



(12)发明专利申请

(10)申请公布号 CN 109741237 A

(43)申请公布日 2019. 05. 10

(21)申请号 201811631613.X

(22)申请日 2018.12.28

(71)申请人 中国科学院半导体研究所

地址 100083 北京市海淀区清华东路甲35号

(72)发明人 于双铭 李鸿龙 吴南健

(74)专利代理机构 中科专利商标代理有限责任公司 11021

代理人 任岩

(51)Int.Cl.

G06T 1/20(2006.01)

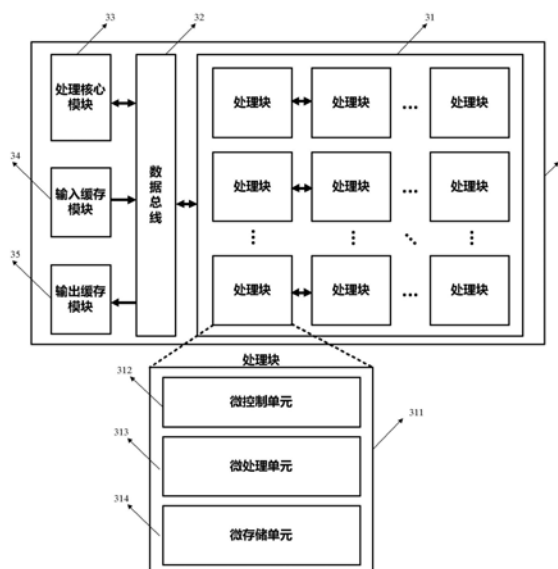
权利要求书2页 说明书10页 附图2页

(54)发明名称

大规模图像数据处理系统及方法

(57)摘要

本发明公开了一种大规模图像数据处理系统及方法,该系统包括:高速视觉芯片核心处理模块,用于对输入的图像数据进行层次化多级并行的深度学习算法处理,包括:处理核心模块和处理块阵列,该处理块阵列包含 $M \times N$ 个连接可配置的处理块, M 、 N 均为正整数,每个处理块包含微处理单元阵列;其中,处理核心模块用于整个处理块阵列的连接配置、运行控制、以及状态监控,同时能实现全局级图像数据处理;处理块阵列中的各个处理块用于多种不同分辨率图像的并行处理;微处理单元阵列包含 P 个微处理单元, P 为正整数,用于实现像素级数据的并行处理。该系统具有结构简单、低功耗、高效、数据吞吐率高、目标识别精度高的综合性能。



1. 一种大规模图像数据处理系统,其特征在于,包括:

高速视觉芯片核心处理模块(3),用于对输入的图像数据进行层次化多级并行的深度学习算法处理,该高速视觉芯片核心处理模块(3)包括:处理核心模块(33)和处理块阵列(31),该处理块阵列(31)包含 $M \times N$ 个连接可配置的处理块(311), M 、 N 均为正整数,每个处理块(311)包含微处理单元阵列(313);

其中,所述处理核心模块(33)用于整个处理块阵列(31)的连接配置、运行控制、以及状态监控,同时能实现全局级图像数据处理;

所述处理块阵列(31)中的各个处理块(311)之间通过处理核心模块(33)的配置而呈现不同的连接关系,用于多种不同分辨率图像的并行处理;

所述微处理单元阵列(313)包含 P 个微处理单元, P 为正整数,采用单指令多数据流(SIMD)工作模式,用于实现像素级数据的并行处理。

2. 根据权利要求1所述的大规模图像数据处理系统,其中,每个所述处理块(311)还包含:微控制单元(312),用于实现图像数据的块级并行处理及处理块(311)的运行控制;

所述微处理单元阵列(313)中的微处理单元用于实现SIMD操作,SIMD操作的控制指令通过所述微控制单元(312)实现;和/或,

所述处理核心模块(33)采用精简指令集计算机(RISC)指令。

3. 根据权利要求2所述的大规模图像数据处理系统,其中,所述微控制单元(312)包含:指令存储器、译码单元、数据执行单元、寄存器组以及数据选择器;

其中,指令存储器用来存储当前处理块(311)的执行指令,微控制单元(312)和微处理单元阵列(313)共用一套指令,微控制单元(312)中寄存器数据和微处理单元阵列(313)中寄存器数据共享,能根据处理任务并行度及复杂度在微控制单元(312)和微处理单元阵列(313)之间进行切换,译码单元根据取到的指令进行指令解析,根据指令类型去选择微控制单元(312)或微处理单元阵列(313)去操作数据进行运算。

4. 根据权利要求2所述的大规模图像数据处理系统,其中,每个所述处理块(311)还包含:微存储单元(314),所述微控制单元(312)能够以字节或半字访问微存储单元(314)中的全部数据;

优选的,每个处理块(311)中,微处理单元的总数据带宽和微存储单元(314)的数据带宽相匹配。

5. 根据权利要求1所述的大规模图像数据处理系统,其中,所述微处理单元包含:寄存器组、数据执行单元以及数据选择器;

其中,所述微处理单元的数据执行单元分为两个分支:一个分支为16bit通用计算单元,包含加法、减法、移位和逻辑运算的操作,用于实现计算机视觉以机器学习模型中的数据操作类型;另一个分支为8bit MAC计算单元,用于实现深度学习模型中卷积操作。

6. 根据权利要求1所述的大规模图像数据处理系统,还包括:

数据源输入模块(1),用于接收数据源图像传感器输入的图像数据;

数据存储模块(4),用于存储从数据源图像传感器输入的图像数据;

FPGA控制模块(2),与数据源输入模块(1)、数据存储模块(4)、以及高速视觉芯片核心处理模块(3)分别连接,用于图像数据的接收、缓存、分发、状态控制、模块参数配置和数据读出;以及

通讯模块(5), 设置于FPGA控制模块(2)和上位机之间, 用于FPGA控制模块(2)和上位机的数据交互, 将高速视觉芯片核心处理模块(3)处理后的结果数据输出到上位机, 上位机将配置数据输入给FPGA控制模块(2);

优选的, 所述数据源输入模块(1)通过光纤实现数据输入, 所述通讯模块(5)通过光纤实现数据输出。

7. 根据权利要求6所述的大规模图像数据处理系统, 其中, 所述高速视觉芯片核心处理模块(3)对应的所述数据源图像传感器为256bit图像存储器, 同时采用256bit数据带宽进行数据传输, 所述高速视觉芯片核心处理模块(3)采用流水的方式传输和处理图像数据, 将图像数据传输过程和图像数据处理过程分为两部分, 在处理当前图像数据的同时, 将下一幅图像数据输入进行处理, 将数据传输时间最小化。

8. 根据权利要求1所述的大规模图像数据处理系统, 其中, 对于单幅原始图像像素超过该大规模图像数据处理系统的最大分辨率的情况, 所述高速视觉芯片核心处理模块(3)中的处理块(311)采用滑动窗口的方式进行图像读取, 每次从原始图像中选取特定大小分辨率的图像块输送至高速视觉芯片核心处理模块(3)进行处理, 同时采用上下左右各重叠特定数量像素的方式滑动窗口, 保证目标的主体部分能够出现在至少其中一块图像块中。

9. 根据权利要求1至8任一项所述的大规模图像数据处理系统, 其中, 所述高速视觉芯片核心处理模块(3)还包括:

数据总线(32), 与所述处理块阵列(31)连接, 用于数据传输;

输入缓存模块(34), 通过所述数据总线(32)将缓存的输入数据输出至所述处理块阵列(31)中; 以及

输出缓存模块(35), 通过所述数据总线(32)缓存所述处理块阵列(31)的输出数据。

10. 一种基于权利要求1至9中任一项所述的大规模图像数据处理系统实现大规模图像实时处理的方法, 其特征在于, 包括:

待处理的图像数据输入至高速视觉芯片核心处理模块(3)中;

处理核心模块(33)启动处理块阵列(31)来进行图像处理, 根据图像处理的要求和对应的算法配置处理块阵列(31)中各个处理块(311)的连接关系, 并对处理块阵列(31)进行运行控制和状态监控, 同时实现全局级图像数据处理; 以及

在每个处理块(311)中, 微处理单元阵列(313)采用单指令多数据流(SIMD)工作模式, 实现像素级数据的并行处理。

大规模图像数据处理系统及方法

技术领域

[0001] 本公开属于数据处理和人工智能技术领域,涉及一种大规模图像数据处理系统及方法。

背景技术

[0002] 随着人工智能的兴起,机器的触觉、视觉等成为研究的重要方向,其中,图像目标识别作为机器视觉的一个重要研究方向,现已被广泛应用于医疗、农业、工业、国防以及交通等领域,例如目标跟踪、视频监控、信息安全、自动驾驶、图像检索、遥感图像分析、以及无人机导航等。

[0003] 目标识别是将图像或者视频中的目标与不感兴趣的部分区分开,判断是否存在目标,若目标存在则确定目标的位置,由于识别背景、环境及目标的复杂性,机器视觉在目标识别的准确度和高效性具有一定的困难,尤其是随着图像分辨率越来越高且图像数据的规模越发庞大且复杂的情况下,如何实现高速、实时的数据处理更显得尤为困难。

[0004] 现有的图像处理系统一般只能处理规模较小的数据,并且处理过程中如果需要同时采用不同的算法无法实现并行处理,对应大规模图像数据处理存在处理速度慢、准确度较低、算法不兼容等问题。

发明内容

[0005] (一)要解决的技术问题

[0006] 本公开提供了一种大规模图像数据处理系统及方法,以至少部分解决以上所提出的技术问题。

[0007] (二)技术方案

[0008] 根据本公开的一个方面,提供了一种大规模图像数据处理系统,包括:高速视觉芯片核心处理模块3,用于对输入的图像数据进行层次化多级并行的深度学习算法处理,该高速视觉芯片核心处理模块3包括:处理核心模块33和处理块阵列31,该处理块阵列31包含 $M \times N$ 个连接可配置的处理块311, M 、 N 均为正整数,每个处理块311包含微处理单元阵列313;其中,处理核心模块33用于整个处理块阵列31的连接配置、运行控制、以及状态监控,同时能够实现全局级图像数据处理;处理块阵列31中的各个处理块311之间通过处理核心模块33的配置而呈现不同的连接关系,用于多种不同分辨率图像的并行处理;微处理单元阵列313包含 P 个微处理单元, P 为正整数,采用单指令多数据流(SIMD)工作模式,用于实现像素级数据的并行处理。

[0009] 在本公开的一些实施例中,每个处理块311还包含:微控制单元312,用于实现图像数据的块级并行处理及处理块311的运行控制;微处理单元阵列313中的微处理单元用于实现SIMD操作,SIMD操作的控制指令通过微控制单元312实现;和/或,处理核心模块33采用精简指令集计算机(RISC)指令。

[0010] 在本公开的一些实施例中,微控制单元312包含:指令存储器、译码单元、数据执行

单元、寄存器组以及数据选择器；其中，指令存储器用来存储当前处理块311的执行指令，微控制单元312和微处理单元阵列313共用一套指令，微控制单元312中寄存器数据和微处理单元阵列313中寄存器数据共享，能根据处理任务并行度及复杂度在微控制单元312和微处理单元阵列313之间进行切换，译码单元根据取到的指令进行指令解析，根据指令类型去选择微控制单元312或微处理单元阵列313去操作数据进行运算。

[0011] 在本公开的一些实施例中，每个处理块311还包含：微存储单元314，微控制单元312能够以字节或半字访问微存储单元314中的全部数据；

[0012] 优选的，每个处理块311中，微处理单元的总数据带宽和微存储单元314的数据带宽相匹配。

[0013] 在本公开的一些实施例中，微处理单元包含：寄存器组、数据执行单元以及数据选择器；其中，微处理单元的数据执行单元分为两个分支：一个分支为16bit通用计算单元，包含加法、减法、移位和逻辑运算的操作，用于实现计算机视觉以机器学习模型中的数据操作类型；另一个分支为8bit MAC计算单元，用于实现深度学习模型中卷积操作。

[0014] 在本公开的一些实施例中，大规模图像数据处理系统，还包括：数据源输入模块1，用于接收数据源图像传感器输入的图像数据；数据存储模块4，用于存储从数据源图像传感器输入的图像数据；FPGA控制模块2，与数据源输入模块1、数据存储模块4、以及高速视觉芯片核心处理模块3分别连接，用于图像数据的接收、缓存、分发、状态控制、模块参数配置和数据读出；以及通讯模块5，设置于FPGA控制模块2和上位机之间，用于FPGA控制模块2和上位机的数据交互，将高速视觉芯片核心处理模块3处理后的结果数据输出到上位机，上位机将配置数据输入给FPGA控制模块2；

[0015] 优选的，数据源输入模块1通过光纤实现数据输入，通讯模块5通过光纤实现数据输出。

[0016] 在本公开的一些实施例中，高速视觉芯片核心处理模块3对应的数据源图像传感器为256bit图像存储器，同时采用256bit数据带宽进行数据传输，高速视觉芯片核心处理模块3采用流水的方式传输和处理图像数据，将图像数据传输过程和图像数据处理过程分为两部分，在处理当前图像数据的同时，将下一幅图像数据输入进行处理，将数据传输时间最小化。

[0017] 在本公开的一些实施例中，对于单幅原始图像像素超过该大规模图像数据处理系统的最大分辨率的情况，高速视觉芯片核心处理模块3中的处理块311采用滑动窗口的方式进行图像读取，每次从原始图像中选取特定大小分辨率的图像块输送至高速视觉芯片核心处理模块3进行处理，同时采用上下左右各重叠特定数量像素的方式滑动窗口，保证目标的主体部分能够出现在至少其中一块图像块中。

[0018] 在本公开的一些实施例中，高速视觉芯片核心处理模块3还包括：数据总线32，与处理块阵列31连接，用于数据传输；输入缓存模块34，通过数据总线32将缓存的输入数据输出至处理块阵列31中；以及输出缓存模块35，通过数据总线32缓存处理块阵列31的输出数据。

[0019] 根据本公开的另一个方面，提供了一种基于本公开的大规模图像数据处理系统实现大规模图像实时处理的方法，该方法包括：待处理的图像数据输入至高速视觉芯片核心处理模块3中；处理核心模块33启动处理块阵列31来进行图像处理，根据图像处理的要求和

对应的算法配置处理块阵列31中各个处理块311的连接关系,并对处理块阵列31进行运行控制和状态监控,同时实现全局级图像数据处理;以及在每个处理块311中,微处理单元阵列313采用单指令多数据流(SIMD)工作模式,实现像素级数据的并行处理。

[0020] (三)有益效果

[0021] 从上述技术方案可以看出,本公开提供的大规模图像数据处理系统及方法,具有以下有益效果:

[0022] (1)高速视觉芯片核心处理模块采用层次化多级并行处理架构,其中,处理核心模块用于整个处理块阵列的连接配置、运行控制、以及状态监控,同时能实现全局级图像数据处理,能够实现O(1)的全局级图像数据处理能力,处理块阵列包含 $M \times N$ 个处理块,处理块间通过处理核心模块的配置而设置其连接关系,进而处理块阵列能够实现O($M \times N$)的图像“块级”并行加速,每个处理块中的P个微处理单元能够实现O($M \times N \times P$)的“像素级”数据并行处理,从而实现大规模数据的多级并行处理;该系统可实现在视觉芯片上运行深度学习算法,完成在大规模图像数据中对多种目标实时高速地进行智能检测、识别、分类等功能,具有结构简单、低功耗、高能效、数据吞吐率高、目标识别精度高的综合性能,可应用于需要进行大规模图像数据实时目标识别或其他大规模数据处理的场合,数据吞吐率最高可达3.2Gbps,满足高速图像目标识别的应用需求;

[0023] (2)像素级图像处理和块级图像处理在时间上是分开的,微控制单元和微处理单元阵列共用一套指令,可以实现微控制单元对整个执行过程的控制,实现微控制单元寄存器中数据和微处理单元阵列中寄存器数据共享,实现根据处理任务并行度及复杂度在微控制单元和微处理单元之间的切换,同时可以减少硬件资源开销;

[0024] (3)一个处理块中的微处理单元的总数据带宽和微存储单元的数据带宽相匹配,因此单个微处理单元只能访问微存储单元中相对应位置的有限数据,但可以直接获取临近微处理单元中寄存器(比如左右各3个微处理单元单元)的数值。微处理单元的这种数据访问关系是基于目标检测算法的图像的局域计算只涉及到临近几个像素的计算分析的基础上进行设计的,通过这种设计在减少微处理单元对微存储单元的访问带宽的同时,还丝毫不影响微处理单元获取数据的性能;

[0025] (4)微处理单元中的数据执行单元分为两个分支,一个分支为16bit通用计算单元,包含加法、减法、移位和逻辑运算的操作,用于实现计算机视觉以机器学习模型中的数据操作类型;另一个分支为8bit MAC计算单元,用于实现深度学习模型中卷积操作,深度学习模型中为减少计算量并降低权重大小将网络模型量化为8bit数据类型,由于微处理单元中寄存器组数据带宽为16bit,为充分利用微处理单元寄存器数据带宽,每个数据执行单元中包含两个MAC单元,每个MAC包含一个8bit乘法器和32bit累加器。对于32bit累加器的输出结果通过移位方式将数据缩放到8bit数据带宽,用于匹配后续卷积操作数据带宽。通过这种数据执行单元的分支设计,将处理器的处理能力提升1倍。

附图说明

[0026] 图1为根据本公开一实施例所示的大规模图像数据处理系统的结构示意图。

[0027] 图2为根据本公开一实施例所示的FPGA控制模块的具体结构示意图。

[0028] 图3为根据本公开一实施例所示的高速视觉芯片核心处理模块的具体结构示意

图。

【0029】 【符号说明】

- [0030] 1-数据源输入模块；
[0031] 2-FPGA控制模块；
[0032] 21-中央控制模块； 22-数据输入接口模块；
[0033] 23-视觉芯片接口模块； 24-数据存储器接口模块；
[0034] 25-数据输出接口模块；
[0035] 3-高速视觉芯片核心处理模块；
[0036] 31-处理块阵列；
[0037] 311-处理块；
[0038] 312-微控制单元； 313-微处理单元阵列；
[0039] 314-微存储单元；
[0040] 32-数据总线； 33-处理核心模块；
[0041] 34-输入缓存模块； 35-输出缓存模块；
[0042] 4-数据存储模块； 5-通讯模块；
[0043] 6-时钟产生模块。

具体实施方式

[0044] 本公开提供了一种大规模图像数据处理系统及方法,采用层次化多级并行处理架构来设置高速视觉芯片核心处理模块,可实现在视觉芯片上运行深度学习算法,完成在大规模图像数据中对多种目标实时高速地进行智能检测、识别、分类等功能,具有结构简单、低功耗、高能效、数据吞吐率高、目标识别精度高的综合性能,可应用于需要进行大规模图像数据实时目标识别或其他大规模数据处理的场合,数据吞吐率最高可达3.2Gbps,满足高速图像目标识别的应用需求。

[0045] 本公开的大规模图像数据处理系统,包括:高速视觉芯片核心处理模块,用于对输入的图像数据进行层次化多级并行的深度学习算法处理,该高速视觉芯片核心处理模块为层次化多级并行处理架构,包括:处理核心模块和处理块阵列,该处理块阵列包含 $M \times N$ (M 、 N 的个数不进行限制, M 和 N 可相同或不同)个连接可配置的处理块, M 、 N 均为正整数,每个处理块包含微处理单元阵列;其中,处理核心模块用于整个处理块阵列的连接配置、运行控制、以及状态监控,同时能实现全局级图像数据处理;处理块阵列中的各个处理块之间通过处理核心模块的配置而呈现不同的连接关系,用于多种不同分辨率图像的并行处理;微处理单元阵列包含 P 个微处理单元, P 为正整数,可以与 M 或 N 相同或不同,采用单指令多数据流(SIMD)工作模式,用于实现像素级数据的并行处理。

[0046] 在本公开的一些实施例中,每个处理块还包含:微控制单元,用于实现图像数据的块级并行处理及处理块的运行控制;微处理单元阵列中的微处理单元用于实现SIMD操作,SIMD操作的控制指令通过微控制单元实现;和/或,处理核心模块采用精简指令集计算机(RISC)指令。

[0047] 在本公开的一些实施例中,对于单幅原始图像像素超过该大规模图像数据处理系统的最大分辨率的情况,高速视觉芯片核心处理模块中的处理块采用滑动窗口的方式进行

图像读取,每次从原始图像中选取特定大小分辨率的图像块输送至高速视觉芯片核心处理模块进行处理,同时采用上下左右各重叠特定数量像素的方式滑动窗口,保证目标的主体部分能够出现在至少其中一块图像块中。

[0048] 为使本公开的目的、技术方案和优点更加清楚明白,以下结合具体实施例,并参照附图,对本公开进一步详细说明。本公开中,说明书中的“处理器”为包含本公开所描述的高速视觉芯片核心处理模块的硬件结构,可以是视觉芯片,可以是其他处理器。

[0049] 在本公开的第一个示例性实施例中,提供了一种大规模图像数据处理系统。

[0050] 图1为根据本公开一实施例所示的大规模图像数据处理系统的结构示意图。

[0051] 参照图1所示,本实施例所示的大规模图像数据处理系统,包括:数据源输入模块1、FPGA控制模块2、高速视觉芯片核心处理模块3、数据存储模块4、通讯模块5、以及时钟产生模块6。

[0052] 下面结合附图来详细介绍本实施例中大规模图像数据处理系统的各个模块的组成以及模块内部组件之间的关系。

[0053] 参照图1所示,本实施例中,大规模图像数据处理系统(下面有些地方简称系统)的操作流程如下:首先,输入图像数据通过数据源输入模块1从外部高速存储器传输到FPGA控制模块2中,进而存储在数据存储模块4中。然后基于FPGA控制模块2实现图像的采样、切块、处理等操作,用于完成图像切块及图像金字塔构建,然后依次将切分出来的每一个图像块传输到高速视觉芯片核心处理模块3内部进行图片处理,本实施例中,采用该高速视觉芯片核心处理模块3实现目标识别的功能,则每一个图像块传输到高速视觉芯片核心处理模块3内部进行目标检测算法的执行处理,下面的相应设置均是以实现目标识别的功能为例进行说明。

[0054] 本实施例中, $N=M$,处理块阵列包含 $M \times M$ 个连接可配置的处理块,每个处理块包含P个微处理单元。

[0055] 图2为根据本公开一实施例所示的FPGA控制模块的具体结构示意图。

[0056] 参照图2所示,本实施例中,FPGA控制模块2包括:中央控制模块21、数据输入接口模块22、数据存储器接口模块24、视觉芯片接口模块23、以及数据输出接口模块25。

[0057] 中央控制模块21分别与数据输入接口模块22、数据存储器接口模块24、视觉芯片接口模块23、以及数据输出接口模块25相连接。中央控制模块21用于控制整个FPGA控制模块2的运行模式和工作状态,调配和分发图像数据在系统中的传输通道,控制各个接口模块分别与FPGA控制模块2以外的各个模块相连接,进行数据交互。数据输入接口模块22与数据源输入模块1相连接,用于接收输入的图像数据源。数据存储器接口模块24与数据存储模块4相连接,用于发送和接收图像数据存储以及读出数据存储模块4,并且发送控制信号给数据存储模块4。视觉芯片接口模块23与高速视觉芯片核心处理模块3相连接,用于向高速视觉芯片核心处理模块3发送控制信号和图像输入数据,接收来自高速视觉芯片核心处理模块3的处理结果数据。数据输出接口模块25与通讯模块5相连接,用于将待输出的数据进行打包和编码,发送给通讯模块5,使通讯模块5上传给上位机。

[0058] 图3为根据本公开一实施例所示的高速视觉芯片核心处理模块的具体结构示意图。

[0059] 参照图3所示,本实施例中,高速视觉芯片核心处理模块3为层次化多级并行处理

架构,包括:处理块阵列31、处理核心模块33、数据总线32、输入缓存模块34、以及输出缓存模块35,其中,处理块阵列31包含 $M \times M$ 个连接可配置的处理块311,每个处理块311包含微处理单元阵列313,微处理单元阵列313中包含 P 个微处理单元。

[0060] 其中,处理核心模块33为第一层次,负责整个处理块阵列的运行控制及状态监控,同时能实现全局的图像处理任务;处理块阵列31为第二层次,在处理核心模块33的配置下处理块之间呈现不同的连接关系(可以对应不同的算法)实现图像块级并行加速处理;微处理单元阵列313中 P 个微处理单元为第三层次,实现图像像素级并行加速处理。

[0061] 处理块阵列31作为该大规模图像数据处理系统的处理核心,主要包含 $M \times M$ 个连接可配置的处理块311,处理块之间通过处理核心模块33的配置而呈现不同的连接关系。

[0062] 每个处理块311包含:一个微控制单元312、 P 个微处理单元形成的微处理单元阵列313、以及一个微存储单元314, P 为正整数。微处理单元313主要用于实现像素级数据的并行处理,采用单指令多数据流(SIMD)工作模式。

[0063] 本实施例中,用于目标识别的算法对应设置微处理单元阵列的结构中,周围相邻6个微处理单元(左右侧各三个)单元相连,实现彼此间的数据交互。

[0064] 微控制单元312用于实现图像数据的块级并行处理及处理块311的运行控制。微处理单元阵列313中的微处理单元用于实现SIMD操作,SIMD操作的控制指令通过微控制单元312实现。

[0065] 处理核心模块33采用精简指令集计算机(RISC)指令,负责整个处理块阵列的运行控制及状态监控,同时能实现全局的图像处理任务。

[0066] 由上可知,高速视觉芯片核心处理模块3中的微处理单元阵列313能够实现 $O(M \times M \times P)$ 的像素级数据并行处理(第三层次),处理块311能够实现 $O(M \times M)$ 的图像块级并行加速(第二层次),处理核心模块33能够实现 $O(1)$ 的全局级图像数据处理能力(第一层次)。

[0067] 输入缓存模块34和输出缓存模块35主要是实现处理块阵列31与外部接口的数据交互以及高速视觉芯片核心处理模块3内部各存储器间的数据交换,能够有效的实现图像亚采样、图像转置、完整图像及任意位置图像块灵活配置。

[0068] 处理块阵列31是高速视觉芯片核心处理模块3的计算核心,主要包含 $M \times M$ 个连接可配置的处理块311,通过对连接的配置能够处理不同大小的图像。处理块阵列31拥有两个层次的处理单元,块级并行的处理块311和像素级并行的微处理单元阵列313。

[0069] 图3给出了处理块311的电路结构,参照图3所示,一个完整的处理块311主要包含:微控制单元312、微处理单元阵列313、以及相应的局部微存储单元314。

[0070] 微控制单元312主要包含:指令存储器、译码单元、数据执行单元、寄存器组、以及必要的选择器。其中,指令存储器用来存储当前处理块311的执行指令,由于像素级图像处理和块级图像处理在时间上是分开的,微控制单元312和微处理单元阵列313共用一套指令,能够实现微控制单元312对整个执行过程的控制,实现微控制单元312寄存器中数据和微处理单元阵列313中寄存器数据共享,实现根据处理任务并行度及复杂度在微控制单元312和微处理单元313切换,同时可以减少硬件资源开销。译码单元根据取到的指令进行指令解析,根据指令类型去选择特定的执行器(微处理单元或微控制单元312)去操作数据进行运算。寄存器组的数据有两个来源:数据执行单元、微存储单元。微控制单元312能够以字节或半字访问微存储单元314中的全部数据。

[0071] 微处理单元是高速视觉芯片核心处理模块3的像素级并行处理单元,用于对像素级图像处理过程进行并行加速。在一优选实施例中,设置一个处理块311中的微处理单元的总数据带宽和微存储单元314的数据带宽相匹配。因此单个微处理单元只能访问微存储单元314中相对应位置的有限数据,但可以直接获取临近微处理单元中寄存器(左右各3个微处理单元)的数值(这里并未示意微处理单元阵列中各个微处理单元的连接关系)。微处理单元这种数据访问关系是基于目标检测算法的图像的局域计算只涉及到临近几个像素的计算分析的基础上进行设计的。通过这种设计可以减少微处理单元对微存储单元314的访问带宽,同时丝毫不影响微处理单元获取数据的性能。

[0072] 本实施例中,一个微处理单元主要包含:一个 16×16 bit的寄存器组、16bit数据执行单元、以及必要的选择器。当然,具体数据执行单元、寄存器组中的位数可根据实际需要进行适应性设置,不局限于本实施例。

[0073] 微处理单元中寄存器组的功能和微控制单元312中寄存器组的功能类似,主要用于存储来自微存储单元314的数据和数据执行单元运算的结果,不同的是微处理单元中寄存器组的数据来源除了微控制单元312的两个来源外,还可以直接来自微控制单元312及临近微处理单元的寄存器组。

[0074] 微处理单元主要用来实现SIMD操作,控制指令通过微控制单元312单元实现,微处理单元的数据执行单元功能相比微控制单元312主要更偏向于数据计算方面。微处理单元的数据执行单元主要包含加法、减法、乘加、与、或、异或、取反等操作。数据执行单元的第一操作数来自寄存器组,第二操作数来自寄存器组、立即数或邻近微处理单元的寄存器组。数据执行单元计算的结果同样写入寄存器组进行后续的计算或者写回微存储单元314。

[0075] 微处理单元中的数据执行单元分为两个分支:一个分支为16bit通用计算单元,包含加法、减法、移位和逻辑运算等操作,主要用于实现计算机视觉以机器学习模型中的数据操作类型;另一个分支为8bit MAC计算单元,用于实现深度学习模型中卷积操作。深度学习模型中为减少计算量并降低权重大小将网络模型量化为8bit数据类型,由于微处理单元中寄存器组数据带宽为16bit,为充分利用微处理单元中寄存器数据带宽,每个数据执行单元中包含两个MAC单元,每个MAC包含一个8bit乘法器和32bit累加器。对于32bit累加器的输出结果通过移位方式将数据缩放到8bit数据带宽,用于匹配后续卷积操作数据带宽。通过这种数据执行单元分为两个分支的设计,将处理器的处理能力提升1倍。

[0076] 由于高速视觉芯片核心处理模块3拥有多层次并行处理结构(微处理单元阵列313,处理块阵列31,处理核心模块33),能够有效面对一个完整的图像检测算法(往往需要涉及到多种计算机视觉和深度学习算法)的高并行度算法部分的并行化加速要求,也能够适应低并行度算法的计算要求,可以有效避免单纯SIMD形式的2维处理阵列面对低并行度算法的资源浪费情况。

[0077] 处理块311之间具有灵活的连接方式,可灵活配置成不同形式的处理块阵列31。通过配置处理块311之间的连接关系,可以将处理块阵列31配置成 $0.25M \times 4M$ 、 $0.5M \times 2M$ 、 $M \times M$ 、 $2M \times 0.5M$ 、 $4M \times 0.25M$ 等多种形式,通过使用处理核心模块31、输入缓存模块34和输出缓存模块35,对多种分辨率图像($0.25MN \times 0.25MN$ 、 $0.5MN \times 0.5MN$ 、 $MN \times MN$ 、 $2MN \times 2MN$ 、 $4MN \times 4MN$ 等)进行切分,并传入相应的处理单元实现并行化处理。

[0078] 高速视觉芯片核心处理模块3具有完成大规模图像实时处理的功能,本实施例以

目标识别为例,具体操作流程如下所述:输入到高速视觉芯片核心处理模块3的图像块首先会存放在输入缓存模块34中,等待数据传输完毕后,处理核心模块33将数据转移到处理块阵列31中的微存储单元314中。等待数据完全传输至微存储单元314中后,处理核心模块33将会启动处理块阵列31来完成目标检测算法第一阶段的执行,等待第一阶段处理完毕,处理核心模块33会将第一阶段的结果进行分析,若算法第一阶段给出的结果认为图像块中不含有疑似目标,则算法执行结束,并将该结果通过通讯模块5保存至上位机中高速存储器。若算法第一阶段给出的结果认为该图像块含有疑似目标,则处理核心模块33将启动处理块阵列31执行算法的第二阶段,并将第二阶段产生的目标的位置区域以及大小信息通过通讯模块5保存至上位机中的高速存储器。处理核心模块33控制着高速视觉芯片核心处理模块3内所有操作的执行流程,同时参与第一阶段结果分析和判断等需要获取高速视觉芯片核心处理模块3全局性数据才能完成的计算过程。

[0079] 输入缓存模块34主要用于存储输入图像数据。由于使用的数据总线32带宽较大,若输入缓存模块34使用单片的存储器结构时,在对图像区域进行数据采集传输过程中无法实现精确位置的数据获取,本实施例通过采用4片并行的存储器组合方式构建输入缓存模块34。

[0080] 输出缓存模块35主要用来存取高速视觉芯片核心处理模块3输出的结果信息,以待传输到处理器外部。由于处理结果仅包含图像中目标位置信息或当前完成处理的图像块中是否含有目标等比较简明的信息,数据传输过程较为简单,因此输出缓存模块35可以直接选用单片256bit带宽的存储器模型。

[0081] 高速视觉芯片核心处理模块3的图像数据输入端与FPGA控制模块2的输出端相连接,端口位宽为64bit。高速视觉芯片核心处理模块3的图像数据输出端与FPGA控制模块2的输入端相连接,端口位宽为64bit。数据吞吐率最高可达3.2Gbps。

[0082] 本实施例中,高速视觉芯片核心处理模块3采用流水的方式传输和处理数据,将图像数据传输过程和数据处理过程分为两部分,在处理当前图像的同时,将下一幅图像输入到处理单元中,将数据传输时间最小化,但由于从图像存储器中读取数据时带宽只有8bit。这种方式在传输大尺寸图像时会占用较多的时钟周期,当图像处理算法较为高效时会出现等待下一幅图像数据的情况。因此本文提出的高速视觉芯片核心处理模块3采用256bit图像存储器,同时采用256bit数据带宽进行数据传输,这样的设置能够实现:即使面临耗时较少的图像处理算法时也能够及时完成图像数据的传输,实现大量图像数据的连续处理。

[0083] 下面来参照图1和图2介绍本实施例所示的大规模图像数据处理系统中各个模块之间的关系。

[0084] 本实施例中,FPGA控制模块2包括1块高性能FPGA芯片,该FPGA芯片采用Altera公司Arria V系列芯片,与数据源输入模块1的输出端相连接,与高速视觉芯片核心处理模块3的输入端和输出端均相连,与数据存储模块4的数据输入端和数据输出端均相连,与通讯模块5的输入端和输出端均相连。

[0085] 优选的,数据源输入模块1通过光纤实现数据输入,通讯模块5通过光纤实现数据输出。

[0086] FPGA控制模块2包括视觉芯片接口模块23,用于和高速视觉芯片核心处理模块3进行数据发送、指令参数发送、数据接收和状态机控制。视觉芯片接口模块23提供一路数字时

钟信号给高速视觉芯片核心处理模块3,时钟信号频率为50MHz,视觉芯片接口模块23向高速视觉芯片核心处理模块3发送的图像数据和指令参数数据、接收处理结果数据和控制信号都与该时钟信号同步。

[0087] FPGA控制模块2包括数据存储器接口模块24,该数据存储器接口模块24包括DDR存储控制器,与数据存储器模块4连接,为数据存储器模块4提供数据读写接口,用于和数据存储器模块4进行数据发送、数据接收和协议控制。读写接口共用一套地址信号。

[0088] 数据存储器接口模块24与数据源输入模块1、及数据存储器模块4连接。数据存储器接口模块24产生读写使能信号和读写地址指针,可以分时对数据存储器模块4进行读写,当数据源输入模块1有数据传来时,数据存储器接口模块24通过乒乓读写的方式将数据写入数据存储器模块4;当数据源输入模块1没有数据传来时,数据存储器接口模块24通过乒乓读写的方式读取数据,并将所读取的数据传给FPGA控制模块2。

[0089] 数据存储器模块4包括至少2个DDR存储单元,数据存储器模块4的输入端口与FPGA控制模块2的输出端口相连,数据存储器模块4的输出端口与FPGA控制模块2的输入端口相连,每个数据通道为16bit。图像数据和指令参数数据通过数据源输入模块1经由FPGA控制模块2输入到数据存储器模块4中。

[0090] 通讯模块5包括2个双向光纤接口,用于FPGA控制模块2通过通讯模块5向上位机发送处理结果数据。

[0091] 时钟产生模块6的输入端与FPGA控制模块2的输出端相连,时钟产生模块6的输出端与高速视觉芯片核心处理模块3的输入端相连接,用于给高速视觉芯片核心处理模块3产生时钟频率为200MHz的时钟信号。该时钟信号用于高速视觉芯片核心处理模块3内部的图像处理算法运行的时钟。

[0092] 本公开中,对于单幅原始图像像素超过该大规模图像数据处理系统的最大分辨率的情况,高速视觉芯片核心处理模块中的处理块采用滑动窗口的方式进行图像读取,每次从原始图像中选取特定大小分辨率的图像块输送至高速视觉芯片核心处理模块进行处理,同时采用上下左右各重叠特定数量像素的方式滑动窗口,保证目标的主体部分能够出现在至少其中一块图像块中。

[0093] 例如,大规模图像数据处理系统所处理的单幅原始图像像素数目众多,系统无法实现对整幅图像的一次性处理,因此本实施例中,采用滑动窗口的方式,每次从原始图像中选取 256×256 大小图像输送到系统进行处理,同时为避免出现将目标切分至多个图像块而出现漏检的情况,采用上下左右各重叠16个像素的方式滑动窗口,保证目标的主体部分能够出现在至少其中一块图像块中。同时为适应 256×256 图像块的处理,通过处理块311之间的连接方式配置为 8×2 处理块阵列31模式(同时微处理单元阵列313为 256×2)。

[0094] 高速视觉芯片核心处理模块3作为数据的处理核心,每次从数据缓存中读取一定大小的数据块进行处理,同时将产生的处理结果(有效目标图像块、目标的位置信息及大小等)输出到数据存储器模块4中,作为最终结果保存。通过目标检测算法和多级并行高速视觉芯片核心处理模块3的协同设计,使得该系统可以实现高速的目标检测过程。同时该系统可作为近传感器端的边缘型处理系统,只产生处理的结果,从而将无关的背景信息进行有效的抑制,极大程度上减少数据传输带宽。

[0095] 在本公开的第二个示例性实施例中,提供了一种基于本公开的大规模图像数据处

理系统实现大规模图像实时处理的方法,该方法包括:待处理的图像数据输入至高速视觉芯片核心处理模块3中;处理核心模块33启动处理块阵列31来进行图像处理,根据图像处理的要求和对应的算法配置处理块阵列31中各个处理块311的连接关系,并对处理块阵列31进行运行控制和状态监控,同时实现全局级图像数据处理;以及在每个处理块311中,微处理单元阵列313采用单指令多数据流(SIMD)工作模式,实现像素级数据的并行处理。

[0096] 综上所述,本公开提供了一种大规模图像数据处理系统及方法,采用层次化多级并行处理架构来设置高速视觉芯片核心处理模块,处理核心模块33为第一层次,负责整个处理块阵列的运行控制及状态监控,同时能实现全局的图像处理任务;处理块阵列31为第二层次,在处理核心模块33的配置下处理块之间呈现不同的连接关系(可以对应不同的算法)实现图像块级(不同大小的图像)并行加速处理;微处理单元阵列313中P个微处理单元为第三层次,实现图像像素级并行加速处理;可实现在视觉芯片上运行深度学习算法,完成在大规模图像数据中对多种目标实时高速地进行智能检测、识别、分类等功能,具有结构简单、低功耗、高能效、数据吞吐率高、目标识别精度高的综合性能,可应用于需要进行大规模图像数据实时目标识别或其他大规模数据处理的场合,数据吞吐率最高可达3.2Gbps,满足高速图像目标识别的应用需求。

[0097] 需要说明的是,在附图或说明书正文中,未绘示或描述的实现方式,均为所属技术领域普通技术人员所知的形式,并未进行详细说明。

[0098] 还需要说明的是,本公开可以借助于包括有若干不同元件的硬件以及借助于适当编程的计算机来实现。本公开的各个部件实施例可以以硬件实现,或者以在一个或者多个处理器上运行的软件模块实现,或者以它们的组合实现。本领域的技术人员应当理解,可以在实践中使用微处理器或者数字信号处理器(DSP)来实现根据本公开实施例的相关设备中的一些或者全部部件的一些或者全部功能。本公开还可以实现为用于执行这里所描述的方法的一部分或者全部的设备或者装置程序(例如,计算机程序和计算机程序产品)。这样的实现本公开的程序可以存储在计算机可读介质上,或者可以具有一个或者多个信号的形式。这样的信号可以从因特网网站上下载得到,或者在载体信号上提供,或者以任何其他形式提供。

[0099] 类似地,应当理解,为了精简本公开并帮助理解各个公开方面中的一个或多个,在上面对本公开的示例性实施例的描述中,本公开的各个特征有时被一起分组到单个实施例、图、或者对其的描述中。然而,并不应将该公开的方法解释成反映如下意图:即所要求保护的本公开要求比在每个权利要求中所明确记载的特征更多的特征。更确切地说,如下面的权利要求书所反映的那样,公开方面在于少于前面公开的单个实施例的所有特征。因此,遵循具体实施方式的权利要求书由此明确地并入该具体实施方式,其中每个权利要求本身都作为本公开的单独实施例。

[0100] 以上所述的具体实施例,对本公开的目的、技术方案和有益效果进行了进一步详细说明,所应理解的是,以上所述仅为本公开的具体实施例而已,并不用于限制本公开,凡在本公开的精神和原则之内,所做的任何修改、等同替换、改进等,均应包含在本公开的保护范围之内。

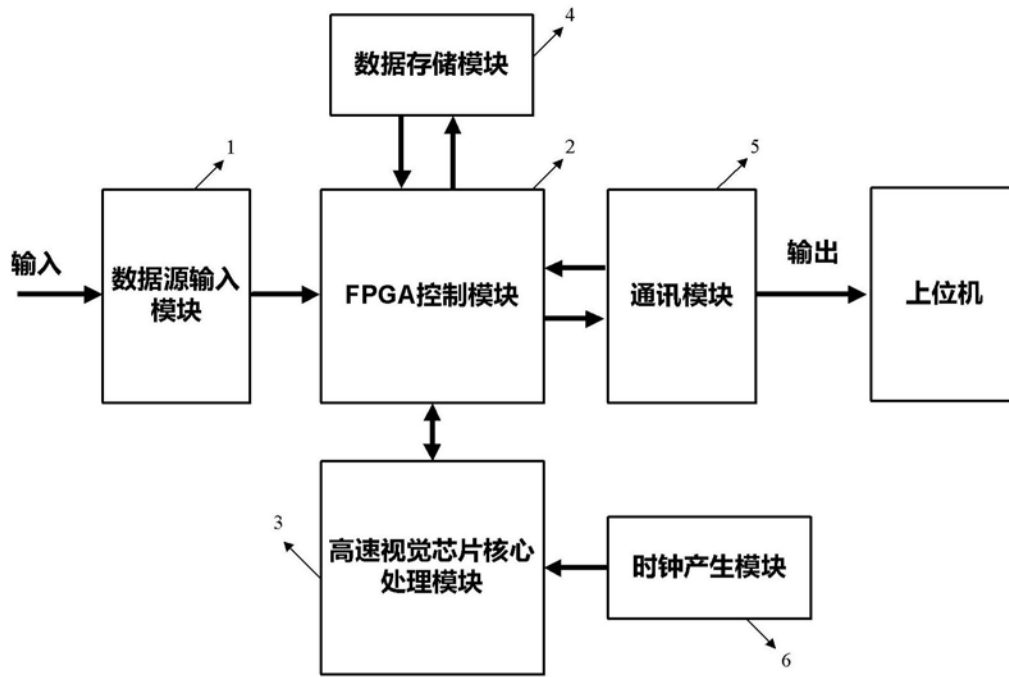


图1

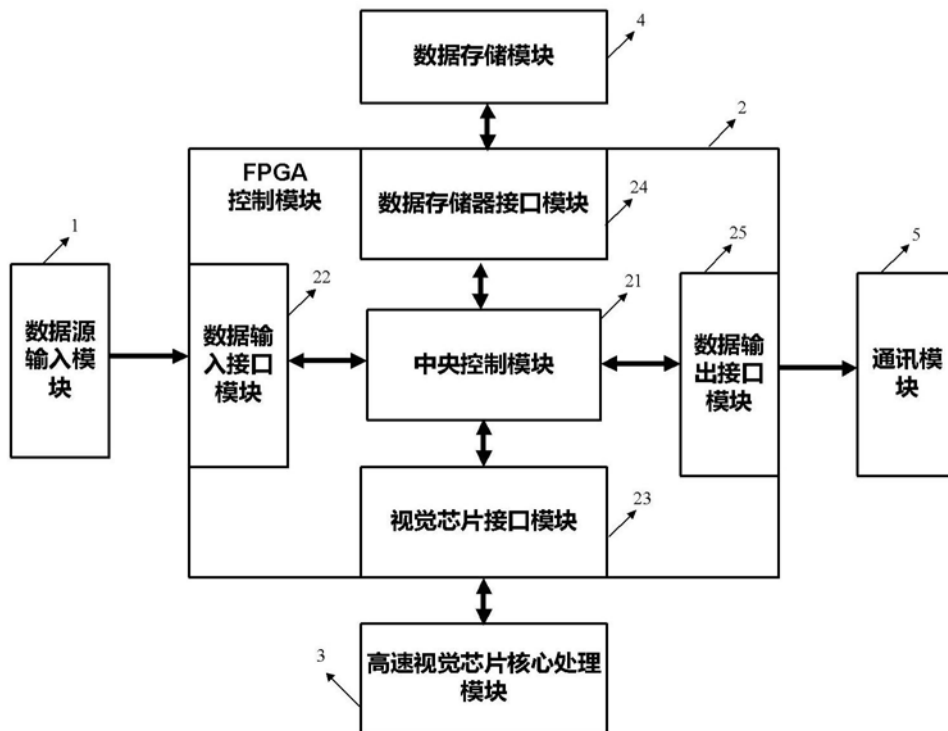


图2

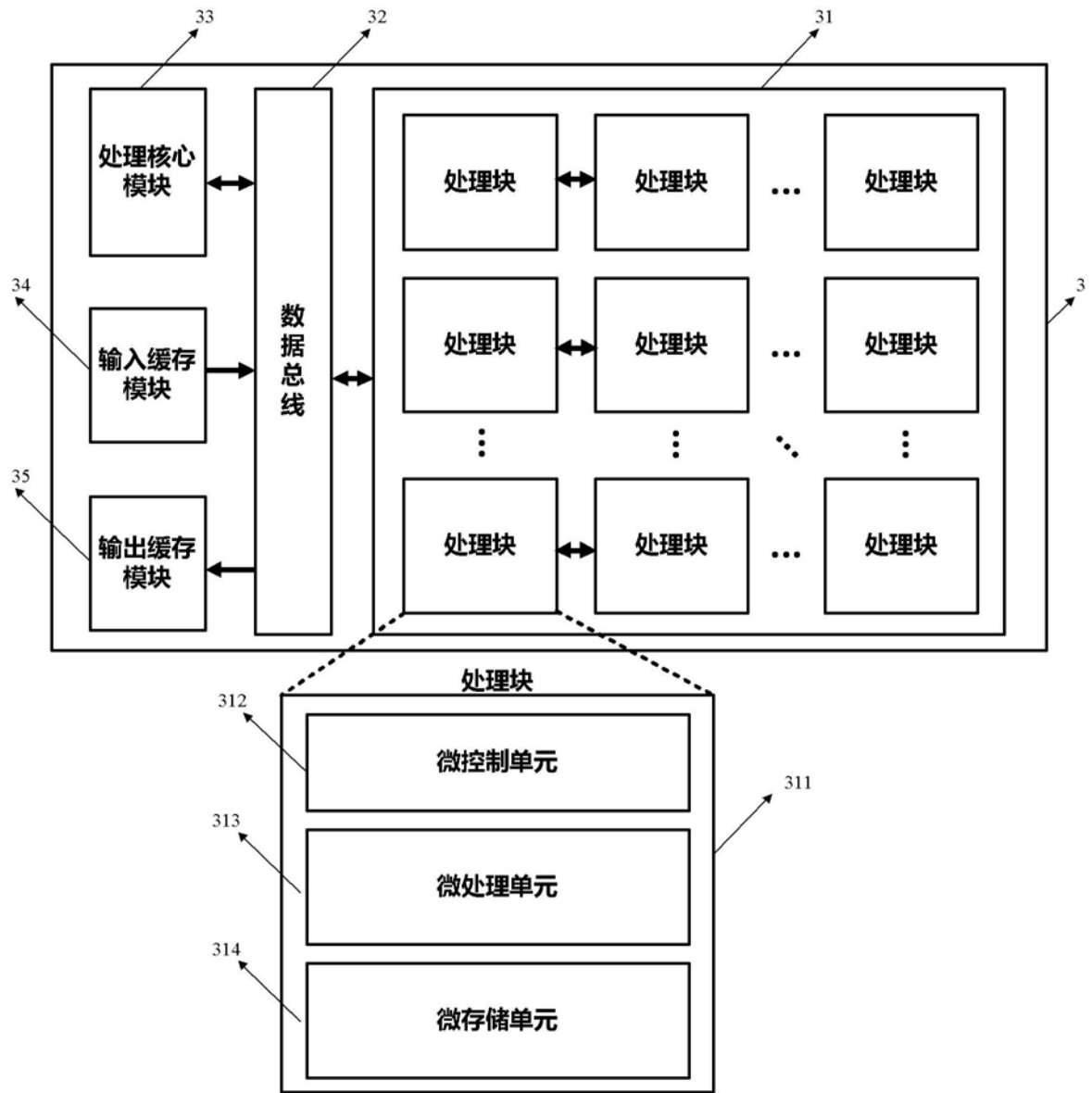


图3