Sergio Sanz-Rodríguez, Fernando Díaz-de-María, Mehdi Rezaei

Low-complexity VBR controller for spatial-CGS and temporal scalable video coding

LOW-COMPLEXITY VBR CONTROLLER FOR SPATIAL-CGS AND TEMPORAL SCALABLE VIDEO CODING

Sergio Sanz-Rodríguez, Fernando Díaz-de-María

Department of Signal Theory and Communications
Universidad Carlos III, Leganés (Madrid), Spain
{sescalona, fdiaz}@tsc.uc3m.es

Mehdi Rezaei

Faculty of Electrical and Computer Engineering
University of Sistan & Baluchestan, Iran
mehdi.rezaei@ieee.org

ABSTRACT

This paper presents a rate control (RC) algorithm for the scalable extension of the H.264/AVC video coding standard. The proposed rate controller is designed for real-time video streaming with buffer constraint. Since a large buffer delay and bit rate variation are allowed in this kind of applications, our proposal reduces the quantization parameter (QP) fluctuation to provide consistent visual quality bit streams to receivers with a variety of spatio-temporal resolutions and processing capabilities. The low computational cost is another characteristic of the described method, since a simple lookup table is used to regulate the QP variation on a frame basis.

Index Terms— Rate Control, Variable Bit Rate, Scalable Video Coding, H.264/AVC

1. INTRODUCTION

Nowadays, the modern RTP/IP-based transmission systems, such as Internet and wireless networks, are becoming more and more popular in video communications. For this kind of channels, scalable video coding (SVC) is able to provide bit rate adaptation to varying channel conditions as well as to receiving devices with heterogeneous display and computational capabilities. SVC allows to extract from a high-quality bit stream, either one or a subset of bit streams with lower spatio-temporal resolutions or reduced qualities that can be decoded by any target receiver.

The scalable extension of H.264/AVC video coding standard has been recently standardized [1] and evaluated [2]. It provides both coding efficiency and decoding complexity similar to that achieved using the single-layer coding. H.264/SVC supports spatial, temporal and quality scalable coding. With the spatial scalability, a layered coding approach is used to encode different pictures sizes of an input video sequence. The base layer provides an H.264/AVC compatible bit stream with the lowest spatial resolution, while higher picture sizes are encoded by the enhancement layers. In addition, the redundancies between consecutive layers can be exploited via inter-layer prediction tools in order to improve the coding efficiency.

Each spatial layer is capable of supporting temporal scalability by using hierarchical prediction structures, that go from those very efficient ones using hierarchical B pictures to those low-delay structures with zero structural delay. The pictures which belong to the base temporal layer can only use the previous pictures of the same layer as references. The pictures of an enhancement temporal layer can be bidirectionally predicted by pictures of a lower layer. The number of temporal layers in a spatial layer is determined by the group of pictures (GOP) size, defined in SVC as the distance between two consecutive I or P pictures, also named as key pictures.

With the quality or signal-noise ratio (SNR) scalability, different reconstruction fidelities with the same spatio-temporal resolution can be provided in the bit stream. The H.264/SVC standard defines two types of SNR scalability: coarse grain (CGS) and medium grain (MGS) scalable coding. The first one is a special case of spatial scalability with identical pictures sizes. The second one employs a multilayer approach within a spatial layer to provide a finer bit rate granularity in the rate-distortion (R-D) space.

Then, if we assume a video application with different connection qualities and heterogeneous target receivers, a set of valid scalable bit streams with a variety of bit rates for each available spatio-temporal resolution must be provided. To this end, a RC algorithm is included in the SVC encoder so that the high-quality bit stream and the remaining sub bit streams comply with the hypothetical reference decoder (HRD), a normative part of the standard that describes a set of requirements to transmit and decode bit streams. Although several RC schemes have been proposed for H.264/SVC [3, 4, 5, 6], neither of them supports full scalability. The RC adopted by the Joint Video Team (JVT) group [3] is partially developed, since only the base spatial layer encoding is supported. On the other hand, other more complete schemes support spatial and quality (CGS) scalability [4, 5, 6], but the HRD compliance in receivers with lower frame rates is not guaranteed.

In this paper a variable bit rate (VBR) controller with buffer constraint for spatial, CGS and temporal scalable video coding is proposed for real-time video streaming. A long-term bit rate adaptation is feasible in this kind of applications in order to improve the visual quality in comparison to the constant bit rate (CBR) schemes required by low-delay environments [4, 6]. To this purpose, the provided QP variation on a frame basis is smooth enough to minimize the quality fluctuation. Another characteristic of the proposed RC scheme is its low computational cost, since a simple lookup table is employed to determine the QP value of each spatial layer.

The paper is organized as follows. In Section 2 an overview of the proposed RC scheme is given. In Section 3 the QP selection method is described. Section 4 shows and analyzes the experimental results. Finally, the conclusions and future work are outlined in Section 5.

2. ALGORITHM OVERVIEW

For a given real-time video streaming service, it is assumed that the SVC encoder involves D spatial (or quality for the CGS case) layers with identifiers $\{0, \ldots, d, \ldots, D-1\}$, and each one of them contains $T^{(d)}$ temporal layers with identifiers $\{0, \ldots, t, \ldots, T^{(d)}-1\}$. Furthermore, assuming that most of receivers do not work with very low frame rates, a parameter $t_{min}^{(d)}$ is defined as the minimum available temporal layer identifier for the d^{th} spatial layer.

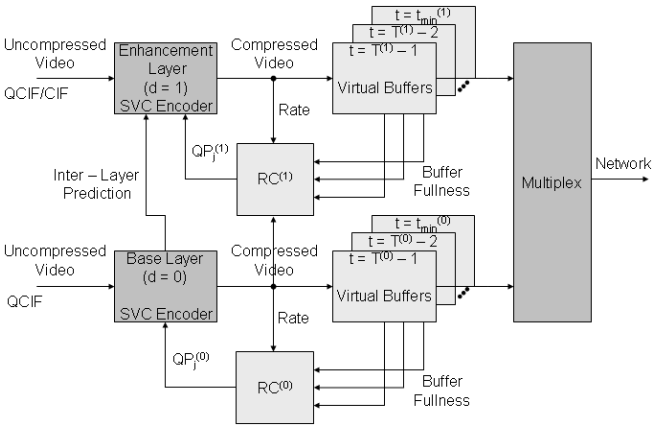

Fig. 1. Block diagram of the proposed SVC rate control scheme for two spatial layers ($D = 2$).

The proposed RC scheme is illustrated in Fig. 1. A rate controller, $RC^{(d)}$, and a set of virtual buffers associated to temporal layers from $t_{min}^{(d)}$ to $T^{(d)} - 1$ are incorporated in each spatial layer. Placed between the encoder and the network, each virtual buffer simulates the encoder buffering process of a sub bit stream encoded with given target bit rate and frame rate $\left(R^{(d,t)}, f^{(d,t)}\right)$, where

$$R^{(d-i,t-j)} \leq R^{(d,t)} \leq R^{(D-1,T^{(D-1)}-1)}, \tag{1}$$

with $i = 0, \ldots, d$ and $j = 0, \ldots, t$, and

$$f^{(d,t)} = f_{in}^{(d)} \times 2^{-(T^{(d)}-1-t)}, \tag{2}$$

where $f_{in}^{(d)}$ represents the input sequence frame rate. Note that $R^{(d,t)}$ must be higher than those bit rates of lower spatio-temporal pictures, since they are also included in the sub bit stream (d, t).

Then, if we assume that the j^{th} picture belonging to the spatio-temporal layer (d, t) is to be encoded, the aim of $RC^{(d)}$ is to obtain an appropriate QP value, $QP_j^{(d)}$, on a frame basis such that the QP fluctuation is minimized to improve the visual quality, while the buffer fullness $V^{(d,k)}$, with $k = max(t_{min}^{(d)}, t), \ldots, T^{(d)} - 1$, are maintained in secure levels after encoding the picture. All of these buffers must be checked to determine the QP value, since a good behavior is not ensured if only one of them is monitored.

Two kinds of signals are used as inputs to $RC^{(d)}$: 1) the fullness of the corresponding virtual buffers, and 2) the access unit (AU) bit rate, with an AU consisting of all representations of a picture with different spatial layer identifier for a given time instant. A QP increment, $\Delta QP^{(d)}$, is computed from these data and the $QP_j^{(d)}$ value is then determined using the following expression:

$$QP_j^{(d)} = \begin{cases} \overline{QP}_{GOP}^{(d)} + \Delta QP^{(d)} & \text{if key picture} \\ QP_{j-1}^{(d)} + \Delta QP^{(d)} & \text{otherwise,} \end{cases} \tag{3}$$

where $\overline{QP}_{GOP}^{(d)}$ is the average QP value of the last encoded GOP. In the case of CGS scalability, $QP_j^{(d)}$ is bounded with respect to the QP value of same picture size of the lower spatial layer:

$$QP_j^{(d)} = min\left(QP_j^{(d-1)}, QP_j^{(d)}\right). \tag{4}$$

Thus, a higher quality for the enhancement d^{th} layer is ensured. Note that $QP_j^{(d)}$ must be an integer value between 0 to 51 for H.264/SVC standard.

3. QP INCREMENT SELECTION

The aim of this section is to describe those details of the RC algorithm which determine the desired QP increment, $\Delta QP^{(d)}$, to be used in (3) for the current j^{th} picture with identifier (d, t).

3.1. Access unit bit rate

Assuming that the picture coding order in SVC is on an AU basis [1], the number of bits generated by the current AU obeys:

$$AU^{(d,t)} = \sum_{m=0}^{d} b_j^{(m,t)}, \tag{5}$$

where $b_j^{(m,t)}$ is the amount of bits generated by the j^{th} picture with identifier (m, t) belonging to the AU.

3.2. Average complexity of a layer (d,t)

The proposed RC method defines $\overline{X}^{(d,t)}$ as the average complexity of all the encoded pictures belonging to the t^{th} temporal layer until the d^{th} spatial layer. It is updated using the following expression:

$$\overline{X}^{(d,t)} = \alpha \sum_{m=0}^{d} \left(Q_j^{(m)} b_j^{(m,t)}\right) + (1-\alpha)\overline{X}^{(d,t)}, \tag{6}$$

where $Q_j^{(m)}$ is the quantization step value associated to $QP_j^{(m)}$, and α is a forgetting factor which is set to 0.875 in the experiments.

3.3. Lookup-table-based rate controller

In order to determine the $\Delta QP^{(d)}$ value for the current picture, the lookup table shown in Table 1 is used. Specifically, a set of QP increments, $\Delta QP^{(d,k)}$, for $k = max(t_{min}^{(d)}, t), \ldots, T^{(d)} - 1$, is obtained from the normalized buffer fullness and output AU bit rate signals, $A^{(d,k)}$ and $B^{(d,k)}$, respectively. Letters VH (Very High), H (High), M (Medium), L (Low) and VH (Very Low) represent different buffer and rate regions, which are delimited by the thresholds given in Fig. 2. The lookup table and thresholds values have been optimized in such a way that the average fullness in each buffer is about 40% of the buffer size during the encoding process, since the overflow risk is more critical than that of underflow when a key picture is encoded. The final QP increment, $\Delta QP^{(d)}$, is computed by rounding the arithmetic mean of the obtained $\Delta QP^{(d,k)}$ values. Afterward, $QP_j^{(d)}$ is determined using the expressions (3) and (4).

Finally, after encoding the j^{th} picture with identifier (d, t), signals $A^{(d,k)}$ and $B^{(d,k)}$ are updated as follows:

1. Compute $AU^{(d,t)}$ using eq. (5);
2. Update $\overline{X}^{(d,t)}$ using eq. (6);
3. For $k = max(t_{min}^{(d)}, t)$ to $T^{(d)} - 1$, do

a. $$V^{(d,k)} = V^{(d,k)} + AU^{(d,t)} - \frac{R^{(d,k)}}{f^{(d,k)}}; \tag{7}$$

b. $$T^{(d,k)} = \frac{R^{(d,k)}}{f^{(d,k)}} \left(\frac{\overline{X}^{(d,t)} \sum_{n=0}^{k} N^{(d,n)}}{\sum_{n=0}^{k} \left(\overline{X}^{(d,n)} N^{(d,n)}\right)} \right); \tag{8}$$

c. $$A^{(d,k)} = \frac{V^{(d,k)}}{BS^{(d,k)}}; \tag{9}$$

d. $$B^{(d,k)} = \frac{AU^{(d,t)}}{T^{(d,k)}}; \tag{10}$$

Table 1. Lookup table for $\Delta QP^{(d,k)}$ selection.

$\mathbf{B^{(d,k)}}$ \ $\mathbf{A^{(d,k)}}$	**VH**	**H**	**M**	**L**	**VL**
VH	4	3	2	1	0
H	3	2	1	0	-1
M	2	1	0	-1	-2
L	1	0	-1	-2	-3
VL	0	-1	-2	-3	-4

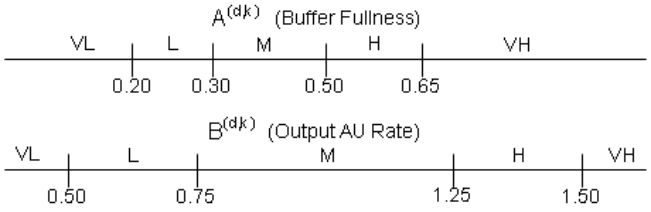

Fig. 2. Threshold values for the $A^{(d,k)}$ and $B^{(d,k)}$ signals.

where $BS^{(d,k)}$ is the buffer size of the layer (d,k), $T^{(d,k)}$ is the amount of target bits to encode the AU in the sub bit stream (d,k), and $N^{(d,n)}$ is the total number of pictures in the n^{th} temporal layer within the GOP.

4. EXPERIMENTS AND RESULTS

The Joint Scalable Video Model (JSVM) reference software JSVM 11 [7] was used to implement the proposed RC algorithm. In order to produce VBR bit streams with full scalability the H.264/SVC encoder was configured as follows:

- **Number of pictures:** 900
- **GOP size/Intra period:** 8/8
- **GOP structure**: Hierarchical B pictures
- **Number of spatial layers:** $D=3$
  - **d=0:** QCIF, $f_{in}^{(0)}=12.5$ Hz ($T^{(0)}=3$)
  - **d=1:** CIF, $f_{in}^{(1)}=12.5$ Hz ($T^{(1)}=3$)
  - **d=2:** CIF, $f_{in}^{(2)}=25$ Hz ($T^{(2)}=4$)

Two sets of color video sequences with a variety of complexities were employed in the experiments. The first one consisted of the following well-known sequences: "Crew", "Foreman", "News", "Paris", "Silent" and "Soccer". These sequences were linked several times to reach the number of frames to be encoded. The second set consisted of the following sequences with scene changes: "Soccer-Mobile-Foreman", "Airshow" (documentary), "Ice" (cartoon), "Nature" (documentary) and "The Lord of the Rings" (movie). The four last sequences were decompressed from high-quality DVD disks and then downsampled to obtain the required QCIF and CIF formats.

In order to assess the performance of our proposal, it was compared to constant QP encoding as an example of constant quality within a scene. Firstly, all sequences were encoded several times with constant QP. For the base spatial layer, $QP_j^{(0)}$ was set to $\{26,28,30,32,34\}$, and the QP values for the second and third enhancement layers were selected for achieving about the same and 3 dB higher qualities, respectively. Then, the spatio-temporal bit rates obtained for each encoding were used as target bit rates, $R^{(d,t)}$, by the RC scheme. The minimum available temporal layer identifiers, $t_{min}^{(d)}$, were set to 1 in order to provide valid streams with temporal resolutions of 6.25 Hz (t=1), 12.5 Hz (t=2) and 25 Hz (t=3). The initial QP values, $QP_1^{(d)}$, were the same as those used by the constant QP encoding, and the buffer sizes, $BS^{(d,t)}$, were set to $3\times R^{(d,t)}$ bits (3 s) with target fullness equal to the 40% of the buffer size.

Table 2. Average results achieved by the proposed RC. Incremental measurements are given with respect to constant QP.

Layer (d,t)	$\mathbf{\Delta\mu_{PSNR}}$ **(dB)**	$\mathbf{\Delta\overline{\sigma}^2_{PSNR,j}}$ **(dB)**	**Bit Rate Error (%)**	**Mean Buffer Level (%)**
(0,1)	0.05	0.31	0.15	40.89
(0,2)	0.07	0.28	0.00	39.83
(1,1)	0.06	0.24	-0.01	40.07
(1,2)	0.06	0.22	0.05	39.81
(2,1)	0.05	0.24	-0.37	38.68
(2,2)	0.04	0.22	0.00	39.13
(2,3)	0.04	0.25	-0.46	36.34

Luminance mean peak SNR gain based on the Bjøntegaard recommendation [8], $\Delta\mu_{PSNR}$, and output bit rate error were some of the metrics employed to compare both methods. The average results on all the test video sequences for target μ_{PSNR} values close to 35 dB (d=0), 35 dB (d=1) and 38 dB (d=2) are summarized in Table 2. As it can be observed, the proposed RC achieves similar performance to that of constant QP encoding. Furthermore, the average results of the output bit rate error, which are close to zero percent, and the mean buffer level, which are about the target fullness, provided by our proposal indicate a good long-term adjustment to the target rate. Tables 3 and 4 show a detailed comparison of the RC scheme and constant QP encoding for two representative sequences, "News" and "The Lord of the Rings", respectively. Observing these results, we conclude that while in stationary sequences our proposal obtains results close to the constant quality case, in those video sequences with time-varying complexity the mean quality becomes even higher than that achieved by the constant QP encoding.

Representative behaviors of the buffer fullness, PSNR and QP evolutions are illustrated in Fig. 3, in which the high-quality encoded version (d=2) of "The Lord of the Rings" was selected as an example of video sequence with time-varying complexity. For each temporal layer, the high correlation between the constant QP encoding and the proposed scheme in PSNR terms is mainly due to the smooth QP fluctuation owing to the proposed VBR algorithm. The buffer fullness within the spatial layer is properly used to improve the mean quality without incurring in overflows or underflows in most sequences exhibiting high temporal heterogeneity.

Although the constant QP encoding generally provides a smaller PSNR variance (see Fig. 3), we should be aware that the buffer constraint is not taken into account, i.e., very high instantaneous bit rate variations are allowed to reduce the dynamic range of PSNR in the scene changes. Therefore, the PSNR variance is not an appropriate metric to evaluate the visual quality consistency within a scene when unconstrained buffer methods are used as reference schemes. The PSNR local variance was, instead, selected for this purpose:

$$\sigma^2_{PSNR,j}=\frac{1}{W}\sum_{i=j-W/2}^{j+W/2-1}\Big(PSNR_i-\mu_{PSNR,W}\Big)^2, \qquad (11)$$

where W denotes the window size in number of pictures, and $\mu_{PSNR,W}$ is the mean PSNR for the given window size. W is set to 2^{t+1} pictures in the experiments, which is short enough to minimize the influence of the PSNR leaps between scenes. Thus, small

Table 3. Constant QP vs. proposed RC for "News".

Layer (d,t)	Algorithm	$\mu_{\mathbf{PSNR}}$ (dB)	$\overline{\sigma}^2_{\mathbf{PSNR,j}}$ (dB)	Bit Rate (kbps)	$\Delta\mu_{\mathbf{PSNR}}$ (dB)
(0,1)	**Const. QP**	34.85	0.02	63.89	0.02
	Proposed	34.78	0.16	63.38	
(0,2)	**Const. QP**	34.91	0.02	71.93	0.06
	Proposed	34.86	0.14	71.15	
(1,1)	**Const. QP**	34.63	0.01	167.37	0.01
	Proposed	34.63	0.08	166.81	
(1,2)	**Const. QP**	34.64	0.01	191.76	0.03
	Proposed	34.66	0.07	191.00	
(2,1)	**Const. QP**	37.46	0.02	266.99	0.01
	Proposed	37.41	0.19	263.90	
(2,2)	**Const. QP**	37.52	0.02	309.99	0.04
	Proposed	37.51	0.15	306.60	
(2,3)	**Const. QP**	37.60	0.02	343.87	0.01
	Proposed	37.62	0.14	342.42	

Table 4. Constant QP vs. proposed RC for "The Lord of the Rings".

Layer (d,t)	Algorithm	$\mu_{\mathbf{PSNR}}$ (dB)	$\overline{\sigma}^2_{\mathbf{PSNR,j}}$ (dB)	Bit Rate (kbps)	$\Delta\mu_{\mathbf{PSNR}}$ (dB)
(0,1)	**Const. QP**	35.34	0.65	80.93	0.11
	Proposed	35.52	1.27	81.84	
(0,2)	**Const. QP**	35.22	0.64	109.65	0.20
	Proposed	35.45	1.20	110.17	
(1,1)	**Const. QP**	35.03	1.08	199.68	0.15
	Proposed	35.23	1.45	201.91	
(1,2)	**Const. QP**	34.84	1.05	283.51	0.17
	Proposed	35.06	1.38	286.36	
(2,1)	**Const. QP**	37.65	0.97	359.86	0.11
	Proposed	37.76	1.25	361.88	
(2,2)	**Const. QP**	37.44	0.90	523.38	0.10
	Proposed	37.57	1.15	528.45	
(2,3)	**Const. QP**	37.45	0.86	699.79	0.15
	Proposed	37.57	1.22	698.94	

$\sigma^2_{PSNR,j}$ values indicate smooth short-term PSNR fluctuations and, therefore, good visual quality. In order to summarize the results in a unique measurement, the mean value of the PSNR local variance, $\overline{\sigma}^2_{PSNR,j}$, was computed. As shown in Tables 2, 3 and 4, the differences in terms of the mean PSNR local variance are not significant; therefore, the RC algorithm is capable of achieving similar visual qualities to those of constant QP encoding.

Similarly to [9], our proposal is also characterized by its low complexity, since few operations are required for the QP selection when compared to the methods based on R-D modeling, such as [3, 5, 6].

5. CONCLUSIONS AND FURTHER WORK

In this paper, a simple lookup-table-based VBR controller for H.264/SVC has been presented for real-time video streaming. The proposed RC provides buffer-constrained bit streams with spatial, CGS and temporal scalability. In comparison to constant QP encoding, the experimental results indicate that our proposal achieves similar mean PSNR and good visual quality consistency within a video scene in every considered spatio-temporal resolution, while the buffer overflows are underflows are avoided. Furthermore, a tight long-term adjustment to the target bit rate is ensured as well.

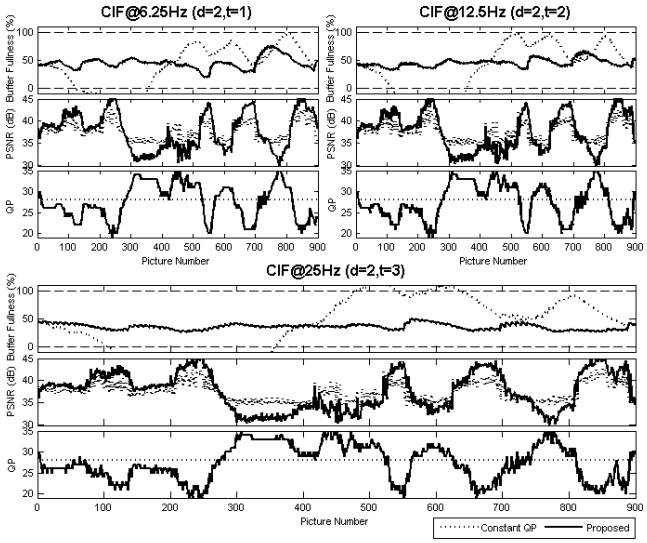

Fig. 3. Encoder buffer fullness, PSNR and QP evolutions. Layers (d=2,t={1,2,3}) from "The Lord of the Rings". High-quality plots corresponding to every layer (d,t) are available in [10].

As future work, fuzzy techniques could be applied, as in [9], in order to reduce even more the QP variation, specially when the buffer fullness fluctuates between two consecutive buffer regions.

6. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1103–1120, Sept. 2007.

[2] M. Wien, H. Schwarz, and T. Oelbaum, "Performance Analysis of SVC," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 17, no. 9, pp. 1194–1203, Sept. 2007.

[3] A. Leontaris and A. M. Tourapis, "Rate control for the Joint Scalable Video Model (JSVM)," *Video Team of ISO/IEC MPEG and ITU-T VCEG, JVT-W043, San Jose, California*, April 2007.

[4] T. Anselmo and D. Alfonso, "Buffer-based constant bit-rate control for scalable video coding," in *Picture Coding Symposium, 2007. PCS 2007*, Nov. 2007.

[5] L. Xu, W. Gao, X. Ji, D. Zhao, and S. Ma, "Rate control for spatial scalable coding in SVC," in *Picture Coding Symposium, 2007. PCS 2007*, Nov. 2007.

[6] Y. Liu, Y.C. Soh, and Z.G. Li, "Rate control for spatial/CGS scalable extension of H.264/AVC," May 2007, pp. 1746–1750.

[7] J. Vieron, M. Wien, and H. Schwarz, "JSVM 11 software," *24th Meeting: Geneva, Doc. JVT-X203*, July 2007.

[8] G. Bjøntegaard, "Calculation of average PSNR differences between RD curves," *VCEG contribution, VCEG-M33, Austin*, April 2001.

[9] M. Rezaei, M.M. Hannuksela, and M. Gabbouj, "Semi-fuzzy rate controller for variable bit rate video," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 5, pp. 633–645, May 2008.

[10] [Online], "http://www.tsc.uc3m.es/∼sescalona/graphs.rar," .