OPTIMAL RATE ADAPTATION IN THE SCALABLE EXTENSION OF H.264/AVC WITH COMBINED SCALABILITY

Livio Lima, Massimo Mauro, Riccardo Leonardi

Department of Information Engineering
University of Brescia
email: {firstname.lastname}@ing.unibs.it
web: www.ing.unibs.it/tlc/

ABSTRACT

Adaptation for scalable video coding is one of the recent challenges in video distribution over modern networks, which are heterogeneous both in terms of available bandwidth and user end terminal capability. In case of SNR/spatial combined scalability an interesting issue concerns how to remove data from different spatial resolutions for the adaptation to different rates. Scalable Video Coding offers the possibility to adapt the content following the "quality layer" abstraction. In this work we present a new method to optimally define quality layers for a scalable bitstream in a combined scalability scenario using Integer Linear Programming and distortion models. The performances of the proposed approach are comparable with the state-of-the-art methods, but they are obtained with a significant complexity reduction and augmented flexibility.

1. INTRODUCTION

In the last years, scalable video coding emerged as a promising technology for efficient distribution of videos through heterogeneous networks, and it has been recently standardized as scalable extension of the H.264/AVC standard [1], hereafter indicated as SVC. An useful overview of the SVC extension can be found in [2]. The main advantage of SVC is that it offers the flexibility to decode video at different "working points" in terms of spatial, temporal and quality resolution from a unique coded representation, simply decoding only a subset of the original bitstream.

In a heterogeneous environment the scalable video content needs to be *adapted* to meet different end terminal capability requirements in terms of display area or fluctuations of the available bandwidth (*network* or *rate adaptation*). In particular, rate adaptation for scalable video is one of the recent challenges in video distribution over the network. Roughly speaking, the adaptation problem concerns the extraction of the "best" subset of the scalable coded representation in order to minimize a distortion measure of the decoded video sequence. The rate adaptation problem can thus be considered as a more general *rate-constrained optimization problem*, with the distortion as *target function*. Additional constraints could be potentially introduced.

SVC offers the possibility to optimally adapt the content following the "quality layer" abstraction. A quality layer can be arbitrarily defined via the *priority_id* field of each Network Abstraction Layer Unit (NALU), the elementary unit of a SVC bitstream. NALUs with the same value of *priority_id* belong to the same quality layer and have the same importance in the adaptation process. In a combined scenario, where different spatial and SNR layers are defined, an interesting issue concerns how to assign quality layers between different scalability layers in order to optimize the decoding process.

The first method (currently adopted in SVC reference software) proposed for SVC adaptation using quality layers is described in [3]. This approach does not consider the possibility of a multi-layer quality layer generation, since it always assigns higher priority to the NALUs of lower spatial resolution with respect to the NALUs of higher spatial resolutions. An extension to this method has been proposed in [4] to enable a "multi-layer mode" generation. The main drawbacks of these approaches are that both require partial decodings of the bitstream for distortion measurement, and they do not enable the addition of further constraints to the adaptation problem. Similar problems can also be found in other works, e.g. [5], [6] and [7].

Multi-layer adaptation is also proposed in [8] for Fine Grain Scalability (not used anymore) and quality scalability. This approach enables a minimum flexibility to weight different spatial resolutions, but has the same problem of a time-consuming distortion evaluation step. Recently, in [9] the authors proposed an adaptation method based on distortion models for Medium Grain Scalability (MGS) introducing the possibility to control the quality between different groups of pictures. However, this method still requires minimal bitstream decoding, does not support spatial scalability and does not provide any method to control the quality inside a single group of picture.

The objective of this work is to extend the method proposed in [10], which addressed only SNR scalability, in order to support a multi-resolution combined scenario. The adopted approach estimates the distortion through a suitable model and uses Integer Linear Programming (ILP) models to solve the problem of multi-layer *priority_id* assignment. We show that the method has RD performance comparable to the state-of-the-art. At the same time, the proposed method does not require bitstream decoding and enables the flexibility to include additional constraints.

The paper is structured as follows. In Section 2 we overview the fundamental concepts of the SVC standard and the high-level description of coded data. In Section 3 we present the method proposed to solve the adaptation problem with combined scalability. Experimental results are provided in Section 4, while in Section 5 conclusions are drawn.

2. SVC ESSENTIALS

Essentially, in SVC a video sequence is processed by layers. The lower layer is called base layer (BL), and it is independently coded using an H.264/AVC coding scheme, generating a part of the SVC bitstream that can be decoded by H.264/AVC compatible decoders. All the other layers are called enhancement layers (EL). In SVC there are three types of enhancement layers: spatial EL provides spatial scalability, Coarse Grain Scalability (CGS) EL and Medium Grain Scalability (MGS) EL provide quality scalability. Temporal scalability is achieved by hierarchical B-frames decomposition within each layer [2], processing the layer in Group of Pictures (GOP) where the GOPs are separated by the so-called key-pictures.

To increase coding efficiency, enhancement layers are predicted from the base layer or from other enhancement layers using inter-layer prediction tools. These tools introduce additional coding modes to the classical inter- and intra- modes of H.264/AVC. Three inter-layer prediction tools have been introduced in SVC: inter-layer motion prediction, inter-layer residual prediction and inter-layer intra prediction. Briefly, with inter-layer motion prediction motion information (motion vectors and MB partition information) from other layers can be reused, after the suitable scaling. With inter-layer residual prediction, the residual signal of other layers is used (with appropriate scaling) to predict the residual signal of the current layer. With inter-layer intra prediction it is possible to predict the intra-signal from the intra-MBs in the reference layer.

CGS and MGS use the inter-layer prediction tools in a similar way, without any scaling of reference layer information, but with some differences in the prediction of the key-pictures. Moreover, they use different signaling. In fact, CGS is conceptually similar to spatial scalability with each layer having the same spatial resolution. CGS does not provide flexible SNR extraction, since the number of available rates is equal to the number of layers and, as with spatial scalability, it is possible to switch between layers only at Instantaneous Decoder Refresh (IDR) pictures. MGS has been introduced to increase flexibility, with the possibility to discard quality levels at a picture level and to distribute enhancement layer transform coefficients among different NALUs (called MGS vectors) in order to enable finer extraction. Since MGS enables higher flexibility, we consider in our work only the combined scenario with spatial scalability and MGS.

With MGS scalable coding, the process of motion-compensated prediction could introduce drift. Drift describes the effect of unsynchronized motion-compensated prediction loops between encoder and decoder, e.g., because quality refinement packets (used for the prediction at the encoder) have been discarded from the bitstream. With MGS drift is controlled by means of the key-picture concept. For each picture a flag is transmitted, which signals whether the base quality reconstruction or the enhancement layer reconstruction of the reference pictures is employed for motion-compensated prediction. All pictures of the coarsest temporal level are transmitted as key pictures, thus, no drift is introduced in these pictures. In contrast to that, all temporal refinement pictures typically use the reference with the highest available quality for motion-compensated prediction, enabling high coding efficiency but introducing drift.

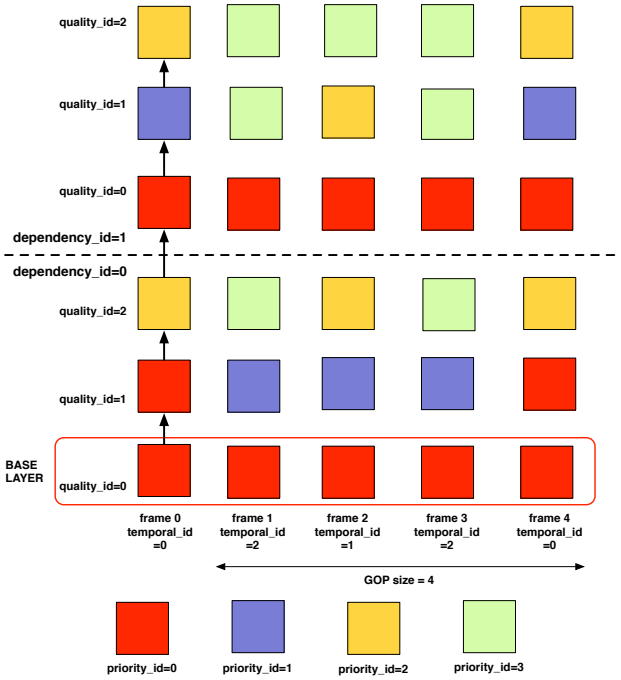

Figure 1: **Example of *priority_id* assignment for two spatial resolutions.**

Coded video data are organized into Access Unit (AU), where each AU contains the data for a single picture. Within an AU, data are distributed into NALUs, each one identified by the following fields of the NALU header: *dependency_id* for the spatial resolution, *temporal_id* for the temporal level and *quality_id* for the quality level. Additionally, the *priority_id* field can be used to define the "level of importance", i.e., the quality layer. If the *priority_id*s are assigned, the adaptation can be performed by discarding NALUs in decreasing order of *priority_id*. An example of the relation between *dependency_id*, *temporal_id*, *quality_id* and *priority_id* is shown in Fig. 1.

Within an AU, SVC enables each layer to be inter-predicted from any layer with lower *dependency_id* and/or *quality_id*, generating several prediction structure possibilities. The choice of the prediction path affects the coding performance and the robustness to drift. Nevertheless, most of the application using SVC adopt the most efficient solution (in terms of coding efficiency), predicting each layer from the next lower layer in terms of *dependency_id* or *quality_id*. This is also the only configuration considered in this work, and it is represented in Fig. 1 by the arrows in frame 0.

Fig. 1 is also useful to clarify two aspects. First, it can be noted that the value 0 of *priority_id* (higher importance) includes all the NALUs of resolution 1 with *quality_id*=0. This is due to a SVC decoder constraint, which requires all these NALUs to decode the resolution 1. The second aspect is that the *priority_id* assignment has been performed in multi-layer mode, i.e., some NALUs with *dependency_id* = 1 have higher importance than some other NALUs with *dependency_id* = 0. This approach can improve the rate-distortion performance of the higher resolution with respect to the non multi-layer approach proposed in [3].

3. PROPOSED METHOD

As introduced in Section 1, we formulate the *priority_id* assignment problem for combined scalability scenario as an optimization problem subject to some constraints. The proposed approach is based on an Integer Linear Programming (ILP) model. ILP is a common approach used to solve optimization problems since it can offer high flexibility and computational advantages. In Section 3.2 is shown as the ILP approach can be used to determine which NALUs have to be discarded given a fixed value of available rate, while in Section 3.3 is presented how the resolution of this subproblem is used to generate quality layers.

3.1 Distortion model

The main goal of the adaptation is the minimization of the decoded sequence distortion. As previously introduced, in order to reduce the complexity, our approach estimates the distortion through a model, conversely from the approaches described in [3], [4], [5], [6], [7]. Furthermore, the use of a distortion model enables the regeneration of quality layers at each point of the distribution network, where the original sequence, required to measure the distortion, is not available.

In literature several models useful to estimate the distortion on a single frame are available. However, our objective is to minimize the overall sequence (or GOP) distortion. We need therefore to take into account the drift effect (see Sec. 2). Assuming combined MGS/spatial scalability, for each frame i, the lower spatial resolution (*dependency_id*=0) distortion contribution on the entire low resolution sequence (or GOP) is given by:

$$D_i^0 = D_{i,MAX}^0 - \sum_{q=1}^{Q} D_{i,q}^0 \tag{1}$$

where $D_{i,MAX}^0$ is the maximum distortion when only the NALU with *quality_id*=0 (which is required and not discardable) is available, while $D_{i,q}^0$ is the distortion reduction obtained decoding also the NALU with *quality_id*=q. Each contribution $D_{i,q}^0$ is estimated using the following model:

$$D_{i,q}^0 = \mathbf{D}_F \mathbf{W}_D \tag{2}$$

where D_F is the *Distortion on Frame* and represents the distortion contribution within the frame, and W_D is the *Drift Weight* that models the drift effect of the NALU. Basically, in our model, D_F depends on the difference between the Quantization Parameter (QP) of the considered NALU and the QP of the NALU of the lower quality level. W_D depends on the number of prediction paths between the current frame and other frames and on their relative *depth*, i.e., the number of intermediate levels in the hierarchical B-frame decomposition in which the prediction is propagating. According to this model, NALUs of higher temporal levels will have a higher W_D.

The distortion contributions of the higher spatial resolutions (*dependency_id*$= s$, $s > 0$) are given by:

$$D_i^s = (D_{i,MAX}^s - \sum_{q=1}^{Q} D_{i,q}^s) - (\alpha D_i^{s-1} + \beta) \tag{3}$$

where the term $\alpha D_i^{s-1} + \beta$ models the distortion contribution of the NALUs of lower spatial resolution (*dependency_id*$= s-1$) on the resolution s. This contribution depends on the inter-layer prediction, which through the upsampling filters uses the intra and inter signals of lower spatial resolution as predictor for higher spatial resolution. It has to be noted that is always true using the inter-layer prediction in "forced mode", i.e., each MB of the higher spatial resolution is forced to reuse the information of lower spatial resolution. This mode could reduce the coding performance for some sequence but drastically reduces the coding complexity.

The model for the inter-layer prediction effect has been experimentally derived through an extensive set of experiments on different video sequence. The results show that a good approximation of the distortion contribution can be obtained with $\alpha = 1$ and $\beta = 0$.

3.2 ILP model

In this section we show how to solve, using the ILP approach, the subproblem $SP(R)$ of identifying which NALUs have to be discarded given a maximum available rate R. The general ILP model is given by:

$$\text{(ILP):} \qquad Z = \max \textstyle\sum_{s=0}^{S} \mathbf{c_s} \mathbf{x_s}$$

$$\text{subject to} \begin{cases} A_s \mathbf{x_s} \leq b_s & s = 0, ..., S \\ B\mathbf{x} \leq d & \\ \mathbf{x} \geq 0 & \text{integer} \end{cases} \tag{4}$$

The unknown $\mathbf{x} = \{\mathbf{x_s}\}$ with $\mathbf{x_s} = \{x_{i,q}^s\}^T$ is a vector of binary variables, one for each NALU, that indicates if the NALU has to be maintained ($x_{i,q}^s = 1$) or discarded ($x_{i,q}^s = 0$). The vector $\mathbf{c} = \{\mathbf{c_s}\}$ with $\mathbf{c_s} = \{c_{i,q}^s\}$ represents the distortion contributions. Each contribution $c_{i,q}^s$ represents the distortion for the NALU estimated by the model (2). Consequently, the term $\mathbf{c_s}\mathbf{x_s}$ represents the sum of all the distortion contributions of the quality enhancement NALUs of spatial resolution s:

$$\mathbf{c_s}\mathbf{x_s} = \sum_{i=0}^{N-1} \sum_{q=1}^{Q} x_{i,q} c_{i,q} \tag{5}$$

where N is the number of frames, while Q is the maximum value of *quality_id*. Two considerations have to be done at this point. First, the minimization of the distortion (1) is equal to the maximization of the sum of contributions $D_{i,q}^0$. Second, the objective function Z addresses only the distortion at the higher spatial resolution. Nevertheless, the model can be easily extended in order to consider any objective function.

Within each picture i, the SVC standard defines that a NALU with *quality_id*=x can be decoded only if all the NALU with *quality_id*$< x$ are available. This set of constraints ($A_s \mathbf{x_s} \leq b_s$ in model (4)) can be represented as:

$$-x_{i,q}^s + x_{i,q+1}^s \leq 0 \qquad \forall i = 0, ..., N-1, \forall s = 0, ..., S \tag{6}$$

In rate adaptation problems, the most critical constraint is represented by the budget constraint ($B\mathbf{x} \leq d$ in model (4)), i.e., the maximum number of bits R available to represent the encoded video. This constraint can be represented as:

$$\sum_{i=0}^{N-1} \sum_{q=0}^{Q} x_{i,q}^s r_{i,q}^s \leq R \tag{7}$$

where $r^s_{i,q}$ is the rate (in bits) of each NALU and it is straightforward to obtain.

In addition to the main budget constraint, further constraints can be considered as, for example, the distortion control over the sequence. If D^s_i is the distortion for frame i at spatial resolution s, these additional constraints can be expressed as:

$$\alpha \frac{\sum_{i=0}^{i=N-1} D^s_i}{N} \leq D^s_i \leq \beta \frac{\sum_{i=0}^{i=N-1} D^s_i}{N} \qquad \forall i = 0,...,N-1 \quad (8)$$

where D^s_i is given by equation (3) and $\alpha < 1, \beta > 1$.

Usually, the complexity to solve the problem (4) is high since the problem is in general NP-hard. Discussions on resolution strategies for ILP problems is out of the scope of this paper. It can be shown that, due to particular properties of matrix constraints A and B, our problem can be solved using a Linear Programming (LP) approach, i.e., considering the unknown $\mathbf{x}$ as continuous. LP problems are much less complex and they can be efficiently solved thanks to the *simplex method*.

Furthermore, it has to be noted that the solution of the subproblem $SP(R)$ is "optimal", i.e. it gives the best Rate-Distortion tradeoff, for the particular distortion model considered. Consequently, a more accurate model could improve the overall system performance.

3.3 Algorithm

Let us assume to have found the optimal solution of the subproblem $SP(R)$, described in Section 3.2. The proposed algorithm easily generates quality layers through multiple resolutions of the subproblem $SP(R)$ at different rate points. The steps of the algorithm are the followings:

1. estimation of distortion vector $\mathbf{c}$ using the model (2)
2. choice of a set of K rates $d = [d_0,...,d_K]$, starting from the rate required to include all the NALUs with *quality_id*=0 (d_0) to the rate of the full SVC stream (d_K)
3. resolution of K subproblems $SP(d_k)$, obtaining K solution vectors $\mathbf{x}_k$
4. let $\mathcal{N}_k$ be the set of NALUs with related binary variable equal to 1 in $\mathbf{x}_k$. The *priority_id* value equal to k is assigned to the NALUs that belong to the set $\{\mathcal{N}_k - \mathcal{N}_{k-1}\}$, with $\mathcal{N}_{-1} = \emptyset$

The quality layer k is represented by the NALUs with the value of *priority_id* equal to k.

4. EXPERIMENTAL RESULTS

The proposed method has been compared with [4], included in the SVC reference software (JSVM). We call the two methods "New_MLQL" and "JSVM_MLQL" respectively. We have considered also the method proposed in our previous work [10], which has been modified to support the presence of more than one spatial layer. This method works in an "independent mode", namely considering different spatial layers as independent and assigning always higher priority to lower spatial layers. We call it "New_QL". The three methods have been compared both in terms of RD perfomance and from a complexity point of view, i.e., the time required for the *priority_id* assignement. The number of *priority_id* is always set to 64, the maximum allowed.

Sequence	SET1		SET2		
	$\Delta\mu_{PSNR}$ [dB]	T gain	$\Delta\mu_{PSNR}$	$\Delta\sigma_{PSNR}$	T gain
City (4CIF)	-0.14	432	-0.75	-36%	93
Crew (4CIF)	-0.03	447	-0.29	-32%	83
Harbour (4CIF)	-0.03	478	-0.73	-47%	86
Ice (4CIF)	-0.09	418	-0.75	-49%	86
Soccer (4CIF)	-0.07	532	-0.64	-39%	89

Table 1: **Performance comparison between JSVM_MLQL and New_QL methods.**

Sequence	SET1		SET2		
	$\Delta\mu_{PSNR}$ [dB]	T gain	$\Delta\mu_{PSNR}$	$\Delta\sigma_{PSNR}$	T gain
City (4CIF)	-0.13	305	-0.32	-24%	55
Crew (4CIF)	0.00	318	-0.19	-24%	57
Harbour (4CIF)	0.04	333	-0.24	-34%	55
Ice (4CIF)	-0.05	332	-0.33	-35%	57
Soccer (4CIF)	-0.06	325	-0.30	-21%	60

Table 2: **Performance comparison between JSVM_MLQL and New_MLQL methods.**

Two set of experiments have been carried out, without (SET1) and with (SET2) the additional set of constraints for the control on PSNR fluctuations. In each set the methods have been tested on different test sequences, considering MGS scalability in different configurations in terms of GOP size. In each experiment the full SVC stream has been generated with the values of *priority_id* obtained by the three algorithms and successively the optimality of *priority_id* generation has been evaluated adapting and decoding the stream at different rates following the quality layer description.

Tables 1 and 2 show a summary of the experimental results for New_QL and New_MLQL methods respectively. A negative $\Delta\mu_{PSNR}$ indicates a PSNR performance loss of the proposed method, *T gain* indicates the gain in terms of execution time for the *priority_id* assignment and $\Delta\sigma_{PSNR}$ in SET2 is the reduction of the PSNR fluctuation measured as σ_{PSNR}, i.e., the standard deviation of the PSNR values in the reconstructed video sequence. It can be noted that for SET1 experiments the performance of the proposed methods are comparable to [4], with a slightly better performance of New_MLQL as also shown in Fig. 2. The complexity reduction is considerable in all cases, of a factor of about 400 on average.

In SET2 experiments, the proposed method shows a performance loss and a decreasing of the computational gain. This last effect is due to the greater complexity of the ILP problem with the additional contraints. The effect on RD performances is explained in Fig. 3, where the mean σ_{PSNR} and the PSNR frame-by-frame (at a particular rate point) are shown. PSNR fluctuations are well controlled and peaks are eliminated, at a cost of a reduced average PSNR. Tests show that all the methods can work in real-time in all cases on a modern laptop (Intel Core 2 Duo 2.2 Ghz, 2GB RAM).

5. CONCLUSIONS

In this paper we have presented a new method to optimally define quality layers for an SVC stream in a combined MGS/spatial scalabilty scenario. Our approach makes use

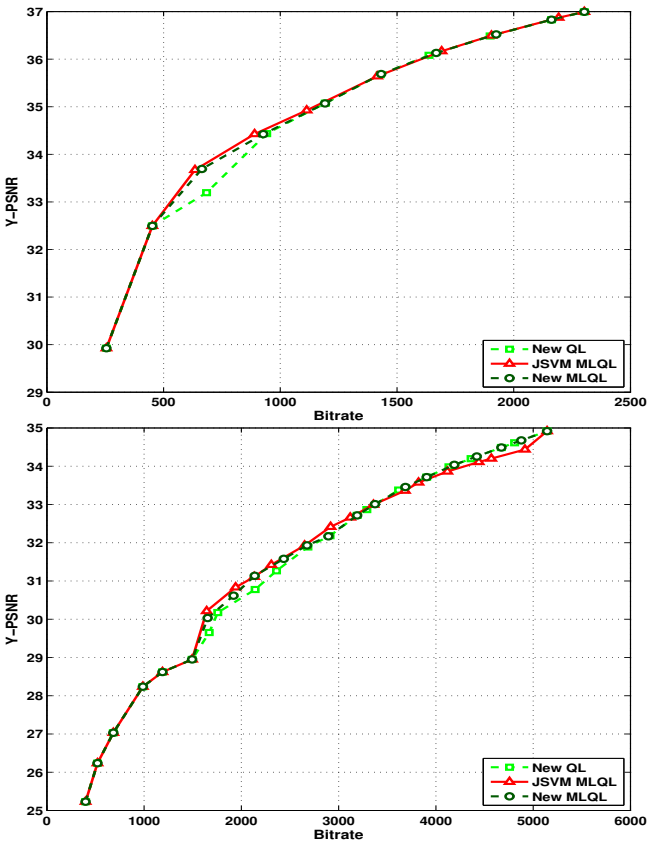

Figure 2: **Experiments without distortion control (SET1).**

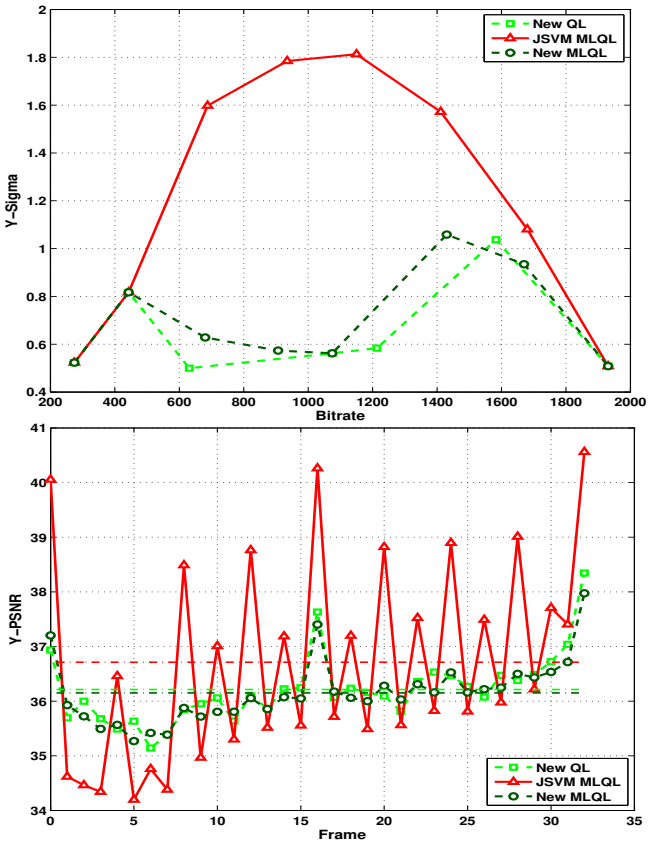

Figure 3: **Experiments with distortion control (SET2).**

of Integer Linear Programming and distortion models. Two different methods have been evaluated: a multi-layer mode and an independent mode assignement, with the first giving slightly better results. The two methods have been compared with the multi-layer assignement method of the JSVM software. We have shown that our approach provides RD extraction performances which are comparable with the JSVM method, while being 300-400 times faster. This strong complexity reduction is given from the use of the distortion model which does not require bitstream decoding to compute distortion contributions of every NALU. A further advantage of the proposed approach is the possibility to insert any additional constraints to the original formulation, exploiting the great flexibility of the ILP modelization. This flexibility allows to adapt the problem to many different situations and application contexts.

Current research is focusing on the improvement of the distortion model, and on the evaluation of the proposed approach in High Definition and videoconferencing scenarios.

REFERENCES

[1] Advanced Video Coding for Generic Audiovisual Services, ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG-4 AVC), ITU-T and ISO/IEC JTC 1, Version 8 : Consented in July 2007.

[2] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the Scalable Video Coding extension of the H.264/AVC standard", *IEEE Transactions on CSVT*, vol. 17, pp. 1103–1120, Sept. 2007.

[3] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux, "Optimized rate-distortion extraction with quality layers in the Scalable extension of H.264/AVC", *IEEE Transactions on CSVT*, vol. 17, pp. 1186–1193, Sept. 2007.

[4] M. Mathew, K. Lee, and W.-J. Han, "Multi layer quality layers", ITU-T VCEQ JVT-S043, Geneva, April 2006.

[5] E. Maani, and K. Katsaggelos, "Optimized bit extraction using distortion modeling in the Scalable extension of H.264/AVC", *IEEE Transactions on Image Processing*, vol. 18, pp. 2022-2029, Sept. 2009.

[6] C. Gu, D. Zhao, and X. Ji, "Fast Rate Allocation Based on Distortion Estimation Modeling in Scalable Video Coding", in *Proc. SPIE*, 2008.

[7] T. Rusert, and J.-R. Ohm, "Backward Drift Estimation with Application to Quality Layer Assignment in H.264/AVC Based Scalable Video Coding", in *Proc. ICASSP 2007*, Honolulu, Hawaii, USA, 2007.

[8] T. Cong Thang, J.W. Kang, J.-J. Yoo, and Y.M. Ro, "Optimal multi-layer Adaptation of SVC Video over Heterogeneous Environments", *Advances in Multimedia*, vol. 2008, Article ID 739192, 2008.

[9] R. Li, J. Sun, and W. Gao, "Fast weighted algorithms for bitstream extraction of SVC medium grain scalable video coding", in *Proc. ICME 2010*, Singapore, 2010.

[10] L. Lima, M. Mauro, T. Anselmo, D. Alfonso, and R. Leonardi, "Optimal rate adaptation with integer linear programming in the scalable extension of H.264/AVC", submitted to *Proc. ICIP 2011*.