

一种新型 2-D 卷积器的 FPGA 实现

桑红石, 廖定彬, 江 威

(华中科技大学 图像识别与人工智能研究所 多谱信息处理技术国家级重点实验室, 湖北 武汉 430074)

摘 要: 提出了一种基于 FPGA 实现的新型 2-D 卷积器结构. 该结构占用硬件资源少, 外部带宽只需 1 pixels/clock. 与传统的卷积器相比, 不仅资源耗用减少 95% 以上, 并且可以依据系统实时性要求调整吞吐率, 带来了系统设计的灵活性. 在 Altera EP4SGX230KF40 FPGA 上实现结果表明, 对于 5×5 大小的图像卷积, 达到 90 Mpixels/s 的吞吐率, 满足大多数图像处理系统的实时性要求.

关键词: 2-D 卷积器; 结构设计; 图像卷积; FPGA

中图分类号: TN403

文献标识码: A

文章编号: 1000-7180(2011)09-0027-04

A Novel FPGA-Based 2-D Convolver

SANG Hong-shi, LIAO Ding-bin, JIANG Wei

(National Key Lab of Sci. & Tech. on Multi-Spectral Information Processing, Institute for Pattern Recognition & Artificial Intelligence, Huazhong University of Sci. and Tech., Wuhan 430074, China)

Abstract: This paper presents a novel 2-D convolver targeted at field programmable gate array (FPGA). The design features its low external memory bandwidth of 1 pixels/clock, and consumes comparatively little on-chip resources, reducing 95% of that compared with traditional convolvers. Also, it brings flexibility to system design by configurable throughput rate based on the real-time requirements. When implemented on Altera EP4SGX230KF40 FPGA, the throughput rate could reach up to 90 Mpixels/s for 5×5 image convolution, and it could satisfy most real-time image processing applications.

Key words: 2-D convolver; architecture design; image convolution; FPGA

1 引言

2-D 卷积在图像处理领域中有着广泛的用途, 如空间滤波, 图像增强^[1-3]等, 但由于卷积计算所需要的乘-加运算量特别大, 当处理较大的图像时, 采用商用 DSP 器件很难达到实时性能. 随着 FPGA 技术的飞跃发展, 图像处理中的诸多算法, 特别是基于窗口的处理算法常常将自身固有的 2-D 数据复用和 FPGA 的并行性结合用以硬件加速^[4-5], 加之 FPGA 内嵌的大量硬布线乘法器, 特别适合于卷积计算的硬件实现. 因此, 文中提出一种基于 FPGA 实现的新型 2-D 图像卷积器.

外部带宽, 片上存储器容量和吞吐率是 2-D

卷积器的三个设计指标. 文献[1]给出了采用多条双口 RAM 组成的延迟线缓存图像像素的传统图像卷积器硬件结构—FB^[2], 当图像或者卷积核较大时, 该结构会占用大量的片上存储器资源, 带来较大的面积和功耗开销. 文献[2]中提出的 MWPB 以多条输入 FIFO 替代了延迟线的使用, 在达到 1 clocks/pixel 的吞吐率情况下, 其外部带宽没有超过 2 pixels/clock, 在硬件资源开销和外部带宽之间实现了较好的折中. 但由于该设计使用了较多的输入 FIFO, 使得 I/O 配置变得复杂、引脚数增多, 同时, 为达到 1 clock/pixel 的吞吐率, MWPB 特有的输出流水线也会耗用额外的 FPGA 存储器资源.

文中针对已有设计结构中片上存储器资源开销

较大、FIFO 使用过多等不足,提出以部分 2-D 数据复用和 Z 字形扫描路径为特点的 2-D 卷积器硬件结构,并将设计方案在 FPGA 上实现。

2 基本原理和特点

2-D 图像卷积计算如式(1)所示,图像 I 和卷积模板 T 的大小分别为 $M \times N$ 和 $R \times S$, $O(x, y)$ 为输出图像,每个输出图像像素为模板和图像中对应 $R \times S$ 子块的点乘之和。

$$O(x, y) = \sum_{i=1}^{i=R} \sum_{j=1}^{j=S} T(i, j) \cdot I(x+i-1, y+j-1) \quad (1)$$

其中, $\forall (x, y) \in (M-R+1) \times (N-S+1)$ 。

对式(1)亦可以进行分解为:

$$P^i(x, y) = \sum_{j=1}^{j=S} T(i, j) \cdot I(x+i-1, y+j-1) \quad (2)$$

$$O(x, y) = \sum_{i=1}^{i=R} P^i(x, y), \text{ 其中:}$$

$$P^i(x, y) = \sum_{j=1}^{j=S} T(i, j) \cdot I(x+i-1, y+j-1) \quad (3)$$

这说明 2-D 卷积计算可以由 1-D 卷积计算来完成,若将所需的 1-D 卷积结果相加,即可以得到 2-D 卷积结果。这也是本设计硬件结构的基本思想。

卷积的计算过程也就是卷积计算窗口在图像中的滑动过程,存在行/列方向上的 2-D 数据复用,如图 1 所示。若卷积计算窗口在行方向上只滑动 MW 个位置,继而转为在下一行上滑动 MW 个位置,到图像最后一行后转先到前一行未计算的卷积窗口继续滑动 MW 个位置,这样重复下去直到结束全部计算,便是所谓的行主导 Z 字形扫描方式,后面的硬件开销分析表明该扫描方式会急剧减少片上存储器资源的开销。由于行方向上每次扫描的第 MW 个卷积计算窗口中的数据均不能达到完全的数据复用,本文称之为部分 2-D 数据复用。

3 2-D 卷积器硬件结构

3.1 卷积器结构

FPGA 中内嵌了大量的乘法器,使用和卷积模板一样大小的乘法器阵列并行计算可以极大地提高运算速度。文中设定图像和卷积模板的灰度级均为 8 位,图像的大小为 1024×1024 。图 2 给出了基于上

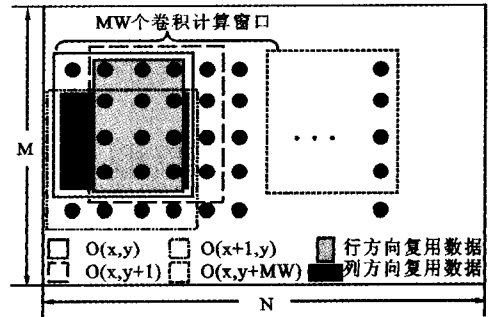


图 1 卷积计算过程示意图

述思想的卷积器硬件结构,其中移位寄存器组 RAR 由 S 个 8 位寄存器组成,用以缓存图像的 S 列像素,每个像素均被广播到 R 个处理单元 (Processing Element, PE)。该卷积器包含 $R \times S$ 大小的处理单元阵列,每个单元包含一个 $8b \times 8b$ 的乘法器和一个用来保存预先输入的卷积系数的 8 位寄存器。每行 S 个计算单元的乘法计算结果用加法器树 (Adder Tree, AT) 相加后写入到由 R-1 条深度为 MW 的双口 RAM 构成的中间结果缓存 (RIS) 中。一条 P 深度的 FIFO 用来匹配卷积器的数据处理速度和外部的输入带宽。

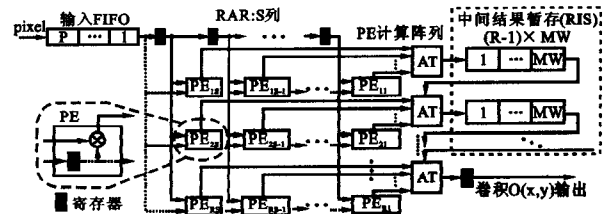


图 2 卷积器硬件结构

当图像第一行的前 S 列像素移入 RAR 后,计算出的 $P^1(1, 1)$ 被写入 RIS。下一个时钟第 S+1 列像素移入 RAR, 得到 $P^1(1, 2)$ 也被写入 RIS。同理, RAR 中移入 $MW+S-1$ 个像素后, $P^1(1, MW)$ 被写入 RIS。此后, 图像的第二行数据开始移入直到前 S 列像素均写入 RAR 后, $P^1(2, 1)$ 和 $P^2(1, 1) + [P^1(1, 1)]$ 被同时计算出来并分别写入到 RIS 中, 其中 $[]$ 内的是之前保存在 RIS 中的, 可以直接从构成 RIS 的双口 RAM 读出来。重复相似的操作, 当第 R 行的前 S 列像素移入 RAR 后, $P^1(R, 1), P^2(R-1, 1) + [P^1(R-1, 1)], \dots, P^R(1, 1) + [\sum_{k=1}^{k=R-1} P^k(1, 1)]$ 均同时计算出来并被写入到 RIS 中。这样, 第一个有效卷积结果 $O(1, 1)$ 将在下一个时钟出现在输出端口。此后, 其他的卷积结果便有规则、周期性的输出。这样实现了公式(2)。需要提出的是, 每次转行填充 RAR 时的前 S-1 个时钟周期内

均为无效的中间计算结果,这会降低硬件的使用效率,并导致吞吐率的下降。

若图像以行为主的方式存储在外部存储器中,当整幅图像存放完成后,文中的卷积器才能开始计算,这对于没有足够存储空间的低成本嵌入式实时系统是不方便的,可以将图 2 所示的卷积器结构转置并简单的修改,就可以实现式(3),得到类似 MWPB 的数据输入输出格式.这样针对以行/列为主的图像存放格式的不同应用系统,就可以灵活的选择实现式(3)或者式(2)的卷积器硬件结构.另外,图像的列数 N 需要被扩展为 $INT \cdot MW + S - 1$ 列,其中 INT 为正整数且满足 $(INT - 1) \cdot MW + S - 1 < N \leqslant INT \cdot MW + S - 1$,若有扩展的列,则填充为 0 以便卷积器能正常的工作。

3.2 性能分析

图像像素按照行主导 Z 字形扫描方式串行移入 RAR,因此本设计所需要的外部带宽为 1 pixels/clock,且 I/O 配置简单.每行移入的 $MW + S - 1$ 个像素会计算出 MW 个有效中间结果数据,那么,用于中间结果暂存的双口 RAM 的深度即为 MW .硬件的使用效率为 $MW / (MW + S - 1)$,其倒数即为吞吐率, $1 + (S - 1) / MW$ clocks/pixel.显而易见,RIS 的存储容量和吞吐率均为 MW 的函数,图 3 给出了 5×5 和 9×9 卷积窗口下它们的关系.当 MW 足够大时,吞吐率会逐渐收敛到 1 clocks/pixel,此时消耗的中间结果存储器容量也相应增大.根据系统的实时性要求,选择不同的 MW 来调整卷积器的吞吐率,增加了设计的灵活性,同时也减少了片上的资源开销。

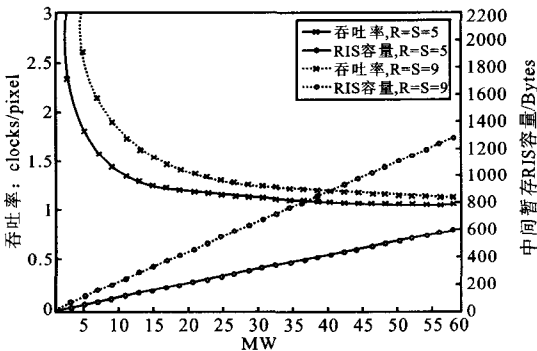


图 3 吞吐率,RIS 存储器容量和 MW 的关系

值得注意的是,MW 增大到一定程度时,继续增加对吞吐率的改善并不明显.因此,当存储器资源紧缺时,依据图 3 选择合适的 MW 将可以在卷积器性能和硬件开销之间获得满意的折中,使得设计更为优化.以 5×5 的图像卷积为例,取 MW 为 20 时,

吞吐率为 1.2 clocks/pixel,但是消耗的存储器资源仅为 200 bytes,而文献[1]中传统的 FB 结构却需要 4K bytes,也就是说,文中以 20% 的吞吐率的损失,减少了 95% 的片上存储器开销.因此,本设计在资源耗用和吞吐率之间取得了较好的折中。

4 实现结果与讨论

文中选取常用的 5×5 和 9×9 两种卷积核在 Altera EP4SGX230KF40 FPGA 上进行了实现,MW 分别取为 10 和 15,FIFO 的深度依文献[2]中取为 8.为便于比较,表 1 归纳了 FB,MWPB 和本设计在外部带宽 (pixels/clock),资源耗用 (Bytes) 以及吞吐率 (clocks/pixel) 三个方面的特性.其中资源耗用包含寄存器和存储器两个部分,具体计算卷积核时,两部分均以位 (bit) 为单位合并度量.为不失一般性,和本文一样,MWPB 寄存器中包含用于保存卷积系数的寄存器,但未计入其特有输出流水线占用的 $[Sx(S-1)/2] \times W$ 的存储器容量.W 为中间结果平均位宽,卷积核分别为 2.5,2.7 Bytes.

在文献[3]提出的对卷积器综合性能评估方法基础上,添加对外部带宽这一因素的考虑,即保证外部带宽和资源耗用最小的同时,具有最大的吞吐率的卷积器结构为最优,取性能评估值为三者的乘积,那么最小者对应的结构为最优化的卷积器.从表 1 可以看出,相对于 FB,本设计的片上资源耗用降低了 95%.与 MWPB 比较,本文卷积器结构的性能评估值在两种卷积核下分别降低了 24.4% 和 7.6%,因此,本设计在两种卷积核下均要优于 MWPB.值得一提的是,MWPB 中特有的用于输出流水线的存储器容量在和两种卷积核下分别为本结构中 RIS 暂存容量的 25% 和 30%,若是将之计算进表 I 中的资源消耗上,则文中结构相对于 MWPB 将会具有更大的优势。

采用 Verilog HDL 对上述三种卷积器结构进行了硬件实现,直接利用 MegaWizard Plup-In Manager 优化设计中需要的所有双口 RAM 和 FIFO,并用 QuartusII 10.0 对整个设计进行了综合.表 2 列出了三种结构的资源消耗报告和最高工作频率 (F,MHz),前者细分为了寄存器 (Reg,bit),查找表 (LUT) 和存储器 (Mem,Bytes) 三个部分。

MWPB 使用了较多的输入 FIFO,占用了较多的逻辑资源,较之文中结构,其用于缓存输入图像的寄存器数量是文中的 $R + S - 1$ 倍,从表 2 可以看出,MWPB 占用的逻辑资源几乎是本设计的 3 倍。

相对于 FB,文中卷积器的存储器容量在两种卷积核下均降低了 99%以上,且最高工作频率分别提高了 18. 8MHz 和 27. 9MHz. 同时,本设计可以分别达到

90 Mpixels/s 和 71. 4 Mpixels/s 的吞吐率,满足大多数图像处理系统的实时性要求.

表 1 不同结构卷积器的特性对比

设计	吞吐率	资源耗用		带宽	5×5				9×9			
		寄存器	存储容量		资源耗用	带宽	吞吐率	性能评估	资源耗用	带宽	吞吐率	性能评估
FB	1	2R×S	(R-1)×(N-S)+P	1	33072	1	1	33072	66320	1	1	66320
MWPB	1	(2R+S-1)×S	(R+S-1)×P	(R+S-1)/S	1136	1. 8	1	2045	2960	1. 89	1	5595
文中	1+(S-1)/MW	(R+1)×S	(R-1)×MW× W+P	1	1104	1	1. 4	1546	3376	1	1. 53	5168

表 2 不同结构卷积器的 FPGA 实现对比

项目	5×5			9×9		
	FB	MWPB	文中	FB	MWPB	文中
Reg	451	1 039	316	1 265	2 771	799
LUT	208	676	191	495	1 507	545
Mem	32 756	98. 25	108	65 472	239. 5	332
Total	<1%	<1%	<1%	<1%	2%	<1%
F	107. 2	115. 7	126. 1	81. 3	96. 4	109. 2

5 结束语

文中提出了一种低带宽需求,硬件资源开销小的新型 2-D 卷积器结构. 该结构可以根据系统实时性要求调整吞吐率,在资源开销和吞吐率之间取得了较好的折中. 通过与典型卷积器结构的 FPGA 实现对比,文中设计降低了 95%的片上资源开销,在 5×5 图像卷积时达到 90 Mpixels/s 的吞吐率,是一种性能较高的 2-D 卷积器.

参考文献:

[1] Bosi B, Bois G, Savaria Y. Reconfigurable pipelined 2

-D convolvers for fast digital signal processing[J]. IEEE Trans. on Very Large Scale Integr. (VLSI) Syst., 1999, 7(3): 229-308.

[2] Zhang Hui, Xia Mingxin, Guangshu Hu. A multiwindow partial buffering scheme for FPGA-based 2-D convolvers[J]. IEEE Trans. on Circuits Syst. II: Exp. Briefs, 2007, 54(2): 200-204.

[3] Cardells-Tormo F, Molinet P. Area-efficient 2-D shift-variant convolvers for FPGA-based digital image processing[J]. IEEE Trans. on Circuits and Syst. II: Exp. Briefs, 2006, 53(2): 105-109.

[4] 王德俊,王建立. 基于 FPGA 的实时图像改进直方均衡化算法[J]. 微电子学与计算机, 2010, 27(5): 87-89.

[5] 邓军,杨银堂. 一种基于并行处理技术的插值滤波算法及其 FPGA 实现[J]. 微电子学与计算机, 2010, 27(11): 82-85.

作者简介:

桑红石 女,(1970-),副教授. 研究方向为数字图像处理算法 VLSI 实现与并行处理器架构.

廖定彬 男,(1987-),硕士研究生. 研究方向为数字信号处理算法与 VLSI 设计.