A CONTENT BASED VIDEO TRAFFIC MODEL USING CAMERA OPERATIONS

Paul Bocheck and Shih-Fu Chang

Department of Electrical Engineering and Image Technology for New Media Center
Columbia University, New York, N.Y. 10027, USA

bocheck@itnm.columbia.edu, sfchang@itnm.columbia.edu

ABSTRACT

We present our recent work on *content based video* (CBV) traffic modeling of variable bit rate (VBR) sources. CBV approach differs from previous works in that it is not based only on matching of various statistics of the original source, but rather on modeling and mapping its visual content into the corresponding bit rate. We show that CBV model is fully compatible with current and future compression algorithms including those of very low bit rate video coding. We introduce the separation principle between the visual content and encoder dependent bit rate mapping. We construct and verify two experimental CBV models for basic camera operations. The results obtained show that the CBV model can closely match the various statistics of MPEG-2 VBR stream.

1. INTRODUCTION

Over the last few years it become clear that the video component of the emerging multimedia technology will play an important role in future communication and storage systems. This underlines the key importance of video traffic modeling in the development of future multimedia systems [1].

Generally, the VBR video rate depends on *video style*, *content* and *compression technique* used. For example, the different video styles (videophone, movie, news, sport, etc.) can have distinct statistics of scene length, etc. Also, the captured video content, such as scene background, static or moving objects, and the camera motion are very important in terms of resulting bit rate. Because of compatibility reasons, streams with various different compression and coding techniques will most probably be supported in future multimedia systems. Logically, the traffic model taking advantage of video content information should be able to match and predict the video rate better than current models. The content based approach to the modeling of VBR video streams is one of the research directions aimed at exploring this important area.

CBV model generated VBR rate prediction information could be utilized in the end-to-end QOS management, connection admission control and video stream scheduling algorithms to effectively and dynamically allocate resources such as buffers or bandwidth to the video streams. Since these resources are usually allocated on per stream basis, the use of traffic models which assume large number of sources to be multiplexed will not give appropriate result. This apply especially to video servers where only limited number of streams is multiplexed and efficient disk retrieval scheduling, buffer management and network interface scheduling determine the optimal admission strategy and the final cost per stream.

2. VIDEO CONTENT

The typical video sequence is usually described as a collection of independent video shots, also called *scenes*. Each scene by itself is an ordered set of video *frames* depicting a real-time, continuous action [2]. From this perspective, the scene could be seen as a sampled and encoded projection of real-time 3D world.

With the availability of advanced image processing and editing technology, the typical video sequence consists not only of simple static scenes but various visual effects are present in the final video sequences. We say that the scene is composed of several different *epochs*, each containing one of these visual effect primitives (Figure 1). Epoch i of the length τ_i is described in terms of *global epoch content descriptor* Ψ_i. We identified the following visual effects, connected with camera operations to be included in descriptor Ψ_i: static scene, panning, zooming, and translation.

At each epoch we identify a set of *virtual objects* $O = \{O_j: j=1,2,...N\}$. By virtual object we mean the spatially segmented

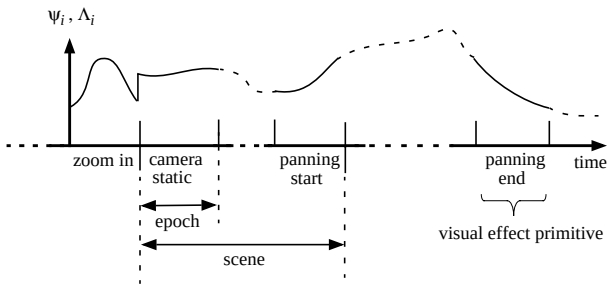

Figure 1. Scene decomposition

frame region having some similar features. The simplest example of such segmentation is fixed block-based frame segmentation used in MPEG-2. More advanced coding techniques such as region-based or object-based are able to decompose the frame sequence into more complicated regions of various shapes and sizes [3]. In many cases the virtual object will resemble the real object but sometimes virtual object will not directly correspond to the real objects. One of the characteristics of virtual object is its size, which is very important in terms of influence of the object on the encoded bit rate. We call the largest object in the epoch the *background object* O_1. At epoch i virtual object j is described in terms of *local epoch content descriptor* $\Lambda_{i,j}$. The following features should be included in the $\Lambda_{i,j}$ descriptor: size, position, shape, complexity (intensity, color, texture), and local motion vector. Note that global descriptor Ψ_i operates on the set O of N virtual objects at epoch i, while local descriptor $\Lambda_{i,j}$ operates on single virtual object O_j only.

3. CONTENT BASED MODEL

Compressed video has a very complicated structure and it is very difficult to model its bit rate accurately [4]. The previous attempts to characterize VBR video streams by various stochastic models have not been fully successful. The main idea behind the CBV modeling is that it targets the natural source of compressed video traffic: the video content.

Since the video stream rate depends on both scene characteristic and compression technique, it is desirable to model them independently. This *separation principle* of the content based video model is schematically depicted on Figure 2. The CBV model consists of two independent parts: (1) *epoch model* and (2) *traffic model*. The epoch model generates the *frame descriptors* F_k on frame by frame basis. It takes into account the cumulative influence of global and local epoch content descriptors. The traffic model is coding standard dependent. It generates the bit rate based on frame descriptors. It also creates the appropriate frame type structure and assembles the stream according to the particular frame ordering.

In terms of targeted applications CBV model can function as a synthetic traffic generator by arranging the epoch content descriptors in the list form called *video content template*. This way, specific video styles may be generated. In the case of real-time video, the epoch content descriptors are generated from video stream by *video content analyzer* or supplied directly by digital camera.

3.1. Epoch Model

The function of the epoch model is to model the video epoch in both spatial and temporal dimension. The output from the epoch model, the frame descriptor, is used subsequently by the traffic model to generate the corresponding bit rate. Based on new global descriptor Ψ_i, the model may change its internal state. The local descriptors $\Lambda_{i,j}$ are used mainly as state parameters. At each state the different mathematical model, best describing the current epoch is selected. Each global descriptor Ψ_i can have several parameters describing the current state.

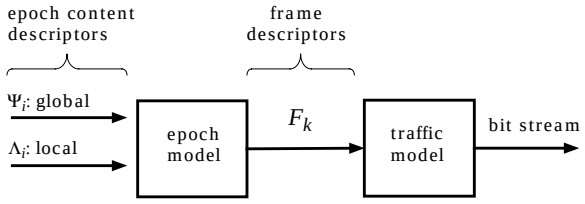

Figure 2. Separation principle of content based model

The parameters included in the frame descriptors F_k generally depend on the coding technique. Currently, for MPEG-2 we identified two components of the frame descriptor k:

$$F_k \rightarrow \{R_k, M_k\} \qquad (1)$$

where R_k and M_k are *frame complexity* of the background object O_1 and *frame motion coefficients* respectively. Then each epoch is modeled using two independent stochastic processes $R=\{R_k;\ k=1,2,...\}$ and $M=\{M_k;\ k=1,2,...\}$. We call R and M *epoch reference processes.* The selection of the actual stochastic process for R and M with corresponding parameters depends on current state of scene content model. We are investigating an innovative approach that the frame complexity and motion coefficient could be approximated by using random walk (RW) with the drift, and auto-regressive AR or ARMA models. More complex models could be chosen, such as DAR or TES if necessary [5].

The following parameters were selected in the Ψ_i, and $\Lambda_{i,1}$ descriptors for each epoch i:

$$\Psi_i \rightarrow \{s, M\}_i \qquad \Lambda_{i,1} \rightarrow \{R\}_i \qquad (2)$$

where s is a camera motion state, M and R are epoch reference processes with the following parameters:

$$M, R \rightarrow \begin{cases} \textit{initial value, step, autocorrelation} & \text{for RW} \\ \textit{mean, variance, correlation} & \text{for AR(1)} \end{cases} \qquad (3)$$

3.2. MPEG-2 Traffic Model

For each single frame descriptor F_k (Eq. 1), the traffic model creates three values $S_{I,k}$, $S_{P,k}$, and $S_{B,k}$ representing the sizes of modeled I, P, and B frames respectively. For simplicity, we assume these values are arranged in the sequences $S_I=\{S_{I,k};\ k=1,2,...\}$, $S_P=\{S_{P,k};\ k=1,2,...\}$, and $S_B=\{S_{B,k};\ k=1,2,...\}$. Based on the specific MPEG-2 standard, the video stream is then properly assembled from S_I, S_P, and S_B sequences to reflect correct frame ordering. Assuming a GOP (Group Of Pictures) of size 12 with 4 subgroups each starting with I or P reference picture, the sequence would be then:

$$S_{I,1},S_{B,2},S_{B,3},S_{P,4},S_{B,5},S_{B,6},...,S_{P,10},S_{B,11},S_{B,12},S_{I,13},...$$

The values of $S_{I,k}$, $S_{P,k}$, and $S_{B,k}$ frames are modeled as follows. Since I frames are intra-frame coded using the DCT transformation, they are directly related to the current frame complexity coefficient, denoted as R_k. Then for each frame k, its compressed I-frame size, denoted as $S_{I,k}$, is modeled as:

$$S_{I,k} = f_I(R_k) = C_I R_k \qquad (4)$$

where f_I is MPEG-2 specific mapping function simplifying to scaling constant C_I.

On the other hand, P and B frames are coded using the motion compensation. Their frame size is the combination of frame complexity (R_k) and frame motion coefficient, denoted as M_k. We approximate the size of P and B frames, denoted as $S_{P,k}$ and $S_{B,k}$ respectively for each frame k as:

$$S_{P,k} = f_P(R_k, M_k) = R_k M_k \quad (5)$$

$$S_{B,k} = f_B(S_{P,k}, M_k) = S_{P,k} M_k + S_{P,k}(1 - M_k)\beta \quad (6)$$

Where f_P and f_B are MPEG-2 specific mapping functions, R_k is frame complexity coefficient, and β=0.5 is a empirically estimated scaling coefficient. Eq. 6 expresses the observed nonlinear dependency of $S_{B,k}$ frame size on M_k. The intuition behind this non-linearity is following: it was observed that for low motion, the advantage of bidirectional motion compensated coding was not significant, while for higher motion the substantial compression gain was observed. Note however, that for very high motion, the B frame size approaches the P frame size (Figure 3).

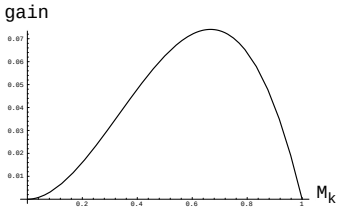

Figure 3. Nonlinear compression gain of B frames relative to size of P frame and its dependency on frame motion coefficient M_k

4. MODEL VERIFICATION

We evaluated our model by comparing the first and second order characteristics of the original and modeled trace. To be able to test our model, we created the video content template of the original source sequence-1, depicted on the left of Figure 4. In our initial experiments we estimated epoch content descriptor parameters as in Eq. 2, 3. We have chosen the random walk for its relative low computation requirements and high correlation between close samples. We approximated the initial frame complexity coefficient, denoted as R_1, and step size, denoted as Δ_R as:

$$R_1 = m_R, \ \Delta_R^2 = \sigma_R^2 / \tau$$

where m_R and σ_R^2 are mean and variance of I frame sizes in the epoch, and τ is the epoch length. For the still camera motion we set δ=0. For each epoch, motion M_1 and motion step size were evaluated similarly by normalizing each frame size in a GOP with respect to its I frame (the first frame in GOP) and taking average and variation of such normalized P frames.

The trace comparison of different frame types is depicted in Figure 4. In the model trace we can identify similar periods corresponding to epochs in the original source. Autocorrelation of I, P and B frames, depicted in Figure 5 is also very similar.

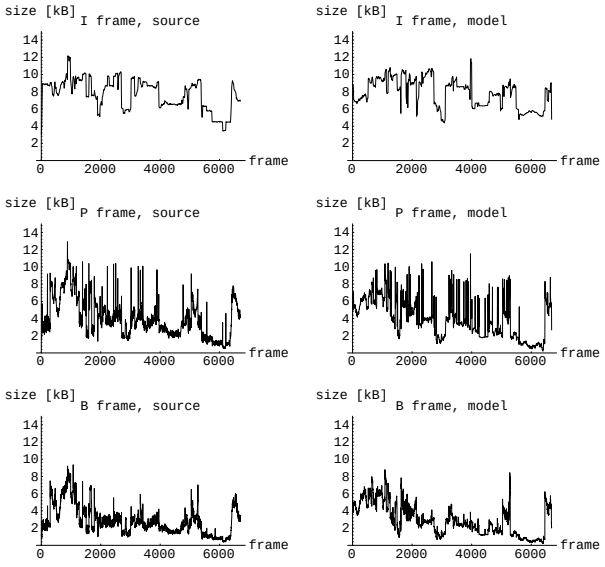

Figure 4. Trace of I, P, and B frames of sequence-1 and model

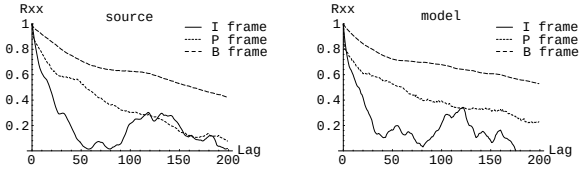

Figure 5. Autocorrelation of I, P, and B sequences for source sequence-1 (left) and model (right)

We can confirm its slow decaying characteristic, as reported in [6]. To further evaluate the model, we simulated the ATM multiplexer loaded with several sources, either real or modeled. The results are depicted in Figure 6. Four cases of 100, 120, 140 and 200 sources correspond to load ρ=0.47, 0.57, 0.66, and 0.95. The bit error rate (BER) of the model closely matched the bit error rate of the source for the low buffer values and all four utilizations. Note that for high multiplexer loads the model estimates the change of the slope of the bit error rate characteristics well. Observed multiple-slope characteristics,

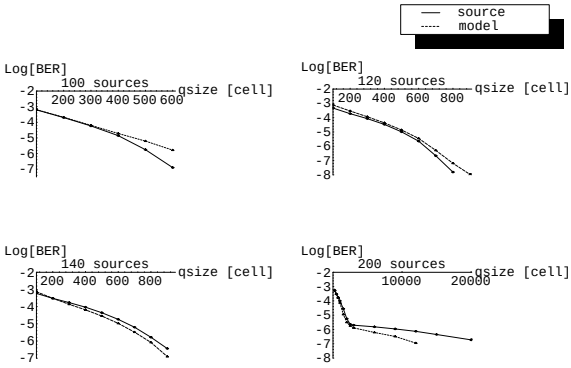

Figure 6. ATM queuing simulation (source sequence-1 and model)

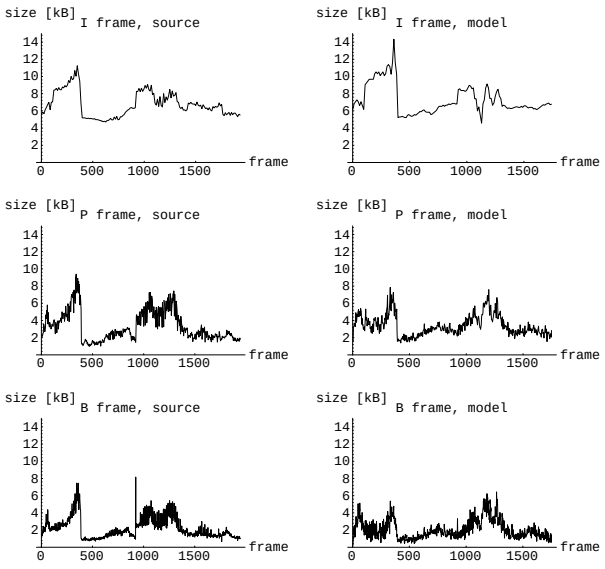

Figure 7. Trace of I, P, and B frames of sequence-2 and model including panning camera operation

appearing in cases of high link utilizations were analyzed in [7].

The content model including the panning camera operation was similarly evaluated and compared with the original sequence-2, selected from five long epochs containing the camera panning operations (Figure 7). Using the global scene motion model [8,9] we extracted the panning speed for each P frame in the sequence (Figure 8). At this stage we used panning speed as a frame motion coefficient M_k. The epoch complexity was modeled similarly as in the previous case using the random walk. The resulting trace, histogram and autocorrelation are depicted on the Figures 7, 9, and 10 respectively. Even though the original trace

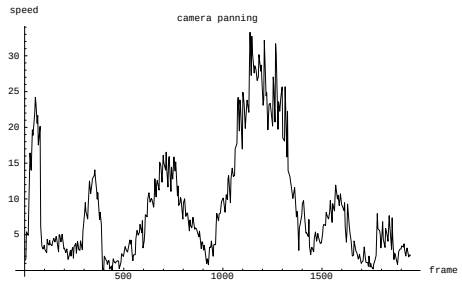

Figure 8. Panning speed of the original sequence-2

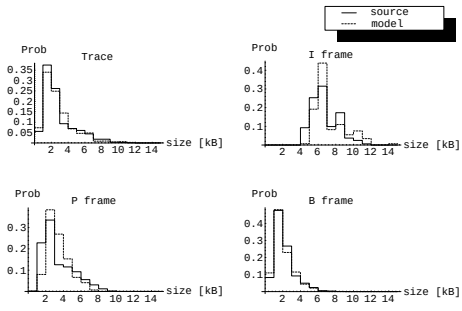

Figure 9. Histogram of the sequence-2 and model.

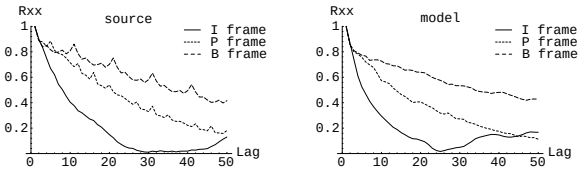

Figure 10. Autocorrelation of the sequence-2 and model.

sequence-2 was relatively short (1944 frames), the modeled stream reassemble the original stream trace well. Also, both original and model histograms autocorrelation coefficients match each other well.

5. CONCLUSION

We presented a framework of video traffic modeling based on the new approach of using the information about the visual content. An experimental model for camera operations was presented and statistically verified. Based on obtained results we argue that this new technique can significantly increase the accuracy of VBR video models. In the future, we would like to incorporate into the current model several other video and object characteristics.

The proposed approach to VBR video traffic modeling also has great synergy with recent work on content-based image/ video search and retrieval [9].

REFERENCES

[1] V. S. Frost and B. Melamed, “Traffic Modeling For Telecommunications Networks”, IEEE Communications Magazine, March 1994

[2] D. Arijon, “Grammar of the film Language”, Los Angeles: Silman-James Press, 1976

[3] H. Sanderson and G. Crebbin, “Image segmentation for compression of images and image sequences”, IEE Proc.-Vis. Image Signal Process., Vol. 142, No. 1, February 1995

[4] P. Pancha and M. E. Zarki, “MPEG Coding for Variable Bit Rate Video Transmission”, IEEE Communications Magazine, May 1994

[5] D.P.Heyman and T.V.Lakshman, “Source Models for VBR Broadcast-video Traffic”, IEEE 1994

[6] J. Beran, R. Sherman, M. S. Taqqu, and W. Willinger, “Long-Range Dependence in Variable-Bit-Rate Video Traffic”, IEEE Transactions on Communications, Vol. 43, No. 2/3/4, February/March/April 1995

[7] A. Baiocci, N. B. Melazzi, M. Listanti, A. Roveri, and R. Winkler, “Loss Performance Analysis of an ATM Multiplexer Loaded with High-Speed ON-OFF Sources”, IEEE Journal on Selected Areas in Communications”, Vol. 9, No. 3, April 1991

[8] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, “Hierarchical Model-Based Motion Estimation“, 2nd ECCV, 1992, pp. 237-252

[9] J. Meng and S.-F. Chang, “Tools for Compressed-Domain Video Indexing and Editing, “SPIE Conference on Storage and Retrieval for Image and Video Database, Vol. 2670, San Jose, Feb. 1996