



(12)发明专利申请

(10)申请公布号 CN 111506434 A

(43)申请公布日 2020.08.07

(21)申请号 202010610916.4

(22)申请日 2020.06.30

(71)申请人 腾讯科技(深圳)有限公司

地址 518057 广东省深圳市南山区高新区
科技中一路腾讯大厦35层

(72)发明人 严石伟 李明耀 丁凯 蒋楠

(74)专利代理机构 深圳翼盛智成知识产权事务
所(普通合伙) 44300

代理人 李汉亮

(51)Int.Cl.

G06F 9/50(2006.01)

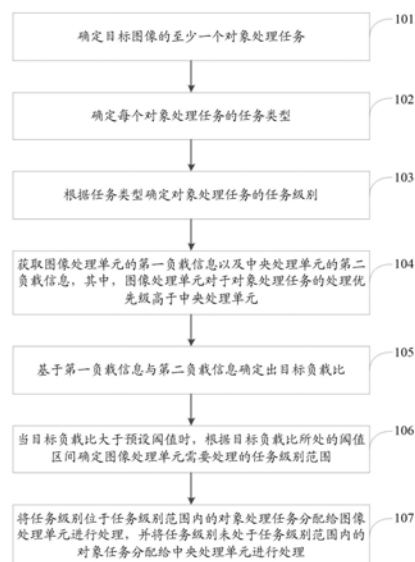
权利要求书3页 说明书18页 附图6页

(54)发明名称

一种任务处理方法、装置及计算机可读存储
介质

(57)摘要

本申请实施例公开了一种任务处理方法、装置及计算机可读存储介质。本申请实施例通过确定目标图像的至少一个对象处理任务；确定每个对象处理任务的任务类型；根据任务类型确定对象处理任务的任务级别；获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息；基于第一负载信息与第二负载信息确定出目标负载比；当目标负载比大于预设阈值时，根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围；将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理，并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。可以提高任务处理的效率。



1. 一种任务处理方法,其特征在于,包括:
 - 确定目标图像的至少一个对象处理任务;
 - 确定每个所述对象处理任务的任务类型;
 - 根据所述任务类型确定所述对象处理任务的任务级别;
 - 获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,所述图像处理单元对于对象处理任务的处理优先级高于所述中央处理单元;
 - 基于所述第一负载信息和所述第二负载信息确定出目标负载比;
 - 当所述目标负载比大于预设阈值时,根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任务级别范围;
 - 将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给所述中央处理单元进行处理。
2. 根据权利要求1所述的方法,其特征在于,所述将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理的步骤,包括:
 - 获取对象处理任务对应的子任务队列,所述子任务队列中包括多个子任务;
 - 将相同子任务属性的子任务进行任务聚合,生成任务聚合后的目标子任务队列;
 - 将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合,并通过所述图像处理单元对矩阵聚合后的目标子任务队列进行处理。
3. 根据权利要求2所述的方法,其特征在于,在生成任务聚合后的目标子任务队列的步骤之后,还包括:
 - 当检测到所述目标子任务队列的子任务数量满足预设数量时,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤;
 - 当检测到所述目标子任务队列的子任务数量不满足预设数量时,在预设时间内等待相同子任务属性的子任务进行任务聚合,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤。
4. 根据权利要求2所述的方法,其特征在于,所述将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤,包括:
 - 获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;
 - 将多个数据向量进行聚合处理,得到数据矩阵;
 - 将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列。
5. 根据权利要求4所述的方法,其特征在于,所述通过所述图像处理单元对矩阵聚合后的目标子任务队列进行处理的步骤,包括:
 - 根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;
 - 将所述共享数据库、所述数据矩阵和目标特征矩阵通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述数据矩阵和目标特征矩阵进行处理。
6. 根据权利要求2所述的方法,其特征在于,所述通过所述图像处理单元对矩阵聚合后的目标子任务队列进行处理的步骤,包括:

根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;

将所述共享数据库和矩阵聚合后的目标子任务队列通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述矩阵聚合后的目标子任务队列进行处理。

7.根据权利要求1所述的方法,其特征在于,所述任务级别包括第一任务级别,第二任务级别以及第三任务级别;

所述根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任务级别范围的步骤,包括:

将大于所述预设阈值的阈值区间分为第一阈值区间和第二阈值区间,所述第二阈值区间大于所述第一阈值区间;

若所述目标负载比所处的阈值区间为所述第一阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别和第二任务级别;

若所述目标负载比所处的阈值区间为所述第二阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别。

8.根据权利要求1所述的方法,其特征在于,所述基于所述第一负载信息与所述第二负载信息确定出目标负载比的步骤,包括:

对所述第一负载信息和所述第二负载信息进行比值计算,得到所述第一负载信息与所述第二负载信息之间的目标负载比。

9.根据权利要求1-8任一项所述的方法,其特征在于,还包括:

当所述目标负载比小于预设阈值时,将所有对象处理任务分配至所述图像处理单元进行处理。

10.一种任务处理装置,其特征在于,包括:

第一确定单元,用于确定目标图像的至少一个对象处理任务;

第二确定单元,用于确定每个所述对象处理任务的类型;

第三确定单元,用于根据所述任务类型确定所述对象处理任务的级别;

获取单元,用于获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,所述图像处理单元对于对象处理任务的优先级高于所述中央处理单元;

第四确定单元,用于基于所述第一负载信息与所述第二负载信息确定出目标负载比;

第五确定单元,用于当所述目标负载比大于预设阈值时,根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的级别范围;

分配单元,用于将任务级别位于级别范围内的对象处理任务分配给所述图像处理单元进行处理,并将任务级别未处于级别范围内的对象任务分配给所述中央处理单元进行处理。

11.根据权利要求10所述的装置,其特征在于,所述分配单元,包括:

获取子单元,用于获取对象处理任务对应的子任务队列,所述子任务队列中包括多个子任务;

聚合子单元,用于将相同子任务属性的子任务进行任务聚合,生成任务聚合后的目标子任务队列;

处理子单元,用于将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合,并通

过所述图像处理单元对矩阵聚合后的目标子任务队列进行处理。

12. 根据权利要求11所述的装置,其特征在於,所述分配单元,还包括:

第一执行子单元,用于当检测到所述目标子任务队列的子任务数量满足预设数量时,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤;

第二执行子单元,用于当检测到所述目标子任务队列的子任务数量不满足预设数量时,在预设时间内等待相同子任务属性的子任务进行任务聚合,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤。

13. 根据权利要求11所述的装置,其特征在於,所述处理子单元,用于:

获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;

将多个数据向量进行聚合处理,得到数据矩阵;

将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列。

14. 根据权利要求11所述的装置,其特征在於,所述处理子单元,用于:

获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;

将多个数据向量进行聚合处理,得到数据矩阵;

将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列;

根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;

将所述共享数据库、所述数据矩阵和目标特征矩阵通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述数据矩阵和目标特征矩阵进行处理。

15. 一种计算机可读存储介质,其特征在於,所述计算机可读存储介质存储有多条指令,所述指令适于处理器进行加载,以执行权利要求1至9任一项所述的任务处理方法的步骤。

一种任务处理方法、装置及计算机可读存储介质

技术领域

[0001] 本申请涉及计算机技术领域，具体涉及一种任务处理方法、装置及计算机可读存储介质。

背景技术

[0002] 随着信息技术的高速发展，传统零售、电子商务逐渐难以满足人们的消费需求，以大数据、云计算、人工智能等技术为代表的创新科技在零售领域展现出越来越重要的作用，衍生出智慧零售方式。在智慧零售领域中，可以通过人工智能技术中的深度学习计算框架实时获取用户信息进行处理，从而得到客流信息。

[0003] 相关技术中，通过深度学习计算框架对用户信息进行处理时，需要同时对多个图像处理的任務进行处理，为了保证运算速度，可以将多个任务集中在GPU (Graphics Processing Unit, 图形处理器) 上进行针对性计算处理。

[0004] 在对相关技术的研究和实践过程中，本申请的发明人发现，现有技术中，将多个图像处理的任務全部集中在GPU上进行处理，会导致GPU负载量过大，造成任务处理阻塞，影响任务处理速度，从而导致任务处理效率过低。

发明内容

[0005] 本申请实施例提供一种任务处理方法、装置及计算机可读存储介质，可以提高任务处理效率。

[0006] 本申请实施例提供了一种任务处理方法，包括：

- 确定目标图像的至少一个对象处理任务；

- 确定每个所述对象处理任务的任務类型；

- 根据所述任務类型确定所述对象处理任务的任務级别；

- 获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息，其中，所述图像处理单元对于对象处理任务的處理优先级高于所述中央处理单元；

- 基于所述第一负载信息与所述第二负载信息确定出目标负载比；

- 当所述目标负载比大于预设阈值时，根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任務级别范围；

- 將任務级别位于任務级别范围内的对象处理任务分配给所述图像处理单元进行处理，并将任務级别未处于任務级别范围内的对象任务分配给所述中央处理单元进行处理。

[0007] 相应的，本申请实施例还提供了一种任务处理装置，包括：

- 第一确定单元，用于确定目标图像的至少一个对象处理任务；

- 第二确定单元，用于确定每个所述对象处理任务的任務类型；

- 第三确定单元，用于根据所述任務类型确定所述对象处理任务的任務级别；

- 获取单元，用于获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息，其中，所述图像处理单元对于对象处理任务的處理优先级高于所述中央处理单元；

第四确定单元,用于基于所述第一负载信息与所述第二负载信息确定出目标负载比;

第五确定单元,用于当所述目标负载比大于预设阈值时,根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任务级别范围;

分配单元,用于将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给所述中央处理单元进行处理。

[0008] 在一些实施例中,处理子单元,可以用于:获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;将多个数据向量进行聚合处理,得到数据矩阵;将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列;根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;

将所述共享数据库和矩阵聚合后的目标子任务队列通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述矩阵聚合后的目标子任务队列进行处理。

[0009] 在一些实施例中,第五确定单元,可以包括:

划分子单元,用于将大于所述预设阈值的阈值区间分为第一阈值区间和第二阈值区间,所述第二阈值区间大于所述第一阈值区间;

第一确定子单元,用于若所述目标负载比所处的阈值区间为所述第一阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别和第二任务级别;

第二确定子单元,用于若所述目标负载比所处的阈值区间为所述第二阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别。

[0010] 在一些实施例中,第四确定单元,可以包括:

计算子单元,用于对所述第一负载信息和所述第二负载信息进行比值计算,得到所述第一负载信息与所述第二负载信息之间的目标负载比。

[0011] 在一些实施例中,该任务处理装置还可以包括:

处理单元,用于当所述目标负载比小于预设阈值时,将所有对象处理任务分配至所述图像处理单元进行处理。

[0012] 相应的,本申请实施例还提供了一种计算机可读存储介质,所述计算机可读存储介质存储有多条指令,所述指令适于处理器进行加载,以执行如上所述的任务处理方法。

[0013] 本申请实施例通过确定目标图像的至少一个对象处理任务;确定每个对象处理任务的任务类型;根据任务类型确定对象处理任务的任务级别;获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息;基于第一负载信息与第二负载信息确定出目标负载比;当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围;将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。以此,通过确定对象处理任务的任务类型,监测不同的处理单元之间形成的负载比,在负载比达到一定条件时,根据负载比将不同任务级别的对象处理任务分配至不同的处理单元进行并行处理,缓解处理单元的处理压力,极大的提升了任务处理的效率。

附图说明

[0014] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0015] 图1为本申请实施例提供的任务处理系统的场景示意图。

[0016] 图2为本申请实施例提供的第一种任务处理方法的流程示意图。

[0017] 图3为本申请实施例提供的第二种任务处理方法的流程示意图。

[0018] 图4为本申请实施例提供的一种任务处理方法任务调度示意图。

[0019] 图5为本申请实施例提供的一种任务处理方法子任务聚合示意图。

[0020] 图6为本申请实施例提供的第一种任务处理装置的结构框图。

[0021] 图7为本申请实施例提供的第二种任务处理装置的结构框图。

[0022] 图8为本申请实施例提供的第三种任务处理装置的结构框图。

[0023] 图9为本申请实施例提供的第四种任务处理装置的结构框图。

[0024] 图10为本申请实施例提供的第五种任务处理装置的结构框图。

[0025] 图11为本申请实施例提供的第六种任务处理装置的结构框图。

[0026] 图12为本申请实施例提供的服务器的结构示意图。

具体实施方式

[0027] 下面将结合本申请实施例中的附图,对本申请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申请的一部分实施例,而不是全部的实施例。基于本申请中的实施例,本领域技术人员在没有作出创造性劳动前提下所获得的所有其他实施例,都属于本申请保护的范围。

[0028] 本申请实施例提供一种任务处理方法、装置、计算机可读存储介质及终端。具体地,本申请实施例提供适用于计算机设备的任务处理装置。其中,该计算机设备可以为终端或服务器等设备,服务器可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器。终端可以是智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等,但并不局限于此。终端以及服务器可以通过有线或无线通信方式进行直接或间接地连接,本申请实施例在此不做限制。

[0029] 请参阅图1,图1为本申请实施例所提供的任务处理系统的场景示意图,包括摄像头设备、和网络设备(该任务处理系统还可以包括多个摄像头设备,摄像头设备具体个数在此处不作限定),摄像头设备与网络设备之间可以通过通信网络连接,该通信网络,可以包括无线网络以及有线网络,其中无线网络包括无线广域网、无线局域网、无线城域网、以及无线个人网中的一种或多种的组合。网络中包括路由器、网关等等网络实体,图中并未示出。摄像头设备可以通过通信网络与网络设备进行信息交互,比如摄像头设备可以实时拍摄视频流,将该视频流通过通信网络发送至网络设备中。

[0030] 该任务处理系统可以包括任务处理装置,该任务处理装置具体可以集成在网络设

备如终端或服务器等设备中,如图1所示,该网络设备接收摄像头发送的视频流,在对视频流进行处理时接收对象处理任务确定对象处理任务的任务类型,根据对象处理任务的任务类型确定对象处理任务的任务级别,同时获取图像处理单元的第一负载信息和中央处理单元的第二负载信息,并计算第一负载信息与第二负载信息的比值,确定负载比,根据对象处理任务的任务级别和图像处理单元与中央处理单元的负载比,将对象处理任务分配至图像处理单元和中央处理单元进行处理。可以同时利用图像处理单元和中央处理单元对对象处理任务进行处理,加快任务处理速度,从而提高任务处理效率。

[0031] 需要说明的是,图1所示的任务处理系统的场景示意图仅仅是一个示例,本申请实施例描述的任务处理系统以及场景是为了更加清楚的说明本申请实施例的技术方案,并不构成对于本申请实施例提供的技术方案的限定,本领域普通技术人员可知,随着任务处理系统的演变和新业务场景的出现,本申请实施例提供的技术方案对于类似的技术问题,同样适用。

[0032] 基于上述问题,本申请实施例提供第一种任务处理方法、装置及计算机可读存储介质,可以有效提高服务器的任务处理速度。以下分别进行详细说明。需说明的是,以下实施例的描述顺序不作为对实施例优选顺序的限定。

[0033] 本申请实施例提供一种任务处理方法,该方法可以由终端或服务器执行,本申请实施例以任务处理方法由服务器执行为例来进行说明。

[0034] 如图2所示,图2为本申请实施例提供的第一种任务处理方法的流程示意图。该任务处理方法的具体流程可以如下:

101、确定目标图像的至少一个对象处理任务。

[0035] 其中,本申请实施例的场景可以应用在大型商场或超市中,可以在该大型商场或超市的主要位置设置监控摄像头,用于实时拍摄顾客的行为信息,生成视频,通过对视频中的顾客行为信息进行处理,为商场和店铺提供客流数据以及顾客身份和游逛轨迹等信息。其中,在对顾客行为信息进行处理时,系统后台服务通过对软件开发工具包(SDK,Software Development Kit)有效封装、请求任务的合理调度、输出数据的高效、高可用的处理为客户输出最终的客流、身份等信息。

[0036] 该对象处理任务可以为图像处理任务,图像处理可以是指通过提取摄像头设备获取的视频的视频帧进行相应的图像处理。

[0037] 102、确定每个对象处理任务的任务类型。

[0038] 在接收到对象处理任务之后,可以确定出对象处理任务的任务类型。该对象处理任务的任务类型可以为图像处理任务的类别,比如,图像处理任务可以包括人脸、人体检测,人脸、人体跟踪,人脸、人体优选,人脸配准,人体关键点,人脸人体绑定,人脸人体提特征、提属性、检索等的多种图像处理类型。

[0039] 103、根据任务类型确定对象处理任务的任务级别。

[0040] 其中,对象处理任务的任务类型可以包括多种,每一种任务类型的计算复杂度都不相同,例如,人脸、人体轨迹内各抓拍提特征、提属性(人脸可以是获取性别、年龄;人体可以是获取提袋、衣着)、轨迹检索等。SDK在完成提特征、提属性、跟踪和检索等功能时,需要考虑人脸的多个要素,因此人脸人体提特征、提属性、检索的任务的计算复杂度较高,可以为计算密集型任务。对于人脸人体检测、绑定:人脸检测是获取人脸框,人体检测是获取人

体框,绑定则是将获取到的人脸与对应的人体进行绑定,确定为同一人,因此,检测和绑定的任务的计算复杂度一般,因此人脸人体检测、绑定可以为一般计算密集型任务。而对于人脸人体优选、关键点、配准:优选是完成人脸质量分计算,关键点是获取人脸或人体坐标,配准是获取人脸五点信息,因此,人脸人体优选、关键点、配准的任务的计算复杂度较低,可以为非计算密集型任务。

[0041] 在一些实施例中,步骤根据任务类型的计算复杂度确定对象处理任务的任务级别,可以包括:

基于该任务类型将该对象处理任务分为第一任务级别、第二任务级别以及第三任务级别,其中,该第一任务级别优于该第二任务级别,该第二任务级别优于该第三任务级别。

[0042] 其中,可以基于该任务类型的计算复杂度将该对象处理任务分为第一任务级别、第二任务级别以及第三任务级别。该第一任务级别的计算复杂度可以高于第二任务级别的计算复杂度,第二任务级别的计算复杂度可以高于第三任务级别的计算复杂度,每一级别等级对应相应类型的处理任务。

[0043] 例如,对象处理任务可以包括跟踪、提特征、提属性、检索、检测、绑定、优选、配准、关键点等类型。其中,跟踪、提特征、提属性、检索类型的任务涉及的计算复杂度较高,则可以将跟踪、提特征、提属性、检索这些任务类型的等级级别确定为第一任务级别;检测、绑定类型的任务涉及的计算复杂度一般,则可以将检测、绑定这些任务类型的等级级别确定为第二任务级别;优选、配准、关键点类型的任务涉及的计算复杂度较低,则可以将优选、配准、关键点这些任务类型的等级级别确定为第三任务级别。例如,如表1所示:

表1

任务类型	级别等级	等级处理措施
跟踪, 特征, 属性, 检索	第一任务级别	集中在 GPU 上进行处理
检测, 绑定	第二任务级别	集中在 GPU 或者 CPU 上进行处理, 其中, 在 GPU 上进行处理时速度更快
优选, 配准, 关键点	第三任务级别	集中在 GPU 或者 CPU 上进行处理, 其中, 在 GPU 或者 CPU 上进行处理的处理速度相同

104、获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,图像处理单元对于对象处理任务的处理优先级高于中央处理单元。

[0044] 其中,该图像处理单元和中央处理单元可以为GPU(Graphics Processing Unit,

图形处理器)和CPU(Central Processing Unit / Processor,中央处理器)。

[0045] GPU又称显示核心、视觉处理器、显示芯片,是一种专门在个人电脑、工作站、游戏机和一些移动设备(如平板电脑、智能手机等)上图像运算工作的微处理器。用途是将计算机系统所需要的显示信息进行转换驱动,并向显示器提供行扫描信号,控制显示器的正确显示,是连接显示器和个人电脑主板的重要元件,也是“人机对话”的重要设备之一。显卡作为电脑主机里的一个重要组成部分,承担输出显示图形的任务。

[0046] CPU是电子计算机的主要设备之一,电脑中的核心配件。其功能主要是解释计算机指令以及处理计算机软件中的数据。电脑中所有操作都由CPU负责读取指令,对指令译码并执行指令的核心部件。

[0047] 在本申请实施例中,由于对象处理任务大都为图像处理任务,主要集中在GPU处理器上进行处理,少量的对象处理任务可以通过CPU协助完成。

[0048] 以此,图像处理单元的第一负载信息可以为GPU的负载,中央处理单元的第二负载信息可以为CPU 的负载。负载指的是处理器在一定时间内同时处理的任务数量,比如,在一定时间内,处理器可以同时处理的任务数量可以为10个,实际处理的任务数量可以为4个,则当前处理器的负载可以为4:10,也即0.4,该负载信息反映了处理器的处理繁忙状态,该负载信息越高,说明处理器越繁忙,该负载信息越低,说明处理器越空闲。

[0049] 105、基于第一负载信息和第二负载信息确定出目标负载比。

[0050] 其中,根据第一处理器的负载以及第二处理器的负载,确定目标负载比,该目标负载比可以体现出图像处理单元和中央处理单元之间的负载相对关系,该目标负载比较高时,说明图像处理单元负载程度较大,该中央处理单元负载程度相对较低,该目标负载比较低时,说明图像处理单元负载程度较低,该中央处理单元负载程度较高。在一些实施例中,步骤根据该第一负载信息和第二负载信息确定出目标负载比,可以包括:

计算该第一负载信息与该第二负载信息的比值,确定该目标负载比。

[0051] 其中,在获取图像处理单元的第一负载,与中央处理单元的第二负载之后,可以计算第一负载与第二负载的比值。例如,图像处理单元的第一负载可以为0.4,中央处理单元的第二负载可以为0.4,则可以确定目标负载比为1。

[0052] 106、当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围。

[0053] 其中,在相关技术中,该对象处理任务均运行在GPU上,而对于系统上的CPU资源基本上没有进行利用,导致对CPU资源的浪费,实际上,一些计算复杂度较低的对象处理任务,例如,跟踪、优先和配准等任务在CPU上进行计算与在GPU上进行计算的速度差不多,所以在GPU繁忙时,可以将一些计算复杂度较低的任务分配至CPU进行计算,增加任务处理的效率,该预设阈值即为界定GPU繁忙的值,例如为1.1,当目标负载比超过该预设阈值时,说明GPU相对于CPU处于负载度更高的情况。

[0054] 进一步的,在目标负载比只高出预设阈值不多的情况下,说明GPU相对CPU负载超出的程度不高,可以将一些计算复杂度低的任务级别的对象处理任务分配至中央处理单元进行处理。在目标负载比高于预设阈值较多的情况下,说明GPU相对CPU负载超出的程度较高,可以将一些计算复杂度低的任务级别的对象处理任务和一些计算复杂度一般的任务级别的对象处理任务均分配至中央处理单元进行处理,更好的缓解GPU的压力,增加信息计算

的效率。

[0055] 在一些实施例中,假设该第一负载可以为0.3,第二负载可以为0.6,则可以确定目标负载比为0.5,小于预设阈值。在目标负载比小于预设阈值的情况下,说明GPU负载低于CPU负载,该GPU处于相对空闲的状态,则可以将一些计算复杂度低的任务级别的对象处理任务,计算复杂度一般的任务级别的对象处理任务,以及计算复杂度高的对象处理任务分配至图像处理单元进行处理,在保证图像处理单元正常工作的情况下,提高了任务处理的效率。

[0056] 在一些实施例中,步骤根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任務级别范围,可以包括:

(1)将大于预设阈值的阈值区间分为第一阈值区间和第二阈值区间,第二阈值区间大于所述第一阈值区间;

(2)若目标负载比所处的阈值区间为第一阈值区间,则确定所述图像处理单元需要处理的任務级别范围包括第一任务级别;

(3)若目标负载比所处的阈值区间为所述第二阈值区间,则确定图像处理单元需要处理的任務级别范围包括第一任务级别和第二任务级别。

[0057] 其中,可以将大于预设阈值的阈值区间划分为第一阈值区间和第二阈值区间,其中,第一阈值区间可以小于第二阈值区间。比如,预设阈值可以为1.1,则可以设置第一阈值区间为(1.1, 2),第二阈值区间可以为 $[2, +\infty)$,其中(1.1, 2)表示大于1且小于2的阈值区间, $[2, +\infty)$ 表示大于等于2的阈值区间。

[0058] 基于此,当目标负载比处于第一阈值区间时,表示图像处理单元的负载大于中央处理单元的负载,但图像处理单元的使用率仍处于正常状态。此时,可以确定图像处理单元需要处理的任務级别范围包括第一任务级别和第二任务级别。

[0059] 在一些实施例中,当目标负载比处于第二阈值区间时,表示图像处理单元的负载远大于中央处理单元的负载,图像处理单元的使用率超出正常状态。此时,可以确定图像处理单元需要处理的任務级别范围包括第一任务级别。

[0060] 107、将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。

[0061] 其中,当目标负载比处于第一阈值区间时,可以将第一任务级别和第二任务级别的对象处理任务分配至图像处理单元进行处理,将第三任务级别的对象处理任务分配至中央处理单元进行处理,通过图像处理单元与中央处理单元共同对对象处理任务进行处理,提高任务处理效率。

[0062] 又或者,当目标负载比处于第二阈值区间时,可以将第一任务级别的对象处理任务分配至图像处理单元进行处理,将第二任务级别的对象处理任务和第三任务级别的对象处理任务分配至中央处理单元进行处理。通过将低任务级别的对象处理任务调度至中央处理单元进行处理,实现多任务并行处理时对计算资源的合理利用,提高处理效率。

[0063] 在一些实施例中,步骤将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理,可以包括:

(1.1)获取对象处理任务对应的子任务队列,该子任务队列中包括多个子任务;

(1.2)将相同子任务属性的子任务进行任务聚合,生成任务聚合后的目标子任务队列;

(1.3) 将该目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合,并通过该图像处理单元对矩阵聚合后的目标子任务队列进行处理。

[0064] 其中,对象处理任务可以为多个对象处理任务,每个对象处理任务可以包括多个子任务。基于此,将所有对象处理任务对应的子任务进行集合,可以得到子任务队列。

[0065] 例如,对象处理任务可以包括第一对象处理任务、第二对象处理任务以及第三代处理任务,该第一对象处理任务可以包括第一子任务,第二子任务,第三子任务;第二对象处理任务可以包括第四子任务,第五子任务,第六子任务;第三对象处理任务可以包括第七子任务,第八子任务,第九子任务等。则可以基于所有对象处理任务对应的子任务确定子任务队列,子任务队列中包括:第一子任务,第二子任务,第三子任务,第四子任务,第五子任务,第六子任务,第七子任务,第八子任务,第九子任务。

[0066] 在确定子任务队列之后,可以将子任务队列中相同子任务属性的子任务进行任务聚合,生成聚合后的目标子任务队列。其中,子任务属性可以表示子任务的类型。

[0067] 例如,子任务队列中可以包括第一子任务,第二子任务,第三子任务,第四子任务,第五子任务等,该第一子任务的子任务属性可以为第一属性,该第二子任务的子任务属性可以为第一属性,该第三子任务的属性可以为第一属性,该第四子任务的属性可以为第二属性,该第五子任务的属性可以为第三属性。基于此,可以确定第一子任务、第二子任务以及第三子任务具有相同的子任务属性,可以基于子任务属性,将子任务属性相同的子任务进行聚合,可以得到任务聚合后的目标子任务队列,目标子任务队列包括第一子任务,第二子任务,第三子任务。

[0068] 在确定目标子任务队列之后,可以将目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合,并通过图像处理单元对矩阵聚合后的目标子任务队列进行处理。其中,数据矩阵可以为子任务需要进行处理的多个数据,特征矩阵可以为每个数据的特征信息。例如,每张图像可以包括多个像素点,每个像素点为一个数据,一个像素点数据包括多个参数信息(颜色值,亮度值等),则这多个参数信息可以为每个数据的特征信息。

[0069] 在一些实施例中,步骤通过该图像处理单元对矩阵聚合后的目标子任务队列进行处理,可以包括:

(2.1) 根据该目标子任务队列中的子任务属性,获取该目标子任务队列对应的共享数据库;

(2.2) 将该共享数据库和矩阵聚合后的目标子任务队列通过显存带宽传输至该图像处理单元,基于该共享数据库对该矩阵聚合后的目标子任务队列进行处理。

[0070] 其中,共享数据库可以包括目标任务队列中的子任务在进行计算处理时所需的数据信息。不同任务属性的子任务,所需要的共享数据库可以不同。

[0071] 在相关技术中,每处理一个目标任务队列中的子任务,需要获取一次共享数据库。当目标任务队列中的子任务数量较多时,相应的需要获取多次共享数据库,在加载显存数据至显存带宽时造成显存带宽拥堵,导致任务处理堵塞等待时间过长,影响任务处理的效率。而本申请实施例通过将目标任务队列中的子任务按照预设数量进行聚合处理,减少共享数据库获取次数,在加载显存数据至显存带宽时,保证显存带宽的畅通,降低任务处理堵塞等待的时间,可以提高任务处理效率。

[0072] 在确定目标子任务队列对应的共享数据库之后,可以将目标子任务队列以及目标

子任务队列对应的共享数据库通过显存带宽传输至图像处理单元,之后在图像处理单元中,可以基于共享数据库对目标子任务队列中的子任务进行处理。

[0073] 该显存带宽是指显示芯片与显存之间的数据传输速率,它以字节/秒为单位。显存带宽是决定显卡性能和速度最重要的因素之一。显存,也被叫做帧缓存,它的作用是用来存储显卡芯片处理过或者即将提取的渲染数据。现今,高密度运算由GPU在显卡上完成,由此更加重了对显存的依赖。由于显存在显卡上所起的作用,显然显存的速度和带宽直接影响到显卡的整体速度。

[0074] 在一些实施例中,在步骤生成任务聚合后的目标子任务队列之后,还可以包括:

(3.1) 当检测到该目标子任务队列的子任务数量满足预设数量时,执行将该目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤;

(3.2) 当检测到该目标子任务队列的子任务数量不满足预设数量时,在预设时间内等待相同子任务属性的子任务进行任务聚合,执行将该目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤。

[0075] 其中,预设数量可以为进行聚合处理的最多子任务数量。例如,目标子任务队列中的子任务数量可以为5,预设数量可以为5,此时,则可以将目标子任务队列中的5个子任务进行聚合处理。其中,可以将5个子任务的数据矩阵进行聚合处理,得到聚合后的数据矩阵;将5个子任务的特征矩阵进行聚合处理,得到聚合后的特征矩阵。

[0076] 又例如,目标子任务队列的子任务数量可以为2,预设数量可以为5,此时目标子任务队列的任务数量不满足预设数量,为了提高处理单元的并行处理效率,可以设定预设时间,在预设时间内等待相同子任务属性的子任务进行聚合。预设时间可以为20ms(毫秒),在等待相同子任务属性的子任务进行聚合时,可以存在多种情况:第一种情况,等待时间到达10ms时,相同属性的子任务数量达到了预设数量,则停止等待,将目标子任务队列中的子任务进行聚合处理;第二种情况,等待时间到达预设时间20ms时,相同属性的子任务数量仍未达到预设数量,则停止等待,将目标子任务队列中的子任务进行聚合处理。

[0077] 在一些实施例中,步骤将该目标子任务队列的所有数据矩阵和特征矩阵进行矩阵聚合,可以包括:

(4.1) 获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;

(4.2) 将多个数据向量进行聚合处理,得到数据矩阵;

(4.3) 将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据该数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列。

[0078] 其中,数据向量表示每个子任务需要进行处理的数据,其中每个数据向量对应一个特征矩阵,特征矩阵包括每个数据向量的特征信息。

[0079] 在获取目标任务队列中的每个子任务的数据向量之后,可以将所有子任务的数据向量进行聚合,得到目标子任务队列的数据矩阵;在获取到每个数据向量的特征矩阵之后,可以将所有数据向量的特征矩阵进行聚合处理,得到目标特征矩阵,根据数据矩阵和目标特征矩阵,得到进行矩阵聚合处理后的目标子任务队列。

[0080] 在一些实施例中,步骤通过该图像处理单元对矩阵聚合后的目标子任务队列进行处理,可以包括:

(5.1)根据该目标子任务队列中的子任务属性,获取该目标子任务队列对应的共享数据库;

(5.2)将该共享数据库、该数据矩阵和目标特征矩阵通过显存带宽传输至该图像处理单元,基于该共享数据库对该数据矩阵和目标特征矩阵进行处理。

[0081] 其中,根据目标子任务队列中的子任务属性,获取目标子任务队列对应的共享数据库,这一步骤已在上述步骤中进行说明,在此不多做赘述。

[0082] 在确定目标子任务队列中的子任务对应的共享数据库之后,将对数据向量进行聚合处理得到的数据矩阵,对特征矩阵进行聚合处理得到的目标特征矩阵,以及共享数据库传输至图像处理单元,在图像处理单元中,通过共享数据库对数据矩阵和目标特征矩阵进行相应的计算处理。

[0083] 本申请实施例通过确定目标图像的至少一个对象处理任务;确定每个对象处理任务的类型;根据任务类型确定对象处理任务的级别;获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息;基于第一负载信息与第二负载信息确定出目标负载比;当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的级别范围;将级别位于级别范围内的对象处理任务分配给图像处理单元进行处理,并将级别未处于级别范围内的对象任务分配给中央处理单元进行处理。以此,通过确定对象处理任务的类型,监测不同的处理单元之间形成的负载比,在负载比达到一定条件时,根据负载比将不同级别的对象处理任务分配至不同的处理单元进行并行处理,缓解处理单元的处理压力,极大的提升了任务处理的效率。

[0084] 根据上述介绍的内容,下面将举例来进一步说明本申请的任务处理方法。在本实施例中,将以该任务处理装置具体集成在服务器中为例进行说明。请参考图3,图3为本申请实施例提供的第二种任务处理方法的流程示意图。具体流程可以如下:

201、服务器获取对象处理任务,确定对象处理任务的类型。

[0085] 其中,本申请实施例可以应用于智慧零售系统中,通过摄像头采集人脸人体信息,对人脸人体信息进行处理,确定客流信息。其中对人脸人体信息进行处理时,包括人脸检测、人体检测、跟踪、优选以及人脸配准、人体关键点、人脸人体绑定、提特征、提属性、检索等完整的一个AI功能线进行处理。该AI功能线可以为一个任务队列,该任务队列中可以包括多个对象处理任务,通过对多个对象处理任务进行处理,以完成对人脸人体信息的处理。

[0086] 以此,服务器可以从请求任务队列中获取任务,作为对象处理任务。在获取到对象处理任务之后,可以进一步的确定对象处理任务的类型。对象处理任务的类型参见上述实施例,在此不多做赘述。

[0087] 202、服务器根据对象处理任务类型确定对象处理任务的级别。

[0088] 其中,服务器可以首先获取各个样本任务类型与各个样本级别等级的对应关系。

[0089] 例如,样本任务类型可以包括:跟踪、特征、属性、检索、检测、绑定、优选、配准、关键点等,样本级别等级可以包括第一任务级别、第二任务级别、第三任务级别等,样本任务类型与样本级别等级的对应关系可以为:跟踪、特征、属性、检索对应第一任务级别,检测、绑定对应第二任务级别,优选、配准、关键点对应第三任务级别。

[0090] 在一实施方式中,第一任务级别的计算复杂度高于第二任务级别的计算复杂度,第二任务级别的计算复杂度高于第三任务级别的计算复杂度,也即表示处理器在处理任务

类型为跟踪、特征、属性、检索的计算压力大于任务类型为检测、绑定的计算压力,处理任务类型为检测、绑定的计算压力大于任务类型为优选、配准、关键点的计算压力。

[0091] 203、服务器获取图像处理单元的负载信息和中央处理单元的负载信息,确定目标负载比。

[0092] 其中,图像处理单元可以为GPU处理单元,中央处理单元可以为CPU处理单元,图像处理单元对对象处理任务的处理效率可以优于中央处理单元对对象处理任务的处理效率。

[0093] 在获取到图像处理单元的负载信息,中央处理单元的负载信息之后,可以计算图像处理单元的负载信息与中央处理单元的负载信息的比值,从而可以确定目标负载比,目标负载比可以表示图像处理单元和中央处理单元处理任务的占比情况,目标负载比可以体现出图像处理单元和中央处理单元之间的负载相对关系,该目标负载比较高时,说明图像处理单元负载程度较大,该中央处理单元负载程度相对较低,该目标负载比较低时,说明图像处理单元负载程度较低,该中央处理单元负载程度较高。

[0094] 请参考图4,图4为本申请图4为本申请实施例提供的一种任务处理方法任务调度示意图。如图4所示,图4左侧上半部分表示图像处理单元的请求任务队列,下半部分表示中央处理单元的请求任务队列,图4左侧最上方“t1,t2,t3,t4,t5,t6,t7,t8”表示时间段,时间段下方的矩形框表示处理单元的负载状态,也即负载信息。见图4下方的矩形框,不同的矩形框分别表示不同的负载状态,负载状态可以包括第一负载状态,第二负载状态,第三负载状态以及第四负载状态,其中,第一负载状态表示处理单元的负载程度很低,第二负载状态表示处理单元的负载程度较低,第三负载状态表示处理单元的负载程度较高,第四负载状态表示处理单元的负载程度很高。

[0095] 参见图4,在t1时间段内,图像处理单元的请求任务队列中没有对象处理任务,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元和中央处理单元均为第一负载状态。

[0096] 其中,在t2时间段内,图像处理单元的请求任务队列中包括对象处理任务:检测1,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第二负载状态,中央处理单元为第一负载状态。

[0097] 在t3时间段内,图像处理单元的请求任务队列中包括对象处理任务:配准1、检测2,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第二负载状态,中央处理单元为第一负载状态。

[0098] 其中,在t4时间段内,图像处理单元的请求任务队列中包括对象处理任务:优选1、配准2、检测3、检测4,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第三负载状态,中央处理单元为第一负载状态。

[0099] 其中,在t5时间段内,图像处理单元的请求任务队列中包括对象处理任务:特征1、优选2、配准3、配准4、检测5,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第四负载状态,中央处理单元为第一负载状态。

[0100] 其中,在t6时间段内,图像处理单元的请求任务队列中包括对象处理任务:检索1、特征2、优选3、优选4、配准5,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第四负载状态,中央处理单元为第一负载状态。

[0101] 其中,在t7时间段内,图像处理单元的请求任务队列中包括对象处理任务:检索2、

特征3、特征4、优选5,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第三负载状态,中央处理单元为第一负载状态。

[0102] 其中,在t8时间段内,图像处理单元的请求任务队列中包括对象处理任务:检索3、检索4、特征5,中央处理单元的请求任务队列中没有对象处理任务,此时,图像处理单元为第二负载状态,中央处理单元为第一负载状态。通过图4左侧图像处理单元中的任务队列和中央处理单元中的任务队列,可以确定图像处理单元的负载状态和中央处理单元负载状态。在每一时间段内,可以通过图像处理单元的负载状态与中央处理单元的负载状态,确定图像处理单元与中央处理单元的目标负载比。

[0103] 204、服务器判断目标负载比是否大于第一预设阈值。

[0104] 其中,在确定目标负载比之后,服务器可以判断目标负载比是否大于第一预设阈值,第一预设阈值可以为1.1,在此不做限定。

[0105] 例如,目标负载比可以为1.5,第一预设阈值可以为1.1,则可以确定目标负载比大于第一预设阈值,可以执行步骤205;又例如,目标负载比可以为0.5,第一预设阈值可以为1.1,则可以确定目标负载比小于第一预设阈值,可以执行步骤208。

[0106] 205、服务器判断目标负载比是否大于第二预设阈值。

[0107] 其中,当确定目标负载比大于第一预设阈值之后,可以继续判断目标负载比是否大于第二预设阈值,第二预设阈值可以为2,其中,第二预设阈值大于第一预设阈值。

[0108] 例如,目标负载比可以为2.5,第二预设阈值可以为2,则可以确定目标负载比大于第二预设阈值,可以执行步骤206;又例如,目标负载比可以为1.5,第二预设阈值可以为2,则可以确定目标负载比小于第二预设阈值,可以执行步骤207。

[0109] 206、服务器根据对象处理任务的任务级别,按照预设第一分配规则将对象处理任务分配至图像处理单元与中央处理单元进行处理。

[0110] 其中,预设第一分配规则可以包括:将第一任务级别对应任务类型的对象处理任务分配至图像处理单元进行处理,将第二任务级别和第三任务级别对应任务类型的对象处理任务分配至中央处理单元进行处理。

[0111] 例如,参见图4中,左侧t5、t6时间段时,满足条件图像处理单元与中央处理单元的负载比大于第一预设阈值,且大于第二预设阈值。在t5时间段内,包括对象处理任务:特征1、优选2、配准3、配准4、检测5,其中,特征1对应第一任务级别,检测5对应第二任务级别优选2、配准3和配准4对应第三任务级别。按照第一分配规则对对象处理任务进行分配,也即执行图4中的任务调度,见图4右侧,进行任务调度之后,将特征1、检测5分配至图像处理单元进行处理,并将优选2、配准3、配准4分配至第二任务处理单元进行处理。进行任务调度后,图像处理单元的负载状态由第四负载状态变为第二负载状态,中央处理单元的负载状态由第一负载状态变为第二负载状态,保证了图像处理单元与中央处理单元任务处理的均衡,可以提高任务处理的效率。

[0112] 207、服务器根据对象处理任务的任务级别,按照预设第二分配规则将对象处理任务分配至图像处理单元与中央处理单元进行处理。

[0113] 其中,预设第二分配规则可以包括:将第一任务级别和第二任务级别对应任务类型的对象处理任务分配至图像处理单元进行处理,将第三任务级别对应任务类型的对象处理任务分配至中央处理单元进行处理。

[0114] 例如,参见图4中,在左侧t4和t7时间段,满足条件图像处理单元与中央处理单元的负载比大于第一预设阈值,且小于第二预设阈值。在t4时间段内,包括对象处理任务:优选1、配准2、检测3、检测4,其中,优选1、配准2对应第三任务级别,检测3、检测4对应第二处理等级,按照第二分配规则对对象处理任务进行分配,也即执行图4中的任务调度,见图4右侧,进行任务调度之后,将优选1、配准2分配至图像处理单元进行处理,并将检测3、检测4分配至中央处理单元进行处理。进行任务调度后,图像处理单元的负载状态由第三负载状态变为第二负载状态,中央处理单元的负载状态由第一负载状态变为第二负载状态。

[0115] 208、服务器根据对象处理任务的任务级别,按照预设第三分配规则将对象处理任务分配至图像处理单元与中央处理单元进行处理。

[0116] 其中,预设第三分配规则可以包括:将第一任务级别、第二任务级别和第三任务级别对应任务类型的对象处理任务分配至图像处理单元进行处理。

[0117] 例如,参见图4中,在左侧t8时间段,满足条件图像处理单元与中央处理单元的负载比小于第一预设阈值。在t8时间段内,包括对象处理任务:检索3、检索4、特征5,其中,检索3、检索4、特征5对应第一任务级别,按照第二分配规则对对象处理任务进行分配,也即执行图4中的任务调度,见图4右侧,进行任务调度之后,将检索3、检索4、特征5分配至图像处理单元进行处理。进行任务调度后,图像处理单元的负载状态保持第二负载状态,在图像处理单元负载状态正常时,提高任务处理效率。

[0118] 在一些实施例中,将第一任务级别对应类型的对象处理任务在图像处理单元进行处理时,可以获取对象处理任务对应的子任务队列,将同一类型的对象处理任务的子任务进行聚合,当聚合到预设数量的子任务后,对预设数量的子任务进行矩阵聚合,得到聚合后的目标子任务队列,之后对目标子任务队列进行处理。

[0119] 其中,请参考图5,图5为本申请实施例提供的一种任务处理方法子任务聚合示意图。对象处理任务可以对应一个子任务队列,子任务队列中可以包括多个子任务,比如,多个子任务可以为Task1~TaskN。

[0120] 在获取到子任务队列之后,判断子任务队列中子任务数量是否满足预设数量,预设数量可以为进行任务聚合的最多子任务数量,预设数量可以为10,在此不做限定。

[0121] 若子任务队列中的子任务数量满足预设数量,则将子任务队列中的所有子任务进行聚合处理。其中,对子任务队列中的所有子任务进行聚合处理可以包括:获取每个子任务对应的多个数据向量,也即图5所示的 $U_{11} \dots U_{1m} \dots U_{n1} \dots U_{nm}$,其中,数据向量可以表示每个子任务中进行处理的不同数据,将多个数据向量进行矩阵聚合,可以得到进行聚合处理后的数据矩阵。获取每个数据向量对应的特征矩阵,也即图5所示的 $F_{11} \dots F_{1m} \dots F_{p1} \dots F_{pm}$,特征矩阵可以表示为数据向量的数据信息等,将所有数据向量的特征矩阵进行矩阵聚合处理,可以得到聚合处理后的特征矩阵。

[0122] 进一步的,在进行矩阵聚合处理之后,可以获取子任务队列中的子任务对应的共享数据库,将共享数据库以及聚合处理后的数据矩阵以及聚合处理后的特征矩阵加载至图像处理单元的显存带宽,进而通过图像处理单元对对象处理任务的子任务队列进行计算处理。其中,共享数据库也即图5所示的显存数据,包括用于处理任务时的各种计算数据。

[0123] 在一些实施例中,若子任务队列中的子任务数量小于预设数量,则可以设定聚合等待时间,聚合等待时间可以为20ms。例如,子任务队列中的子任务数量可以为5个,预设数

量可以为10个,则在确定子任务队列中的子任务数量小于预设数量之后,可以等待设定聚合时间,等待与子任务队列中的子任务属性相同的其他子任务。若在设定等待时间之前,子任务数量达到了预设数量,则停止等待,将所有子任务进行聚合处理,具体聚合处理步骤参见上述说明,在此不多做赘述。若在设定等待时间到达时,子任务数量仍未达到预设数量,则停止等待,将已获取的子任务进行聚合处理,具体聚合处理步骤参见上述说明,在此不多做赘述。通过将对象处理任务中的多个子任务进行聚合处理,可以减小处理单元中显存带宽与算力之间的差距,提高任务处理效率。

[0124] 本申请实施例通过确定目标图像的至少一个对象处理任务;确定每个对象处理任务的类型;根据任务类型确定对象处理任务的级别;获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息;基于第一负载信息与第二负载信息确定出目标负载比;当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围;将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。以此,通过确定对象处理任务的类型,监测不同的处理单元之间形成的负载比,在负载比达到一定条件时,根据负载比将不同任务级别的对象处理任务分配至不同的处理单元进行并行处理,缓解处理单元的处理压力,极大的提升了任务处理的效率。

[0125] 为便于更好的实施本申请实施例提供的任务处理方法,本申请实施例还提供一种基于上述任务处理方法的任务处理装置。其中名词的含义与上述任务处理方法中相同,具体实现细节可以参考方法实施例中的说明。

[0126] 请参阅图6,图6为本申请实施例提供的第一种任务处理装置的结构框图,该任务处理装置可以应用于诸如智能手机、平板电脑、笔记本电脑、台式计算机、智能音箱、智能手表等终端,以及诸如可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器,该装置包括:

- 第一确定单元301,用于确定目标图像的至少一个对象处理任务;
- 第二确定单元302,用于确定每个所述对象处理任务的类型;
- 第三确定单元303,用于根据所述任务类型确定所述对象处理任务的级别;
- 获取单元304,用于获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,所述图像处理单元对于对象处理任务的处理优先级高于所述中央处理单元;
- 第四确定单元305,用于基于所述第一负载信息与所述第二负载信息确定出目标负载比;
- 第五确定单元306,用于当所述目标负载比大于预设阈值时,根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任务级别范围;

分配单元307,用于将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给所述中央处理单元进行处理。

[0127] 在一些实施例中,请参阅图7,图7为本申请实施例提供的第二种任务处理装置的结构框图。分配单元307,可以包括:

获取子单元3071,用于获取对象处理任务对应的子任务队列,所述子任务队列中包括

多个子任务；

聚合子单元3072,用于将相同子任务属性的子任务进行任务聚合,生成任务聚合后的目标子任务队列；

处理子单元3073,用于将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合,并通过所述图像处理单元对矩阵聚合后的目标子任务队列进行处理。

[0128] 在一些实施例中,请参阅图8,图8为本申请实施例提供的第三种任务处理装置的结构框图。分配单元307,还可以包括:

第一执行子单元3074,用于当检测到所述目标子任务队列的子任务数量满足预设数量时,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤;

第二执行子单元3075,用于当检测到所述目标子任务队列的子任务数量不满足预设数量时,在预设时间内等待相同子任务属性的子任务进行任务聚合,执行将所述目标子任务队列的数据矩阵和特征矩阵进行矩阵聚合的步骤。

[0129] 在一些实施例中,处理子单元3073,可以用于:获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;将多个数据向量进行聚合处理,得到数据矩阵;将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列。

[0130] 在一些实施例中,处理子单元3073,可以用于:获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;将多个数据向量进行聚合处理,得到数据矩阵;将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列;根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;将所述共享数据库、所述数据矩阵和目标特征矩阵通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述数据矩阵和目标特征矩阵进行处理。

[0131] 在一些实施例中,处理子单元3073,可以用于:获取目标子任务队列中每个子任务对应的多个数据向量,以及每个数据向量对应的特征矩阵;将多个数据向量进行聚合处理,得到数据矩阵;将多个特征矩阵进行聚合处理,得到目标特征矩阵,根据所述数据矩阵和目标特征矩阵得到矩阵聚合后的目标子任务队列;根据所述目标子任务队列中的子任务属性,获取所述目标子任务队列对应的共享数据库;

将所述共享数据库和矩阵聚合后的目标子任务队列通过显存带宽传输至所述图像处理单元,基于所述共享数据库对所述矩阵聚合后的目标子任务队列进行处理。

[0132] 在一些实施例中,请参阅图9,图9为本申请实施例提供的第四种任务处理装置的结构框图。第五确定单元306,可以包括:

划分子单元3061,用于将大于所述预设阈值的阈值区间分为第一阈值区间和第二阈值区间,所述第二阈值区间大于所述第一阈值区间;

第一确定子单元3062,用于若所述目标负载比所处的阈值区间为所述第一阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别和第二任务级别;

第二确定子单元3063,用于若所述目标负载比所处的阈值区间为所述第二阈值区间,则确定所述图像处理单元需要处理的任务级别范围包括第一任务级别。

[0133] 在一些实施例中,请参阅图10,图10为本申请实施例提供的第五种任务处理装置

的结构框图。第四确定单元305,可以包括:

计算子单元3051,用于对所述第一负载信息和所述第二负载信息进行比值计算,得到所述第一负载信息与所述第二负载信息之间的目标负载比。

[0134] 在一些实施例中,请参阅图11,图11为本申请实施例提供的第六种任务处理装置的结构框图。该任务处理装置还可以包括:

处理单元308,用于当所述目标负载比小于预设阈值时,将所有对象处理任务分配至所述图像处理单元进行处理。

[0135] 本申请实施例公开了一种任务处理装置,通过第一确定单元301,确定目标图像的至少一个对象处理任务;第二确定单元302,确定每个对象处理任务的任务类型;第三确定单元303,根据所述任务类型确定对象处理任务的任务级别;获取单元304,获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,图像处理单元对于对象处理任务的处理优先级高于中央处理单元;第四确定单元305,基于第一负载信息与第二负载信息确定出目标负载比;第五确定单元306,当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围;分配单元307,将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。以此,通过确定对象处理任务的任务类型得到任务级别,监测不同的处理单元之间形成的负载比,在负载比达到一定条件时,根据负载比将不同级别的对象处理任务分配至不同的处理单元进行并行处理,缓解处理单元的处理压力,极大的提升了任务处理的效率。

[0136] 本申请实施例还提供一种计算机设备,如图12所示,其示出了本申请实施例所涉及的服务器的结构示意图,具体来讲:

该计算机设备可以包括一个或者一个以上处理核心的处理器401、一个或一个以上计算机可读存储介质的存储器402、电源403和输入单元404等部件。本领域技术人员可以理解,图12中示出的计算机设备结构并不构成对计算机设备的限定,可以包括比图示更多或更少的部件,或者组合某些部件,或者不同的部件布置。其中:

处理器401是该计算机设备的控制中心,利用各种接口和线路连接整个计算机设备的各个部分,通过运行或执行存储在存储器402内的软件程序和/或模块,以及调用存储在存储器402内的数据,执行计算机设备的各种功能和处理数据,从而对计算机设备进行整体监控。可选的,处理器401可包括一个或多个处理核心;可选的,处理器401可集成应用处理器和调制解调处理器,其中,应用处理器主要处理操作系统、用户界面和应用程序等,调制解调处理器主要处理无线通信。可以理解的是,上述调制解调处理器也可以不集成到处理器401中。

[0137] 存储器402可用于存储软件程序以及模块,处理器401通过运行存储在存储器402的软件程序以及模块,从而执行各种功能应用以及数据处理。存储器402可主要包括存储程序区和存储数据区,其中,存储程序区可存储操作系统、至少一个功能所需的应用程序(比如声音播放功能、图像播放功能等)等;存储数据区可存储根据服务器的使用所创建的数据等。此外,存储器402可以包括高速随机存取存储器,还可以包括非易失性存储器,例如至少一个磁盘存储器件、闪存器件、或其他易失性固态存储器件。相应地,存储器402还可以包括存储器控制器,以提供处理器401对存储器402的访问。

[0138] 计算机设备还包括给各个部件供电的电源403,可选的,电源403可以通过电源管理系统与处理器401逻辑相连,从而通过电源管理系统实现管理充电、放电、以及功耗管理等功能。电源403还可以包括一个或一个以上的直流或交流电源、再充电系统、电源故障检测电路、电源转换器或者逆变器、电源状态指示器等任意组件。

[0139] 计算机设备还可包括输入单元404,该输入单元404可用于接收输入的数字或字符信息,以及产生与用户设置以及功能控制有关的键盘、鼠标、操作杆、光学或者轨迹球信号输入。

[0140] 尽管未示出,计算机设备还可以包括显示单元等,在此不再赘述。具体在本实施例中,计算机设备中的处理器401会按照如下的指令,将一个或一个以上的应用程序的进程对应的可执行文件加载到存储器402中,并由处理器401来运行存储在存储器402中的应用程序,从而实现前述实施例提供的各种方法步骤,如下:

确定目标图像的至少一个对象处理任务;

确定每个所述对象处理任务的任务类型;

根据所述任务类型确定所述对象处理任务的任务级别;

获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息,其中,所述图像处理单元对于对象处理任务的处理优先级高于所述中央处理单元;

基于所述第一负载信息与所述第二负载信息确定出目标负载比;

当所述目标负载比大于预设阈值时,根据所述目标负载比所处的阈值区间确定所述图像处理单元需要处理的任务级别范围;

将任务级别位于任务级别范围内的对象处理任务分配给所述图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给所述中央处理单元进行处理。

[0141] 本申请实施例公开了一种任务处理方法、装置及计算机可读存储介质。本申请实施例通过确定目标图像的至少一个对象处理任务;确定每个对象处理任务的任务类型;根据任务类型确定对象处理任务的任务级别;获取图像处理单元的第一负载信息以及中央处理单元的第二负载信息;基于第一负载信息与第二负载信息确定出目标负载比;当目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围;将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。以此,通过确定对象处理任务的任务类型,监测不同的处理单元之间形成的负载比,在负载比达到一定条件时,根据负载比将不同任务级别的对象处理任务分配至不同的处理单元进行并行处理,缓解处理单元的处理压力,极大的提升了任务处理的效率。

[0142] 本领域普通技术人员可以理解,上述实施例的各种方法中的全部或部分步骤可以通过指令来完成,或通过指令控制相关的硬件来完成,该指令可以存储于计算机可读存储介质中,并由处理器进行加载和执行。

[0143] 为此,本申请实施例提供一种计算机可读存储介质,其中存储有多条指令,该指令能够被处理器进行加载,以执行本申请实施例所提供的任一种任务处理方法中的步骤。例如,该指令可以执行如下步骤:

确定目标图像的至少一个对象处理任务;确定每个对象处理任务的任务类型;根据任务类型确定对象处理任务的任务级别;获取图像处理单元的第一负载信息以及中央处理单

元的第二负载信息,其中,图像处理单元对于对象处理任务的处理优先级高于所述中央处理单元;基于第一负载信息与第二负载信息确定出目标负载比;当所述目标负载比大于预设阈值时,根据目标负载比所处的阈值区间确定图像处理单元需要处理的任务级别范围;将任务级别位于任务级别范围内的对象处理任务分配给图像处理单元进行处理,并将任务级别未处于任务级别范围内的对象任务分配给中央处理单元进行处理。

[0144] 以上各个操作的具体实施可参见前面的实施例,在此不再赘述。

[0145] 其中,该计算机可读存储介质可以包括:只读存储器(ROM,Read Only Memory)、随机存取记忆体(RAM,Random Access Memory)、磁盘或光盘等。

[0146] 由于该计算机可读存储介质中所存储的指令,可以执行本申请实施例所提供的任一种任务处理方法中的步骤,因此,可以实现本申请实施例所提供的任一种任务处理方法所能实现的有益效果,详见前面的实施例,在此不再赘述。

[0147] 以上对本申请实施例所提供的任务处理方法、装置及计算机可读存储介质进行了详细介绍,本文中应用了具体个例对本申请的原理及实施方式进行了阐述,以上实施例的说明只是用于帮助理解本申请的方法及其核心思想;同时,对于本领域的技术人员,依据本申请的思想,在具体实施方式及应用范围上均会有改变之处,综上所述,本说明书内容不应理解为对本申请的限制。

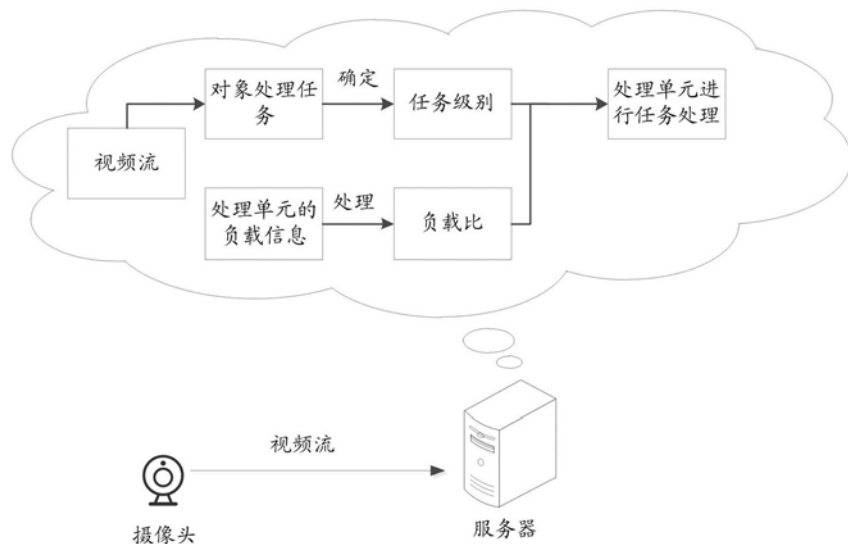


图1

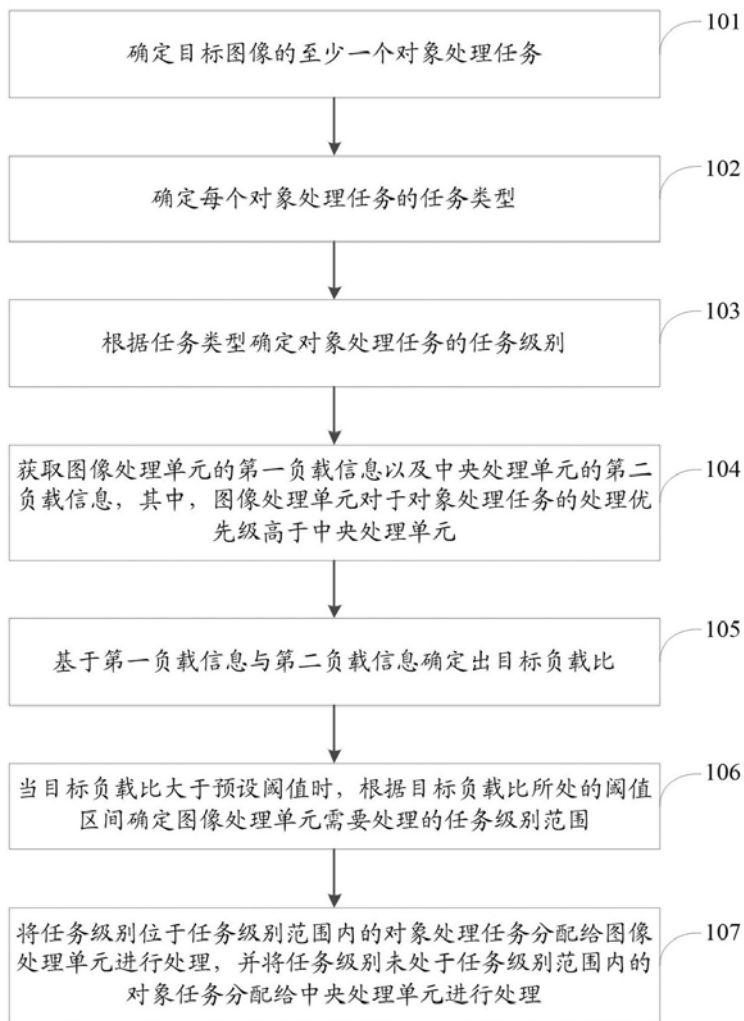


图2

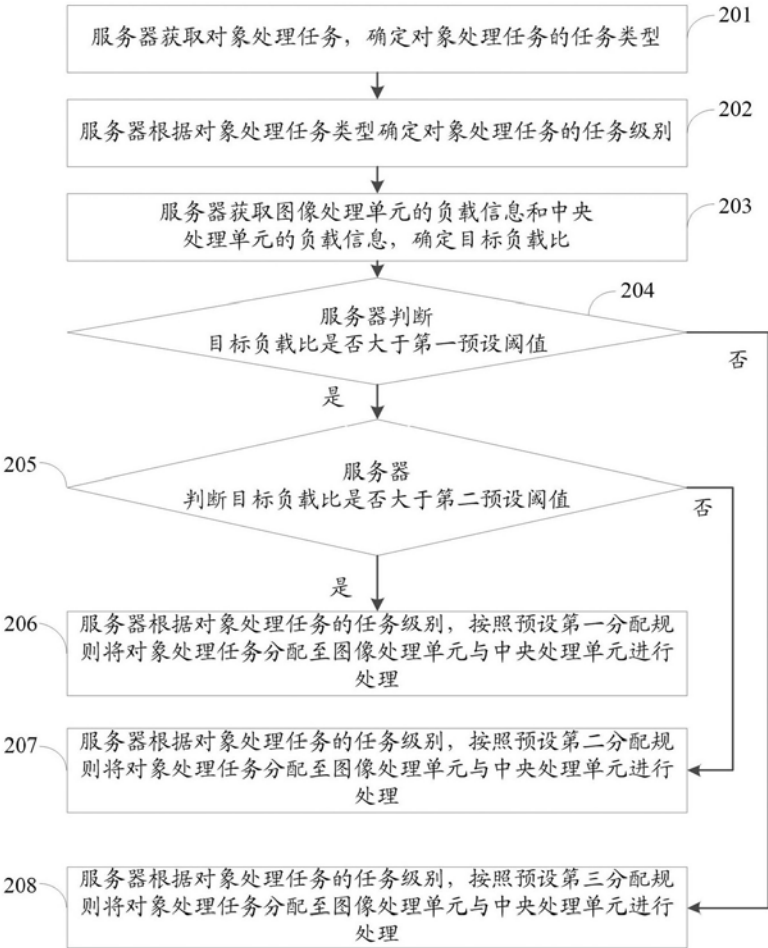


图3

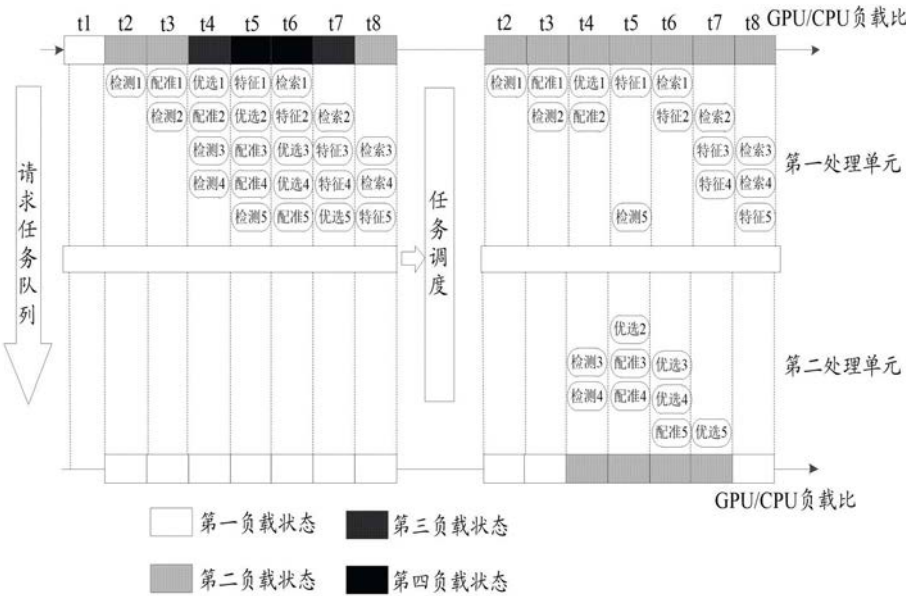


图4

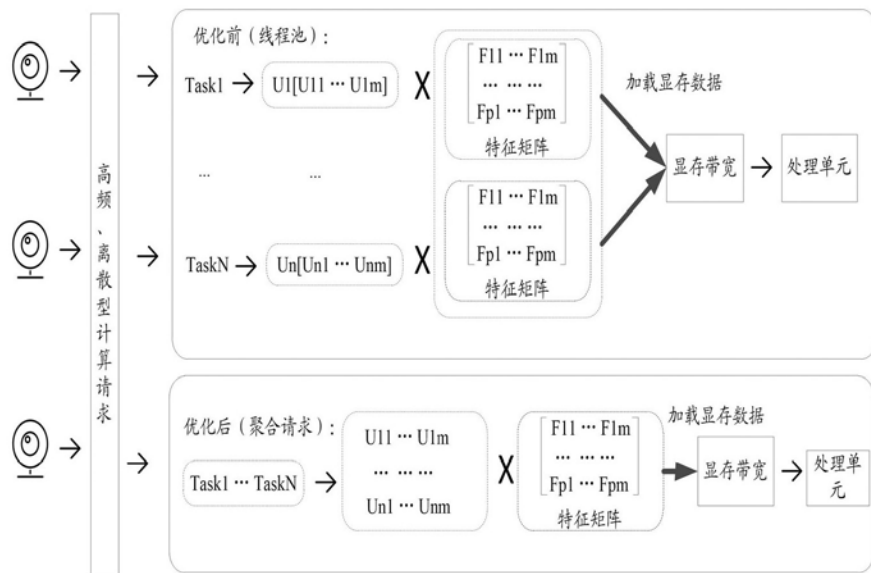


图5

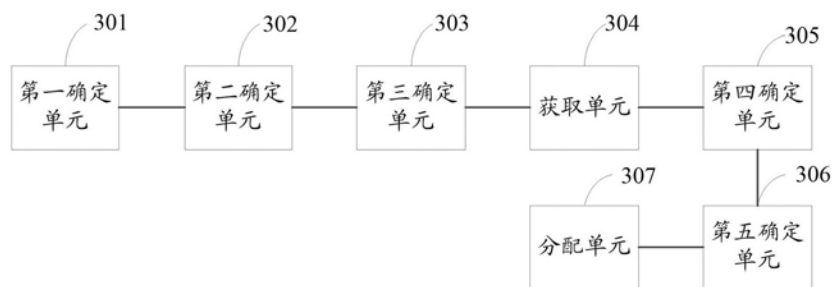


图6

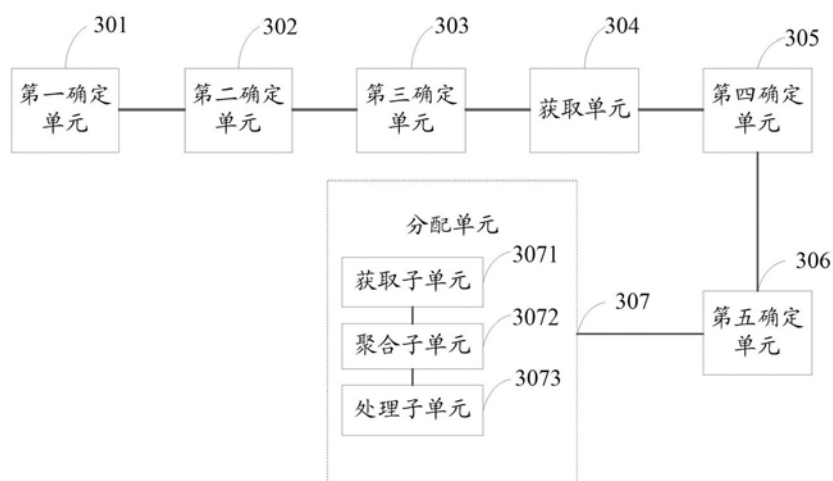


图7

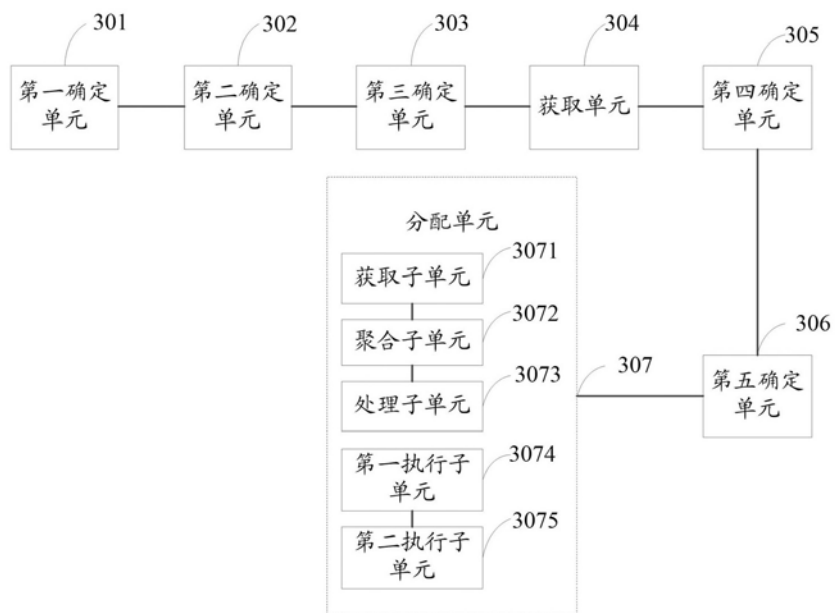


图8

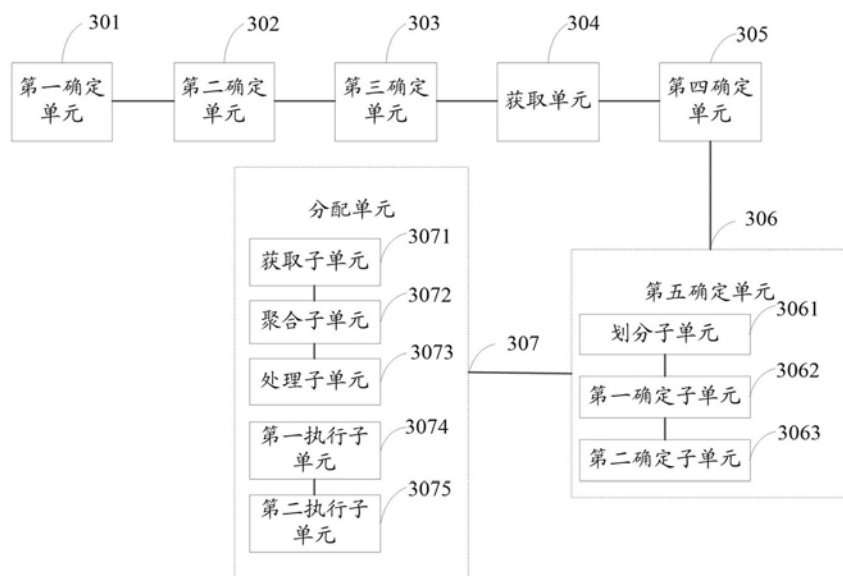


图9

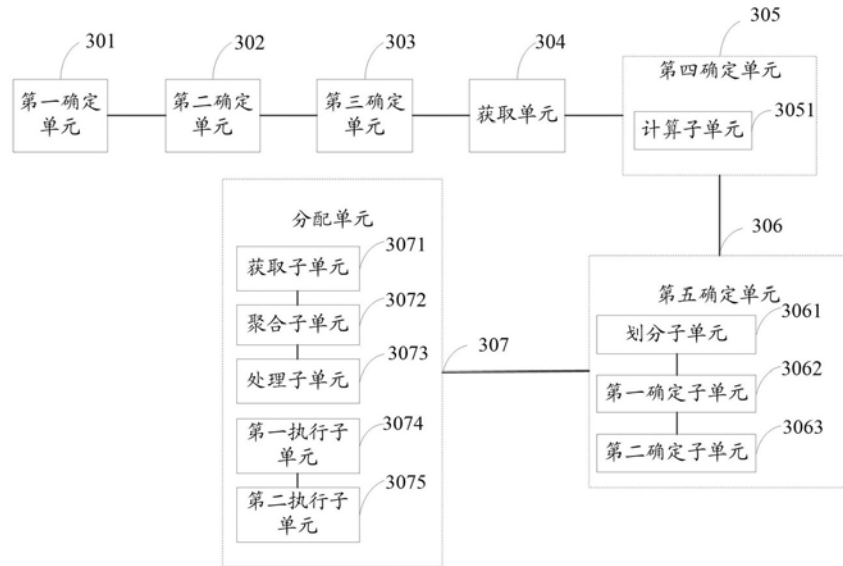


图10

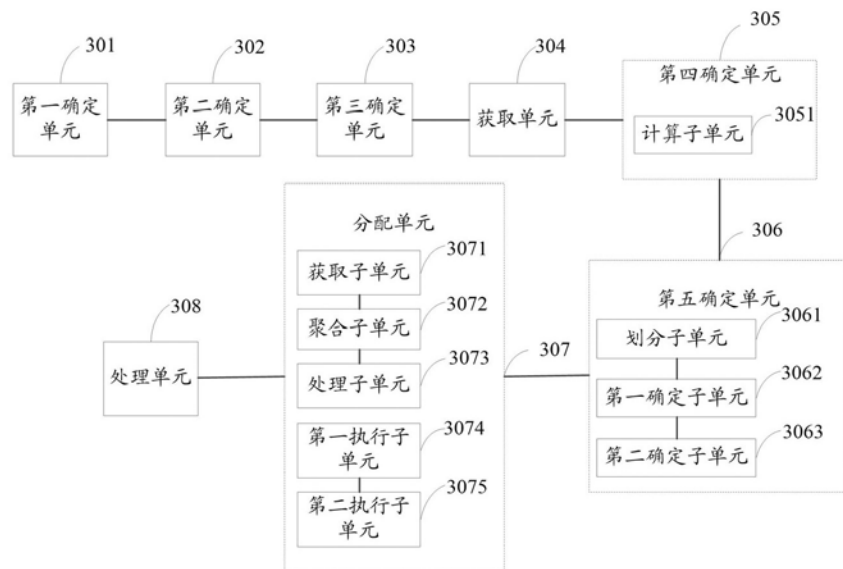


图11

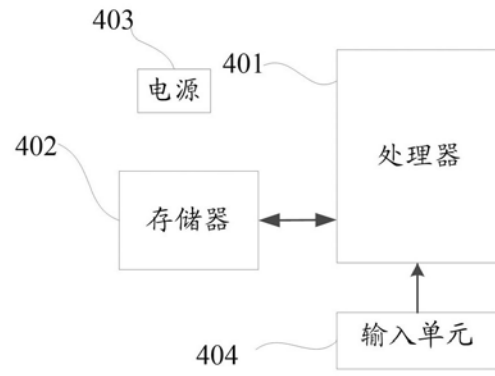


图12