EUSIPCO 2013 1569743407

Lossless Embedded Compression Using Multi-mode DPCM & Averaging Prediction for HEVC-like Video Codec

Li Guo, Dajiang Zhou, and Satoshi Goto
Graduate School of Information, Production and Systems, Waseda University, Japan
Email: guoli@toki.waseda.jp

ABSTRACT

Video encoders and decoders for HEVC-like compression standards require huge external memory bandwidth, which composes a significant portion of the codec power consumption. To reduce the memory bandwidth, this paper presents a new lossless embedded compression algorithm and a high-throughput hardware architecture. Firstly, hybrid spatial-domain prediction is proposed to combine the merits of DPCM scanning and averaging. The prediction is then enhanced with multiple modes to accommodate various image characteristics. Finally, efficient residual regrouping is used to improve the compression performance based on semi-fixed length (SFL) coding. Experimental results show the proposed scheme can reduce data traffic by an average of 57.6%. The average compression ratio is 2.49, improved by at least 13.2% relative to the state-of-the-art algorithms. Although the complexity is increased, by applying several optimizations the hardware cost of designed architecture increases slightly. This work can be implemented with 54.2k gates cost for compressor and 46k gates for decompressor, which can support 4k×2k@120fps decoder.

Index Terms— Embedded compression, lossless reference frame recompression, multi-mode DPCM and averaging prediction, HEVC, H.264/AVC

1. INTRODUCTION

In video codec systems, including encoders and decoders for HEVC, H.264/AVC, MPEG-2, etc., usually a large external DRAM is required for buffering mass data such as the reference frames. As a result of the huge bandwidth requirement, power consumption of memory access occupies a significant part of system power [1]. Therefore, techniques to overcome memory bandwidth problem play an extremely important role. Much work has been done from various points of view, including reusing the overcalled reference frame data on various levels and improving the DRAM access efficiency by optimized memory controller architectures [2], etc.

Embedded compression (EC), also known as *reference frame recompression*, is another effective solution to the DRAM bandwidth problem. It has been widely discussed in previous work and demonstrated in several video decoder chips [3]. To reduce the memory data traffic, EC works as an additional layer between the codec core and DRAM controller, which compresses the reference frames before storing them into DRAM and decompresses the data fetched back.

The existing EC schemes can be briefly divided into two categories. One is lossy EC based on fixed compression ratio (CR). However, fixing CR inevitably leads to image quality degradation [4], while the error propagation caused by the quality loss of reference frames can be an even more critical issue. The other category is lossless EC. Lossless EC schemes must be based on variable compression ratio. As a result, special memory organization is required to store and fetch the sizes of the compressed data partitions [5].

The data compression of lossless EC is usually composed of prediction and entropy coding. For the prediction stage, both spatial domain [3][6]and frequency domain [7] prediction are widely used. For the entropy coding stage, although variable length coding (VLC) is widely used, there are many difficulties in implementing high-throughput VLC hardware. So latest EC schemes [3][6] employed two-step entropy coding: to first separate the residuals into small groups, and then perform fixed length coding based on the local residual feature of each group.

This paper presents a new lossless EC algorithm. Firstly, hybrid spatial-domain prediction is proposed to combine the merits of DPCM scanning and averaging. Prediction is then enhanced with multiple modes to accommodate the various image characteristics. Finally, efficient residual regrouping is used to improve the EC performance based on SFL coding.

The rest of this paper is organized as below. The proposed multi-mode DPCM scanning and averaging (MDA) scheme is explained in Section 2. Hardware implementation is described in Section 3. Section 4 shows the experimental results of several previous works and proposal. The conclusions are drawn in Section 5.

2. PROPOSED ALGORITHM

2.1. Spatial Correlation Analysis

Among the existing spatial-domain prediction algorithms, hierarchical average and copy (HAC) [6] shows competitive

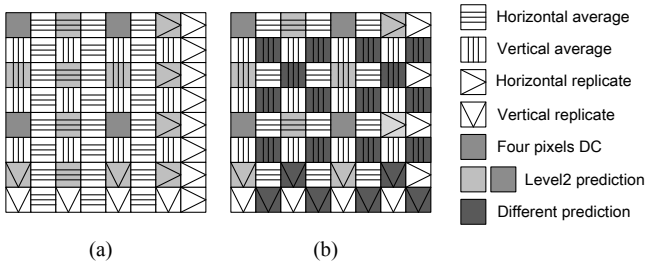

Fig. 1. HAC prediction [6] and its transpose, (a) HAC, 2-pixel or 4-pixel distance pixels are used in level 2, (b) HAC transpose and different predicted pixels compared with HAC

performance in embedded compression. Spatial correlation between the neighboring pixels in natural sequences has been discussed. Based on this statistic, HAC prediction mainly uses average prediction of two neighboring pixels. For some pixels which can't be predicted with adjacent pixels' value, residuals are computed by pixels from longer distance, as is shown in Fig. 1 (a). However, their spatial correlation is much weaker than that of neighboring pixels. The statistical features of HAC with different prediction distance are shown in Fig. 2.

The entropy of HAC residuals calculated by neighboring pixels' prediction is only 3.7677, by prediction of 2-pixel distance pixels is up to 4.7594 and by 4-pixel distance DC prediction is 5.23. So level-2 prediction of HAC negatively affects the performance of compression.

Except for prediction distance problem, HAC is also not suitable for multi-mode extension. HAC transpose and the different predicted pixels compared with HAC are shown in Fig. 1(b). There are only 20 pixels with different prediction technique between HAC 8×8 prediction block and its transpose. Moreover, the predicted values of most pixels in level 2 are still the same. So the addition of HAC transpose mode is not promising in improving compression ratio.

2.2. Multi-mode DPCM & Averaging Prediction

Although average prediction of two neighboring pixels is better than directly replicate, some pixels are inevitably predicted by 2-pixel or 4-pixel distance pixels, which may lower the compression performance. So it is a tradeoff between better prediction mode and shorter prediction distance.

As shown in Fig. 3(a), prediction mode (PMode) 0 is a basic mode of the proposed prediction scheme. The predictions of pixels in row 0 are calculated by horizontal DPCM scanning. Then 28 pixels in even columns are predicted by DPCM scanning in vertical direction. Horizontal predictions are used for pixels in odd columns. In columns 1, 3 and 5, pixels are predicted by the average of left and right neighboring pixels. In the last column7, 7 pixels are predicted by replicating the value of left pixel. The remaining one top-left pixel is presented by original 8-bit value. So the proposed PMode 0 uses only adjacent pixels for prediction, avoiding the problem of predicted by 2-pixel or 4-pixel distance pixels in HAC. The average prediction is defined as follows:

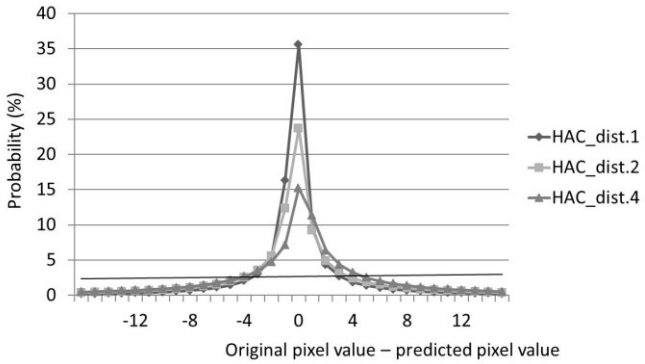

Fig. 2. The probability distribution of HAC [6] residuals predicted by pixels from different distance

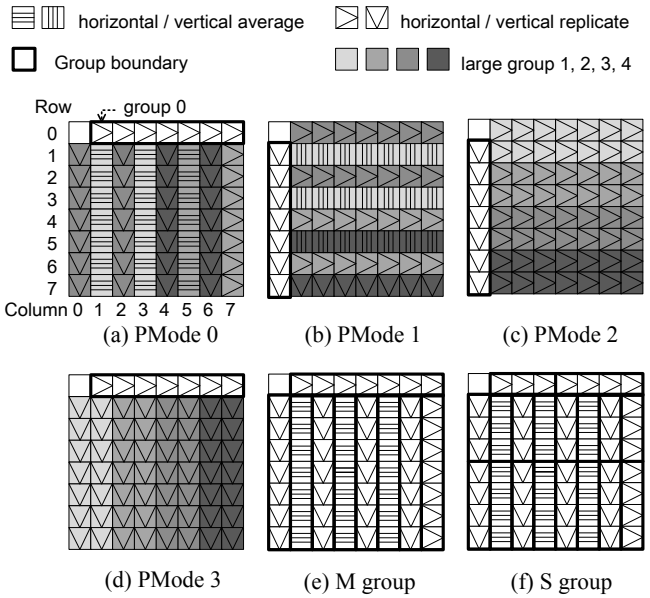

Fig. 3. Proposed MDA prediction and regrouping schemes. (a)~(d) prediction modes and minimum grouping division (one M group0 and 4 L group1~4), (e) only M group, (f) only S group

$$\mathrm{Pr}_i = \frac{P_{i-1} + P_{i+1}}{2} \qquad (1)$$

where Pr_i is predicted value of current pixel i, P_{i-1} and P_{i+1} are two adjacent pixels' value in horizontal or vertical direction.

In natural sequences, the features of blocks are variable, including gradual changed block, smooth block, etc. Therefore, only one prediction mode is not generally suitable for different feature blocks. Multi-mode predictions are proposed to accommodate variable image characteristics. PMode 1 is the transpose of PMode 0, which has 43 different predicted pixels, shown in Fig. 3(b). Vertical averaging and horizontal DPCM scanning are utilized in prediction PMode 1. However, the directions of averaging and DPCM scanning prediction are quadrature in PMode 0 or PMode 1. So these two modes are not suitable for those gradual changed block. So simple horizontal and vertical DPCM scanning modes (PMode 2 and PMode 3) are added, shown in Fig. 3(c) (d).

Table 1. The probability of four MDA prediction modes taken as selected prediction mode

Qp	PMode 0	PMode 1	PMode 2	PMode 3
22	29.34%	26.88%	22.80%	20.97%
27	29.68%	27.23%	22.13%	20.96%
32	30.83%	28.15%	21.10%	19.91%
37	32.12%	29.34%	19.62%	18.93%
Average	30.49%	27.90%	21.41%	20.19%

We calculate the probability of these four prediction modes taken as selected mode. The experimental results are shown in Table 1. The probability is the average of 18 test sequences with all frames. The detail experiment condition is shown in Section 4. According to the experimental results, it is obvious that the addition of every mode is reasonable. So these four prediction modes are combined as Multi-mode DPCM & Averaging (MDA) prediction.

2.3. Residual Regrouping for MDA

2.3.1. Basic Residual Grouping

Residuals need to be grouped before significant bits truncation (SBT) or semi-fixed length (SFL) coding. Better compression performance can be obtained by efficient residual grouping scheme. To meet the feature of SFL and SBT coding, we combine the residuals with similar distribution into one group. So usually the residuals predicted by the same prediction technique are grouped together. One 8×8 prediction block can be divided into 9 groups shown in Fig. 3(e), according to the same prediction technique of averaging or DPCM scanning. In order to make group size the same, there are 7 pixels in every group. Such a 7-pixel group is named as middle size group (M). However, if only one or two residuals in one middle group are very large, two small size groups (3 or 4 pixels/group, S) are more efficient than one middle group. A grouping example with only small groups is presented in Fig. 3(f). On contrary, for some smooth blocks, the coding modes (CM) of SFL coding in one block are similar. So large size group (L) is introduced to combine two middle groups together as shown in Fig. 3(a) ~ (d), and the bits of coding mode CM can be saved. These three are the basic residual groups (small, middle and large).

2.3.2. Residual Regrouping

Using small groups, less number of bits is needed to present residuals, compared with middle or large groups. However, the overhead of coding mode CM is twice more than middle group, and 3.6 times more than large group. Therefore, the regrouping scheme of four small groups within one large group is proposed to reduce the overhead bits of small groups. In most case, the four coding modes of small groups within one large group are similar or even the same, so the small groups with the same coding mode can be regrouped together.

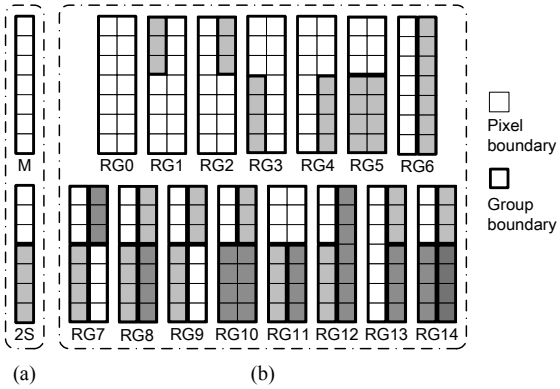

Fig. 4. Regrouping schemes, (a) 2 RG modes for M group 0, (b) 15 RG modes for L group1~4, the same gray scale means same CM

Semi-fixed-length Code Table

mm		0	1	2	3~4	5~8	9~16	17~32	>=33
M(mode)		0	1	2	3	4	5	6	7
D (Coding)	0		0	00	000	0000	00000	000000	Original 8-bit value
	±1		1	S1	SS1	SSS1	SSSS1	SSSSS1	
	±2			10	S10	SS10	SSS10	SSSS10	
	±3				S$\bar{S}$1	SS$\bar{S}$1	SSS$\bar{S}$1	SSSS$\bar{S}$1	
	±4				100	S100	SS100	SSS100	
	±5					S$\bar{S}$01	SS$\bar{S}$01	SSS$\bar{S}$01	
	±6					S$\bar{S}$10	SS$\bar{S}$10	SSS$\bar{S}$10	
	±7					S$\bar{S}\bar{S}$1	SS$\bar{S}\bar{S}$1	SSS$\bar{S}\bar{S}$1	
	±8					1000	S1000	SS1000	
	...								
	±16						10000	S10000	
	...								
	±32							100000	

"S" is the sign bit of residuals, and "$\bar{S}$" is logic negation of S.
XX: additional trailing bit T is used to indicate the sign bit.

Fig. 5. Semi-fixed-length code table

As Fig. 3(a) ~ (d) shows, one 8×8 block can be divided into 4 large groups and one middle group. For the middle group 0, two small groups can be regrouped if their coding modes are the same, as Fig. 4(a) shown. 1 bit additional flag is enough for indicating regrouping modes (RGM). For the large group 1~4, there are 15 regrouping modes shown in Fig. 4(b). A 3-bit flag is required for RG0, while 4 bits are needed for other regrouping modes.

2.4. Semi-Fixed Length Coding

For entropy coding part, semi-fixed length (SFL) coding proposed in [3] is used. Based on the minimum and maximum residual values in one group, a coding mode (CM) is decided. Then the residuals are coded (to be D) according to semi-fixed-length code table shown in Fig.5. Since a trailing bit (T) is added, the presenting range for M between 1 and 6 is $[-2^{M-1}+1, 2^{M-1}]$ or $[-2^{M-1}, 2^{M-1}-1]$.

2.5. Overall Processing Flow

Fig. 6 shows the overall processing flow of the proposed EC scheme. Reference frames are divided into 8×8 blocks. The

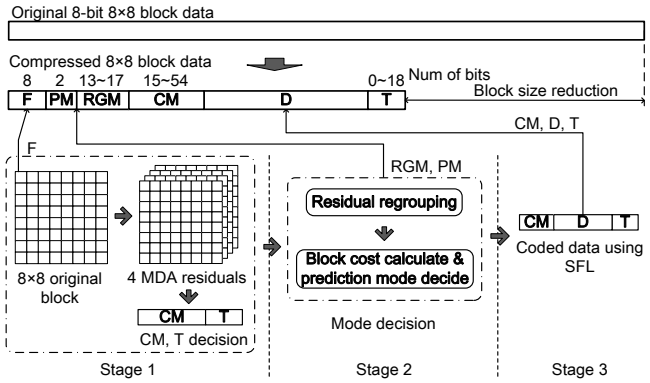

Fig. 6. Overall processing flow of proposed MDA

processing flow can be divided into three stages of residual calculating, prediction mode decision and coding. In stage 1, four MDA residual blocks are calculated by subtracting the predicted value of MDA from original pixel value. The top-left pixel is presented with its original 8-bit value (F). Then coding mode (CM) and trailing bit (T) for small size groups are produced. In the mode decision stage 2, after small groups with the same CM are regrouped, bit cost of all modes are evaluated. The one with minimum bit cost is chosen as prediction mode (PM). In stage 3, all the residuals calculated by PM and grouped by RGM, are coded with SFL code. After SFL coding, all data including F, PM, RGM, CM, D and T are merged into a bit stream.

3. HARDWARE IMPLEMENTATION

The proposed MDA hardware composes of a compressor and a decompressor. Based on the feature of proposed MDA, several optimizations are considered to reduce complexity. A 3-stage pipelined architecture is designed for MDA compressor, as shown in Fig. 6 and Fig. 7. In order to reduce the computational complexity, two steps hardware architecture of prediction mode decision (stage1, 2) and coding (stage3) is proposed. In stage 1, all residuals for four prediction modes are calculated before coding mode decision. Since some pixels have the same prediction technique within four prediction modes, predicted value can be reused to reduce computational complexity by 33%. Also a simplified coding mode decision algorithm can further reduce complexity. In stage 2, after residual regrouping, the bit cost of all prediction modes are evaluated, and the minimum one is taken as the selected prediction mode (PM). In stage 3, residuals of selected mode are coded by SFL, so the high complexity part only needs to process one mode. Therefore, although multiple prediction modes are included, the complexity increase of proposed architecture is much less than being proportional to the number of modes. Based on the pipeline, one 8×8 block can be finished in 6 cycles, so throughput of 10.7 samples per cycle can be achieved.

Fig. 8 shows the 3-stage pipeline architecture of MDA decompressor. In stage 1, the input data are shifted to 5 regis-

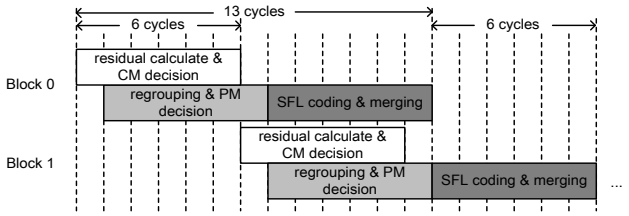

Fig. 7. Pipeline stages of MDA encoder

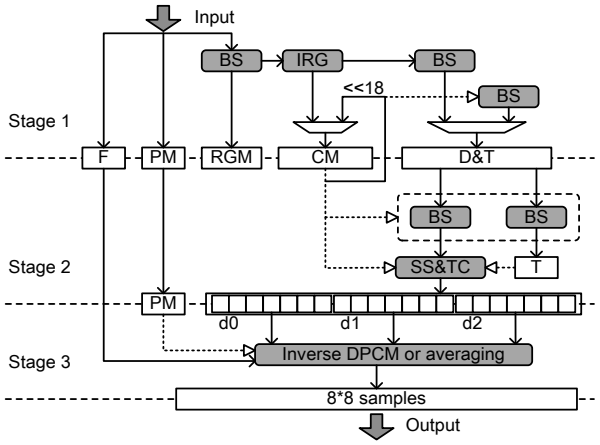

Fig. 8. 3-stage pipeline architecture of MDA decoder. BS denotes "barrel shifter". IRG denotes "inverse regrouping". SS&TC denotes "sub-block splitting and trailing bit compensation".

ters, including F, PM, RGM, CM and D&T. The coding mode (CM) for each S group is decoded according to regrouping mode (RGM). In stage 2, after coded residuals are separated to M group by barrel shifters (BS), they are further split and decoded to independent residuals. Finally, in stage 3, samples are reconstructed by inverse DPCM scanning or averaging. As a result, only 3 cycles are required for decoding 8×8 block, and the throughput is up to 21.3 pixels/cycle.

4. EXPERIMENTAL RESULTS

To evaluate the efficiency of proposed MDA algorithm, it is integrated with reference software HM 8.0 decoder. All reconstructed frames of 18 test sequences are coded (shown in Fig.9). The configuration of low delay coding main mode is used. Quantization parameters (Qp) are 22, 27, 32 and 37.Together with proposed algorithm, the performance of three previous works [3][6][7] is also simulated as comparison. Unify block size to 8×8. Both Compression Ratio (CR) and Data Reduction Ratio (DRR) are widely used factors to evaluate the efficiency of EC scheme. The original data size divided by compressed data size is defined as CR. DRR is the percent of reduced data size compared with original data size.

Since only EC schemes for luma block are clearly described in previous works, so we compare average CR and DRR of all 18 sequences for luma blocks between proposal and previous works [3][6][7] in Table 2. Show average results of YUV under 4:2:0 in Table 3. In addition, the average CR

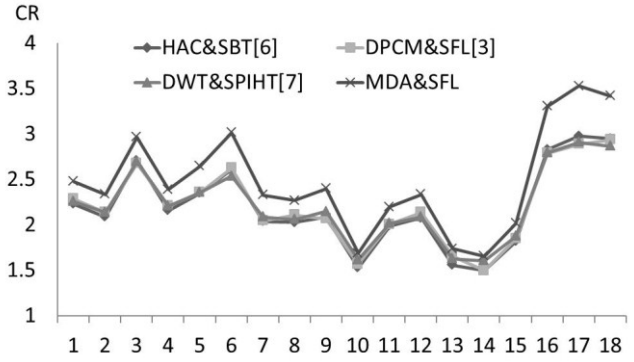

Fig. 9. Experimental results CR comparison. (18 sequences from class A to E: Traffice, PeopleOnStreet, Kimono, ParkScene, Cactus, BasketballDrive, BQTerrace, BasketballDrill, BQMall, PartyScene, RaceHorses, BasketballPass, BQSquare, BlowingBubble, RaceHorse, vidyo1, vidyo3 and vidyo4)

Table 2. Average CR and DDR comparison

Qp	HAC+SBT [6]		DPCM+SFL [3]		DWT+SPIHT [7]		MDA+SFL Proposed	
	CR	DRR %	CR	DRR %	CR	DRR %	CR	DRR %
22	2.06	48.87	2.09	49.78	2.09	50.15	2.33	54.22
27	2.17	51.69	2.20	52.51	2.20	52.76	2.47	57.16
32	2.25	53.65	2.27	54.24	2.27	54.43	2.55	58.88
37	2.31	55.22	2.31	55.49	2.31	55.76	2.60	60.12
Avg	2.20	52.36	2.22	53.01	2.22	53.28	2.49	57.60

Table 3. Average CR and DRR for only luma and 4:2:0 YUV block

Qp	CR				DDR %			
	22	27	32	37	22	27	32	37
Luma	2.33	2.47	2.55	2.60	54.22	57.16	58.88	60.12
4:2:0	2.58	2.73	2.82	2.89	58.96	61.35	62.88	64.11

Table 4. Results comparison with previous architecture

	HAC&SBT	DWT& SPIHT	Proposed MDA&SFL	
	Comp. or Decomp. [1]		Comp.	Decomp.
Unit	16×8	16×16	8×8	
CMOS tech. (nm)	180	180	65	
Max. freq. (MHz)	180	10	300	
Throughput(Gpixles/s)	0.9 or 2.6	0.005	3.2	6.4
Throughput(pixels/cycle)	5.1 or 14.2	0.45	10.7	21.3
Gate count (k)	36.1	26.9	54.15	45.97
TPUA [2] (10^{-5}pixels/cycle/gate)	14.1(com.) or 39.3(dec.)	1.7	19.8	46.3

[1] Compressor and decompressor can't be used at the same time.
[2] TPUA[6] : (throughput / gate count), is the evaluation criterion to consider both hardware cost and throughput.

results of four Qp are compared between 18 test sequences in Fig. 9. It is obvious that proposed EC performs much better in every sequence. On average, CR of proposed algorithm can achieve up to 2.49 with no quality degradation and no bitrate increment. Compared with previous HAC & SBT [6], proposed scheme can improve CR by 13.2%. The average data reduction ratio of MDA is about 57.6%.

The detail hardware implementation results of proposed MDA architecture are shown in Table 4. Compared with HAC&SBT [6] and DWT&SPIHT [7], which can't compress and decompress at the same time, the separate design of compressor and decompressor is able to fit the characteristic of codec better. So although the complexity of proposal is larger, the efficiency of TPUA can be higher. With reasonable gate cost increase, the proposed EC architecture is able to achieve higher throughput and maximum frequency than previous works [6][7], which can meet the requirement of 4k×2k @120fps video decoder.

5. CONCLUSION

This paper proposes a new lossless embedded compression algorithm to reduce external memory bandwidth. Multi-mode hybrid spatial-domain prediction with only neighboring pixels is proposed. Before SFL coding, efficient residual regrouping is performed to further improve compression ratio. Experimental results show that the data reduction ratio of proposed EC is about 57.6% on average with no quality degradation. Average compression ratio of 2.49 is achieved, improved by at least 13.2% compared with other previous works. The proposed architecture achieves a throughput of 10.7 pixels/cycle with 54.2k gates cost for EC compressor, while decompressor needs 46k gates 21.3 pixels/cycle throughput, which can support 4k×2k @120fps decoder

ACKNOWLEDGEMENT

This research was supported in part by the University Joint Research Project of STARC, Japan.

REFERENCES

[1] M. Budagavi and M. Zhou, "Video coding using compressed reference frames," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp.1165-1168, 2008.
[2] J. Zhu, P. Liu, and D. Zhou, "An SDRAM controller optimized for high definition video coding application," in Proc. IEEE Int. Symp. Circuits Syst., pp. 3518–3521, 2008.
[3] D. Zhou, J. Zhou, X. He, J. Zhu, J. Kong, P. Liu, and S. Goto, "A 530 Mpixels/s 4096x2160@60fps H.264/AVC High Profile Video Decoder Chip," IEEE J. Solid-State Circuit, vol.6, no.4, pp. 777-788, 2011.
[4] L. Song, D. Zhou, X. Jin, and S. Goto, "A constant rate bandwidth reduction architecture with adaptive compression mode decision for video decoding," EUSIPCO, pp.2017-2021, 2010.
[5] X. Bao, D. Zhou, P. Liu and S. Goto, "An advanced hierarchical motion estimation scheme with lossless frame recompression and early-level termination for beyond high-definition video coding", IEEE Trans. on Multimedia, vol.14, no.2, pp.237-249, 2012.
[6] J. Kim and C.-M. Kyung, "A lossless embedded compression using significant bit truncation for HD video coding", IEEE Trans. Circuits Syst. Video Techn., vol.20, no.6, pp. 848-860, 2010.
[7] C.-C. Cheng, P.-C. Tseng, and L.-G. Chen, "Multimode embedded compression codec engine for power-aware video coding system," IEEE Trans. Circuits Syst. Video Techn., vol.19, no.2, pp. 141-150, 2009.