Content-Based MPEG Video Traffic Modeling

Ali M. Dawood and Mohammed Ghanbari, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a video model to generate VBR MPEG video traffic based on the scene content description. Long sessions of nonhomogeneous video clips are decomposed into homogeneous video shots. The shots are then classified into different classes in terms of their texture and motion complexity. Each shot class was uniquely described with an autoregressive model. Transitions between the shots and their durations have been analyzed. Unlike many classical video source models, this model may be used to generate traffic of any type of video scenes ranging from a low complexity video conferencing to a highly active sport program. The performance of the model is evaluated by measuring the mean cell delay when the generated video traffic is fed to an ATM multiplex buffer.**

***Index Terms*—Image classification, image content, image motion analysis, image texture analysis, video modeling.**

I. Introduction

THE VALIDITY of network simulations depends on the accuracy of the traffic model and in particular that of video traffic which generates most of it. Video model is an aid for designing and testing future communication networks that will carry multiplexed video traffic. It is an essential tool in estimating many networking issues such as the delay arising from statistical multiplexing, the buffer size required for multiplexing and the bandwidth required for carrying video [1]. The performance of the video model in simulating networks depends on how close the video model is tuned to the real video traffic. Real video traffic can definitely be used for this purpose, but it limits the network simulation to the available pre-encoded traffic only. A synthetic video model has the advantage that it can be easily adjusted to model different kinds of video sources; hence, image sequence storage and encoding hardware are no longer needed, not to mention the time saving. Generally, video models must be versatile, in characterizing video traffic, to represent a large range of video sources by varying only a few parameters.

In the past decade, several models for variable bit rate (VBR) video have been proposed. A good survey of video models can be found in [2]–[5]. The parameters of these models are normally obtained by matching certain statistical characteristics of a real video sequence and the model under consideration. Particular attention is given to matching the mean, variance and correlation between the bits-per-frame of the model and real traffic.

Maglaris *et al.* have used a 10-s-long sequence of video and proposed two models for video data [6]. The first model is a first order auto-regressive (AR) model which has proven to be useful for queuing simulation but is not appropriate for queuing analysis due to its mathematical complexity. The second model is a discrete-state continuous-time M-state Markov chain, which is more suitable for queuing analysis. Both models aimed to produce video traffic similar to a video-phone source showing the head and shoulders of a talking person. Sen *et al.* [7] extended this model in order to model video sources with scene changes. Their model basically consists of two M-state Markov chains with transitions between the states of each chain and between the two chains. Each chain represents one activity level of a video. Grunenfelder *et al.* [8] used an auto-regressive with moving average (ARMA) process in order to model the bit rate of a conditional replenishment (CR) video codec without a frame buffer at the cell level. The parameters of the ARMA process are determined from the coded traffic statistics.

Heyman *et al.* [9] used a 30-min-long video-conferencing sequence with no scene change. Their main conclusions were 1) the number of ATM cells per coded frame follows a gamma distribution, 2) a second order AR model fits the data well, but it does not produce enough large values to be a good model for traffic studies, and 3) a two-state Markov chain model does not provide enough accuracy in the model and larger number of states is required. Hughes *et al.* [10] have proposed an M-state discrete-time discrete-state Markov model for the CR video traffic. The cell per frame was approximated by a truncated geometric distribution. An additional state was added to the single Markov chain to represent the scene change. The Markov chain was also extended to $N \times M$ states in order to describe the aggregate traffic of N identical statistically independent multiplexed sources. A three-state Markov model has been proposed by Shim *et al.* [11] to model the bit rate of a video source with scene change. State 0 is modeled as a two AR process (2-AR). A transition to state 1 represents a scene change and the same with state 2, as it is assumed that the effect of the scene change lasts for two frames. The model was proposed to discuss the issue of call admission control (CAC) in ATM networks.

These models were derived for non-MPEG video. MPEG video models should include the regular group of picture (GOP) structure. Pancha *et al.* [12] have studied the statistics of MPEG where they observed that gamma distribution fits the empirical distribution of the packetized bits/frame extremely well for video sources with low bit rates. Their

Manuscript received August 26, 1998; revised November 8, 1998. The associate editor coordinating the review of this paper and approving it for publication was Dr. Samir Kapoor. A. Dawood was sponsored by Etisalat College of Engineering, a division of Emirates Telecommunications Corporation (ETISALAT), United Arab Emirates.

The authors are with Multimedia Communication Research Laboratory, Department of Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ, U.K. (e-mail: ghan@essex.ac.uk).

Publisher Item Identifier S 1520-9210(99)01583-7.

results show that the distribution of the ATM cells per frame for high-quality video does not appear to follow any smooth distribution. They extended their studies [13] to include the traffic characterization of MPEG video source at the frame, slice, and macroblock levels. Heyman *et al.* [14] used a 10-min MPEG-2 VBR coded movie sequence. They showed that the number of bits/frame distribution of I-frames has a lognormal distribution and its autocorrelation follows a geometric function. They concluded that there is no specific distribution that can fit P- and B-type frames. Wu *et al.* [15] proposed the correlated state transition (CST) scheme for the modeling of MPEG video at slice level. CST demonstrated a good performance in capturing the autocorrelation of the data.

Frame-level MPEG modeling was proposed by Krunz *et al.* [16]. The model was based on the decomposition of an MPEG coded stream into I-, P-, and B-frames. Each frame type was modeled separately. The lognormal distribution was found to be the best match for all of the three types. Then, frames are periodically generated according to the GOP structure of the MPEG. Although the autocorrelation was calculated for each frame type, it was not considered in modeling.

Ni *et al.* [17] proposed a mini-source based discrete-time Markov modulated deterministic process (D-MMDP) to model the macro-frame (M-frame equivalent to a GOP) smoothed VBR MPEG-2 traffic. Although this has the advantage of eliminating the cyclical variation in the MPEG video pattern, it is at the expense of coarsening the time scale. Typically, a GOP has duration of half a second, which is considered long for high-speed networks.

Review of these works shows that all of the methods follow the general *classical* modeling, where the mean and variance of real video are matched to an AR model or any known distribution function. Also in these classical models, the characterized video was homogeneous. The nature of the video content and its length have not been taken into consideration, i.e., whether it is a 10-s video-phone or a 30-min movie! Video material in real life has diverse characteristics in terms of its nature (video-phone, news, movie, sport, etc.) and content (active quiz show or nonactive news session). Obviously, modeling a video by considering its nature (style) and content will result in a better representation of the video traffic. This has motivated our research work which departs from the classical video modeling.

Bocheck and Chang [18] have also proposed a content-based video (CBV) traffic model which uses camera operations such as panning, zooming, and static scene as visual effect primitives to model the video. The scenes were considered to be composed of several epochs, each containing one of the visual effect primitives. The CBV model was made of two independent parts, the epoch and the traffic model. The former was used to model the video epochs in both spatial and temporal dimensions. Its output was used subsequently by the traffic model to generate the corresponding bit rate. Two parameters were used for the epoch model, namely, global and local descriptors, which were used in defining the epoch model state. At each state, a different mathematical model best describing the current epoch was selected. Frame descriptor, which is the output of the epoch model, includes parameters that depend on the coding techniques.

In this paper, we propose a VBR MPEG video model that can represent any style of video based on the description of its content, by using statistics of the video shots within a long video clip. Shot durations and the transition probabilities between the shots are investigated. In this work we assume an image sequence has already been segmented into homogeneous shots.

II. Decomposition of Video

A. The Hierarchy of Video

A video can be represented in a hierarchical form, where the top level is the program and the bottom level is a single video frame. In this paper, we use *Video-Clip* to represent a self-sufficient piece of video program, such as a film, a comedy episode, or a TV news bulletin. A video-clip can be temporally segmented into meaningful units called *Stories*. In drama films, story can be considered as a piece of video that involves a single element of the dramatic action with a fixed number of characters and a common location. For example, the coverage of an individual news topic can be called a story within the news session video-clip. Stories can be further segmented into SHOTS, which represent a continuous action from the start to end of a single camera operation. There may be panning, or zooming in order to follow the subject, but there is no abrupt change of viewpoint. At the next level, shot consists of a number of group of pictures (GOP), and each GOP has a well-defined structure in terms of its individual video *Frames*.

B. Shot Characterization

Most of the work on video modeling which was surveyed in Section I has been tested for homogeneous video. No considerations were given to the video content and its length except in [18]. Video programs in real life have diverse content in terms of image texture and motion.

In fact, homogeneous video is another name for the video shot described in the previous subsection. A video-clip is nonhomogeneous since it comprises several shots of different characteristics. Therefore modeling a video-clip should start from modeling the shot, then advancing from one shot to another to represent the whole clip.

In this paper, texture and motion are used to classify shots into various groups. Each combination of texture and motion is graded subjectively into a number of levels. The more levels, the more accurate the model, but at the expense of the model's complexity. For moderate complexity, three levels of texture and three levels of motion are chosen, namely, low, medium, and high, resulting in nine different types of shots: LL, LM, LH, ML, MM, MH, HL, HM, and HH, where the first letter represents the texture level, and the second one represents the motion level, and L, M, and H stand for low, medium, and high, respectively.

As an example, Fig. 1 shows nine single frames of various shots. The first column shows the shots with slow motion and

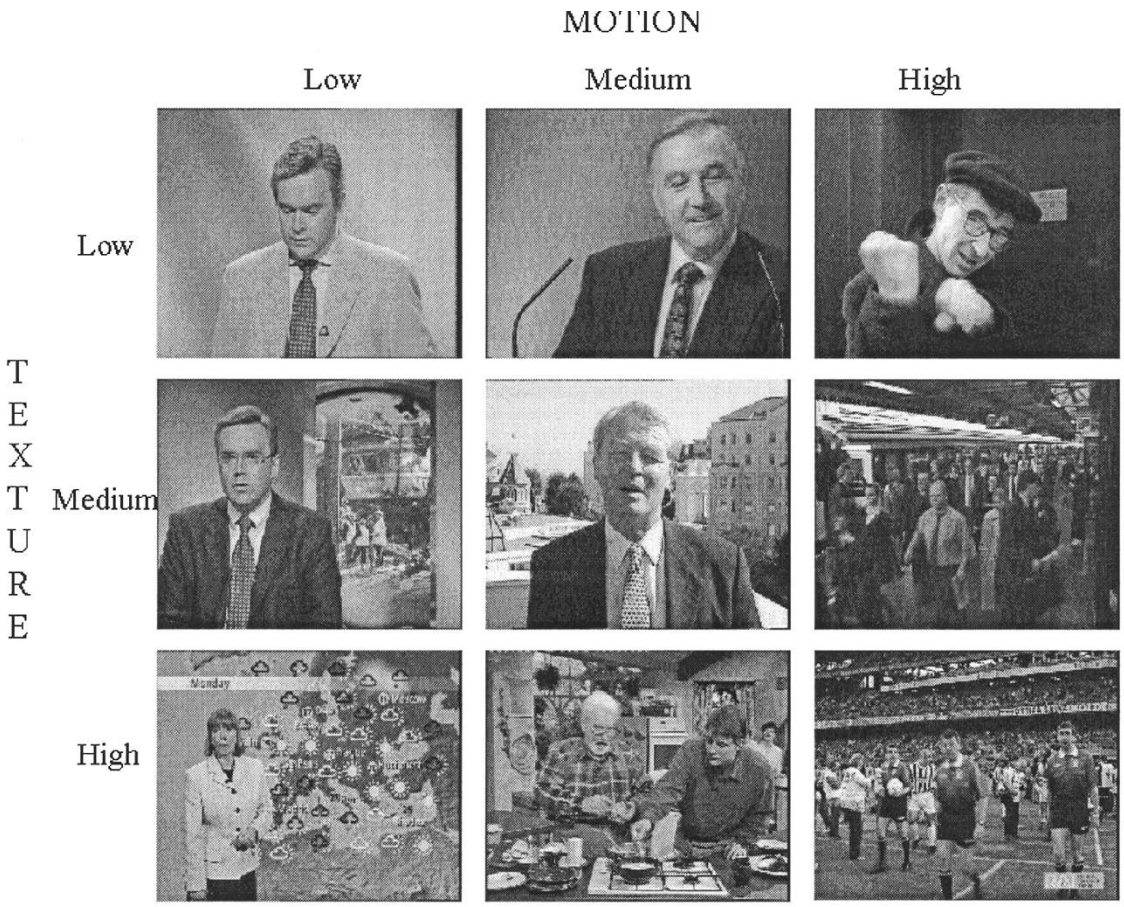


Fig. 1. Single frame from each shot.

the last column with high motion. Similarly in the first row shots have low texture and in the bottom row with high texture. Combinations of rows and columns specify the level of motion and texture in each shot. Table I lists the bit rate of each shot when coded with a VBR MPEG encoder. The encoded ATM cells per frame profile (53-byte cells with a payload of 48 bytes) of the shots is shown in Fig. 2 following the shot order presented in Table I.

To measure the texture level of the encoded frames, for each frame type the average magnitudes of the DCT coefficients of the luminance/block (8×8 pixels) were calculated during the encoding process and then averaged over the shot. In our experiment we have used a GOP structure of $M = 3$ and $N = 12$, that is a distance of three frames between the anchor I- or P-frames and twelve frames between the I-frames. The MPEG bit stream was also made variable by using fixed quantizer scale codes of 3, 3, 4 for I-, P-, and B-frames, respectively. Figs. 3–5 show the relation between the average DCT coefficient magnitude and the generated ATM cells per frame for I, P, and B, respectively. Relation between DCT coefficient magnitudes and the bit rate is very distinct for I-frames, shown in Fig. 3, as expected, but due to the motion these relations are not so clear cut for P- and B-frames as shown in Figs. 4 and 5. Hence using bit rate of I-frames, one can classify images in terms of their texture.

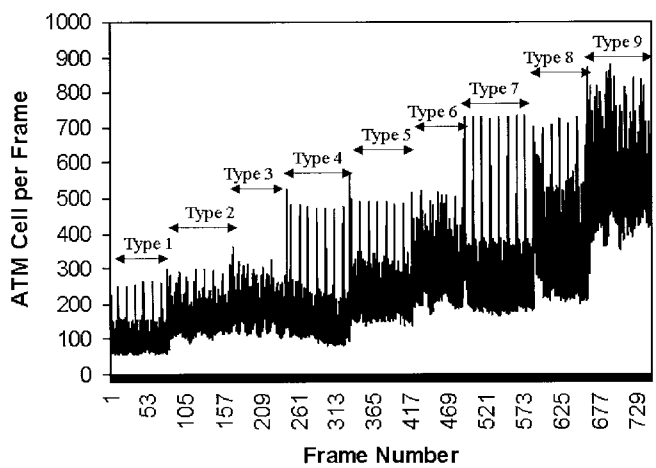


Fig. 2. ATM cells per frame for the nine different shots.

On motion classification, for each frame type the magnitude of motion vectors (MV)/macroblock were also extracted during the encoding process and averaged over the shot, in order to measure the motion level of the encoded shot. Since I-frames are intraframe coded, the average MV magnitudes were calculated for the P- and B-frames only within every shot. Figs. 6 and 7 depict the relation between the magnitude of MV and the generated cells per frame for P- and B-frame, respectively. Again, in terms of motion, classification clearly

TABLE I
MEAN BIT RATE (MBR) IN Mbit/s FOR EACH CODED SHOT, WITH A PICTURE SIZE OF 352 × 288 (PIXELS) AT 25 Hz

Texture	Motion	MBR(Mbps)
L	L	0.883
L	M	1.364
L	H	1.525
M	L	1.441
M	M	1.994
M	H	2.464
H	L	2.427
H	M	3.080
H	H	4.737

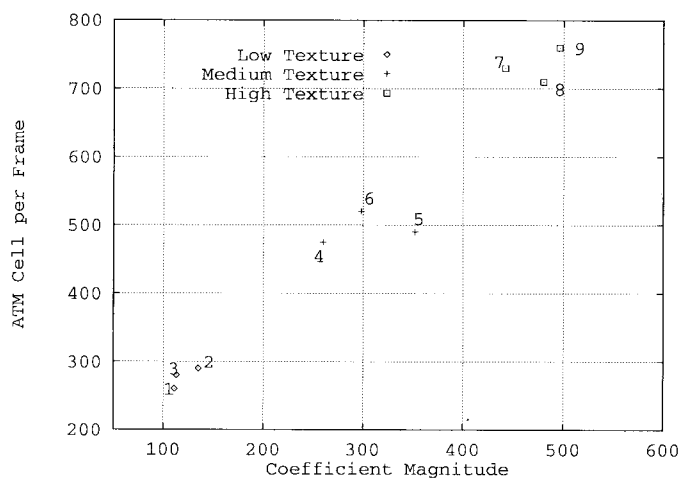


Fig. 3. Average DCT coefficient magnitude of I-pictures for nine shots.

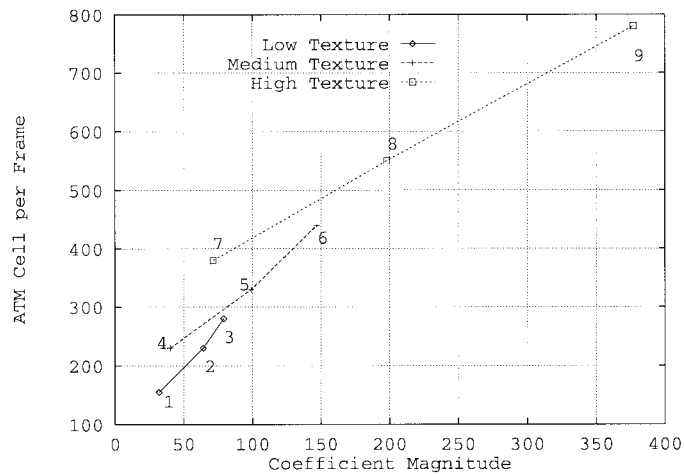


Fig. 4. Average DCT coefficient magnitude of P-pictures for nine shots.

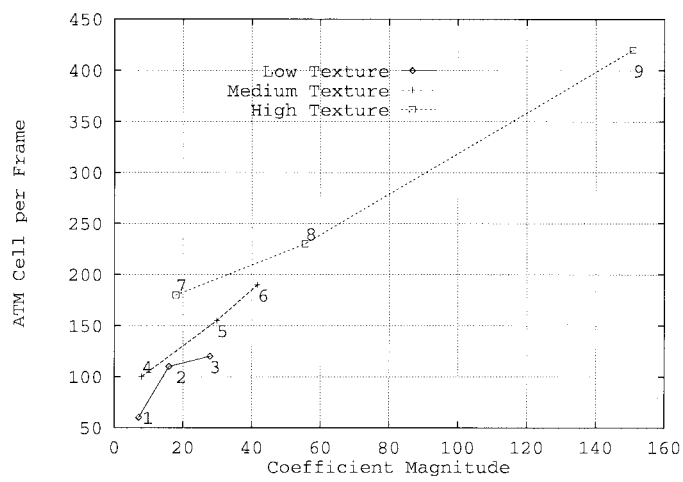


Fig. 5. Average DCT coefficient magnitude of B-pictures for nine shots.

shows the effect of motion on the bit rate, but due to the influence of DCT coefficient (texture), the classification is not so distinct. However, if results of I-frames are combined with those of P- and B-frames, then a very reliable classification can be made. For example, the class of the texture can be known from the I-frames, then P- and B-frames would help on the motion-based classification. Therefore, results obtained in Figs. 3–7 emphasize the fact that the generated number of cells per frame is directly affected by the texture and motion level of the encoded video shot.

C. Characterization of Real Video Clip

The above shot classification and analysis were applied to a 30-min BBC news bulletin, which was found to be a good example of a video-clip containing shots with different motion and texture levels. The news shots were classified subjectively into nine types according to their motion and texture levels. In case subjective classification for some complex shots such as a shot with camera panning, camera shaking, and cartoon seemed to be difficult, the encoded cells per frame profile was used instead. There were 228 shots detected in the 30-min news session, which is found to be close to what has been reported in [19]. Fig. 8 depicts the histogram of shot durations in frames. The mean shot duration was found at 177 frames (equivalent to 7.1 s at 25 frames/s) and the distribution curve was observed to follow a second-order gamma distribution. The frequency of occurrences of each shot type is tabulated in Table II. The temporal correlation between the shots was determined by measuring the probability that one shot type follows the next type. This is listed in Table III and is used as the transition probabilities in our model for defining the temporal dependencies among the shots. This transition probability table shows that a LL shot is more likely to follow another LL shot, and shots with high motion are almost unlikely to follow LL, unless they have lower texture. Also from HH type we see that it is rare for a high texture and high motion (HH) shot to follow a low motion and low texture (LL) shot, but it normally transits via a medium texture or medium motion shot as shown in Table III.

III. COMPOSITION OF VIDEO CLIPS

A. Content-Based Model (CBM)

To model video based on this concept, we assume that video has already been segmented into shots using temporal

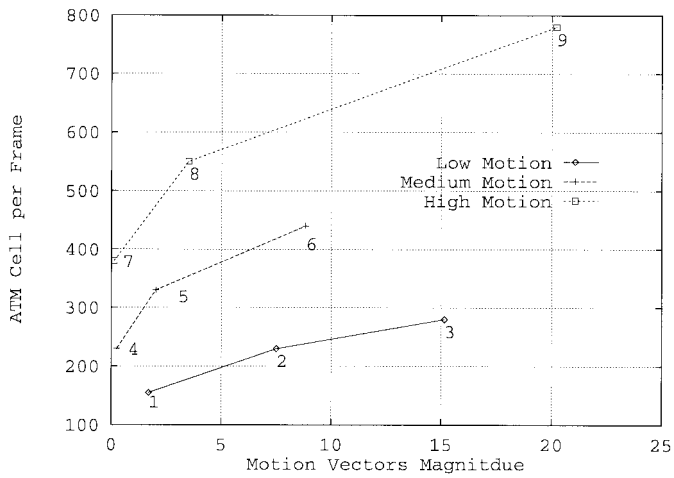


Fig. 6. Average of MV magnitudes of P-pictures for nine shots.

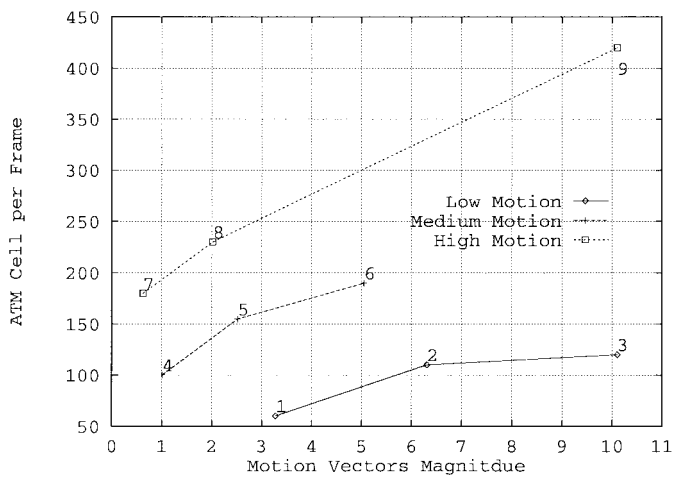


Fig. 7. Average of MV magnitudes of B-pictures for nine shots.

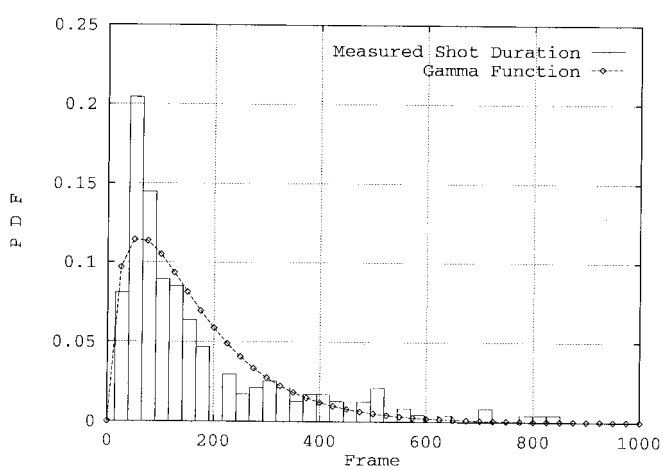


Fig. 8. Histogram of shot duration (frames).

segmentation procedures, such as time constrained clustering [19], [20], self-organizing scene clustering [21], and histogram comparison [22].

In Section II-B, we classified video shots into nine different types based on their texture and motion complexity levels.

TABLE II
NUMBER OF EACH SHOT TYPE IN 30-MIN NEWS

Texture	Motion	Frequency
L	L	43
L	M	30
L	H	18
M	L	48
M	M	33
M	H	26
H	L	20
H	M	7
H	H	3

TABLE III
SHOT TYPES TRANSITION PROBABILITY (%) FOR 30-MIN NEWS

From	To								
	LL	LM	LH	ML	MM	MH	HL	HM	HH
LL	46.51	25.58	2.33	6.98	11.63	0.00	2.33	4.65	0.00
LM	17.02	27.66	19.15	8.51	10.64	0.00	8.51	6.38	2.13
LH	0.00	40.00	15.00	15.00	10.00	5.00	5.00	5.00	5.00
ML	16.67	16.67	3.33	16.67	23.33	6.67	10.00	6.67	0.00
MM	21.21	15.15	3.03	12.12	9.09	6.06	9.09	21.21	3.03
MH	0.00	14.29	28.57	14.29	14.29	0.00	0.00	28.57	0.00
HL	5.56	0.00	5.56	38.89	5.56	0.00	27.78	16.67	0.00
HM	7.69	15.38	7.69	11.54	34.62	0.00	3.85	19.23	0.00
HH	0.00	0.00	0.00	0.00	0.00	66.67	0.00	33.33	0.00

Each type has a different overall mean bit rate as expected. The lowest bit rate is for low texture and the low motion (LL) shot, and the highest bit rate is when both texture and motion are high, HH, as shown in Table IV. However, there are combinations of texture and motion which generate almost similar rates. For example, low texture and high motion (LH) is almost identical to medium texture but low motion (ML), and many others. The implication of this is that in fact we do not need nine types, and some types generate similar bit rates, but nevertheless, transition from one type to another depends on texture and motion, as shown in Table III. Hence we will keep the nine types for this reason. Also note that in almost all bit rates (all shot types), the percentage of bit rate in B-frames from the mean bit rate is almost constant. In Table IV, this percentage varies between 2.4–3.14% of the coded bit rate. Similarly, for every type, the percentage of bit rate assigned to P-frames is fairly constant from 5.5 to 6.6%. However, what changes is the bit rate assigned to I-frames, which is again not much. This is about 5.6 to 10.5%. In this table the mean bit rate (third column) is divided into I-, P,, and B-frames with a GOP structure of $N = 12$ and $M = 3$. There are one I-, three P,- and eight B-frames in a GOP, with a GOP frequency of $\frac{12}{25}$ of frame rate.

TABLE IV
MEAN BIT RATE (MBR) AND THE PERCENTAGE OF I-, P-, AND B-FRAME BIT RATES FOR NINE DIFFERENT SHOTS CODED AT 25 HZ, WITH A GOP STRUCTURE OF $N = 12$ AND $M = 3$

Tex	Mot	MBR(Mbps)	μ I%	σ/μ I%	μ P%	σ/μ P%	μ B%	σ/μ B%
L	L	0.883	10.35	2	6.17	2.9	2.4	7
L	M	1.364	7.48	3.8	5.93	6.2	2.83	8
L	H	1.525	6.46	8.23	6.64	7.4	2.76	13
M	L	1.441	11.6	0.82	5.618	5	2.44	5.4
M	M	1.994	8.64	0.91	5.82	4.6	2.73	6.6
M	H	2.464	7.42	2.23	6.28	5.8	2.71	9.2
H	L	2.427	10.58	0.26	5.51	2.1	2.61	4.6
H	M	3.080	8.11	1.57	6.28	3.72	2.62	8.6
H	H	4.737	5.64	6.82	5.79	6.7	3.12	8

The implications are that irrespective of the mean bit rate/shot, one can define a simple relation between the bits assigned to I-, P-, and B-frames. Hence, for a given mean bit rate per shot, the respective bitsper frame type can be calculated. Another notable implication of this table is the relation between the shot type and mean bit rate. Therefore, if the number of shots and shot type in a video clip is known, then one can define the mean bit rate for each shot from the overall mean bit rate. This is significant, since in VBR video, normally the overall bit rate is defined, not bits for individual shots.

After classification of video clip into shots, and determination of bit rate for each shot and the proportion of I, P, and B bit rates, we can model the video. Hence a shot can be defined as a vector

$$S_k(AR\_I_i, AR\_P_i, AR\_B_i, t_k)$$

where $k = 1, 2, \cdots, N$ represents the kth shot in a video clip of N shots, $i = 1, \cdots, 9$ represents the ith shot type, t_k is the duration of the kth shot, and AR_I_i, AR_P_i, and AR_B_i are the I-, P-, and B-frames autoregressive model parameters (determined according to [6]), respectively, for the ith shot type.

The following steps summarize the whole procedure for a synthetic generation of video traffic based on the following model.

1) Define the number of shots (N) in the video-clip.
2) Specify the shot type, and derive the mean bit rate of each shot from the overall mean bit rate.
3) Specify the shot duration, according to the statistics and Gamma function.
4) Using the mean and variance, calculate the autoregressive (AR) model's parameters for the kth shot [6].
5) Go to step 3 for the kth + 1 shot.

B. Homogeneous Shot Modeling

Since shots are homogeneous and homogeneous video is well represented by an autoregressive (AR) model, then we model each shot with an AR, considering the GOP structure (e.g., the sequence of I, P, and B in a GOP).

The first-order AR process is defined as

$$\lambda(i) = a\lambda(i-1) + bw(i) \tag{1}$$

where $\lambda(i)$ represents the generated bits for the ith frame, a and b are constants, and $w(i)$ is an independent Gaussian random process with a mean η and variance 1. The steady-state average bit rate $E(\lambda)$ and the autocovariance of bits per me $C(j)$ are given by [6]

$$E(\lambda) = \frac{b}{1-a}\eta \tag{2}$$

$$C(j) = \frac{b^2}{1-a^2}a^j \qquad j = 0, 1, 2, 3 \cdots . \tag{3}$$

Three AR models are used for each shot, one for each frame type, considering the values in Table IV. Note that although in Table IV the percentage of mean bit rate for frame types is fairly constant, the standard deviation (square root of the variance) inside the shot varies considerably. Also for each shot and frame type, the ratio of the standard deviation to mean bit rate varies with texture and motion activity.

C. CBM Simulation Results

To show the performance of the proposed model, a virtual video-clip (VVC) was edited from 11 shots, covering all of the nine types of interest. Since bit rate within a shot is fairly constant, only a small portion of the real shot (3–4 s) was included in the VVC. This was done to reduce the simulation time. The VVC was then made 32-s-long (800 frames) and the shots were arranged to form a realistic story similar to a news video clip, according to the following scenario.

1) At the start, a news reader talks for approximately 3.5 s; the background is plain with a picture hanging on it. The background is thus medium textured, and the shot is classified as medium texture and low motion, ML.
2) Then, a reporter with a closed-up shot reports in open-air with a plain sky background. The movement is more than the news reader. This shot lasts for 3.5 s, and it is classified as low texture and medium motion, LM.
3) A person is interviewed in the open area for approximately 3.5 s. He is standing on the road with a flag waving on the side. The texture and the motion are rated as medium, MM.
4) The news reader then returns with a plain background, and talks for 3.2 s, LL shot.
5) In the next shot, two persons demonstrate a cooking recipe for 3 s; the texture is high and the motion is medium, HM.
6) The next shot is a person aggressively banging on a door for 3 s. The classification is low texture and high motion, LH.
7) The news reader with the plain background comes back for 1.5 s (LL).
8) Then, a medium textured scene from a train station lasts for 2.7 s. It shows passengers walking on the platform. It is considered MH.

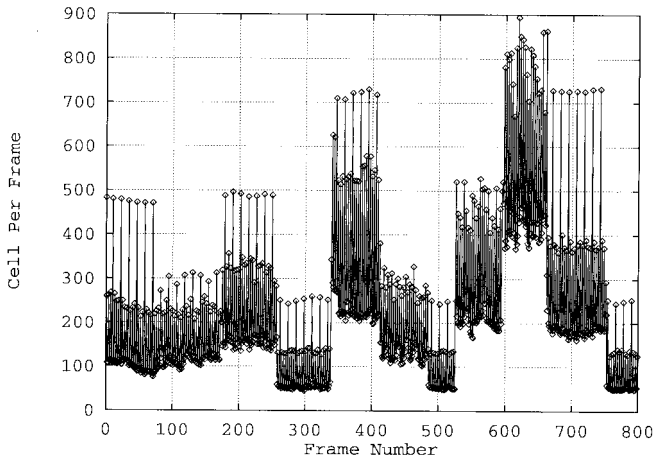


Fig. 9. VVC cells per frame—real.

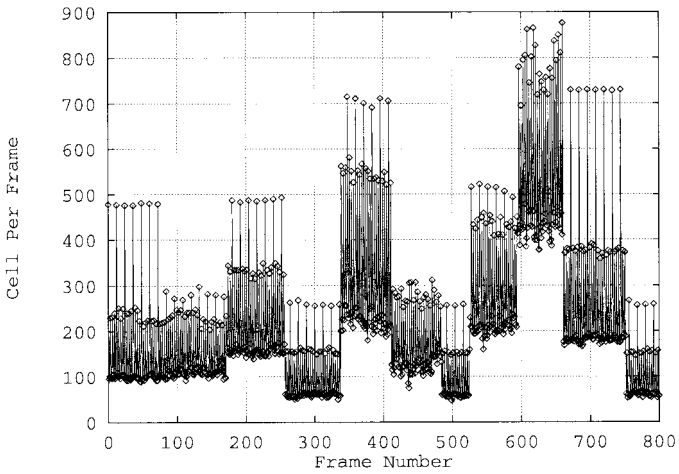


Fig. 10. VVC cells per frame—CBM.

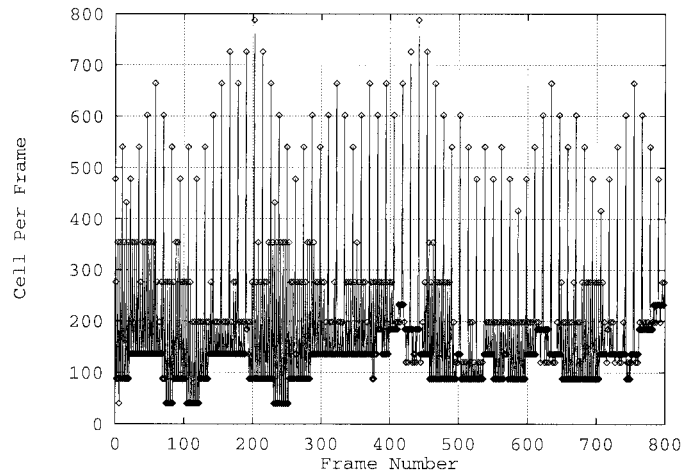


Fig. 11. VVC cells per frame—classical AR.

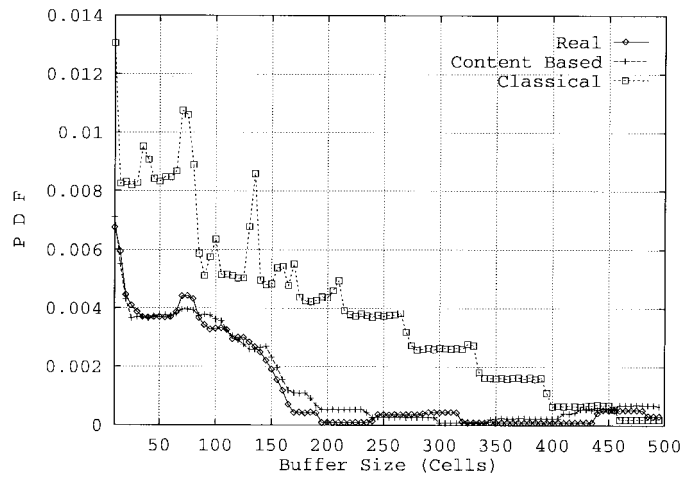


Fig. 12. Cell delay of 70% loaded ATM buffer.

9) After that, a 2.5-s football scene shows a highly textured stadium and running players. The shot is therefore of HH type.
10) Next is a weather lady talking about the weather in Europe for 3.5 s. The background is a synthetic map with cloud and sun symbols making it a highly textured background but the motion is low, i.e., HL shot.
11) Finally, the news reader with a plain background is talking for 2 s to summarize and close the session (LL).

The VVC is then VBR MPEG encoded with a fixed quantizer scale codes of 3, 3, and 4 for I-, P-, and B-frames, respectively. The cells per frame profile is shown in Fig. 9.

In order to evaluate the performance of our proposed model, the VVC was also modeled with the classical autoregressive method. In this model the statistics of the whole 800 frames were used in the modeling. Note that in CBM, we use autoregressive (AR) model for each shot. Fig. 10 shows the cell-rate profile of the CBM. Here we have used the exact durations and order of shot appear in the VVC. Fig. 11 shows the classical AR model. We have also evaluated the network behavior in response to the three forms of generated traffic. Each video traffic was packed into ATM cells, and the mean delay of the ATM multiplexer for network loads of 70 and 90% were calculated, as shown in Figs. 12 and 13, respectively.

Inspection of Figs. 9–13 shows how the content-based video model (CBM) closely follows the real nonhomogeneous MPEG traffic, while the classical method without the consideration of video content, fails to achieve such performance.

D. Robustness of Content-Based Modeling

Since CBM is based on the subjective description of the video content, the shot classification may vary from person to person depending on his/her perception and definition of the low, medium, and high motion and texture levels of every individual shot within the video clip. In order to study the robustness of the CBM model to different opinions in shot classification, some shots were classified in a different type. For example, a LM shot was assumed to be LL, or a HM to be MM. Table V illustrates the 11 shots with the new order, where most of them have been classified in a different class.

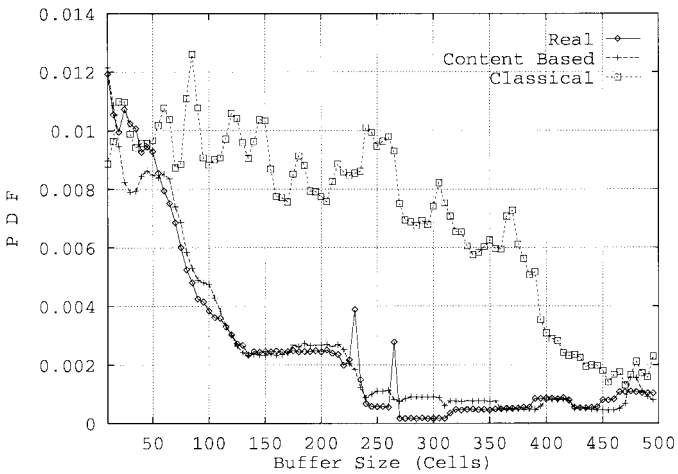


Fig. 13. Cell delay of 90% loaded ATM buffer.

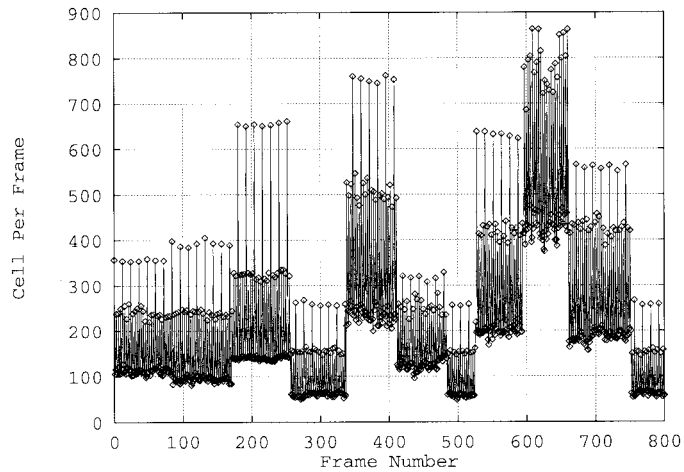


Fig. 14. VVC CBM—with a different classification.

TABLE V
RECLASSIFICATION OF SHOTS DUE TO A DIFFERENT OPINION

No.	Original	Re-classified
1	ML	MM
2	LM	LL
3	MM	ML
4	LL	LL
5	HM	MM
6	LH	LM
7	LL	LL
8	MH	MM
9	HH	HH
10	HL	HM
11	LL	LL

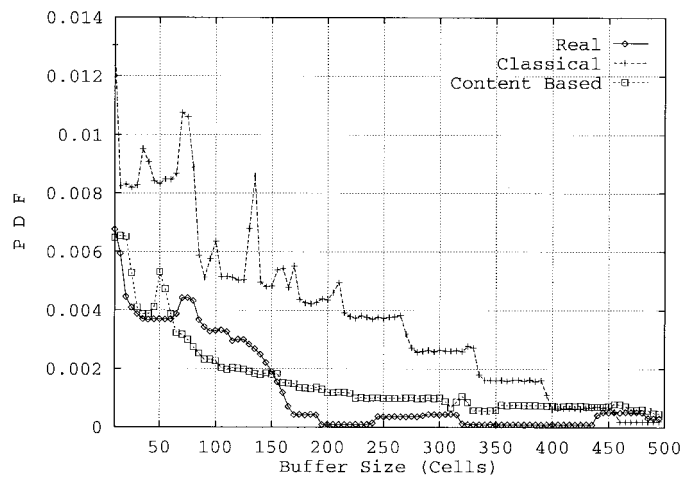


Fig. 15. Cell delay of 70% loaded ATM buffer.

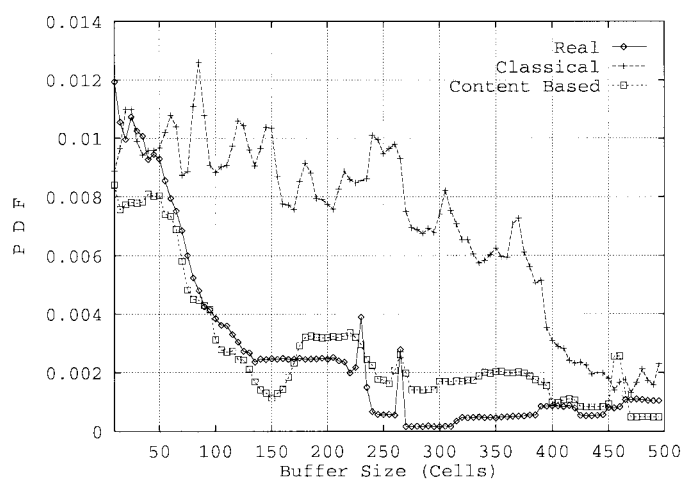


Fig. 16. Cell delay of 90% loaded ATM buffer.

The cells per frame of the CBM generated traffic with this new description is shown in Fig. 14. The network behavior was evaluated, as in Section III-C, to compare the performance of the CBM when making incorrect shot description with the classical AR model. Figs. 15 and 16 show the mean delay of the cells in an ATM buffer for network loads of 70 and 90%, respectively. Although the mean delay curve does not fit well with that of the real data it is still far better than the classical AR model's curve. These two graphs therefore, verify the robustness of the CBM in modeling video clips containing a number of shots even if it is based on some incorrect shot description. This is useful when running the CBM model with probabilistic transitions between the shot types, as will be seen in the next section.

IV. A PROBABILISTIC VERSION OF CBM

The simulation made in Section III-C was to emphasize the idea of the CBM scheme in video modeling, where the transition between the shots and the duration of each shot were made deterministic according to the scenario in the VVC. In order to derive a more realistic content-based video model, these transitions and durations should be made probabilistic, based on the video clip shots characteristics. The transition probabilities can be obtained from Table III, and the shot durations are based on gamma function of Fig. 8. In this

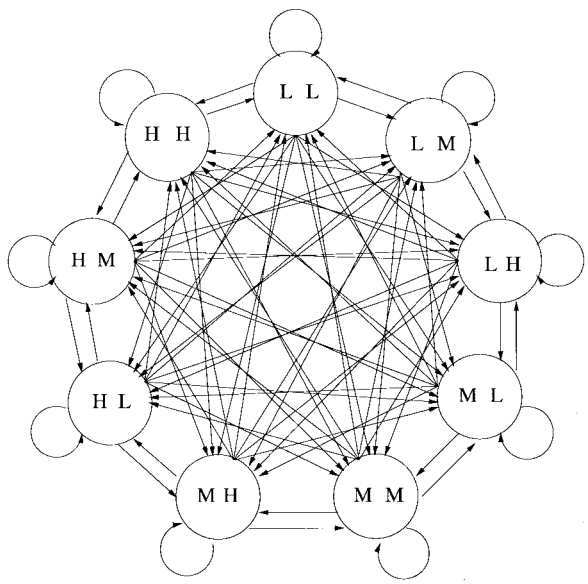


Fig. 17. Transition probabilities between all states.

section, the probabilistic results of the CBM will be shown and compared against the deterministic ones.

A. The Nine_State Model

Since the shots were classified into nine types, a nine_state model is used to represent the CBM, where each state represents a single shot type. The transition is allowed to occur between any two states; hence, there will be nine transitions from each state. A nine_state model with all the transition probabilities is shown in Fig. 17. Note, according to Table III some of the transitions may not occur (zero transition probability); hence, these transitions can be excluded from the diagram.

B. A Real News Video Clip

A 3000-frame (2-min) video clip was extracted from the 30-min news program. There were 21 shots found in the 2-min video clip, and the probability density function (PDF) of the durations (shot lengths in frames) of these shots was found to follow a gamma function with $\alpha = 2$ and $\beta = 70$. The transition probabilities between the states are listed in Table VI. It can be seen that many transitions have zero entries due to the limited number of shots available in the 2-min video clip. This limitation causes some of the transitions to be eliminated from the diagram, and, similarly, some states as well.

Assuming that the starting state is LL, then transition to the next states can be obtained from Table VI. Video traffic is generated by following these steps.

1) Start from an initial state.
2) Find the duration of the shot with a gamma function of $\alpha = 2$ and $\beta = 70$.
3) According to the type of the shot, use Table IV to calculate the autoregressive (AR) model's parameters.
4) Run AR model for the duration of the shot given in step 2.
5) Transit to the next state according to Table VI.
6) Go to step 2.

TABLE VI
SHOT TYPE TRANSITION PROBABILITY (%) FOR 2-MIN NEWS

From	To								
	LL	LM	LH	ML	MM	MH	HL	HM	HH
LL	0.0	0.0	0.0	50.0	0.0	0.0	50.0	0.0	0.0
LM	50.0	0.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0
LH	0.0	33.3	66.7	0.0	0.0	0.0	0.0	0.0	0.0
ML	0.0	25.0	0.0	50.0	25.0	0.0	0.0	0.0	0.0
MM	0.0	0.0	25.0	0.0	0.0	25.0	25.0	25.0	0.0
MH	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.0	0.0
HL	0.0	33.3	0.0	0.0	66.7	0.0	0.0	0.0	0.0
HM	0.0	0.0	0.0	0.0	0.0	0.0	100.0	0.0	0.0
HH	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

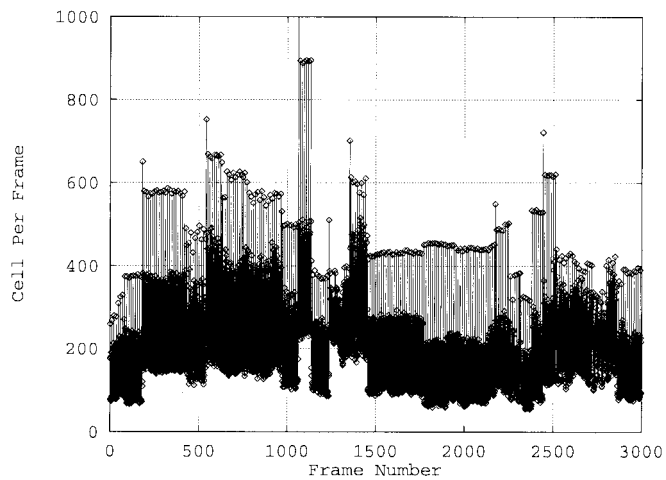


Fig. 18. Real traffic.

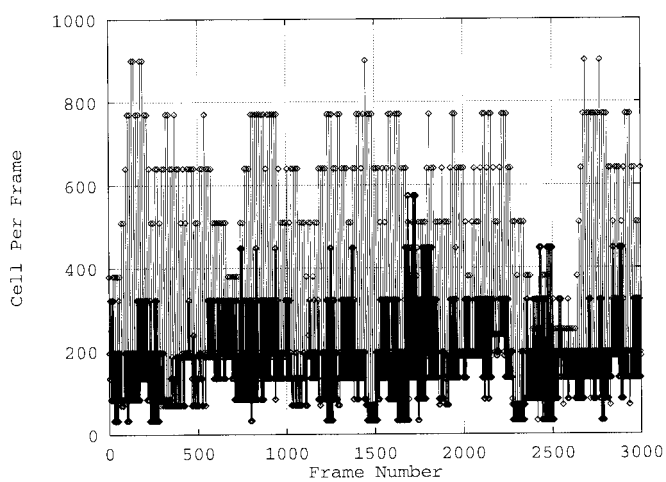


Fig. 19. Generated traffic with a classical AR model.

C. Simulation Results of Nine_State Model

The 2-min video clip shown in Fig. 18 was modeled with three different methods: 1) a classical AR method as shown in Fig. 19, where the AR parameters were derived from the

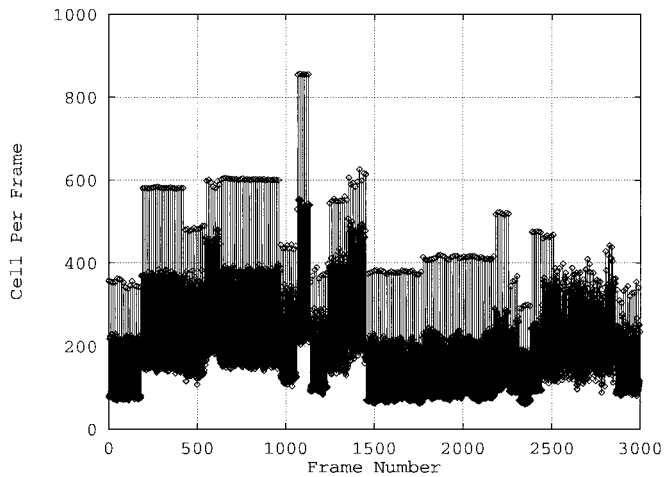


Fig. 20. Generated traffic with exact shot description and durations.

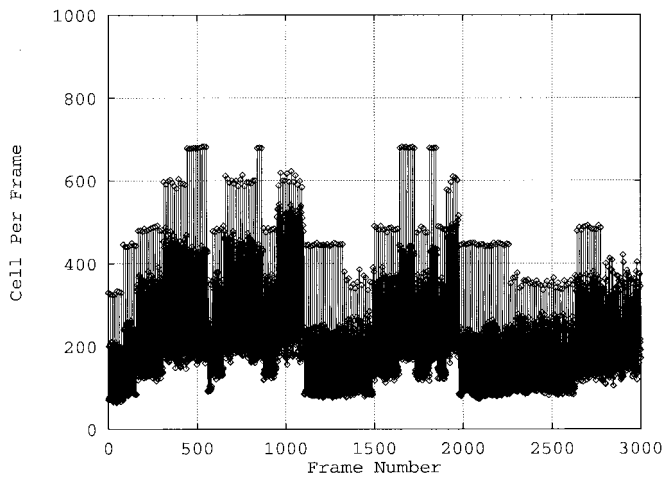


Fig. 21. Generated traffic with a nine_ state model.

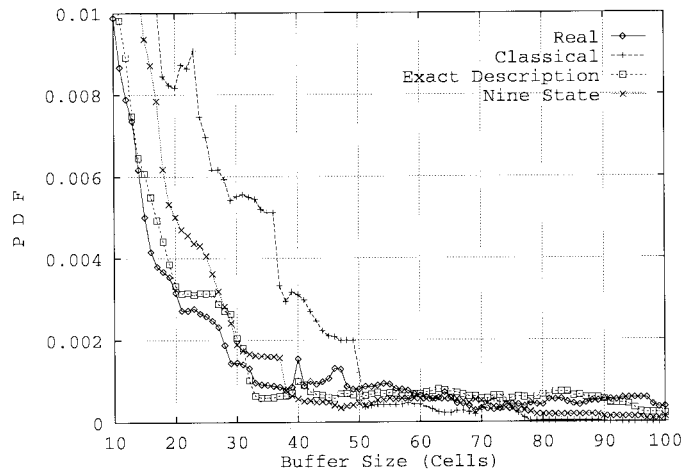


Fig. 22. Cell delay of 70% loaded ATM buffer.

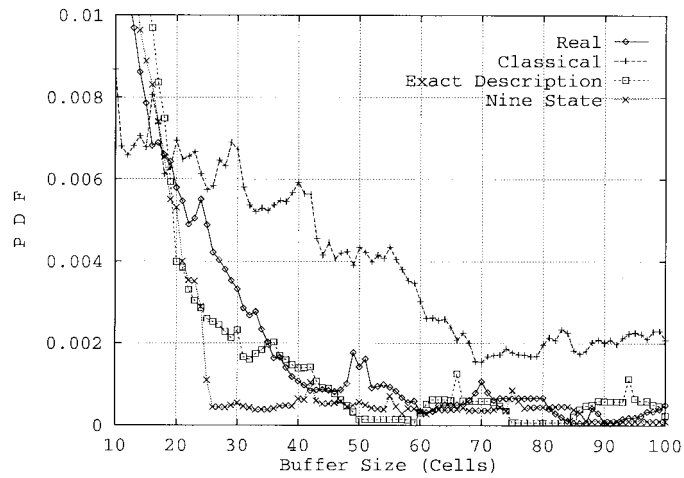


Fig. 23. Cell delay of 90% loaded ATM buffer.

overall characteristics of the real video traffic, 2) content-based, but using the exact descriptions and durations of the shots as shown in Fig. 20, and 3) applying the nine_state model with the state transitions of Table VI and durations of the shots following a gamma process of $\alpha = 2$ and $\beta = 70$, as shown in Fig. 21.

We have also evaluated the network performance with these traffics, where each model's generated traffic has been fed into an ATM multiplexer with a given service rate. Figs. 22 and 23 show the mean delay of the ATM multiplexer with network loads of 70 and 90%, respectively, for the real and three modeled video traffics. The worst performance was observed to be for the classical method which does not consider the video content information. The description based model shows a better performance especially when the video content has been described exactly. The nine_state model, which is a purely statistical model has much better performance than the classical model.

V. Conclusions

In this paper various aspects of video modeling were reviewed, and the demand for a new approach of modeling was outlined. We proposed a content-based model (CBM) that can generate MPEG video traffic for a wide range of applications, such as video conferencing, TV news, and sport programs. The approach was based on the subjective description of the video content. Video-clip content was decomposed into homogeneous shots, where each shot is modeled with three-autoregressive (3-AR) models for I-, P-, and B-frames. The bit/frame within each shot is derived from the mean bit rate and specific ratios between I-, P-, and B-frames.

Shots were classified into nine types according to their texture and motion, and a nine_state model was proposed to represent the nine types. The model was made to switch between one shot type to another in a probabilistic way based on a transition table.

Performance comparison between the classical and the CBM approaches indicate the efficiency of the CBM for modeling a nonhomogeneous video-clip. The CBM robustness was tested by modeling a video-clip based on some incorrect description.

The network behavior have confirmed the superior performance of the CBM over the classical method in general, and shown that the nine_state model is a realistic solution for the CBM type of video modeling.

REFERENCES

[1] L. Alparone, F. Argenti, L. Capriotti, and G. Benelli, "Models for ATM video packet transmission," *Eur. Trans. Telecommun. Related Technol.,* vol. 3, no. 5, pp. 491–497, Sept./Oct. 1992.

[2] J. Bae and T. Suda, "Survey of traffic control scheme and protocols in ATM networks," *Proc. IEEE,* vol. 79, pp. 170–189, Feb. 1991.

[3] J. P. Cosmas, G. H. Petit, R. Lehnert, C. Blondia, K. Kontovassilis, O. Casals, and T. Theimer, "A review of voice, data and video traffic models for ATM," *Eur. Trans. Telecommun.,* vol. 5, no. 2, pp. 139–154, Mar./Apr. 1994.

[4] V. S. Frost and B. Melamed, "Traffic modeling for telecommunications networks," *IEEE Commun. Mag.,* vol. 32, pp. 70–81, Mar. 1994.

[5] I. W. Habib and T. N. Saadawi, "Multimedia traffic characteristics in broadband networks," *IEEE Commun. Mag.,* vol. 30, pp. 48–54, July 1992.

[6] B. Maglaris, D. Anastassiou, P. Sen, G. Karlsson, and J. D. Robbins, "Performance models of statistical multiplexing in packet video communications," *IEEE Trans. Commun.,* vol. 36, pp. 834–844, July 1988.

[7] P. Sen, B. Maglaris, N. Rikli, and D. Anastassiou, "Models for packet switching of variable-bit-rate video sources," *IEEE J. Select. Areas Commun.,* vol. 7, pp. 865–869, June 1989.

[8] R. Grunenfelder, J. P. Cosmas, S. Manthorpe, and A. Odinma-Okafor, "Characterization of video codecs as autoregressive moving average processing and related queuing system performance," *IEEE J. Select. Areas Commun.,* vol. 9, pp. 284–293, Apr. 1991.

[9] D. P. Heyman, A. Tabatabi, and T. V. Lakshman, "Statistical analysis and simulation study of video teleconference traffic in ATM networks," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 2, pp. 49–59, Mar. 1992.

[10] C. J. Hughes, M. Ghanbari, D. E. Pearson, V. Sefridis, and J. Xiong, "Modeling and subjective assessment of cell discard in ATM video," *IEEE Trans. Image Processing,* vol. 2, pp. 212–222, Apr. 1993.

[11] C. Shim, I. Ryoo, J. Lee, and S. Lee, "Modeling and call admission control algorithm of variable bit rate video in ATM networks," *IEEE J. Select. Areas Commun.,* vol. 12, pp. 332–344, Feb. 1994.

[12] P. Pancha and M. El-Zarki, "A look at the MPEG video coding standard for variable bit rate video transmission," presented at IEEE INFOCOM'92, Conf. Computer Communications, Florence, Italy, 1992.

[13] ——, "MPEG coding for variable bit rate video transmission," *IEEE Commun. Mag.,* vol. 32, pp. 54–66, May 1994.

[14] D. P. Heyman, A. Tabatabi, and T. V. Lakshman, "Statistical analysis of MPEG-2 coded vbr video traffic," in *6th Int. Workshop on Packet Video,* Portland, OR, Sept. 1994.

[15] J. C. Wu, Y. W. Chen, and K. C. Jiang, "Modeling and performance study of MPEG video sources over ATM networks," in *Proc. IEEE Int. Conf. Communications,* Seattle, WA, 1995, vol. 3.

[16] M. Krunz, R. Sass, and H. Hughes, "Statistical characteristics and multiplexing of MPEG streams," in *Proc. IEEE Int. Conf. Computer Communications, INFOCOM'95,* Boston, MA, Apr. 1995, vol. 2, pp. 455–462.

[17] J. Ni, T. Yang, and D. H. K. Tsang, "Source modeling, queuing analysis, and bandwidth allocation for VBR MPEG-2 video traffic in ATM," *Proc. Inst. Elect. Eng., Commun.,* vol. 143, no. 4, pp. 197–205, Aug. 1996.

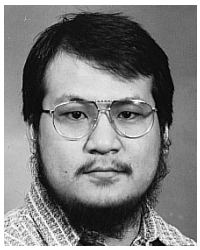
[18] P. Bocheck and S. Chang, "A content based video traffic model using camera operations," in *Proc. IEEE Int. Conf. Image Processing, ICIP'96,* Lausanne, Switzerland, Sept. 1996, vol. 2, pp. 817–820.

[19] M. M. Yeung and B. Yeo, "Video visualization for compact presentation and fast browsing of pictorial content," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 7, pp. 771–785, Oct. 1997.

[20] B. Yeo and M. M. Yeung, "Analysis and synthesis for new digital video applications," in *Proc. IEEE Int. Conf. Image Processing, ICIP'97,* Oct. 1997, vol. 1, pp. 1–4.

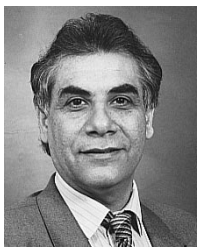
[21] Y. Ariki and Y. Saito, "Extraction of TV news articles based on scene cut detection using DCT clustering," in *Proc. IEEE Int. Conf. Image Processing, ICIP'96,* Lausanne, Switzerland, Sept. 1996, vol. 3, pp. 847–850.

[22] H. Yu, G. Bozdagi, and S. Harrington, "Feature-based hierarchical video segmentation," in *Proc. IEEE Int. Conf. Image Processing, ICIP'97,* Oct. 1997, vol. 2, pp. 498–501.



Ali M. Dawood was born in Sharjah, United Arab Emirates (UAE), in 1971. He received the BTEC diploma and the B.Eng. degree in communications in 1991 and 1994, respectively, from Etisalat College of Engineering, a division of Emirates Telecommunications Corporation—ETISALAT, UAE. He received the M.Sc. degree in telecommunication and information systems from the University of Essex, Colchester, U.K., in 1995. He is currently pursuing the Ph.D. degree in the Department of Electronic Systems Engineering at the University of Essex.

He has been employed by ETISALAT, since 1994, as a candidate for teaching assistance, where both the M.Sc. and Ph.D. degrees were under the sponsorship of Etisalat College of Engineering. His research interests include MPEG video traffic source modeling, networks performance evaluation, and high-level video processing for multimedia applications.



Mohammed Ghanbari (M'78–SM'97) received the B.Sc. degree in electrical engineering from Aryamehr University of Technology, Tehran, Iran, in 1970 and the M.Sc. degree in telecommunications and Ph.D. degree in electronics engineering, both from the University of Essex, U.K., in 1976 and 1979, respectively.

After working almost ten years in industry, he started his academic career as a Lecturer in the Department of Electronic Systems Engineering, University of Essex in 1988 and was promoted to Senior Lecturer, Reader, and then Professor in 1993, 1995, and 1996, respectively. His research interests are video compression and video networking. He is best known for his pioneering work on two-layer video coding for ATM networks.

Dr. Ghanbari is the co-recipient of the 1995 A. H. Reeves Premium Prize for the year's best paper published in the *IEE Proceedings* on the theme of digital coding. He was also a co-investigator of the European MOSAIC Project, studying the subjective assessment of picture quality, which resulted to ITU-R Recommendation 500. He is the co-author of *Principles of Performance Engineering*(London, U.K.: IEE Press, 1997). He has been a member of the organizing committee of several international conferences and workshops. He was the Chairman of the Steering Committee of the *1997 International Workshop on Audio Visual Services over Packet Networks, AVSPN'97,* and Guest Editor to 1997 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, Special Issue on Multimedia Ttechnology and Aapplications. Currently, he represents the University of Essex as one of the six U.K. academic partners in the Virtual Centre of Excellence in Digital Broadcasting and Multimedia.