OPTIMAL RATE ADAPTATION WITH INTEGER LINEAR PROGRAMMING IN THE SCALABLE EXTENSION OF H.264/AVC

Livio Lima [#], *Massimo Mauro* [#], *Tea Anselmo* [*], *Daniele Alfonso* [*], *Riccardo Leonardi* [#]

[#] University of Brescia, Department of Information Engineering, Brescia, Italy
[*] STMicroelectronics, Advanced System Technology, Agrate Brianza, Italy

ABSTRACT

Adaptation for scalable video 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. 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 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 strong complexity reduction and augmented flexibility.

Index Terms— Video coding, Rate-distortion, Integer Linear Programming.

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 (*user adaptation*) 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.

In SVC coded data are organized into Network Abstraction Layer Units (NALUs), whose "importance" can be defined with the *priority_id* field (non-normative) of the NALU header. A group of NALUs with the same value of *priority_id* is called *quality layer*. Efficient adaptation can be performed by discarding NALUs according to the value of *priority_id*, in such a way that less important NALUs are discarded before. Therefore, adaptation performance depends on the optimality in which quality layers are generated.

The first method (currently adopted in SVC reference software) proposed for SVC adaptation using quality layers is [3], that first orders NALUs by their Rate-Distortion "slope" and then group them in order to obtain the desired number of quality layers. The same approach is used in similar works: [4] uses improved measurement of the distortion contributions, while [5] and [6] decrease the number of decoding processes to calculate distortion contributions. The main drawbacks of these approaches are that all of them require partial decoding of the bitstream for distortion measurement, and they do not enable the addition of further constraints to the adaptation problem.

The objective of this work is to propose a new approach to generate quality layers with performance comparable to the state-of-the-art. At the same time, the proposed method does not require bitstream decodings and enables the flexibility to include additional constraints. The adopted approach estimates the distortion through a model and uses Integer Linear Programming (ILP) models to solve the problem of *priority_id* assignment. At the moment, the proposed method supports only SNR scalability.

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

2. SVC ESSENTIALS

Essentially, in SVC video sequence is processed at different spatial resolutions (spatial levels), each one encoded using a coding scheme based on H.264/AVC. To increase coding efficiency, higher spatial resolutions are predicted from lower resolutions using inter-layer prediction tools. Tempo-

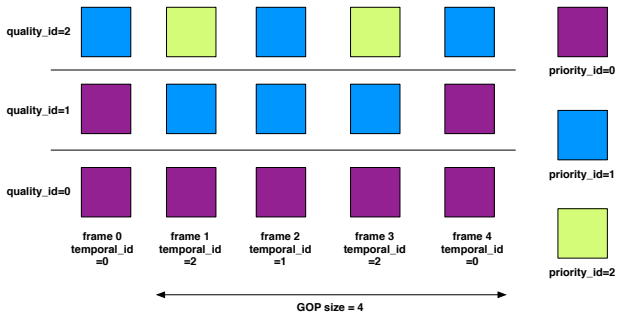

Fig. 1. **Example of *priority_id* assignment for a single spatial resolution (*dependency_id=0*).**

ral scaling is achieved by hierarchical B-frames decomposition. For quality scaling, two modes are provided by SVC: Coarse Grain Scalability (CGS) and Medium Grain Scalability (MGS). CGS is conceptually similar to spatial scalability where each level has 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 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 picture level and to distribute enhancement layer transform coefficients among different NALUs (called *MGS vectors*) in order to enable finer extraction.

With MGS scalable coding, the motion-compensated prediction possibly introduces *drift*. Drift describes the effect of not synchronized 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 the use 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. The importance of drift on the sequence distortion will be described in the following section.

As previously introduced, coded video data are organized 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. With an optimal assignment of the *priority_id* the rate 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 Figure 1.

3. PROPOSED METHOD

As described in Section 1, the adaptation problem can be considered as a more general optimization problem. The objective is to extract a subset of NALUs minimizing the overall distortion of the decoded video and satisfying a rate constraint. Furthermore, additional constraints could be included (e.g., a limitation on the fluctuations of the distortion among different frames).

Optimization theory includes several techinques to solve such problems, depending on particular properties of the problem itself. In our method the problem is modeled and solved with an ILP approach. ILP is a common approach in other disciplines as the Operation Research 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

ILP model requires the knowledge of rate and distortion contributions for each NALU. Rate contributions are straightforward to obtain, while the exact knowledge of the distortion contributions requires multiple bitstream decoding, as done in works [4], [5] and [6]. In order to limit the complexity of quality layers generation, in our approach we adopt a model to estimate the distortion contribution of each NALU. 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 (and quite accurate) models to estimate the distortion on a single frame are available. However, our objective is to minimize the overall sequence (or GOP) distortion; thus, we need to take the drift effect into account. To do this, the distortion contribution of each NALU is estimated using the following model:

$$D_N = D_f W_D \tag{1}$$

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, D_f depends on the difference between the Quantization Parameter (QP) of the NALU and the QP of the NALU of 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.

Differently from other approaches, the motion features of each sequence are not considered in this model, because this

requires partial bitstream decodings and increases the model complexity. Our objective is to maintain the model as simple as possible, ensuring a high computational efficiency.

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:

$$\text{(ILP):} \qquad Z = \min \mathbf{cx}$$

$$\text{subject to} \begin{cases} A\mathbf{x} \geq b \\ B\mathbf{x} \geq d \\ \mathbf{x} \geq 0 \qquad \text{integer} \end{cases} \tag{2}$$

The unknown $\mathbf{x} = \{x_{t,q}\}^T$ is a vector of binary variables, one for each NALU, that indicates if the NALU has to be maintained ($x_{t,q} = 1$) or discarded ($x_{t,q} = 0$). The vector $\mathbf{c} = \{c_{t,q}\}$ represents the distortion contributions. Each contribution $c_{t,q}$ represents the distortion for the NALU estimated by the model (1). Consequently, the objective function $\mathbf{cx}$ represents the overall distortion on the decoded sequence:

$$\mathbf{cx} = \sum_{i=0}^{N-1} \sum_{q=0}^{Q} x_{i,q} c_{i,q} \tag{3}$$

where N is the number of frames, while Q is the maximum value of *quality_id*.

Within each picture i, 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 ($A\mathbf{x} \geq b$ in model (2)) can be represented as:

$$x_{i,q} - x_{i,q+1} \geq 0 \qquad \forall i = 0, ..., N-1 \tag{4}$$

In rate adaptation problems, the most critical constraint is represented by the budget constraint ($B\mathbf{x} \geq d$ in model (2)), i.e. the maximum number of bits R available to represent the encoded video. This constraint can be represented as:

$$-\sum_{s=0}^{S} \sum_{i=0}^{N-1} \sum_{q=0}^{Q} x_{i,q} r_{i,q} \geq -R \tag{5}$$

where $r_{i,q}$ is the rate (in bits) of each NALU.

In addition to the main budget constraint, further constraints can be considered as, for example, the distortion control over the sequence. If D_i is the distortion for frame i, these additional constraints can be expressed as:

$$\alpha \frac{\mathbf{cx}}{N} \leq D_i \leq \beta \frac{\mathbf{cx}}{N} \qquad \forall i = 0, ..., N-1 \tag{6}$$

where $D_i = \sum_{q=0}^{Q} x_{i,q} c_{i,q}$ and $\alpha < 1, \beta > 1$.

In general, the complexity to solve the problem 2 is high since the problem is in general NP-hard. Discussion on resolution strategies for ILP problems is out of the scope of this paper. It can be shown as, 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*.

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 (1)
2. choice of a set of K rates $d = [d_0, ..., d_K]$, starting from the rate of SVC base layer (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 $\mathcal{N}_k$ 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 to [3], included in the SVC reference software. Two set of experiments have been performed, without (SET1) and with (SET2) the additional set of constraints (6). In each set the methods have been tested on different test sequences, considering MGS scalability in different configurations in terms of GOP size, intra period, number of MGS vectors. The two methods have been compared from two points of view: the rate distortion performance and the complexity in terms of computational time required for *priority_id* generation. The number of *priority_id* is always set to 64, the maximum allowed.

Table 1 shows a summary of the performed experiments. A negative Δ_{PSNR} indicates a performance loss of the proposed method, *T gain* indicates the gain in terms of execution time 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 method (JSVM_ILP) are comparable to

Sequence	SET1		SET2		
	Δ_{PSNR} [dB]	T gain	Δ_{PSNR}	$\Delta_{\sigma PSNR}$ [%]	T gain
Bus (CIF)	-0.06	30	-0.74	-60	8
Football (CIF)	-0.05	25	-0.7	-37	8
Foreman (CIF)	-0.21	28	-0.74	-45	8
Mobile (CIF)	-0.14	27	-0.7	-47	8
City (4CIF)	-0.21	106	-0.76	-40	30
Crew (4CIF)	-0.09	120	-0.51	-28	33
Harbour (4CIF)	-0.03	156	-0.86	-45	64
Soccer (4CIF)	-0.08	135	-0.47	-26	56
Ice (4CIF)	-0.1	100	-0.63	-15	43

Table 1. **Summary of performed experiments.**

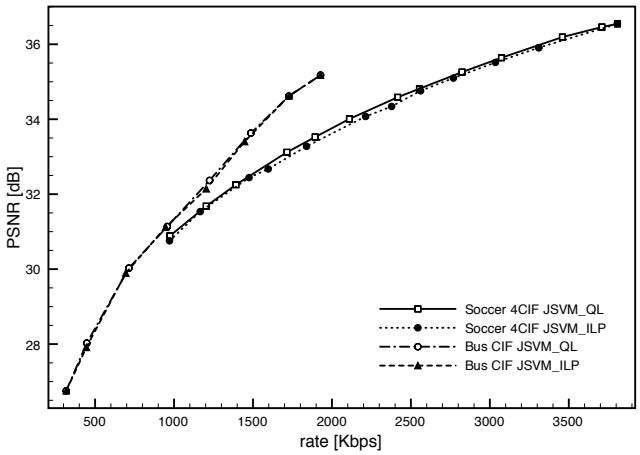

Fig. 2. **Examples without distortion control (SET1).**

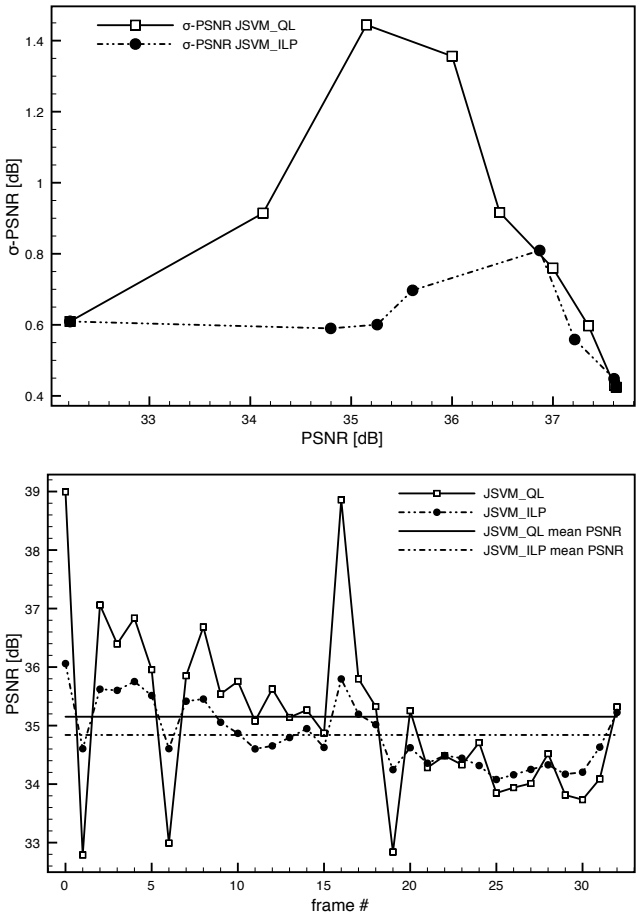

Fig. 3. **Experiment with distortion control (SET2), Crew sequence.**

[3](JSVM_QL), as also shown in Figure 2, with considerable complexity reduction. 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 Figure 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. Finally, tests show that with a CPU Intel Core 2 Duo 2.2 Ghz quality layers are generated in real time for a 4CIF 60Hz sequence.

5. CONCLUSIONS

The proposed work presents a new method to optimally define quality layers for a SVC stream using Integer Linear Programming and distortion models. We have shown as the use of our approach leads to a great reduction of computational complexity while maintaining good RD performances. Moreover, the adoption of an ILP model enables high flexibility in order to adapt the problem to different requirements and to various application contexts.

6. 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 Transaction on CSVT*, vol. 17, no. 9, pp. 1103–1120, 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 Transaction on CSVT*, vol. 17, no. 9, pp. 1186–1193, 2007.

[4] E. Maani and K. Katsaggelos, "Optimized bit extraction using distortion modeling in the Scalable extension of H.264/AVC," *IEEE Transaction on Image Processing*, vol. 18, no. 9, pp. 2022–2029, 2009.

[5] C. Gu, D. Zhao, and X. Ji, "Fast rate allocation based on distortion estimation modeling in scalable video coding," in *Proc. of SPIE*, 2008.

[6] 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. of ICASSP*, 2007.