An Efficient Inter Mode Decision Approach for H.264 Video Coding

Xuan Jing and Lap-Pui Chau
School of Electrical and Electronic Engineering
Nanyang Technological University,Singapore
Email: jingxuan@pmail.ntu.edu.sg  elpchau@ntu.edu.sg

Abstract

Variable size block motion estimation is a very important technique for video coding. The new H.264 standard employs 7 different size block types which can significantly improve the coding performance compared with the previous video coding standards. On the other hand, the computational complexity of H.264 encoder increases dramatically due to the various coding modes used. In this paper, an efficient inter mode decision approach is presented. The objective is to reduce the number of candidate block types in the motion estimation while maintaining the coding efficiency. Experimental results show that the proposed method can save the computation cost by up to 42% at the same PSNR and bitrate.

1. Introduction

The JVT H.264 video coding standard [1] achieves much higher coding efficiency than the previous standards such as H.261 and H.263. This is mainly due to the fact that the H.264 encoder employs more complicated approaches in the coding procedure. One important approach is variable size block motion estimation and mode decision. Traditionally, motion estimation is performed only on the macroblock (MB) level thus each 16×16 macroblock will be assigned one motion vector which can lead to a minimum block matching error. However, when the macroblock contains multiple objects and every object moves in different directions or when the macroblock lies on the boundary of a moving object, only one motion vector for the whole block will not be enough to represent true motions and it will result in serious prediction error. In order to improve the prediction accuracy, H.264 enables 7 different block sizes in motion estimation which are 16×16, 16×8, 8×16, 8×8, 8×4, 4×8 and 4×4. When conducting mode decision for a macroblock, firstly, rate distortion optimized (RDO) motion estimation is carried out for each 8×8 sub-partition and its mode decision is done by minimizing the Lagrangian functional:

\[ J(s,c,\text{MODE} | QP, \lambda_{\text{MODE}}) = \text{SSD}(s,c,\text{MODE} | QP) + \lambda_{\text{MODE}} \cdot R(s,c,\text{MODE} | QP) \]  

where \( QP \) is the macroblock quantization parameter, \( \lambda_{\text{MODE}} \) is the Lagrange multiplier for mode decision, SSD is the sum of the squared differences between the original block \( s \) and its reconstruction \( c \) and MODE is one of the potential prediction modes. Similarly, motion estimation will be performed for 16×16, 16×8 and 8×16 modes and the final mode decision for the macroblock is made by examining the Lagrangian cost function (1). Although the rate distortion performance of H.264 encoder is much better than the previous standards, the additional computation burden caused by exhaustively searching all combinations of different modes makes it difficult to achieve real time implementations. Besides various fast motion estimation algorithms, many fast mode decision strategies have been proposed recently to further reduce the computation load of H.264. In [2] Tu et al proposed a merging procedure for 8×8 blocks by checking the distances between their motion vectors. In [3] and [4] early termination was used to reduce the number of potential prediction modes. With the help of an edge map, Lim et al in [5] assigned proper potential modes for each macroblock and its sub-partitions according to their homogeneity properties. Different from the above mentioned methods, this paper proposed a very efficient classification method which can reduce the average number of block types used while maintaining the same coding performance.

The rest of the paper is organized as follows. In Section 2 we will present the proposed inter mode decision method. Section 3 illustrates the simulation results for performance comparison. Finally, a conclusion is given in Section 4.
2. Efficient inter mode decision

Variable size block motion estimation has been studied for many years. The main concept is to use large blocks for homogeneous areas and small blocks for areas containing complex motions such that the resulting entropy for the residue and motion vectors can be minimized. This has been successfully implemented into the H.264 standard recently which results in a better rate distortion performance. However, exhaustively examining all combinations of different coding modes for every macroblock is not desirable although it is very accurate. Since if a macroblock is belong to a homogeneous area there is little chance that it will be split into smaller size blocks according to the Lagrangian functional. Therefore it is not necessary to examine smaller size block types for these macroblocks.

Previously, Huang et al [6] classified the macroblocks into background, shade motion and edge motion blocks by a two-stage classifier. In their method, motion vectors for background blocks were directly set to zero. For shade motion blocks pixel subsampling technique was used to reduce the computation of motion estimation. The variable size block matching was only conducted for edge motion blocks. In order to classify macroblocks as homogeneous or non-homogeneous, an edge map was used in [5]. If the amplitude of the edge vectors within a block is less than a threshold, it will be classified as homogeneous. Further, modes for smaller size blocks will be disabled for homogeneous blocks.

Without using multi-stage classifier or edge detection, our proposed efficient inter mode decision method only depends on the absolute differences between consecutive frames. Generally, the absolute frame difference contains lots of information about the motions in successive frames. Large amplitudes will appear on the moving edges or boundaries of moving objects while small amplitudes in homogeneous areas. So if the amplitudes in a macroblock are small, it is most likely that this macroblock belongs to a homogeneous region and using only larger block sizes in motion estimation will be accurate. Otherwise, this macroblock may contain complex motions and using more block types can achieve better rate distortion performance. In the proposed efficient inter mode decision algorithm, we first subtract the current frame by its previous frame. Then the sum of absolute difference for the current macroblock can be calculated using (2).

\[
SAD = \sum_{i=1}^{N} \sum_{j=1}^{N} |x_{i,j} - y_{i,j}|
\] (2)

where \(x_{i,j}\) and \(y_{i,j}\) denote the gray levels of pixels at location \((i,j)\) of the current frame and its previous frame respectively and \(N\) is the macroblock size which is 16. If the SAD is less than a threshold (TH) this macroblock is considered as in a homogeneous region and only large block types \((16 \times 16, 16 \times 8, 8 \times 16)\) will be used in its motion estimation. Otherwise it belongs to a moving edge region and additional four block types \((8 \times 8, 8 \times 4, 4 \times 8\) and \(4 \times 4\)) will also be enabled. The proposed algorithm is summarized as follows:

**Step 1:** Subtract the current frame by its previous frame and calculate the SAD for the current macroblock using (2).

**Step 2:** Compare the SAD with a threshold \((TH)\).
- If \(SAD<TH\) go to step 3;
- Otherwise go to step 4;

**Step 3:** Conduct rate distortion optimized (RDO) motion estimation and mode decision using \(16 \times 16, 16 \times 8, 8 \times 16\) block types.

**Step 4:** Conduct rate distortion optimized (RDO) motion estimation and mode decision using all \(7\) block types \(16 \times 16, 16 \times 8, 8 \times 16, 8 \times 8, 8 \times 4, 4 \times 8\) and \(4 \times 4\).

Note that, in calculating SAD the previous original frame is used as the reference rather that the previous reconstructed frame. This is because coding errors in the reconstructed frame may affect the accuracy of this classifier. By employing such classification algorithm, computational complexity of the original encoder is greatly reduced while the coding efficiency is maintained. In the next section simulation results will be discussed to demonstrate the performance of our method.

3. Simulation results

The proposed mode decision approach was tested using the first 100 frames from four testing video sequences (Foreman, Carphone, Miss America and Mother & Daughter) all in QCIF format 176×144. In these test videos, Foreman and Carphone have relatively large motions while the other two have moderate motions. The experiment was carried out on the JVT JM6.0 encoder [7]. In the motion estimation, 5 reference frames were enabled with the maximum search range 32 and the motion vector resolution was 1/4 pixel. CABAC (Context Adaptive Binary Arithmetic Coding) was adopted as the entropy coding method. The Hadamard transform was used to transform DCT coefficients. The frame rate was 30 fps and the frame coding structure was IBBP. The
experiments was conducted for four quantization parameters QP=28,32,36 and 40.

The PSNR and bitrate comparisons between the original encoder and the one with the proposed efficient inter mode decision method are tabulated in Table 2. The coding results are very similar. Compared with the original encoder, the PSNR degradations of the proposed algorithm are no more than 0.08 dB and in most cases bitrate savings are achieved. On average, the proposed scheme achieved 0.2% bitrate saving at the cost of 0.04 dB PSNR drop. In other words, the rate distortion performances of the two methods are the same. The computational complexity comparisons are illustrated in Table 1. Instead of using 7 block types for every macroblock in motion estimation, only 4 to 5 block types were actually employed. Thus averagely about 36% computations are saved. Figure 1 and Figure 2 give examples of the classification results and the final mode decisions for Foreman and Mother & Daughter sequences. In Figure 1(b) and 2(b) the white blocks are the macroblocks with SAD values larger than the threshold (TH), they represent the areas which may contain complex motions. Figure 1(c) 2(c) and 1(d) 2(d) demonstrate the decoded frames of the original encoder and the modified one with their coding modes respectively.

4. Conclusion

In this paper, an efficient inter mode decision algorithm for H.264 standard has been proposed. The motivation is to predict the potential block types before conducting the rate distortion optimized motion estimation. A classifier based on the absolute frame difference has been employed to decide which block types are to be enabled. Experimental results show that the proposed method achieves large computation savings while maintaining the same rate distortion performance.

5. References


Table 2. Comparisons for PSNR and Bitrate

<table>
<thead>
<tr>
<th>Sequence</th>
<th>QP=28 ΔPSNR(dB)</th>
<th>QP=32 ΔPSNR(dB)</th>
<th>QP=36 ΔPSNR(dB)</th>
<th>QP=40 ΔPSNR(dB)</th>
<th>QP=28 ΔBitrate(%)</th>
<th>QP=32 ΔBitrate(%)</th>
<th>QP=36 ΔBitrate(%)</th>
<th>QP=40 ΔBitrate(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QP=28</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.04</td>
<td>-0.01</td>
<td>0.9</td>
<td>-0.5</td>
<td>0.3</td>
<td>-0.6</td>
</tr>
<tr>
<td>Carphone</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QP=28</td>
<td>-0.08</td>
<td>0</td>
<td>0</td>
<td>-0.01</td>
<td>-0.1</td>
<td>0.2</td>
<td>-0.5</td>
<td>-1.1</td>
</tr>
<tr>
<td>Miss America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QP=28</td>
<td>-0.08</td>
<td>-0.06</td>
<td>-0.03</td>
<td>-0.03</td>
<td>-0.5</td>
<td>-0.8</td>
<td>0.7</td>
<td>-0.7</td>
</tr>
<tr>
<td>Mother &amp; Daughter</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QP=28</td>
<td>-0.07</td>
<td>0</td>
<td>-0.01</td>
<td>-0.08</td>
<td>0.4</td>
<td>-0.3</td>
<td>-0.2</td>
<td>-1.0</td>
</tr>
</tbody>
</table>

Table 1. Complexity comparisons

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Average Blocktypes</th>
<th>Δ Complexity(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreman</td>
<td>5.04</td>
<td>-28</td>
</tr>
<tr>
<td>Carphone</td>
<td>4.48</td>
<td>-36</td>
</tr>
<tr>
<td>Miss America</td>
<td>4.08</td>
<td>-42</td>
</tr>
<tr>
<td>Mother &amp; Daughter</td>
<td>4.24</td>
<td>-39</td>
</tr>
</tbody>
</table>
Figure 1. Samples of Foreman

(a) (b)

(c) (d)

Figure 2. Samples of Mother & Daughter

* (a) Original frame (b) The classification result based on the SAD thresholding (c) Decoded frame with corresponding macroblock modes by the original H.264 encoder (d) Decoded frame with corresponding macroblock modes by the proposed method