DYNAMIC RESOURCE ALLOCATION VIA VIDEO CONTENT AND SHORT-TERM TRAFFIC STATISTICS

Min Wu, Robert A. Joyce, and S.-Y. Kung

Department of Electrical Engineering
Princeton University, Princeton, NJ 08544
{minwu, robjoyce, kung}@ee.princeton.edu

ABSTRACT

Dynamic resource allocation is critical in the transmission of VBR video. Our study shows that content is one of the major factors that controls the bandwidth of the video bitstream, yet content alone may not be sufficient in predicting future traffic and in determining how much resource to request. A new framework of traffic prediction is proposed in this paper, taking into account both content features and available short-term bandwidth statistics.

1. INTRODUCTION

Transmission of digital video over bandwidth-limited shared networks will become increasingly important in future Internet and wireless communication. This is a very challenging problem, as we need to cope with the ever changing system parameters, such as the number of data sources and receivers, the bandwidth required by each source stream, and the topology of the network itself. An optimal resource allocation system must dynamically consider global strategies (network-wide management) as well as local strategies (e.g., access control for individual connections). In this paper, we shall focus on the local strategies.

Bandwidth allocation and management for individual streams generally must be done at the "edges" of the network, in order to conserve computational resources on internetwork switches. While offline systems could compute the exact bandwidth characteristics of a stream in advance, in many applications on-line processing is desired or even required. To keep delay and computational requirements low, any information used to make bandwidth decisions should be directly available in the compressed video stream. It is desirable to have a resource management system that can accurately estimate the required bandwidth in real-time.

Resource Renegotiation For VBR Video The resource management of VBR video, which will become increasingly popular due to its consistent perceptual quality, will be studied in this paper. The hallmark of VBR video is that its bandwidth undergoes both short- and long-term changes, in reaction to the complexity—and therefore, compressibility—of the underlying video. Allocating a single constant amount of bandwidth for a VBR stream will yield one of two results: inefficient use of network resources, due to over-allocated bandwidth, or large endpoint (and possibly internetwork) buffers. The bandwidth requests made by the VBR source must be periodically renegotiated in order to obtain high network utilization and low delay.

Conventional approaches renegotiate resources according to changes in bitstream level statistics [1]. The connection between previous traffic and the future traffic are parametrically modeled in work such as [2, 3], and references therein. Content-based approaches have been introduced, motivated by the high correlation between long-term traffic characteristics and video content [4, 5]. We shall show that while content is a major factor in determining the bandwidth, content alone may not be sufficient for predicting future traffic and in estimating how much resource to request. More precisely, we shall look at two issues: (1) at which points the bandwidth should be renegotiated, and (2) how much bandwidth to ask for at any given point.

Bandwidth Renegotiation Points The on-line determination of bandwidth renegotiation points in VBR video generally falls into three categories: deterministic, traffic-based, and content-based. Deterministically setting the renegotiation points is the simplest method: bandwidth requests are made every n frames, where n is an empirically determined balance between request overhead and correlation of frame bitrates. Traffic-based renegotiation, mentioned above, occurs when a stream violates a previously negotiated bandwidth request, or when utilization drops below some level. Although traffic-based renegotiation tracks the real bandwidth more closely, a single complex frame can cause the requested bandwidth to remain elevated for some time. A more "natural" set of renegotiation points is the set of shot boundaries. By studying the bits used per frame in VBR video, one sees that the most dramatic changes occur at the

This work was supported in part by a New Jersey State R&D Excellence Award, and by Mitsubishi Electric ITA.

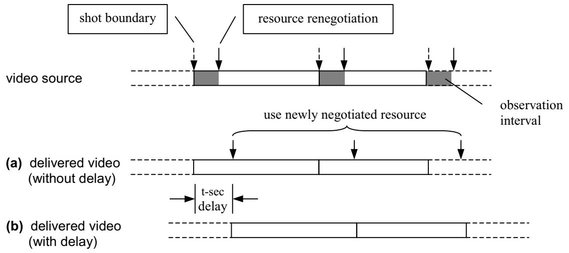

shot boundary
resource renegotiation
video source
observation interval
use newly negotiated resource
(a) delivered video (without delay)
t-sec delay
(b) delivered video (with delay)

Fig. 1. Traffic prediction scenarios with different delay.

beginning of new camera shots [5]. Within a single shot, the traffic characteristics are relatively constant[1].

There exist many approaches to finding shot boundaries in the compressed domain; we use the cut detector described in [6]. This method uses a windowed relative threshold on the sum of absolute pixel differences, and allows for fast, on-line computation of renegotiation points.

Bandwidth Requests Per Interval The next step is to determine how much resource to request at each renegotiation point, without introducing significant delay. For natural renegotiation points such as shot boundaries, previous traffic generally cannot help to determine how much resource to request as the traffic pattern has changed. With the requirement of online processing in mind, one can predict the traffic for the entire shot based on a short observation of the beginning part, as illustrated in Figure 1. Renegotiation is performed after the observation, and if granted, the video will be transmitted using the newly reserved bandwidth. Note that the observation will inevitably introduce a short delay in renegotiation. The video may be transmitted without delay, as in Fig. 1(a); with this approach, unexpected bursty traffic may occur in the shaded period, but it could be smoothed out by a network buffer if t is small. For applications tolerating a short-delay, the video may be transmitted with t-second delay according to Fig. 1(b) so that the video traffic is always under control. In this paper, we shall focus on case (b). Our proposed framework may be extended to case (a).

A content-based prediction approach has been proposed by Bocheck *et al*, consisting of training and testing stages[5]. In the training stage, content features are quantized into a small number of levels (e.g., slow/medium/fast motion), and every possible combination of significant features is labeled as one content class for which the typical traffic pattern is computed. After training, the content class of each shot in the test video is identified by extracting the same features, and the typical traffic pattern of the class is used as the predicted traffic for that shot. However, we notice some potential weaknesses of this approach. First, the specific prediction structure via classification can only feasibly incorporate a limited number of coarsely quantized features; each feature is weighted equally, rather than by its relevance to traffic. Second, prediction based solely on content may not be applicable for video streams produced with different encoding algorithms or parameters. In addition, some useful and readily available information, such as the exact bandwidth statistics of the video in the observation periods, are not incorporated.

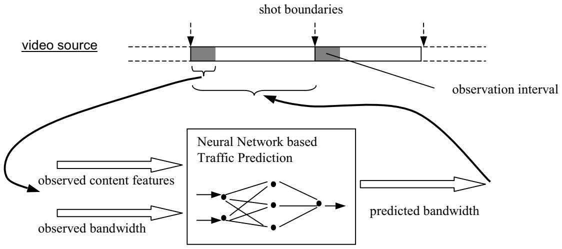


Fig. 2. Neural network based traffic prediction.

2. NN-BASED TRAFFIC PREDICTION

We propose a new framework for traffic prediction, taking into account both the content features and the bandwidth statistics of the video in the observation periods. Prediction results determine how much bandwidth resource to request for the shot. Our goal is not only to enhance the prediction accuracy, but also to evaluate the contribution by various inputs; systems with different tradeoffs for different applications can be constructed based on the evaluation. Although the problem of predicting long-term or future traffic based on short-term traffic can be handled via parametric modeling, it is not easy to come up with a simple and effective parametric model when incorporating content features. For this reason, we use a neural network (NN) to accomplish the prediction task (Figure 2). The input to the neural network consists of content features and traffic descriptors from the observation period. The output is the traffic descriptor for the entire shot. We use a 2-layer neural network and apply a back-propagation approach in supervised training [7].

Content Features Four content features are extracted from the initial frames of a shot to form the content-derived inputs of the neural network. The first, I frame spatial complexity, directly affects peak bandwidth requirements for future I frames in the shot (and indirectly, P and B frames). The spatial complexity can be estimated using a weighted sum of the magnitudes of the AC coefficients for each block.

Motion vectors from adjacent P frames are subtracted to form "acceleration" vectors, the mean magnitude of which forms our second content feature,

$$\overline{\|\text{accel}\|} = \frac{1}{MN} \sum_{i,j} \|\vec{m}_k(i,j) - \vec{m}_{k-1}(i,j)\| \qquad (1)$$

[1] If a shot has a sudden change in content features, the change can be considered a boundary as far as renegotiation is concerned. For simplicity, we will ignore such intrashot "boundaries".

where $\vec{m}_k(i,j)$ is the forward motion vector for macroblock (i,j) of frame k, and M and N are the frame dimensions in macroblocks. A high value of this mean indicates that the motion in the video is not simple, and that the residue frames will become increasingly complex (thus requiring more bits). Similarly, the mean magnitude of the motion vectors offers a measure of how much motion compensation is needed (and therefore, how complex the residue frames are likely to be). Finally, the (spatial) covariance of the x and y motion vector components is measured.

These features were selected from a set of candidate compressed-domain content features according to the following evaluation. In the first step, video shots are classified into k traffic clusters based on a specific traffic descriptor. Classification can be done via K-means, E-M, or other algorithms. In the second step, a consistency measure $\mathcal{C}$ for each feature is computed:

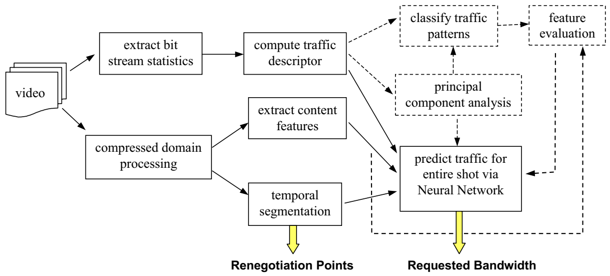
$$\mathcal{C} = \frac{\text{mean inter-class distance}}{\text{mean intra-class distance}} \tag{2}$$

A good feature will have small intra-class distance and large inter-class distance, yielding a large consistency measure[2]. The above four features give the largest $\mathcal{C}$ values (in decreasing order as presented) among about a dozen candidate features. Other novel features could be used, insofar as they give large $\mathcal{C}$ values.

Video Traffic Descriptors Many traffic descriptors have been proposed in the literature. Among them, peak rate and average rate are two very simple ones, but they do not capture the traffic patterns over different time scales. To overcome this problem, Knightly *et al* proposed the D-BIND descriptor for deterministic service, which provides a performance guarantee for the worst case [8]. D-BIND, or the *deterministic bounding interval dependent* model, is essentially a vector containing the maximum allowed arrival rate for various intervals. It is defined as follows: Let $A[\tau, \tau + t]$ be the cumulative number of bits arriving during the t-length interval beginning at time τ. The tightest bound over all time, called the *empirical envelope*, is

$$B^*(t) = \sup A[\tau, \tau + t] \tag{3}$$

A piecewise-linear bounding function B_{W_T} is constructed, where $W_T = \{(q_k, t_k) | k = 1, 2, ..., p\}$ is the vector of bit arrival and interval pairs. Given a set of t_k, the tightest function is denoted $B^*_{W_T}$. The D-BIND descriptor is usually expressed in terms of arrival rates, $R_T = \{(r_k, t_k) | k = 1, 2, ..., p\}$, where $r_k = q_k / t_k$. This descriptor captures both the short-term burstiness and the long-term traffic characteristics of a video segment, while being relatively simple to implement in admission control and policing.



video
extract bit stream statistics
compute traffic descriptor
classify traffic patterns
feature evaluation
principal component analysis
extract content features
compressed domain processing
temporal segmentation
predict traffic for entire shot via Neural Network
Renegotiation Points
Requested Bandwidth

Fig. 3. Overall structure of the VBR resource predictor.

As a proof-of-concept, we use D-BIND descriptors and deterministic service in our tests, though the proposed framework is applicable to others policies. Fixing $[t_1, ..., t_p]$, D-BIND can be described by a vector $[r_1, ..., r_p]$. r_1 through r_4 of the short-term observed traffic are the remaining inputs to our neural network (Fig. 2). When describing the entire shot, the dimensionality of D-BIND is large and the prediction complexity goes up. Such an increase is rather wasteful as there is some redundancy in D-BIND. For example, r_k approaches the mean bitrate for large k. In order to lessen redundancy and reduce prediction complexity, we apply principal component analysis (PCA) to D-BIND and use the first N principal components as traffic descriptors. The neural network will then predict these N values. The overall system structure is illustrated in Figure 3.

3. EXPERIMENTAL RESULTS

We shall demonstrate the performance of our proposed framework by comparing the link utilization with a previously proposed bitstream level approach. We also evaluate the contribution of video content and bandwidth statistics of the short observation periods to traffic prediction. Our experiments are performed on a 13175-frame video (about 7 minutes) digitized from cable television at 30 fps. The video consists of a fast-action documentary segment from “The Oprah Winfrey Show” and clips of the ABC series “The Practice.” It is encoded via MPEG-1 VBR of a fixed quantization step size, with an average bit rate of 2.1Mbps.

Link Utilization The R-VBR scheme, a heuristic renegotation algorithm using D-BIND descriptors, was proposed in [1]. It raises the reserved bandwidth (described by D-BIND) by a factor α when the real bandwith exceeds the current reservation, and lowers it by a factor β when the real bandwidth remains below the reserved resource for K frames. The average R-VBR renegotiation frequency is determined by (α, β, K). In contrast, our proposed scheme uses the shot boundaries, obtained from content-based temporal segmentation, as renegotiation points. 177 shots are identifed in the sample video. Bandwidth reservations are comprised of two D-BIND principal components from our neural network output. The neural network is trained by 100

[2] The consistency measure used here only considers features that are related to traffic descriptors in a monotonic way.

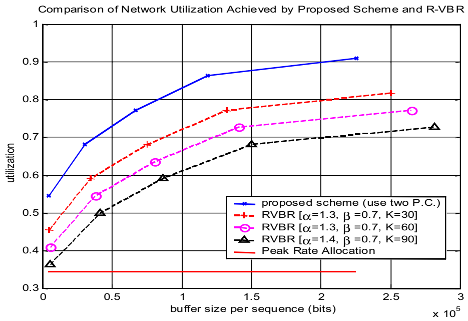


Fig. 4. Network utilization for multiplexed sources.

sweeps with data from the first 50 shots.

Link utilization is obtained by trace-driven simulation, similar to that described in [5]. Multiple video sources, based on the above described sample video but with random starting points, are multiplexed into a T3 line (link speed of 45 Mbps). The simulation results are shown in Figure 4. With the three sets of parameters speficied, renegotiation requests from R-VBR were generated at average intervals of 0.81, 1.54, and 2.23 seconds. The corresponding utilizations are shown in the dashed curves. The horizontal line shows the utilization if the peak bandwidth were allocated to each sequence. The upper solid curve is the utilization of our proposed scheme, which renegotiates once every 2.48 seconds on average. Our proposal outperforms the R-VBR scheme of similar renegotiation frequency by 18%, and by 9% against the R-VBR with tripled reneogtiation frequency.

MSE of Traffic Prediction We compared the prediction MSE under four different strategies, keeping in mind that overestimation of shot D-BIND descriptors could lower utilization, while underestimation would degrade QoS. With respect to renegotiation points, we consider: (A) using equal-length request intervals (one request every 75 frames, which is the average shot length), and (B) using shot boundaries from temporal segmentation. We also consider three different neural network inputs for traffic prediction, all based on statistics from the observation intervals: (I) the 4 content features, (II) the 4-dimensional D-BIND, and (III) both content and D-BIND. The MSE values are shown in Figure 5. Comparing the two leftmost columns, (A-III) and (B-III), we observe that (B-III) gives much smaller MSE, meaning that content-based renegotiation points are by far superior to non-content-based ones. Comparing the three rightmost columns, we see that short-term traffic (B-II) gives better prediction than content features alone (B-I). We also find that using both the content and short-term bandwidth of the observation periods (B-III) is only marginally better than using short-term bandwidth alone (B-II). This implies that most of the useful traffic information in content features is already inherent in very short-term bandwidth statistics.

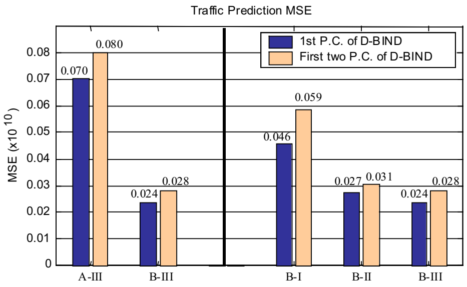


Fig. 5. Traffic prediction MSE using different strategies (D-BIND principal values are on the order of 10^5 bits).

4. CONCLUSION AND FUTURE RESEARCH

We have proposed a new framework for resource allocation of VBR video. According to our preliminary experiments, we found that (1) for determining optimal renegotiation points, a content-based approach is preferred over non-content-based methods; (2) in traffic prediction, using short-term bandwidth statistics as neural network inputs is more effective than using content. Further progress could be made by examining less conservative traffic descriptors and admission control policies, as well as a wider variety of video content and bitrates.

Acknowledgement The authors wish to thank Prof. Bede Liu of Princeton University for his enlightening discussions and support.

5. REFERENCES

[1] H. Zhang and E. W. Knightly, "RED-VBR: A new approach to support delay-sensitive VBR video in packet-switched networks," in *Proc. NOSSDAV*, 1995, pp. 258–272.

[2] S. Chong, S. Li, and J. Ghosh, "Predictive dynamic bandwidth allocation for efficient transport of real-time VBR video over ATM," *IEEE J. Sel. Areas of Comm.*, vol. 13, no. 1, pp. 12–23, 1995.

[3] M. R. Izquierdo and D. S. Reeves, "A survey of statistical source models for variable bit-rate compressed video," *Multimedia Systems*, vol. 7, no. 3, pp. 199–213, 1999.

[4] A. M. Dawood and M. Ghanbari, "MPEG video modelling based on scene description," in *Proc. IEEE ICIP*, 1998, vol. 2, pp. 351–355.

[5] P. Bocheck and S.-F. Chang, "Content-based VBR traffic modelling and its application to dynamic network resource allocation," Research Report 48c-98-20, Columbia Univ., 1998.

[6] B.-L. Yeo and B. Liu, "Rapid scene analysis on compressed video," *IEEE Tr. Circuits and Systems for Video Tech.*, vol. 5, no. 6, pp. 533–544, 1995.

[7] S.-Y. Kung, *Digital Neural Networks*, Prentice Hall, 1993.

[8] E. W. Knightly and H. Zhang, "D-BIND: An accurate traffic model for providing QoS guarantees to VBR traffic," *IEEE Tr. Networking*, vol. 5, no. 2, pp. 219–231, 1997.