VLSI FRIENDLY EDGE GRADIENT DETECTION BASED MULTIPLE REFERENCE FRAMES MOTION ESTIMATION OPTIMIZATION FOR H.264/AVC

Zhenyu Liu, Yiqing Huang, Yang Song, Satoshi Goto and Takeshi Ikenaga

The Graduate School of IPS, Waseda University
Lab N355, 2-7, Hibikino, Wakamatsu 808-0135, Kitakyushu, JP
phone&fax: +(81)93-692-5319, email: liuzhenyu@aoni.waseda.jp, http://liuzhenyu73.googlepages.com/

ABSTRACT

In H.264/AVC standard, motion estimation can be processed on multiple reference frames (MRF) to improve the video coding performance. The computation is also increased in proportion to the reference frame number. Many software oriented fast multiple reference frames motion estimation (MRF-ME) algorithms have been proposed. However, for the VLSI real-time encoder, the heavy computation of fractional motion estimation (FME) makes the integer motion estimation (IME) and FME must be scheduled in two macro block (MB) pipeline stages, which makes many fast MRF-ME algorithms inefficient. In this paper, one edge gradient detection based algorithm is provided to reduce the computation of MRF-ME. The image being rich of texture and sharp edges contains much high frequency signal and this nature makes MRF-ME essential. Through analyzing the edges' gradient, we just perform MRF-ME on those blocks with sharp edges, so the redundant ME computation can be efficiently reduced. Experimental results show that average 26.43% computation can be saved by our approach with the similar coding quality as the reference software. This proposed algorithm is friendly to hardwired encoder implementation. Moreover, the provided fast algorithms can be combined with other fast ME algorithms to further improve the performance.

1. INTRODUCTION

The superior performance of the latest international video coding standard, H.264/AVC, mainly comes from the new techniques, which include 1/4-pixel accurate variable block size motion estimation (VBSME) with multiple reference frames (MRF), intra prediction (IP), context-based adaptive variable length entropy coding (EC) and in-loop deblocking (DB), etc.

According to the analysis in [1], 89.2% computation power is consumed by ME part. MRF is the main issue that leads to the huge computation complexity. The required computation is in direct ratio to the reference frame number. However, the performance of MRF algorithm mainly depends on the nature of sequences. For some test sequences, MRF-ME greatly improves the coding quality, but this is not always. In other words, most computation consumed by MRF-ME is wasted. In order to reduce the redundant computation of MRF-ME, many interesting and powerful algorithms have been provided [1][2][3]. One excellent work is provided in [1], which provides four criterions to early terminate the motion search on MRFs. These algorithms efficiently reduce 30%-80% redundant computation in the software. Another kind of promising scheme is reducing the search areas on MRFs depending on the MVs' strong correlations in consequent pictures [2][3].

This paper was supported by CREST, JST.

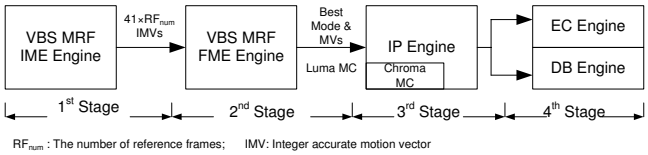


Figure 1: Block diagram of 4-stage hardwired H.264 encoder

Most of these algorithms are software oriented and the restrictions of VLSI implementation [4] are not fully considered. The main adverse impact of hardwired encoder is its MB pipelining architecture. A brief overview of the MB pipelining dataflow is introduced first. As shown in Fig. 1, in the first MB stage, IME engine processes on all reference frames. The integer motion vectors (MV) of 41 blocks in MB on all reference frames are achieved and dispatched to the second FME stage. Through 1/4-pixel accurate ME and precise RD-cost evaluation, FME engine finds the best candidates and the corresponding reference frames and decides the best inter prediction mode. The post inter/intra mode decision IP and chroma MC are implemented at the third stage. EC and DB are processed in parallel in the fourth stage.

Based on the hardwired encoder structure, the performance of those provided fast MRF-ME algorithms are reviewed and analyzed. All the early termination criterions of [1] must be used in the second FME stage. At this moment, the macro block in FME stage has already finished its MRF-IME search. That means the computation load of IME, which is the most computation intensive part, can not be saved. For the hardwired encoder implementation, motion vector composition based algorithms [2][3] have following drawbacks: (1)This algorithm consumes much hardware cost because of its MV composition. For example, in reference [2], 4×4 block based MVs on each frame must be kept. If the frame size is 720×480 with 128×128 search range and 5 reference frames, totally 1.65Mb memories are required. For the accuracy of MV composition, the multiplication, which increases the hardware cost, is also applied. (2)This algorithm just simplifies the computation of IME and does not reduce the burden of FME engine.

Through the mathematical analysis [5][6][7], aliasing caused by the picture high frequency is the main issue that deteriorates the prediction efficiency. Sub-pel interpolation and MRF techniques adopted by H.264 mainly aim to compensate the prediction error caused by aliasing. In this paper, based on the 2-D gradient measurement with Sobel operator, we can accurately analyze the frequency nature of image

and dynamically adjust the reference frame number. Consequently, the redundant computation is efficiently eliminated. Moreover, our algorithm is compatible with other fast block matching, fast MRF-ME and fast inter/intra mode decision algorithms [8][9]. Combined with these fast algorithms, our algorithm can achieve better performance.

The rest of this paper is organized as follows. In section 2, the impact of aliasing to prediction error, which is caused by the high frequency signal, is first briefly introduced. And then, the spatial edge gradient effect to its frequency spectrum is mathematically analyzed. Based on the conclusion of section 2, the fast MRF-ME algorithm is proposed in section 3. Section 4 shows some experimental results to demonstrate our algorithm. Conclusions are drawn in section 5.

2. IMPACT OF EDGE GRADIENT TO MRF-ME ALGORITHM

By mathematical analysis, aliasing is the main component that deteriorates the prediction efficiency [5, 6, 7]. H.264/AVC adopts 1/4-pel accurate interpolation and MRF techniques to compensate the prediction error caused by aliasing. In this section, we first give a brief overview of the impact of aliasing and this analysis reveals that the high frequency spectrum is the main issue. And then the edge gradient contribution to high frequency spectrum is mathematically analyzed.

In order to simplify the mathematical description, the analysis is restrict to one spatial dimension signal [7]. $l_t(x)$ and $l_{t-1}(x)$ denote the spatial-continuous signals at time instance t and $t-1$. $l_t(x)$ is a displaced version of $l_{t-1}(x)$ and the distance is d_x, which can be expressed as $l_t(x) = l_{t-1}(x-d_x)$. Their frequency domain signals are denoted as $L_t(\omega)$ and $L_{t-1}(\omega)$. These continuous image signals are sampled by the sensor array before digital processing. The interval of the spatial samplers is denoted as s_x. The displacement error can be expressed as $\Delta_x = d_x - \text{round}(d_x/s_x)\cdot s_x$. Aliasing dose not exist if Nyquist-Shannon sampling precondition, i.e., $L_{t-1}(\omega)=0$ for $|\omega| \geq \omega_s/2$, where ω_s is the sampling frequency, is satisfied. However, because no spatial-limited signals can be band limited and the low-pass filter of the sampling system is not ideal, the precondition of Nyquist-Shannon sampling theorem cannot be fulfilled.

According to [7], with the normalized sampling frequency, i.e., $\omega_s = 2\pi$, the magnitude of prediction error signal caused by aliasing can be described as (1)

$$|E_t(\omega)| = 2\cdot|A_{t-1}(\omega)|\cdot|\sin(\Delta_x\cdot\pi)| \quad (1)$$

where $A_{t-1}(\omega)=L_{t-1}(\omega+2\pi)+L_{t-1}(\omega-2\pi)$.

According to (1), two important conclusions can be drawn:

1. Because of the item $|A_{t-1}(\omega)|$, aliasing is cause by the high frequency signals in $L_{t-1}(\omega)$, where $|\omega|\geq\pi$.
2. According to the item $|\sin(\Delta_x\cdot\pi)|$, the impact of aliasing vanishes at full pixel displacements and is maximum at half pixel displacements.

Conclusion 1 states that the image being rich of high frequency signals is prone to be affected by the aliasing problem. Conclusion 2 explains the necessity of MRF-ME during prediction processing: If the displacement error $\Delta_{x,t-1}$ between the current $s_t(x_n)$ and the previous $s_{t-1}(x_n)$ image is sub-pel and the more previous image $s_{t-k}(x_n)$ has the full-pel displacement, i.e., $\Delta_{x,t-k}=0$, $s_{t-k}(x_n)$ is preferred to be chose as the reference because its aliasing problem dose not exist any more.

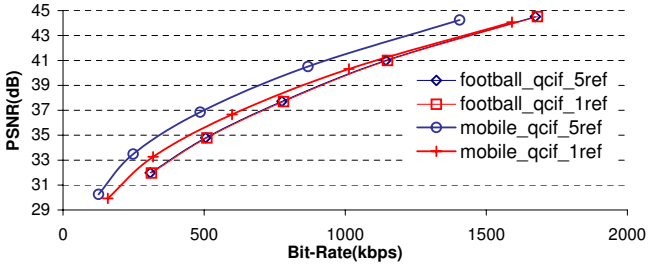


Figure 2: Rate-distortion curves of 'mobile_qcif' and 'football_qcif' with 1 and 5 reference frames

Now, we can explain why 'Mobile' sequence is so sensitive to MRF-ME. The main reason is that many textures are contained in this video sequence. Sharp edges in the spatial domain generate rich high frequency signals in the frequency domain. Even though the 2-D Wiener filter interpolation algorithm in H.264/AVC can alleviate the error of aliasing, its effect can not compare with the reference image with the full-pel displacement.

Seven reasons of MRF-ME achieving better prediction results, such as 'uncovered background', 'Alternating camera angles' and 'Camera shaking' etc, are listed in [2]. In fact, through our experiments, aliasing is the main issue making MRF-ME essential. For example, the rate-distortion curves of 'football_qcif' and 'mobile_qcif' standard sequences with 1 and 5 reference frames are shown in Fig.2. Even though 'football_qcif' has large and complex motions, MRF-ME can not achieve noticeable coding gain. However for 'mobile_qcif', because of its sophisticated texture, the peak signal to noise ratio (PSNR) differences between searching five reference frames and searching only one reference frame are about 1.4-1.5dB.

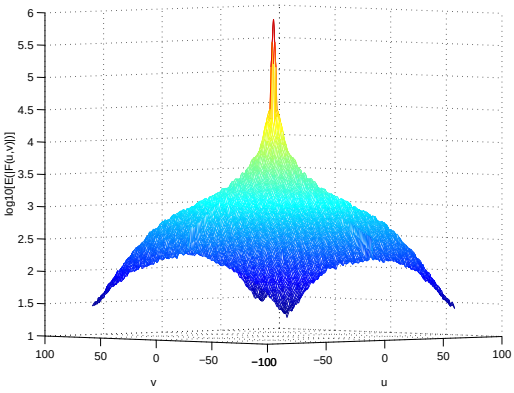
The 2-D Fourier spectrum amplitude of these two sequences are shown in Fig.3. It is obvious that the high frequency signal of 'mobile_qcif' is much more abundant than its counterpart 'football_qcif'. The intuitive approach is dynamically adjusting the reference frame number depending on the image's Fourier spectrum analysis. However, this method has two demerits: (1) Fourier transform is not a powerful tool for short length signals. It can not efficiently reveal the local frequency nature of a signal. (2) The computation complexity of 2-D Fourier transform is very high.

In fact, we can derive a signal's frequency spectrum features through analyzing its gradient amplitude. In theory, the signal frequency spectrum spreads linearly with its gradient amplitude.

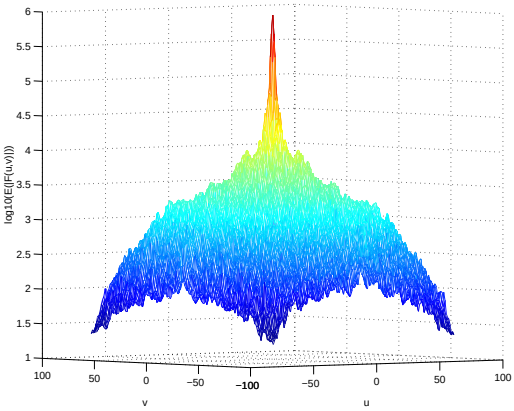
Proof: Suppose $l(x)$ is a Lebesgue integrable function, $L(\omega)$ is denoted as its Fourier transform. In particular, if we squeeze a function in x domain a times, it can be expressed as $l(ax)$. The derivative of $l(ax)$ is

$$\frac{\partial l(ax)}{\partial x} = a l'(ax) \quad (2)$$

(2) demonstrates that if $l(ax)$ is concentrated around 0, its



(a) football qcif frequency



(b) mobile qcif frequency

Figure 3: Frequency spectrum analysis of 'mobile_qcif' and 'football_qcif' sequence (128 frames and *hamming window* adopted)

gradient amplitude is increased linearly to a. According to the scaling property of Fourier transform, if $l(x) \Longleftrightarrow L(\omega)$ denotes that $l(x)$ and $L(\omega)$ are a Fourier transform pair, we can derive

$$l(ax) \Longleftrightarrow \frac{1}{|a|} L(\frac{\omega}{a}) \tag{3}$$

Namely, the frequency spectrum span of $l(ax)$ also expends in proportional to a. In summary, if a image contains a lot of sharp edges, it must be rich of high frequency spectrum. From the analysis of (1), we can see that, in this case, MRF-ME is a essential and efficient approach to reduce the prediction error.

3. EDGE GRADIENT DETECTION BASED FAST MRF-ME ALGORITHM

Sobel operator is widely used to perform a 2-D spatial gradient measurement on an image and also emphasizes regions of high spatial frequency that correspond to edges. In this paper, we apply it as the edge detector. Another reason of adopting Sobel operator is that this operator is also applied in the fast inter/intra mode decision algorithms [8][9][10]. So, applying the same operator makes it easy to combine our algorithm with these fast mode decision schemes. Sobel edge detector uses a pair of 3×3 convolution masks, G_x and G_y. G_x estimates the gradient in the x-direction and the other, G_y, estimates the gradient in the y-direction. G_x and G_y can be written as

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} G_y = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} \tag{4}$$

P denotes the input picture and the convolution between P and G_x derives the gradient in horizontal direction. At the coordinate $[i, j]$ of luminance picture, the gradient in $\vec{x}$ is defined as

$$\begin{aligned} dx_{i,j} = &-p_{i-1,j-1} - 2p_{i-1,j} - p_{i-1,j+1} + \\ &p_{i+1,j-1} + 2p_{i+1,j} + p_{i+1,j+1} \end{aligned} \tag{5}$$

In the same way, we can achieve the gradient in $\vec{y}$

$$\begin{aligned} dy_{i,j} = &p_{i-1,j-1} + 2p_{i,j-1} + p_{i+1,j-1} - \\ &p_{i-1,j+1} - 2p_{i,j+1} - p_{i+1,j+1} \end{aligned} \tag{6}$$

A $M \times N$ block, where M and N are 16 or 8, is defined as a homogeneous one, if all edge gradient amplitudes, $|dx_{i,j}|$ and $|dy_{i,j}|$, in this block are less than a predefined threshold Thd_H. This decision procedure can be expressed as

$$\begin{cases} |dx_{i,j}| < Thd_H \text{ and } |dy_{i,j}| < Thd_H & homogeneous \\ \quad {\scriptstyle \forall i,j \in positions\ in\ the\ block} & \\ otherwise & nonhomogeneous \end{cases} \tag{7}$$

How to define the value of Thd_H is another critical issue because this threshold directly affects the performance of our algorithm. It is assumed that the prediction error e is a stationary jointly Gaussian source of zero mean and variance σ^2. The distortion of quantization [11] is label as D, which can be approximated as

$$D = \frac{QP^2}{3} \tag{8}$$

In order to simplify the analysis of the variance of prediction error e, we still focus on one spatial dimension signal. From Fig. 4, it can be observed that e can be approximated as

$$e \approx \Delta_x \cdot {l_{t-1}}'(x_n) \tag{9}$$

Where, the displacement error Δ_x is a random variable with zero mean, $\Delta_x \in (-s_x/2, s_x/2)$, s_x is the spatial sampling distance. In consequent, σ^2 should be linear to $({l_{t-1}}'(x_n))^2$

$$\sigma^2 \approx a \cdot ({l_{t-1}}'(x_n))^2 \tag{10}$$

According to the analysis of [12], the rate distortion function of a memoryless Gaussian source of variance σ^2 with respect to the squared-criterion is

$$R(D) = \begin{cases} 0.5 \log \frac{\sigma^2}{D} & 0 \le D \le \sigma^2 \\ 0 & D \ge \sigma^2 \end{cases} \tag{11}$$

It should be noticed that when $D \ge \sigma^2$, $R(D)$ becomes zero. A more intuitive explanation is that, when the amplitude of residues is less than the threshold of quantization, these

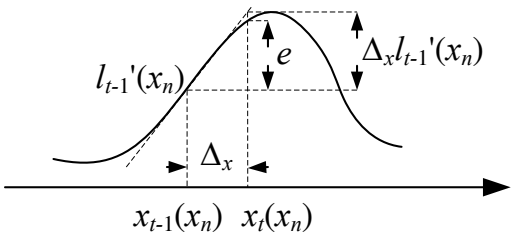


Figure 4: 1-D prediction error analysis

residues vanish and no rate cost is required. From (8), (10)and (11), Thd_H should be increased linearly with QP. With exhaustive experiments on various video sequences, Thd_H is set as

$$Thd_H = 5QP \tag{12}$$

The pseudo code of fast inter ME algorithm of encoding one macro block is shown in the following, as Fig. 5. Our changes to the reference software have been highlighted with the bold italic font.

```
Edge_Gradient_Detection ( Current Macro-Block )

Mode= P16X16 or P16X8 or P8X16:
Loop 16x16 or 16x8 or 8x16 Blocks
    Loop Reference Frames ( frame number=1 for homogenous;
                            frame number=5 otherwise)
        Integer and Fractional Motion Estimation {
            Compute Cost of 16x16 or 16x8 or 8x16 Block;
        }
    End Loop
    Accumulate Cost of 16x16 or 16x8 or 8x16 Blocks as Cost of Macroblock;
End Loop

Mode=P8X8:
Loop 8x8 Blocks
    Loop Sub-Partition Modes
        Loop Reference Frames ( frame number=1 for homogenous;
                                frame number=5 otherwise)
            Loop Sub-Blocks
                Integer and Fractional Motion Estimation {
                    Compute Cost of Sub-Blocks;
                }
            End Loop
        End Loop
        Accumulate Cost of Sub-Blocks as Cost of 8x8-Block;
    End Loop
    Accumulate Cost of 8x8-Blocks as Cost of Macroblock;
End Loop
```

Figure 5: Pseudo coding of our fast MRF-ME algorithm

This algorithm can be easily integrated to the hardwired encoder engine as illustrated in Fig. 6. During loading in the current MB pixels, the edge gradient detection is performed on the current MB. The edge detection results can be applied to guide the reference frame number of the IME processing of current MB. When the current MB is transferred to the FME stage to do the refine search, its edge information is also delivered from IME stage to eliminate the redundant computation of FME stage.

4. EXPERIMENTAL RESULTS

The proposed fast MRF-ME algorithm is embedded into JM11.0 provided by JVT. The simulation conditions are shown bellow.

1. MV search range is ± 16 for QCIF and ± 32 for CIF

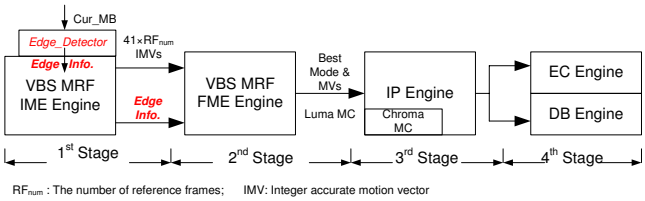


Figure 6: VLSI architecture with fast MRF-ME algorithm

Table 1: Coding performance comparisons

Sequence	ME_r(%)	BDPSNR(dB)	BDBR(%)
Foreman_qcif	-20.51	-0.038	0.82
Mobile_qcif	-07.71	-0.023	0.43
News_qcif	-17.26	-0.017	0.28
Football_qcif	-30.06	-0.020	0.28
Foreman_cif	-43.59	-0.098	2.27
Mobile_cif	-07.81	-0.012	0.21
Stefan_cif	-14.55	-0.007	0.14
Football_cif	-45.56	-0.025	0.40
Container_cif	-50.85	0.002	-0.06

2. RD optimization is enabled
3. Reference frame number is 5
4. MV is 1/4 pel accurate
5. GOP is IPPP
6. Encoded 200 frames

Four QCIF and five CIF format standard sequences are experimented with quantization parameters 20, 24, 28, 32 and 36. In order to evaluate the ME speedup of our algorithm, ME speedup ME_r is defined as

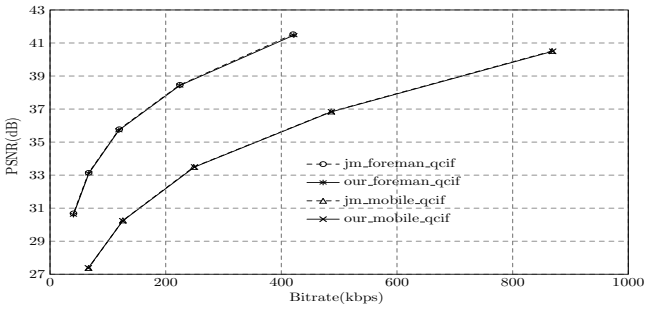
$$ME_r = \frac{T_f - T_{jm}}{T_{jm}} \times 100\% \tag{13}$$

where, T_f denotes the the ME time of our algorithm and T_{jm} is the time taken by JM11.0.

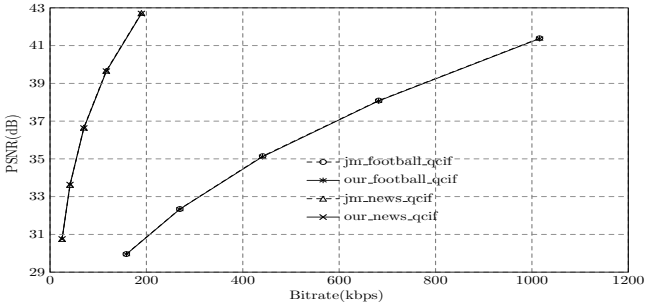
The rate-distortion curve comparisons are shown in Fig. 7. As our algorithm provides the almost the same coding efficiency, it is hard to distinguish our algorithm's curves from the reference ones in most cases. At high bit-rate, the degradation of PSNR is slightly noticeable for "foreman_qcif" and "foreman_cif" and is less than 0.1dB.

The experimental results are shown in Table 1. BDBR (Bjonteggard Delta BitRate) and BDPSNR (Bjonteggard Delta PSNR) [13], which are respectively average difference of bit-rate and PSNR between two methods, are used and they are derived from the simulation results when QP = 20; 24; 28; 32. It is observed that our algorithm has reduced the ME time by 26.43% on average. The speedup performance of our method depends on the nature of the video sequence. Sharp edges of 'Mobile_qcif/cif' make them very sensitive to MRF-ME algorithm. For these sequences, just 7.71%-7.81% ME time can be saved by our approach. For those sequences not depending on MRF-ME algorithm, such as "Football_cif" and "Container_cif", 45.56% and 50.85% ME time can be saved, respectively.

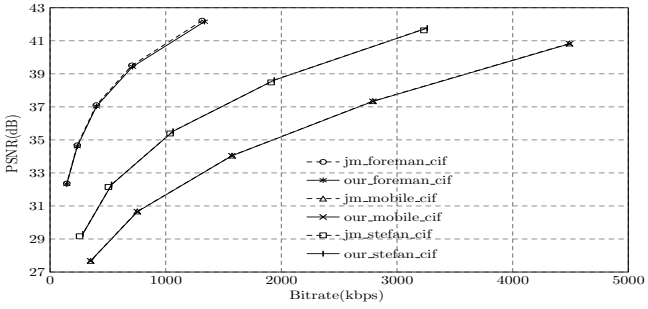
Our fast reference frame selection scheme can save a lot of unnecessary ME operations while maintaining the video quality almost identical to full search scheme. It is different from the conventional fast block matching algorithms,



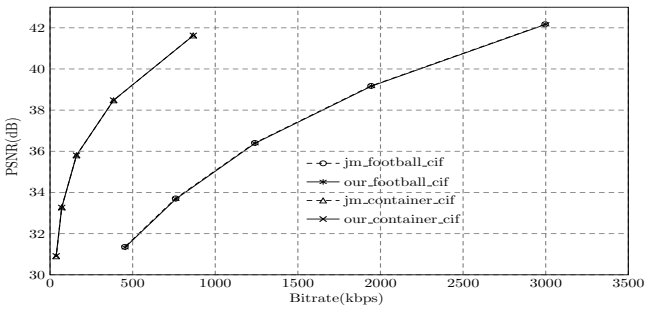
(a) foreman_qcif and mobile_qcif RD curve comparisons



(b) football_qcif and news_qcif RD curve comparisons



(c) foreman_cif, mobile_cif and stefan_cif RD curve comparisons



(d) football_cif and container_cif RD curve comparisons

Figure 7: RD curve comparisons for QCIF and CIF sequences

such as four-step search [14], diamond search [15] and successive elimination [16]. These algorithms focus on reducing the search candidates and simplifying the matching criterion. In fact, the methods of [2, 3] also can be categorized as fast block matching algorithms. Our method is also different from the methods provided in [1], which early terminates the ME processing at FME stage. The soul of our method is reducing the redundant computation through the spatial analyzing of raw video sequences. One interesting matter is that our algorithm and other fast algorithms are orthogonal. For example, before ME processing, our algorithm decides whether to use multiple reference frame and then conventional fast search methods can be adopted during the search procedure.

5. CONCLUSIONS

Fully considering the limitations of MB-pipelined hardware architectures, we propose the VLSI friendly fast algorithms for MRF-ME in H.264/AVC: By analyzing the edge gradient of current picture, we can estimate its frequency spectrum nature. For those blocks being rich of sharp edges, multiple reference frames are adopted during their ME processing, otherwise, just the previous reference frame is needed. In this paper, the mathematical analysis shows that the threshold for edge detection should be proportional to the quantization parameter. Through experiments, it is defined as $5QP$. Experimental results show that average 26.43% computation can be saved with almost the same coding quality as the reference software. Moreover, the provided scheme can be combined with other fast ME algorithms to further improve the performance.

REFERENCES

[1] Y. Huang, B. Hsieh, S. Chien, S. Ma, and L. Chen, "Analysis and complexity reduction of multiple reference frames motion estimation in h.264/avc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.16, no.4, pp.507–522, April 2006.

[2] Y. Su and M. Sun, "Fast multiple reference frame motion estimation for h.264/avc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.16, no.3, pp.447–452, March 2006.

[3] M. Chen, Y. Chiang, H. Li, and M. Chi, "Efficient multi-frame motion estimation algorithms for mpeg-4 avc/jvt/h.264," *Proceedings of the 2004 International Symposium on Circuits and Systems*, pp.737–740, May 2004.

[4] T. Chen, S. Chien, Y. Huang, C. Tsai, C. Chen, T. Chen, and L. Chen, "Analysis and architecture design of an hdtv720p 30 frames/s h.264/avc encoder," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.16, no.6, pp.673–688, June 2006.

[5] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE Journal on Selected Area in Communications*, vol.SAC-5, no.7, pp.1140–1154, August 1987.

[6] B. Girod, "Efficiency analysis of multihypothesis motion compensated prediction for video coding," *IEEE Transactions on Image Processing*, vol.9, no.2, pp.173–183, February 2000.

[7] T. Wedi and H.G. Musmann, "Motion- and aliasing-compensated prediction for hybrid video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.13, no.7, pp.577–586, July 2003.

[8] F. Pan, X. Lin, S. Rahardja, K.P. Lim, Z.G. Li, D. Wu and S. Wu, "Fast mode decision algorithm for intraprediction in h.264/avc video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.15, no.6, pp.813–821, July 2005.

[9] D. Wu, F. Pan, K.P. Lim, S. Wu, Z.G. Li, X. Lin, S. Rahardja and C.C. Ko, "Fast intermode decision in h.264/avc video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.15, no.6, pp.953–958, July 2005.

[10] G. Sullivan, T. Wiegand, and K.P. Lim, "Joint model reference encoding methods and decoding concealment methods," September 2003. JVT-I049.

[11] H. Gish and J.N. Pierce, "Asymptotically efficient quantizing," *IEEE Transactions on Information Theory*, vol.14, no.5, pp.676–683, September 1968.

[12] T. Berger, *Rate Distortion Theory*, Prentice Hall, 1971.

[13] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," April 2001. VCEG-M33.

[14] L.M. Po and W.C. Ma, "A novel four-step search algorithm for fast block motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.6, no.3, pp.313–317, June 1996.

[15] J.Y. Tham, S. Ranganath, M. Ranganath, and A.A. Kassim, "A novel unrestricted center-biased diamond search algorithm for block motion estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol.8, no.4, pp.369–377, August 1998.

[16] W. Li and E. Salari, "Successive elimination algorithm for motion estimation," *IEEE Transactions on Image Processing*, vol.4, no.1, pp.105–107, January 1995.