ANALYSIS AND REDUCTION OF REFERENCE FRAMES FOR MOTION ESTIMATION IN MPEG-4 AVC/JVT/H.264

Yu-Wen Huang[1,*], *Bing-Yu Hsieh*[1], *Tu-Chih Wang*[1], *Shao-Yi Chien*[1],
Shyh-Yih Ma[2], *Chun-Fu Shen*[2], *and Liang-Gee Chen*[1]

1. DSP/IC Design Lab., Graduate Institute of Electronics Engineering and
Department of Electrical Engineering, National Taiwan University, yuwen@video.ee.ntu.edu.tw
2. Vivotek Incorporation, steve@vivotek.com

ABSTRACT

In the new video coding standard, MPEG-4 AVC/JVT/H.264, motion estimation is allowed to use multiple reference frames. The reference software adopts full search scheme, and the increased computation is in proportion to the number of searched reference frames. However, the reduction of prediction residues is highly dependent on the nature of sequences, not on the number of searched frames. In this paper, we present a method to speed up the matching process for multiple reference frames. For each macroblock, we analyze the available information after intra prediction and motion estimation from previous one frame to determine whether it is necessary to search more frames. The information we use includes selected mode, inter prediction residues, intra prediction residues, and motion vectors. Simulation results show that the proposed algorithm can save up to 90% of unnecessary frames while keeping the average miss rate of optimal frames less than 4%.

1. INTRODUCTION

Joint Video Team (JVT) gathered experts from ISO/IEC MPEG-4 Advanced Video Coding (AVC) and ITU-T H.264 to develop the latest standard. The new standard significantly outperforms previous ones in bit-rate reduction. Compared to MPEG-4 advanced simple profile, up to 50% of bit-rate reduction can be achieved. Such improvement mainly comes from the prediction part [1]. Motion estimation at quarter-pixel accuracy with variable block sizes and multiple reference frames greatly reduces prediction errors. Even if inter-frame prediction cannot find a good match, intra prediction will make it up instead of directly coding the texture.

The reference software, JM4.3 [2], adopts full search for both inter and intra prediction. Although there are seven kinds of block size (16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4) for motion compensation, the sum of absolute difference (SAD) of a 4x4-block can be reused for the SAD calculation of a larger block. Thus, variable block size motion estimation (ME) does not lead to much increase in computation. Intra prediction allows 4 modes for 16x16-blocks and 9 modes for 4x4-blocks. The computational load can be estimated as the SAD calculation of thirteen 16x16-blocks and extra operations for interpolation, which are quite small compared to inter prediction. As for the multiple reference frames ME, it contributes to the heaviest computational load. The required operations are proportional to the number of searched frames. Nevertheless, the decrease of prediction residues depends on the nature of sequences. Sometimes the prediction gain is very significant, but sometimes a lot of computation is wasted without any benefits. In this paper, we present an effective method to accelerate the multiple reference frames ME without significant loss of video quality. The rest of this paper is organized as follows. In Section 2, we will analyze the statistics of selected mode, residues, and motion vectors for multiple reference frames. In Section 3, we will describe our fast algorithm. Simulation results will be shown in Section 4. Finally, Section 5 gives a conclusion.

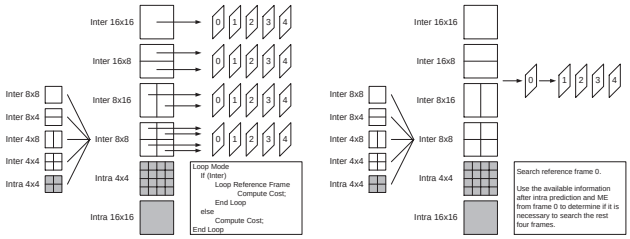

Fig. 1. Searching steps of intra/inter prediction with multiple reference frames in H.264 reference software and our method.

2. ANALYSIS

The left side of Fig. 1 shows the searching steps in H.264 reference software. The prediction of a macroblock (MB) is performed mode by mode with full search scheme. The allowed modes are inter16x16, inter16x8, inter8x16, inter8x8, intra4x4, and intra16x16. Note that the inter8x8 mode can be further partitioned into smaller blocks. Given an inter-mode, the reference software carries out the matching process reference frame by reference frame. The best mode is chosen by minimizing a Lagrangian cost function, which considers both 2-D 4x4 Hadamard transformed SAD (SATD) and number of bits required to code the side information. The right side of Fig. 1 illustrates our method. In Table 1, we can see that 80% of the optimal motion vectors (MVs) determined by the reference software belong to the nearest reference frame. Therefore, we first adopt exhaustive search for intra-modes and inter-modes from previous frame. Next, we analyze the available information including selected mode, intra prediction residues, inter prediction residues, and MVs, to determine if it is helpful to search more frames. Intuitively, the prediction gain of multiple reference frames mainly results from occluded and uncovered objects.

[*] The author thanks SiS Education Foundation for financial support.

Table 1. Statistics of Reference Frames.

Sequences	Previous Frame	Others
Coastguard	75%	25%
Container	91%	09%
Foreman	76%	24%
Hall Monitor	92%	08%
Mobile Calendar	36%	64%
Mother and Daughter	92%	08%
Silent	91%	09%
Stefan	65%	35%
Table Tennis	87%	13%
Weather	90%	10%
Average	80%	20%

CIF size, search range [-16, +16], 5 reference frames, QP=30.

Table 2. Statistics of Selected Modes.

Sequences	16x16		16x8		8x16		8x8		Intra	
Coastguard	57	92	06	81	06	78	30	39	01	08
Container	92	95	01	66	01	61	04	20	02	02
Foreman	64	88	08	68	08	66	17	35	03	06
Hall Monitor	90	98	01	53	01	61	07	44	01	14
Mobile Calendar	49	50	06	28	07	28	37	20	01	11
Mother and Daughter	89	97	03	77	03	76	04	27	01	04
Silent	83	98	03	78	03	79	10	37	01	12
Stefan	47	84	06	63	05	63	38	39	04	07
Table Tennis	76	97	04	71	04	73	13	41	03	08
Weather	87	98	01	65	02	64	09	37	01	02
Average	73	90	04	65	04	65	17	34	02	07

CIF size, search range [-16, +16], QP=30.

A|B is defined as follows.

A: % of MBs when only prediction from previous frame is allowed.

B: % of A keeping the same mode and ref. frame after 5 frames are searched.

We first treat the saving of computation for multiple reference frames in the view point of compression. After prediction, residues are transformed, quantized, and then entropy coded. If we can detect that the transformed and quantized coefficients are very close to zero in the first reference frame, we can turn off the matching process from the rest frames since more computation will not cause any reduction in prediction errors. This concept is very simple and effective. Moreover, DCT, Q, IQ, and IDCT can also be saved by early detection of all-zero quantized coefficients. The quantization steps of 4x4-residues are described in the following equations:

$$qp\_per = QP/6 \quad (1)$$

$$qp\_rem = QP\%6 \quad (2)$$

$$qp\_bits = qp\_per + 15 \quad (3)$$

$$qp\_const = (1 << q\_bits)/6 \quad (4)$$

$$QM[i][j] = (|TR[i][j]| \times quant\_coef[qp\_rem][i][j] + qp\_const) >> q\_bits \quad (5)$$

where QP is quantization parameter (0-51), QM is 4x4-quantized magnitude, TR is 4x4-transformed residues, and $quant\_coef$ is a 3-D 6x4x4-matrix. If inequality (6) holds,

$$|TR| < (2^{q\_bits} - qp\_const)/quant\_coef \equiv f(QP) \quad (6)$$

the quantized magnitude will be zero, which means the threshold becomes a function of QP and can be implemented as a look-up table. Besides, TR is not available before transformation, so we assume residues are Laplacian distributed and find the relation between SAD or SATD and TR. The threshold is directly applied on SAD or SATD. The detailed derivation is omitted in this paper.

The mode decision result after intra prediction and ME from previous frame is also a very important cue. In Table 2, A|B is defined as follows. A is the percentage of a mode after intra prediction and ME from previous frame. B is the percentage of A that optimal reference frame and mode remain unchanged after 5 frames are searched. We can see that 73%, 4%, 4%, 17%, and 2% of the MBs are selected as 16x16, 16x8, 8x16, 8x8, and intra, respectively, when only previous one frame is searched. After the rest 4 frames are searched, 90% of the 16x16-MBs still remain as the optimal selection. As for 16x8-MBs, 8x16-MBs, 8x8-MBs, and intra MBs, the percentages that the optimal mode and reference frame do not change are 65%, 65%, 34%, and 7%, respectively. This means that 73%×90% + 4%×65% + 4%×65% + 17%×34% + 2%×7% = 76.82% of MBs need only previous one

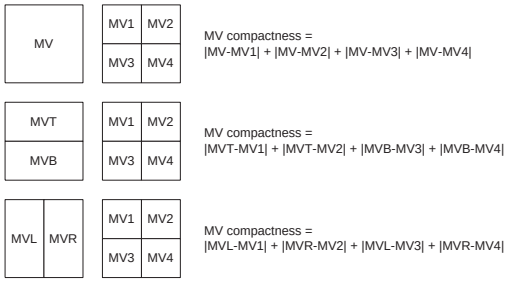

Fig. 2. Definition of MV compactness of a MB.

reference frame, which is quite consistent with the results in Table 1. Furthermore, when a MB is split into smaller blocks for motion compensation using only previous one frame, it means that the motion is discontinuous. In this case, the MB may cross the object boundaries, where occlusion and uncovering often occur. Thus, there is a greater possibility that the best matched candidates belong to the other 4 reference frames. When the intra-mode has better prediction than inter-modes from previous frame, the MB may belong to uncovered parts or new objects. The best candidates are very likely to be found in other 4 reference frames, too.

Now we try to find the correlation between MV distribution and optimal reference frames. After ME from previous one frame, we have one MV for 16x16-MB, two MVs for 16x8-MB, two MVs for 8x16-MB, and four MVs for 8x8-MB. If the best mode is 16x16, 16x8, or 8x16, the definition of MV compactness for each of the 3 modes is shown in Fig. 2, respectively. Next, we keep on searching the other 4 frames. If the optimal frame or mode of a MB does not change after ME from 5 frames, we classify these MBs as type I. Otherwise, we classify them as type II. The average MV compactness of type I and type II for each sequence is shown in Table 3. The MV compactness of MBs with optimal reference frame belonging to the rest 4 frames tends to be larger than that of MBs predicted by previous frame. Therefore, if the MV compactness of a MB after ME from previous frame is very small, we should stop searching the rest 4 frames.

The texture is also taken into consideration. The reduction of residues by applying multiple reference frames is more significant

Table 3. Statistics of Average MV Compactness.

Sequences	16x16		16x8		8x16	
Coastguard	08.9	12.6	09.9	12.5	07.8	09.5
Container	07.4	12.3	07.7	09.2	07.1	08.9
Foreman	08.3	16.4	10.1	13.2	10.5	13.2
Hall Monitor	08.4	14.5	09.8	12.7	10.7	12.5
Mobile Calendar	07.1	11.8	07.4	11.2	09.5	10.9
Mother and Daughter	06.9	10.5	07.7	11.4	10.0	11.6
Silent	07.2	11.9	10.4	14.2	11.6	16.2
Stefan	08.4	22.4	11.8	16.5	12.7	17.3
Table Tennis	08.6	17.5	08.1	13.4	10.4	16.7
Weather	07.7	11.8	07.1	08.5	07.2	10.7
Average	07.9	14.2	09.0	12.3	09.8	12.8

CIF size, search range [-16, +16], QP=30.

The unit of MV compactness is quarter pixel.

I|II is defined as follows.

I: optimal ref. frame and mode do not change after ME from 5 frames.

II: optimal ref. frame and mode change after ME from 5 frames.

Table 4. Statistics of Average SATD of Intra Prediction.

Sequences	16x16		16x8		8x16	
Coastguard	3599	5168	4932	5587	4846	5150
Container	3103	8218	4813	5242	5960	6010
Foreman	2104	3094	3103	3269	2562	3290
Hall Monitor	2341	3875	4015	2149	4182	1950
Mobile Calendar	6165	8371	6843	8239	6290	8723
Mother and Daughter	1482	2472	2488	2388	2250	2501
Silent	2881	2904	3010	2604	3227	3256
Stefan	3723	5981	5795	6113	5975	6188
Table Tennis	3025	3920	3773	3420	3725	3827
Weather	4648	4324	3985	3745	4286	4669
Average	3307	4833	4276	4276	4330	4556

CIF size, search range [-16, +16], QP=30.

P|Q is defined as follows.

P: optimal ref. frame and mode do not change after ME from 5 frames.

Q: optimal ref. frame and mode change after ME from 5 frames.

at object boundaries, where occlusion and uncovering often occur. The texture of object boundaries should be more complex than other flat regions. We use SATD after intra prediction to represent the complexity of texture of a MB. In Table 4, intra prediction and ME from previous one frame is first applied for each MB. Then we focus on the MBs having the best mode as 16x16, 16x8, and 8x16. Again, we keep on searching the rest 4 frames and classify the MBs into two types. MBs that do not change optimal reference frame and mode are classified as type P, while MBs with different reference frame and mode are denoted as type Q. It is clear that the SATD of intra prediction for type P is smaller than that for type Q. Therefore, if a MB is significantly textured, we should search more frames. However, it is also clear from Table 4 that this threshold value should be adaptive with different scenes or sequences.

There are some exceptions in Table 4, such as 16x8 and 8x16 modes for Hall Monitor, 16x8 mode for Silent. These are due to the complicated texture of stationary background that only requires one reference frame. In these cases, if we do not want to lose video quality, the threshold for SATD of intra prediction should be small enough, which will cause waste of computation for highly-textured stationary background. Fortunately, we can use the MV compactness to prevent these situations. Our algorithm will be described in the next section. Let us summarize the analysis as follows. After intra prediction and ME from previous one frame,

- If 16x16 mode is selected, the optimal reference frame tend to be unchanged.
- If inter-modes with smaller blocks are selected, searching more frames tend to be helpful.
- If MVs of larger blocks are similar to MVs of smaller blocks, it is likely that no occlusion or uncovering occurs in MB, so one reference frame may be enough.
- If MVs of larger blocks are more different from MVs of smaller blocks, MB often crosses object boundaries and thus requires more reference frames.
- If the texture of a MB is very complicated, it may require more reference frames.

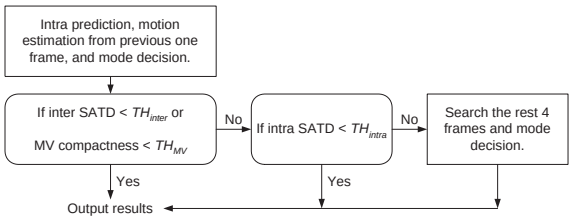

Fig. 3. Fast algorithm for multiple reference frames ME.

3. PROPOSED ALGORITHM

According to the above analysis, we propose a fast algorithm for multiple reference frame ME to save the computation of full search and to maintain the same video quality. The steps are shown in Fig. 3. Note that TH_{inter} is a function of QP and is implemented as a look-up table. TH_{MV} is empirically obtained. TH_{intra} must be adaptive with different scenes. Currently, we use the number of intra MBs in previous coded frame to detect scene change. If more than 10% of the MBs are intra-coded, we will adjust TH_{intra}. The detailed derivation of TH_{intra} is omitted due to the limited space. As shown in Fig. 4, we connected ten standard sequences together to show the dynamic adjustment of TH_{intra} according to the number of intra MBs in previous frame.

4. SIMULATION RESULTS

Figure 5 compares the rate distortion curves of the reference software and the proposed algorithm. It is shown that the maximum peak signal to noise ratio (PSNR) drop is 0.2 dB (Hall Monitor). The average PSNR drop is less than 0.05 dB, so that the two curves for each sequence are hardly distinguishable. Table 5 shows the miss detection rate and the false alarm rate of the proposed algorithm. Note that miss detection of optimal reference frames leads to the degradation of PSNR, and the false alarm results in waste of computation. In fact, TH_{intra} provides an easy trade-off between speed and quality. Adjusting TH_{intra} cannot decrease the miss detection rate and false alarm rate at the same time. The higher the TH_{intra}, the more the computation is saved. The lower the

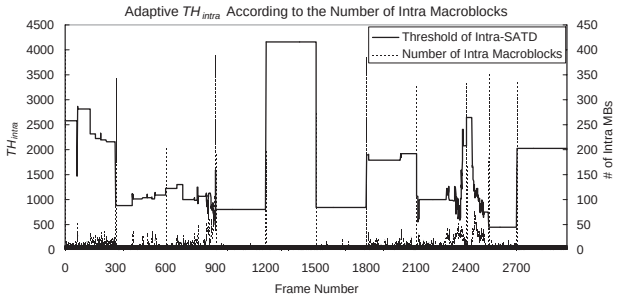

Fig. 4. Adaptive TH_{intra} according to the number of intra MBs.

Table 5. Miss Detection and False Alarm Rates.

Sequences	Miss Detection	False Alarm
Coastguard	6.21%	41.15%
Container	1.90%	24.96%
Foreman	2.57%	47.42%
Hall Monitor	5.37%	12.83%
Mobile Calendar	5.86%	31.60%
Mother and Daughter	2.67%	24.57%
Silent	3.73%	19.65%
Stefan	5.62%	42.36%
Table Tennis	2.41%	25.15%
Weather	2.64%	12.30%
Average	3.90%	28.20%

CIF size, search range [-16, +16].

TH_{intra}, the less the PSNR drop is achieved. The average miss detection rate is only 3.90%, which means 96.10% of MBs can find the optimal reference frame. The average false alarm rate is 28.20%, which means we should further improve our algorithm to save the 28.20% of computation while keeping the miss detection rate from rising in the mean time. Given a budget of computation resources, how to select TH_{intra} will also be our future work. Figure 6 shows the number of average searched frames for the reference software and the proposed algorithm. It is shown that 10%-67% of ME operations can be saved.

5. CONCLUSION

We proposed a simple and effective fast algorithm for multiple reference frames motion estimation. We first analyzed the available information after intra prediction and motion estimation from previous one frame. Then we applied several threshold values on the available information to determine if it is necessary to search more frames. Experimental results showed that our method can save 10%-67% of ME computation depending on sequences while keeping the quality nearly the same as full search scheme.

6. REFERENCES

[1] *Committee Draft of Joint Video Specification (ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC),* July, 2002.

[2] *Joint Video Team (JVT) software JM4.3,* October, 2002.

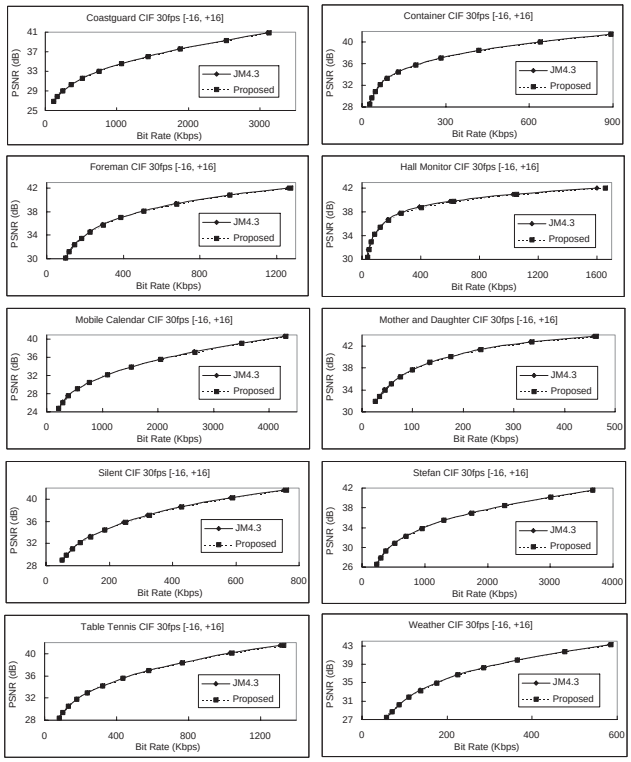

Fig. 5. Rate distortion curves of various sequences.

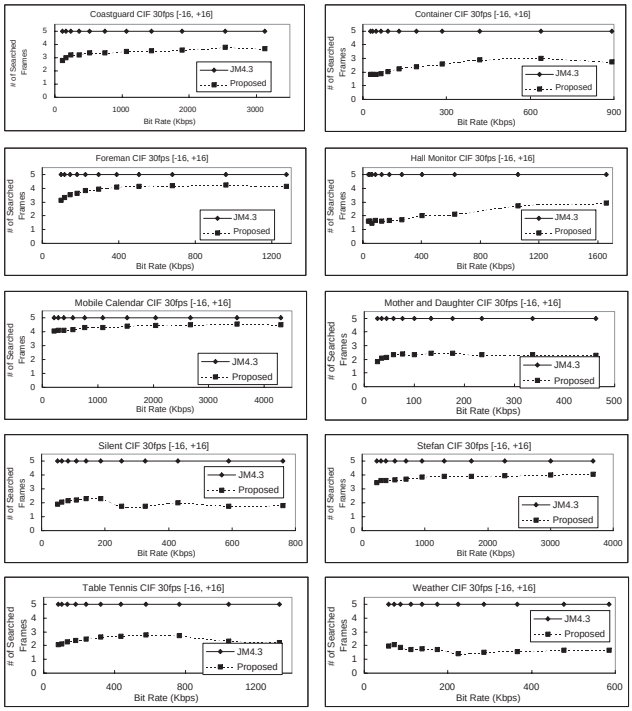

Fig. 6. Average searched frames for various sequences.