Rate Control for Video Telephony Applications

Nisanth MP, Srividya Narayanan, G Nageswara Rao, D Jayachandra

Abstract—**Video Telephony based on the 3G-324M standard has been one of the most demanding multimedia applications in the last few years, with 3G networks emerging across the world. The key requirement of a video telephony system is sending good quality audio, video and control information in real time across the network with bandwidth restrictions as low as 64 kbps. There are several challenges in video coding (say MPEG4/H.263 or H.264 standard) for the stringent requirement of the H324M standard. If a proper rate control scheme is not a part of the video encoder, the number of bits generated per second for a sequence of video frames varies widely, depending on the information content of the source (i.e. the complexity of the video sequence – such as motion content, distribution of texture frequency content, scene changes, frame drops by the source etc.). But, any increase in the instantaneous (per second) bit rate of the compressed data would lead to packet losses and/or undesirable quality and/or delay at the decoding terminal. Generic constant bit rate (CBR) control algorithms ensure the bit consumption only over several seconds. They allow bit rate to shoot up during a single second although this overshoot will be compensated in the coming seconds. This is unacceptable for the real-time requirements of Video Telephony. Our paper describes the rate control algorithm implemented in Aricent MPEG4SP video encoder used for Video Telephony, which ensures that the compressed data bits is generated strictly within the given bandwidth and the quality of received video is good. This new second-based rate control algorithm (VT-RC) is based on the Scalable Rate Control scheme suggested in the informative Annex L of MPEG4 Part 2 standard, with many modifications to improve RD performance as well as to suit low delay requirements of VT.**

Index Terms* — *Macroblock (MB), Ratecontrol (RC), RemainingBitsInSecond, VideoTelephony (VT)

I. Introduction

VT is a real time application. The time taken for data transfer between the two persons doing VT is very sensitive. There should be no/negligible delay between the first person's recording of visuals and its receipt at the remote end, for the conversation to be natural and effective. The bit rate (or bandwidth - BW) consumed by text or even voice communication is very small (compared to the currently available bandwidth over varied geographies). So, text/voice communication can happen in real time without delays (lags). But video data requires much higher bit rate for digital representation as the information content of a video sequence is much higher compared to a voice recording of same duration for similar human perceptual expectations – or quality. So, sending video data over similar n/w in real time is a very constrained act.

To reduce the amount of data used for representing video sequence, a number of efficient compression algorithms have been proposed over time, and some of them were also standardized for seamless interoperability (like MPEG4 Part2, H.264, VC1 etc.). These compression schemes work by removing temporal, spatial, statistical as well as psycho-visual redundancies in the video sequence. So, the inherent RD performance of these algorithms is highly source dependent – i.e. the compression will result in different bit rates for different sequences, and even the bit rate may vary widely over the time for a single continuous recording, depending on the information content distribution in the video sequence over space (over the spatial positions of the same frame) and time (over the sequence of frames). This distribution is ideally represented by the amount of data (bit rate) in the compressed stream, after the redundancies are removed.

The challenge of VT is that the bandwidth available for video data should be fully utilized for maximum quality, but at no point in time, the bit rate should exceed this limited available bandwidth. If, say, the number of bits, for one second of video, exceeded the bit rate (bits per second), the extra bits will cause a corresponding delay in the availability of frames at the receiving person (Video Decoder), or packet loss in the channel. Either way, the QoS of VT is affected. If the delay/packet loss is high, VT is rendered ineffective. So, ideally, the compressed video sequence should consume Constant Bit Rate (CBR) at every instant of time. But, the natural behavior of the compression algorithms does not ensure this. This calls for an efficient RC algorithm integrated to the encoding algorithm.

Some of the commonly used rate control models for video coding are scalable rate control (SRC) based on a quadratic model [1], TM5 MB based rate control [3], TMN-8 [4] and Fixed QP. In each of these models, the output bit rate can be controlled by losing information content in an intelligent manner so that the quality of the video is not affected badly. Major decision in an RC algorithm would be the selection of varying quantization, not coding (skipping) MBs or partially coding MBs (entropy coding only some of the quantized transform coefficients.) and skipping frames.

This requires analysis of the information content of the source sequence. One way is to analyze the whole sequence and use this info to allocate bits for the whole sequence (target bitrate X duration), then for each GOV, then for each frame, then for each slice, and then for each MB etc. in a hierarchical manner. But this approach is not possible in a real time situation, where video data up to current time only is available to the encoder. (Encoder does not have any information about the complexity of future video frames).

The SRC in informative Annex L of MPEG4 part 2 standard [1] is targeted at the above mentioned scenario, which forms the basis for the RC scheme proposed in this paper. A brief review of SRC is given in part A of section II. This SRC was modified to remove some of its limitations before implementing it in Aricent MPEG4 SP Reference Encoder. These modifications [2] are briefed in part B of section II. This modified SRC is adapted for VT applications by incorporating a second based rate control scheme (VT-RC) as described in part C of section II. Part D of section II describes a few special cases covered by VT-RC. Section III briefs the experimental results for these RC schemes. Section IV concludes the paper and discusses some possible improvements for VT-RC.

II. Second Based Modified SRC

A. Scalable Rate Control:

The scalable rate control [1] scheme assumes a quadratic model for the encoder rate quantization function as follows.

$$R = S\left(\frac{X_1}{Q} + \frac{X_2}{Q^2}\right) \qquad (1)$$

The encoding bit count is denoted as R . The mean of absolute difference (*MAD*) is used as the encoding complexity measure and is denoted by *S*. *Q* denotes the quantization parameter and X_1 and X_2 are the model parameters.

The target bit rate is computed based on the bits available and the bits consumed by the last encoded frame.

$$T = \max(T_{\min}, 0.95R_r / N_r + 0.05R_{prev}) \qquad (2)$$

where T is the target bits, $T_{\min}$ is the minimum guaranteed target bits, R_r is the number of bits remaining for encoding the sequence (or segment), N_r is the number of frames remaining to be encoded in the sequence (or segment) and R_{prev} is the number of bits consumed by the previous frame. This target bits is then adjusted according to the buffer status to avoid the possibility of overflow and underflow. The target bits obtained thus is substituted for the encoding bit count *R* in equation (1) and solved for the quantization parameter *Q*. The value of *Q* obtained is clipped between 1 and 31. Moreover *Q* is limited to vary within 25 percent of the previous *Q* to maintain uniformity in quality. After encoding the frame, the model parameters X_1 and X_2 are updated, based on the encoding results of the current frame, using linear mean square estimation.

A rate control method at the level of macro blocks (MB) is also applied in order to achieve the VOP target more accurately and to adapt the quantization parameter to the MB variance and other perceptually relevant measures. Again a quadratic model similar to that of equation 1 is used. However the quantization parameter of an MB is restricted to be within ±2 of that of the previous MB.

B. Modified SRC:

SRC in Annex L doesn't consider the complexity of the current frame while allocating target bits for the frame. In the ARM platform based implementation of MPEG4SP Encoder, motion estimation is done for the complete frame, before starting to code that frame. Hence the complexity data and coding type decision data for the whole frame is available before hand for determining the rate control parameters at frame and MB level.

$$T = c\frac{R_r}{N_r} \qquad (3)$$

where $c = \max(0.8, \min(1.2, MAD_P / MAD_{AVG}))$ and MAD_{AVG} is the average complexity of the previous P-VOPs. Thus the target is allowed to change within 20% based on the relative change in complexity compared to the average complexity of the previous P-VOPs. So, in our RC scheme, a complex frame gets more bits for coding it [2].

When an inter MB with high encoding complexity or and Intra MB follows an inter MB with low encoding complexity, it would also have a relatively lower quantization scale than required, as the maximum increase allowed is only 2 units. As a result it consumes an inordinately high number of bits at the cost of other MB following it. To solve this problem, instead of using just the present MAD to solve equation (1) for MB level RC, a linear combination of the MADs of the present and following MBs are used [2]:

$$MAD = cMAD(i) + (1-c)MAD(i+1) \qquad (4)$$

Also, RD efficiency of the tools used in I frame is different from those in predicted (P) Frames. So, if an I-frame needs to be of similar quality as a P frame, it has to be coded using more bits. But, SRC in Annex L assumes I frames are very rare and mostly, only at the start of the sequence. So, it doesn't consider this separate RC requirement of I frames. But in VT like scenarios, I frames are quite frequent, as (a) the channel is erroneous and has packet loss forcing insertion of I frames more frequently, and (b) the low bit-rate requirement often becomes too stringent for the video complexity and quantization error increases as it propagates, to force I frames. So, in order to ensure smooth temporal variation in quality (to avoid flickering), and since the quality of the P frames after the I-frame depends on that I frame's quality, a separate RC scheme for Intra (I) frames is employed in the Modified SRC [2].

Consider a segment with N_r frames out of which the last frame is an I-VOP (current frame) and the rest are P-VOPs. The target for the I-VOP is allocated based on the following formula.

$$T = \frac{R_r}{\left[1 + \frac{(N_r - 1)MAD_P}{kMAD_I}\right]} \qquad (5)$$

Where MAD_P is the average complexity of the P-VOPs preceding the I-VOP. For the I-VOP, MAD_I is defined as the mean of the *MAD*s over the macroblocks. The MAD over a macroblock is defined as the mean of the difference from the average value over the macroblock. A scaling factor *k* is used as the encoding complexities MAD_P and MAD_I cannot be compared directly as they are calculated in entirely different ways. A nominal value of *k = 2* is found to be suitable for most sequences. However a better performance is obtained by varying *k* based on the complexity of the P-VOPs. Based on experiments over a wide range of sequences and bit rates the following formula was developed for calculating *k*.

$$k = \begin{cases} (6.75 - MAD_P / 2)/2.5 & \text{if } k < 6 \\ 1.5 & \text{otherwise} \end{cases} \qquad (6)$$

Thus k varies within a range of 1.5 to 2.5 depending on MAD_P. When MAD_P is high, then it is given greater weight i.e. MAD_P/k is higher whereas when MAD_P is low it is given less weight, i.e. MAD_P/k is lower.

C. Second Based RC:

In Modified SRC (as well as in original SRC), the rate and buffer requirement for various bitrates is as per the Virtual Buffer Verifier (VBV) model [6]. VBV model parameters (Buffer size, Initial buffer position etc.) determine the RD performance and experienced delay. In those schemes, the video sequence segment considered for RC is, generally, a GOV. Remaining_frames_in_GOV and Remaining_bits_in_GOV are tracked and used as Nr and Rr respectively. Let's see how the following scenario will be handled by M-SRC:

Assume that VBV buffer parameters take default values: for a 64kbps (bit-rate) SP L0 stream, VBV_size ≈ 164kb; VBV_initial_position ≈ 100kb. The maximum delay can be as large as ≈ 2.5 seconds (VBV_size/bit-rate). (Figure 1)

Even if we set VBV_size = bit rate; VBV_initial_position = 0.5 * VBV_size, at the decoder input, the delay could be as large as ≈ 1 second. This is the case, if we neglect Video Complexity Verifier (VCV) and Video Memory Verifier (VMV) dependencies [6], i.e. if VCV and VMV positions stay constant. If they are also changing in an adverse way, as typical in performance and memory constraint decoders like mobile devices, the delay before the decoded frame gets displayed on the screen could be still higher. This is an un-acceptable situation in VT (Figure 1). [(1) In Fig.1, delay for M-SRC with VBV size = bit rate, exceeds 1 second because, even after the decoder buffer under-flowed, the encoder generated frames of size, more than drain rate. (2) the rapid fluctuations in bits per frame plot is due to frame drops by BW adaptation algorithm.]

In a VBV model based RC, the direct solution to reduce delay is to reduce the VBV buffer size, still lower. But this would adversely affect the RD performance, as it reduces the additional bits that can be given to more complex frames. The above problem can be alleviated by the proposed second based RC. In VT-RC, the video sequence segment size used is Frame-rate (i.e. 1 second duration). Remaining_frames_in_Second and Remaining_bits_in_Second are tracked and used as N_r and R_r respectively. If there are remaining bits in a previous second, the same is not carried forward to the current second. Carry over bits is even limited to a few frames (3 or 5) in some cases, where stricter control is required.

The VBV buffer size value is set to the bit rate configured to encode, rather than the default standard defined values. At the starting, the buffer is assumed to be at 50% of its capacity and at any point, our intention is to maintain the buffer at this level. If available bits for a frame is too less for its complexity, that frame is skipped. For the next frame, there will be enough bits so that it can be coded with good quality. Thus dynamically varying the frame rate (within allowed limits), subjective quality of the video can be further enhanced [5].

The first I frame is normally encoded with higher quality in most rate control algorithms. Here, we use relatively lesser bits to encode the first I frame and allocate more for the P frame following an I-frame so that error propagation is minimized. Though this might lead to PSNR drop for the first I frame, the overall PSNR difference would be negligible and this can achieve tighter control of the bit rate. Per second rate control is not applied to I frames.

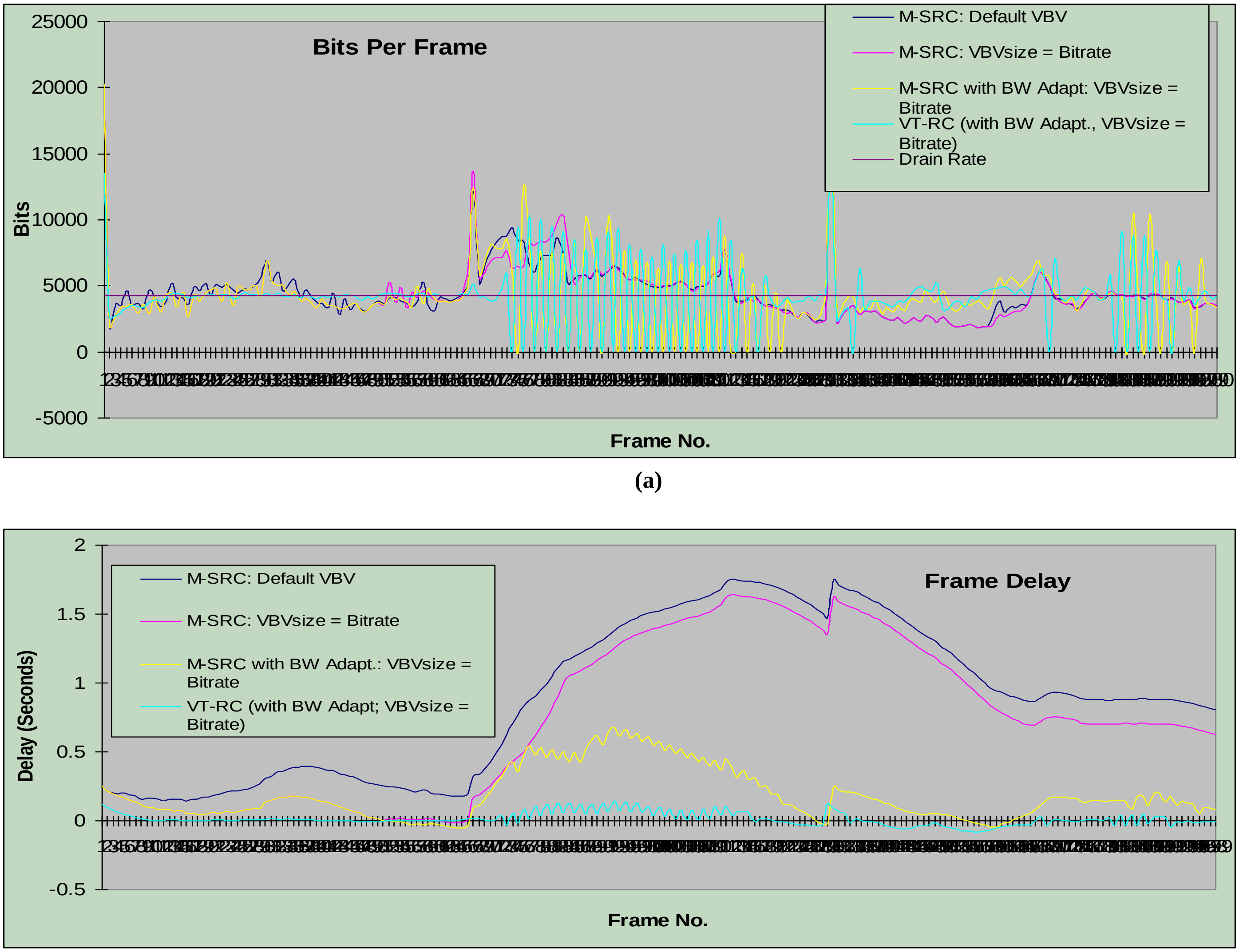
MB based rate control is used for the P frames, to tightly control the target bits. In an MB based rate control, the quadratic model is modified to include the rate variations across a frame. Quantization values of the previous MBs are used, which are stored in a circular buffer of size 20. After every encoded MB, the MB modeling parameters are updated based on the quantization value and the number of texture bits. In addition, based on the buffer fullness, the MB quantization level is adjusted such that it selects higher values of quantization when the decoder buffer is tending towards under flowing and lower values of quantization while it is overflowing.

These modifications ensure that average bits consumed by a few frames (say 3 to 5) is closer to a constant drain-rate while (consumed bits)/ (frame complexity) ratio is still taken in to consideration. All these modifications in VT-RC, along with the handling of special cases described in the next part, achieve good quality video meeting low delay requirements.

D. Special Cases:

The following special cases have been taken care of in second based RC scheme:

1. Dropping of frames for quality and bit rate – Since it is difficult to attain motion smoothness for high motion at low bitrates, frame rate control needs to be judiciously applied. Frames can be dropped along the second interval instead of dropping all the last frames towards the end of the second to achieve the bit rate and



(a)

(b)

Fig. 1. (a) Bits per Frame, **(b)** Delay plots for 200 frames of Swans sequence, 15fps, 64kbps

quality. The drop mechanism is based on buffer fullness, excess bits used, remaining bits and frames in a second. In addition if the frame corresponds to the next second, based on timestamp value, the frame is dropped to ensure real time encoding.

2. Scene change detection – Scene change is assumed if more than 30% of the MBs need to be coded as Intra MB [7]. However, in VT, insertion of an I-frame would lead to overshoot of the instantaneous bit rate. I frame is allowed only if the buffer is 60% empty, and it's the first I frame of that second and if the RemainingFramesInSec is more than 3. Such frames can be coded as P frames rather than I frame so as to maintain the bit rate within the second.
3. Stuffing bits – For very low bit rate scenarios during the VT call, at times the network expects a minimum amount of say 30000 bits to sustain the call. Also, if low complexity frames arrive continuously, the decoder buffer may overflow. In such scenarios, we can code the stuffing pattern repeatedly to achieve the bit rate. This can be done at frame level or at MB level if packet mode is used for encoding. As MPEG4 standard does not allow the concept of empty packets, we need to stuff each MB with the stuffing pattern to account for the minimum bit rate in packet mode.
4. Intra-refresh – A motion adaptive intra –refresh mechanism is used to ensure that error propagation at the decoder end is reduced.
5. Dynamic change of the bit rate / frame rate or insertion of I frames from the application affects the second based rate control for that particular second alone.

Thus, VT-RC ensures that the bit rate is within the limits during every second boundary. This includes implementation of dropping optimal number of frames for maintaining quality and/or bit rate. It takes into account several special cases also, like scene changes, bit stuffing requirements, frame drops by the source etc.

III. Experimental Results for various RC Schemes

Table 1 compares the bit rates for different RC schemes:

Seq	Fps	Bit rate	TM5	TM5 frame drop	SRC	Sec SRC
Fore	15	45	47.95	45.21	45.98	44.94
Fore	8	42	43.58	43.58	42.90	41.98
Foot ball	15	45	85.11	85.19	46.73	44.92

Table 1

Comparison of achieved frame rate for various RC schemes is shown in the Table 2:

Seq	Fps	Bit rate	TM5	TM5 frame drop	SRC	Sec SRC
Fore	15	45	15	12	15	15
Fore	8	42	8	8	8	8
Foot ball	15	45	14	13	10	10

Table 2

Comparison of PSNR and frame skips for different RC schemes is shown in the Table 3:

Seq	Fps	Bit rate	TM5	TM5 frame drop	SCR	Sec based SCR
fore	15	45	30.68	30.52	30.93	30.48
Foot ball	15	45	17.05	17.05	18.5	18
Fore	8	42	31.43	31.43	31.89	31.43

Table 3

The following chart (Figure 2) depicts the per second bit rate for the various RC schemes for foreman 45kbps bit rate and 15 fps as the target frame rate. With the second based SRC, there are no sudden peaks and the bit rate is constant in all the seconds unlike the TM5 or the SRC algorithms.

Although for some streams the PSNR values is lesser in SRC with second based rate control, the subjective quality is similar to the SRC model. Also second based SRC ensures that at every given second, the rate is within the target bitrate and there are no sudden overshoots for I frames. TM5 uses MB based RC and is useful in applications where motion estimation is done at MB level rather than frame level. But this does not achieve strict bit rate control for VT requirements as the frame complexity is not considered for rate distortion modeling.

IV. Conclusion and Future Scope

(A) Current implementation of RC for VT uses fixed 1-Second boundaries. That is, if the first frame’s time-stamp is ‘X’ seconds, this RC will try to ensure the bit-consumption to be with in Bit-rate, for the Frame-rate number of frames from ‘X’ seconds to ‘X + 1’ seconds, from ‘X + 1’ seconds to ‘X + 2’ seconds, from ‘X + 2’ seconds to ‘X + 3’ seconds, and so on.

(B) This scheme ensures that the maximum delay (due to bit-rate over shoot), any frame may suffer, is only a fraction (that can be adjusted in RC) of a second.

(C) Also, this delay will not propagate indefinitely: Zero-delay will be restored within the next few frames, or in worst case, before the next second boundary. i.e. Zero-delay will be restored within less than a second duration in any case.

But, consider some situations like these:

A. *Smoother Variation in Temporal Subjective Quality*

(1) In any 1-second duration, say from ‘X’ seconds to ‘X + 1’ seconds, if initial frames are of small information content (low motion, texture contents), they consume, say, about drain-rate bits per frame. Then, near the end of that second (near ‘X+1’ boundary), if some frame of high information content comes (a scene change and as a result, an I-frame; or high motion content), there will not be enough bits available in that second. This will cause that frame and consequent frames in that second to be skipped.

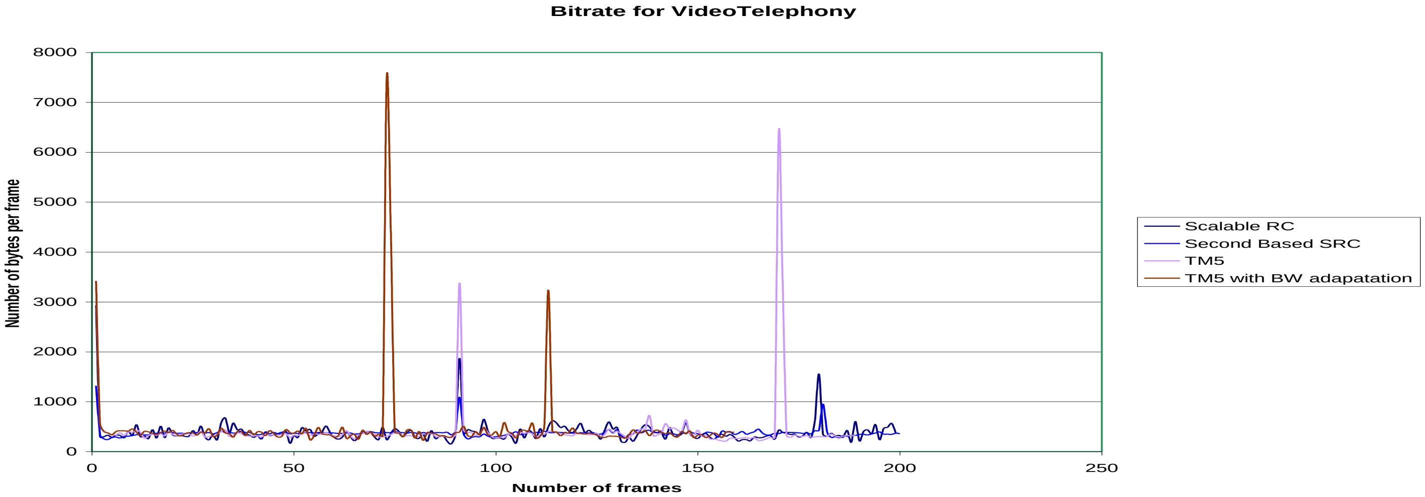


Fig. 2 Bits per frame plot comparing TM5, M-SRC & VT-RC

(2) If the sequence is complex, and the available bit-rate is too limited, then frame skips are expected. But, with the current algorithm, the probability of frame skips is higher towards the end of the 1-second boundaries compared to the same near the starts.

In the above two scenarios, conditions A, B and C listed above are ensured. But, temporal subjective quality is affected, even though the average PSNR over the second or during the complete sequence is as good as it can get, with the available bit-rate.

A smoother gradation in temporal quality can be achieved by using a sliding window for the 1-second duration instead of fixed second boundaries. A circular array can keep track of the bit-consumption for the past frame-rate number of frames at any instant. From the left out available bits in the current running second (‘Bit-rate’ MINUS ‘the running sum of the array’), the bit-target for the current frame can be decided based on its complexity. This way, probability of frame-skip is evenly distributed over the entire sequence, improving the temporal subjective quality.

Also, this probability itself can be reduced by controlling the target bit allocation for a frame at any instant, such that, there is minimum amount of bits left for a complex frame (I-frame, or a high motion P) expected at the next time-stamp. (This is at the expense of a slight reduction in average PSNR, and this amount can be programmable).

B. *Smoother Variation in Spatial Subjective Quality*

(3) Consider a case where, within a frame, there is a considerable gradation in information content, like in a sky-land kind of scene, where the top (sky) part is more or less plain and the bottom (land) part is with lots of complexity.

In the current VT-RC, the target bits allocated for the current frame (TB_c) is a function of “current frame’s SAD-sum”, “left out bits in the current second” and “left out frames in the current second”. TB_c is translated to Quant-Scale for the current frame (QS_c) which is clipped to within 20% of the previous frame’s starting QS. The QS for each MB (slice) is then updated based on the available bits and bit consumption, but clipped to within ±2 of previous MB’s (Slice’s) QS.

In the above mentioned scenario (3), what happens is as the following: The resulting QS_c is quite high for the plain sky region and low for the complex land region. The QS gets reduced, and more and more bits are allotted to the plain sky MBs, as coding progresses from top to bottom. But, by the time the land MBs at the bottom starts getting coded, the QS would have become quite low, and more than ideal bits are already consumed for the low information sky MBs. This will result in a clumsy situation, where there will be lesser bits than required for coding the complex land-part while the current-QS being very low.

This will result in more bits than allowed for the current frame, while leaving poor quality land-part at the bottom. VT-RC fares much better than other RC schemes, like TM5, in this case, but there is further scope for improvement.

Since the distribution of complexity (SAD/variance) in the frame is available before the start of encoding, this problem can be solved in a generic way as follows:

1) Allocate bits for each MB according to its complexity, before starting to encode current frame.
2) Estimate the actual QS required for MBs based on this, and make a sequence of this QS of MBs in raster-scan (coding) order.
3) Modify this QS sequence such that DPCM of QS does not suffer slope overload distortion. (Scaling down the amplitude about the mean - if required - might serve the purpose.)

C. *Strictly Meeting the Target Bits*

Selectively skipping non-zero higher frequency quantized coefficients can ensure strict meeting of target bits for MBs, in the following scenarios where target-meeting is otherwise difficult:

(1) MB-SAD is high, but QS is limited by +2 increment.
(2) QS as demanded by MB-SAD is allotted, but actual bits consumed after coding the MB far exceeds the allotted target bits.

Implementing a per second bit rate control together with the sliding window approach would result in stringent bit rate control with best possible quality for real time video telephony.

References

[1] “Rate Control”, ISO/IEC 14496-2:1999/Amd.1:2000(E), Annex L.

[2] K. Ramkishor and James Mammen, “Technical Report : Rate Control for MPEG4 Video”

[3] TM5 Rate Control : ISO/IEC/JTC1/SC29/WG11, Coded representation of pictures and audio information, MPEG phase 2 test model 5, Apr. 1993 [http://www.mpeg.org/MPEG/MSSG/tm5/index.html]

[4] J. Ribas-Corbera and S. Lei, “Rate control in DCT video coding for low delay communications,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[5] K. Ramkishor, James. P. Mammen, “Bandwidth Adaptation for MPEG-4 Video Streaming over the Internet”, DICTA2002: Digital Image Computing Techniques and Applications, 21--22 January 2002, Melbourne, Australia.

[6] "Video Buffering Verifier", ISO/IEC 14496-2:1999/Amd.1:2000(E), Annex D.

[7] Ramkishor, K.; Raghu, T.S.; Suman, K.; Gupta, P.S.S.B.K., “Adaptation of video encoders for improvement in quality”, Circuits and Systems, 2003. ISCAS apos;03. Proceedings of the 2003 International Symposium on Volume 2, Issue , 25-28 May 2003 Page(s): II-692 - II-695 vol.2

[8] H. J. Lee, T. H. Chiang, and Y. Q. Zhang, “Scalable rate control for MPEG-4 video,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 878–894, Sept. 2000.