

引用格式: 贺飞龙, 蒋林, 刘新闯, 等. 一种提前终止单元划分的帧内预测动态可重构实现[J]. 微电子学与计算机, 2020, 37(2): 14-19. [HE F L, JIANG L, LIU X C, et al. An intra-prediction dynamic reconfigurable implementation of early termination unit partitioning[J]. Microelectronics & Computer, 2020, 37(2): 14-19.]

一种提前终止单元划分的帧内预测动态可重构实现

贺飞龙¹, 蒋林², 刘新闯³, 山蕊³, 王昱¹, 吴皓月³

(1 西安邮电大学 计算机学院, 陕西 西安 710121; 2 西安科技大学 集成电路实验室, 陕西 西安 710054;

3 西安邮电大学 电子工程学院, 陕西 西安 710121)

摘 要: 针对专用硬件实现高效视频编码(High Efficiency Video Coding, HEVC)帧内预测算法资源占用大, 且硬件资源不能重复利用、灵活性差的问题. 提出一种可重构的视频阵列处理器, 能够根据当前视频序列的特点进行帧内预测算法的动态映射. 首先, 分析 HEVC 帧内预测算法的特点和重构的可行性, 以提前终止编码块划分的阈值作为处理器进行硬件重构的依据. 其次, 以计算出来的参数驱动可重构阵列处理器进行硬件重构. 最后, 在重构的阵列处理器上进行帧内预测算法映射. 通过在 4×4 的可重构阵列上进行 Planar 和 DC 两种预测模式实现, 结果表明: 与专用硬件实现方法相比资源减少了 65%, 与多核处理器实现方法相比延时降低了 32%.

关键词: 高效视频编码; 帧内预测; 可重构; 阵列处理器

中图分类号: TP302

文献标识码: A

文章编号: 1000-7180(2020)02-0014-06

An intra-prediction dynamic reconfigurable implementation of early termination unit partitioning

HE Fei-long¹, JIANG Lin², LIU Xin-chuang³,
SHAN Rui³, WANG Yu¹, WU Hao-yue³

(1 College of Computer Science, Xi'an University of Posts and Telecommunications, Xi'an 710121, China;

2 Laboratory of Integrated circuit, Xi'an University of Science and Technology, Xi'an 710054, China;

3 College of Electronic Engineering, Xi'an University of Posts and Telecommunications, Xi'an 710121, China)

Abstract: The High Efficiency Video Coding (HEVC) intra prediction algorithm for the dedicated hardware has a large resource occupation, and the hardware resources cannot be reused and the flexibility is poor. A reconfigurable video array processor is proposed, which can dynamically map the intra prediction algorithm according to the characteristics of the current video sequence. Firstly, the characteristics of HEVC intra prediction algorithm and the feasibility of reconstruction are analyzed. The threshold of early termination of coding block partition is determined as the basis for processor hardware reconstruction. Second, the reconfigurable array processor is driven by the calculated parameters for hardware reconstruction. Finally, intra prediction algorithm mapping is performed on the reconstructed array processor. By performing Planar and DC prediction mode experiments on a 4×4 reconfigurable array, the results show that the resource is reduced by 65% compared with the dedicated hardware implementation method, and the latency is reduced by 32% compared with the multi-core processor implementation.

Key words: high efficiency video coding; intra prediction; reconfigurable; array processor

收稿日期: 2019-05-23; 修回日期: 2019-07-06

基金项目: 国家自然科学基金资助项目(61834005, 61772417, 61802304, 61602377, 61874087, 61634004); 陕西省科技统筹创新工程项目(2016KTZDGY02-04-02); 陕西省重点研发计划(2017GY-060); 陕西国际科技合作计划(NO. 2018KW-006)

1 引言

HEVC 视频标准采用了 H. 264 的基于块的混合编码思想,摒弃了固定大小的宏块,采用灵活性较高的四叉树递归划分方式,增加了多角度帧内预测模式,以提升预测的精细度。

灵活的编码块、丰富的预测模式在提升编码质量的同时势必会带来高编码复杂度、高功耗。如何权衡编码复杂度和编码效率就显得特别重要。当前在对编码质量影响不大的情况下,简化算法过程,概括起来有三个方向:1)提前终止编码块的划分或者预判编码块类型^[1];2)预测模式的裁剪与算法的优化^[2];3)率失真函数的优化。优化统一后的帧内预测算法的计算量仍然巨大,若单从算法层面优化根本无法满足实时视频编码的需求。文献[3][4]在优化了的编码算法基础上用 FPGA 进行加速。这样做虽然可以在确保编码质量的基础上有效缩减编码时间,但其占用较大的硬件资源和专用的 DSP。文献[5]基于多核处理器实现解码器,虽然有效利用了多核的并行性,但其使用了多达 64 个处理元,硬件成本较高。

为了充分提高资源利用率,缩减帧内预测硬件消耗,以解决帧内预测算法所占资源较大,且硬件编程灵活性差的问题^[6-9]。本文提出一种可重构的视频阵列处理器,充分利用可重构阵列处理器动态配置的特点,实现帧内预测算法的动态映射。

2 帧内预测算法及重构可行性分析

HEVC 使用编码单元 CU 作为基本的编码块,最大的编码单元(Largest Coding Unit,LCU)为 64×64 ,采用递归分割四叉树的方法继续分割成 32×32 、 16×16 、 8×8 的块大小,这种从最大块到最小块递归的划分方式虽然提高了预测的精细度和准确度。但是,在率失真代价计算时,同样也提升了从最小块到最大块递归计算的复杂度。再加上丰富的帧内预测模式,那么对于整个帧内编码块的划分,计算量非常巨大。文献[3][4]通过设定阈值提前终止或跳过编码块的递归划分,在编码质量损失不大情况下,有效缩短了编码时间。

在编码块类型确定以后,帧内预测会根据其类型(64×64 、 32×32 、 16×16 、 8×8)决定 PU 块的类型。由于帧内预测块的划分形式只有 $2N \times 2N$ 、 $N \times N$ 两种,且 $N \times N$ 的划分只有在当前编码块为 8×8 时才进行,所以,如果当前 CU 块为 64×64 、 $32 \times$

32 、 16×16 时,当前 PU 块为相同的块大小。而当 CU 块为 8×8 时,PU 块要通过代价函数的计算,确定是一个 8×8 的块,还是 4 个 4×4 的块。

提前终止或跳过编码块的方法确实有效缩短了遍历 LCU 到最小编码单元(Smallest Coding Unit, SCU)的所用的时间,但 HEVC 计算的复杂度,除了灵活的编码块带来的,还有丰富的帧内预测模式,遍历所有块大小的同时,还要遍历 35 种预测模式。文献[7][8]提出了根据视频序列的特点对帧内预测模式进行裁剪的方法,既能保证一定的编码质量,又缩短了编码时间。

可重构的体系结构针对不同应用的适应性特点,既具有通用处理器的灵活性,又具有专用硬件的高性能^[10],是实现类似视频算法这种灵活性和高性能统一的合适平台。可重构体系结构根据其重新配置的粒度可以划分为细粒度、粗粒度和混合粒度。现场可编程门阵列(Field Programmable Gate Array, FPGA)属于细粒度可重构体系结构,其位级的配置结构,带来巨大的配置信息,配置时间较长,不利于处理实时应用。粗粒度的可重构体系结构(Coarse-grained reconfigurable architecture, CGRA)以功能部件作为基本配置单元相较 FPGA 有较少的配置信息,能保证硬件结构的快速改变。在对编码算法优化的基础上,结合 CGRA 的特点,提出一种类 CGRA 的可重构阵列处理器,对帧内预测算法进行重构实现,可以有效减少硬件资源的消耗。

通过对 HEVC 帧内预测算法和可重构处理器的特点进行分析,结合目前已提出的优化方法得出,可以像优化算法一样,在硬件实现帧内预测算法时,并不需要把所有的帧内预测模式和不同块都进行实现。将 HEVC 算法在可重构阵列处理器上实现是解决灵活性和资源利用率问题的一种方法,关键问题是根据需要建立相应的性能模型以驱动阵列处理器进行重构。通过分析,从 CU 块的划分到帧内预测算法的实现这一过程是适合在可重构视频阵列处理器上实现的。

3 动态可重构处理器结构

为了满足帧内预测算法不同预测模式计算的需求,提出了一种可重构的视频阵列处理器,以 4×4 规模大小的阵列结构为例如图 1 所示。主要部件包括 16 个同构的处理单元(Processing Element, PE),指令、数据存储单元,主控单元,其中,主控单元通过 H 型网络^[11]完成对 PE 阵列的控制。

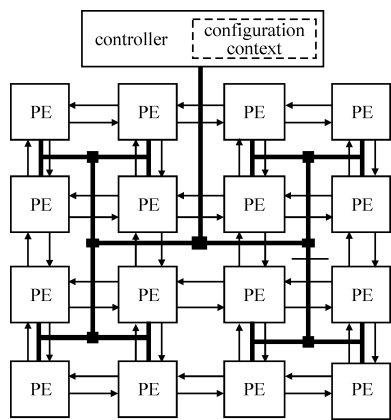


图 1 4×4 规模的可重构阵列处理器

3.1 灵活的数据交互方式

如图 2 所示,具体说明了相邻 PE 数据交互的方式. 相邻 PE 通过四个方向的邻接互连共享寄存器进行数据的交互. 同时,为了避免处理元同一时刻收发数据的冲突,设置了两组共享寄存器. 而对于不相邻的处理元之间,采用行列交叉的方式实现统一编址的存储器,进行数据传输. 以两种方式实现灵活的数据交互方式,以满足不同帧内预测模式数据传输的需要.

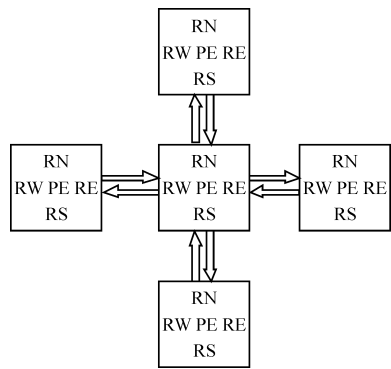


图 2 PE 邻间互连图

3.2 可配置的阵列规模

对于不同的帧内预测模式,所需 PE 阵列的规模是不同的,如果采用不可伸缩的阵列结构,势必在某些算法中会带来 PE 资源的浪费. 为了提高 PE 的利用率,阵列规模必须是可控的. 全局控制器通过 H 型网络(如图 1 粗线所示)对整个 PE 阵列进行控制,通过下发特殊指令,完成对整个 PE 阵列规模的重构,可以是 $M \times N$ ($M, N=1, 2, 3, 4$) 任意规模大小的阵列结构,以满足不同帧内预测模式对阵列结构的需求. 所需 PE 个数除了由帧内预测模式决定,还与硬件资源约束有关. 在硬件资源不受约束或约束不大的情况下,可以通过配置较大规模的 PE 阵

列来提升算法的执行速度. 相反,在硬件资源受限的情况下,通过合理规划 PE 阵列规模和 PE 功能,以达到资源占用和执行速度的折中.

3.3 配置信息的快速下发

为了满足动态重构的需求,全局控制器能够在算法执行的任意时刻通过 H 型网络向任意 PE 下发配置文本,且下发时间只需两个时钟周期. 在 PE 阵列计算的过程中,为了实现对阵列执行状态的监控,全局控制器可以通过反馈网络对每个 PE 的执行状态进行采集.

4 帧内预测重构实现

4.1 帧内预测重构过程

基于第二部分介绍的可重构视频处理器,可以实现下面的重构过程:首先计算出当前 CU 块的方差,与设定的阈值进行比较,如果大于该阈值,表示该 CU 块要继续递归划分,切换到其子块继续进行方差的计算,直至方差小于阈值或 CU 块递归到 8×8 的块大小;然后确定停止划分时 CU 块的类型,若不是 8×8 的块,控制器通过 H 型下发网络直接下发相应的配置信息对阵列进行当前 CU 块大小的帧内预测的配置,如是 8×8 的块,配置阵列进行 8×8 和 4×4 帧内预测,最后以绝对误差和(Sum of Absolute Difference, SAD)作为代价函数,确定是以 8×8 的块还是 4 个 4×4 的块. 执行流程如图 3 所示:

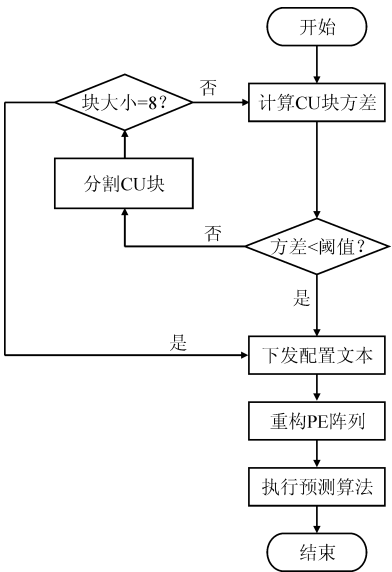


图 3 帧内预测重构流程图

4.2 CU 块阈值计算过程

如图 4 所示为方差计算映射图. 公式(1)为其计

算公式. 为了提高运算速度, 方差计算 PE 阵列结构以功能部件流水的形式进行组织. 第一列 PE 被配置成取数 PE, 然后, 将取出的数通过共享寄存器传给第二列的 PE, 在第二列 PE 接到传来的数后, 第一列 PE 继续进行下一轮取数, 而第二列 PE 进行上一轮的求和运算, 以类似的方法, 完成第三列的求平均值和第四列的求方差. 以 16×16 CU 块方差计算过程为例, 详细介绍其计算过程. 为了充分利用多核的并行性, 将一个 16×16 CU 块分割成 4 个 8×8 的 CU 块, 依次定义成 CU0、CU1、CU2、CU3, 每个块的处理过程映射到每一行. 第二列 PE 完成各自子块的求和运算, 第四行的 ADD 除了完成当前块的求和, 还负责将整列求和并计算出平均值, 通过共享寄存器传给第三列完成每个子块的方差, 第四列将子块的方差进行求和得到整个块的方差.

$$var = \frac{1}{N^2} \sum_{i=0}^N \sum_{j=0}^N (ave - I_{(i,j)})^2 \quad (1)$$

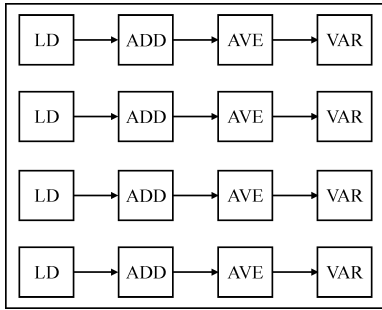


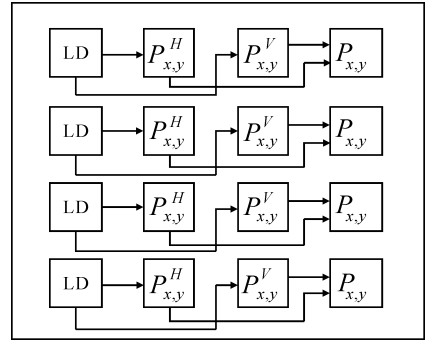
图4 方差计算映射图

4.3 帧内预测实现过程

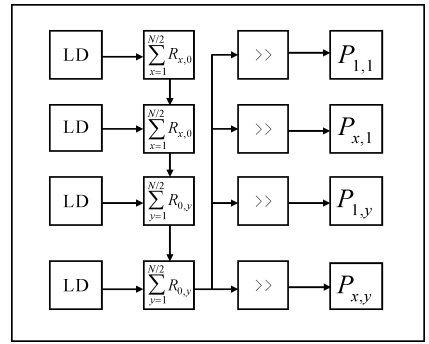
如图5为 PE 阵列实现帧内预测算法的结构图. 图(a)为 Planar 模式的映射结构图, PE 功能配置是以 Planar 公式(2)为依据, N 的值由 PU 块的大小决定, 由公式可知 $P_{x,y}^V$ 和 $P_{x,y}^H$ 之间没有数据依赖, 是可以并行计算的, 所以, 在配置 PE 阵列结构时, 与图4不同. 应把第二、三列 PE 配置成与第四列 PE 进行数据传输的形式, 但由于这两列 PE 不是邻接 PE, 所以, 将其配置成通过统一编址的存储器进行数据传输. 由于采用的是 4×4 阵列结构, 对于一个 4×4 的预测块只需一次流水就可以处理完成, 8×8 的预测块需要 4 次, 依次往上叠加. 图(b)为 DC 模式的 PE 阵列结构图, 公式(3)为其计算公式. 分析可得, dcValue 的值影响到整个预测值的计算, 把求和运算拆分成两部分并行计算, 可以有效加快预测值的计算. 以这样的方法对第二列 PE 进行配置. 同时, 通过 H 型网络对 PE 的运算状态进行检

测, 若检测到 dcValue 的完成标志, 就对二、三列 PE 进行二次配置, 实现动态重构, 提高 PE 阵列的利用率.

$$\begin{aligned} P_{(x,y)}^H &= (N-x) \cdot R_{(0,y)} + x \cdot R_{(N+1,0)}, \\ P_{(x,y)}^V &= (N-y) \cdot R_{(x,0)} + y \cdot R_{(0,N+1)}, \\ P_{(x,y)} &= (P_{(x,y)}^V + P_{(x,y)}^H + N) >>> (\log_2(N) + 1). \end{aligned} \quad (2)$$



(a) Planar 模式映射图



(b) DC 模式映射图

图5 Planar 模式和 DC 模式映射图

$$dcValue = \left(\sum_{x=1}^N R_{<x,0>} + \sum_{y=1}^N R_{<0,y>} + N \right) >>> (\log_2(N) + 1),$$

$$P_{<1,1>} = (R_{<1,0>} + R_{<0,1>} + 2 \cdot dcValue + 2) >>> 2,$$

$$P_{(x,1)} = (R_{<x,0>} + 3 \cdot dcValue + 2) >>> 2, \quad x = 2, 3, \dots, N, \quad (3)$$

$$P_{(1,y)} = (R_{<0,y>} + 3 \cdot dcValue + 2) >>> 2, \quad y = 2, 3, \dots, N,$$

$$P_{(x,y)} = dcValue, \quad x, y = 2, 3, \dots, N.$$

5 实验结果及性能分析

通过在 4×4 的可重构阵列规模上, 以 CU 块方差为性能参数驱动阵列进行不同块的 Planar 模式和 DC 模式的重构实验, 实现在不改变阵列结构的情况下, 通过下发配置信息, 动态改变阵列功能, 实现 Planar 和 DC 两种帧内预测模式的重构. 既能实

现并行加速效果,又有灵活调度的特点. 实验所设计的可重构阵列处理器用 Verilog HDL 语言描述,以标准测试序列 akiyo. qicf 作为测试序列在 Questasim 10.1d工具下进行功能仿真验证,采用90 nm CMOS 的技术工艺进行综合.

5.1 仿真验证

为验证算法映射的正确及统计阵列的性能参数,截取部分仿真结果进行分析说明,如图 6 所示为处理 4×4 块 DC 模式仿真过程,PE00 负责从 DIM 中读数,对于 4×4 的块需要在水平方向取 8 个参考像素,当取数完成后给 PE02 给个握手信号,当 PE02 检测到握手信号“6666”时,将 1 号地址和 9 号

地址的参考像素读到 PE02 的 R1 和 R2,完成公式(3)中第 2 个算式,首先将 R1 和 R2 相加后存在 R7,由于在 PE01 中已经完成了 *dcValue* 的计算,此时可以通过共享存储将 *dcValue* 的值读到 R15,2 倍 *dcValue* 的值存储在 R13,将 R7 和 R13 的相加后送到 R12,再加常量 2,右移两位得到值 120,

由图 6 中 1 号圈可知结果正确. 计算公式(3)中第 3 个算式时,首先计算 3 倍 *dcValue* 的值,前面已经得到 2 倍 *dcValue* 的值存在 R13,因此,只需要将 R15 与 R13 相加即可得到 3 倍 *dcValue* 的值,然后加常量 2 及参考像素 128 后右移两位得到值 116,由图 6 中 2 号圈可知结果正确,依次可验证算法的正确性.

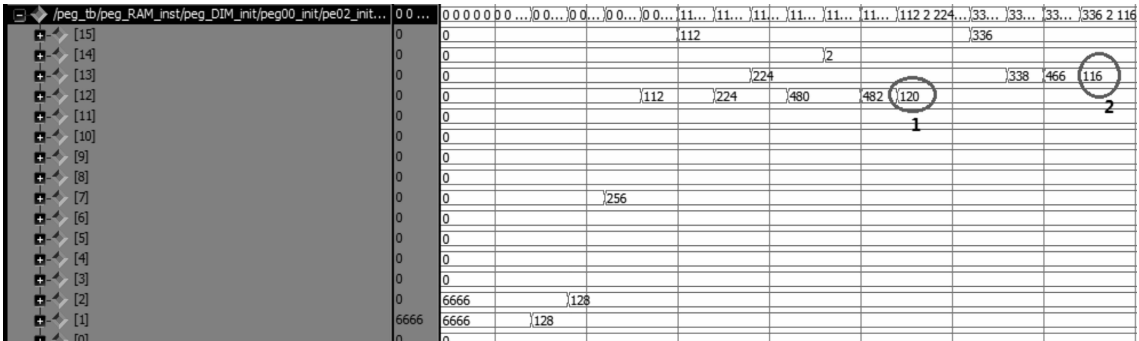


图 6 4×4 块 DC 模式仿真图

5.2 性能分析

实验结果如表 1、表 2、表 3 所示. 其中表 2 和表 3 是对部分性能的统计. 由表 1 得阵列处理器最大频率 357 MHz,资源占用为 25.2 kgate. 处理 4×4 的块 Planar 模式需要 77 个时钟周期,DC 模式需要 56 个时钟周期. 处理 8×8 的块 Planar 模式需要 191 个时钟周期,DC 模式需要 121 个时钟周期.

表 1 性能对比

	文献[9]	文献[8]	本文
模式	Planar、DC	Planar、DC	Planar、DC
工艺(nm)	90	—	90
门个数(k)	70.4	—	25.2
功耗(mW)	—	—	45.7053
最大频率(MHz)	357	—	357
处理时间	4×4	—	77、56
	8×8	285、229	191、121

表 2 吞吐量统计

		吞吐量 (cycle/px))	计算时间 (cycle)
Planar	4×4	4.81	77
	8×8	2.98	191
DC	4×4	3.50	56
	8×8	1.89	121

表 3 面积效率和能量效率统计

		性能 (Gbps)	面积效率 (Gbps/kgate)	能量效率 (Gbps/mW)
Planar	4×4	1.19	0.047	0.026
	8×8	1.92	0.076	0.042
DC	4×4	1.63	0.064	0.035
	8×8	3.02	0.120	0.066

由表 1 实验数据得出在实现与文献[9]相同算法情况下,硬件消耗仅为 35%. 虽然处理延时较文献[9]专用硬件有所增加,但本文提出的在可重构阵列处理器上的实现方案,通过下发配置能实现两种算法的灵活调度. 而且本文是将 HEVC 处理过程中 CU 块的划分到帧内预测作为一个统一的处理过程,而非仅仅实现帧内预测 Planar 模式和 DC 模式两种算法. 将单个算法的处理延时作为性能评价不够准确. 按照本文提出的方法, CU 块通过阈值提前终止,不需要遍历从最大块到最小块,从而节省了遍历所有块的时间开销. 若单比较 Planar 模式和 DC 模式的性能,较文献[8]同样用 16 个多核阵列实现的结果相比,处理延时减少了 32%.

6 结束语

针对硬件实现 HEVC 帧内预测算法所占资源较大、资源利用率不高、灵活性不足的问题,在传统阵列处理器的基础上,结合可重构处理器的特点和现已存在的帧内预测优化算法的特点,提出了一种由 CU 块阈值驱动的动态可重构阵列处理器。通过实验得出,与专用硬件实现帧内预测算法相比,能有效降低资源消耗,实现同一阵列资源的重复利用,提高了资源利用率。与多核处理器实现的方案相比,不仅具有算法动态映射的灵活性,而且算法本身处理速度也有所提高。因此,将 HEVC 算法在可重构阵列处理器上实现是解决灵活性和资源利用率问题的一种有效方法。

参考文献:

- [1] TRANG D L D, KIM K R, CHANG I J, et al. Texture characteristic based fast algorithm for CU size decision in HEVC intra coding[C]//2017 7th International Conference on Integrated Circuits, Design, and Verification (ICDV). Hanoi: IEEE, 2017: 88-93.
- [2] YI H, QIN H. The Optimization of HEVC Intra Prediction Mode Selection[C]// 2017 4th International Conference on Information Science and Control Engineering (ICISCE). Changsha: IEEE, 2017: 1743-1748.
- [3] ABDELLAH S, YUCEF S, LAMINE D. Lamine. Implementation of HEVC intra 4×4 prediction on FPGA[C]//2015 Science and Information Conference (SAD). London, UK: IEEE, 2015: 1160-1164.
- [4] AZGIN H, MERT A C, KALALI E, et al. An efficient FPGA implementation of HEVC intra prediction[C]//2018 IEEE International Conference on Consumer Electronics (ICCE). Las Vegas, NV: IEEE, 2018: 1-5.
- [5] TANG F, ZHAI Z Y. Implementation and optimization of HEVC decoder based on multi-core processor

[J]. Computer Engineering and Design, 2017, 38(1): 75-80.

- [6] 谢晓燕,徐卫芳,刘帆. HEVC 帧内预测 Planar 和 DC 模式算法的并行化设计[J]. 电视技术, 2015, 39(5): 4-8.
- [7] ZHU H, ZHOU W, QING D, et al. Efficient intra prediction VLSI architecture for HEVC standard[C]//2013 IEEE International Conference of IEEE Region 10 (TENCON 2013). Xian: IEEE, 2013: 1-4.
- [8] SHEN J, YANG J. Fast intra prediction algorithm based on early termination coding unit partitioning[J]. Journal of Computer Applications and Software, 2016, 33(5): 169-172.
- [9] GUO L, WANG X D, XU B W, et al. Fast algorithm for intra prediction mode decision and coding unit partitioning based on HEVC[J]. Journal of Computer Applications, 2018, 38(4): 1157-1163.
- [10] AKBARI O, KAMAL M, AFZALI-KUSHA A, et al. PX-CGRA: Polymorphic approximate coarse-grained reconfigurable architecture[C]//2018 Design, Automation & Test in Europe Conference & Exhibition (DATE). Dresden: IEEE, 2018: 413-418.
- [11] 张雪婷,蒋林,邓军勇,等. 可重构视频阵列处理器中全局控制器的设计与实现[J]. 微电子学与计算机, 2017, 34(11): 75-79.

作者简介:

贺飞龙 男,(1993-),硕士研究生.研究方向为专用集成电路设计. E-mail:15002902717@163.com

蒋林 男,(1970-),教授.研究方向为专用集成电路设计.

刘新闻 男,(1993-),硕士研究生.研究方向为专用集成电路设计.

山蕊 女,(1986-),博士,讲师.研究方向为集成电路设计.

王昱 女,(1993-),硕士研究生.研究方向为计算机视觉及图像处理.

吴皓月 女,(1995-),硕士研究生.研究方向为微电子学与固体电子学.