Early Termination Schemes for Fast Intra Mode Decision in High Efficiency Video Coding

Hao Zhang
School of Information Science and Engineering
Central South University
Changsha, Hunan 410083 China
Email: hao@csu.edu.cn

Zhan Ma
Dallas Technology Lab
Samsung Telecommunications America
Richardson, TX 75082 USA
Email: zhan.ma@sta.samsung.com

Abstract—The emerging High-Efficiency Video Coding video coding standard has shown the significant coding performance improvement compared to the H.264/AVC with the cost of huge complexity increase. Hence, HEVC fast encoding algorithms are highly demanded for real-time applications. In this paper, we propose several early termination schemes for fast intra prediction in HEVC. More specifically, variation of coding mode costs are used to terminate current coding unit (CU) mode decision as well as TU size selection, where the CU costs are derived at the rough mode decision phase using Hadamard transform. Neighboring modes’ costs are also used to skip full rate-distortion optimized quantization (RDOQ). For all available test sequences provided by the JCT-VC, it demonstrates about 32% encoding time reduction for All Intra configuration with BD-RATE increase about 1.1%. Our proposal is complementary to some of other fast intra prediction methods such as fast rough mode decision. We expect more encoder speedup by integrating our method with other published works.

I. INTRODUCTION

The emerging High-Efficiency Video Coding standard (HEVC), developed under the efforts of Joint Collaborative Team on Video Coding (JCT-VC), has demonstrated significant performance improvement on compression ratio over the H.264/AVC [1], [2]. It is noted that part of the coding efficiency improvement comes from the newly introduced recursive tree structure and larger block transforms [1], [3], [4]. Block structure definitions in the previous standards are extended in HEVC, noted as Coding Unit (CU), Prediction Unit (PU) and Transform Unit (TU). Each CU can be recursively split into multiple sub-CUs and each leaf sub-CU can be split into multiple PUs as well. To further improve the compression efficiency, HEVC also adopts recursive TUs at each PU level for residual coding. A brief illustration of the recursive CU, PU and TU is shown in Figure 1.

Intra prediction in HEVC is significantly improved from the H.264/AVC with 35 prediction modes, as illustrated in Figure 2. As we can see, increased spatial intra prediction modes lead to much higher compression efficiency while requires a huge mount of computational power for optimal mode selection with full Rate-Distortion Optimized Quantization (RDOQ). Hence, a three-step method is adopted in the HEVC reference software HM [5] for intra encoding complexity reduction. A pre-process, known as the rough mode decision (RMD), is performed to select a small set of candidate modes for the normal computational intensive RDOQ. Please note that Hadamard transform based cost evaluation is used in the RMD process. Then, recursive quadtree based TU selection (RQT) is performed on the best mode from the previous RDOQ process to determine the best TU partitions. However, such three-step encoding process could reduce the coding complexity partially, and further fast intra mode decision mechanisms are still required for real time implementations.

In this paper, we propose several novel early termination schemes to speed up the intra prediction in HEVC. First, after the rough mode decision, variation of the mode costs (referred as Hadamard cost which is derived based on the Hadamard transform in the RMD process) are computed. If the cost exceeds the pre-defined threshold, CU is directly split into next depth without continuing the RDOQ and RQT at current depth. During each recursive CU split process, the accumulated costs are used to predict the overall cost. The computation would be terminated if the predicted cost exceeds the current minimum cost. A similar prediction method is used to early terminate the TU size selection. Finally, more complexity reduction is achieved after comparing Hadamard costs of adjacent intra modes.

The remainder of this paper is organized as follows. Section II presents a brief literature review of the fast intra prediction algorithms for HEVC. Section III describes the proposed early termination schemes. Experiments are performed in Section IV to demonstrate the effectiveness of the proposed solution, while Section V concludes this paper.
II. FAST INTRA MODE DECISION FOR HEVC: A REVIEW

Recently, a few fast intra mode decision (MD) algorithms are developed for HEVC. Some of them are combined with the fast inter mode decision algorithms. For example, Tan et al. propose a fast RQT algorithms for both intra and inter mode decision [6]. The number of candidate modes is reduced to three for the rough mode decision. However, for each candidate mode, the transform depth is explored to determine the optimal TU size. Teng et al. propose a Merge-and-Split decision process and reduce the coding complexity by using the inheritance property of zero-blocks [7]. For non-zero TUs, it uses rate and distortion predictions to early terminate the merge or split process. Aforementioned algorithms only deal with RQT and do not consider the fast termination of CU or mode selection processes.

Zhao et al. compared various settings with different mode candidates for the full RDOQ [8]. It proposed a smaller candidate set that does not lead to noticeable quality loss. Choi et al. recently propose a tree pruning algorithm for fast CU decision [9]. It skips the further CU splitting if the current CU chooses the SKIP mode. Such SKIP mode based early termination can not be applied to the intra prediction. Jiang et al. employ a gradient based method to speed up CU decision [10]. Based on the gradient directions for each CU, only a small set of modes are selected for further intra prediction. Hence, a large portion of modes are relieved from computation.

Moreover, Bayesian decision rule is used by Shen et al. to make non-split decisions where Bayesian risk and feature vectors need to be calculated [11]. A content adaptive fast intra mode decision method is proposed by Tian and Goto based on texture complexity analysis [12]. Mode filtering is used to reduce the number of candidates. Down-sampling is used to generate 16x16 blocks from 64x64 based LCUs to calculate the texture complexities. Small PU size is skipped for low complexity PU while and large PU size is skipped for high complexity PU.

Besides, we recently propose a new method for fast intra prediction mode decision at each CU depth [13]. First, a 2:1 downsampld Hadamard transform is used for the rough mode decision. Furthermore, a gradual search is employed to reduce the number of modes for Hadamard cost calculation. Finally, an early termination method is applied to speed up the RDOQ process. In this paper, we focus on the early termination of both CU and TU process based on Hadamard costs, adjacent modes and cost prediction. The proposed scheme is complementary to our previous work and some other published works in the literature.

III. EARLY TERMINATION SCHEMES FOR FAST INTRA PREDICTION ALGORITHM

Our proposed method mainly employs the Hadamard costs to accelerate the intra mode decision process. The Hadamard cost is calculated by combining the sum of absolute Hadamard transformed difference (SATD) with estimated mode bits consumption. Overall algorithm is composed of three sub-algorithms, including early termination of the CU and TU split, early stop of the intra prediction mode decision at certain CU depth, and early RDOQ skip of the certain modes.

A. Early Termination of CU and TU Split

Given a 2Nx2N CU, Rate-Distortion (R-D) costs are recursively calculated in the RDOQ process. The cost of the current level \( n \) (referred as \( C_n \)) is always compared with the sum of costs of its four NxN sub-CUs (noted as \( C_{n+1} \)). The CU level with the smaller RD cost is selected for next step. These four sub-CUs are encoded in a raster-scan order, i.e., from the left-top sub-CU to the right-bottom CU. We use \( C_{n+1,i} \) to represent the R-D cost for the \( i \)-th sub-CU (\( 1 \leq i \leq 4 \)) and use \( AccuC_{n+1,i} \) to denote the accumulated R-D cost after encoding the \( k \)-th \( (k \in [1, 2, 3, 4]) \) NxN sub-CU. Hence, we have

\[
AccuC_{n+1,i} = \sum_{i=1}^{k} C_{n+1,i}.
\]

Instead of using the real R-D cost after encoding, proposed scheme would use the predicted cost \( PredC_{n+1,k} \) as an approximation of \( C_{n+1} \). The split process will be early terminated if the following inequality holds, i.e.,

\[
PredC_{n+1,k} > \alpha_k C_n
\]

Here \( \alpha_k \) are the parameters. In this paper, we set \( \{\alpha_1, \alpha_2, \alpha_3, \alpha_4\} = \{1.5, 1.2, 1.1, 1\} \) for all sequences.

In HM, Hadamard cost is calculated for each 8x8 or 4x4 block and the total cost is the summation of the costs for all the blocks. As for each sub-CU, we obtain Hadamard cost by summing up all the blocks within this sub-CU. Please note that Hadamard cost is available in the rough mode decision step without requiring actual encoding the CU and its sub-CUs. We use \( AccuHC_{n,k} \) to represent the accumulated Hadamard
cost at level \( n \) after encoding the \( k \)-th sub-CU. \( \text{PredC}_{n+1,k} \) is then derived using
\[
\text{PredC}_{n+1,k} = \min \left\{ \frac{4}{k} \frac{\text{AccuHC}_{n+1,k}}{\text{AccuHC}_{n+1,k}} \right\} \text{AccuHC}_{n+1,k}. \tag{3}
\]
Here the ratio between the \( k \)-th accumulated and the total Hadamard cost is used to approximate the ratio between the accumulated R-D cost (up to \( k \)-th sub-CU) and the total RD cost in RDOQ. Note that \( \text{AccuHC}_{n+1,k} \) denotes the total Hadamard cost for all four CUs at level \( n+1 \). It is also noted that \( \text{AccuHC}_{n+1,k} \) is actual accumulated R-D costs after encoding \( k \)-th CU.

Combining Eq. (1)-(3), we have
\[
\min \left\{ \frac{4}{k} \frac{\text{AccuHC}_{n+1,k}}{\text{AccuHC}_{n+1,k}} \right\} \sum_{i=1}^{k} C_{n+1,i} > \alpha_k C_n. \tag{4}
\]
If Eq. (4) holds for any \( k \), CU splitting is terminated for level \( n+1 \). As we can see, for an \( n \)-th level 2Nx2N CU, we might only need to encode the first \((n+1)\)-th level NxN sub-CU, i.e., \( k = 1 \), to skip the CU splitting. There is also the case that we have to encode all sub-CU to decide the splitting, i.e., \( k = 4 \). In such case, it is the same as the normal R-D optimized encoding. However, simulations show that a normal takes the value less than 4, resulting the complexity reduction by skipping the CU splitting.

TU split has the similar recursive structure as CU split and is determined after selecting the CU size. Hence, we can use the actual encoding R-D cost to decide whether the TU size is required further splitting, i.e.,
\[
\frac{4}{k} \sum_{i=1}^{k} C_{n+1,i} > \beta_k C_n, \tag{5}
\]
where \( \beta_k \) are also parameters, which are set to \( \{\beta_1, \beta_2, \beta_3, \beta_4\} = \{1.2, 1.1, 1, 1\} \) in this paper.

B. Early Termination of CU Cost Calculation

In Sec. III-A, proposed scheme is to early stop the splitting process using the Hadamard cost. In this part, we use the variation of the Hadamard cost to jump to next CU level. Let Hadamard cost of each NxN sub-CU (of a 2Nx2N CU) \( HC_{n,i} \), \( 1 \leq i \leq 4 \) which is the sum of all the 4x4 or 8x8 Hadamard costs within this \( i \)-th sub-CU. We use the scaled mean absolute deviation (SMAD) of the Hadamard cost as follows:
\[
\text{SMAD}_n = \frac{\sum_{i=1}^{4} |HC_{n,i} - \mu_{HC_n}|}{4\mu_{HC_n}} \tag{6}
\]
where \( \mu_{HC_n} \) represents the mean of the \( HC_{n,i} (1 \leq i \leq 4) \). We have found that if \( \text{SMAD}_n > \gamma \), these four NxN sub-CUs would be quite different in terms of the residual distribution, and smaller CU size will be preferred. Therefore, full RDOQ at current CU level is terminated and the encoding process is directly jumping into the next CU level. Threshold \( \gamma \) is found to be related to the quantization parameter (QP). In our simulations, \( \gamma = 0.3 \) when \( \text{QP} \leq 25 \), \( \gamma = 0.8 \) when \( \text{QP} \geq 25 \). That is because for larger QP values, the difference in residual distribution might be less important to the total R-D cost.

C. Early RDOQ Termination

In HM, \( M \) modes with the least Hadamard costs, are categorized together as \( \Psi \) after the rough mode decision. These \( M \) candidates are further encoded using the RDOQ to derive the best one. In \( \Psi \), modes are ordered with their Hadamard costs from the lowest to the highest, i.e., \( HC(m_i) \leq HC(m_{i+1}) \) \((1 \leq i \leq M)\). The final best mode \( m_{\text{opt}} \) would be chosen from these \( M \) modes for the following RQT process to determined best TU size. Here, we propose an early RDOQ termination scheme based on mode adjacency and Hadamard cost. As shown in [8], the first two mode candidates in \( \Psi \) have the high probability to be selected as the best mode \( m_{\text{opt}} \). Therefore, these two modes always go through the full RDOQ check in our scheme. Moreover, let \( |m_i - m_j| \) to denote the distance between mode \( m_i \) and \( m_j \) [13]. Two modes \( m_i \) and \( m_j \) are considered adjacent if the following conditions are met:

1) Neither \( m_i \) nor \( m_j \) is Planar or DC mode;
2) \(|m_i - m_j| \leq 2\).

If these two conditions are not met, \( m_i \) and \( m_j \) are not considered as the adjacent modes. With these definitions, we define the following early termination scheme:

1) Use set \( S \) to keep all the modes that have gone through the full RDOQ process. Initialize \( S = \{m_1, m_2\} \) and perform full RDOQ for these best two modes.
2) For each of the rest \( M-2 \) modes, say \( m_i \), if it is the adjacent mode to any mode in \( S \) and \( HC(m_i) > \delta HC(m_j) \) RDOQ for this mode would be skipped. Otherwise, it is encoded with RDOQ and added into \( S \). Here \( m_1 \) is the mode with the smallest Hadamard cost and threshold \( \delta \) is set to 1.05 in this paper.

D. Integrated Fast Intra Prediction

In this section, we integrate the above early termination schemes into a full fast intra prediction method, as shown in Figure 3. Such integrated fast algorithm using early terminations are implemented into the HEVC reference software HM7.0 to demonstrate the performance.

IV. EXPERIMENTAL RESULTS

Experimental results for various test sequences are shown in this section. Our proposed fast intra method is compared with the default algorithm in HM7.0, following the common conditions defined in [14]. All intra encoder setting is simulated to demonstrate encoder performance. Class A (4Kx2K), B (1080p), C (WVGA), D (QWVGA) and E (720p) sequences are all used for performance verification. BD-Rate performance and encoder time reduction are shown in Table I. Figure 4 plots rate-distortion curves with four samples. On average, our proposed solution achieves 32% encoding time reduction for all intra coding with 1.1% BD-Rate increase. In addition, we also perform the simulation using the optional Class F screen content sequences, where our proposed fast algorithm provides 36.5% encoding time reduction with 1.4% BD-Rate increase.
Fig. 3: Proposed Fast Intra Prediction Algorithm Using Early Termination Schemes

**TABLE I: Coding efficiency and complexity reduction for proposed fast intra prediction on HM7.0**

<table>
<thead>
<tr>
<th></th>
<th>Luma-Y</th>
<th>Chroma-U</th>
<th>Chroma-V</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic</td>
<td>1.2%</td>
<td>0.8%</td>
<td>0.7%</td>
<td>70%</td>
</tr>
<tr>
<td>PeopleOnStreet</td>
<td>1.3%</td>
<td>1.1%</td>
<td>0.9%</td>
<td>68%</td>
</tr>
<tr>
<td>Nebuta</td>
<td>0.9%</td>
<td>1.1%</td>
<td>1.5%</td>
<td>70%</td>
</tr>
<tr>
<td>SteamLocomotive</td>
<td>1.3%</td>
<td>1.0%</td>
<td>1.4%</td>
<td>61%</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kimono</td>
<td>1.1%</td>
<td>1.3%</td>
<td>1.7%</td>
<td>63%</td>
</tr>
<tr>
<td>ParkScene</td>
<td>0.7%</td>
<td>0.5%</td>
<td>0.6%</td>
<td>70%</td>
</tr>
<tr>
<td>Cactus</td>
<td>1.0%</td>
<td>0.6%</td>
<td>0.5%</td>
<td>72%</td>
</tr>
<tr>
<td>BasketballDrive</td>
<td>1.1%</td>
<td>1.3%</td>
<td>1.4%</td>
<td>67%</td>
</tr>
<tr>
<td>BQTerrace</td>
<td>0.8%</td>
<td>0.9%</td>
<td>1.1%</td>
<td>69%</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BasketballDrill</td>
<td>1.3%</td>
<td>0.9%</td>
<td>0.8%</td>
<td>71%</td>
</tr>
<tr>
<td>BQMall</td>
<td>1.2%</td>
<td>0.6%</td>
<td>0.6%</td>
<td>65%</td>
</tr>
<tr>
<td>PartyScene</td>
<td>0.9%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>74%</td>
</tr>
<tr>
<td>RascalHorses</td>
<td>0.9%</td>
<td>0.7%</td>
<td>0.6%</td>
<td>69%</td>
</tr>
<tr>
<td>D</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BasketballPass</td>
<td>1.3%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>67%</td>
</tr>
<tr>
<td>BQSquare</td>
<td>1.1%</td>
<td>0.4%</td>
<td>0.4%</td>
<td>70%</td>
</tr>
<tr>
<td>BlowingBubbles</td>
<td>1.2%</td>
<td>0.0%</td>
<td>0.0%</td>
<td>73%</td>
</tr>
<tr>
<td>RaceHorses</td>
<td>1.0%</td>
<td>0.5%</td>
<td>0.3%</td>
<td>73%</td>
</tr>
<tr>
<td>E</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FourPeople</td>
<td>1.4%</td>
<td>1.3%</td>
<td>1.2%</td>
<td>62%</td>
</tr>
<tr>
<td>Johnny</td>
<td>1.5%</td>
<td>2.4%</td>
<td>2.3%</td>
<td>61%</td>
</tr>
<tr>
<td>KristenAndSara</td>
<td>1.6%</td>
<td>2.0%</td>
<td>1.8%</td>
<td>60%</td>
</tr>
<tr>
<td>F</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BasketballDrillText</td>
<td>1.2%</td>
<td>0.7%</td>
<td>0.6%</td>
<td>71%</td>
</tr>
<tr>
<td>ChinaSpeed</td>
<td>1.3%</td>
<td>0.8%</td>
<td>0.7%</td>
<td>63%</td>
</tr>
<tr>
<td>SlideEditing</td>
<td>1.1%</td>
<td>0.5%</td>
<td>0.5%</td>
<td>68%</td>
</tr>
<tr>
<td>SlideShow</td>
<td>1.9%</td>
<td>1.7%</td>
<td>1.7%</td>
<td>44%</td>
</tr>
<tr>
<td>Overall</td>
<td>1.1%</td>
<td>0.9%</td>
<td>0.9%</td>
<td>68%</td>
</tr>
</tbody>
</table>

V. CONCLUSION

We propose a novel early-termination based fast intra decision method in this paper. Compared with the reference software HM7.0, the proposed algorithm achieves considerable encoding time reduction with negligible bitrate increment and PSNR loss. Our solution use Hadamard costs to early terminate the CU and TU selection as well as mode selection to avoid computational RDQ process. Hadamard cost can be derived in the rough mode decision process as implemented in current HM software without introducing additional overhead. We expect more encoder speed up when combined with our previous work in [13] or other complementary works in the literature. As the future study, we will investigate how to optimize various parameters used in this paper. Meanwhile, we would like to open our source code at (URL: http://vision.poly.edu/~zma03/opensrc/iscas_sourceHM7.zip) for the community. Any comments, bugfix, and further implementations are more than welcome.

REFERENCES


