IMPROVED FEEDBACK COMPENSATION MECHANISMS FOR MULTIPLE VIDEO OBJECT ENCODING RATE CONTROL

Paulo Nunes[1], *Fernando Pereira*[2]

[1]Instituto Superior de Ciências do Trabalho e da Empresa – Instituto de Telecomunicações
[2]Instituto Superior Técnico – Instituto de Telecomunicações
E-mail: {paulo.nunes, fernando.pereira}@lx.it.pt

ABSTRACT

This paper proposes new buffer and video object distortion feedback compensation mechanisms for efficiently dealing with deviations between the ideal and the actual behavior of video scene encoders when jointly encoding multiple arbitrarily shaped video objects in the context of compliant low-delay object-based MPEG-4 video coding. The proposed solution computes target buffer occupancies for each encoding time instant based on the amount and complexity of the video data to encode, and the bit allocation for each encoding time instant is feedback adjusted according to deviations relatively to this ideal behavior. Additionally, each video object bit allocation is also feedback adjusted based on the relative distortion of the various video objects in the scene. The proposed solution outperforms the non-normative MPEG-4 reference rate control algorithm for a wide range of bit rates and spatio-temporal resolutions, for typical test sequences.

Index Terms — *object-based video coding, multiple video object rate control, MPEG-4 video.*

1. INTRODUCTION

It is widely accepted that, for pleasant visual consumption, the video data should be coded with approximately constant quality or, at least, with smoothly changing quality (both spatially and temporally). Due to the varying scene complexity, hybrid video coding schemes, such as the MPEG-4 video coding scheme [1], produce, typically, a variable number of bits per each encoding time instant, even for slightly changing video quality. Therefore, in buffer/delay constrained constant bit rate (CBR) video encoding, the bit rate variability is handled through a smoothing bitstream buffer in order to achieve a constant average bit rate measured over short periods of time. In this context, the rate controller is faced with two conflicting goals: 1) keep the bitstream buffer occupancy within permitted bounds, which typically requires finely adjusting the encoding parameters, e.g., macroblock (MB) quantization parameter (QP), to produce more/less encoding bits according to the buffer occupancy tendency; 2) adjust the encoding parameters aiming to maximize the subjective quality of the decoded video, which typically requires slowly changing the QP (both spatially, between adjacent MBs, and temporally, between successive encoding time instants).

To accomplish these goals, the rate controller needs: 1) to properly allocate the available bit rate taking into account the video data complexity and the proper video buffering verifier mechanism constraints; 2) to compute the coding parameters that would lead to the estimated bit allocations.

For low-delay video encoding under reasonable encoding complexity, feedback rate control assumes an important role since it is not usually viable to encode the same content multiple times (e.g., for each set of possible encoding parameters) as a pure feedforward rate control solution would require. While feedback methods react *a posteriori* to deviations relatively to the ideal encoding behavior, feedforward methods infer in advance to encoding the results of a given set of encoding decisions. Therefore, for feedback rate control methods, adequate compensation mechanisms are needed for dealing with deviations between the idealized and the actual encoder behavior.

In order to control the output of a multiple video object (MVO) encoder in a way that the video rate buffering verifier (VBV) mechanism is not violated and the quality of the decoded video objects (VOs) is maintained approximately constant (along time and among the several VOs composing the video scene for each encoding time instant), the bit rate allocation should take simultaneously into account the following three aspects: 1) VO coding complexity based on the video object planes (VOPs) content and VO temporal resolution; 2) VO distortion feedback compensation; and 3) Dynamic VBV target buffer occupancy according to the scene coding complexity. Up to the authors' knowledge, these three aspects are not simultaneously handled in published MVO rate control (RC) solutions [2–8], leading to the inefficient use of the coding resources, i.e., buffer size and bit rate.

In this context, this paper proposes an improved bit allocation strategy for low-delay MVO encoding using the VBV mechanism feedback information and the VO encoding quality measured as the mean square error (MSE) between the original and the reconstructed video data. Relatively to the typically used benchmarking for object-based rate control algorithm – the MPEG-4 Video Verification Model (VM) [2], the proposed solution can provide more efficient bit allocation leading to higher average peak signal-to-noise (PSNR) and smooth quality variations, which improve the user experience.

The rest of the paper is organized as follows: Sections 2 and 3 describe, respectively, the proposed bit allocation and VBV control algorithms; Section 4 presents some results evaluating the performance of the proposed solution in comparison with the MPEG-4 VM [2]; and, finally, Section 5 draws some conclusions.

2. BIT ALLOCATION FOR MVO ENCODING

2.1. Group of Scene Planes (GOS) Bit Allocation

The GOS regards the set of all encoding time instants between two random access points, typically encoded with a constant number of bits (in CBR scenarios). GOSs may be composed by VOs with different VOP rates. In the case of a single VO, a GOS becomes a Group of Video Object Planes (GOV). The rate control algorithm aims at allocating a nominal number of bits to each GOS (luminance, chrominances and relevant auxiliary data), $\overline{T}_{GOS}$, that is proportional to the GOS duration, i.e.,

$$\bar{T}_{GOS}[m] = R[m] \times \left(t_{GOS}[m+1] - t_{GOS}[m] \right), \; m = 1, \ldots, N_{GOS} \quad (1)$$

where $R[m]$ and $t_{GOS}[m]$ are, respectively, the average target bit rate and the starting time instant for GOS m, and N_{GOS} is the number of GOSs in the sequence. Deviations from the expected results are compensated through the following feedback compensation equation

$$T_{GOS}[m] = \bar{T}_{GOS}[m] + K_{GOS} \sum_{k=1}^{m-1} \left(\bar{T}_{GOS}[k] - S_{GOS}[k] \right) \quad (2)$$

where $S_{GOS}[k]$ is the number of bits used to encode the GOS k, and K_{GOS} is given by

$$K_{GOS} = 1 / \min\left(\alpha_N, N_{GOS} - m + 1 \right) \quad (3)$$

with $\alpha_N = \max\left[1, \left\lceil 3 B_S / \bar{T}_{GOS}[m] \right\rceil \right]$; B_S is the VBV buffer size.

The rationale for setting $K_{GOS} \leq 1$ is to avoid large quality fluctuations between adjacent GOSs, notably when scene changes occur inside a given GOS, and the bit allocation error compensation would penalize essentially the upcoming GOS, if $K_{GOS} = 1$. With this approach, GOS bit allocation deviations are smoothed through α_N GOSs, if the buffer size is sufficiently large to accommodate these bit rate variations.

2.2. Scene Plane (SP) Bit Allocation

The SP represents the set of VOPs of all VOs to be encoded at a particular encoding time instant (not all VOs of a given scene need to have VOPs to be encoded in every SP). At the SP-level, in order to obtain approximately constant quality along a GOS, each SP should get a nominal target number of bits that is a fraction of the GOS target (2), proportional to the amount and complexity of the VOs to be encoded in that particular time instant. The coding complexity of a given VOP n in SP p of GOS m is, given by

$$X_{VOP}[n] = \alpha_T[n] \cdot \omega[n], \quad n = 1, \ldots, N_{VO} \quad (4)$$

where N_{VO} is the number of VOs in the scene and $\alpha_T[n]$ and $\omega[n]$ are, respectively, the coding type weight ($T \in \{I, P, B\}$) and coding complexity weight – reflecting the texture, shape, and motion (if applicable) coding complexity (see [9] for details) – of VO n in SP p of GOS m. Notice, that $\alpha_T[n] = 0$, if VO n does not have a VOP in SP p of GOS m.

The coding complexity of a given SP p of GOS m is defined as the sum of its VOP complexities defined according to (4), i.e.,

$$X_{SP}[p] = \sum_{n=1}^{N_{VO}} X_{VOP}[n][p], \quad p = 1, \ldots, N_{SP}[m] \quad (5)$$

Therefore, the GOS m coding complexity is the sum of its SP complexities defined according to (5), i.e.

$$X_{GOS}[m] = \sum_{p=1}^{N_{SP}[m]} X_{SP}[p], \quad m = 1, \ldots, N_{GOS} \quad (6)$$

where $N_{SP}[m]$ is the number of scene planes in the GOS m, with $N_{SP}[m] = \left\lceil t_{GOS}[m] \times SR \right\rceil$, where t_{GOS} is the GOS duration and SR is the scene rate.

The nominal target number of bits allocated for each SP in a given GOS is set by the following equation (the GOS index has been dropped for simplicity)

$$\bar{T}_{SP}[p] = T_{GOS} \cdot X_{SP}[p] / X_{GOS}, \quad p = 1, \ldots, N_{SP} \quad (7)$$

The actual SP target is given by the feedback equation

$$T_{SP}[p] = \bar{T}_{SP}[p] + \frac{X_{SP}[p]}{\sum_{k=p}^{N_{SP}} X_{SP}[k]} \sum_{k=1}^{p-1} \left(\bar{T}_{SP}[k] - S_{SP}[k] \right) \quad (8)$$

Notice that, since the proposed solution concerns low-delay encoding scenarios, both X_{SP} and X_{GOS} can only be computed with data from the current or past encoding time instants. In order to obtain approximately constant distortion among consecutive encoding time instants for each VO, the different VOP coding weights are also adapted at the beginning of each GOV, of a given VO, through the following equation

$$\alpha_I = \left(\bar{b}_I / \bar{b}_P \right) \left(\bar{D}_I / \bar{D}_P \right)^{\gamma_T} \quad (9)$$

where $\bar{b}_I$, $\bar{D}_I$, $\bar{b}_P$, and $\bar{D}_P$ are, respectively, the average number of bits per pixel and the average pixel distortion for I- and P-VOPs, computed over window sizes W_I and W_P of past I- and P-VOPs encoding results, and γ_T is a parameter that controls the impact of the average distortion ratios on the estimation of the VO coding weight ($W_I = 3$, $W_P = N_{SP} - 1$, and $\gamma_T = 0.5$).

2.3. Video Object Plane Bit Allocation

At the VOP-level, i.e., inside each SP, in order to obtain approximately constant quality among the several VOPs composing the SP, each VOP should get allocated a nominal target number of bits that is a fraction of the SP target (8), proportional to the relative complexity of the VOP to be encoded in that particular time instant. Therefore, the nominal target number of bits for the VO n VOP in a given SP p of GOS m is given by

$$\bar{T}_{VOP}[n] = T_{SP} \cdot X_{VOP}[n] / X_{SP}, \quad n = 1, \ldots, N_{VO} \quad (10)$$

For MVO encoding, it is important to guarantee that the spatial quality among the different VOs in the scene is kept approximately constant, i.e., an important goal is to encode all the objects in the scene with approximately constant quality. This goal can hardly be achieved when only a pure feedforward approach is used to compute the VO weights used to distribute the SP target among the several VOPs in the given SP. This is the approach followed in [2], where there is no compensation for deviations on the bit rate distribution among the several VOPs for a given encoding time instant. Therefore, it is important to update the VO coding complexity weights along time and to compensate the bit allocation deviations through the feedback adjustment of these parameters in order to meet the requirement of spatial quality smoothness.

In this paper, the following compensation mechanism is proposed aiming at reducing the deviations in the average distortion among the several VOs composing the scene for a given SP. For this purpose, the SP average luminance pixel distortion is defined as the weighted sum of the various VOs distortions, i.e.,

$$D_{SP}[p] = \sum_{k=1}^{N_{VO}} \left(N_{PIX}[k] \cdot D_{VO}[k] \right) \Big/ \sum_{k=1}^{N_{VO}} N_{PIX}[k] \quad (11)$$

where $N_{PIX}[k]$ is the number of pixels in VO k VOP in SP p.

Using (11) as the reference target SP distortion, a complexity weight adjustment is computed for each VO

$$\phi_D[p][n] = \left(\frac{S_{VOP}[p-1][n] / \alpha_T[p-1][n]}{\sum_{k=1}^{N_{VO}} S_{VOP}[p-1][k]} \times \frac{D_{VOP}[p-1][n]}{D_{SP}[p-1]} \right)^{\gamma_D} \quad (12)$$

where γ_D is a parameter to control the impact of ϕ_D in the VOP bit allocation feedback compensation (typically, $0.1 \leq \gamma_D \leq 0.5$; in this paper, $\gamma_D = 0.2$ has been used).

From (4) and (12), the VO complexity is feedback-adjusted as

$$\eta_D[n] = \phi_D[n] \cdot X_{VOP}[n] \quad (13)$$

and subsequently

$$T_{VOP}[n] = T_{SP}\frac{\eta_D[n]}{\sum_{k=1}^{N_{VO}}\eta_D[k]} + \frac{\eta_D[n]}{\sum_{k=n}^{N_{VO}}\eta_D[k]}\sum_{k=1}^{n-1}\left(T_{SP}\frac{\eta_D[k]}{\sum_{k=1}^{N_{VO}}\eta_D[k]} - S_{VOP}[k]\right) \quad (14)$$

2.4. Macroblock Bit Allocation

The MB is the smallest coding unit for which the QP can be changed. At the MB-level, i.e., inside each VOP, in order to obtain approximately uniform quality among the several non-transparent MBs, each MB should get a nominal target number of bits that is a fraction of the VOP target (14), proportional to the relative complexity of the MB to be encoded in that particular VOP, X_{MB} – in this paper, the MB prediction error mean absolute difference (MAD) is used. Therefore, the nominal and the actual target number of bits for each MB in a given VOP, respectively $\overline{T}_{MB}[i]$ and $T_{MB}[i]$, are given by

$$\overline{T}_{MB}[i] = T_{VOP}\cdot X_{MB}[i]\Big/\sum_{k=1}^{N_{MB}} X_{MB}[k], \quad i = 1,\ldots,N_{MB} \quad (15)$$

$$T_{MB}[i] = \overline{T}_{MB}[i] + K_{MB}\sum_{k=1}^{i-1}\left(\overline{T}_{MB}[k] - S_{VOP}[k]\right) \quad (16)$$

where N_{MB} is the number of MBs in the VOP being encoded and

$$K_{MB} = \max\left[X_{MB}[i]\Big/\sum_{k=i}^{N_{MB}} X_{MB}[k], 1\Big/\min\left[N_{MB}-i+1, 16\right]\right] \quad (17)$$

The rationale for (17) is the following: at the beginning of the VOP encoding, the bit allocation errors are compensated at most along the subsequent 16 MBs (this value has been set empirically), avoiding slow reactions for MBs with low complexities, i.e., low MADs; as the encoding proceeds to the last MBs, the K_{MB} factor distributes the accumulated MB bit allocation error through the remaining MBs to be encoded.

For a fine allocation of bits inside each VOP, the rate control algorithm computes a QP for each MB, taking into account the complexity of the several MBs to encode using the following MB-level rate-distortion function

$$R_{MB}(Q) = \left(a/Q_{MB}\right)\cdot X_{MB} \quad (18)$$

where *a* is the model parameter estimated after each MB encoding.

3. ADDING A NOVEL VIDEO BUFFERING VERIFIER CONTROL COMPENSATION

In order to efficiently use the available buffer space, for each SP of each GOS, the target VBV buffer occupancy (immediately before removing all SP VOPs from the VBV buffer) is computed by

$$B_T[p] = B_S - \left(T_{GOS}\frac{\sum_{k=1}^{p-1} X_{SP}[k]}{X_{GOS}} - \left(t_{SP}[p] - t_{SP}[1]\right)\times R\right) - B_L \quad (19)$$

where B_S is the VBV buffer size, T_{GOS} is the target number of bits for encoding the whole GOS *m* given by (2), $X_{SP}[k]$ is the SP *k* complexity given by (5), X_{GOS} is the GOS *m* complexity given by (6), $t_{SP}[p]$ is the time instant of SP *p*, R is the average output target bit rate for GOS *m*, and B_L is the VBV underflow margin as explained in the following paragraphs.

Since at the beginning of each GOS all VOPs are Intra coded, this will typically lead to the highest level in terms of encoder rate buffer occupancy, as Intra coded VOPs require usually more bits for achieving the same spatial quality than Inter coded VOPs. Consequently, in terms of VBV occupancy, this will correspond to the highest occupancy immediately before removing the first VOPs of a GOS and the lowest VBV occupancy immediately after removing these VOPs from the VBV buffer. Therefore, in nominal terms, the available VBV margin is defined by the available encoder rate buffer space immediately after adding the encoded bits of the first SP in the GOS, or, in terms of VBV occupancy, by the occupancy of the VBV buffer immediately after removing the bits of the first SP VOPs.

Since VBV buffer underflow (encoder rate buffer overflow) is more critical than VBV buffer overflow (encoder rate buffer underflow), it is convenient to unequally distribute the VBV margin over these two critical zones. Therefore, at the beginning of each GOS, the VBV margin is computed as follows

$$B_M = B_S - T_{GOS}\cdot X_{SP}[1]/X_{GOS} \quad (20)$$

The nominal free space in the buffer (20) is unequally divided as $B_L = \beta_{VBV}\times B_M$ and $B_U = (1-\beta_{VBV})\times B_M$, with $\beta_{VBV} = 0.9$.

Based on (19), the target number of bits used to encode the corresponding SP (8) is further adjusted by a multiplicative factor

$$K_{VBV} = \begin{cases} 1-\alpha_{VBV}\left(\left(B_T - B\right)/B_T\right) & \Leftarrow B \le B_T \\ 1+\alpha_{VBV}\left(\left(B - B_T\right)/\left(B_S - B_T\right)\right) & \Leftarrow B > B_T \end{cases} \quad (21)$$

where $\alpha_{VBV} = 0.25$ is a controller parameter set empirically

The rationale for (21) is to decrease the SP bit allocation if the VBV buffer is approaching underflow (i.e., too many bits have been generated by the encoder in the past) and to increase the SP bit allocation if the VBV buffer is approaching overflow (i.e., too few bits have been generated by the encoder in the past). Therefore, the bit allocation given by (8) is adjusted as follows

$$T_{SP}^{SBC}[p] = T_{SP}[p]\times K_{VBV} \quad (22)$$

In some extreme cases, notably for small buffer sizes, this soft SP-level VBV control may lead to SP bit allocations near imminent violations of the VBV mechanism; therefore, whenever this situation occurs, a further adjustment is performed in order to guarantee that the SP bit allocation will keep the VBV occupancy within the nominal VBV operation area defined by

$$\beta_L\times B_S \le B \le \beta_U\times B_S, \text{ with } \beta_L = 0.05 \text{ and } \beta_U = 1.0 \quad (23)$$

4. EXPERIMENTAL RESULTS

In this section, the performance for MVO encoding of the proposed rate control solution (so called IST solution) is compared with the MPEG-4 VM5 rate control algorithm initially described in [2]. Two random access conditions are tested: 1) one random access point (I-VOP) every second – label IP = 1s; and 2) one single random access point at the beginning of the sequence (IPP…) – label IP = 10s. The VBV buffer size is set numerically to *R*/2 bits (*R* - target bit rate). Four representative test sequences at 30 Hz and with 300 frames have been selected: *Stefan* and *Bream* with 2 VOs; and *Coastguard* and *News* with 4 VOs. These sequences can be grouped according to their motion activity into: 1) high-motion video sequences (*Stefan* and *Coastguard*), and 2) low-motion video sequences (*Bream* and *News*).

The two rate control solutions are compared in terms of the so called average scene quality, measured as the luminance Average Scene PSNR between the original and the reconstructed video frames at the decoder using the tool for compactly comparing two PSNR curves developed by the ITU-T Video Coding Experts Group [10] (see Table I). The so called Scene PSNR Variation is also used for assessing the quality smoothness between the various VOs in the scene; it is computed as the ratio between the Average Scene PSNR Difference and the Average Scene PSNR, where the

first is the weighted sum of the absolute difference between each VO PSNR and the Scene PSNR for each SP (weighted by the relative size of each VO).

Table I illustrates the PSNR gains and bit rate reductions for the proposed solution in two different conditions: I) Proposed MVO RC solution with VM5 VBV control, i.e., without Sec. 3; II) Proposed MVO RC solution with the VBV control proposed in Sec. 3 (IST label). These results support the following conclusions:

- Both cases (I and II) have higher PSNR gains (thus also bit rate reductions) for IP = 1s tests due to the more efficient bit allocation and the finer QP (MB-level) control as illustrated also in Figure 1 (top) for *Stefan* under various encoding conditions. PSNR gains for case II can be as high as 2.6 dB.
- For the less demanding scenarios, i.e., IP = 10s and low-motion sequences, VM5 performs slightly better than case I and close to case II, in terms of Average Scene PSNR, due to the high coding quality of the easy-to-code background VOs. However, the Scene PSNR Variation is lower (smoother quality) for cases I and II, as illustrated in Figure 2 for *Bream* (case II).
- Case II provides an additional PSNR gain, relatively to case I, of approximately 0.9 dB on average (5% less bit rate), reducing also the amount of skipped frames as illustrated in Figure 1 (bottom) for *Stefan* (shown as severe PSNR drops).

5. CONCLUSION

This paper proposes two improved feedback mechanisms (VO distortion feedback and video rate buffer feedback) for low-delay MVO MPEG-4 encoding. The proposed solution is compared with the MPEG-4 VM5 solution [2] with the main conclusion that the proposed MVO RC solution clearly outperforms the benchmarking solution in terms of the average quality and quality smoothness, resulting in a more efficient use of the available resources, i.e., the buffer space and the target bit rate.

Table I – Average PSNR and bit rate gains [CIF@15Hz]

Sequence	PSNR [dB]				Bit Rate [%]			
	IP = 1s		IP = 10s		IP = 1s		IP = 10s	
	I	II	I	II	I	II	I	II
Stefan	1.98	2.57	1.91	2.15	-30.4	-38.8	-34.5	-37.7
Coastguard	0.61	1.13	0.58	0.55	-12.6	-21.8	-11.8	-10.9
Bream	-0.33	0.80	-0.13	-0.06	6.2	-15.5	2.5	1.1
News	1.07	2.50	-0.78	-0.49	-54.7	-33.9	16.8	8.9
	0.83	**1.75**	**0.40**	**0.54**	**-22.9**	**-27.5**	**-6.8**	**-9.7**

6. REFERENCES

[1] ISO/IEC 14496-2:2001, "Information Technology – Coding of Audio-Visual Objects – Part 2: Visual (2nd Ed.)", 2001.

[2] MPEG Video, "MPEG-4 Video Verification Model 5.0", Doc. N1469, Maceió MPEG meeting, Nov. 1996.

[3] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 Rate Control for Multiple Video Objects", IEEE Trans. on CSVT, vol. 9, no. 1, pp. 186-199, Feb. 1999.

[4] J. Ronda, M. Eckert, F. Jaureguizar, and N. Garcia, "Rate Control and Bit Allocation for MPEG-4", IEEE Trans. on CSVT, vol. 9, no. 8, pp. 1243-1258, Dec. 1999.

[5] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable Rate Control for MPEG-4 Video", IEEE Trans. on CSVT, vol. 10, no. 6, pp. 878-894, Sep. 2000.

The work presented was developed within VISNET I and VISNET II, European Networks of Excellence (http://www.visnet-noe.org).

[6] P. Nunes and F. Pereira; "Scene Level Rate Control Algorithm for MPEG-4 Video Encoding", Proc. of VCIP'01, San Jose, CA, USA, vol. 4310, pp. 194–205, Jan. 2001.

[7] Y. Sun and I. Ahmad, "A Robust and Adaptive Rate Control Algorithm for Object-based Video Coding", IEEE Trans. on CSVT, vol. 14, no. 10, pp. 1167-1182, Oct. 2004.

[8] Y. Sun and I. Ahmad, "Asynchronous Rate Control for Multi-Object Videos", IEEE Trans. on CSVT, vol. 15, no. 8, pp. 1007-1018, Aug. 2005.

[9] P. Nunes and F. Pereira, "Rate Control for Scenes with Multiple Arbitrarily Shaped Video Objects", Proc. of PCS'97, Berlin, Germany, pp. 303–308, Sep. 1997.

[10] G. Bjontegaard, "Calculation of Average PSNR Differences Between RD-curves", VCEG-M33, Austin, TX, USA, Apr. 2001.

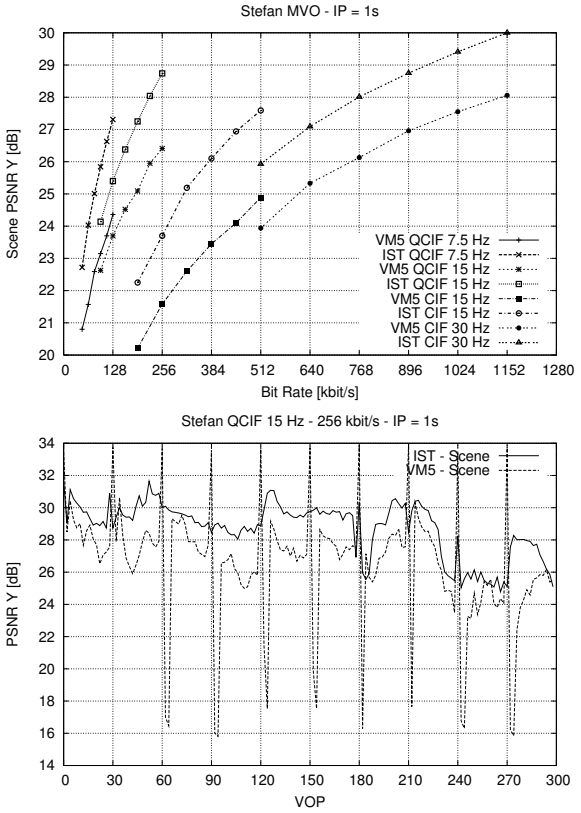

Figure 1. *Stefan* (IP = 1s): (top) Average Scene PSNR versus bit rate; (bot.) Scene PSNR evolution QCIF@15 Hz [256 kbit/s]

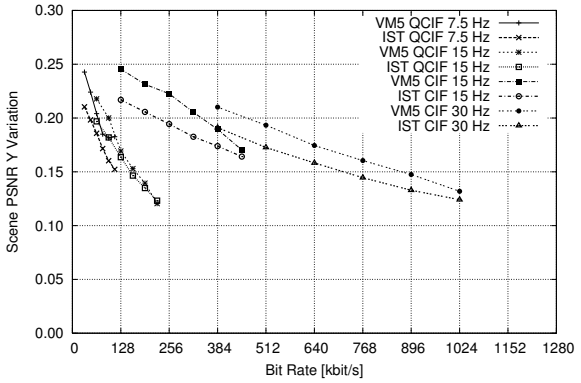

Figure 2. *Bream* (IP = 10s) Scene PSNR Variation versus bit rate