COMPLEXITY MODELING OF SCALABLE VIDEO DECODING

Zhan Ma, Yao Wang

Dept. of ECE, Polytechnic University, Brooklyn, NY 11201
Emails: zma03@vision.poly.edu, yao@poly.edu

ABSTRACT

This paper addresses the computational complexity of scalable video decoding using emerging scalable extension of H.264/AVC (SVC) standard compliant decoder. Scalable functionalities provided by SVC standard encompass temporal, spatial, quality enhancements and their combinations. The complexity model for decoding a bit stream with only temporal, spatial, or quality scalability are developed first. We then extend to a more general model for decoding a bit stream with arbitrarily combined scalability. Comparison with the number of clock cycles used in SVC decoding on a PC shows that the proposed model is very accurate.

Index Terms— Computational complexity, complexity modeling, scalable video decoding

1. INTRODUCTION

Due to the advances of network bandwidth and wireless access techniques, and pervasive multimedia content service over heterogenous network infrastructures to diverse users (terminals), multimedia stream with scalable features is demanded to satisfy the different requirements, by adapting its scalable functionalities [1]. Because of its friendly network interface and high coding efficiency, H.264/AVC [2] promises the dominant status for video service industry in coming decades, thus JVT (Joint Video Team) experts decide to extend the H.264/AVC to provide scalable functionalities to face the diverse requirements via a single bitstream which is generated following the syntax, semantics and operations defined in scalable video coding extension of H.264/AVC [3].

Along with massive researches on mobile computing and wide deployments of wireless video service, computational complexity problem is raised up for video processing on power limited device since large amount of data transformations involve. How to predict or model the computational complexity consumption of video processing attracts more and more attentions from industry and academia. In [4], authors analyze the computational complexity of software based H.264/AVC [2] baseline profile decoder by its decoding subfunctions, and estimate the time complexity on DSP and general purpose computer via the frequency of use of decoding subfunctions. He *et al.* [5] propose a power-rate-distortion (P-R-D) framework for typical video encoder with fully scalable coding scheme.

Because the SVC standard has a coding efficiency comparable to the non-scalable H.264/AVC standard, it is expected that it may be widely adopted by mobile multimedia applications, where battery energy consumption is of paramount concern. In this paper, we model the decoding computational complexity of the SVC decoder. Given a particular implementation platform of the decoder, the complexity derived from the model can be translated into power consumption. One important motivation for developing such model is to enable a receiver of a SVC bit stream to determine which spatial, temporal and quality layers to decode to achieve a desired tradeoff between the decoded video quality and decoding power consumption.

The paper is organized as follows, the complexity model of SVC will be described in Section 2, and then the experimental verifications of analytical model is conducted in Section 3. Section 4 concludes the paper and give the future direction of this work.

2. COMPLEXITY MODEL OF SVC

SVC bitstream can be produced either by the individual scalable tools, or via the combination of supported scalable functionalities. In order to give an insightful understanding of decoding time complexity, we firstly analyze the individual scalable tool, i.e., temporal, spatial, and quality, respectively, and then combine them together to obtain a general decoding complexity function in terms of the numbers of decoded spatial, temporal and quality layers. Please refer to [3] for more information about the details of SVC. The common symbols used in deriving our model are presented in Table 1.

2.1. Temporal Scalability

Temporal scalability could be efficiently provided via hierarchical B pictures. We take the popular dyadic prediction structure with one picture reference (depicted in Fig. 1) to consider time complexity for decoding the temporal scalable

This material is based upon work supported by the National Science Foundation under Grant No. 0430145.

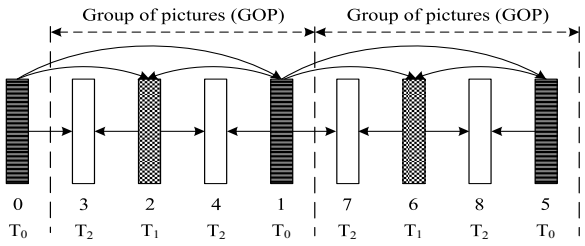


Fig. 1. Dyadic hierarchial B pictures for temporal scalability. The numbers below pictures specify the coding order, and symbol T_k indicates the temporal layer identifier.

Table 1. Symbols for Analytical Complexity Model

Notion	Description
$C_{\rm I}$/$C_{\rm P}$/$C_{\rm B}$	Average macroblock decoding complexity of I-/P-/B-picture
$C_{\rm S}$/$C_{\rm Q}$	Average macroblock decoding complexity at spatial/quality enhancement layers
T/D/Q	Total layer number for temporal-/spatial-/quality-scalability
t/d/q	Layer index for temporal-/spatial-/quality-scalability
M	Number of macroblocks per picture

bitstream. We analyze the bitstream with temporal scalability only, i.e., temporal scalability at a fixed spatial resolution without quality enhancements.

With the dyadic prediction structure of hierarchical B pictures, for each GOP at layer T, there are a total of (2^T-1) B-pictures to be decoded, and there is only one key picture in this GOP to be decoded (depicted in Fig. 1), which can be either an I-picture or P-picture. Assuming that α portion of key pictures are coded as I-pictures and there are M macroblocks per picture, then the complexity for decoding a GOP from base (t=0) to T-th layer is

$$C_{\rm GOP,TS}(T,M) = M(\alpha\, C_{\rm I}+(1-\alpha)C_{\rm P}+(2^T-1)C_{\rm B}). \quad (1)$$

The $C_{\rm I}$, $C_{\rm P}$ and $C_{\rm B}$ in Equation (1) represent the average complexity of decoding one macroblock in I-, P-, and B-pictures, respectively.

2.2. Spatial Scalability

In order to provide resolution diversity, spatial scalability is supported by SVC. In addition to the intra-layer prediction, such as motion-compensated prediction and intra prediction, the inter-layer prediction is employed to improve coding efficiency. In the decoding part, after processing the base layer bitstream to get the lower resolution video, the reconstructed pictures are up-sampled as reference for enhancement layer decoding prediction. The reference picture up-sampling ratio is selected as 2 in this paper, which means that the number of macroblocks at current spatial layer will be 4 times the macroblock number of its preceding layer. Let M_d represent the number of macroblocks in spatial layer d, then M_d = $4M_{d-1}$ = 4^dM_0, where M_0 indicates the number of macroblocks per picture in the base layer.

Generally, the decoding complexity of a spatial enhancement block depends on the underlying picture type. In order not to introduce too many model parameters, we use a single parameter $C_{\rm S}$ to denote the average decoding complexity of each macroblock in spatial enhancement layers. Note that $C_{\rm S}$ includes the complexity for upsampling as well as decoding inter/intra-layer prediction error. Assuming that at the d-th spatial layer, there are 2^t pictures in each GOP, the complexity for decoding a GOP in d-th spatial layer is

$$C_{\rm GOP,SSenh}(d,t,M_0) \;=\; 2^tM_dC_{\rm S} = 2^t4^dM_0C_{\rm S}. \quad (2)$$

Assume that at the base layer (d=0), a GOP is coded using hierarchical B pictures with dyadic prediction structure, with complexity indicated in Eq. (1), then the complexity for decoding each GOP from base to D-th layer is

$$\begin{aligned}
&C_{\rm GOP,SS,TS}(D,t,M_0)\\
&= C_{\rm GOP,TS}(t,M_0) + \sum_{d=1}^{D} C_{\rm GOP,SSenh}(d,t,M_0)\\
&= C_{\rm GOP,TS}(t,M_0) + \sum_{d=1}^{D} 2^t4^dM_0C_{\rm S}\\
&= M_0\left\{\alpha\, C_{\rm I} + (1-\alpha)C_{\rm P} + (2^t-1)C_{\rm B}\right.\\
&\qquad \left. + \frac{4}{3}(4^D-1)2^tC_{\rm S}\right\}. \qquad (3)
\end{aligned}$$

2.3. Quality Scalability

Coarse-grain quality (CGS) and medium-grain quality scalabilities (MGS) are employed in SVC [3] to support quality scalability. CGS can be referred as a special case of spatial scalability with identical picture size for enhancement and base layers. The same switchable inter- and intra- layer prediction mechanisms are provided in CGS without upsampling and inter layer deblocking operations compared to spatial scalable case. For the inter layer prediction of CGS, a refinement of texture information is achieved by requantizing the residual texture signal in the enhancement layer with a small quantization step (QP) related to the preceding CGS layer. The main difference for decoding different CGS layer is using different but related quantization step (QP). Furthermore, CGS provides discrete rate points corresponding to the coded layers. In order to support a flexible bitstream adaption, MGS is included in SVC, with a modified high level signaling compared to the CGS. Based on *key picture* concept, it is allowed to trade off the coding efficiency and drift for hierarchical prediction structures. From the decoder point of view, high

level syntax signaling overhead could be ignored compared to the large amount of data transformations after parsing, i.e., inverse transform, reference reconstruction etc. In order to simplify the addressed problem, the popular CGS scheme is considered in this part and following combined cases.

Assuming for a specified q-th layer of Q-layer quality scalable bitstream at a given spatial, temporal resolution, macroblock decoding complexity at quality enhancement layer could be abstracted as $C_{\rm Q}$. The complexity for decoding any enhancement quality layer in a GOP with 2^t pictures and M macroblocks per picture is

$$C_{\rm GOP,QSenh}(t,q,M) = 2^t M C_{\rm Q}. \tag{4}$$

Assuming the base layer GOP is coded using the hierarchical B picture with dyadic prediction structure, the decoding complexity for each GOP from the base to Q-th layer is

$$\begin{aligned} &C_{\rm GOP,TS,QS}(t,Q,M) \\ &= C_{\rm GOP,TS}(t,M) + \sum_{q=1}^{Q} C_{\rm GOP,QSenh}(t,q,M) \\ &= M(\alpha\, C_{\rm I} + (1-\alpha)C_{\rm P} + (2^t-1)C_{\rm B} + 2^t Q C_{\rm Q}). \end{aligned} \tag{5}$$

2.4. Combined Scalability

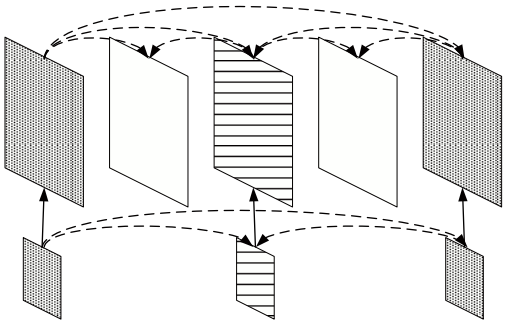


Fig. 2. Multilayer structure with inter-layer prediction for combined spatio-temporal scalability.

In general, SVC bitstream delivered to diverse users will be coded with combination of temporal, spatial and quality scalable functionalities. Assuming for spatial layer d, there are $T(d)$ temporal layers, and upon them there are $Q(d,T(d))$ quality layers. The pictures at the d-th spatial layer are predicted from preceding $(d-1)$-th layer or their intra-layer references adaptively. Note that the number of temporal layer could be different for each spatial layer, however, we assume that the temporal layer number at higher spatial layer is not less than the number of temporal scalability at its preceding or relative lower spatial points, i.e., $T(d) \geq T(d-1)$. Thus, the inter layer operations are processed upon $2^{T(d-1)}$ pictures from $(d-1)$-th to d-th spatial layer, the rest $(2^{T(d)} - 2^{T(d-1)})$ pictures at d-th layer will predicted using hierarchical B scheme with dyadic prediction structure (depicted in Fig. 2). Then, the complexity for decoding d-th spatial layer is

$$\begin{aligned} &C_{\rm GOP,ComS}(d, M_0) \\ &= C_{\rm GOP,SSenh}(d, T(d-1), M_0) \\ &+ \ (2^{T(d)} - 2^{T(d-1)}) 4^d M_0 C_{\rm B} \\ &+ \ 2^{T(d)} 4^d Q(d,T(d)) M_0 C_{\rm Q} \\ &= 2^{T(d-1)} 4^d M_0 C_{\rm S} + (2^{T(d)} - 2^{T(d-1)}) 4^d M_0 C_{\rm B} \\ &+ \ 2^{T(d)} 4^d Q(d,T(d)) M_0 C_{\rm Q} \\ &= 4^d 2^{T(d)} M_0 \left\{ \eta(d)\, C_{\rm S} + (1-\eta(d)) C_{\rm B} \right. \\ &\qquad \left. + Q(d,T(d)) C_{\rm Q} \right\}, \end{aligned} \tag{6}$$

with $\eta(d) = 2^{T(d-1)-T(d)}$. The total complexity for decoding from spatial base layer to spatial layer D, within spatial layer d, decoding up to $T(d)$ temporal layers, and $Q(d,T(d))$ quality layers, is

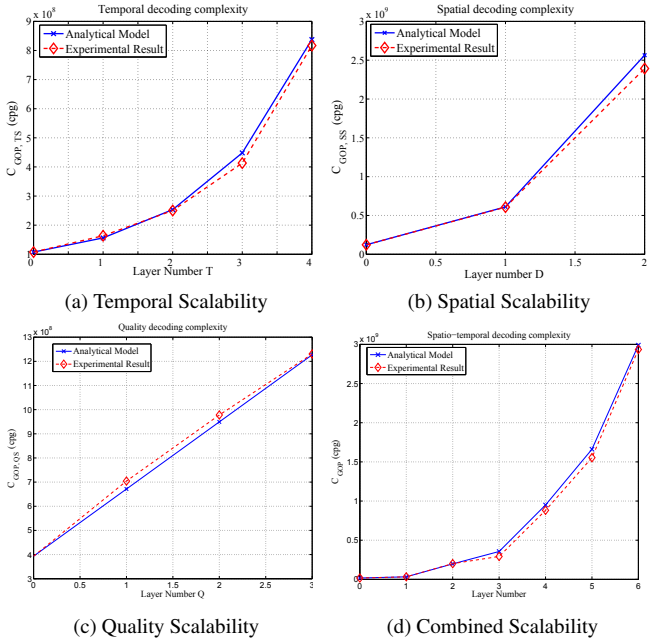
$$\begin{aligned} C_{\rm GOP} &= M_0 \left[\alpha\, C_{\rm I} + (1-\alpha) C_{\rm P} + (2^{T(0)} - 1) C_{\rm B} \right] \\ &+ \ M_0 2^{T(0)} Q(0,T(0)) C_{\rm Q} + \sum_{d=1}^{D} C_{\rm GOP,ComS} \\ &= M_0 \left\{ \left[\alpha\, C_{\rm I} + (1-\alpha) C_{\rm P} + (2^{T(0)} - 1) C_{\rm B} \right] \right. \\ &+ \ \sum_{d=0}^{D} 2^{T(d)} 4^d Q(d,T(d)) C_{\rm Q} \\ &+ \ \left. \sum_{d=1}^{D} 2^{T(d)} 4^d (\eta(d) C_{\rm S} + (1-\eta(d)) C_{\rm B}) \right\}. \end{aligned} \tag{7}$$

In the special case of $T(d)$=$T(d-1)$+1, and $Q(d,T(d))$=Q for all d, the above model is simplified into,

$$\begin{aligned} C_{\rm GOP} &= M_0(\alpha\, C_{\rm I} + (1-\alpha) C_{\rm P} + (2^{T(0)} - 1) C_{\rm B}) \\ &+ \frac{8^{D+1}-1}{7} 2^{T(0)} M_0 Q C_{\rm Q} \\ &+ 4\frac{8^{D}-1}{7} 2^{T(0)} M_0 (C_{\rm S} + C_{\rm B}). \end{aligned} \tag{8}$$

3. MODEL VERIFICATION

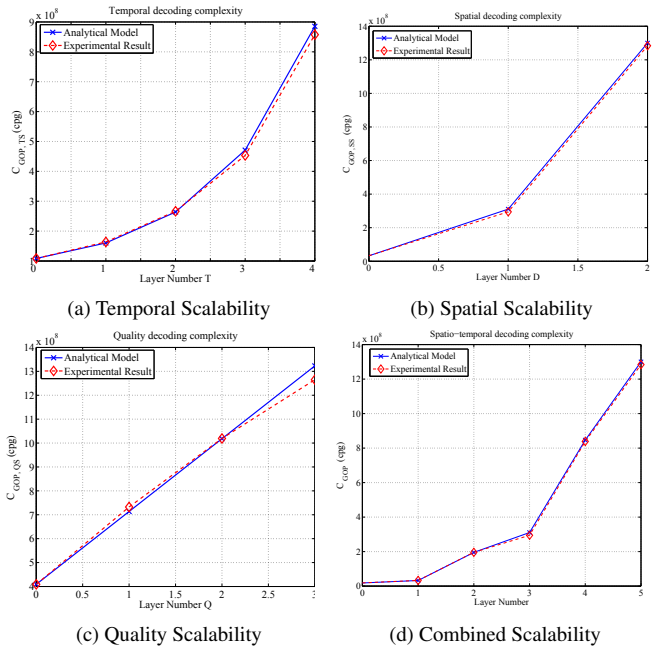
To validate the proposed complexity model, we measure the number of clock cycles used by a PC running the SVC decoder, by using the Intel Vtune Analyzer [6]. Clock cycle per GOP (cpg) is selected to represent the GOP decoding complexity. The SVC reference software JSVM [7] is used to both generate the test bitstreams and to decode the bitstreams with different number of spatial, temporal and quality layers. We also use the VTune to determine the basic model parameters, $C_{\rm I}$, $C_{\rm P}$, $C_{\rm B}$, $C_{\rm S}$, $C_{\rm Q}$. For example, to determine $C_{\rm I}$, we



(a) Temporal Scalability (b) Spatial Scalability

(c) Quality Scalability (d) Combined Scalability

Fig. 3. GOP decoding complexity for sequence Soccer. a) T ranging from 0 to 4 at CIF resolution without quality enhancements; b) D from 0 to 2; c) quality Q from 0 to 3 at CIF resolution; d) spatial layer ranging from QCIF to 4CIF with GOP size from 2 to 8. In this case, C_{I} = 124377 cycles, C_{P} = 173526 cycles, C_{B} = 152188 cycles, C_{S} = 499452 cycles and C_{Q} = 87716 cycles.



(a) Temporal Scalability (b) Spatial Scalability

(c) Quality Scalability (d) Combined Scalability

Fig. 4. GOP decoding complexity for sequence Crew. a) T ranging from 0 to 4 at CIF resolution without quality enhancements; b) D from 0 to 2; c) quality Q from 0 to 3 at CIF resolution; d) spatial layer from QCIF to 4CIF with GOP size from 2 to 8. In this scenario, C_{I} = 120791 cycles, C_{P} = 274102 cycles, C_{B} = 130787 cycles, C_{S} = 205556 cycles and C_{Q} = 102062 cycles.

generate a bitstream with all I pictures, and then C_{I} is overall complexity divided by the number of macroblocks. To determine C_{P}, we generate a bitstream in IBP structure, the IPP bitstream could be easily extracted, then C_{P} is the total P picture complexity divided by the number of macroblocks in P pictures, where the P picture complexity is the overall complexity minus the complexity for decoding I pictures. Fig. 3 and Fig. 4 compare the measurement data with the analytical model for the video sequence "Soccer" and "Crew". As can be seen, the model matches with the measurement data very well.

4. CONCLUSION

This paper models the decoding complexity of received SVC bitstream in terms of the number of decoded spatial, temporal, and quality layers. The model uses the decoding complexity for different types of macroblocks as the basic parameters, and is expected to be applicable for both software and hardware decoders. The complexity model can be translated into actual power consumption or instruction count according to the platform architecture. Comparison with the measured clock cylcles on a PC running the SVC decoder shows that the proposed model is very accurate for modeling the software decoding complexity. For future work, we plan to investigate the relation between the macroblock decoding parameters (C_{I},C_{P}, etc.) and the characteristics of the underlying sequence and possibly codec operating parameters.

5. REFERENCES

[1] J.-R. Ohm, "Advances in Scalable Video Coding," *Proc. of the IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.

[2] H.264/AVC, *Draft ITU-T Rec. and Final Draft Intl. Std. of Joint Video Spec. (H.264/AVC)*, Joint Video Team, Doc. JVT-G050, Mar. 2003.

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

[4] M. Horowitz, A. Joch, F. Kossentini, and A. Hallapuro, "H.264/AVC Baseline Profile Decoder Complexity Analysis," *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 704–716, July 2003.

[5] Z. He, Y. Liang, and *et al.*, "Power-Rate-Distortion Analysis for Wireless Communication under Energy Constraints," *IEEE Trans. on Circuits Syst. Video Technol.*, vol. 15, no. 5, pp. 645–658, May 2004.

[6] Intel VTune Performance Analyzer [online], Available: http://www.intel.com/cd/software/products/asmo-na/eng/vtune/239-144.htm.

[7] Joint Scalable Video Model (JSVM), *JSVM Software*, Joint Video Team, Doc. JVT-X203, Geneva, Switzerland, June 2007.