

A Content Based Approach to VBR Video Source Modeling

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 are presenting a new framework of content based video (CBV) modeling, suitable for scalable variable bit rate (VBR) video traffic. Our approach differs from previous works in that it is not based only on matching of various statistics of the original source, but rather it includes characterization and classification of the video content. We show, that CBV model is fully compatible with existing compression algorithms and also with the current research directions in the very low bit rate video compression. As an example, we introduce MPEG-2 VBR video traffic model based on the ordered list of scene descriptors forming the template of the video stream. The model is verified in terms of trace, first and second order statistics. Results of simulation of ATM network interface multiplexer indicate that the CBV model can generate video traffic which closely matches the important performance parameters and statistics of real VBR MPEG-2 stream. Based on our results we argue that introduction of video content into the modeling of video streams allows us to specify the different classes of important features of video sources resulting in more accurate VBR video traffic model. The CBV model is able to generate the VBR traffic with specific video content and therefore it can be successfully used in the performance estimation and simulation of storage retrieval, buffering, scheduling and admission control of video servers and networks.

I. Introduction

In order to make future communication and storage systems able to support VBR video, it is very important to study and understand its characteristics. By determining new distinctive features of VBR traffic, its source and influence on system performance we can

improve system design of future multimedia and video systems. VBR video has also dynamic storage and retrieval requirements compared to constant bit rate video (CBR). By dynamic we mean that amount of information per any time unit varies over the time. Note that even though the CBR could be considered variable bit rate on the frame basis, its rate is more or less constant assuming the group of pictures (GOP) as the time unit.

Generally, the amount of required information per frame depends on video content and compression technique. We can expect that the future communication networks will transport video streams of various styles: videophone, movie, news, sport, etc. All these video styles have also different statistics: for example, scene length distribution. Also, even though we expect further improvements in digital compression technology, streams with several different coding standards will most probably be supported because of compatibility reasons. Therefore, the VBR traffic models being able to capture both the real video styles and various compression techniques are needed. The content based approach to the modeling of VBR video streams is one of the research directions aimed at solving this problem.

Being able to characterize the VBR video stream also reflects as the ability to predict in some probabilistic sense its future behavior. Such information could be utilized in admission or scheduling algorithms to effectively allocate dynamic resources such as buffers or bandwidth to 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.

II. Video semantics

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 [1]. From this perspective, the scene could be seen as a sampled and encoded projection of real-time 3D world. With availability of advanced image processing and editing technology, the typical video sequence consists not only of static real-world scenes: various artificial effects such as camera operations (camera movement, object following, zooming, panning), picture in picture and graphics embedding, etc. are present in the video sequences. Also, the assumption of scene independence

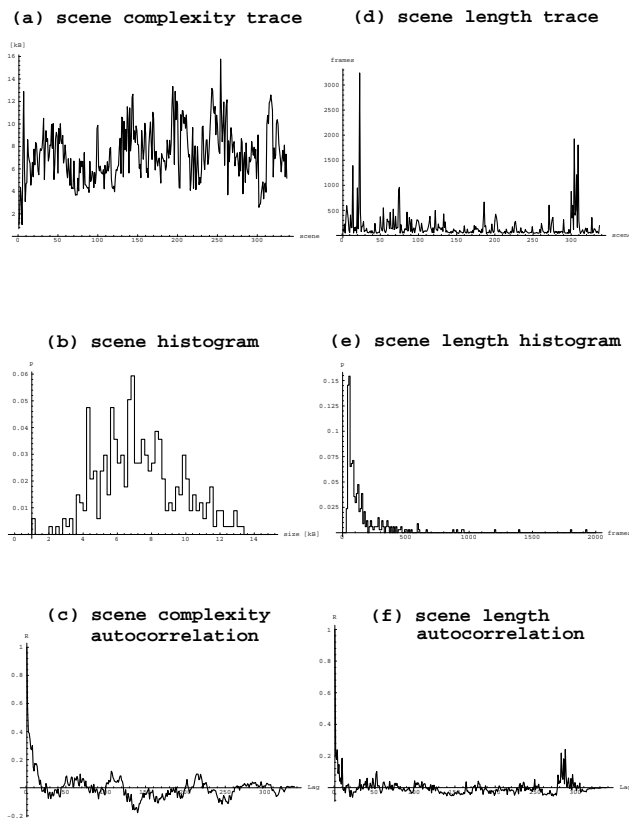


Figure 1. Scene complexity and scene length statistic of movie Forrest Gump (57000 frames)

does not seem to hold in most video sequences. The natural explanation is that the sequential real-time scenes are usually captured in the same visual background such as room, forest, etc. Therefore, the scenes tend to be correlated. This can be observed on the scene complexity

and scene length statistics which are shown on Figure 1. For this purpose we defined the scene complexity as the size of the first intraframe DCT-coded frame of the scene. Figure 1a depicts the scene complexity trace indexed by scene number. The Figure 1b depicts its histogram which can be approximated by normal distribution. Figure 1c depicts its corresponding normalized autocorrelation. We can readily see that scenes are in fact correlated. This intra-scene correlation can be one of reasons of observed phenomena of long-term correlation of VBR traffic.

Another important observation is directly coming from the scene length trace, depicted on Figure 1d. Observed heavy tail of the scene length histogram in Figure 1e can also explain the long term autocorrelation and fractal characteristics of VBR video. Note the bell-like shape of marginal distribution: there is a low probability of having very short scenes, while the probability of having very long ones decays relatively slow. This distribution can be very well matched by lognormal distribution. On the contrary to the scene complexity the scene lengths seem to be not correlated (see Figure 1f).

Each scene can be further decomposed into a set of virtual objects $O = \{O_i; i=1,2,\dots,N\}$. By virtual object we mean spatially or temporary segmented region within the frame sequence having some similar features such as color, texture or motion. In some cases the virtual object can reassemble the real objects but sometimes virtual object will not directly correspond to the real objects. The simplest example of fixed segmentation is block-based segmentation of MPEG-2. More advanced coding techniques such as region-based or object-based are able to decompose the frame sequence into regions of various shapes and sizes. Each of the virtual objects is then associated with a set of object descriptors: object model (2D or 3D), shape, size, color, complexity (texture), activity and 2D or 3D motion. As we already mentioned, the proposed decomposition is similar to techniques used in second generation low bit rate video coding techniques [11,12]. Note, that by definition, the scene segmentation is based on non-continuous sample paths of scene or object characteristics. Therefore, we can assume, that during the scene duration segmented objects will change their corresponding parameters only continuously.

Some scene operations such as camera movement, zooming or panning affect some descriptors of all objects. For example, during the camera movement all virtual objects will be offset by global scene motion.

III. Content based approach

The VBR coded video has a very complicated structure. The previous attempts to characterize VBR video streams by various stochastic models without understanding the nature of the coding process have not been fully successful and have only limited usability. Such models focussed only on matching the trace statistics or queuing analysis. The introduction of video content into the VBR video modeling allows us to create realistic VBR video traffic based not only on global statistical parameters but also on the video style, scene or object description. Since the particular video stream is a combination of scene characteristic and coding algorithm specific mapping, it is desirable to separate them by identifying the independent descriptor variables characterizing the scene, frame, and objects.

We propose the following technique which allow us to accommodate both current and future coding algorithms. The new technique of content based modeling is based on the following principle. The model consists of two independent parts: (1) general scene/object model and (2) mapping function. This separation principle is schematically depicted on Figure 2. The scene model generates the frame content descriptors on frame by frame basis. It takes into account the cumulative influence of global scene operators (zooming, panning, etc.) and descriptors of all objects on each frame descriptor. The mapping function is coding standard dependable. It takes into account the specific coding standard used and maps video content to actual output bit rate. For example, it creates the appropriate frame type and assemble the stream according to particular frame ordering. In the case of MPEG-2, this corresponds to I, P and B specific frame ordering. Such division is very natural and better facilitate the understanding of the VBR video stream behavior.

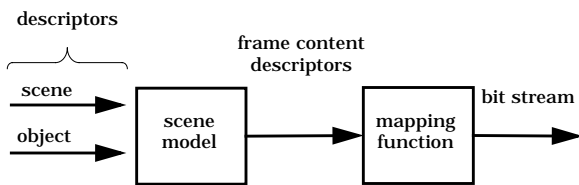


Figure 2. Separation principle of content based model

The selected parameters included in the scene/object or frame descriptors generally depend on the coding technique. For example, for MPEG-2 model we propose

the following parameters to be included in the scene descriptor Ψ :

$$\Psi = \{\tau, \chi, \mu\} \quad (1)$$

where τ is scene length, χ is scene complexity and μ is scene motion. Detailed discussions about particular descriptor parameters are included in section IV.

We are investigating an innovative approach that scene complexity and motion could be approximated by using random walk or auto-regressive AR model [2]. This approach comes naturally from the definition of a scene and its high temporal frame correlation. The changes in the bit rate during the scene are due to object complexity changes, object movement, or due to camera operations (affecting all objects). Each frame in the scene sequence can be decomposed into two components: the scene complexity and object movement. In the special case of MPEG-2, the frame complexity could be estimated from the DCT coefficients of I frames and the object movement speed from the motion vectors and residual error of motion compensated P and B frames. Therefore, the scene could be described by two stochastic processes R and M . The process R corresponds to scene complexity and M corresponds to the motion. Note that I frame sequence, which is intra-frame coded using DCT transform depends on R , while both P and B frame sequences depend on R and M . To simplify the MPEG-2 modeling, which is based only on static macroblock-based motion compensation and also taking into account individual objects, we can divide the scenes into collection of sub-scenes, corresponding to individual object activity, and treat them similarly as scenes itself. In this case, the difference is only in time scale. Example of different scenes with various levels of activity extracted from source trace is in Figure 6.

IV. MPEG-2 VBR video stream analysis

Figure 3 depicts three streams of MPEG-2 VBR coded video sequence corresponding to three scalable layers: low, medium and high. This video trace was created using Columbia University's MPEG-2 software encoder. The trace consists of approximately 7000 frames of the movie Ben Hur. A GOP (Group Of Pictures) of size N consists of subgroups of M pictures starting with I or P reference picture. We selected $N=12$ and $M=3$ and the following scalability options: spatial for the second layer and SNR for the third layer. We found that subjective visual quality of the low and medium layers is comparable to the VHS while high layer is comparable to S-VHS quality. Since each layer is generated from the

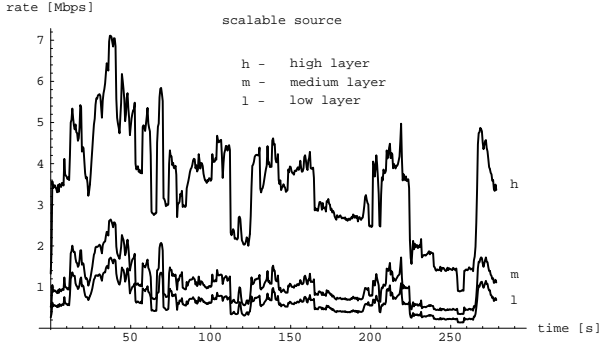


Figure 3. Scalable MPEG-2 VBR trace

same video source, the same scene descriptor would be used to model different layers with different model parameters. Scalable stream characteristics are given in following Table 1.

Table 1:

	\bar{x}	σ^2	$\frac{\max\{x_i\}}{\bar{x}}$
Low	0.6632	0.1818	3.6494
Medium	1.0870	0.4001	3.5083
High	3.5176	2.1060	2.5445

In this paper we are focusing on the low layer only, but we expect to apply our results to the higher layers.

The left side of Figure 4 depicts trace resulting from separation of different frame types: I, P, and B. I frame trace can be described as collection of time periods with approximately constant or continuously varying bit rate. These time intervals match observed scenes or they correspond to sudden changes inside the scene: the object influence. Since I frames are intra-frame coded using the DCT transformation, we can expect that for each frame k , its compressed I-frame size, denoted as $S_I=\{S_{I,k}; k=1,2,\dots\}$, directly corresponds to scene complexity, denoted as R_k . Namely, for MPEG-2 holds the following: $S_{I,k}=f_I(R_k)=R_k$, where f_I is coding algorithm specific mapping function. On the other hand, P and B frames are coded using the motion compensation. Their frame size is the combination of scene complexity (R_k) and motion, denoted as M_k for each frame k . High peaks observed on the trace of P and B frames are results of scene changes or instantaneous object changes during the scene. The

size of P or B frames depends on how well the motion prediction algorithm can be applied to the scene. Therefore, defining scene motion coefficient as M_k , $0 \leq M_k \leq 1$, we can express the size of P frames as $S_{P,k}=f_P(R_k, M_k)$ and the size of B frames as $S_{B,k}=f_B(R_k, M_k)$ for each frame k . Where f_P and f_B are coding algorithm specific mapping functions, $S_P=\{S_{P,k}; k=1,2,\dots\}$ and $S_B=\{S_{B,k}; k=1,2,\dots\}$ are sequences of P and B frames. Note that based on N and M parameters and MPEG-2 frame ordering, resultant MPEG-2 sequence has to be assembled from S_I , S_P and S_B sequences.

V. A content-based VBR source model for MPEG-2

We propose the following model for generating MPEG-2 sequences. Each scene is generated independently of each other. The scene complexity $R=\{R_n; n=1,2,\dots\}$ and motion $M=\{M_n; n=1,2,\dots\}$ are independent stochastic processes. We call R and M intra-scene reference processes for complexity and motion respectively. In terms of selecting the actual stochastic process for R and M , we have chosen the random walk for its relative low computation requirements and high correlation between close samples. More complex models could be chosen, such as AR, DAR or TES if necessary [2, 3, 4, 5, 6]. The following were selected as

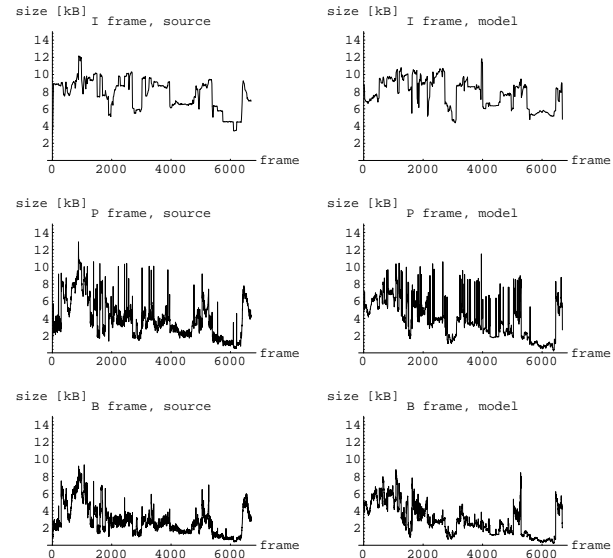


Figure 4. Trace of I, P, and B frames of real (left) and model (right) sequences

scene descriptors (Eq.1): $\tau=\{sceneLength\}$, $\chi=\{mean, stepSize\}$, and $\mu=\{mean, stepSize\}$. Means of complexity and motion are normalized to indicate 0 as lowest frame size and no motion, and 1 to indicate maximum frame size or motion. The following is the detailed description:

1. Intra-scene complexity reference frame sequence R is modeled as random walk $R=\{R_k; k=1,2,\dots\}$ with mean R_I and step size D_R .
2. Intra-scene motion reference sequence M is modeled as random walk $M=\{M_k; k=1,2,\dots\}$ with mean M_I and step size D_M .
3. S_I , I frame sequence defined on set of all frames is then in the form $S_I = \{S_{I,k}=R_k; k=1,2,\dots\}$. Namely, the I frames are modeled by intra-frame complexity only.
4. S_P , P frame sequence defined on set of all frames is approximated as follows:
 $S_P=f_P(R,M)=\{S_{P,k}=R_kM_k; k=1,2,\dots\}$.
 Namely, motion compensation effect is modeled by the multiplicative motion factor M_k .
5. S_B , B frame sequence, defined on set of all frames is approximated as follows:
 $S_B=f_B(R,M)=\{S_{B,k}=S_{P,k}M_k+S_{P,k}(1-M_k)b, k=1,2,\dots\}$, where $b=0.5$ is MPEG-2 specific coefficient obtained empirically from source trace. The second term in MPEG-2 specific approximation of f_B is due to fact that for the same motion, compression achieved by B frame referenced to P frame is much lower then compression achieved by P frame referenced to I frame. Also, for high motion, the advantage of B frame coding is not substantial.
6. The resultant sequence T is obtained by sub-sampling of S_I , S_P , and S_B and correct ordering of I, P, and B frames according to MPEG-2 standard. In case of $N=12$, $M=3$, the sequence would be as follows:
 $S_{I,1}, S_{B,2}, S_{B,3}, S_{P,4}, S_{B,5}, S_{B,6}, \dots, S_{P,10}, S_{B,11}, S_{B,12}, S_{I,13}, \dots$

7. The effect of scene change is modeled as follows: if the scene change occurs on I frame, no change to the sequence is made. If the scene change occurs during the P frame, P frame will be replaced with corresponding $S_{I,k}$ value. If the scene change occurs on B frame, both consecutive B frames are replaced with corresponding $S_{P,k}$.

To model intra-scene objects, each scene is further divided into the set of sub-scenes and modeled using the above algorithm.

In our current experiment, scene cut locations are identified manually. Automatic scene change detection methods using compressed data only have been proposed in [13] and will be incorporated into our system. We assume independent content among different scenes at this stage. But as we have shown in section II, there exists some long-term scene dependence. We are incorporating this relationship into the model.

VI. Statistical verification and simulation

We evaluated our model by comparing the first and second order characteristics of the original and modeled trace [7]. To be able to test our model, we created the scene descriptor table corresponding to the original source. First we normalized the source such that minimum and maximum frame sizes corresponded to 0 and 1 respectively. Scene complexity and motion activity can be estimated by examining the DCT coefficients and motion vectors of the compressed streams. Access to the low level coefficients requires some processing of the

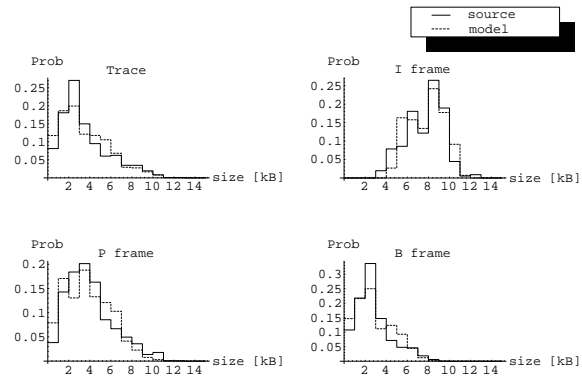


Figure 5. Histogram comparison of real and model sequences

streams. For time efficiency, our initial experiments estimated them as follows. Scene complexity $R_I = m_R$ and step size $\Delta_R^2 = \sigma_R^2 / \tau$, where m_R and σ_R^2 are mean and variance of size of I frames in the scene, and τ is the scene length. For each scene, motion M_I 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 on Figure 4. On the model trace we can identify similar periods corresponding to scenes in the original source. On traces of P and B frames we can identify high peaks as the result of scene changes, very similar to the original trace. Examples of different scenes are depicted in Figure 6. Three different scenes are shown: first with

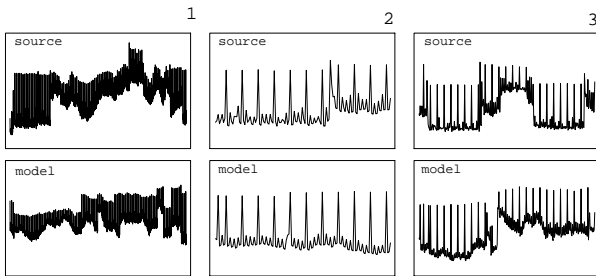


Figure 6. Trace of different scenes. At the first row are scenes from source, below are corresponding scenes from model

high motion, second and third with intra-scene object. As depicted on Figure 5, the histograms of trace and model match each other closely. Also, autocorrelation of I, P and B frames, depicted on Figure 7 is very similar. We can confirm its slow decaying characteristic, as reported in [8]. To further evaluate the model, we simulated the ATM multiplexer loaded with several sources, either real or modeled. The results are depicted on Figure 8. Four

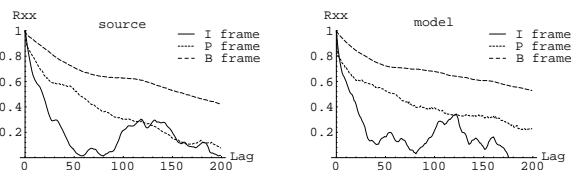


Figure 7. Autocorrelation of I, P, and B sequences for real (left) and model (right) sequences

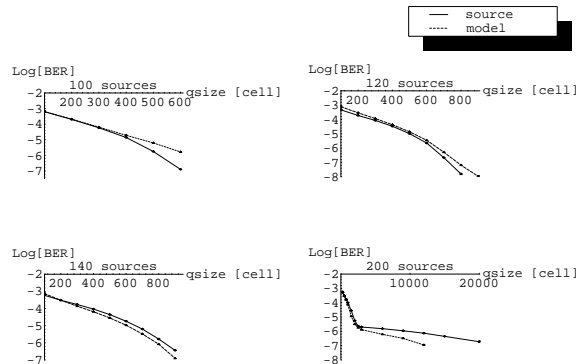


Figure 8. ATM queuing simulation

cases of 100, 120, 140 and 200 sources correspond to load $\rho = 0.47, 0.57, 0.66,$ 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. The differences at high buffer values (for low utilization, model slightly overestimates the BER, while for high utilization the BER is little underestimated) may be due to non-uniform spacing of sources during the frame interval (we used uniform random generator for the starting time of each source within the frame interval). 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, appearing in cases of high link utilizations were analyzed in [9].

VII. Conclusion and continuing work

We presented a new framework of content based approach to modeling of VBR video sources. This approach is based on the fact that typical video compression algorithms such as MPEG-2 in order to encode the original video stream have explored a variety of high-correlated features in both spatial and temporal scale. Identification of various feature descriptors and their influence on the final bit rate is the key point in the content-based video modeling. We outlined the separation principle which isolates the independent video features from their specific use during the compression. Such separation allows us to accommodate both current (MPEG-2) and future coding algorithms (e.g. region-based, object-based or model-based coding) into the same framework by specification of only algorithmic, coding specific part of the model. The video feature part of the model could remain the same. The classification of

higher level features such as camera movement, zooming, panning, etc., and understanding of their projection into the resulting video trace would allow us to better predict behavior of the video stream in real time applications. In this paper, an experimental model, based on subset of such features, the scene duration, complexity and motion was presented and verified in terms of first and second order statistics. The obtained results show that such simple scene descriptor model can very well match VBR video statistics of real sources. In order to further compare the model characteristics with the real trace, we simulated ATM network multiplexing 100 to 200 video sources in the burst mode. The results indicated, that for the burst mode of independent sources, the results match each other well for the low values of buffer sizes. Because of the relatively small size of the video trace, the difference at high buffer size values, which is less than the order of magnitude, depends on the phase sources are multiplexed during the frame period.

In the future, by identifying and separating other independent video features, we would like to accommodate new content-based video coding techniques into the same framework. The proposed approach to VBR video traffic modeling also has great synergy with recent work on content-based image/video search and retrieval [10].

Acknowledgments

This work was supported in part by the National Science Foundation under a CAREER award (IRI-9501266) and the ADVENT project of Columbia University.

References

[1] D. Arijon, "Grammar of the film Language", Los Angeles: Silman-James Press, 1976

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

[3] K. Chandra, A. R. Reibman, "Modeling two-layer MPEG-2 Video Traffic", "AT&T Bell Laboratories", 1995

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

[5] D. Geist and B. Melamed, "TESStool: An Environment for Visual Interactive Modeling of Autocorrelated Traffic", IEEE ICC'92

[6] P. Pancha and M. El Zarki, "MPEG Coding For Variable Bit Rate Video Transmission", IEEE Communications Magazine, May '94

[7] D. M. Lucantoni, M. F. Neuts, and A. R. Reibman, "Methods for Evaluation of VBR Video Traffic Models", IEEE/ACM Transactions on Networking, Vol. 2, No. 2, April 1994

[8] 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

[9] A. Baiocchi, 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

[10] S.-F. Chang, "Compresses-Domain Techniques for Image/Video Indexing & Manipulation", Invited Paper, IEEE ICIP'95, Washington DC, October 1995

[11] F. Eryurtlu, A.M. Kondo, and B.G.Evans "Very low-bit-rate segmentation-bases video coding using contour and texture prediction", IEE Proc.-Vis. Image Signal Process., Vol. 142, No. 5, October 1995

[12] 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

[13] 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.