Adaptive Granularity selection in Reference Picture Memory Compression

Yanzhuo Ma, Lijuan Kang
ISN Laboratry
Xidian University
Xi’an, Shaanxi, China
yzma@mail.xidian.edu.cn

Abstract—Memory size occupation and memory access bandwidth are capital issues for high resolution video codec design because of the large scale of data, and the high power consumption especially for wireless terminals. An adaptive and random access-obeyed reference pictures memory compression (RPMC) scheme based on as small granularity as 2x2 blocks for 2-bit truncation is proposed in this paper to solve the problems. The impact of granularity in RPMC to memory access bandwidth during motion estimation and/or motion compensation is analyzed firstly. Then an adaptive RPMC scheme based on 2x2 block size is proposed, based on min-max scalar quantization (MMSQ). Finally, the results based on HM are provided which show that compared with the common used 4x4 block-based methods, little performance loss is introduced. At the same time, based on the adaptively merge scheme, the average memory access bandwidth is saved than the 4x4 block based method by up to 17%.

Keywords- Video processing; reference pictures memory compression (RPMC); memory bandwidth; power consumption; granularity

I. INTRODUCTION

Nowadays, in video coding or processing such as frame insertion, motion estimation (ME) is a popular tool to exploit the inter-frame correlation, which needs one or more encode-decoded frames to be stored in buffer for reference. As the frame size and bit-depth of common video materials are increasing, the codec buffer size is increasing accordingly. Nevertheless, limited by the hardware techniques, the frequently access to the reference frames causes large memory accessing bandwidth and high power consumption. [1]

Vary kinds of works have been carried out to alleviate these memory-power consumptions. A traditional and effective method for memory access bandwidth reduction is by using cache [2], which can avoid some of the repeat access, but is extremely expensive for the whole frame or even lines of uncompressed high definition video data. Thus, Reference pictures memory compression (RPMC) scheme could be considered to decrease both the memory access bandwidth and buffer size. To maintain the random access feature of reference frame buffer, the RPMC scheme needs to be based on certain unit size, and satisfy a fixed compression ratio, e.g. the most widely used one, from 10-bit/pixel to 8-bit/pixel, which is shown in Fig .1.

As simple random access methods are concerned, the simplest RPMC scheme is fixed rounding, which equally introduces distortions to each pixel, which has been introduced in HEVC proposal as a simple solution [3].To exploit the high correlation between adjacent pixels in one unit, the min-max scalar quantization (MMSQ) method [4] [5] based on 4x4 blocks was proposed by M. Budagavi etc., which is simple and efficient. MMSQ is compatible with none of the existed standards, so A. D. Gupte etc. proposed to transport the error of MMSQ from encoder to decoder, to avoid drift effect while using in h.264/AVC.[1] Z. Ma proposed a two-layer method, which applies a check board pattern down sample to get a base layer, and present the other samples as enhanced layer by their residuals with the interpolation values from the base layer.[6] This method is flexible, however, due to the down sampling, it could introduce aliasing artifact. J. Kim etc. proposed the method using Huffman based variable length coding table to achieve lossless compression. [7] This kind of methods has been utilized by some hardware designers. [8],[9] However, this vlc table based method could not obey the principle of random access, which could increase complex computation and delay. Another technique of offset compensation is proposed to reduce the compression loss [10]. A luma-chroma combined block based compression method provides more budget to luma data to decrease compression loss. [11]

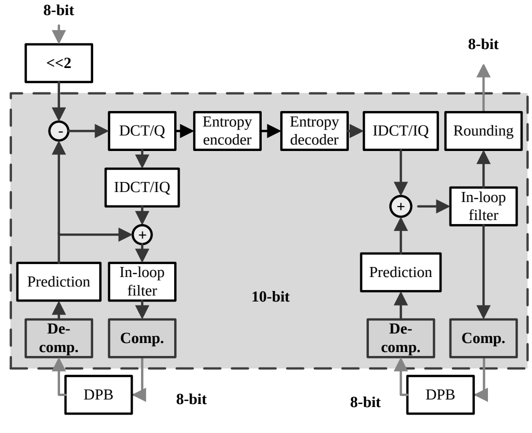


Figure 1. RPMC in hybrid encoding framework

The above methods mostly compress video data based on blocks not smaller than 4x4 pixels, which did not consider about the align cost during data reading and processing, which could introduce extra bandwidth waste. That is, the larger the granularity of the compression units is, the more memory access bandwidth will be wasted when the align/burst parameters of the memory are low, which will be analyzed in the next section. In this paper, considering the impact of granularity in RPMC, an adaptive RPMC scheme based on smaller block size (2x2), is proposed to lower the redundant memory bandwidth accessing especially when the align/burst parameters are low.

In next section, the impact of granularity in RPMC is analyzed. The proposed RPMC scheme based on 2x2 block is described. Then the performance of the proposed scheme is shown by experimental results. Conclusions are given at last.

II. Granularity in RMPC

During the process of motion estimation (ME) and motion compensation (MC), a rectangular area in each reference picture is generally accessed. Burst memory access is considered an efficient way in this case. In fact, since the compression units are generally stored in raster scan order, granularity in RPMC together with the align/burst parameter of memory simultaneously affect the memory bandwidth and random access performance, especially when CPU access the memory without cache. This is based on the following facts.

a) The align address is determined by the compression unit size and the align parameter of memory. That is, the align address internal is the least common multiple (LCM) of them:

I=LCM(Palign, Nunit)≈max{Palign, Nunit} bitswhere Nunit is the bit budget for one compression unit.

b) The basic access unit size is determined by compression unit size and burst parameter of memory. That is, the basic access unit size is S=LCM(Pburst, Nunit)≈max{Pburst, Nunit}bits

where Nunit is the bit budget for one compression unit. If S is determined by Pburst, the minimum access area is a rectangular with the width larger than height; while if S is determined by Nunit, the minimum access area is with a square shape similar with the compression unit. However, the height of the minimum access area will always be the height of the compression unit.

c) The mismatch of align address/basic access unit size with MC rectangular area leads to redundant memory accesses anyhow.

According to the MV, the redundant access without caches could be calculated as follows.

$$\mathrm{N_{redundant}} = \sum_{i,j} p_{i,j} \cdot N_{ri,j}\ (pixels) \tag{3}$$

where pi,j is the probability of MV's value be (i, j), in 1/M pixel; Nri,j is the redundant accessed data amount when MV is (i, j). Nri,j is determined by the differentiation between the actual access area and the size of the block doing MC.

$$\begin{aligned}\mathrm{N_{rdt}} &= \sum_{i,j} p_{i,j} \cdot (N_{ai,j} - N_b) \\ &= \sum_{i,j} p_{i,j} \cdot (w_{ai} \times h_{aj} - w_b \times h_b)\ (pixels)\end{aligned} \tag{4}$$

where wai and hai are the width and height of the actually accessed area, and wb and hb are the width and height of the block doing MC which are fixed by the video coded stream. As the argument (b) denounced, the minimum access area could be with a rectangular or a square shape. The actually accessed area could be not a square but a rectangular area because of the minimum access area or the asymmetry of MV's component.

d) When the Pburst is small (Palign is not larger than Pburst, and normally increases along with Pburst), the redundant memory access is mainly affected by the compression unit size. The smaller the unit size is, the less the redundant memory access is, and the more cost effective. An example is shown in Fig .2.

e) When the Pburst is large, the redundant memory access is affected simultaneously by the parameters of physical memory and compression unit size. The more the unit size is smaller than the Pburst, the more redundant memory access is.

f) The argument (e) is a deduction based on the premise of all the blocks are arranged directly in raster scan pattern. However, the small blocks can be rearranged as larger blocks, in 2x2 or 4x4 patterns, and the larger blocks can be laid out in raster scan then. This rearrangement can deduce the redundant memory access of small blocks when Pburst is large.

The fixed rounding scheme is based on the smallest compression unit, i.e. one pixel [7]. Compared with it, the methods based on 4x4 blocks[8,9] increase the memory access bandwidth remarkably (about 40%) when the align/

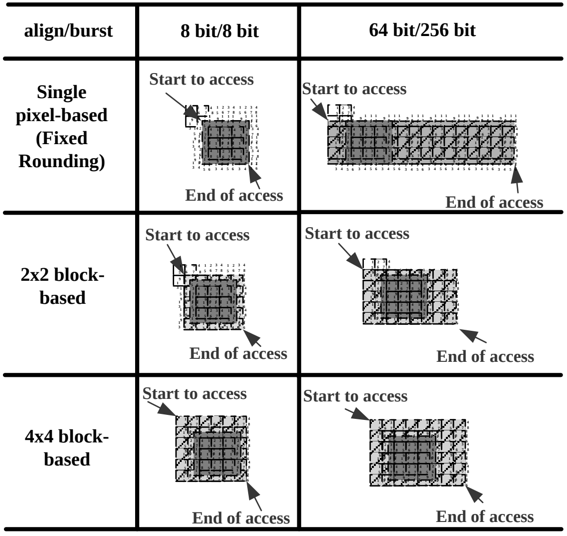


Figure 2. Example of accessing an 8x8 block in reference picture which is not aligned either with 4x4 blocks or with 2x2 blocks while Align/Burst parameters are 8bit/8bit and 64bit/256bit. The really accessed areas are presented in rectangular region for each scheme.

burst parameters are low. What’s more, Flatness features of different block sizes are different, too. It is intuitive that smaller blocks are more flat and with more probability to be lossless compressed.

Based on the aforementioned facts, larger blocks tend to lead to more redundant memory access, and smaller blocks can help to achieve low memory access bandwidth no matter how much the align and burst parameters of memory are. Moreover, smaller blocks are more flat than the bigger ones, and can be compressed almost as efficiently as the bigger ones. Therefore, to tradeoff the compression efficiency and memory access bandwidth consumption, the 2x2 block size would be a flexible and reasonable solution.

III. 2x2 Adaptive Scaling for Reference Pictures Memory Compression

Since the granularity of RFMC affect the access bandwidth seriously, the smaller the block is the lower the access bandwidth can be. When the compression ratio is lower, the access unit can be designed in a smaller size.

This contribution proposes to compress the pixel values by 2x2 unit size for compression from 10-bit depth to 8-bit depth, which is effective based on the following 2 reasons:

a) The flatness of pixel values is much better than 4x4 ones, as shown in Table 1. If we divide the pictures to blocks with size of 2x2, there are more than 80% of them in which the differences between the maximum value and minimum value are below 64. However, if we divide the pictures into 4x4 blocks, there are only 74.27% of them whose differences are below 128. Therefore, the encode efficiency decrease can be controlled just at the same level as 4x4 unit.

b) The 2x2 unit brings smaller granularity and can be reorganized as larger ones, such as 4x4, 8x8, etc. Smaller granularity can bring less redundant access bandwidth with low align/burst parameters. Reorganized as larger units can obtain efficiency with high align/burst parameters or with caches.

B. Compression Algorithm

When the reference pictures are to be compressed from 10-bit/pix to 8-bit/pix, the scaling algorithm is implemented as the concrete process shown in Fig .3. Other compression ratios are similar. In this algorithm, the pixel value will be stored with no more than 2-bit lost in most cases (0-bit lost when S=0, 1-bit lost when S=1, 2-bit lost when S=2，3-bit lost but with offset when s=3).

The de-scaling process is shown in Fig .4, which is inverse progress of those shown in Fig .3.

In fact, this compression is lossless when the R=max-min<64. While 64<=R<128 and 128<=R<256, the residual will be stored separately, which could be accessed by an address table.

C. Reorganization of reference frame memory

The 2x2 unit brings smaller granularity and can be reorganized as larger ones, such as 4x4, 8x8, etc. Smaller granularity can bring less redundant access bandwidth with low align/burst parameters. Reorganized as larger units can obtain efficiency with high align/burst parameters or with caches.

```
Compress (scaling) process
//calculate and store the max and min value in the 2x2 block
min=minivalue(block), max=maxvalue(block);
//calculate the range of values in the block, and the quantizer Q
R=max-min+1;
Q=ceil(R/26);
//for each pixel except the minimum one, do quantization
if(R<28){
for (each pel in block except the minimum one){
   P= (Vorg-min)/Q; }
}
else{ //if R is too large, use fixed rounding,
   O=0;
   for(each pel in block){
     P=Vorg/23;
     O+= Vorg -Vorg/23×23; }
O/=8;}
```

Figure 3. Compress (scaling) process for 10-to-8 bits

```
Decompress (de-scaling) process
//read the first 2 bits, obtain the flag, then get the
corresponding value of s // u(2)

if（s==3）//s:[0..3] flag:11, s=3
{
      for(i=0;i<4;i++)
        d[i]=p[i]<<3+(offset<<1);
}
else//flag not 3, which means s=0 ,1 or 2
{
  //get the values in turn: 10-bit min,2-bit minidx,  3*6-
    bit p[i] (0<=i<4 and i!=minidx);
      d[minidx]=min;
    for(i=0;i<4&&i!=minidx;i++)
          d[i]=(p[i]<<s)+min;
}
```

Figure 4. aCompress (scaling) process for 10-to-8 bits

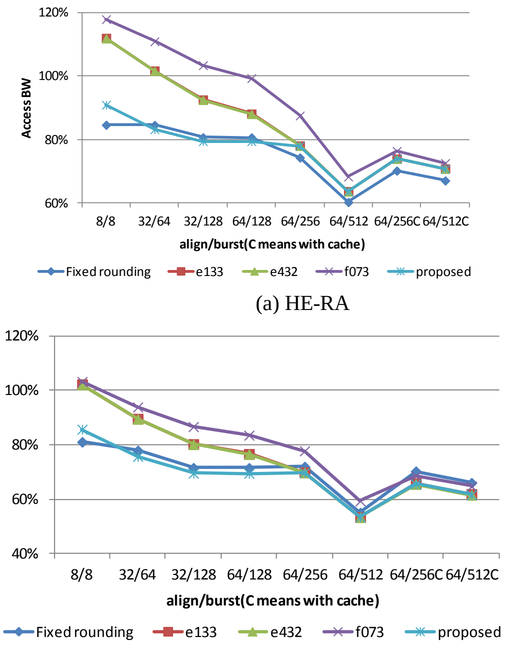
By analyzing the size of unit to be compressed better for bandwidth’s reduction, we make use of the 2x2 unit’s flexibility and reorganize them when align/burst parameter is high.

The principle of reorganization is to match the unit size and burst size. That is, the combined-unit produced by reorganization should be as large as the burst size.

IV. Experimental Results

To evaluate the effectiveness of the method and the impact to video compression efficiency, the experiments are all implemented on HEVC Model(HM) [12, 13]. In the experiments, we tested and compared the fixed rounding, several compression schemes based on 4x4 blocks [5, 10, 11], and the proposed scheme based on 2x2 blocks.

In terms of memory access bandwidth, when the values of align/burst parameters are high, those based on 4x4 blocks have achieved remarkable memory bandwidth decrease with the cost of encoding efficiency slightly dropped off, as shown in Fig .5. But when the parameters’ values are low, they caused high memory access bandwidth, even higher than no compress scheme. And compare to the fixed rounding scheme, they cause about 40% more accesses.



(a) HE-RA

(b) HE-LD

Figure 5. The relative access bandwidth with different memory parameters

The proposed method decreased the memory access bandwidth by 15~17% compared with the 4x4 block based methods when burst parameter is smaller than 64, and about 9% when burst parameter is 128. For large burst parameters the memory bandwidth can be kept the same as the 4x4 block based compression methods by the non-normative trick that merges four 2x2 blocks to one 4x4 block as the basic unit for store and access. And compared with no compression scheme, on average, 9.1%~29.26% with RA, 14.32%~46.65% with LD of bandwidth is decreased by the proposed method.

V. CONCLUSIONS

A 2x2 block based RPMC scheme is proposed in this paper. By reducing the granularity of compression unit, memory access bandwidth has been distinctly decreased with little compression efficiency loss.

It is suitable for portable devices application. And by considering about the residual of compression in while MC being implemented, it is compatibility with the existed standards.

ACKNOWLEDGMENT

The authors would like to thank the support from the National Natural Science Foundation of China (61471277), the Programmer of Introducing Talents of Discipline to Universities (B08038) and the Fundamental Research Funds for the Central Universities (72115612).

REFERENCES

[1] Ajit D. Gupte, Bharadwaj Amrutur, Mahesh M. Mehendale, Ajit V. Rao, and Madhukar Budagavi, “Memory bandwidth and power reduction using lossy reference frame compression in video encoding,” IEEE Trans. circuits and systems for video technology, vol.21, no. 2, Feb. 2011, pp. 225-230. doi: 10.1109/tcsvt.2011.2105599

[2] Daniel F. Finchelstein, Vivienne Sze, and Anantha P. ChandrakasanMulticore, “Processing and efficient on-chip caching for h.264 and future video decoders,” IEEE Transitions on circuits and systems for video technology, vol.19, Nov. 2009, pp. 1704 - 1713 . doi: 10.1109/tcsvt.2009.2031459

[3] Ken McCann, Benjamin Bross, Shun-ichi Sekiguchi, Woo-Jin Han, “HM3: High Efficiency Video Coding (HEVC) test model 3 encoder description,” JCTVC-E602, Geneva, Mar. 2011.

[4] M. Budagavi and Z. Minhua, “Video coding using compressed reference frames,” in Proc. Int. Conf. Acou., Speech Signal Process. Apr. 2008, pp. 1165–1168.

[5] Takeshi Chujoh, Komukai-Toshiba-cho, Saiwai-ku, “Adaptive scaling for reference pictures memory compression,” JCT-VC Document, JCTVC-E133, Geneva, Mar. 2011.

[6] Zhan Ma and Andrew Segall, “Frame buffer compression for low-power video coding,” IEEE International Conference on Image Processing, ICIP2011, sept. 2011, pp. 757 – 760.

[7] Jaemoon Kim and C.M Kyung, “A lossless embedded compression using significant bit truncation for HD video coding”, IEEE Trans. CSVT, Vol. 20, Jun. 2010, pp.848-860. doi: 10.1109/tcsvt.2010.2045923

[8] Silveira D., Povala G., Amaral L., Zatt B., Agostini L., Porto M., “An energy-efficient hardware design for lossless reference frame compression in video coders,” Proc. IEEE International Conference on Electronics, Circuits, and Systems, ICECS 2013 Dec. 2013, pp. 573-576. doi: 10.1109/icecs.2013.6815479

[9] Dieison S., Guilherme P., L ívia A., Bruno Z., Luciano A., Marcelo P., “Memory bandwidth reduction for H.264 and HEVC encoders using lossless reference frame coding,” IEEE International Symposium on Circuits and Systems (ISCAS), Jum. 2014, pp. 2624-2627. doi: 10.1109/iscas.2014.6865711

[10] Dzung Hoang, “Unified scaling for 10-bit to 8-bit reference frame compression,” JCT-VC Document, JCTVC-E432, Geneva, Mar. 2011.

[11] Shan Liu, Ximin Zhang, Shawmin Lei, “Joint Luma-Chroma adaptive reference picture memory compression,” JCTVC-F073, Torino, Jul. 2011.

[12] Information on http://phenix.int-evry.fr/jct/

[13] Frank Bossen, “Common test conditions and software reference configurations,” JCT-VC Document, JCTVC-E700, Geneva, Mar. 2011.