(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112015548 A
(43)申请公布日 2020.12.01

(21)申请号 202010774691.6

(22)申请日 2020.08.04

(71)申请人 成都云图睿视科技有限公司
地址 611700 四川省成都市高新区应龙北二路900号

(72)发明人 孟莹

(74)专利代理机构 成都明涛智创专利代理有限公司 51289
代理人 刘晓政

(51)Int.Cl.
G06F 9/50(2006.01)
G06F 9/48(2006.01)
G06N 3/063(2006.01)

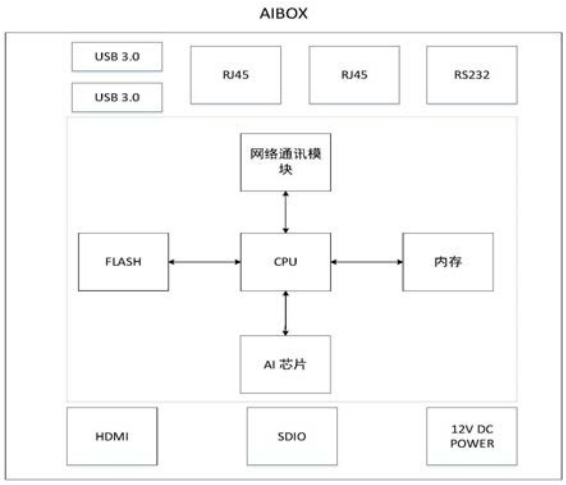
权利要求书1页 说明书4页 附图3页

(54)发明名称

一种用于人工智能视频应用的边缘计算平台

(57)摘要

本发明公开了一种用于人工智能视频应用的边缘计算平台，其特征在于：包括平台硬件部分，所述硬件部分包括一个多核多线程CPU、一个内存、一个FLASH闪存、一个网络通讯模块、一个或多个AI芯片以及硬件接口，所述内存、FLASH闪存、网络通讯模块和AI芯片均与多核多线程CPU电性连接，本发明的有益效果是：本发明一种用于人工智能视频应用的边缘计算平台，通过采用硬件神经网络加速器和专用的软件开发工具包，以及多个加速器协同工作，可以将人工智能视频应用推向实时化。



CN 112015548 A

1.一种用于人工智能视频应用的边缘计算平台,其特征在于:包括平台硬件部分,所述硬件部分包括一个多核多线程CPU、一个内存、一个FLASH闪存、一个网络通讯模块、一个或多个AI芯片以及硬件接口,所述内存、FLASH闪存、网络通讯模块和AI芯片均与多核多线程CPU电性连接,其中,

多核多线程CPU用于完成摄像头视频流数据的解码,AI芯片的任务请求,多任务的执行和调度等功能;

FLASH用于存储平台的配置参数和AI模块运行的数据结果;

AI芯片用于支持多路高清摄像头高帧率实时处理,可以支持多线程异步工作模式,以及多个芯片协同工作模式;

网络通讯模块用于与云端连接获取相关数据。

2.根据权利要求1所述的一种用于人工智能视频应用的边缘计算平台,其特征在于:所述硬件接口包括2个USB 3.0接口、2个以太网RJ45接口、一个显示HDMI接口、一个串行RS232接口、一个SD卡接口、一个直流12V电源接口。

3.根据权利要求1所述的一种用于人工智能视频应用的边缘计算平台,其特征在于:所述网络通讯模块包括一个以太网控制器和一个4G模块。

4.根据权利要求1所述的一种用于人工智能视频应用的边缘计算平台,其特征在于:所述网络通讯模块是一个PCI总线应用的集成的千兆级以太网控制器,该终端引出两路RJ45接口,其中一个具有POE功能,即可以通过以太网线向终端设备供电。

5.根据权利要求1所述的一种用于人工智能视频应用的边缘计算平台,其特征在于:所述AI芯片为深度神经网络推理的神经计算芯片。

6.根据权利要求1-5任意一项所述的一种用于人工智能视频应用的边缘计算平台的工作方法,其特征在于:包括一下步骤:

S1、通过一个中央交换机,该终端可以与多个摄像头组成一个局域网,通过局域网,终端可以接收多路摄像头RTSP视频流数据,并在终端内解码视频流数据,调用AI芯片进行图片的检测与识别,将输出的场景结构化数据发送给云端或存在终端FLASH中,云端服务器可以将数据结果分发给用户;

S2、同时云端也可以存取磁盘录像机的视频数据,用于服务器端数据的分析和网络模型的训练。

7.根据权利要求6所述的一种用于人工智能视频应用的边缘计算平台的工作方法,其特征在于:所述S1的视频处理中存在单路视频单个视频处理单元、多路视频单个视频处理单元和多路视频多个视频处理单元的三种情况。

8.根据权利要求7所述的一种用于人工智能视频应用的边缘计算平台的工作方法,其特征在于:所述单路视频单个视频处理单元处理时多个AI任务依次向VPU发出推导请求,VPU推导完成将计算结果返回给CPU,下一个AI任务再向VPU发出推导请求,如此进行,此时VPU不会出现忙的情况,因为AI任务是依次执行的,在VPU推导的时候,此时CPU处于等待状态;所述多路视频单个视频处理单元处理时SHELL外壳部分为每一路视频流开启一个处理线程,每个处理线程相互独立;所述多路视频多个视频处理单元处理时通过VPU的API函数可以指定对应的VPU处理对应的视频流。

一种用于人工智能视频应用的边缘计算平台

技术领域

[0001] 本发明涉及人工智能视频技术领域,具体为一种用于人工智能视频应用的边缘计算平台。

背景技术

[0002] 由于卷积神经网络的理论突破,深度学习理论再次用于解决各大技术难题,由于其强大的表达能力,迅速地超过了传统的机器学习方法,同时也促进了许多人工智能技术快速落地,广泛地应用于智能监控、自动驾驶、金融、教育等领域。

[0003] 正由于神经网络强大地表征能力,也决定了其在工业多场景应用中网络的复杂性,所以它不仅需要更多的数据去训练神经网络模型,也需要强大的算力,在边缘端部署还需要考虑实时性和低成本的要求。如果AI模型的推导在云端进行,但受限于现场通信带宽和时延等要求,所以在大多数情况下, AI模型需要在边缘端部署,如果在边缘端采用采用昂贵的GPU设备进行推导,那么成本不划算,所以需要在边缘端找到一个实时的和低成本的解决方案。

[0004] 针对上述不足,本发明的目的在于设计一种用于人工智能视频应用的边缘计算平台。

发明内容

[0005] 本发明的目的在于提供一种用于人工智能视频应用的边缘计算平台,以解决上述背景技术中提出的问题。

[0006] 为实现上述目的,本发明提供如下技术方案:一种用于人工智能视频应用的边缘计算平台,包括平台硬件部分,所述硬件部分包括一个多核多线程 CPU、一个内存、一个FLASH闪存、一个网络通讯模块、一个或多个AI芯片以及硬件接口,所述内存、FLASH闪存、网络通讯模块和AI芯片均与多核多线程CPU电性连接,其中,

[0007] 多核多线程CPU用于完成摄像头视频流数据的解码,AI芯片的任务请求,多任务的执行和调度等功能;

[0008] FLASH用于存储平台的配置参数和AI模块运行的数据结果;

[0009] AI芯片用于支持多路高清摄像头高帧率实时处理,可以支持多线程异步工作模式,以及多个芯片协同工作模式;

[0010] 网络通讯模块用于与云端连接获取相关数据。

[0011] 作为优选,所述硬件接口包括2个USB3.0接口、2个以太网RJ45接口、一个显示HDMI接口、一个串行RS232接口、一个SD卡接口、一个直流12V 电源接口。

[0012] 作为优选,所述网络通讯模块包括一个以太网控制器和一个4G模块。

[0013] 作为优选,所述网络通讯模块是一个PCI总线应用的集成的千兆级以太网控制器,该终端引出两路RJ45接口,其中一个具有POE功能,即可以通过以太网线向终端设备供电。

[0014] 作为优选,所述AI芯片为深度神经网络推理的神经计算芯片。

[0015] 一种用于人工智能视频应用的边缘计算平台的工作方法，其特征在于：包括一下步骤：

[0016] S1、通过一个中央交换机，该终端可以与多个摄像头组成一个局域网，通过局域网，终端可以接收多路摄像头RTSP视频流数据，并在终端内解码视频流数据，调用AI芯片进行图片的检测与识别，将输出的场景结构化数据发送给云端或存在终端FLASH中，云端服务器可以将数据结果分发给用户；

[0017] S2、同时云端也可以存取磁盘录像机的视频数据，用于服务器端数据的分析和网络模型的训练。

[0018] 作为优选，所述S1的视频处理中存在单路视频单个视频处理单元、多路视频单个视频处理单元和多路视频多个视频处理单元的三种情况。

[0019] 作为优选，所述单路视频单个视频处理单元处理时多个AI任务依次向VPU 发出推导请求，VPU推导完成将计算结果返回给CPU，下一个AI任务再向VPU 发出推导请求，如此进行，此时VPU不会出现忙的情况，因为AI任务是依次执行的，在VPU推导的时候，此时CPU处于等待状态；所述多路视频单个视频处理单元处理时SHELL外壳部分为每一路视频流开启一个处理线程，每个处理线程相互独立；所述多路视频多个视频处理单元处理时通过VPU的API 函数可以指定对应的VPU处理对应的视频流。

[0020] 与现有技术相比，本发明的有益效果是：本发明所述一种用于人工智能视频应用的边缘计算平台，通过采用硬件神经网络加速器和专用的软件开发工具包，以及多个加速器协同工作，可以将人工智能视频应用推向实时化。

附图说明

[0021] 图1为本发明硬件结构示意图和硬件接口图；

[0022] 图2为本发明一般化功能应用图；

[0023] 图3为本发明AI核心模块处理单路视频和单个VPU的流程图；

[0024] 图4为本发明AI核心模块处理多路视频和单个VPU的流程图。

具体实施方式

[0025] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例仅仅是本发明一部分实施例，而不是全部的实施例。基于本发明中的实施例，本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其他实施例，都属于本发明保护的范围。

[0026] 请参阅图1-4，本发明提供一种技术方案：一种用于人工智能视频应用的边缘计算平台，包括平台硬件部分，所述硬件部分包括一个多核多线程CPU、一个内存、一个FLASH闪存、一个网络通讯模块、一个或多个AI芯片以及硬件接口，所述内存、FLASH闪存、网络通讯模块和AI芯片均与多核多线程CPU 电性连接，其中，

[0027] 多核多线程CPU用于完成摄像头视频流数据的解码，AI芯片的任务请求，多任务的执行和调度等功能；

[0028] FLASH用于存储平台的配置参数和AI模块运行的数据结果；

[0029] AI芯片用于支持多路高清摄像头高帧率实时处理，可以支持多线程异步工作模

式,以及多个芯片协同工作模式;

[0030] 网络通讯模块用于与云端连接获取相关数据。

[0031] 其中,所述硬件接口包括2个USB3.0接口、2个以太网RJ45接口、一个显示HDMI接口、一个串行RS232接口、一个SD卡接口、一个直流12V电源接口。

[0032] 其中,所述网络通讯模块包括一个以太网控制器和一个4G模块,其上网方式可以是以太网和4G无线上网两种方式,一个或多个AI芯片。

[0033] 其中,所述网络通讯模块是一个PCI总线应用的集成的千兆级以太网控制器,该终端引出两路RJ45接口,其中一个具有POE功能,即可以通过以太网线向终端设备供电。

[0034] 其中,所述AI芯片为深度神经网络推理的神经计算芯片,其采用的神经计算芯片是IntelMovidiusXMA2485,其具有高性能、低功耗的特点,能支持每秒1万亿次运算(TOPS),可以支持多路高清摄像头高帧率实时处理,可以支持多线程异步工作模式,以及多个芯片协同工作模式,单芯片可以同时接受多个推理请求;适用于该AI芯片的软件工具开发包可以优化神经网络模型,在保证模型精度的条件下,进一步提高性能,它具有高性能、低功耗的特点,能支持每秒1万亿次运算(TOPS),芯片内置4G内存,可以支持多路高清摄像头高帧率实时处理,可以支持多线程异步工作模式,以及多个芯片协同工作模式,单芯片可以同时接受多个推理请求;适用于该AI芯片软件工具开发包可以优化神经网络模型,在保证模型精度的条件下,可以进一步提高性能;借助于AI芯片强大的性能,可以在边缘侧加速深度学习推理,将人工智能应用推向实时化。

[0035] 一种用于人工智能视频应用的边缘计算平台的工作方法,包括一下步骤:

[0036] S1、通过一个中央交换机,该终端可以与多个摄像头组成一个局域网,通过局域网,终端可以接收多路摄像头RTSP视频流数据,并在终端内解码视频流数据,调用AI芯片进行图片的检测与识别,将输出的场景结构化数据发送给云端或存在终端FLASH中,云端服务器可以将数据结果分发给用户;

[0037] S2、同时云端也可以存取磁盘录像机的视频数据,用于服务器端数据的分析和网络模型的训练。

[0038] 其中,所述S1的视频处理中存在单路视频单个视频处理单元、多路视频单个视频处理单元和多路视频多个视频处理单元的三种情况。

[0039] 其中,所述单路视频单个视频处理单元处理时多个AI任务依次向VPU发出推导请求,VPU推导完成将计算结果返回给CPU,下一个AI任务再向VPU 发出推导请求,如此进行,此时VPU不会出现忙的情况,因为AI任务是依次执行的,在VPU推导的时候,此时CPU处于等待状态;所述多路视频单个视频处理单元处理时SHELL外壳部分为每一路视频流开启一个处理线程,每个处理线程相互独立;所述多路视频多个视频处理单元处理时通过VPU的API函数可以指定对应的VPU处理对应的视频流。

[0040] 具体的,使用本发明时,该平台的一般化应用场景概况如下:通过一个中央交换机,该终端可以与多个摄像头组成一个局域网,通过局域网,终端可以接收多路摄像头RTSP视频流数据,并在终端内解码视频流数据,调用AI 芯片进行图片的检测与识别,将输出的场景结构化数据发送给云端或存在终端FLASH中,云端服务器可以将数据结果分发给用户;同时云端也可以存取磁盘录像机的视频数据,用于服务器端数据的分析和网络模型的训练;

[0041] 本平台在单路视频单个VPU(VisionProcessingUnit)的情况下,多个 AI任务依次向VPU发出推导请求,VPU推导完成将计算结果返回给CPU,下一个AI任务再向VPU发出推导请求,如此进行,此时VPU不会出现忙的情况,因为AI任务是依次执行的,在VPU推导的时候,此时CPU处于等待状态(同步模式);在异步模式下,当前处理线程进行VPU推导,其他线程可以正常进行,不会造成其他线程的阻塞;等所有任务完成后再从视频流中取下当前时刻图片,进行下一次AI任务处理;

[0042] 本平台在多路视频单个VPU的情况下,SHELL外壳部分为每一路视频流开启一个处理线程,每个处理线程相互独立,每个线程执行的任务,但是此时会出现视频流数大于VPU个数,所以可能出现两个线程同时申请VPU推导请求,或者当线程申请时VPU处于忙碌状态,此时VPU会形成一个处理队列,将新的请求添加到队列末尾,依次处理每个请求,并将处理结果返回给相应的线程,其他未得到处理的线程则等待,直到返回该线程的处理结果;

[0043] 平台也支持多路视频多个VPU,通过VPU的API函数可以指定对应的VPU 处理对应的视频流,其中每个VPU的处理过程多路视频单个VPU的情况相似,这样通过多个VPU,可以实现真正的并行推导多路视频流,进一步扩大吞吐量,加快视频的处理速度。

[0044] 尽管已经示出和描述了本发明的实施例,对于本领域的普通技术人员而言,可以理解在不脱离本发明的原理和精神的情况下可以对这些实施例进行多种变化、修改、替换和变型,本发明的范围由所附权利要求及其等同物限定。

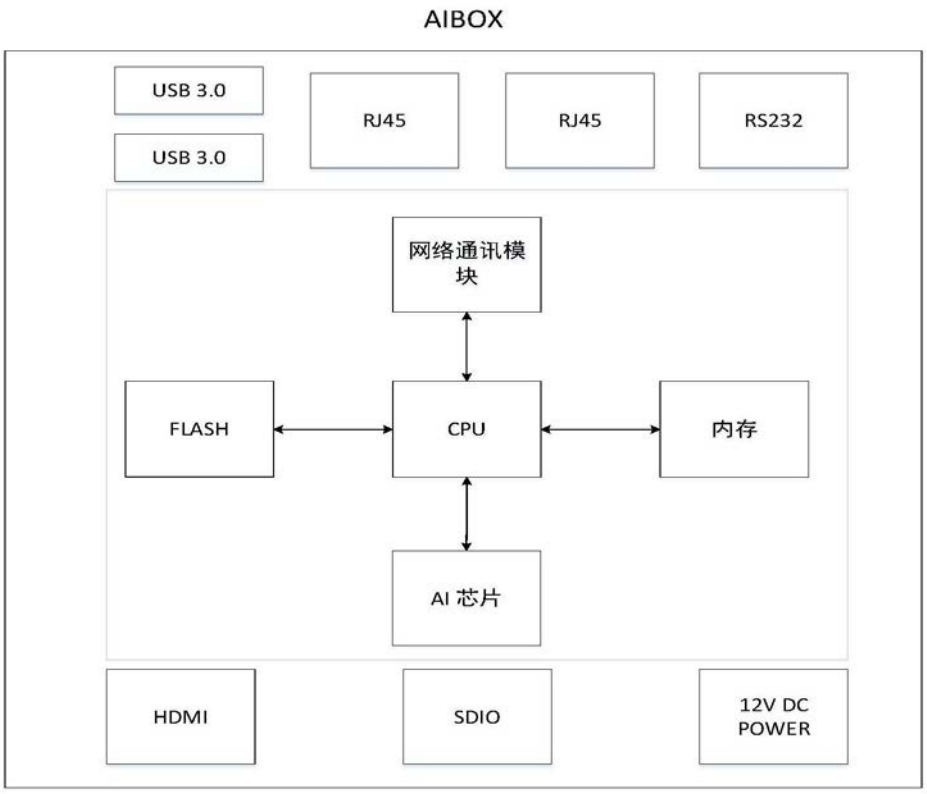

AIBOX
USB 3.0
USB 3.0
RJ45
RJ45
RS232
网络通讯模块
FLASH
CPU
内存
AI 芯片
HDMI
SDIO
12V DC POWER

图1

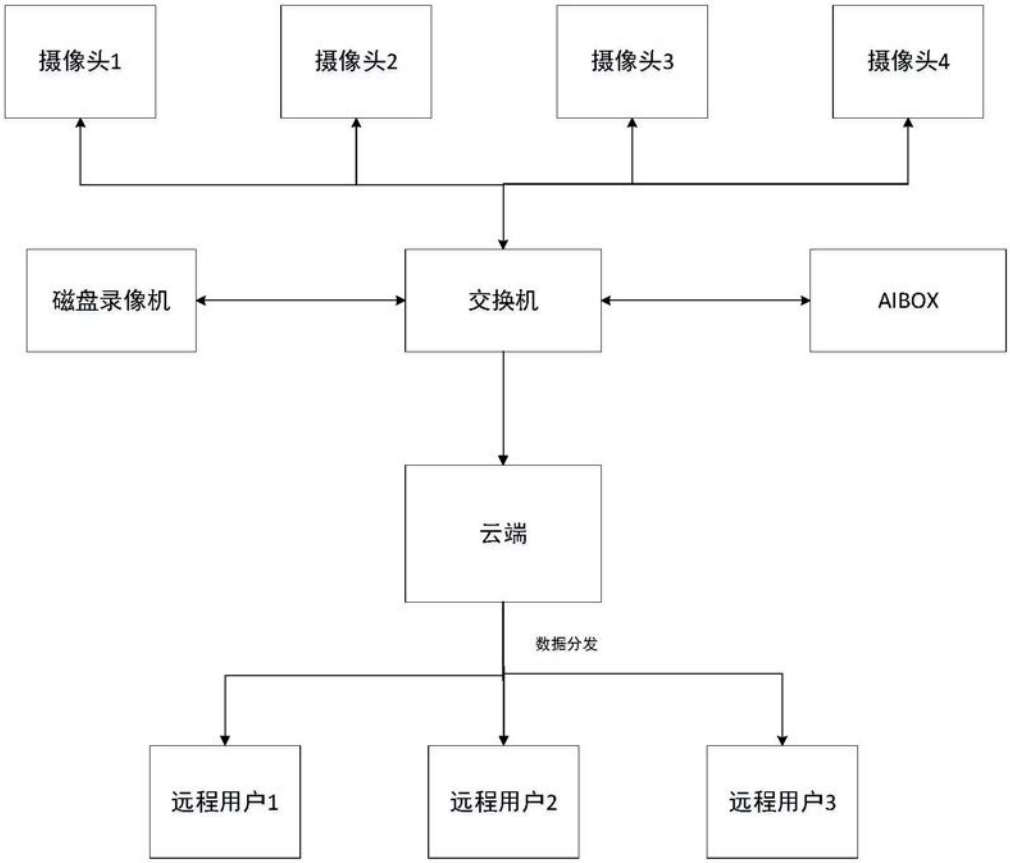

摄像头1
摄像头2
摄像头3
摄像头4
磁盘录像机
交换机
AIBOX
云端
数据分发
远程用户1
远程用户2
远程用户3

图2

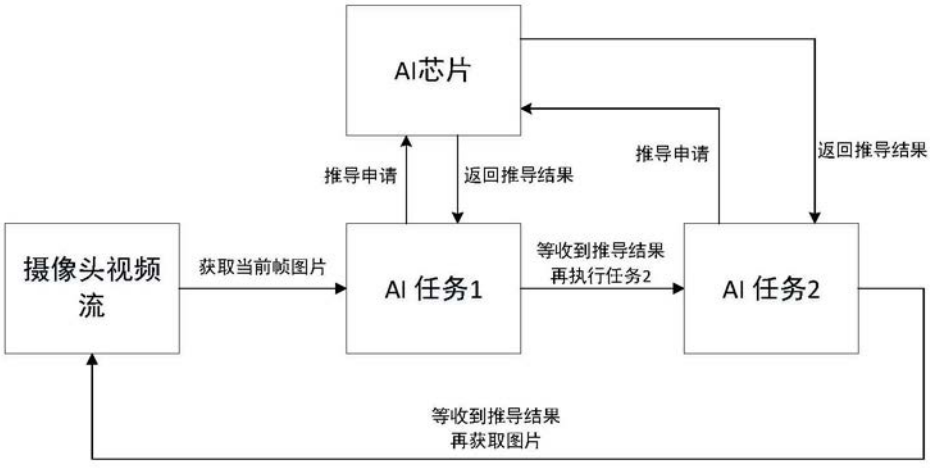

AI芯片
推导申请
返回推导结果
推导申请
返回推导结果
摄像头视频流
获取当前帧图片
AI 任务1
等收到推导结果
再执行任务2
AI 任务2
等收到推导结果
再获取图片

图3

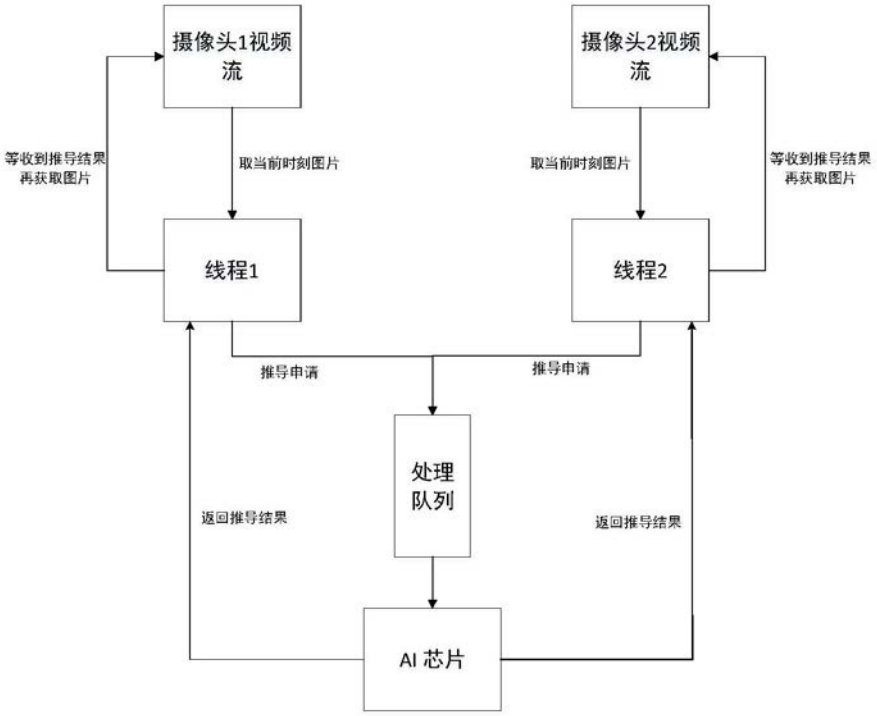

摄像头1视频流
摄像头2视频流
等收到推导结果再获取图片
取当前时刻图片
取当前时刻图片
等收到推导结果再获取图片
线程1
线程2
推导申请
推导申请
处理队列
返回推导结果
返回推导结果
AI 芯片

图4