Block-Based Parallel Intra Prediction Scheme for HEVC

Jie Jiang
Institute of Intelligence Control and Image Engineering, Xidian University, Xi'an, China
Email: supergirl.jj@gmail.com

Baolong Guo and Wei Mo
Institute of Intelligence Control and Image Engineering, Xidian University, Xi'an, China
Email: {blguo, wmo}@xidian.edu.cn

Kefeng Fan
Guilin University of Electronic Technology, Guilin 541004, China
Email：fankf@cesi.ac.cn

***Abstract* — Advanced video coding standards have become widely deployed in numerous products, such as multimedia service, broadcasting, mobile television, video conferences, surveillance systems and so on. New compression techniques are gradually included in video coding standards so that a 50% compression rate reduction is achievable every ten years. However, dramatically increased computational complexity is one of the many problems brought by the trend. With recent advancement of VLSI (the Very Large Scale Integration) semiconductor technology contributing to the emerging digital multimedia word, this paper intends to investigate efficient parallel architecture for the emerging high efficiency video coding (HEVC) standard to speed up the intra coding process, without any prediction modes ignored. Parallelism is achieved by limiting the reference pixels of the 4 × 4 subblocks, allowing the subblocks to use different direction modes to predict the residuals. Experimental implementations of the proposed algorithm are demonstrated by using a set of video test sequences that are widely used and freely available. The results show that the proposed algorithm can achieve a satisfying intra parallelism without any significant performance lose.**

***Index Terms* — HEVC, intra coding, parallel architecture, multiple directions.**

I. INTRODUCTION

Continuous emergence of video coding standards and the growth in development and implementation technology for them have undoubtedly created a completely new world of multimedia. So far, contributions to video coding technology have mainly focused on improving coding efficiency. The challenges remain: not only to find efficient coding algorithms which require high performance but also to speed up the coding process.

The ongoing video coding standard, High Efficiency Video Coding (HEVC) [1], is getting more attention due to its high compression efficiency. However, the computational complexity of HEVC would be 2-10 times higher than its counterpart, which is considered an obstacle to implement it in real-time. Therefore, many research works focus on how to reduce the computational complexity.

The purpose of these works is to design and evaluate the performance of new methods to reduce encoder complexity, while keeping the quality of reconstructed video sequences for intra coding. The works generally fall into two categories.

1. Fast mode decision approaches with early termination using adaptive thresholds or optimized Lagrangian rate distortion optimization (RDO) function [2-4].

2. Parallel architectures to speed up the intra prediction process [5-14].

With recent advancement of VLSI (the Very Large Scale Integration) semiconductor technology contributing to the emerging digital multimedia word, research on parallel architectures gets more attention. In this paper, we focus on the second case, and present a block based parallel architecture to speed up the intra prediction for HEVC.

The remaining parts of this paper are organized as follows: Section II reviews the state of art within the field of parallel architectures. Section III introduces the spatial prediction in HEVC. Section IV presents the proposed scheme, including 2X parallel intra prediction and its expansion to 4X parallelism. Experimental results are presented within Section V. Finally, we conclude this paper in section VI.

II. RELATED WORK

The main image and intra frame of video compression extensively adopts the block-based structure from prediction and transform to entropy coding, where the coding of one block is dependent on the availability of its left, upper-left, and upper-right blocks. Such a highly dependent structure is not quite suitable for parallelization, especially for ASIC (Application Specific Integrated Circuit) solutions. Even so, when dual-core and quad-core computers are available, there are still many efforts on parallelizing the encoding and decoding from different aspects, as described below.

1. GOP (Group of Pictures) approaches: Barbosa [5] and Vander [6] propose to partition a sequence into some GOPs. The correlation between GOPs is low, and it can not only limit error propagation, but also support parallel coding processing. However, it needs to get the data of all the pictures in a GOP before parallelism. When the GOP has too

doi:10.4304/jmm.7.4.289-294

many pictures, it will lead to serious delay, which is not convenient for the real-time video coding applications.

2. Frame approaches: Chen [7] proposes to realize the parallelism coding at the frame level. This approaches is limited to the correlation between frames, therefore, the speedup cannot be linearly increased corresponding to the number of process cores.

3. Pipeline approaches: Gulati [8] and Klaus [9] propose to organize the prediction, transform, and entropy coding of macroblocks (MBs) as a pipeline and assign them to multiple cores for parallel computing. This class of approaches can achieve limited parallelisms if workloads are unbalanced at different cores. Two times speedup is reported for high definition sequences on general-purpose quad-core computers in [9].

4. Slice partitioning (SP) approaches: Rodriguez [10] proposes to partition an image into some regions that are referred to as slices. The coding of slices can be carried out independently by different cores. They provide good parallelism but would result in a significant loss on coding performance if there are too many slices.

5. MB-reordering (MR) approaches: Despite these efforts on parallelizing existing coding algorithms, due to strong dependency among blocks in intra prediction and the filtering process of top/left reconstructed samples, the intra luma prediction process of small blocks is a challenge for parallelization. The original process handles blocks in serial, which is not efficient. Efficient architectures have been reported in [11-14], which propose to process MBs in the wave front order so that MBs in each diagonal line can be coded concurrently when neighboring MBs are available. Owing to the fine-granularity parallelism at the MB level, these approaches can achieve good parallelism and are more widely used at present.

Huang's work [11] has bubbles between Intra 4 × 4 predictions because of the low throughput of reconstruction process so that the prediction has to wait for the completion of reconstruction. Lee's work [12] perfectly pipelines the intra prediction and reconstruction process; however, it requires that both intra prediction and reconstruction have exact equal processing cycles. It also reduces some prediction modes in some blocks in order to enforce pipelining; hence, the video quality is degraded. Jin's work [13] proposes both partially and fully pipelined architectures for intra 4 × 4 prediction and has the same drawback as the approach in [12]. Moreover, the architectures add dependency graph process in order to improve gains; however, this increases hardware overhead. It takes 25 cycles to process each block, which is too long for high throughput reconstruction. Similar to Huang's work, Suh's work [14] proposes an efficient parallel architecture followed by a redundancy reduction algorithm to speed up the intra 4 × 4 prediction. However, the approaches above all have drawbacks either with the pipelining architecture or in compression gains.

In this paper, we propose a block-based intra prediction parallel algorithm to solve the problem above.

III. SPATIAL PREDICTION IN HEVC

ISO-IEC/MPEG and ITU-T/VCEG recently formed the joint collaborative team on video coding (JCT-VC). The JCT-VC aims to develop the next-generation video coding standard, called high efficiency video coding (HEVC). A joint proposal to the high efficiency video coding standardization effort have been partially adopted into the HEVC Test Model (HM). The major improvements for intra coding come from two tools described below.

A. Intra Prediction

The nine intra prediction modes supported in H.264/AVC with different directionalities is not flexible enough to represent complex structures or image segments with different directionalities. HEVC extends the set of directional prediction modes of H.264/AVC, providing increased flexibility and more accurate predictions for the sample values. The increased prediction accuracy provides significant reductions in residual energy of the intra coded blocks and improvements in coding efficiency [15].

It can support from 16 to 34 prediction directions for different prediction size, and the prediction can achieve 1/8 pixel accuracy. Table I shows a 64 × 64 prediction unit contains different subblock size and its corresponding prediction modes. The 34 directions of the intra prediction modes are illustrated in Fig. 1.

TABLE I

A 64 ×64 PREDICTION UNIT CONTAINS DIFFERENT SUB-BLOCK SIZE AND ITS CORRESPONDING PREDICTION MODES

Amount	Size	Modes
256	4×4	17
64	8×8	34
16	16×16	34
4	32×32	34
1	64×64	5

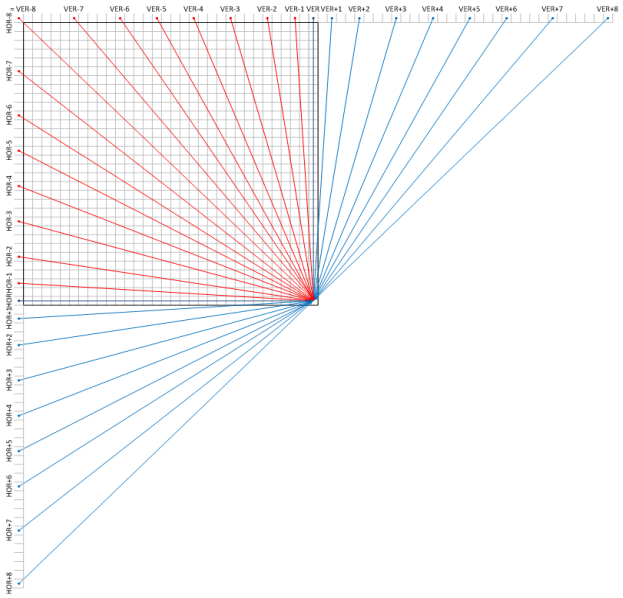

Figure 1. 34 Prediction modes for HEVC

It improves the coding efficiency significantly through permitting block prediction in an arbitrary direction by indicating the prediction angle.

B. Quadtree-based coding structure

Coding efficiency can be significantly improved by utilizing macroblock structures with sizes larger than 16 × 16 pixels, especially at sequences with high resolutions [16]. To achieve a more flexible coding scheme, HEVC utilizes a quadtree-based coding structure [17] with support for macroblocks of size 64 × 64, 32 × 32, 16 × 16, 8 × 8, and 4 × 4 pixels. HEVC separately defines three block concepts: coding unit (CU), prediction unit (PU) and transform unit (TU). After the size of largest coding unit (LCU) and the hierarchical depth of CU have been defined, the overall structure of codec is characterized by the various sizes of CU, PU and TU in a recursive manner. This allows the codec to be readily adapted for various kinds of content, applications, or devices that have different capabilities.

The CU splitting process is described in Fig. 2, using a Split flag to ensure whether the current CU needs to split into a smaller size or not.

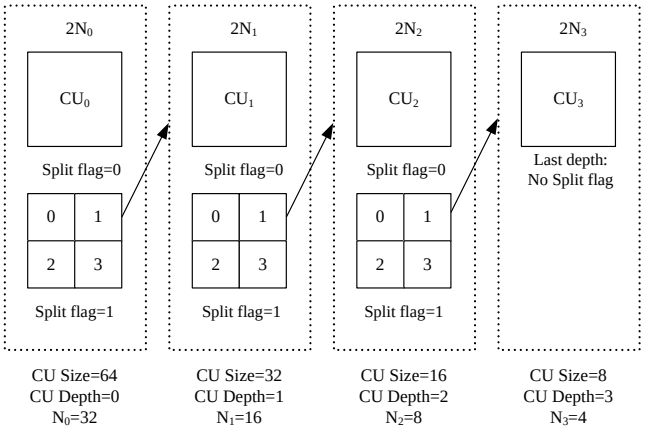

Figure 2. The CU splitting process

The coding efficiency improvements are more visible at higher resolutions, where bit rate reductions reach around 50% and 35% for the low delay and random access experiments, respectively.

C. Observation and Motivation

Intra prediction is currently achieved by partitioning a largest coding unit (LCU) into one or more blocks through a recursive splitting process. Each block is predicted spatially and subsequently refined, and the prediction is performed sequentially using neighboring reconstructed blocks. The prediction process requires that the causal neighbors of the current block must be completely reconstructed before processing the current block. It results in a set of serial dependencies, and these serial dependencies result in significant complexity for both the encoder and decoder processes.

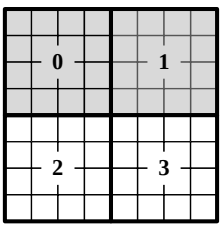

Figure 3. An 8 ×8 block with four 4 ×4 blocks

Take an 8 ×8 block with four 4 ×4 blocks as an example in Figure 3, where block0 is the first 4 ×4 block in decoding order, Block1 is the second one and so forth. Intra predictors for block0 are computed from the top and left neighboring MBs. As one can see, there is a strong dependency (serialization) in computation of intra predictors for general 4 × 4 intra mode. Indeed, the block1 depends on the reconstruction samples of block0. The block2 in turn depends on the reconstructed samples block0 and block1 and its left neighbor block. The block3 depends on the reconstructed samples of block0, block1 and block2. This dependency makes a parallelization of intra prediction to be challenging especially for ASIC implementations. Therefore, it is necessary to minimize dependencies and improve performance respectively.

IV. Block Based Parallel Intra Prediction

In HEVC, for the inter prediction has greater complexity, however, it can realize motion estimation without referring to the neighbor blocks, the parallelism can be easily realized, While the intra prediction process requires that the causal neighbors of the current block must be completely reconstructed before processing the current block. It results in a set of serial dependencies, which makes a parallelization of intra prediction to be a challenge, especially for the 4 ×4 intra prediction.

A. 2X parallel intra prediction

The parallel intra-prediction approach divides a coding unit into two partitions: all blocks in the first partition are predicted without reference to each other; similarly, all blocks in the second partition are predicted with reference to the first partition but also without reference to each other. It consists of three steps:

1. Partitioning the coding units into two sets
2. Predicting the first set blocks (shaded blocks in Fig. 3) in parallel using the already reconstructed block neighbors
3. Predicting the second set blocks (white blocks in Fig. 3) in parallel using already reconstructed block neighbors and also the neighbors from the first set blocks

Take an 8 ×8 block with four 4 ×4 blocks in Fig. 3 as an example, block0 and block1 are in the first set, while block2 and block3 are in the second set. With the partition above, then we define that all blocks in a single partition be processed in parallel. It means that the blocks in the first partition are predicted without reference to each other (or blocks in the second partition). Blocks in the second partition are also predicted without reference to other blocks in the second partition. In all cases, a block could use the pixel values from neighboring coding unit for intra prediction.

Fig. 4 shows the difference between HEVC and the proposed scheme. To predict the first set of blocks, we use pixel values from the upper and left coding unit's boundaries. For example, to predict block1 in Fig. 4 B, we used reconstructed pixel values from the dark gray pixels. Its left reference pixels are different from that in HEVC showed in Fig. 4 A. It utilizes the already reconstructed pixel values of the left coding unit's boundaries. Therefore, block1 and block0 can be predicted in parallel.

Following the prediction of first set of blocks, we then add any transmitted residual to the predicted value to generate a reconstructed first set of blocks. After

reconstructing the first set of blocks, we proceed to predict the second set of blocks. As mentioned before, these blocks are predicted in parallel and use the reconstructed pixel values from blocks in the first partition and the left coding unit's boundaries, as showed in Fig. 4 D.

In the proposed scheme, none of the prediction modes are ignored. That means it support all the prediction directions mentioned in Part A, Section III. Therefore, it can realize the parallism without significant performance loss.

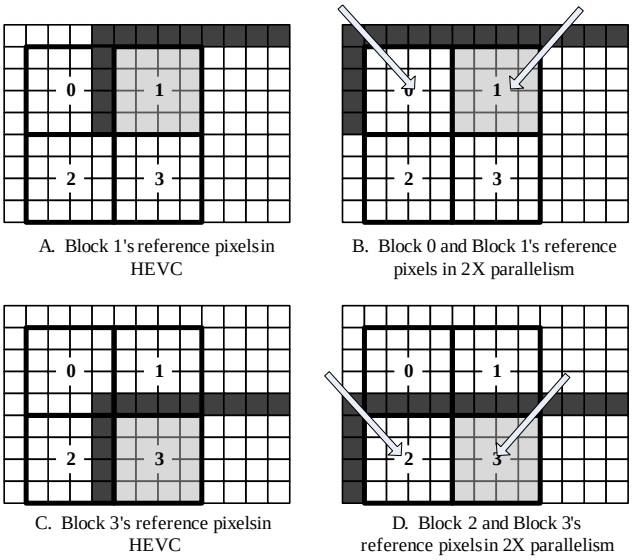

Figure 4. The blocks' reference pixels

By employing the partitioning strategy, we can achieve a direct increase in parallelism. This parallelism is shown graphically in Fig. 5. The 4 × 4 intra prediction within an 8 × 8 coding unit requires 4 sequential steps while our partitioning approach results in only two sequential steps. This is an increase in parallelism by a factor of 2X. In general, the increase in parallelism is N/2, where N is the number of blocks within the macro-block.

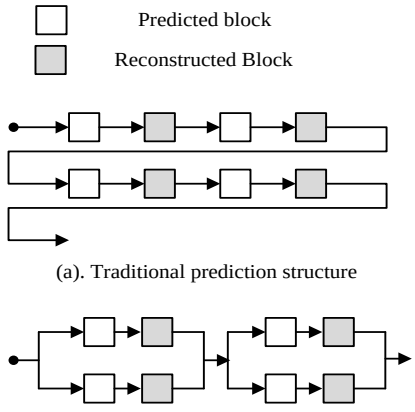

Figure 5. Processing order for the 4×4 blocks within an 8×8 intra predicted unit

B. 4X parallel intra prediction

The method can also be extended the parallelism to be 4X. In this extension, all of four blocks are treated as the first set blocks. They are restricted to have the same prediction mode, and are simply predicted using their 8 × 8 block neighbors, as shown in Fig. 6. However, each 4 × 4 block can have its own predict mode, residual and transform. In this case, the distances between the referenced pixels and predicting pixels increase, therefore it will lead to a bit rate increase, as will be seen in the results later.

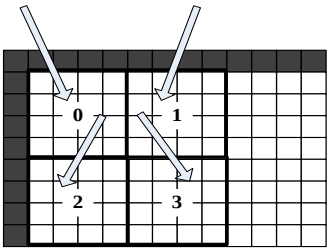

Figure 6. The blocks' reference pixels for 4X parallelism

V. EXPERIMENTAL RESULTS

The parallel intra prediction technology has been integrated into the HEVC reference software HM3.0 [18] and tested the recommended configuration on 18 sequences listed in Table II.

The test sequences include two 4K sequences, five 1080p sequences, four WVGA sequences, four WQVGA sequences, and three 720P sequences, which are widely used in various applications.

We use two common test conditions described in [19], Intra (high efficiency condition) and Intra LoCo (low complexity condition).

The common test conditions are set as follows.

1) 10 seconds length are encoded as intra frames for each sequence.

2) The QP is set at 22, 27, 32, and 37.

3) In-loop deblocking filter and RDO are enabled.

TABLE II

COMPARISON OF THE PERFORMANCE FOR 1DDCT AND 2DDCT

Class A 4K	Class B 1080p	Class C WVGA	Class D WQVGA	Class E 720P
Traffic	Kimono	BasketballDrill	BasketballPass	Vidyo1
PeopleOn Street	ParkScene	BQMall	BQSquare	Vidyo3
	Cactus	PartyScene	BlowingBubbles	Vidyo4
	BasketballDrive	RaceHorses	RaceHorses	
	BQTerrace			

Table III below shows the summary results of 2X parallel intra prediction vs. HM3.0. Table IV shows summary results of 4X parallel intra prediction vs. HM3.0. Y BD-rate means the BD-rate increase for lumu component, while U BD-rate and V BD-rate for the chroma component. Compare Table III with Table IV, we can find that the coding performance of 4X parallism is not as good as 2X parallism. For the 4X parallism case, the distances between the referenced pixels and predicting pixels increase, and that reduces the prediction accuracy, which results in the performance lose.

TABLE III

RD RESULTS OF 2X PARALLEL INTRA PREDICTION WITH VS. HM3.0

	Intra			Intra LoCo		
	Y BD-rate %	U BD-rate %	V BD-rate %	Y BD-rate %	U BD-rate %	V BD-rate %
Class A	0.6	0.1	0	0.5	0.4	0.4
Class B	0.7	0.6	-0.3	0.7	1	0.5
Class C	1.5	1.1	0.8	1.5	1.5	1.3
Class D	1.8	1.4	0.6	1.9	0.6	2.3
Class E	0.7	1.2	-0.5	1	0.8	0.9
All	**1.1**	**0.8**	**0.2**	**1.1**	**0.9**	**1.1**

TABLE IV

RD RESULTS OF 4X PARALLEL INTRA PREDICTION WITH VS. HM3.0

	Intra			Intra LoCo		
	Y BD-rate %	U BD-rate %	V BD-rate%	Y BD-rate%	U BD-rate %	V BD-rate %
Class A	1.1	-0.6	-0.7	1.1	0.3	0.9
Class B	1.2	0.6	0	1.3	0.8	1.1
Class C	2.7	1.9	1	3.3	2.5	2.3
Class D	3.5	1.4	1.3	3.5	2.5	2.1
Class E	1.6	1	0.7	2.1	1.8	0.5
All	**2**	**0.9**	**0.4**	**2.2**	**1.5**	**1.4**

Two example of RD curves are given in Fig. 7 (BasketballPass, LoCo) and Fig. 8 (RaceHorses, HE), showing the proposed performance at four different bit rates. It is observed that the proposed method can achieve the parallism without significant RD performance lose in both low bit rate and high bit rate.

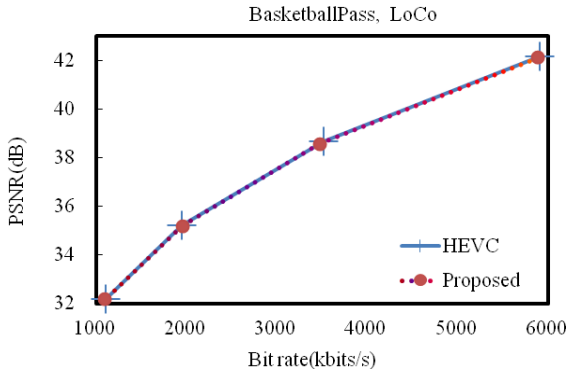

Figure 7. The comparison of RD performance (BasketballPass, LoCo)

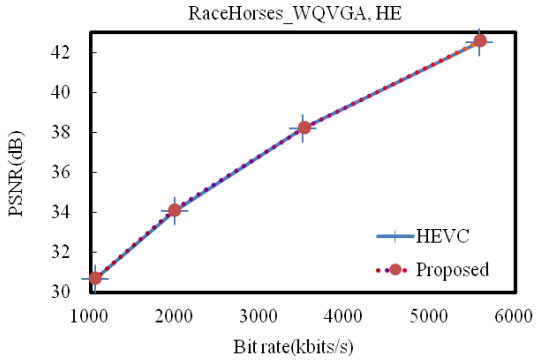

Figure 8. The comparison of RD performance (RaceHorses, HE)

VI. CONCLUSIONS

Due to the various special applications of intra coding frame such as random access point and refresh synchronous frame, intra coding is an important part of video coding. For the intra coding blocks are predicated dependent on the availability of the adjacent blocks in the former rows and columns, so the parallel coding or decoding process cannot be realized among the intra coding blocks, which increases the difficulties for the implementation of real-time coding in high definition and ultra high definition applications. Therefore, how to increase the capacity of parallel computing effectively has considerable value.

In this paper, a parallel prediction scheme for HEVC is proposed. The parallel prediction unit supports the parallelization of intra-coded blocks within the current coding unit in a two-step parallel process. Parallelism is achieved by limiting the reference pixels of the 4 × 4 subblocks, without any prediction modes ignored. Experimental results show that the increased parallelization comes with small losses in coding efficiency. For example, about 1% for HD sequences of 2X parallelism. We assert that this loss is negligible and well justified by the parallelization capability.

We described 2X and 4X parallelism in detail, and in fact, the parallelism can also be expanded to higher parallelism. It can flexibly fit for the various applications.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (No. 60802077; No. 61003196; No. 61172053)

REFERENCES

[1] K. Ugur, K. Andersson and A. Fuldseth et al. “High Performance, Low Complexity Video Coding and the Emerging HEVC Standard”, IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(12), pp: 1688-1697.

[2] G. Sullivan, P. Topiwala, and A. Luthra, "Fast 4x4 Intra prediction Based on The Most Probable Mode in H.264/AVC," IEICE Electronics Express, 2008, 5(19), pp. 782-788.

[3] Yinji Piao, Junghye Min, and Jianle Chen et al, "Encoder improvement of unified intra prediction," JCTVC-C207, Guangzhou, Oct. 2010.
[4] Liang Zhao, Li Zhang, and Xin Zhao, "Further Encoder Improvement of intra mode decision," JCTVC-D283, Daegu, Jan. 2011.
[5] Denilson M.Barbosa, Joao Paulo Kitajama and Wagner Meira JR, "Parallelizing MPEG Video Encoding using Multiprocessors", Proceedings of the XII Brazilian Symposium on Computer Graphics and Image Processing, 1999, pp: 215-222.
[6] E. B. vander, E. G. T. Jaspers, R. H. Gelderblom, "Mapping of H.264 Decoding on a Multiprocessor Architecture", SPIE Conf. on Image and Video Communications and Processing, 2003. 5(7), pp: 707-718.
[7] Y. Chen, E. Li and X. Zhou, "Implementation of H.264 encoder and decoder on personal computers", J. Vis. Commun. Image Representation, 2006, 17(2), pp: 509–532.
[8] A. Gulati and G. Campbell, "Efficient mapping of the H.264 encoding algorithm onto multi-processor DSPs", in Proc. SPIE - IS&T Electronic Imaging, 2005, pp: 94–103.
[9] O. L. Klaus Schoffman, M. Fauster, and L. Boszormeny, "An evaluation of parallelization concepts for baseline profile compliant H.264/AVC decoders", in Proc. Euro-Par—Parallel Processing, 2007, Lecture Notes in Computer Science, pp: 782–791.
[10] A. Rodriguez, A. Gonzalez and M. P. Malumbres, "Hierarchical parallelization of an H.264/AVC video encoder", in Proc. Int. Symp. Parallel Comput. Elect. Eng., 2006, pp: 363–368.
[11] Y. W. Huang, B. Y. Hsieh, and T. C. Chen, "Analysis Fast Algorithm and VLSI Architecture Design for H.264/AVC Intra Frame Coder," IEEE Trans. Circuit and Systems for Video Technology, 2005, 15(3), pp. 378-401, Mar. 2005.
[12] W. Lee, S. Lee, and J. Kim, "Pipelined Intra Prediction Using Shuffled Encoding Order for H.264/AVC," TENCON 2006, pp. 14-17, Nov. 2006.
[13] G. Jin and H. J. Lee, "A Parallel and Pipelined Execution of H.264/AVC Intra Prediction," IEEE International Conference on Computer and Information Technology, CIT'06, pp. 246-250, Sep. 2006.
[14] K. Suh, S. Park and H. Cho, "An Efficient Hardware Architecture of Intra Prediction and TQ/IQIT Module for H.264 Encoder," ETRI Journal, vol. 27, no. 5, pp. 511-524, Oct. 2005.
[15] K. McCann, W.-J. Han, and I.-K. Kim, et al, "Video Coding Technology Proposal by Samsung (and BBC)". JCTVC-A124, Dresden, Germany, Apr. 2010.
[16] S. Ma and C.-C. J. Kuo, "High-definition video coding with super-macroblocks," Proc. SPIE, vol. 6508, part 1, 650816, Jan.2007.
[17] F. Bossen, V. Drugeon and E. Francois, "Video Coding Using a Simplified Block Structure and Advanced Coding Techniques", IEEE Transactions on Circuits and Systems for Video Technology, 2010, 20(12), pp: 1667-1675.
[18] HEVC Reference Software HM-3.0[CP]. https://hevc.hhi.fraunhofer.de/svn/svn_TMuCSoftware/tags/HM-3.0.
[19] F. Bossen, "Common test conditions and software reference configurations," JCTVC-E700, Geneva, Switzerland, Mar. 2011.

Jie Jiang received the B.S. and M.S. degree both in signal and information processing from Xidian University, Xi'an, China, in 2005 and 2008, respectively. She is currently working toward the Ph.D. degree with the Institute of Intelligent Control and Image Engineering (ICIE), Xidian University. Her research interests include video coding and communication.

Baolong Guo received his B.S., M. S. and Ph.D. degrees from Xidian University in 1984, 1988 and 1995, respectively, all in communication and electronic system. From 1998 to 1999, he was a visiting scientist at Doshisha University, Japan. He is currently a full professor with the Institute of Intelligent Control and Image Engineering (ICIE) at Xidian University. His research interests include intelligent information processing, image processing and communication.

Wei Mo was born in Guangxi Province, China, in 1956. He is a professor, Ph.D. supervisor at Xidian University. His research interests include multimedia coding and intelligent information processing.

Fan Kefeng completed his Postdoctoral Research in the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include 3D TV, Smart TV, DRM, digital interface, et al.