Fast Adaptive Inter-Prediction Mode Decision Method for H.264 Based on Spatial Correlation

Bin Feng, Guang-xi Zhu, Wen-yu Liu
Department of Electronic and Information Engineering
Huazhong Univ. of Sci. & Tech.
Wuhan, HuBei Province, 430074, China
Email: bfeng_hust@hotmail.com

Abstract—H.264 defines 7 different coding modes for macroblocks (MBs) in P slices. In order to achieve a coding performance as high as possible, the H.264 encoder calculates rate distortion costs of all possible modes to determine the best mode of a MB. The computation complexity is so large that make it difficult to be used in practical applications especially in real time environment. In this paper, a fast adaptive inter modes decision method is proposed to reduce the complexity of H.264 encoder. Firstly the candidate inter modes used in rate distortion optimization can be limited in a small mode group (MG) by using the characteristics of the motion compensated residual image. Then the two most probable modes of the chosen MG are obtained on the basis of the modes of the up MB and the left MB. By calculating and comparing the rate distortion cost of the two modes, the optimum mode of the MB is determined. The overlapped mode groups and dynamic adjusted thresholds adopted in the proposed method can make the best mode lie within the chosen MG with great possibility which leads to extensive computation reduction with acceptable loss in quality. The experimental results show that the proposed method can save the encoding time up to 65% on average with -0.24dB performance degradation.

I. INTRODUCTION

MPEG-4 Part-10 H.264/AVC [1], the newest video coding standard, is approved by the joint video team of ISO/IEC MPEG and ITU-T VCEG. It employs several complicated techniques including variable block-size motion compensation, motion compensation with quarter-pel accurate motion vectors even outside the picture boundaries, multiple reference pictures and so on. These approaches in encoding procedure enable H.264/AVC to outperform the contemporary video coding standard in rate distortion performance. Compared with the MPEG2, H.264 can save nearly 50% bit rate under the same visual quality [2].

However, the excellent performance of H.264/AVC is achieved at the expense of computation complexity increment. Its encoding time is 16 times more complex than the MPEG-4 simple profile [3]. In the previous standards, only MB block size (16×16 pixels) is used in motion estimation which means that one motion vector is assigned to the whole MB. This method is suitable to the stationary or uniformly moving MB. However, if the MB contains many details, or if the MB lies on the boundary of different objects, only one motion vector will not provide sufficient accuracy for motion compensation and will result in serious performance degradation.

H.264 defines 7 coding modes for MB in P slice, including SKIP mode, 4 modes for inter prediction, and 2 modes for intra prediction. Fig.1 depicts all the available coding modes. An inter coded MB may be divided into MB partitions, i.e. blocks of size 16 × 16, 8 × 16 or 8 × 8 luma samples (and associated chroma samples). If the 8 × 8 partition size is chosen, each 8 × 8 sub-macroblock may be further divided into sub-macroblock partitions of size 8 × 8, 8 × 4 or 4 × 8 or 4 × 4 luma samples (and associated chroma samples). The MB with mode 0 is encoded in SKIP mode and MB with mode 9, 10 is encoded in intra mode. In order to select the best mode for each MB, the encoder exhausts all possible modes in the rate distortion optimization (RDO) framework. The maximum number of computing RD costs values for mode decision is 768 [4]. As a result, the complexity of encoder increases dramatically which makes it difficult to be implemented in real time applications. Experimental results show that motion estimation is the most computation-intensive module, which consumes about one-third of the processing cycles.

Many fast motion estimation algorithms have been proposed recently. In [5], the optimum inter mode is decided based on the predicted relationship among RD cost of different modes. But the RD costs of different modes are not easy to be predicted. Another fast algorithm has also been proposed by Wu. D et al. to use the spatial homogeneity and
the temporal stationarity characteristics of video objects to guide the inter prediction process [6].

In this paper, a fast adaptive inter prediction modes decision method is proposed to reduce the candidate number of inter modes calculated in RDO framework by using the characteristics of the motion compensated residual image and that of the neighboring MB modes. The overlapped mode groups and dynamic adjusted thresholds adopted in the proposed method can make the best mode lie within the chosen MG with great possibility which leads to extensively computation reduction without any perceivable loss in quality. The rest of this paper is organized as follows. In section II, the proposed inter mode decision method is presented. Section III shows and illustrates the experimental results. The conclusions are given in section IV.

II. PROPOSED ALGORITHM

G.R. Kwon et al. proposed an inter prediction method on the basis of the deduction that the MB with best mode 0 (MG1) has the residual errors of zero value, the residual errors of MB with best modes 1 to 3 (MG2) are small, whereas those of MB with best modes 4 to 7 (MG3) are large, and moreover, while the MB with intra mode (MG4) has very large residual values [7]. The residual error is obtained by searching MV for the MB size. Comparing a measure of residual error $N$ with predefined empirical thresholds, the encoder can limit the possible candidate modes into one of the four MGs stated above.

In [7], Kwon et al. considered the key factor that affects the performance of the method is how to obtain more accurate thresholds for the classification of the MGs. But in our opinion, the classification of the MGs is the most important issue of whether the method that utilizes the characteristics of the residual error is efficient. Only when the following condition is satisfied that the classification of the MGs adapts well to the distribution of $N$ in different cases, can the optimum mode of a MB lie in the predicted MG with great possibility.

A. The overlapped mode groups

In order to investigate the distribution of $N$ for different coding modes, we run the H.264 reference software JM85 [8] on many test video sequences. Parts of the results are depicted in Fig.2 and Fig.3, corresponding to “Bus” and “Silent” QCIF videos respectively. Here $N$ denotes the number of pixels of a residual MB after motion compensation that is larger than an empirical threshold, obviously $0 \leq N \leq 256$. It can be seen from the experiment results that:

1. The average value of $N$ for mode 9 and 10 exceeds that of other modes obviously and there are almost no overlapped parts with other modes.

2. The distribution of $N$ for mode 0 is very close to zero and differs greatly to other modes.

3. The distribution for mode 1, 2, 3 and that for mode 8 are so similar that they are not easy to be classified accurately.

4. The dynamic range of the $N$ for mode 1, 2, 3 and that for mode 8 is so large that it will overlap the range of mode 0 with great possibility.

Here mode 8 include mode 4, 5, 6, and 7 actually. That is to say, for a certain value of $N$, the optimum mode of the corresponding MB may fluctuate greatly from SKIP mode to intra modes for different sequences or even different frames. Summing up, for the modes that the dynamic range of $N$ is large, they should belong to several MGs simultaneously. Only in this way, can the best mode of the current MB be guaranteed to belong to the chosen MG with great possibility.

Based on the investigations of the distribution of $N$, three hypotheses were assumed as follows. Here $T_1$, $T_2$, and $T_3$ are empirical thresholds where $0 < T_1 < T_2 < T_3$.

a) The mode whose $N$ is larger than a small threshold $T_1$ should not be the mode 0.

b) The mode whose value of $N$ is less than $T_2$ seems not to be intra modes, including mode 9 and 10.

c) The mode whose value of $N$ is larger than a big threshold $T_3$ can be regarded as intra modes.
Also from the results of many other video sequences, the rationality of the three hypotheses can be verified in most cases. Based on the hypotheses above, the new MGs are proposed as Table I. As the Table I illustrate, the most important trait of these MGs is that each group is overlapped with each other that can sustain the optimum mode that lies in the chosen MG with great possibility. And the problem of error estimation of mode at the border of two modes can be solved at the expense of more encoding time needed. After MGs can be solved, the problem of error estimation of mode at the border of two overlapped with each other that can sustain the optimum cases. Based on the hypotheses above, the new MGs are rationality of the three hypotheses can be verified in most.

<table>
<thead>
<tr>
<th>Mode group</th>
<th>Coding modes</th>
</tr>
</thead>
<tbody>
<tr>
<td>MG1</td>
<td>Mode 0, 1, 2, 3, and 8</td>
</tr>
<tr>
<td>MG2</td>
<td>Mode 1, 2, 3, and 8</td>
</tr>
<tr>
<td>MG3</td>
<td>Mode 1, 2, 3, 8, 9, and 10</td>
</tr>
<tr>
<td>MG4</td>
<td>Mode 9 and 10</td>
</tr>
</tbody>
</table>

B. The prediction of the probable modes

Once the MG is chosen, the method will predict the probable modes of the chosen MG on the basis of the modes of the up MB and the left MB. Based on the analysis of [9], one characteristic for motion estimation scheme of H.264 is adopted in this paper that if a boundary crosses a MB, it is split into smaller blocks [9]. Due to the inherent spatial correlation within a single frame, the direction of the boundary in current MB can be predicted by that of the neighboring MBs. As illustrated in the Fig.4, if the boundary crosses through the up MB in vertical direction, the direction of the boundary in current MB is likely to be vertical too. For the same reason, if the optimum mode of the left MB is 16x8 which means that there is a horizontal boundary in the left MB, the optimum mode of the current MB may be 16x8 too with great possibility. Summing up, the texture information of neighboring MBs is correlated in a frame so that the mode of current MB can be predicted using that of the left MB and up MB.

![Figure 4. The direction correlations in the neighboring MBs](image)

A probability table relates the prediction modes of adjacent MBs to the prediction mode of the current MB. For each combination of adjacent MBs, the probability table includes a list of prediction modes in order of expected occurrence. Let l and u be the mode index of the left MB and up MB respectively. M(l) or M(u) indicates one of the modes{0, 1, 2, 3, 8, 9, 10}. Let p be the mode index of current MB in order of expected occurrence. Many standard video sequences are tested to establish the table Correlation[7][7][7] where Correlation[7][u][p] denotes the

<table>
<thead>
<tr>
<th>Table II. MODE GROUPS PROPOSED</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode group</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>MG1</td>
</tr>
<tr>
<td>MG2</td>
</tr>
<tr>
<td>MG3</td>
</tr>
<tr>
<td>MG4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TABLE II. THE MODES PRIORITY OF CURRENT MB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mode</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>M(l)</td>
</tr>
<tr>
<td>M(u)</td>
</tr>
</tbody>
</table>

pth probable mode of the current MB when the mode of the left MB is M(l) and the mode of up MB is M(u). Parts of the table Correlation[7][7][7] are showed in Table II. Here only three kinds of most possible modes are listed. Using this table, an ordering of the most probable, next most probable prediction modes for current MB is determined. For example, suppose the prediction mode for left MB is mode 1 and for up MB is mode 3. Table II lists the string 1, 0, 8. This indicates that mode 1 is the most probable prediction mode for the current MB. Mode 0 is the next most probable prediction mode, etc. Hence after the MG is determined, the proposed method will look up the modes priority for 7 modes of the current MB in table Correlation[7][7][7] using M(l) and M(u). Combined with the modes of the chosen MG, the modes priority of the chosen modes subset can be deduced. Considering the encoding efficiency and complexity, only the most and next most probable modes in the chosen MG are used to calculate and compare the rate distortion cost. At last, the current MB is encoded as the optimum mode whose rate distortion cost is minimal. The rate distortion function is as follows:

\[
J_i(m) = D_i(m) + \lambda (m) R_i(m)
\]

In the above equation, m is the candidate mode of the determined MG, J_i(m) is the cost function of the ith MB, D_i(m) is the sum of the squared differences between the original MB and its reconstruction, \( \lambda (m) \) is the Lagrange multiplier for mode decision, and \( R_i(m) \) is the bitrate of the MB.

C. The dynamic adjusted thresholds

Let \( E(x,y) \) be the residual value of pixel (x, y) in the ith video frame after motion compensation, and \( F(x,y) \) be the corresponding original value. We denote the reconstruction value in the feedback loop at the encoder as \( \hat{F}_i(x,y,q) \) when the quantization parameter is \( q \). Let \( d(x,y,q) \) be the quantization error of pixel (x, y) in the ith video frame when the quantization parameter is \( q \). Therefore, \( E(x,y) \) is given by

\[
E(x,y) = F(x,y) - \hat{F}_i(x,y,q) = F(x,y) - \hat{F}_i(x-v_x, y-v_y, q) = F(x,y) - \hat{F}_i(x-v_x, y-v_y) + d_i(x-v_x, y-v_y, q)
\]

It should be noted that \( \hat{F}_i \) here represents the index of reference frames, and \( v_x, v_y \) represents the horizontal component and vertical component of motion vector.
that tools such as curve fitting. constants that can be determined by using mathematical results between ability of the terminal. Hence it is sufficient to consider the constraints of the channel capacity and the treatment way. That is to say proposed method should be dynamic adjusted in the same other video sequences. So the thresholds used in the almost linearly. Similar results have been obtained over other video sequences. So the thresholds used in the proposed method should be dynamic adjusted in the same way. That is to say T(q)=kq+b. Here k and b are two constants that can be determined by using mathematical tools such as curve fitting.

III. RESULTS

The proposed algorithm was implemented by modifying the H.264 codec JM85. The performances of the fast adaptive inter modes decision method are tested using the first 100 frames of 6 testing video sequences (Foreman QCIF, Bus QCIF, Football QCIF, Mobile QCIF, Paris QCIF, and Tempete QCIF). The frame rate was 30fps, and the frame coding structure was IPPP. In the motion estimation, two reference frames were enabled with the motion vector resolution of 1/4 pixel and the maximum search range of ±16. The RDO and the Hadamard transform were enabled. The universal variable length coding was used as entropy coding. Experiments were conducted for four quantization parameters QP = 28, 32, 36, and 40. The conditions include Celeron 1.7GHz PC with 384M memory.

The performance comparisons in terms of PSNR and bitrates between the original encoder and the proposed method are showed in TABLE III. ∆PSNR and ∆Bits denote the differences of PSNRs and bitrates of JM85 and the method respectively. A minus sign (-) of ∆PSNR and a plus sign (+) of ∆Bits mean that performance is degraded due to less candidate modes which are calculated in RDO framework. Referring to the TABLE III it can be seen that the PSNR degradation of the proposed method is no more than -0.42dB and on average is -0.24dB which is negligible.

IV. CONCLUSION

A fast adaptive inter modes decision method is proposed in this paper. The characteristics of the residual error image after motion estimation and that of the neighboring MB modes are exploited to reduce the number of candidate modes to be calculated in the RDO procedure. By using overlapped MB groups and dynamic adjusted thresholds, the proposed method can maintain the objective quality at a reasonable level while decrease the computation complexity greatly. Compared with H.264 reference software, experimental results show that the average time savings is up to 65% with slight performance degradation that is -0.24dB on average which is negligible.

REFERENCES