pi_{S_k})} + \frac{B_{T_k}(\mu_{T_k})}{R(\pi_{T_k})}] + E[D_k] \}. \quad (8)$$

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) [10].

5. EXPERIMENTAL RESULTS

The main objective of the experiments presented here is to compare three error protection schemes: (1) UEP-UST, an unequal error protection scheme using the separated packetization scheme, where the shape and texture data are placed in separate packets and therefore can be transmitted over different service channels; (2) UEP-EST, an unequal error protection scheme using combined packetization, where the packets containing both shape and texture data can be transmitted over different service channels; (3) EEP, an equal error protection scheme using combined packetization, where all the packets are transmitted over the same service channel.

Service Class	1	2	3	4
Probability of packet loss	0.2	0.1	0.05	0.01
Transmission rate (Kbps)	315	420	525	630
Cost (microcents/Kilobits)	10	30	60	100

Table 1 Parameters of four service classes

(a) (b)

Figure 1. Reconstructed and composed "Bream" frame *4*

We simulate the DiffServ network as a set of independent time-invariant packet erasure channels, one for each

service class. Packet loss in each class is modeled as a Bernoulli process. In addition, a packet is considered lost if it does not arrive at the decoder on time. A compressed RTP header (*5* Bytes) has been added to each packet [11]. Table 1 shows the parameter setting of the four QoS channels in the experiment. We encode the QCIF "Bream" sequence at *30* fps and transmit it over the simulated DiffServ network. At the decoder, we use two different background VOPs for composition (see Fig. 1). Figure 2 shows the C-D curves for all schemes when we use the backgrounds shown in Fig. 1. As expected, the UEP schemes outperform the EEP scheme in all experiments. UEP-UST outperformed UEP-EST for the first background, and UEP-EST outperformed the UEP-UST for the other background. The reason is twofold: 1) the contrast of the background and foreground in Fig. 1(b) is small, which reduces the importance of shape, and directly reduces the advantage of UEP-UST over UEP-EST; 2) in the UEP-UST scheme, shape and texture are packed in separate packets, which doubles the overhead because an RTP header is required for each packet. Therefore, the UEP-UST approach spends more bits on overhead data than the UEP-EST and EEP schemes. This suggests that the benefits of unequal error protection for shape and texture information in object-based video communications are dependent on the contrast between an object and the background, as well as the amount of overhead required to transmit shape and texture information in separate packets.

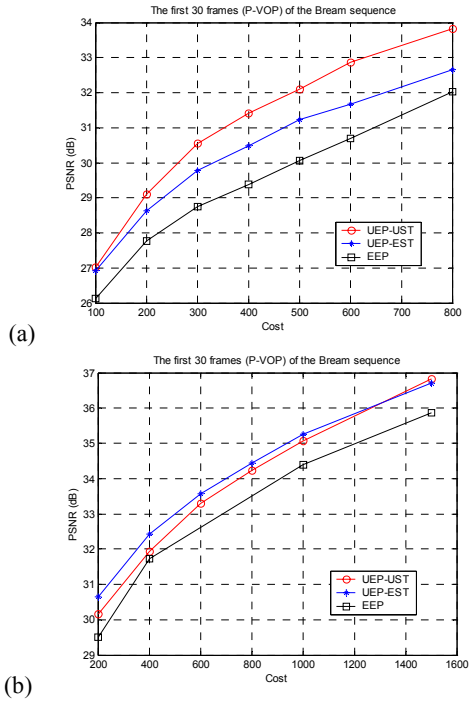

Figure 2. Compare UEP-UST with UEP-EST and EEP; (a) based on Fig.1(b), (b) based on Fig.1(a).

In another experiment, we encode the "Bream" sequence at *15* fps and transmit it over the simulated network, and use the background in Fig. 1(b) at the decoder. The result indicates that the UEP-UST outperformed UEP-EST in most cases. Clearly, reducing the source coding frame rate from *30* fps to *15* fps leads to doubling the bit budget for each frame, thus reducing the impact of the header overhead on performance, and making the benefits of unequal protection on shape and texture salient. It also indicates that the benefits are closely related to parameters such as transmission rate and channel bit rate.

6. CONCLUSIONS

In this paper, we presented an optimal unequal error protection (UEP) scheme for object-based video communications over DiffServ networks, where source coding is jointly designed with packet classification. The performance of two packetization schemes was studied using UEP. Experimental results showed the advantage of the proposed UEP scheme over the equal error protection methods. In addition, the benefits of unequal error protection approaches are highly dependent on the contrast between the object being encoded and the background, as well as the channel bit rate.

REFERENCES

[1] B. E. Carpenter and K. Nichols, "Differentiated Services in the Internet", *Proc. IEEE*, Vol. 90, No. 9, pp. 1479-1494, Sept. 2002.

[2] J. Shin, J. W. Kim, and C. –C. Kuo, "Quality-of-service mapping mechanism for packet video in differentiated services network", *IEEE Trans. Multimedia*, Vol. 3, No. 2, pp. 219-231, June 2001.

[3] P.A. Chou and Z. Miao, "Rate-distortion optimized streaming of packetizaed media", *IEEE Trans. Multimedia*, 2001, Submitted.

[4]. F. Zhai, C. E. Luna, Y. Eisenberg, T. N. Pappas, R. Berry, and A. K. Katsaggelos, "Joint source coding and packet classification for real-time video streaming over differentiated service networks", *IEEE. Trans. Multimedia*, to appear.

[5] H. Wang, G. M. Schuster, and 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, Sept. 2003.

[6] H. Wang and A. K. Katsaggelos, "Robust network-adaptive object-based video encoding", *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004.

[7] K. Stuhlmuller, N. Farber, M. Link, and B. Girod, "Analysis of video transmission over lossy channels", *IEEE J. SAC*, Vol. 18, No. 6, pp. 1012-1032, June 2000.

[8] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience", *IEEE J. SAC*, Vol 18, No. 6, pp. 966-976, June 2000.

[9] H. Wang, Y. Eisenberg, F. Zhai, and A. K. Katsaggelos, "Joint Object-based video encoding and power management for video communications over wireless channels", in *Proc. IEEE International Conference on Image Processing*, Singapore, Oct. 2004.

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

[11] S. Casner and V. Jacobson, "RFC 2508 - Compressing IP/UDP/RTP Headers for Low-Speed Serial Links", http://www.faqs.org/rfcs/rfc2508.html, The Internet Society, 1999.