FAST MULTI REFERENCE FRAME MOTION ESTIMATION FOR HIGH EFFICIENCY VIDEO CODING

Shanshe Wang[1], *Siwei Ma*[2], *Shiqi Wang*[2], *Debin Zhao*[1], *Wen Gao*[2], *Fellow IEEE*

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
[2]Institute of Digital Media, Peking University, Beijing, China
sswang0904@gmail.com, swma@pku.edu.cn, sqwang@jdl.ac.cn, dbzhao@hit.edu.cn, wgao@jdl.ac.cn

ABSTRACT

In this paper, a fast multi reference frame motion estimation algorithm is proposed to reduce the computational complexity for high efficiency video coding (HEVC). Firstly, according to the reference frame distribution, we define the motion complexity for a frame. Based on the motion complexity, an initial reference frame set (RFS) is constructed to reduce the number of the reference frame for motion estimation. Then for each prediction unit (PU), the average distortion per pixel together with the motion vector difference (MVD) in the first reference frame is employed to shrink the RFS in order to early terminate the motion estimation process. In addition, to enhance the robustness of the proposed scheme, an expansion method for RFS is proposed to guarantee the video quality. Experimental results demonstrate that the proposed scheme can significantly save the encoding time. For Low Delay testing configuration, over 30% of time saving can be achieved on average with ignorable performance loss.

Index Terms—High Efficiency Video Coding, motion estimation, reference frame set, motion complexity

1. INTRODUCTION

High efficiency video coding (HEVC) is the latest video coding standard developed by Joint Collaborative Team on Video Coding (JCT-VC) [1]. Many new coding tools are adopted to improve the coding performance. The adaptive quad-tree coding structure based on the coding tree unit (CTU) is one of the key tools [2].

In CTU, three new concepts named coding unit (CU), prediction unit (PU) and transform unit (TU) are introduced to specify the basic processing unit of coding, prediction and transform. CU can have various sizes and allows recursive quad-tree splitting. Given the size of CTU and the maximum hierarchical depth, CU can be expressed in a recursive quad-tree representation adapted to the picture content as illustrated in Fig. 1 (a). Once the splitting of CU hierarchical tree is finished, the leaf node CUs can be further split into PUs. PU is the basic unit for prediction and it allows multiple different shapes to encode irregular image patterns as shown in Fig. 1 (b). The size of PU is limited to that of CU with square or rectangular shape. However, for intra CU and PU splitting, 2Nx2N and NxN partion mode are used, and NxN partition mode is allowed only when the corresponding CU size is equal to the minimun CU size. TU is defined to represent the basic unit for transform. For inter mode, the size of TU is independent with the size of PU; while for intra mode, the size of TU cannot exceed the size of PU. The size of TU cannot exceed the size of CU for both intra and inter mode. Although this highly flexible coding structure provides the encoder great flexibility to improve the coding performance, it also imposes very high computation burden to the encoder, especially for the multi reference frame motion estimation (ME). Thus, a fast multi reference frame ME algorithm is highly desired.

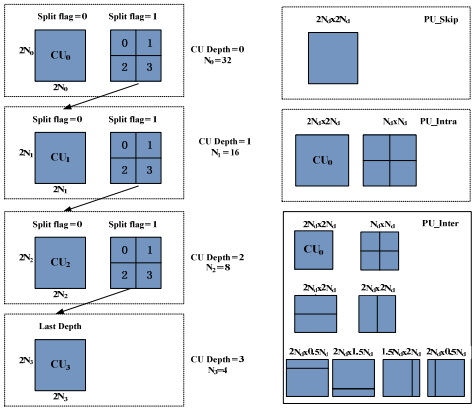

Fig. 1. (a) Recursive CU structure in HEVC. (LCU size= 64, maximum hierarchical depth = 4), (b) PU splitting for skip, intra and inter in HM

Many fast multi reference frame ME algorithms have been proposed for H.264/AVC [3]. In [4], Ho *et al.* proposed an algorithm to minimize the reference frame selection times for the variable size block. Thus the computational complexity of ME was reduced significantly. However, only two kinds of partition size, 16x16 and 8x8, are considered, which is not suitable for the quad-tree coding structure in HEVC. In [5], based on the correlation and continuity of motion vectors among different reference

frames, an adaptive scheme is proposed to select the best reference frame to avoid the exhaustive searching. In [6], by exploiting the motion correlation of adjacent macro-blocks (MB) in both the spatial and temporal domain, Tang *et al.* proposed an optimal reference frame selection scheme. Based on the observation that the best reference frames of the current MB are highly correlated with the characteristic of the motion search region, a fast reference frame selection scheme is proposed in [7], but the quality loss is noticeable. In [8], Huang *et al.* developed context-based adaptive criteria to determine whether it is necessary to do multi reference frame motion estimation. In [9], based on support vector machine (SVM) classifier, a statistical learning method is proposed to model multi reference frame selection as a classification problem, and unnecessary reference frames are eliminated in the searching process.

In this paper, a fast multi reference frame ME algorithm is proposed for HEVC by exploiting both the global and local characteristics of video frames. Firstly, we define the motion complexity for a frame. Based on the motion complexity, an initial reference frame set (RFS) is constructed to reduce the number of reference frame. Then, an adaptive RFS shrinkage scheme is proposed to early terminate the ME process. In addition, a RFS expansion method is provided by adding one more other reference frame to the RFS, which can avoid the local optimization of the initial RFS and guarantee the video quality.

The rest of this paper is organized as follows. Section 2 introduces our proposed fast multi reference frame ME algorithm for HEVC. Experimental results are presented in Section 3. Finally, we conclude the paper in Section 4.

2. PROPOSED FAST MULTI REFERENCE FRAME SELECTION ALGORITHM FOR HEVC

As referred in [10], the reference selection problem for HEVC can be represented as follows,

$$\{s(1),...,s(i),...s(n)\}_{opt} = \bigcup_{i=1}^{n} \arg\min_{s(i)}(D_{f,i} + \lambda_i R_{f,i}) \ s.t. |s(i)| \le r \quad (1)$$

where $s(i)$ is the RFS for frame i and r indicates the maximum number of reference frame. $D_{f,i}$ and $R_{f,i}$ denote the distortion and the coded bits respectively, and λ_i is the Langrage multiplier.

For each potential reference frame, the ME process for a PU can be described as

$$J(ref \mid \lambda_m) = \min_{s(i)}(D_i + \lambda_m R_i) \quad s.t. |s(i)| \le r \quad (2)$$

$$D_i = SA(T)D(s, c(ref, m(ref))) \quad (3)$$

$$R_i = R(MVD(ref) + R(ref)) \quad (4)$$

where D_i denotes the distortion of the encoded PU and R_i denotes the bits used to encode the reference frame index and motion vector difference (MVD) of the PU. SA(T)D is the sum of absolute (transformed) difference, which is a measure of distortion. s and c denote the original video signal and the coded video signal respectively. The task of ME is to achieve the most suitable motion vector for the PU.

In HEVC, the best motion vector for a PU can be obtained by a full search in all reference frames, yet the computation complexity is high in case of many reference frame candidates. Thus, it is necessary to develop a fast algorithm on multi reference frame ME based on the characteristics of the reference frames, which can balance the computational complexity and compression efficiency.

2.1 Motion Complexity

In ME for a PU, each frame in RFS has different probability to be selected as final reference frame. The probability generates reference frame distribution (RFD). Table 1 shows the RFD for some adjacent frames of the sequence *Kimono* with low motion activity. The number of reference frame is set as four. It can be seen that the RFD of the PUs in adjacent frames has great similarity. And there is small probability variation in a temporal local window. While in some sequences with high motion activities e.g. *BQSquare*, the RFD varies much as illustrated in Table 2. So RFD can characterize the motion complexity effectively.

Table 1. Reference frame distribution in *Kimono*

POC	Ref0	Ref1	Ref2	Ref3
6	0.891	0.083	0.018	0.008
7	0.925	0.042	0.031	0.002
8	0.905	0.056	0.037	0.002

Table 2. Reference frame distribution in *BQSquare*

POC	Ref0	Ref1	Ref2	Ref3
9	0.737	0.078	0.092	0.092
10	0.418	0.395	0.143	0.045
11	0.550	0.149	0.228	0.073

Based on the RFD, we define a measurement of motion complexity denoted as C_{ME} via the following two steps.

Firstly, in terms of the temporal correlation with the previous coded frame and its RFD, C_{ME} of the current coding frame is described as

$$C_{ME} = \sum_{i=0}^{k} \alpha_i \times p_i = \sum_{i=0}^{k} \alpha_i \times \frac{N_i}{N} \quad (5)$$

where p_i denotes the probability of i-th reference frame selected as the final reference frame. N_i (i=1,2…,k) denotes the count number of the i-th reference frame referenced by the encoded PUs of the previous frame, N is the sum of N_i. k indicates the number of reference frame. α_i (i=1,2,…,k) refers to the relevant weight of different reference frame, which is correlated with the temporal distance of the reference frames and are set as 1,2,3,4 in our experiments.

Secondly, considering the spatial correlation among the adjacent PUs in the same frame, C_{ME} is further modified as

$$C_{ME} = \alpha \times C_T + (1-\alpha) \times C_S \quad (6)$$

$$\alpha = C_T / (C_T + C_S) \quad (7)$$

where C_T is the C_{ME} of the previous coded frame and C_S is the C_{ME} of coded PUs in the current frame based on (5). α is a variable depending on C_T and C_S of the current PU.

With (5) and (6), C_{ME} is calculated for the test sequences as in Table 3. It can be seen that for the high motion sequences, e.g. *BQSquare*, the value of C_{ME} is always larger compared to the low motion sequences. So the proposed C_{ME} reflects the motion complexity efficiently.

Table 3. C_{ME} of different frames for different sequences

Sequence	Resolution	Frame_6	Frame_20	Frame_140
Racehorse	416x240	1.5777	1.2311	1.4222
BQsquare	416x240	3.1832	2.7984	3.2012
PartyScene	832x480	2.6432	2.4194	2.0765
Kimono	1920x1080	1.2435	1.0850	2.2500

By incorporating C_{ME} into ME, a fast adaptive algorithm of multi reference frame ME is proposed for HEVC. It can significantly reduce the coding complexity with negligible quality loss which will be detailed in the next subsection.

2.2 Proposed Fast Multi Reference Frame Motion Estimation Algorithm

The number of reference frame in the RFS has great influence on the computational complexity of ME. In this section, based on the proposed C_{ME}, we present our proposed algorithm to adaptively adjust the number of reference frame in order to reduce the complexity of ME.

2.2.1 Determination of Initial RFS

For most PUs, it is unnecessary to carry a full search for ME in all reference frames. In our proposed algorithm, an initial RFS is provided based on C_{ME} to avoid implementing ME in all the reference frames. After the C_{ME} is computed based on (5) and (6), then the reference frame with reference index within C_{ME} consists of the initial RFS as shown in Table 4. Table 5 shows the proportion of the PUs in a frame to select the reference frame in the initial RFS as the final reference frame. It can be observed almost all PUs finish ME process in the initial RFS.

Table 4. The mapping from C_{ME} to initial RFS

C_{ME}	[1,2)	[2,3)	[3,4]
Inial RFS	0 1	0 1 2	0 1 2 3

Table 5. C_{ME} of different frames

Sequence	Resolution	POC	C_{ME}	Inial RFS	Proportion
Racehorse	416x240	6	1.5777	0 1	0.974
BQsquare	416x240	6	2.7984	0 1 2	0.955
PartyScene	832x480	6	2.6432	0 1 2	0.985
Kimono	1920x1080	6	1.2435	0 1	0.975

2.2.2 Adaptive Shrinkage of RFS

From the RFD as in table 1 and table 2, it can be also observed most PUs select the most suitable matching unit in the nearest reference frame. So it is of benefit to further shrink the RFS.

As illustrated in (2), (3) and (4), the optimal reference frame selection has great correlation with the distortion and coded bits for the reference frame index and MVD. In this subsection, two variables related to the distortion and MVD are adopted to decide whether to shrink the RFS and thus terminate the ME process ahead of time.

Firstly, Due to the different size of PU, we define the average distortion per pixel of the PU, D_{pixel}, as shown in (8).

$$D_{pixel} = D_1 / S_{Pu} \tag{8}$$

where D_1 refers to the distortion of the PU in the first reference frame and S_{Pu} indicates the size of the PU.

Secondly, since the bits for the reference frame index are relatively constant, another variable, MVD defined as (10), is utilized to facilitate to the early searching termination. MVD can reflect the motion complexity and represent the motion smoothness between the current PU and its adjacent PUs. If MVD is much smaller, it means the current PU has lower motion activity and it also implies it is much more probable that the nearer reference frame would be selected as the final reference frame. So MVD is utilized to further assist the early termination of ME.

$$|MVD| = |MVD\_X| + |MVD\_Y| \tag{9}$$

where MVD_X and MVD_Y are the horizontal and vertical ordinates of MVD.

If the following two inequalities as in (10) are both satisfied, then ME can be terminated.

$$\begin{cases} D_{pixel} < D_0 \\ |MVD| \le MVD_0 \end{cases} \tag{10}$$

where MVD_0 and D_0 are two thresholds.

Based on the information theory [11], the relation between the bit rate, R, and distortion, D, can be denoted as,

$$R(D) = \ln(1/\lambda D) \tag{11}$$

A Taylor expansion of the above equation yields that

$$R(D) \approx \alpha / D \tag{12}$$

In [12], the R-Q model is closely modeled as

$$R \approx (\beta \cdot D)/QP \tag{13}$$

Combine the (12) and (13), the relation between the distortion and QP can be approximately represented as (14). It can be observed that the distortion is highly correlated with QP and increases with QP increasing.

$$D = c \times \sqrt{QP} \tag{14}$$

So the threshold D_0 is modeled adaptively with QP as

$$D_0 = c \times \sqrt{QP} \tag{15}$$

where c is a constant and is set as 0.6 in our experiments.

2.2.3 Expansion Measures for RFS

With the proposed above C_{ME} and RFS, it can be expected to significantly reduce the computational complexity, but a disadvantage of C_{ME} is that it converges to 1 with the

number increasing of coding frames. So an expansion method is added by expanding the reference range from the within RFS to the outside of the RFS. Then above local optimization problem can be avoided.

According to our observation, the distortion ratio of PU in the two adjacent reference frames has evident impact on the final determination of the reference frame. Fig.2 shows the statistical results of D_{ratio} between the second reference frame and the third while the final selected reference frame is the forth. It can be observed that the D_{ratio} of most PUs lies in a small domain and then the probability of expanding the RFS is low when the ratio is within a certain threshold. So we define distortion ratio as follow.

$$D_{ratio} = D_i / D_{i+1} \tag{16}$$

where D_i and D_{i+1} denote the distortion of the PU in the i-th and (i+1)-th reference frame, respectively. If D_{ratio} is beyond a predefined ratio θ which is set as 1.2 in our paper, the RFS should be expanded to include the next reference frame.

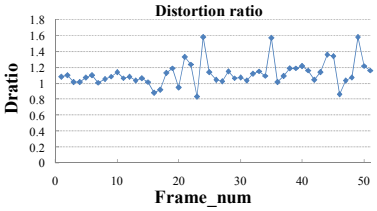

Fig.2. Variance of distortion ratio between adjacent reference frames

2.2.4 Summary of the Proposed Algorithm

For a PU, the proposed algorithm can be summarized as follows.

Summary of the proposed algorithm

Begin

Given the initial parameters for the proposed scheme;

Begin

For each PU in the frame

1) Calculate C_{ME} based on (5) and (6) ;
2) Determine the initial RFS based on mapping scheme as Table 4;
3) Perform the ME on the first reference frame in RFS;
4) Decide whether to shrink the RFS based on (10);
5) If yes, then terminate ME. Else continue ME on other frames;
6) Calculate the distortion ratio based on (16);
7) Determine whether to add a new reference frame into RFS;
8) If yes, add another reference frame. Else terminate ME.

End

End

3 EXPERIMENTAL RESULTS

Experiments were conducted on the HEVC platform HM8.0 for Low_delay_B and Low_delay_P. The maximum number of reference frame is set to 4. When comparing the coding performance difference, we utilize the popular method proposed in [13] to calculate the difference between two R-D curves. The complexity reduction Δt is calculated as

$$\Delta t = \frac{T_{anchor} - T_{pro}}{T_{anchor}} \times 100\% \tag{17}$$

where T_{anchor} and T_{pro} denote the encoding time of original HM anchor and our proposed fast algorithm respectively.

The experimental results tabulated in Table 6 and Table 7 shows the time saving and R-D performance of the proposed algorithm. Compared with the original HM anchor, about 30% of the total encoding time on average is saved. And it can be seen that the BD_PSNR loss of the proposed algorithm is only 0.016dB which is ignorable.

4. CONCLUSIONS

In this paper, based on the proposed motion complexity C_{ME}, we propose a fast multi reference frame motion estimation scheme. An initial reference picture set is created to reduce the number of reference frame for motion estimation. Then an adaptive early termination scheme is provided to save more encoding time. Furthermore, a RPS expansion measure is proposed to avoid the local optimization and quality loss. Experimental results show the proposed algorithm can save up to 30% of total coding time on average while keeping the quality nearly the same as full search scheme.

Table 6. Performance comparison for Low_delay_P

Sequence	Low_delay_P			
	Resolution	BD_Rate	BD_PSNR	Δt
Racehorse	416x240	0.5%	-0.025	32.9%
Basketballpass	416x240	0.4%	-0.022	26.5%
Racehorse	832x480	0.3%	-0.010	22.9%
Vidyo1	1080x720	0.6%	-0.021	34.2%
Vidyo4	1080x720	0.6%	-0.020	34.5%
Kimono	1920x1080	0.3%	-0.008	29.7%
ParkScene	1920x1080	0.4%	-0.013	20.3%
average		0.36%	-0.014	28.7%

Table 7. Performance comparison for Low_delay_B

Sequence	Low_delay_B			
	Resolution	BD_Rate	BD_PSNR	Δt
Racehorse	416x240	0.5%	-0.024	28.5%
Basketballpass	416x240	0.5%	-0.023	29.2%
Racehorse	832x480	0.5%	-0.022	23.4%
Vidyo1	1080x720	0.9%	-0.030	43.7%
Vidyo4	1080x720	0.9%	-0.029	43.6%
Kimono	1920x1080	0.3%	-0.008	27.0%
ParkScene	1920x1080	0.3%	-0.011	21.8%
average		0.54%	-0.019	31.0%

ACKNOWLEDGEMENT

This work was supported in part by Major State Basic Research Development Program of China (973 Program, 2009CB320903), National Science Foundation (61121002, 61103088), National High-tech R&D Program of China (863 Program, SS2012AA010805) and National Sci-Tech Support Plan (2011BAH08B01).

REFERENCES

[1] ITU-T VCEG and ISO/IEC MPEG, "Terms of Reference of the Joint Collaborative Team on Video Coding Standard Development," Document VCEG-AM90 of VCEG and N11112 of MPEG, Jan. 2010.

[2] I. K. Kim, J. Min, T. Lee, W. J. Han, and J. Park, "Block Partitioning Structure in the HEVC Standard", *IEEE Transactions on Circuits and Systems for Video Technology*, in publication.

[3] Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification, ITU-T Rec. H.264 and ISO/IEC 14 496-10 AVC, Joint Video Team, Mar. 2003.

[4] S. Ho, S. Kim, M. Sunwoo, "Fast multiple reference frame selection methods for H.264/AVC," *SiPS*, 8-10 Oct. 2008.Y. Su, Ming-Ting Sun, "Fast Multiple Reference Frame Motion Estimation for H264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 3, pp. 447–452, Oct. 2006

[5] Y. Su, Ming-Ting Sun, "Fast Multiple Reference Frame Motion Estimation for H264/AVC," IEEE Trans. Circuits Syst. Video Technol., vol. 16, no. 3, pp. 447–452, Oct. 2006.

[6] Q. Tang, P. Fu, H. Chen, F. Guo, "A Fast Motion Estimation Algorithm Based on the Adaptive Reference Frame and the Spatial and Temporal Correlations for H.264," *IECON*, Nov. 2007, pp.2399-2402, Taibei.

[7] K. Lee, Gwanggil Jeon, and Jechang Jeong, "Fast Reference Frame Selection Algorithm for H.264/AVC," *IEEE Consumer Electronics Society*, vol. 55, no. 3, pp.773-779, May 2009.

[8] Y. Huang, B. Hsieh, S. Chien, S. Ma, L. Chen, Analysis and Complexity Reduction of Multiple Reference Frames Motion Estimation in H.264/AVC. *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 4, pp.507-522, April 2006.

[9] C. Chiang, S. Lai, "Fast Multi-Reference Frame Motion Estimation via Downhill Simplex Search," *ICME* 2006, pp.121-124.

[10] B Li, J. Xu, H. Li, F. Wu. "Optimized Reference Frame Selection for Video Coding by Cloud," *MMSP*, Hangzhou, China, Nov. 2011

[11]T. G. Cover and J. A. Thomas, Elements of Information Theory, John Wiley and Sons, Inc., 1991.

[12]Y Liu, Z. G. Li, Y. C. Soh, "A Novel Rate Control Scheme for Low Delay Video Communication of H264/AVC Standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 1, pp.67-78, Jan. 2007.

[13] G. Bjontegaard, "Calculation of average PSNR difference between RD-curves," in Proc. ITU-T Q.6/SG16 VCEG 13th Meeting, Austin, TX, Apr. 2001, document VCEG-M33