早稲田大学大学院情報生産システム研究科

博士論文概要

論文題目

Fast Algorithm and VLSI Architecture of HEVC Mode Decision and Reconstruction Loop Based on Data Reuse and Reordering

申　　請　　者
Heming SUN

情報生産システム工学専攻
高位検証技術研究

2016　年　12　月

With the development of the information society, multimedia contents are widely used. Video data occupies the majority of multimedia data and it will dramatically grow when high definition (HD) and ultra-HD video applications are popularized in the near future. In order to relieve the burden of video storage and transmission, video compression technique has been widely used. By encoding, large raw video data is compressed to small binary data. By decoding, the compressed data is decompressed for display. High efficiency video coding (HEVC) is the latest video compression standard which doubles the compression ratio as its predecessor H.264/AVC. However, to reach such high compression ability, many new coding features are adopted. As a result, the encoding/decoding complexity becomes 5.2x/2.1x higher than H.264. So low-complexity algorithms and architectures for HEVC are extremely desired.

Mode decision and reconstruction loop are two indispensable components in the video coding. Mode decision is used to select the best mode which has the smallest rate-distortion (R-D) cost. After choosing the best mode, the reconstruction loop is conducted to generate the reconstructed pixels for the mode decision afterwards. In the mode decision, the residual of the original and predicted picture passes through forward transform and quantization to reduce the data volume. Rate represents the requiring bits for coding quantized transformed residual and the best mode information. After that, de-quantization and inverse transform are conducted to recover the residual and generate the reconstructed picture. Since quantization is lossy, the reconstructed picture is different from the original picture and distortion is used to reflect the degree of difference.

In HEVC, mode decision and reconstruction loop become much more complex and important due to two reasons. The first reason comes from the adoption of large transform in HEVC. The largest transform size in HEVC is 32×32 which is 16x larger than H.264, which will lead to huge hardware consumption. In the state-of-the-art HEVC intra encoder [Pastuszak, TCSVT 2015], transform consumes about 53% of the overall gate counts. Therefore, the area-efficient designs for transform are highly required. Moreover, due to the large transform size, many high-frequency quantized transformed coefficients will be zero. The processing for the zero elements can be optimized. The second reason is that there are many more modes in HEVC than H.264. For intra prediction, 5 prediction units (PU) and 35 prediction modes are supported. Overall, 175 modes are provided in HEVC. However, there are only 2 PUs and 9 prediction modes for intra prediction in H.264. The complexity of calculating the R-D costs for all the modes is high. Therefore, reducing the number of modes is extremely necessary. In this thesis, low-complexity algorithms and architectures for three research topics of the mode decision and reconstruction loop are proposed. Because the simple parallelization will still suffer from high hardware cost such as large area and power consumption, the acceleration based on data reuse and reordering is studied.

The thesis is composed of five chapters.

In Chapter 1, the video encoding diagram is described at first. And then, the position of mode decision and reconstruction loop in the video coding and their relationship are presented. Finally, the motivations of choosing three research topics are given.

In Chapter 2, an area-efficient architecture for transform is presented. A complete transform is composed of row and column transform which require the logical computational part. In addition, a transpose buffer is required to store the results of row transform. For the logical computational part, Chen's algorithm is adopted so that the symmetric property of the transform matrix can be utilized to reduce the number of multiplications and additions. In addition, a reordered parallel-in serial-out (RPISO) scheme is proposed so that the inputs of the butterfly structure could be reused in each clock cycle. As a result, 25% normalized gate counts are saved compared with [Shen, IEICE 2013]. For the transpose buffer part, static random-access memory (SRAM) instead of register is adopted in order to reduce the area consumption. The storing positions in the SRAM are reordered so that it can achieve 100% I/O utilization of SRAM. In addition, two data mapping methods are proposed so that write and read operation can be operated in parallel. As a result, about 62% area consumption can be reduced compared with the SRAM-based transpose buffer in [Zhu, ISCAS 2013].

In Chapter 3, a low-cost system design of the de-quantization and inverse transform is presented. The system can be used to generate the reconstructed pixels in both encoder and decoder. For the de-quantization, the input coefficients are multiplied with scaling parameters. In order to reduce the number of multiplications, the input coefficients are decomposed to base part (baseLevel) and remaining part. The value of base part is not greater than 3, thus the multiplication of baseLevel and scaling parameter can be replaced by look-up-tables (LUTs). For the remaining part, the number of positions with non-zero value is usually not greater than four in one 4×4 block. Therefore, only four multipliers are provided for processing one 4×4 block. In the system, there are three memory operations: read operation of the buffer between de-quantization and inverse transform, write and read operation of the transpose buffer of inverse transform. A specific path is created to detect zero elements by reusing the pixel data. After the detection, the memory operations are skipped for the zero elements. As a result, for the de-quantization, 77% normalized area consumption is reduced compared with [Tikekar, ICIP 2014]. For the memory parts in the system, 29%-86% power consumption is saved compared with not skipping the memory operations for the zero elements.

In Chapter 4, a prediction unit (PU) depth and prediction mode selection algorithm for intra prediction is proposed. At first, a fast preprocessing stage based on a proposed cost model is presented. After estimating the costs for 8×8 PU, the results are reused to predict the costs for larger PUs. Based on the estimated costs of all the PUs, 2 neighboring PU depths out of 5 are selected to do the R-D cost calculation. Still based on the preprocessing results, a prediction mode selection scheme eliminates the necessity to perform fine Hadamard cost calculation in the original HM. A 32×32 PU compensation scheme is also exploited to alleviate the mismatch problem between proposed simplified cost and R-D cost due to the lack of large transform size in proposed cost model. The compensation scheme is able to effectively improve coding performance for high-resolution sequences. In comparison with HM (version 7.0), the proposed algorithm achieves about 52% encoding time reduction, with the corresponding 1.87% BD-bitrate increment. Compared with [Xiong, ISPACS 2012], the coding efficiency becomes worse by 0.62% in terms of BD-bitrate. However, the encoding time reduction is increased by 14% on average and 23% in the best case. Compared with [Zhang, VCIP 2012], the encoding time is increased by 5%. However, the coding efficiency becomes better by 3.23% in terms of BD-bitrate.

In Chapter 5, the conclusion and future work are given. The methods in Chapter 2 and 3 are designed for mode decision and reconstruction loop and contribute to both intra and inter prediction. The proposals in Chapter 4 intend to reduce the complexity of mode decision for intra prediction. In the future, the other components will be designed and the overall system pipeline schedule will be designed.