

# EFFICIENT HIERARCHICAL INTER PICTURE CODING FOR H.264/AVC BASELINE PROFILE

Weixing Wan<sup>1</sup>, Ying Chen<sup>2</sup>, Ye-Kui Wang<sup>3</sup>, Miska M. Hannuksela<sup>3</sup>, Houqiang Li<sup>1</sup>, and Moncef Gabbouj<sup>2</sup>

<sup>1</sup>University of Science and Technology of China

<sup>2</sup>Department of Signal Processing, Tampere University of Technology

<sup>3</sup>Nokia Research Center

## ABSTRACT

Bi-predictive (B) slices are not supported in the Baseline profile of the Advanced Video Coding (H.264/AVC) standard, which results in a decreased coding efficiency compared with other profiles supporting B slices. However, many application standards, such as the mobile multimedia services specified by the Third Generation Partnership Project (3GPP), use only the Baseline profile for H.264/AVC. Therefore, it is worth investigating H.264/AVC coding when only intra (I) and inter (P) slices are supported. In this paper, a content-adaptive Quantization Parameter (QP) cascading scheme for the hierarchical P coding method compatible with Baseline profile of H.264/AVC is proposed. The proposed method is based on a picture-level QP optimization. The proposed method has a significantly better rate-distortion performance than the traditional IPPP coding structure and outperforms hierarchical P coding methods using fixed delta QP settings between temporal levels noticeably with up to 0.53 dB gain in average luminance Peak Signal-to-Noise Ratio (PSNR).

**Index Terms**—H.264/AVC Baseline profile, Hierarchical P, Quantization Parameter

## 1. INTRODUCTION

The Advanced Video Coding (H.264/AVC) standard [1][2] was developed by the Joint Video Team (JVT). Using state-of-the-art coding tools, H.264/AVC achieves a significant improvement in terms of Rate-Distortion (RD) performance compared with earlier standards. It specifies several profiles targeted for different application environments and trade-offs between compression efficiency and computational complexity. The discussions in this paper focus on the compression efficiency of the Baseline profile.

The Baseline profile of H.264/AVC does not support bi-predictive (B) slices, in which a picture may be predicted by two reference picture lists and thus each sub-macroblock partition may be predicted from two reference pictures. In other words, only intra (I) slices and inter (P) slices are supported by the Baseline profile. Using B slices, an

average of 0.5-1 dB Peak Signal-to-Noise Ratio (PSNR) gain can be achieved by adjusting the Quantization Parameter (QP) of the B slices [3]. Compared with many other profiles, the Baseline profile does not support Context-based Adaptive Binary Arithmetic Coding (CABAC), which provides a bit-rate reduction between 5%–15% compared with Context-based Adaptive Variable Length Coding (CAVLC) [1], the only entropy coding tool supported by the Baseline profile.

Regardless of the absence of the support for B slices and CABAC, the Baseline profile has been chosen as the only supported H.264/AVC profile into many application standards, e.g., the Third Generation Partnership Project (3GPP) multimedia services [4][5][6]. Therefore, it is important to investigate how to improve the coding efficiency when only I and P slices are allowed.

The hierarchical B coding structure has been demonstrated as an effective tool for improving coding efficiency, e.g. compared with the traditional IBBP coding structure [7]. In this structure, the importance of pictures at each temporal level differs because of the hierarchically structured temporal prediction chain. Therefore, improved bit-rate saving under the same quality constraint can be achieved by using higher QP values for higher temporal levels. In [7], Schwarz et al. proposed a QP setting method which fixed the difference of QP between each pair of temporal levels without considering the content characteristics change across sequences or pictures. In [8], an exhaustive search method to get the best QP for each temporal level was presented. However, the encoding complexity is much higher than for the method described in [7] because of the exhaustive search. The above methods use picture-level QP optimization, which selects one constant QP value for a whole slice.

In this paper, a hierarchical P coding structure similarly to the hierarchical B coding structure is used for coding H.264/AVC Baseline profile compatible content. To improve the coding efficiency further, we propose a picture-level content-adaptive QP cascading mechanism. In this mechanism, QP delta values in relative to the QP value of the highest temporal level pictures are decided based on the prediction modes of the Macroblocks (MBs).

---

The work of Weixing Wan and Houqiang Li is supported by NSFC General Program under contract No. 60672161, 863 Program under contract No. 2006AA01Z317, and NSFC Key Program under contract No. 60632040. The work of Ying Chen is partly supported by the Nokia Foundation Award granted by Nokia Research Center.

The rest of this paper is organized as follows. In Section 2, an introduction to the hierarchical P coding structure is presented. Section 3 describes the content-adaptive QP cascading mechanism as well as two simple QP cascading mechanisms which are content independent. Experimental results are presented in Section 4, followed by conclusions in Section 5.

## 2. HIERARCHICAL P CODING STRUCTURE

One example of the hierarchical P coding structure (with 4 hierarchical/temporal levels) is shown in Fig. 1. The first picture of a video sequence is an Instantaneous Decoding Refresh (IDR) picture. A picture is called a key picture when all previously coded pictures also precede the picture in display order. As illustrated in Fig. 1, a key picture and all pictures that are temporally located between the current key picture and the previous key picture are considered as a Group Of Pictures (GOP). In the hierarchical P coding structure, multiple references in inter prediction are supported. The prediction relationship is as follows. The key pictures are either intra-coded or inter-coded using previous key pictures as references. The remaining pictures of a GOP are hierarchically predicted and it is possible to use pictures from the past or/and from the future in display order as references. Referring to Fig. 1, picture 1 refers to pictures 0 and 2 which means that the inter prediction reference of each MB or MB partition is either from picture 0 or picture 2, however any MB or MB partition can not simultaneously be predicted from both picture 0 and 2, as only one reference picture list (list 0) is constructed.

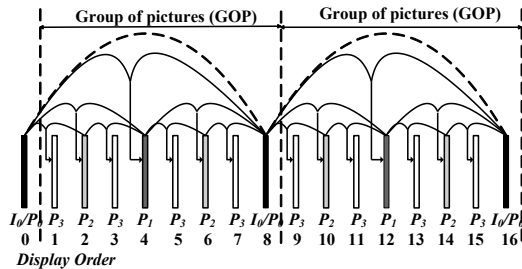


Fig. 1. Dyadic hierarchical P coding with 4 temporal levels

It should be noted that the usage of hierarchical P coding structure can be more flexible than the dyadic structure shown in Fig. 1. Typically to support temporal scalability for each temporal level, a picture is predicted from pictures of a lower or the same temporal level.

## 3. CONTENT-ADAPTIVE QP CASCADING

Let  $T$  be the total number of temporal levels and the QP value for the pictures in the highest temporal level, denoted as the  $QP_{T-1}$ , is set according to the desired target bit-rate, and the QP value for one lower temporal level picture is set

according to a delta QP value and  $QP_{T-1}$ . In this section, we firstly describe the proposed content-adaptive QP cascading mechanism. For comparison, two content independent methods are also introduced.

### 3.1. Content-Adaptive QP Cascading

The difference between the QP value of a picture in a temporal level lower than  $T$  and the input QP value  $QP_{T-1}$  is decided by a scaling-factor, which depends on the prediction modes of the MBs in the picture.

In the hierarchical temporal prediction structures, the motion-compensation prediction can be expressed as the high-pass filtering along the motion trajectory with filter  $\{1, -1\}$  when using inter prediction, or with filter  $\{-1/2, 1, -1/2\}$  when using bi-prediction [9]. The scaling-factor is to balance the residual energies of the whole picture in contrast to the energy of the pictures in a higher temporal level, and thus controls the QP value of the picture. The scaling-factor is derived as the weighted average of the relative energy increase caused by the filtering process which is actually performed during inter predicted motion compensation prediction. After motion estimation, an energy factor of a picture can be calculated as in equation (1):

$$E_{t,m} = \frac{1}{N} \sum_{i=1}^N \alpha_i \quad (1)$$

where  $t$ , ranging from 0 to  $T-1$ , inclusive, denotes the temporal level,  $m$  denotes the picture index within a temporal level  $t$ ,  $N$  is the total number of MBs in a picture, and  $\alpha_i$  represents the weighting factor of the relative energy of the  $i$ -th MB during motion compensation. Note that for simplicity, we assume all MB partitions (if more than one) in an MB are treated with the same intra, inter-P or inter-B modes.

The weighting factor  $\alpha_i$  of an MB depends on the prediction mode of the MB. If an MB is an inter-P MB (with filter  $\{1, -1\}$ ), the factor is  $\sqrt{2}$ . If it is an inter-B MB (with filter  $\{-1/2, 1, -1/2\}$ ), the factor is  $\sqrt{3/2}$ . For an intra MB, the factor is 1. In the Baseline profile, there is no inter-B MB or MB partition. So, for each MB, if it is not intra-coded, all its MB partitions must be inter-P. So, in the H.264/AVC Baseline profile coding, only two factors are in use for all pixels of each MB:  $\sqrt{2}$  and 1, so equation (1) holds for the hierarchical P coding.

After  $E_{t,m}$  is obtained, the corresponding scaling-factor of the  $m$ -th picture with temporal level  $t$  is:

$$SF_{t,m} = \overline{SF_{t+1}} \div E_{t,m} \quad (2)$$

where  $t$  is in the range of 0 to  $T-1$ , inclusive,  $\overline{SF_{t+1}}$  is the average of  $SF_{t+1,j}$  for all values of  $j$  within the current GOP.  $SF_{T-1,j}$  is initialized to 1.0, for any  $j$ .

After the scaling-factor  $SF_{t,m}$  is obtained, the QP value for the corresponding picture, denoted as  $QP_{t,m}$ , can be calculated as in equation (3):

$$QP_{t,m} = QP_{T-1} + 6\log_2 SF_{t,m} \quad (3)$$

where  $QP_{T-1}$  is the input QP value and  $6\log_2 SF_{t,m}$  is the delta QP value. The final value of  $QP_{t,m}$  is rounded and clipped to be an integer value in the range of 0 to 51, inclusive. Note that all the highest temporal level pictures have the same QP value in our method.

The prediction modes of MBs are unknown until the completion of the motion estimation and mode decision processes. Therefore, for each GOP of the target sequence to be encoded, it is first analyzed to find the MB prediction modes by performing the motion estimation and mode decision processes. Then the final QP value for each picture can be calculated as above mentioned.

### 3.2. Content Independent QP Cascading

Two methods that have fixed delta QP values among temporal levels for all sequences are presented as follows. Note that these two methods are not part of the proposed mechanism and are used only for comparison purposes.

- Fixed Scaling-Factor (FSFC): For each P picture, a fixed percentage of the MBs are assumed as P MBs. The final QP of a temporal level is still set as described by the equations above.
- Fixed QP Cascading (FQC): The delta QP between specific two adjacent temporal levels is set to a specific constant value.

## 4. EXPERIMENTAL RESULTS

The presented hierarchical P coding methods were implemented based on the SVC reference software JSVM\_8\_6 [10], which is capable of generating H.264/AVC compatible bitstreams. In our simulation, GOP size was set to 16 and the initial QP values for the highest temporal level pictures were 28, 32, 36 and 40, respectively.

A wide range of sequences were tested. They were “Container”, “Foreman”, “Irene”, “Mobile”, “News”, “Paris”, “Silent”, and “Tempete”. All the sequences were QCIF@30Hz. The following four coding scenarios were compared. Note that in these four scenarios, only the first picture of each sequence was coded using I slice

- Content-Adaptive QP Cascading (CAC): The method mentioned in Section 3.1. In the pre-processing of performing motion estimation and mode decision processes, the QP values for all the pictures were set in a simple way, in which the delta QP between two adjacent levels is set to 2.
- Traditional IPPP coding structure (IPPP): The GOP size was 1 and the number of reference pictures was equal to 2. That is, a sequence was coded as “ $I_0... P_n P_{n+1} P_{n+2} ...$ ”, where  $I_0$  was I picture, any other picture  $P_n$ , was a P picture.  $P_{n+2}$  referred to  $P_{n+1}$  and  $P_n$ , and so on. All pictures in a coded video sequence had the same QP value.

- FSFC: The percentage of P MBs in each picture was considered to be a fixed value. To achieve a good performance for this method, we tested a wide range of percentages from 60% to 100%, and found that when the percentage was set to 100%, the best average performance was achieved. Therefore, FSFC with 100% of the MBs considered as P MBs was chosen for the comparison. In this case,  $E_{tm}$  is  $\sqrt{2}$  and delta QP between two adjacent temporal levels is 3.
- FQC: The delta QP value between adjacent two temporal levels  $t+1$  and  $t$  was fixed, denoted as  $\Delta QP_{t+1}$ . A wide range of  $\Delta QP_{t+1}$  ( $t = 0, 1, \dots, T-2$ ) values were tested to achieve a good performance for this method. Finally, we found that when  $\Delta QP_1$  was set to 4 and  $\Delta QP_{t+1}$  ( $t = 1, 2, \dots, T-2$ ) was set to 1, the best RD performance was achieved. Therefore, FQC with  $\Delta QP_1$  equal to 4 and  $\Delta QP_t$  ( $t > 1$ ) equal to 1 was chosen for the comparison in this paper. Note that the QP setting method for hierarchical B coding in [7] uses the same delta QP values as in this method.

TABLE I  
Performance comparison between CAC and other methods

Sequence	CAC vs IPPP		CAC vs FQC		CAC vs FSFC	
	PSNR (dB)	Bit-rate	PSNR (dB)	Bit-rate	PSNR (dB)	Bit-rate
Container	1.39	-7.2%	0.14	-2.7%	0.28	-5.1%
Foreman	1.31	-19.8%	0.16	-2.9%	0.26	-4.1%
Irene	1.19	-20.0%	0.14	-2.5%	0.21	-3.7%
Mobile	3.16	-43.0%	0.23	-5.2%	0.32	-7.3%
News	1.11	-22.6%	0.38	-5.7%	0.58	-8.6%
Paris	2.20	-28.6%	0.48	-7.5%	0.71	-10.7%
Silent	2.14	-29.7%	0.53	-8.6%	0.73	-11.5%
Tempete	2.13	-34.2%	0.14	-3.2%	0.19	-4.1%
Average	1.84	-25.6%	0.28	-4.8%	0.41	-6.9%

The average RD performances of the four methods are listed in Table I. The results were generated using the Bjontegaard measurement [11], which is based on the bit-rates and average luma PSNR values of the four test points corresponding to four different input QP values.

From Table I, it can be observed that hierarchical P coding outperforms traditional IPPP coding significantly. Compared with traditional IPPP coding, the CAC method brings 25.6% bit-rate saving on average for all tested sequences. Although PSNR fluctuations inside a GOP exist in CAC method, no annoying subjective pumping artifacts occur.

Compared with other QP cascading methods, i.e., FQC and FSFC, the proposed CAC method has noticeable better performance. Up to 8.6% and 11.5% bit-rate savings can be achieved when compared with FQC and FSFC, respectively.

On average, the PSNR gains are about 0.3 dB (compared with FQC) and 0.4 dB (compared with FSFC) respectively.

The tested sequences can be classified according to the motion activity into three types, namely sequences with low motion and almost still background, sequences with low to medium motion, and sequences with high motion.

For the sequences with low motion and almost still background, let us take “*Silent*” as an example. The RD curves are shown in Fig. 2. The average PSNR gain compared with FSFC and FQC is about 0.6 dB.

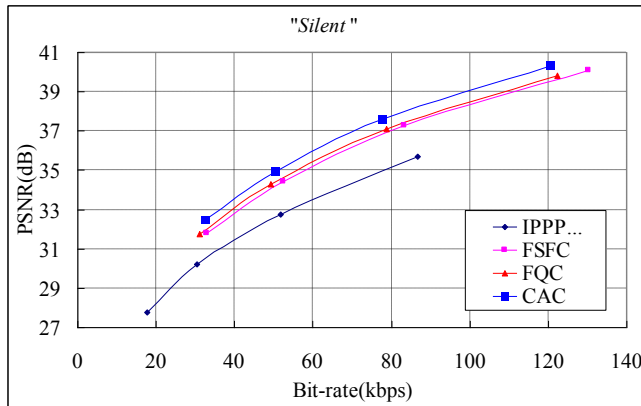


Fig. 2. RD curves for “*Silent*”.

The RD curves for “*News*”, belonging to sequences with low to medium motion, are shown in Fig. 3. Compared with FSFC and FQC, an average of about 0.5 dB PSNR gain was achieved.

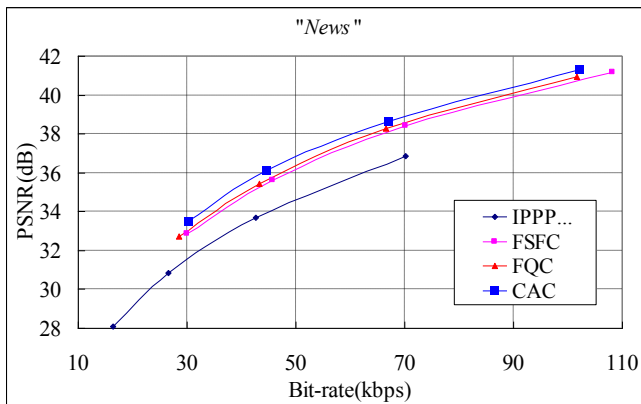


Fig. 3. RD curves for “*News*”.

For the sequences with high motion, the average PSNR gain is a little smaller. However, even in this case, the proposed method CAC outperforms FSFC and FQC by about 0.2dB PSNR gain.

It can be concluded that the proposed CAC method provides greater compression gains for sequences with lower motion activity. The reason for this is that the percentage of I MBs increases for sequences with higher motion activity.

Additionally, it is worth mentioning that constant QP value for all the pictures with the same hierarchical P structure provides even less efficiency than IPPP, which is about 0.3 dB PSNR loss.

## 5. CONCLUSIONS

H.264/AVC Baseline profile lacks the support for bi-predictive (B) slices but supports hierarchical inter prediction structures, which can be used to improve compression efficiency at the cost of increased latency. In this paper, we presented a content-adaptive method for adjusting the Quantization Parameter (QP) for hierarchical inter (P) picture structures. We compared the presented method against traditional non-hierarchical P picture coding (IPPP) as well as two hierarchical P picture coding schemes that selected the QP for a picture based on its temporal level similarly to the methods presented in the literature for hierarchical B picture coding. Simulation results showed that significant gains, more than 25 % bit-rate saving, were achieved over the traditional IPPP coding structure. Compared with the fixed QP setting methods, the proposed method provided a noticeable coding efficiency improvement, about 0.3 dB in luma PSNR on average, which is equivalent to about 5 % bit-rate saving.

## 6. REFERENCES

- [1] T. Wiegand, G. J. Sullivan, G. Bjontegarrd and A. Luthra, “Overview of the H.264/AVC video coding standard,” *IEEE Trans. Circuits Syst. Video Tech.*, Volume 13, pp. 560 - 576, July 2003.
- [2] ITU-T Recommendation H.264, “Advanced video coding for generic audiovisual services,” Nov. 2007.
- [3] M. Flierl and B. Girod, “Generalized B pictures and the draft H.264/AVC video compression standard,” *IEEE Trans. Circuits Syst. Video Tech.*, Volume 13, pp. 587 - 597, July 2003.
- [4] 3GPP TS 26.234, “Transparent end-to-end packet-switched streaming service (PSS) Protocols and codecs”.
- [5] 3GPP TS 26.346, “Multimedia broadcast/multicast service (MBMS); Protocols and codecs”.
- [6] 3GPP TS 26.114, “IP multimedia subsystem (IMS); Multimedia telephony; Media handling and interaction”.
- [7] H. Schwarz, D. Marpe, and T. Wiegand, “Analysis of hierarchical B pictures and MCTF,” *Proc. of IEEE Int. Conference on Multimedia and Expo (ICME)*, pp.1929 - 1932, July 2006.
- [8] D. Prannatha, M. Kim, S. Hahm, B. Kim, K. Lee and K. Park, “Dependent Quantization for Scalable Video Coding,” *The 9th Int. Conference on Advanced Communication Technology*, pp. 222 - 227, Feb. 2007.
- [9] L. Luo, F. Wu, S. Li, and Z. Zhang, “Advanced lifting-based motion threading (MTh) technique for 3D wavelet video coding,” *Proc of SPIE VCIP2003*, Volume 5150, pp. 707 - 718, July 2003.
- [10] J. Vieron, M. Wien, and H. Schwarz, “JSVM 8.6 software,” Joint Video Team Doc. *JVT-U202*, Hangzhou, China, Oct. 2006.
- [11] G. Bjontegaard, “Calculation of average PSNR differences between RD-curves,” *VCEG-M33*, Mar. 2001.