RATE CONTROL INITIALIZATION ALGORITHM FOR SCALABLE VIDEO CODING

Sergio Sanz-Rodríguez, Fernando Díaz-de-María

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

ABSTRACT

In this paper we propose a novel rate control initialization algorithm for real-time H.264/scalable video coding. In particular, a two-step approach is proposed. First, the initial quantization parameter (QP) for each layer is determined by means of a parametric rate-quantization (R-Q) modeling that depends on the layer identifier (base or enhancement) and on the type of scalability (spatial or quality). Second, an intra-frame QP refinement method that allows for adapting the initial QP value when needed is carried out over the three first coded frames in order to take into consideration both the buffer control and the spatio-temporal complexity of the scene.

The experimental results show that the proposed R-Q modeling for initial QP estimation, in combination with the intra-frame QP refinement method, provide a good performance in terms of visual quality and buffer control, achieving remarkably similar results to those achieved by using ideal initial QP values.

Index Terms— Initial QP, rate control, scalable video coding (SVC), H.264/SVC, H.264/advanced video coding.

1. INTRODUCTION

Many video transmission services have benefited from the H.264/ scalable video coding (SVC) standard, which is able to provide bit rate adaptation for varying channel conditions as well as for heterogeneous devices with different display resolutions and computational capabilities. H.264/SVC allows for the extraction of a sub-stream from the high-quality bit stream so that this sub-stream can be decoded by a given target receiver. Furthermore, it provides spatial, quality, and temporal scalability. For spatial scalability, a layered coding approach is used for encoding different picture sizes of an input video sequence. Quality scalability can be either coarse grain scalability (CGS), which is a special case of spatial scalability with identical picture sizes, or medium grain scalability (MGS). Each spatio-quality layer is able to support temporal scalability by using hierarchical group of picture (GOP) structures [1].

In real-time applications, the rate control algorithm (RCA) plays a paramount role. The RCA operates at each spatial or quality layer by selecting, for each coding unit, a proper quantization parameter (QP) value so that the buffer fullness is maintained at secure levels and the distortion is minimized. The former condition is ensured by the hypothetical reference decoder (HRD) [2], which acts on the bit budget of each coding unit by setting two (upper and lower) bounds for the buffer fullness. One of the most difficult problems for the RCA is how to determine an appropriate QP value for the first picture. Although some initial QP estimation methods have been proposed for the base spatial or quality layer [3, 4] and also for enhancement layers [5] in H.264/SVC, none of them actually takes into account the HRD at the beginning of the encoding process.

In this paper we propose a novel rate control initialization scheme for real-time H.264/SVC. In particular, a two-step method is proposed. First, the set of initial QPs (one per layer) is determined by means of a first-frame specific rate-quantization (R-Q) model, which takes three different forms, one for the base layer, and other two for the enhancement layers according to the type of scalability (spatial or CGS/MGS). Second, the estimated QP value is modified intra-frame in order to maintain the buffer level under the proper limits by means of a robust QP refinement algorithm that takes into account both the spatial and temporal complexity of the scene.

The paper is organized as follows. In Section 2, an overview of the RC scheme in H.264/SVC encoder is given. Sections 3 and 4 describe, respectively, the first and second step of the proposed method for initial QP estimation. Section 5 presents and discusses the experimental results. Finally, conclusions are drawn and further work outlined in Section 6.

2. RATE CONTROL SCHEME IN H.264/SVC

Let us assume that the H.264/SVC encoder is composed of D spatial/CGS layers, denoted as $d = \{0, 1, \ldots, D-1\}$, and let us assume that each layer can work with up to $Q^{(d)}$ MGS refinements, denoted as $q = \left\{1, \ldots, q, \ldots, Q^{(d)}\right\}$ (it should be noticed that $q = 0$ denotes the base quality layer for a given spatial/CGS layer). In order to make the notation more compact, hereafter we denote each spatio-quality layer as k instead of referring to it as (d, q), where k can obtained as follows: $k = d \times \left(Q^{(d)} + 1\right) + q$.

As previously stated in our work [6], each layer k involves a RC module and a virtual buffer. The virtual buffer at layer k receives the contributions of layers 0 to k and simulates the encoder buffering process of the corresponding sub-steam. The generation of each HRD-compliant sub-stream depends on two parameters: target bit rate $R_T^{(k)}$ and output frame rate $f_{out}^{(k)}$. It should be noticed that $R_T^{(k)}$ must be higher than those associated with lower layers, i.e., $R_T^{(k-i)} \leq R_T^{(k)}$, with $i = 0, 1, \ldots, k$, since those lower layers form part of the k^{th} sub-stream.

The RCA at each layer must be able to assign a proper QP value to every basic unit (BU), so that the virtual buffer is maintained at secure levels, avoiding both overflow and underflow, without noticeable visual quality degradation. A BU is a group of macroblocks (MBs) in raster scan order that share the same QP value, and whose size can range from a single MB to a whole picture [3].

3. R-Q MODELING FOR INITIAL QP SELECTION

Most RCAs operating in real-time use R-Q models for QP selection. These models are appropriate when the video characteristics are nearly constant in time. Nevertheless, this stationarity condition

is not fulfilled at the beginning of the encoding process nor at a scene change. In order to solve this problem, a specific R-Q model for initial QP estimation is proposed. In particular, three parametric models are proposed, one for the base layer and other two for the enhancement layers (depending on the scalability type), whose parameters are learned from examples as described in the next subsections.

Given a layer k, the proposed models aim to capture the relation among the QP value, denoted as $QP^{(k)}$, the output bit rate $R_{out}^{(k)}$ produced by $QP^{(k)}$, and a spatial complexity measurement. The mean pixel-wise gradient over the sequence is used as complexity measurement:

$$\mu_G^{(k)} = \frac{1}{T}\sum_t G_t^{(k)},$$

where L is the number of pictures in the sequence and $G_t^{(k)}$ is the pixel-wise gradient over the t^{th} luminance picture $I_t^{(k)}$, i.e.:

$$G_t^{(k)} = \frac{1}{M\times N}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(\left|I_{t,ij}^{(k)} - I_{t,(i+1)j}^{(k)}\right| + \left|I_{t,ij}^{(k)} - I_{t,i(j+1)}^{(k)}\right|\right),$$

where M and N are the height and width of the picture, respectively.

3.1. Generation of the Training Data Set

A training data set consisting of pairs $\left\{\mathbf{X}^{(k)}, QP^{(k)}\right\}$, where $\mathbf{X}^{(k)}$ is a two-component vector: $\mathbf{X}^{(k)} = \left(R_{out}^{(k)}, \mu_G^{(k)}\right)$, was obtained from a set of well-known sequences used in the field; specifically: "Akiyo", "Bridge-Far", "Carphone", "Container", "Flower", "Hall", "Highway", "Ice", "News", "Paris", "Silent", "Tempete", "Blue_Sky_720p25", "Pedestrian_720p25", "Riverbed_720p25", "Rush_Hour_720p25", "Parkrun_720p50", "Shields_720p50". Some of these sequences were upsampled and/or downsampled in order to get common intermediate format (CIF), quarter CIF (QCIF), $4\times$CIF (4CIF) and high definition (HD) resolutions. Furthermore, none of these sequences was used in the performance assessment conducted in Section 5. Each training sequence was encoded with constant QP using different configurations defined by the following SVC encoder parameters: QP value, GOP size, number of layers, type of quality scalability, picture sizes, and output frame rates.

Once the training data set was generated, it was divided into two sub-sets, one for the base layer model ($k=0$) and the other for the enhancement layers ($k>0$).

3.2. R-Q Model for the Base Layer

According to the data distribution of the training set associated with the base layer, the QP value can be linearly related to $\ln R_{out}^{(0)}$ and $\ln\mu_G^{(0)}$. As a result, the following R-Q function for initial QP selection is proposed:

$$QP_0^{(0)} = \text{round}\left[a_1\times\ln R_T^{(0)} + a_2\times\ln G_0^{(0)} + a_3\right], \quad (1)$$

where $G_0^{(0)}$ is the mean gradient for the first picture of the base layer and $\{a_1, a_2, a_3\}$ are the model coefficients, whose values were determined by means a robust regression method using the corresponding training data subset. Tables 1 and 2 show the coefficient values for QCIF, CIF, 4CIF and HD formats and three ratios of the input frame rate, $f_{in}^{(0)}$, to the output frame rate, that we denote $\Omega^{(0)}$:

$$\Omega^{(0)} = \frac{f_{in}^{(0)}}{f_{out}^{(0)}}.$$

It is also worth noting that the QP is converted into an integer given its discrete nature in H.264/SVC.

3.3. R-Q Model for the Enhancement Layers

The R-Q function for initial QP estimation in Eq. (1) is no longer appropriate for enhancement layers since neither the bit budget for lower spatio-quality layers nor the inter-layer redundancy are considered in the model. Consequently, a new model is proposed that models the relation between the QP increment with respect to the immediately lower layer, the output bit rate increment and the same gradient-based complexity measurement $\mu_G^{(k)}$. Specifically, the QP and output bit increments are defined, respectively, as follows:

$$\Delta QP^{(k)} = QP^{(k)} - QP^{(k-1)},$$

$$\Delta R_{out}^{(k)} = \frac{R_{out}^{(k)} - R_{out}^{(k-1)}}{R_{out}^{(k-1)}}.$$

For this purpose, the training data set for the enhancement layers was split in two sub-sets, one for spatial scalability and the other for quality scalability. This distinction is motivated by the fact that the output bit rate generated by the SVC encoder using a given $\Delta QP^{(k)}$ is higher for spatial scalability than for quality scalability, since the picture size is increased from one layer to the next and, therefore, the initial QP prediction models should be different from each other.

Similarly to the R-Q model for the base layer, the training data distribution corresponding to spatial scalability shows that $\Delta QP^{(k)}$ can be linearly related to both $\ln\Delta R_{out}^{(k)}$ and $\ln\mu_G^{(k)}$. Thus, the following model is proposed:

$$\Delta QP_0^{(k)} = \text{round}\left[b_1\times\ln\Delta R_T^{(k)} + b_2\times\ln G_0^{(k)} + b_3\right], \quad (2)$$

where $\{b_1, b_2, b_3\}$ are the model coefficients. With respect to quality scalability, the following R-Q function was inferred according to its training data distribution:

$$\Delta QP_0^{(k)} = \text{round}\left[c_1\times\Delta R_T^{(k)} + c_2\right], \quad (3)$$

where $\{c_1, c_2\}$ are the model coefficients. In this case, no positive QP increments are allowed; consequently, since the QP range to be modeled is quite lower than that of the spatial scalability case, neither logarithm nor complexity measurement are required. Table 3 illustrates the coefficient values for both spatial and quality scalability modes and two ratios $\Omega^{(k)}$ of the k^{th} layer output frame rate to that of the immediately lower layer, i.e.,

$$\Omega^{(k)} = \frac{f_{out}^{(k)}}{f_{out}^{(k-1)}}.$$

4. INTRA-FRAME QP REFINEMENT

Although the proposed R-Q modeling is able to provide a good performance for many video sequences, the QP value assigned to the first picture occasionally results in overflow or underflow because neither the HRD constraint nor a motion activity measurement have been taken into account.

To solve this problem, we propose a robust method that allows for refining the estimated initial QP when required by considering these two factors. Furthermore, although this approach is viewed as an HRD-based extension of the proposed R-Q modeling, it can be also employed no matter what algorithm is used for calculating the initial QP and can be implemented for any the video codec.

Regarding HRD, we propose setting both an upper and lower bounds for the buffer occupancy during the encoding of the first picture of each layer. The upper bound acts as a security margin that, if

Table 1. Parameter values for the R-Q model in Eq. (1) developed for the base layer. QCIF and CIF formats.

	QCIF			CIF		
$\Omega^{(0)}$	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$
1	-6.09	5.28	83.97	-5.28	4.84	83.23
2	-6.58	6.17	85.32	-5.81	5.48	86.57
4	-7.26	7.16	88.22	-6.50	6.28	91.01

Table 2. Parameter values for the R-Q model in Eq. (1) developed for the base layer. 4CIF and HD formats.

	4CIF			HD		
$\Omega^{(0)}$	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$	$\mathbf{a_1}$	$\mathbf{a_2}$	$\mathbf{a_3}$
1	-5.65	3.94	98.09	-6.13	5.28	112.64
2	-6.23	4.62	102.52	-6.53	6.28	113.43
4	-6.89	5.42	107.31	-6.93	7.36	113.75

Table 3. Parameter values for the R-Q models in Eqs. (2) and (3) developed for the enhancement layers.

	Spatial Scalability			CGS/MGS	
$\Omega^{(k)}$	$\mathbf{b_1}$	$\mathbf{b_2}$	$\mathbf{b_3}$	$\mathbf{c_1}$	$\mathbf{c_2}$
1	-4.15	1.66	-1.07	-3.91	0.26
2	-4.23	1.63	0.33	-3.02	0.59

carefully chosen, drastically reduces the overflow probability without degrading the video quality. The lower bound helps to reduce the underflow risk while preventing abrupt quality falls in those cases where the QP chosen for the picture is too high for the scene.

In order not to exceed these bounds while encoding the I picture, we propose a small BU size to further refine the QP value within the frame. The idea consists of predicting the total number of bits consumed by the I picture after each BU is encoded and modifying the QP if needed. For this purpose, once a BU is encoded the average amount of BU bits is computed and then multiplied by the total number of BUs in the picture. If the estimated number of picture bits is larger than the upper bound, the QP for the next BU is increased one unit; if it is smaller than the lower bound, the QP is decreased one unit; otherwise, the QP is not modified. Finally, the QP is clipped ± 6 with respect to that of the first BU, which was computed using either Eq. (1) or Eqs. (2) and (3), where the required $QP_0^{(k-1)}$ is computed as the average QP of the I picture of the $(k-1)^{th}$ layer.

Regarding motion activity, let us start by noticing that the average QP value obtained for the first picture may not be suitable for the subsequent frames of the GOP, since it was selected just according to the buffer occupancy, without considering the motion complexity of the scene. For instance, an I picture with simple spatial content could be encoded with a too low average QP for the following high motion P or B pictures, resulting in increased buffer occupancy and few bits remaining for the rest of the GOP. Thus, it would be desirable to further refine this initial average QP selected for the first picture according to both the spatial and temporal complexity of the scene. To this end, we propose the use of the small BU size also for the second and third pictures, so that at least one inter coding is ensured (since the second picture can be either I or P in hierarchical GOP structures). In the case of P or B pictures, the upper bound is computed as $R_T^{(k)}/f_{out}^{(k)}$ in order that the buffer fullness is headed towards secure levels. Furthermore, for these pictures the QP value for the first BU is set to the average QP of the last encoded picture belonging to the same spatial or quality layer.

5. EXPERIMENTS AND RESULTS

Our proposals were used for initializing the RCA for H.264/SVC described in [6]. The algorithms were implemented on the Joint Scalable Video Model (JSVM) H.264/SVC reference software version JSVM 9.16 [7]. Starting out with the design suggested in [8], the following video sequences and H.264/SVC encoder configuration for spatial/CGS testing scenario were used:

- Sequences: "Bus", "Football", "Foreman", "Mobile"
- Number of pictures: 300
- GOP size/Intra period: 8/32 pictures
- Number of spatial/CGS layers: $D=5$, $Q^{(d)}=0$
  - $k=0$: QCIF, $f_{out}^{(0)}=6.25$ Hz
  - $k=1$: QCIF, $f_{out}^{(1)}=12.5$ Hz
  - $k=2$: CIF, $f_{out}^{(2)}=12.5$ Hz
  - $k=3$: CIF, $f_{out}^{(3)}=12.5$ Hz
  - $k=4$: CIF, $f_{out}^{(4)}=25$ Hz
- Symbol mode: CAVLC

With respect to the RC parameters of each layer, buffer size and target buffer fullness were set to 1.5 s and 40% of buffer size, respectively, and the set of target bit rates for each test sequence were those proposed in [8]. The upper and lower bounds of the intra-frame QP refinement method described in Section 4 were set to the amount of bits required to reach 80% and 20%, respectively, of buffer size, and the BU size was set to a row of MBs.

In order to find a proper reference algorithm for comparison purposes, all the sequences were previously encoded with constant QP using the set of QP values that best approached the target bit rates, so that those QPs could be considered as ideal values of initial QP. Thus, comparisons were made between the following algorithms:

1. RCA in [6] + ideal values of initial QP
2. RCA in [6] + R-Q modeling (Section 3)
3. RCA in [6] + R-Q modeling (Section 3) + intra-frame QP refinement (Section 4)

Tables 4 and 5 show a performance analysis of the proposed algorithms for "Mobile", as an example of sequence with high spatial detail, and for "Football", as an example of sequence high motion activity, respectively. The fourth column of both tables shows $QP_{0,ave}^{(k)}$ and $\Delta QP_{0,ave}^{(k)}$, which denote, respectively, the average QP obtained for the first picture of the k^{th} layer and the QP increment between the average QPs of the current and lower layers. The performance of the proposed algorithms for any layer k can be inferred by comparing $QP_{0,ave}^{(k)}$ and $\Delta QP_{0,ave}^{(k)}$ to those of the ideal initial QP method.

As can be seen, for "Mobile" the initial QPs assigned by Algorithms 2 and 3 were very close to those of the ideal case (Algorithm 1). However, for "Football" the initial QP assigned for the base layer was low since the motion complexity is not considered in the model. Although the $\Delta QP_{0,ave}^{(k)}$ values obtained by Algorithms 2 and 3 were also very close to those of Algorithm 1 for most layers, this QP estimation error at the base layer was spread to the remaining layers, thus requiring the proposed refinement for the first inter pictures.

In order to assess the performance of Algorithms 2 and 3 from a quality point of view, the average luminance peak signal-to-noise ratio (PSNR), μ_{PSNR}, and its standard deviation, σ_{PSNR}, were used. The Bjøntegaard recommendation [9] was followed to properly compare the obtained μ_{PSNR} values. As can be observed in the tables,

Table 4. Performance analysis of the proposed RCAs for "Mobile".

Layer (k)	$R_T^{(k)}$ (kbps)	Alg.	$QP_{0,ave}^{(k)}/\Delta QP_{0,ave}^{(k)}$	μ_{PSNR} (dB)	σ_{PSNR} (dB)	Bit Rate Error (%)	#O
0	**48**	**1**	38/-	26.15	1.09	3.18	0
		2	36/-	26.17	1.11	2.93	0
		3	38/-	26.24	1.17	3.11	0
1	**64**	**1**	38/0	26.37	0.87	3.72	0
		2	36/0	26.38	0.82	3.74	0
		3	38/0	26.48	0.88	3.23	0
2	**128**	**1**	44/6	22.09	0.54	3.94	0
		2	41/5	22.10	0.80	4.18	0
		3	43/5	22.06	0.56	3.91	0
3	**256**	**1**	39/-5	25.44	0.76	3.50	0
		2	37/-4	25.47	0.84	3.19	0
		3	39/-4	25.43	0.75	3.75	0
4	**384**	**1**	38/-1	26.32	0.59	2.42	0
		2	36/-1	26.32	0.66	2.61	0
		3	38/-1	26.37	0.60	2.02	0

Table 5. Performance analysis of the proposed RCAs for "Football".

Layer (k)	$R_T^{(k)}$ (kbps)	Alg.	$QP_{0,ave}^{(k)}/\Delta QP_{0,ave}^{(k)}$	μ_{PSNR} (dB)	σ_{PSNR} (dB)	Bit Rate Error (%)	#O
0	**128**	**1**	31/-	33.16	0.78	-0.43	0
		2	22/-	33.08	2.10	-0.34	0
		3	22/-	33.14	1.79	-0.35	0
1	**192**	**1**	32/1	32.53	0.88	1.36	0
		2	21/-1	32.41	1.99	1.36	0
		3	21/-1	32.49	1.57	1.23	0
2	**384**	**1**	38/6	27.60	0.66	-1.00	0
		2	25/4	27.18	2.36	-0.89	15
		3	25/4	27.46	1.59	-0.99	0
3	**512**	**1**	37/-1	28.41	0.56	-0.67	0
		2	24/-1	28.10	2.21	-0.87	9
		3	24/-1	28.32	1.40	-1.02	0
4	**1024**	**1**	34/-3	29.99	0.45	-1.32	0
		2	22/-2	29.80	1.86	-0.82	2
		3	22/-2	30.02	0.96	-0.67	0

Algorithm 3 achieved an average PSNR performance similar to that of Algorithm 1 with just a slight increase of σ_{PSNR}. On the other hand, Algorithm 2 was not able to provide the same quality levels when the initial QP value for the base layer was not properly selected, as occurred in "Football".

The behavior of the buffer fullness was also evaluated. For the particular case of "Football" (see Fig. 1), it was observed that the number of overflows (denoted as #O) for Algorithm 2 is notable at some spatial/CGS layers (see Table 5 for more details). Nevertheless, in Algorithm 3 the buffer occupancy was maintained at secure levels during the encoding process since the proposed intra-frame QP refinement method was also applied to the first P and B pictures.

Finally, the results in terms of target bit rate adjustment showed that, in general, the proposed RCAs achieved bit rate errors similar to those of the reference algorithm (see Tables 4 and 5).

6. CONCLUSIONS AND FURTHER WORK

In this paper a novel rate control initialization algorithm for real-time H.264/SVC has been proposed. The initial QP for every layer is determined by means of a parametric R-Q model that adopts three different forms, one for the base layer, and other two for the enhancement layers according to the type of scalability. Furthermore, a QP refinement method for the first intra and inter pictures of the scene has been proposed to guarantee HRD compliance at the beginning of the encoding process or scene change. This last algorithm, which has proved to work efficiently, can be combined with any initial QP estimation method and can be implemented in any video codec.

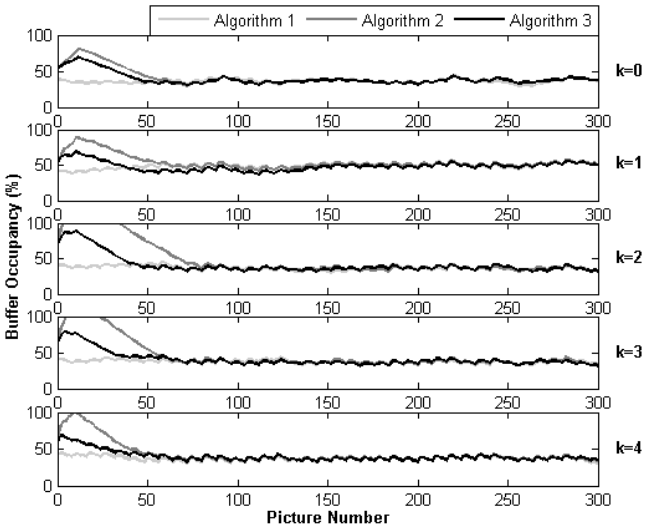

Fig. 1. Buffer occupancy time evolutions corresponding to every layer for "Football".

The experimental results showed that the combination of both approaches achieved good results in terms of visual quality and buffer control. As future work, we plan to include a motion activity measurement in the R-Q models.

7. 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] J. Ribas-Corbera, P.A. Chou, and S.L. Regunathan, "A generalized hypothetical reference decoder for H.264/AVC," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 674–687, 2003.

[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] A. Leontaris and A.M. Tourapis, "Rate control for video coding with slice type dependencies," in *Image Processing, 2008. ICIP 2008. IEEE International Conference on*, 2008, pp. 2792–2795.

[5] J. Yang, Y. Sun, C.S. Kline, and S. Sun, "Adaptive initial quantization parameter selection for H.264/SVC rate control," in *Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on*, 2010, vol. 2, pp. 723 –726.

[6] S. Sanz-Rodríguez and F. Díaz-de María, "RBF-based QP estimation model for VBR control in H.264/SVC," *Circuits and Systems for Video Technology, IEEE Transactions on*, 2011.

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

[8] M. Wien and H. Schwarz, "Testing conditions for SVC coding efficiency and JSVM performance evaluation," *JVT-Q205, 16th JVT Meeting*, Poznan, Poland, July 2005.

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