(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 119096265 A
(43)申请公布日 2024.12.06

(21)申请号 202280095354.1
(22)申请日 2022.04.28
(85)PCT国际申请进入国家阶段日
2024.10.25
(86)PCT国际申请的申请数据
PCT/US2022/026674 2022.04.28
(87)PCT国际申请的公布数据
WO2023/211444 EN 2023.11.02
(71)申请人 创峰科技
地址 美国加利福尼亚州,帕洛阿尔托,海湾东路2479号,110房间
(72)发明人 刘杰 李翔 周扬
(74)专利代理机构 深圳市智圈知识产权代理事务所(普通合伙) 44351
专利代理师 李璇
(51)Int.Cl.
G06T 1/00(2006.01)

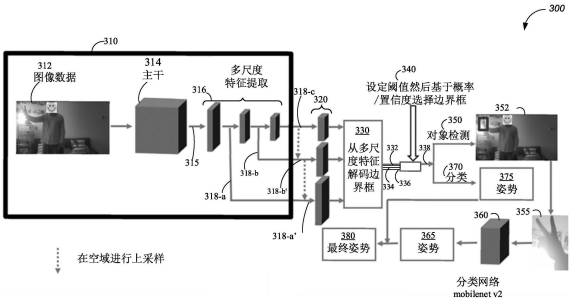
权利要求书2页 说明书15页 附图5页

(54)发明名称

使用轻量级深度学习模型的实时设备上远距离姿势识别

(57)摘要

本申请涉及检测在远距离(例如大于4米)捕获的图像中的姿势。电子设备获取捕获到手部的图像的图像数据,并基于图像数据生成特征图序列。特征图序列具有连续降尺度的特征分辨率。电子设备将特征图序列相组合以生成综合特征图,并根据综合特征图来确定在图像中捕获的手部的手势。在一些实施例中,基于综合特征图识别在图像内的手部区域,并使用手部区域来确定第一姿势向量。将第一姿势向量与根据综合特征图确定的第二姿势向量相组合,以形成最终姿势向量,基于最终姿势向量来确定手势。



CN 119096265 A

1.一种用于检测手势的方法,应用于电子设备,所述方法包括:

获取图像的图像数据,其中,所述图像捕获有手部;

基于所述图像数据生成特征图序列,其中,所述特征图序列具有连续降尺度的特征分辨率;

将所述特征图序列组合以生成综合特征图;以及

根据所述综合特征图确定在所述图像中捕获的所述手部的手势。

2.根据权利要求1所述的方法,其中,所述基于所述图像数据生成特征图序列,包括:使用机器学习模型从所述图像数据中识别一个或多个特征并丰富所述一个或多个特征,以及相对于所述图像数据的分辨率缩小所述特征图序列的特征分辨率。

3.根据权利要求2所述的方法,其中,所述基于所述图像数据生成特征图序列,包括:使用主干网络、基于所述图像数据生成所述特征图序列,所述主干网络包括至少两个卷积层和多个瓶颈结构。

4.根据权利要求3所述的方法,其中:

所述至少两个卷积层中的每个卷积层是步长为2的3×3核;以及

所述至少两个卷积层的第一卷积层之后是所述至少两个卷积层中的第二卷积层。

5.根据权利要求3所述的方法,其中,所述多个瓶颈结构包括:

第一瓶颈结构,被配置为从所述至少两个卷积层的输出中提取一个或多个特征,其中,所述第一瓶颈结构包括:第一1×1逐点卷积层、在所述第一1×1逐点卷积层之后的第一3×3深度卷积层、以及在所述第一3×3深度卷积层之后的第二1×1逐点卷积层和包括跳跃连接结构的加法运算;以及

第二瓶颈结构,被配置为缩小所述第一瓶颈结构的输出,其中,所述第二瓶颈结构包括:第三1×1逐点卷积层、在所述第三1×1逐点卷积层之后的步长为2的第二3×3深度卷积层、以及在所述第二3×3深度卷积层之后的第四1×1逐点卷积层;

其中,所述第二瓶颈结构的输出是多个特征图。

6.根据权利要求5所述的方法,其中,所述第一瓶颈结构和所述第二瓶颈结构中的每个瓶颈结构都具有3层和3的倍数个通道。

7.根据权利要求3所述的方法,其中,每个所述卷积层之后是激活函数。

8.根据前述权利要求中任一项所述的方法,其中,所述生成特征图序列,包括:

生成具有第一特征分辨率的第一特征图;

根据缩放因子将所述第一特征图缩小为第二特征图,所述第二特征图在所述特征图序列中紧跟在所述第一特征图之后;以及

根据所述缩放因子将所述第二特征图缩小为第三特征图,所述第三特征图在所述特征图序列中紧跟在所述第二特征图之后。

9.根据前述权利要求中任一项所述的方法,其中,所述将所述特征图序列相结合以生成综合特征图,包括:

从所述特征图序列中解码出一个或多个边界框;

识别所述一个或多个边界框中满足置信度标准的边界框子集;以及

基于所述一个或多个边界框的所述边界框子集,格式化所述综合特征图。

10.根据前述权利要求中任一项所述的方法,其中,所述手部由距离所述手部至少4米

远的摄像设备捕获在所述图像中。

11.根据前述权利要求中任一项所述的方法,其中,所述电子设备包括安全相机、智能电视、智能扬声器、玩具、智能手表、智能电器或头戴式显示器中的一种。

12.根据前述权利要求中任一项所述的方法,其中,所述图像数据的分辨率为至少1080p,并且所述方法使用所述电子设备的每秒35亿次浮点运算(GFLOPS)或更少的计算资源来实现。

13.根据前述权利要求中任一项所述的方法,其中,根据所述综合特征图确定在所述图像中的所述手势,还包括:

基于所述综合特征图,检测和提取所述图像数据中的手部区域;

基于从所述综合特征图中提取的所述手部区域,确定第一姿势向量;

从与所述图像的所述图像数据相对应的所述综合特征图确定第二姿势向量,所述图像包括所述手部区域和不同于所述手部区域的一个或多个其余区域;

将所述第一姿势向量和所述第二姿势向量相结合,以生成最终姿势向量;以及

基于所述最终姿势向量来确定所述手势。

14.一种电子设备,包括:

一个或多个处理器;以及

存储有指令的存储器,所述指令在被所述一个或多个处理器执行时使所述处理器实现根据权利要求1-13中任一项所述的方法。

15.一种存储有指令的非暂时性计算机可读介质,所述指令在被一个或多个处理器执行时使所述一个或多个处理器实现根据权利要求1-13中任一项所述的方法。

使用轻量级深度学习模型的实时设备上远距离姿势识别

技术领域

[0001] 本申请总体上涉及基于深度学习的姿势识别,包括但不限于使用压缩的图像数据和多尺度检测进行远距离姿势识别的方法、系统和非暂时性计算机可读介质。

背景技术

[0002] 现有的姿势识别方案通常要求在靠近相机(例如距离不到1米)的位置捕获姿势(例如,手势(hand gesture,手部动作)、面部姿势(face gesture,面部动作)等)以准确检测姿势。当在1米以外的距离捕获到姿势时,使用连续的图像、通过形状识别来检测姿势,但这种方案的可靠性较低并且会增加误报(false positive,假阳性)率。使用深度学习算法来增加能够从远程捕获的图像中识别到姿势的距离。然而,现有的深度学习解决方案需要功能强大的处理器和/或具有高分辨率输入(例如,图像数据)的模型,这对于物联网(internet of things,IoT)设备或诸如智能手机等较小的电子设备而言是不实际的。因此,具有能够在计算和/或存储能力有限的IoT设备和/或较小的电子设备中实施的、准确且高效地对在图像数据中捕获的姿势进行检测的系统和方法将是有益的。

发明内容

[0003] 本申请的各实施例涉及使用经缩放和丰富的压缩图像数据来检测所捕获的图像中的一个或多个姿势的姿势识别技术。实施例提供了可以在具有有限计算资源的移动设备或边缘设备上运行的轻量级手部检测模型。轻量级手部检测模型具有轻量级主干(backbone,骨干、骨架)和金字塔检测结构。金字塔检测结构能够对在不同距离(例如,从1米到4米或更远)捕获的图像中的尺寸不同的姿势进行检测,同时具有超低的计算复杂度。通过这些方式,所公开的系统和方法提供了远距离姿势识别方案,该方案进行操作的推理时间与本领域中已知的现有方案相似,但同时使用基本上更少的计算资源(例如,与现有方案使用的5.6-780GFLOPS(giga floating-point operations per second,每秒10亿次浮点运算数)相比,该方案使用每秒3.5GFLOPS)。

[0004] 换句话说,本方案提供了轻量级的端到端深度学习结构,其降低了手部识别的计算复杂度,并且将IoT设备上的推理时间减少了若干个数量级,同时不影响或几乎不影响识别结果的准确性。这些方案以不会明显影响性能的准确性略微下降为代价,实现了实时的设备上部署。在一些实施例中,用于识别和检测姿势的端到端模型与另外的分类模型相结合,以提高姿势识别的可靠性并减少误报。

[0005] 在一个方面,提供了一种用于检测手势的方法。该方法包括:获取图像的图像数据,其中,图像捕获有手部;以及基于图像数据生成特征图序列。特征图序列具有连续降尺度的特征分辨率。该方法还包括组合将特征图序列组合以生成综合特征图,以及根据综合特征图确定在图像中捕获的手部的手势。

[0006] 在一些实施例中,基于图像数据生成特征图序列,包括:使用机器学习模型从图像数据中识别一个或多个特征并丰富这一个或多特征,以及相对于图像数据的分辨率缩小特

征图序列的特征分辨率。

[0007] 在一些实施例中,基于图像数据生成特征图序列,包括:使用主干网络、基于图像数据生成特征图序列,主干网络包括至少两个卷积层和多个瓶颈结构。

[0008] 进一步地,在一些实施例中,至少两个卷积层中的每个卷积层是步长为2的3×3核。至少两个卷积层中的第一卷积层之后是至少两个卷积层中的第二卷积层。此外,在一些实施例中,多个瓶颈结构包括第一瓶颈结构和第二瓶颈结构。第一瓶颈结构被配置为从至少两个卷积层的输出中提取一个或多个特征。第一瓶颈结构包括第一1×1逐点卷积层、在第一1×1逐点卷积层之后的第一3×3深度卷积层、以及在第一3×3深度卷积层之后的第二1×1逐点卷积层和包括跳跃连接结构的加法运算。第二瓶颈结构被配置为缩小第一瓶颈结构的输出。第二瓶颈结构包括第三1×1逐点卷积层、在第三1×1逐点卷积层之后的步长为2的第二3×3深度卷积层、以及在第二3×3深度卷积层之后的第四1×1逐点卷积层。第二瓶颈结构的输出是多个特征图。此外,在一些实施例中,多个瓶颈结构中的每个瓶颈结构都具有3层和3的倍数个通道。在一些实施例中,每个卷积层之后是激活函数。

[0009] 在一些实施例中,生成特征图序列包括:生成具有第一特征分辨率的第一特征图;根据缩放因子将所述第一特征图缩小为第二特征图,该第二特征图在特征图序列中紧跟在第一特征图之后;以及根据缩放因子将第二特征图缩小为第三特征图,该第三特征图在特征图序列中紧跟在第二特征图之后。

[0010] 在一些实施例中,将特征图序列组合以生成综合特征图,包括:从特征图序列中解码一个或多个边界框;识别一个或多个边界框中满足置信度标准的边界框子集;以及基于一个或多个边界框中的边界框子集,格式化(format,编排格式)所述综合特征图。

[0011] 在一些实施例中,手部由距离该手部至少4米远的摄像设备捕获在图像中。

[0012] 在一些实施例中,电子设备包括安全相机、智能电视、智能扬声器、玩具、智能手表、智能电器或头戴式显示器中的一种。

[0013] 在一些实施例中,图像数据的分辨率至少为1080p,该方法使用电子设备的3.5GFLOPS或更少的计算资源来实现。

[0014] 在一些实施例中,根据综合特征图确定在图像中的手势,还包括:基于综合特征图对图像数据中的手部区域进行检测和提取,基于从综合特征图中提取的手部区域来确定第一姿势向量,从与图像的图像数据相对应的综合特征图确定第二姿势向量,将第一姿势向量和第二姿势向量相组合以生成最终姿势向量,以及基于最终姿势向量来确定手势。图像包括手部区域和不同于手部区域的一个或多个其余区域。

[0015] 在另一方面,一些实施方式包括电子系统,电子系统包括一个或多个处理器和存储有指令的存储器,指令在被一个或多个处理器执行时使处理器实现上述任一方法。

[0016] 在又一方面,一些实施方式包括存储有指令的非暂时性计算机可读介质,指令在被一个或多个处理器执行时使一个或多个处理器实现上述任一方法。

[0017] 提及这些示例性实施例和实施方式不是为了限制或定义本公开,而是为了提供示例以帮助理解本公开。在具体实施方式部分中讨论了其他实施例,并提供了进一步说明。

附图说明

[0018] 为了更好地理解所描述的各种实施例,应结合以下附图来参阅下面的具体实施

例,相同的附图标记在所有的附图中指代相应的部分。

[0019] 图1示出了根据一些实施例的具有一个或多个服务器的示例数据处理环境,该一个或多个服务器与一个或多个客户端设备通信耦合。

[0020] 图2示出了根据一些实施例的配置为处理内容数据(例如,图像数据)的电子系统的框图。

[0021] 图3示出了根据一些实施例的使用压缩图像数据以检测和识别一个或多个姿势的姿势识别过程的流程图。

[0022] 图4示出了根据一些实施例的姿势识别过程的图像处理阶段的流程图。

[0023] 图5示出了根据一些实施例的姿势检测和识别方法的流程图。

[0024] 相同的附图标记在所有的附图中指代相应的部分。

具体实施方式

[0025] 现在将详细描述在附图中示出示例的具体实施例。在下面的具体实施例中,阐述了许多非限制性的具体细节,以便帮助理解本文呈现的主题。明显地,对于本领域的技术人员来说,可以使用各种替换方案而不脱离权利要求的范围,并且可以在没有这些具体细节的情况下实施主题。例如,本文呈现的主题可以在具有数字视频能力的多种类型的电子设备上实现,这对于本领域的技术人员来说是明显的。

[0026] 本申请的各实施例涉及使用经缩放和丰富的压缩图像数据来检测所捕获的图像中的一个或多个姿势的姿势识别技术。实施例提供了可以在具有有限计算资源的移动设备或边缘设备上运行的轻量级手部检测模型。轻量级手部检测模型具有轻量级主干和金字塔检测结构。金字塔检测结构能够对在不同距离(例如,从1米到4米或更远)捕获的图像中的尺寸不同的姿势进行检测,同时具有超低的计算复杂度。通过这些方式,所公开的系统和方法提供了远距离姿势识别方案,该方案进行操作的推理时间与本领域中已知的现有方案相似,但同时使用基本上更少的计算资源(例如,与现有方案使用的5.6-780GFLOPS相比,该方案使用每秒3.5GFLOPS)。

[0027] 换句话说,本方案提供了轻量级的端到端深度学习结构,其降低了手部识别的计算复杂度,并且将IoT设备上的推理时间减少了若干个数量级,同时不影响或几乎不影响识别结果的准确性。这些方案以不会明显影响性能的准确性略微下降为代价,实现了实时的设备上部署。在一些实施例中,用于识别和检测姿势的端到端模型与另外的分类模型相结合,以提高姿势识别的可靠性并减少误报。

[0028] 图1是根据一些实施例的示例数据处理环境100,该数据处理环境100具有通信耦合到一个或多个客户端设备104的一个或多个服务器102。一个或多个客户端设备104例如可以是台式计算机104A、平板电脑104B、移动电话104C、头戴式显示器(head-mounted display,HMD)(也称为增强现实(augmented reality,AR)眼镜)104D,或者智能多感测联网家庭设备(例如,监控相机104E、智能电视设备、无人机等)。每个客户端设备104都可以收集数据或用户输入、执行用户应用程序,以及在其用户界面上呈现输出。所收集的数据或用户输入可以在客户端设备104上进行本地处理和/或由服务器102进行远程处理。一个或多个服务器102向客户端设备104提供系统数据(例如,启动文件、操作系统映像和用户应用程序),并且在一些实施例中,当在客户端设备104上执行用户应用程序时处理从(一个或多

个)客户端设备104接收的数据和用户输入。在一些实施例中,数据处理环境100还包括存储器106,该存储器106用于存储与服务器102、客户端设备104以及在客户端设备104上执行的应用程序相关的数据。

[0029] 一个或多个服务器102被配置为能与彼此远离的客户端设备104进行实时数据通信,或能与远离一个或多个服务器102的客户端设备104进行实时数据通信。此外,在一些实施例中,一个或多个服务器102被配置为执行无法由或不会优先由客户端设备104在本地完成的数据处理任务。例如,客户端设备104包括执行交互式在线游戏应用程序的游戏控制台(例如,HMD 104D)。游戏控制台接收用户指令并将其与用户数据一起发送到游戏服务器102。游戏服务器102根据用户指令和用户数据生成视频数据流,并提供视频数据流以供在游戏控制台和与游戏控制台进行相同游戏会话的其他客户端设备上进行显示。又例如,客户端设备104包括联网监控相机104E和移动电话104C。联网的监控相机104E采集视频数据,并将视频数据实时传输到监控相机服务器102。尽管视频数据可选地在监控相机104E上进行预处理,但监控相机服务器102对视频数据进行处理,以识别视频数据中的运动或音频事件,并与移动电话104C共享这些事件的信息,从而使移动电话104的用户能够实时远程监控联网监控相机104E附近发生的事件。

[0030] 一个或多个服务器102、一个或多个客户端设备104和存储器106彼此之间通过一个或多个通信网络108在通信上耦合,一个或多个通信网络108是用于为数据处理环境100内的这些设备和连接在一起的计算机之间提供通信链接的介质。一个或多个通信网络108可以包括诸如电线、无线通信链路或光纤光缆之类的连接。一个或多个通信网络108的示例包括局域网(local area network,LAN)、诸如互联网的广域网(wide area network,WAN)或其组合。一个或多个通信网络108可选地使用任何已知的网络协议实现,包括各种有线或无线协议,诸如以太网、通用串行总线(Universal Serial Bus,USB)、火线(FIREWIRE)、长期演进(Long Term Evolution,LTE)、全球移动通信系统(Global System for Mobile Communications,GSM)、增强型数据GSM环境(Enhanced Data GSM Environment,EDGE)、码分多址(code division multiple access,CDMA)、时分多址(time division multiple access,TDMA)、蓝牙(Bluetooth)、Wi-Fi、基于IP的语音传输(voice over Internet Protocol,VoIP)、Wi-MAX或任何其他合适的通信协议。与一个或多个通信网络108的连接可以直接建立(例如,使用3G/4G连接到无线运营商),或通过网络接口110(例如,路由器、交换机、网关、集线器或智能专用全家控制节点)建立或通过其任意组合建立。这样,一个或多个通信网络108可以表示使用传输控制协议/因特网互联协议(Transmission Control Protocol/Internet Protocol,TCP/IP)组彼此通信的网络和网关的全球集合的因特网。因特网的核心是主要节点或主机计算机之间的高速数据通信线的主干,由数以千计的用于路由数据和消息的商业、政府、教育和其他计算机系统组成。

[0031] 在一些实施例中,深度学习技术应用于数据处理环境100中,以处理由在客户端设备104上执行的应用程序获取的内容数据(例如,视频数据、视觉数据、音频数据),从而识别内容数据中包含的信息,将内容数据与其他数据进行匹配,对内容数据进行分类,或者合成相关的内容数据。内容数据可以广义地包括由客户端设备104的惯性传感器捕获的惯性传感器数据。在这些深度学习技术中,基于一个或多个神经网络来创建数据处理模型以处理内容数据。在应用这些数据处理模型处理内容数据之前,利用训练数据对这些数据处理模

型进行训练。在模型训练之后,移动电话104C或HMD 104D获取内容数据(例如,通过内部相机捕获视频数据)并在本地使用数据处理模型以处理内容数据。

[0032] 在一些实施例中,模型训练和数据处理都是在每个单独的客户端设备104(例如,移动电话104C和HMD 104D)上本地实施的。客户端设备104C从一个或多个服务器102或者存储器106获取训练数据,并应用训练数据来训练数据处理模型。可替代地,在一些实施例中,模型训练和数据处理都在与客户端设备104(例如,客户端设备104A和HMD 104D)相关联的服务器102(例如,服务器102A)处远程实施的。服务器102A从自身、另一服务器102或者存储器106获取训练数据,并应用训练数据来训练数据处理模型。客户端设备104A获取内容数据,将内容数据发送到服务器102A(例如,在应用程序中)以使用经训练的数据处理模型进行数据处理,从服务器102A接收数据处理结果(例如,识别到的手势),将结果呈现在(例如与应用程序相关的)用户界面上,基于姿态在视野中呈现虚拟对象,或者基于结果实现一些其他功能。客户端设备104A本身在将内容数据发送到服务器102A之前不对内容数据进行数据处理或只进行很少的数据处理。此外,在一些实施例中,在客户端设备104(例如,客户端设备104b和HMD 104D)上本地实施数据处理,与客户端设备104关联的服务器102(例如,服务器102B)上远程实施模型训练。服务器102B从自身、另一个服务器102或存储器106获取训练数据,并应用训练数据来训练数据处理模型。经训练的数据处理模型可选地存储在服务器102B或存储器106中。客户端设备104从服务器102B或存储设备106中导入经训练的数据处理模型,使用数据处理模型处理内容数据,并生成呈现在用户界面上的数据处理结果或使用数据处理结果在本地发起一些功能(例如,基于设备姿态渲染虚拟对象)。

[0033] 在一些实施例中,一副AR眼镜104D(也称为HMD)在通信上耦合在数据处理环境100中。AR眼镜104D包括相机、麦克风、扬声器、一个或多个惯性传感器(例如,陀螺仪、加速度计)和显示器。相机和麦克风被配置为从AR眼镜104D的场景捕获视频和音频数据,同时一个或多个惯性传感器被配置为捕获惯性传感器数据。在一些情况下,相机捕获佩戴AR眼镜104D的用户的手势,并利用手势识别过程(例如,图3中的过程300)在本地实时地识别手势。在一些情况下,麦克风记录包括用户的语音命令的环境声音。在一些情况下,由相机捕获的视频数据或静态视觉数据以及由一个或多个惯性传感器测量的惯性传感器数据都被应用于确定和预测设备姿态。由AR眼镜104D、(一个或多个)服务器102或上述两者对由AR眼镜104D捕获的视频、静态图像、音频或惯性传感器数据进行处理,以识别设备姿态。可选地,由(一个或多个)服务器102和AR眼镜104D联合应用深度学习技术以识别和预测设备姿态。设备姿态用于控制AR眼镜104D本身或者与AR眼镜104D执行的应用程序(例如,游戏应用程序)进行交互。在一些实施例中,AR眼镜104D的显示器显示有用户界面,并且所识别或预测的设备姿态用于在用户界面上呈现用户可选择的显示项目(例如,虚拟形象)或与之交互。

[0034] 如上所述,在一些实施例中,深度学习技术被应用于数据处理环境100中,以对由AR眼镜104D捕获的视频数据、静态图像数据或惯性传感器数据进行处理。利用第一数据处理模型,基于这样的视频、静态图像和/或惯性传感器数据来识别和预测2D或3D设备姿态。可视内容可选地使用第二数据处理模型来生成。可选地,由服务器102或AR眼镜104D实施对第一数据处理模型和第二数据处理模型的训练。由服务器102和AR眼镜104D中的每一个独立地实现设备姿态和视觉内容的推理,或者由服务器102和AR眼镜104D联合实施设备姿态和视觉内容的推理。

[0035] 图2示出了根据一些实施例的配置为处理内容数据(例如,图像数据)的电子设备200的框图。电子设备200是服务器102、客户端设备104(例如,图1中的AR眼镜104D)、存储器106或其组合之一。在一示例中,电子设备200是包括姿势识别模块230的移动设备,姿势识别模块230端到端地应用例如图4中的神经网络模型,以在移动设备本地识别手势。电子设备200通常包括一个或多个处理单元202(诸如中央处理单元(central processing unit,CPU)、图形处理单元(graphics processing unit,GPU)、数字信号处理器(digital signal processor,DSP)、神经处理单元(neural processing unit,NPU)、AI处理单元(AI processing unit,APU))、一个或多个网络接口204、存储器206以及用于互连这些组件(有时称为芯片组)的一个或多个通信总线208。电子设备200包括便于用户输入的一个或多个输入设备210,诸如键盘、鼠标、语音命令输入单元或麦克风、触摸屏显示器、触敏输入板、姿势捕获相机或其他输入按钮或控件。此外,在一些实施例中,电子设备200使用麦克风进行语音识别或使用相机260进行姿势识别,以补充或替代键盘。在一些实施例中,电子设备200包括用于捕获例如印刷在电子设备上的图形序列码的图像的一个或多个光学相机(例如,RGB相机260)、扫描仪或光传感器单元。在一些实施例中,电子设备200还包括一个或多个能够呈现用户界面和显示内容的输出设备212,包括一个或多个扬声器和/或一个或多个可视显示器。可选地,电子设备200包括用于确定电子设备200的位置的位置检测设备,诸如全球定位系统(global positioning system,GPS)或其他地理位置接收器。可选地,电子设备200包括惯性测量单元(inertial measurement unit,IMU)280,其集成由多轴惯性传感器捕获的传感器数据,以提供电子设备200在空间中的位置和方向的估计。IMU 280的一个或多个惯性传感器的示例包括但不限于陀螺仪、加速度计、磁力计和测斜仪。

[0036] 可替代地或附加地,在一些实施例中,电子设备200经由一个或多个网络接口204通信地耦合一个或多个设备(例如,服务器102、客户端设备104、存储器106或其组合),其包括一个或多个输入设备210、输出设备212、多个IMU 280或上述其他组件,并向电子系统200提供数据。

[0037] 存储器206包括高速随机存取存储器,例如DRAM、SRAM、DDR RAM或其他随机存取固态存储器设备;可选地包括非易失性存储器,诸如一个或多个磁盘存储设备、一个或多个光盘存储设备、一个或多个闪存设备或者一个或多个其他非易失性固态存储设备。可选地,存储器206可以包括与一个或多个处理单元202远程的一个或多个存储设备。存储器206或者存储器206内的非易失性存储器包括非暂时性计算机可读存储介质。在一些实施例中,存储器206或者存储器206的非暂时性计算机可读存储介质存储有以下程序、模块和数据结构,或者其子集或超集:

[0038] • 操作系统214,包括用于处理各种基本系统服务和用于执行硬件相关任务的程序:

[0039] • 网络通信模块216,用于经由一个或多个(有线或无线)网络接口204和一个或多个通信网络108(诸如因特网、其他广域网、局域网、城域网等),将每个服务器102或客户端设备104连接到其他设备(例如,服务器102、客户端设备104或存储设备106);

[0040] • 用户界面模块218,用于使得能够经由一个或多个输出设备212(诸如显示器、扬声器等)在每个客户端设备104上呈现信息(例如,(一个或多个)应用程序224的图形用户界面、窗口小部件、网站及其网页、和/或游戏、音频和/或视频内容、文本等)。

[0041] • 输入处理模块220,用于检测来自一个或多个输入设备210之一的一个或多个用户输入或交互,并解读检测到的输入或交互;

[0042] • 网络浏览器模块222,用于导航、请求(例如,经由HTTP)并显示网站及其网页,

[0043] 包括用于登录与客户端设备104或另一电子设备相关联的用户账户中的web界面,用于在与用户账户相关联时控制客户端或电子设备,以及用于编辑和查看与用户账户相关联的设置和数据;

[0044] • 由电子系统200执行的一个或多个用户应用程序224(例如,游戏、社交网络应用程序、智能家居应用程序和/或用于控制其他电子设备和查看此类设备捕获的数据的其他基于网络或不基于网络的应用程序);

[0045] • 模型训练模块226,用于接收训练数据并建立用于处理将由客户端设备104收集或获取的内容数据(例如,视频、图像、音频或文本数据)的数据处理模型;

[0046] • 数据处理模块228,用于利用数据处理模型250来处理内容数据,从而识别包含在内容数据中的信息,将内容数据与其他数据进行匹配,对内容数据进行分类或者合成相关的内容数据,其中在一些实施例中,数据处理模块228与多个用户应用程序224中的一个相关联,以响应于从该用户应用程序224接收的用户指令来处理内容数据,并且在一示例中,数据处理模块228被用于实现图3中的姿势识别过程300;

[0047] • 姿势识别模块230,用于压缩图像数据并基于压缩的图像数据生成多尺度特征图(例如下面参考图3和图4所示和所描述的),其中在一些实施例中,由姿势识别模块230和数据处理模块228联合压缩图像数据并生成多尺度特征图,并且多尺度特征图被增加尺度(upscale,提升尺度)和融合,以生成用于姿势识别的综合特征图;

[0048] • 一个或多个数据库240,用于存储至少包括以下一个或多个的数据:

[0049] o设备设置242,包括一个或多个服务器102或客户端设备104的通用设备设置(诸如服务层、设备型号、存储容量、处理能力、通信能力等);

[0050] o一个或多个用户应用程序224的用户账户信息244,诸如用户名、安全问题、账户历史数据、用户偏好和预定义账户设置;

[0051] o一个或多个通信网络108的网络参数246,诸如IP地址、子网掩码、默认网关、DNS服务器和主机名;

[0052] o用于训练一个或多个数据处理模型250的训练数据248;

[0053] o数据处理模型250,用于使用深度学习技术处理内容数据(如,视频、图像、音频或文本数据)其中,数据处理模型250包括如下面参考图3和图4所述的用于实现图像压缩过程的图像压缩模型、用于实现多尺度特征提取过程的特征提取模型和/或一个或多个分类模型和网络;

[0054] o姿势数据库252,用于存储与候选图像相关联的一个或多个姿势(例如,存储在数据库240中);和

[0055] o内容数据和结果254,其由电子系统200(或者与电子系统200通信地耦合的设备例如客户端设备104)分别获取和输出,其中,内容数据由数据处理模型250在客户端设备104进行本地处理或在服务器102进行远程处理,以提供要在客户端设备104上呈现的相关联的结果,内容数据包括候选图像。

[0056] 在一些实施例中,姿势识别模块230还包括图像处理模块232,图像处理模块232被

配置为缩小图像数据、提取一个或多个特征、以及缩放一个或多个特征以提供根据其来生成手势的综合特征向量。下面参考图3和图4描述使用可在资源有限的电子设备中实施的端到端的轻量级神经网络的手势识别的更多细节。

[0057] 可选地,一个或多个数据库240存储在电子系统200的服务器102、客户端设备104和存储器106之一中。可选地,一个或多个数据库240分布在电子系统200的服务器102、客户端设备104和存储器106中的不止一个中。在一些实施例中,以上数据的多于一个副本被存储在不同的设备处,例如,数据处理模型250的两个副本分别被存储在服务器102和存储器106处。

[0058] 以上标识的元素中的每一个都可以存储在一个或多个前述存储设备中,并且对应于用于执行上述功能的一组指令。以上标识的模块或程序(即指令集)不需要被实施为单独的软件程序、过程、模块或数据结构,因此这些模块的各种子集可以在各种实施例中被组合或以其他方式重新排列。在一些实施例中,存储器206可选地存储上述模块和数据结构的子集。此外,存储器206可选地存储以上未描述的附加模块和数据结构。

[0059] 图3示出了根据一些实施例的姿势识别过程300的流程图,该姿态识别过程300用于使用图像数据来检测和识别一个或多个姿势。姿势识别过程300被配置为检测由相机260从远距离拍摄的图像数据中捕获到的姿势。在一些实施例中,姿势和相机260之间的远距离大于预定距离。例如,姿势识别过程300被配置成检测和识别在距离相机260至少4米远处捕获的手势。姿势识别过程300使用有限的计算资源,例如,电子设备200的3.5GFLOPS或更少的计算资源。在一些实施例中,姿势识别过程300被配置成检测和识别与捕获成像设备的距离大于预定距离的姿势。虽然姿势识别过程300被配置成检测位于捕获图像设备的远距离处的姿势,但姿势识别过程300还被配置成检测在相对于捕获成像设备的近距离(例如,与捕获图像数据的相机260距离0.5-1米)处捕获的姿势。

[0060] 姿势识别过程300可选地由以上参考图1和图2描述的一个或多个客户端设备104、服务器102和/或其组合来执行。姿势识别过程300包括图像处理阶段310、特征图上采样过程320、特征图解码过程330、边界框选择过程340、对象检测过程350和/或分类过程370。对象检测过程350之后是裁剪过程和应用分类网络360以识别一个或多个对象检测姿势365。分类过程370基于由边界框选择过程340选择的边界框来识别一个或多个分类姿势375。每个对象检测姿势365是基于整个图像中的相应手部区域355即基于整个图像的相应部分来识别的。相反地,每个分类姿势375直接从整个图像的综合特征图332中确定。对于同一只手的同一姿势,对象检测姿势365可选地比分类姿势375更准确或更不准确,这取决于多个因素(例如,手部与相机的距离)。姿势识别过程300基于一个或多个对象检测姿势365和/或一个或多个分类姿势375来确定最终姿势380。下面详细讨论姿势识别过程300的每个过程。

[0061] 在图像处理阶段310,应用一个或多个机器学习模块来接收图像数据312。在将一个或多个机器学习模块应用于图像数据312之前,可选地对图像数据312进行预处理,如参考图4所描述的。在一些实施例中,图像数据312由电子设备200(图2)的输入设备210(例如,RGB相机260)捕获。可替代地,在一些实施例中,图像数据312经由包括在电子设备200上的网络浏览器模块222和/或一个或多个用户应用224进行获取或下载。可替代地或附加地,在一些实施例中,图像数据312是从通信地耦合到电子设备200的另一设备(例如,膝上型电脑、智能电话、AR眼镜、服务器等)处接收的。图像数据312包括一个或多个姿势。在一些情况

下,图像数据312内的姿势距离捕获该图像数据的电子设备200至少4米远。从图像数据312识别的一个或多个姿势的非限制性示例包括一个或多个手势、面部姿势和身体姿势。以初始分辨率(例如,1080p)接收图像数据312。

[0062] 图像数据312通过在图像处理阶段310中的一个或多个机器学习模块,以压缩图像数据312和/或基于图像数据312生成一个或多个特征图。在一些实施例中,图像数据312在通过一个或多个机器学习模块之前进行降尺度。在一些实施例中,一个或多个机器学习模块包括第一机器学习模块314和第二机器学习模块316,第一机器学习模块314被配置为从图像数据312中识别和丰富(例如,从中提取细节)一个或多个特征,以及(相对于图像数据312的初始分辨率)降低(由第一机器学习模块314生成的)多个特征图的特征图特征分辨率,第二机器学习模块316被配置为基于(压缩的)图像数据312生成(经缩放的)特征图的序列(例如,第一特征图、第二特征图和第三特征图318a-318c)。在一些实施例中,特征图序列被组合成综合特征图,该综合特征图被用于如下面详细讨论的姿势检测和姿势识别。

[0063] 在一些实施例中,第一机器学习模块314是主干网络,例如,包括用于从图像数据312中提取特征的特征提取网络。第一机器学习模块314被配置为对图像数据312进行压缩,同时丰富和/或利用图像数据312中的细节。在一些实施例中,丰富和/或利用图像数据312中的细节意味着提高和/或保留图像数据312中尽可能多的细节,使得图像数据312中的信息不会丢失并且可以在压缩期间恢复。第一机器学习模块314生成据其检测和识别一个或多个姿势的一个或多个特征。

[0064] 在一些实施例中,第二机器学习模块316包括生成特征图序列318的多尺度特征提取模型。基于第一机器学习模块314的输出(例如,由第一机器学习模块314生成的第一特征图)生成特征图序列。在一些实施例中,特征图序列318包括第一特征图318-a、第二特征图318-b和第三经缩放的特征图318-c。在一些实施例中,特征图序列的每个特征图318各自具有在序列中连续降尺度的特征分辨率,因此,特征图序列318对应于多个不同尺度的特征图。例如,第一特征图318-a具有第一特征分辨率,第二特征图318-b具有小于第一特征分辨率的第二特征分辨率,第三特征图318-c具有小于第二特征分辨率的第三特征分辨率。由第一机器学习模块314和第二机器学习模块316生成的特征图在后面参考图4详细讨论。

[0065] 特征图上采样过程320对第二机器学习模型316的输出进行上采样。每个特征图子集都被上采样到各自的之前特征图的较高特征分辨率,并与各自的之前特征图相组合。在一些实施例中,特征图的子集之一(例如,318-b和318-c)是由机器学习模块316生成的。输出的第三特征图318-c被上采样到第二特征图318-b的特征分辨率,并与第二特征图318-b进行组合以形成组合的第二特征图318-b’。第二特征图318-b可以被上采样到第一特征图318-a的特征分辨率,并且与第一特征图318-a进行组合以形成组合的第一特征图318a’。可替代地,在一些实施例中,特征图的子集之一(例如,318-b’)是从上采样的特征图318组合的。例如,组合的第二特征图318-b’被上采样到与第一特征图318-a相同的特征分辨率,并与第一特征图318-a进行组合以形成组合的第一特征图318-a’。这样,特征图序列318经由机器学习模块316被连续地降尺度,特征图子集318-c和318-b均被上采样到与各自的之前特征图318-b和318-a的相同特征分辨率,并分别被融合到各自的之前特征图318-b和318-a中。

[0066] 在一些实施例中,特征图解码过程330被配置为将经更新的不同尺度特征图的序

列组合成综合特征图332,并根据综合特征图332确定一个或多个边界框334。经更新的特征图序列包括经由特征图上采样过程320输出的特征图318-a'、318-b'和318-c。可替代地,在一些实施例中,特征图解码过程330被配置为将不同尺度的特征图318的序列组合成综合特征图332,并根据综合特征图332确定一个或多个边界框334。经更新的特征图序列包括由第二机器学习模块316输出的特征图318-a、318-b和318-c。每个边界框对应于图像数据312中捕获的对象。

[0067] 边界框选择过程340识别一个或多个边界框334中满足置信度标准的的边界框子集338。在一些实施例中,根据置信度标准,边界框选择过程340基于检测到的边界框334是姿势的相应概率或置信度水平336,识别一个或多个边界框中的边界框子集338。例如,在检测到边界框334之后,基于相应的概率或置信度水平336在列表中对边界框334进行排序,并将排序在列表最上方的预定数量的框选择为边界框子集338。在另一示例中,根据置信度标准选择边界框子集338,使得边界框子集338中的每个边界框对应相应的一个预定义姿势的概率或置信度水平336大于阈值。换言之,选择一个或多个最可靠的边界框338来确定图像数据312中的最终姿势380。

[0068] 边界框选择过程340的输出被用于基于对象检测过程350和/或分类过程370来确定最终姿势380。对象检测过程350检测边界框中的一个或多个潜在对象,如图像352中所示的,并且图像数据312被裁剪为手部区域355。也就是说,裁剪处理将图像数据312裁剪为集中于姿势的手部区域355。分类网络360被用于手部区域355,以确定第一姿势向量365。更具体地,基于从综合特征图332中确定的所提取手部区域355,分类网络360确定至少第一姿势向量365。在一些实施例中,分类网络360是本领域中已知的一个或多个网络(例如,诸如mobilenet v1网络、mobilenet v2网络、ShuffleNet网络等)。在一些实施例中,分类网络360是基于预期的姿势任务(例如,特定应用和/或系统所预期的姿势)来选择的,和/或基于用于分类的多个类别(例如,可以由特定应用和/或系统进行分类的不同类型的姿势)来选择的。

[0069] 可选地,在一些实施例中,姿势识别过程300不包括分类网络360(而是仅使用端到端框架来检测和识别姿势)。通过移除分类网络360,姿势识别过程300可以在计算量较少但准确率较低的情况下检测和识别姿势(例如,在准确率和响应时间之间进行权衡)。在一些实施例中,基于特定应用和/或系统的要求,移除分类网络360。例如,在一些实施例中,当电子设备的一个或多个处理器202(图2;例如(一个或多个)CPU、GPU、DSP、NPU、APU)的功能较弱(和/或使用较便宜的处理器)时,移除分类网络360。在一些实施例中,分类网络360的移除将姿势识别过程300的推理时间减少了至少30%。

[0070] 在一些实施例中,基于由边界框选择过程340生成的所选边界框338,分类过程370识别由第二姿势向量375表示的一个或多个分类姿势。分类过程370以相对更直接的方式从综合特征图332中确定第二姿势向量375。也就是说,直接从图像数据312生成第二姿势向量375,而不是从图像数据312裁剪出的手部区域355生成第二姿势向量375。

[0071] 在一些实施例中,分类过程370与对象检测过程350、裁剪过程和分类网络360是并行执行的。通过对一个或多个对象检测姿势365和一个或多个分类姿势375进行组合,来生成图像数据312的最终姿势380。基于综合特征图332(例如,边界框选择过程340的输出)来确定姿势365和姿势375两者。在一些实施例中,姿势365和姿势375分别由第一姿势向量365

和第二姿势向量375表示。通过确定集中在从图像数据312裁剪出的手部区域355上的第一姿势向量365,以及使用分类过程370确定第二姿势向量375,基于综合特征图322来确定最终姿势380。

[0072] 在一些实施例中,第一姿势向量和第二姿势向量(例如,分别为姿势365和姿势375)被组合成最终姿势向量380。例如,在一些实施例中,基于最终姿势向量380(例如,包括多个元素的姿势向量,每个元素表示各自预定义手势的概率)来确定最终姿势380。在一些实施例中,具有最高概率(或满足预定阈值和/或预定置信度水平)的一个或多个对象检测姿势365和/或一个或多个分类姿势375被用于确定最终姿势向量。在一些实施例中,第一姿势向量365和第二姿势向量375以加权方式进行组合。例如,第一姿势向量365和第二姿势向量375具有相等的权重0.5。在一示例中,第一姿势向量365和第二姿势向量375分别具有权重1和0。不应用分类过程370,最终姿势380是从手部区域355得到的。相反,在另一示例中,第一姿势向量365和第二姿势向量375分别具有权重0和1。应用分类过程370来确定最终姿势380,而没有应用对象检测过程350、裁剪过程和分类过程360。

[0073] 在一些实施例中,第一姿势向量365、第二姿势向量375和最终姿势向量380中的每一个都具有预定义数量的元素,该预定义数量的元素对应于相同数量的不同姿势。各姿势向量365、375或380的每个元素指示检测到的姿势对应于各自不同姿势的概率。对于每个姿势向量365、375或380,预定义数量的元素被归一化。在一些实施例中,对于最终姿势向量380,选择对应于最大概率的预定义手势作为最终姿势380。

[0074] 尽管图3示出了确定手势的示例,但在一些实施例中,姿势识别过程300被配置成检测面部姿势、手臂姿势、身体姿势、手势和/或用户执行的其他可分类姿势中的一个或多个。

[0075] 图4示出了根据一些实施例的姿势识别过程的示例图像处理阶段400的流程图。图像处理阶段400示出了上面参考图3描述的图像处理阶段310的一个或多个组件。图像处理阶段400示出了使用第一机器学习模块314(例如,包括主干网络410)、使用第二机器学习模块316(例如,包括多尺度特征提取网络420)来降低接收的图像数据312的分辨率。多尺度特征提取网络420包括网络420A、网络420B、网络420C和网络420D,其被配置成基于图像数据312生成特征图318-aa、特征图318-a、特征图318-b和特征图318-c的序列,每个特征图318具有在特征图318的序列中连续地降尺度的相应特征分辨率。

[0076] 在一些实施例中,接收的图像数据312具有至少1080p的分辨率。在一些实施例中,接收的图像数据312被缩小(例如,通过第一神经网络410)到288×512×3的输入尺寸。图像数据312可以被缩小到任何分辨率或输入尺寸。通过使用包括输入网络430和瓶颈网络440的网络420A,第一神经网络410的输出被进一步缩小。在一些实施例中,神经网络410和神经网络430是3×3核(kernel,卷积核)且步长为2(conv-k3s2)的卷积层。在一些实施例中,网络410和网络430的卷积层之后是激活函数(例如,校正线性激活函数(rectified linear activation function,ReLU))。在一些实施例中,激活函数是恒等函数(identity function)。神经网络410和神经网络430的缩小步长较低,使图3中姿势识别过程300开始时的信息损失的水平较低,并在每个相应输出的小卷积核尺寸范围内保持尽可能多的局部信息。

[0077] 第一机器学习模块314包括第一瓶颈结构(瓶颈-1)和第二瓶颈结构(瓶颈-2)。在

一些实施例中,第一机器学习模块314被称为主干结构。如上文参考图3所述,瓶颈结构(例如,第一机器学习模块314)用于利用尽可能多的细节,以及压缩大型网络(即,第一神经网络410、第二神经网络420和第三神经网络430的输出)。

[0078] 第一瓶颈结构包括第一神经网络结构440和跳跃连接(skip-connection)结构445。在一些实施例中,第一神经网络结构440包括:1×1逐点卷积层440a,在卷积层440a之后的步长为1且3×3核的深度卷积层440b,以及在卷积层440b之后的1×1逐点卷积层440c。在一些实施例中,第一神经网络结构440的每个卷积层(例如,卷积层440a、卷积层440b和卷积层440c)包括激活函数(或恒等函数)。在一些实施例中,跳跃连接结构445被应用于第三神经网络430的输出,并经由加法运算447连接到第一神经网络结构440的输出。跳跃连接结构445被配置成通过保留由第一神经网络结构440生成的特征来帮助收敛。在一些实施例中,跳跃连接结构445是残差块(ResBlock)结构。

[0079] 第二瓶颈结构包括第二神经网络结构450。在一些实施例中,第二神经网络结构450包括:1×1逐点卷积层450a,卷积层450a之后的步长为2的3×3核的深度卷积层440b(其将空间维度降级一半),以及在卷积层450b之后的1×1逐点卷积层450c。在一些实施例中,第二神经网络结构450的每个卷积层(例如,卷积层450a、卷积层450b和卷积层450c)包括激活函数(或恒等函数)。

[0080] 在一些实施例中,为了进一步压缩第一神经网络410、第二神经网络420和第三神经网络430的输出(例如,经缩小的图像数据)同时还保持性能,第一瓶颈结构440和第二瓶颈结构450的通道数为3的倍数而不是6的倍数。基于对各瓶颈结构的一个或多个参数进行微调使得整体准确率和训练效率不降低,选择使用通道数为3的倍数的第一瓶颈结构440和第二瓶颈结构450。

[0081] 在一些实施例中,第二机器学习模块316包括第一瓶颈结构(瓶颈-1)、第二瓶颈结构(瓶颈-2)和第三神经网络结构460。第三神经网络结构460包括多个卷积层(第一卷积层460a、第二卷积层460b和第三卷积层460c)。在一些实施例中,多个卷积层中的第一卷积层460a步长为2(conv-k3s2)。在一些实施例中,多个卷积层中的第二卷积层460b和第三卷积层460c步长为1(conv-k3s1)。在一些实施例中,第三神经网络结构460的每个卷积层(例如,卷积层460a、卷积层460b和卷积层460c)包括激活函数(或恒等函数)。

[0082] 在一些实施例中,第二机器学习模块316被称为多尺度特征提取模型。第二机器学习模块316被配置成对第一机器学习模块314的输出(例如,特征图)进行逐渐降尺度,从而以最小的性能损失降低第一机器学习模块314的输出的整体复杂性(与具有较大核尺寸的输出相比)。例如,在一些实施例中,第二机器学习模块316使用第一瓶颈结构的输出来生成第一特征图318a,使用第二瓶颈结构来生成第二特征图318b,以及使用第三神经网络结构460来生成第三特征图318c(例如,上面参考图3所描述的用于上采样的特征图序列)。第二机器学习模块316(即,多尺度特征金字塔)起着关键作用,因为它在每次缩小(降尺度)后都输出特征图。不同尺寸的特征图可以表示不同的细节水平。例如,图3和图4所示的第二机器学习模块316输出至少三种不同尺寸的特征图。具有较大特征尺寸的特征图(例如,第一特征图318a)允许识别较小的姿势(例如,由较小的手做出的姿势,或与捕获图像数据312的图像设备距离较远的姿势),而具有较小特征尺寸的特征图(例如,第三特征图318c)允许检测较大的边界框(例如,从特征图解码过程330解码出的;图3)。这确保了在远离捕获图像数据

的成像设备处捕获的和在该成像设备附近捕获到的相同姿势将被一致地识别为相同姿势，这将有助于增加姿势识别过程300的准确性，从而提高系统的性能。

[0083] 姿势识别过程的图像处理阶段400的优点是第一机器学习模块314、第二机器学习模块316和第三神经网络结构460保持特征图的超低计算复杂度的能力，这使得能够在具有有限计算资源的电子设备(例如，智能手机、智能手表、智能眼镜和/或具有较弱的处理器的其他电子设备)上进行实时姿势识别。此外，使用姿势识别过程的图像处理阶段400减少了电子设备的功耗，增加了电子设备的电池总寿命。

[0084] 图5示出了根据一些实施例的姿势检测和识别方法的流程图。方法500包括执行上面参考图3和4所描述的姿势识别过程300和图像处理阶段400。如上文所述，姿势识别过程300和图像处理阶段400被配置成在使用有限数量的计算资源(例如，使用3.5GLOPS或更少)的情况下检测和识别图像数据中的一个或多个姿势。有限的计算资源以GLOPS作为测量单位，并被控制在预定义的资源阈值内，使得可以在计算资源低于有限计算资源的许多电子设备上实现姿势识别。姿势检测和识别过程500的操作(例如，步骤)由电子设备(例如，服务器102和/或客户端设备104)的一个或多个处理器202(图2；例如多个CPU、GPU、DSP、NPU、APU)。图5所示的至少一些操作对应于存储在计算机存储器或计算机可读存储介质(例如，存储器206；图2)的指令。操作502至508也可以部分地使用一个或多个处理器和/或使用存储在通信地耦合在一起的一个或多个设备的存储器或计算机可读介质中的指令来执行，该一个或多个设备例如为膝上型电脑、AR眼镜(或其他头戴式显示器)、服务器、平板电脑、安防相机、智能电视、智能扬声器、玩具、智能手表、智能电器、或者可以单独地或结合通信地耦合的电子设备200的相应处理器来执行操作502-508的其他计算设备。

[0085] 方法500包括获取(502)图像的图像数据。该图像捕获到手部。在一些实施例中，手部正在做出姿势。例如，如上文参考图3所示，图像数据312包括由用户的手部做出的姿势。在一些实施例中，图像数据包括相对于捕获图像数据的电子设备至少4米远的姿势。也就是说，手部是由距离该手部至少4米远的相机设备捕获在图像中的。在一些实施例中，图像数据以至少1080p的分辨率被捕获。

[0086] 方法500还包括基于图像数据生成(504)特征图序列。特征图序列具有连续降尺度的特征分辨率。例如，如上文参考图3所示，由图像处理阶段(如上文参考图3和4所述)生成特征图序列(例如，第一特征图、第二特征图和第三特征图，318a-318c)。在一些实施例中，使用第一机器学习模块(例如，上文参考图3和图4所述的第一机器学习模块314)基于图像数据生成特征图序列。第一机器学习模块被配置成从图像数据中识别和丰富(例如，提取和/或生成更显著的细节)一个或多个特征，并相对于图像数据的分辨率降低特征图序列的特征分辨率。

[0087] 在一些实施例中，基于将主干网络应用于图像数据，生成特征图序列。主干网络包括至少两个卷积层和多个瓶颈结构。例如，如图4所示，主干网络可以包括第二神经网络420和第三神经网络430(例如，作为至少两个卷积层)以及第一机器学习模块314，其在图4中由至少第一瓶颈结构(瓶颈-1)和第二瓶颈结构(瓶颈-2)表示。在一些实施例中，主干网络的至少两个卷积层中的每个卷积层具有步长为2的3×3核，至少两个卷积层中的第一卷积层之后是至少两个卷积层中的第二卷积层。例如，第二神经网络420之后是第三神经网络430。在一些实施例中，第一瓶颈结构被配置成从至少两个卷积层的输出中提取一个或多个特

征,第一瓶颈结构包括第一1×1逐点卷积层,在该第一1×1逐点卷积层之后的第一3×3深度卷积层,以及在第一3×3深度卷积层之后的第二1×1逐点卷积层和包括跳过连接结构的加法运算;第二瓶颈结构被配置成缩小第一瓶颈结构的输出,第二瓶颈结构包括第三1×1逐点卷积层,在第三1×1逐点卷积层之后的步长为2的第二3×3深度卷积层,在第二3×3深度卷积层之后的第四1×1逐点卷积层。在一些实施例中,第二瓶颈结构的输出是多个特征图。在一些实施例中,多个瓶颈结构中的每个瓶颈结构都具有3层和3的倍数个通道。参考图4在上文提供了关于至少两个卷积层和多个瓶颈结构的附加信息。

[0088] 在一些实施例中,生成特征图序列包括:生成具有第一特征分辨率的第一特征图;根据缩放因子将第一特征图缩小为第二特征图,该第二特征图在特征图序列中紧跟在第一特征图之后;以及根据缩放因子将第二特征图缩小为第三特征图,该店特征图在特征图序列中紧跟在第二特征图之后。例如,如上文参考图4所示,在一些实施例中,第二机器学习模块316包括第一瓶颈结构(瓶颈-1)、第二瓶颈结构(瓶颈-2)和第三神经网络结构460,其被配置成对第一机器学习模块314的输出逐渐地降尺度,从而以最小的性能损失降低第一机器学习模块314的输出的整体复杂性。具体地,第二机器学习模块316生成第一特征图318a、第二特征图318b和第三特征图318c。在一些实施例中,每个卷积层(上文参考图4所述)之后是激活函数(或恒等函数)。

[0089] 方法500包括组合(506)特征图序列以生成综合特征图。更具体地,在一些实施例中,第一特征图、第二特征图和第三特征图318a-318c被融合在一起。在一些实施例中,在每个接后的特征图在被融合之前被上采样到与其之前的特征图相同的空间尺寸。例如,如上文参考图3所示,对第三特征图318c进行上采样并将其与第二特征图318b融合,对第二特征图318b进行上采样并将其与第一特征图318a融合。

[0090] 在一些实施例中,综合特征图(例如,上文参考图3示出和描述的融合特征图)被解码,以从融合特征图序列中识别一个或多个边界框。在一些实施例中,方法500包括:识别一个或多个边界框中满足置信度标准的边界框子集;以及基于一个或多个边界框中的边界框子集格式化综合特征图。

[0091] 方法500包括根据综合特征图确定(508)图像中的手势。在一些实施例中,基于综合特征图确定手势包括:基于综合特征图检测和提取图像数据中的手部区域(例如,图3中的手部区域355)。基于从综合特征图提取的手部区域355来确定第一姿势向量(例如,对应于图3中的对象检测姿势365)。从与图像的图像数据对应的综合特征图中确定第二姿势向量(例如,对应于图3中的分类姿势375)。图像包括手部区域和不同于手部区域的一个或多个其余区域。组合第一姿势向量和第二姿势向量以生成据其来生成手势的最终姿势向量。要注意的事,第一手势是基于整个图像中的手部区域即基于整个图像的相应部分来确定的。相反地,第二手势是直接根据整个图像的综合特征图确定的。

[0092] 在一些实施例中,方法500被配置成检测图像数据中的姿势,该姿势与捕获图像数据的电子设备距离至少4米远。以上参考图3提供了关于检测和识别过程的附加信息。

[0093] 应当理解,对图5中的操作进行描述的特定顺序仅仅是示例性的,并非旨在表示所述顺序是这些操作可被执行的唯一顺序。本领域的普通技术人员会想到如本文所述的检索候选图像或确定相机姿态的各种方式。另外,应当指出的是,上文参考图3和图4的其他过程的细节同样以类似的方式适用于上文参考图5所述的方法500。为简洁起见,这里不再重复

这些细节。

[0094] 在本文中对各种所述实施方案的描述中所使用的术语只是为了描述特定实施方案的目的,而并非旨在进行限制。如在对各种所述实施方案中的描述和所附权利要求书中所使用的那样,除非上下文另外明确地指示,单数形式“一个”(“a”,“an”)和“该”旨在也包括复数形式。还应当理解,本文中所使用的术语“和/或”是指并且涵盖相关联地列出的项目中的一个或多个项目的任何和全部可能的组合。还将理解的是,术语“包括”(“includes”、“including”、“comprises”和/或“comprising”)在本说明书中使用时是指定存在所陈述的特征、整数、步骤、操作、元件和/或部件,但是并不排除存在或添加一个或多个其他特征、整数、步骤、操作、元件、部件和/或其分组。还将理解的是,尽管术语“第一”、“第二”等在本文中用于描述各种元件,但是这些元件不应受到这些术语限制。这些术语只是用来将一个元件与另一元件区分开。

[0095] 如本文中所使用,根据上下文,术语“如果”任选地被解释为意思是“当.....时”(“when”或“upon”)或“响应于确定”或“响应于检测到”或“根据.....的确定”。类似地,根据上下文,短语“如果确定.....”或“如果检测到[所陈述的条件或事件]”任选地被解释为是指“在确定.....时”或“响应于确定.....”或“在检测到[所陈述的条件或事件]时”或“响应于检测到[所陈述的条件或事件]”或“根据确定检测到[所陈述的条件或事件]”。

[0096] 出于解释的目的,前面的描述是通过参考具体实施例来描述的。然而,上面的例示性论述并非旨在是穷尽的或将本发明限制为所公开的精确形式。根据以上教导内容,很多修改和变型都是可能的。选择并描述这些实施例是为了最好地解释这些技术的原理及其实际应用,从而使得本领域的其他技术人员能够实施。

[0097] 虽然各个附图以特定顺序图示出数个逻辑阶段,但这些与顺序无关的阶段可以被重新排序,并且其他阶段可以被组合或者分解。尽管具体提及一些重新排序或其他分组,但其他方式对于本领域普通技术人员而言将是明显的,因此在本文中呈现的排序和分组并非穷举替选方案。此外,应认识到,这些阶段能够以硬件、固件、软件或其任意组合来实现。

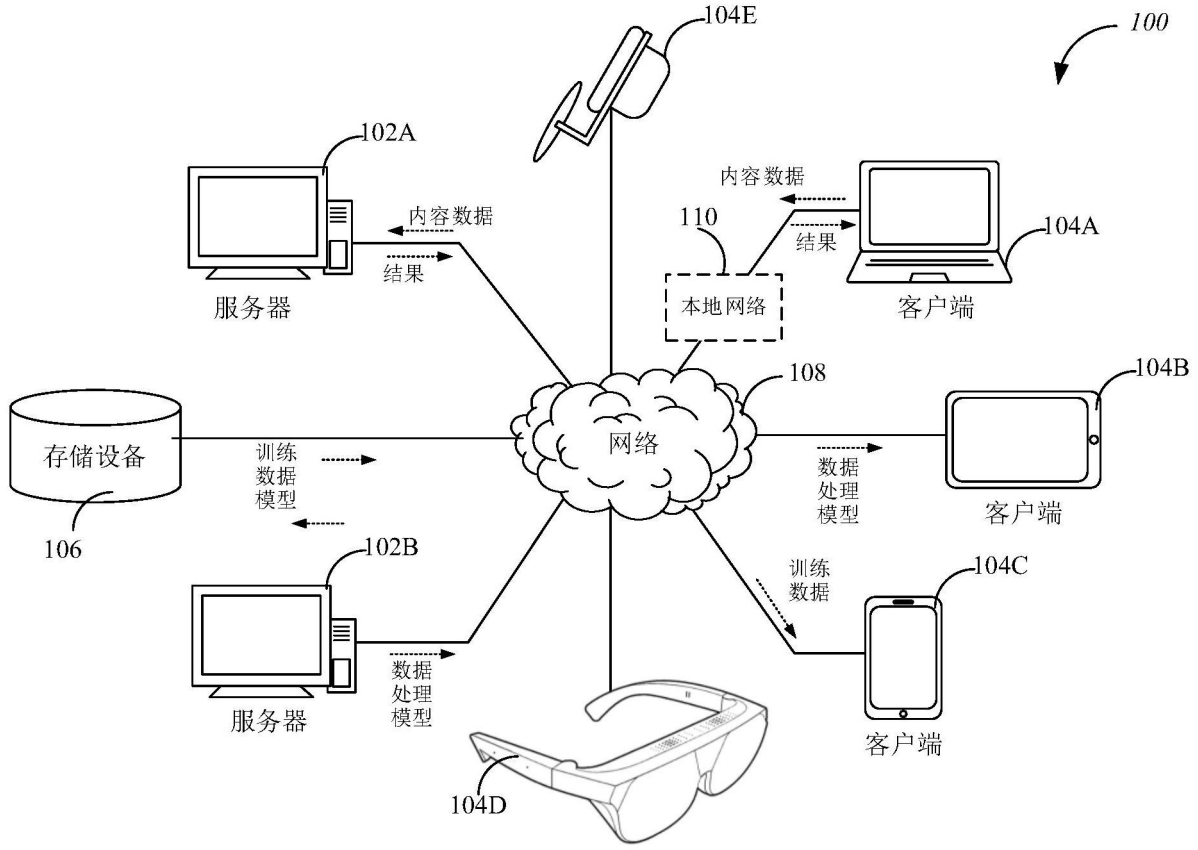

100
104E
102A
内容数据
结果
服务器
内容数据
110
结果
104A
本地网络
客户端
108
104B
存储设备
网络
训练
数据
模型
数据
处理
模型
客户端
106
102B
训练
数据
104C
数据
处理
模型
服务器
客户端
104D

图1

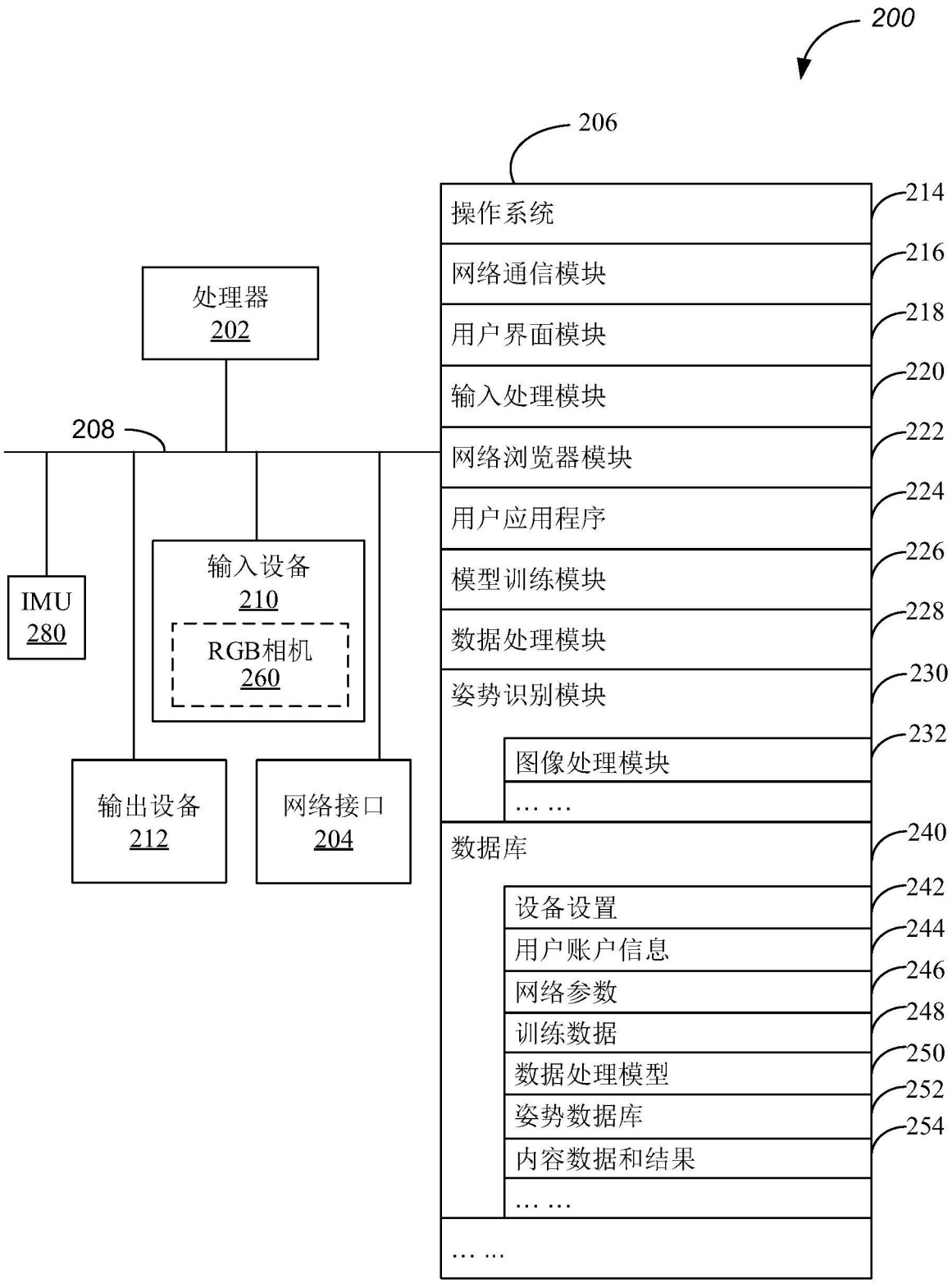

200
206
操作系统
214
网络通信模块
216
用户界面模块
218
输入处理模块
220
网络浏览器模块
222
用户应用程序
224
模型训练模块
226
数据处理模块
228
姿势识别模块
230
图像处理模块
232
… …
数据库
240
设备设置
242
用户账户信息
244
网络参数
246
训练数据
248
数据处理模型
250
姿势数据库
252
内容数据和结果
254
… …
… …
处理器
202
208
IMU
280
输入设备
210
RGB相机
260
输出设备
212
网络接口
204

图2

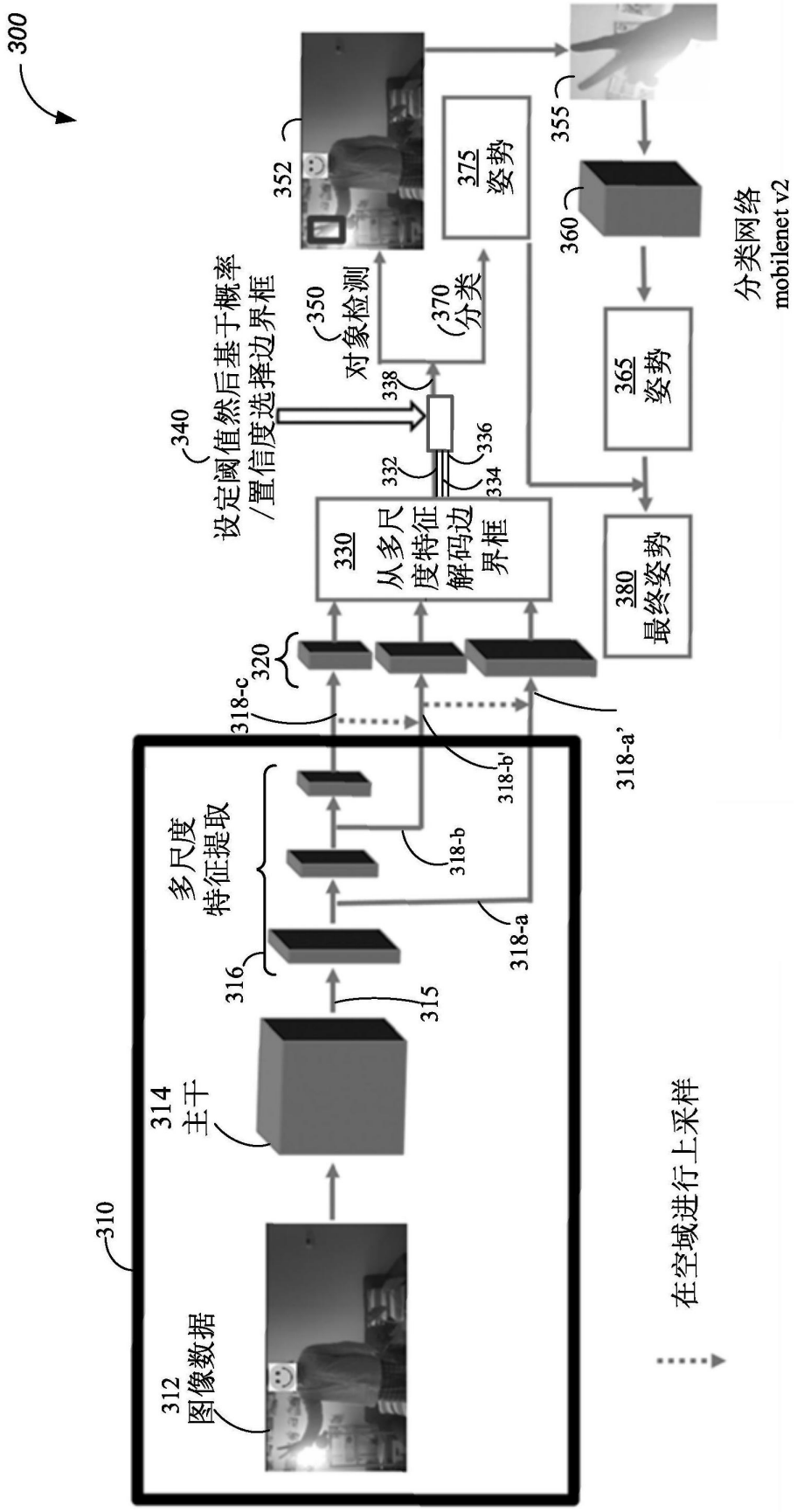

300
312
图像数据
310
314
主干
316
多尺度
特征提取
315
318-a
318-b
318-b'
318-c
320
318-a'
330
从多尺度特征解码边界框
332
334
336
338
340
设定阈值然后基于概率/置信度选择边界框
350
对象检测
352
370
分类
375
姿势
355
360
365
姿势
380
最终姿势
分类网络
mobilenet v2
在空域进行上采样

图3

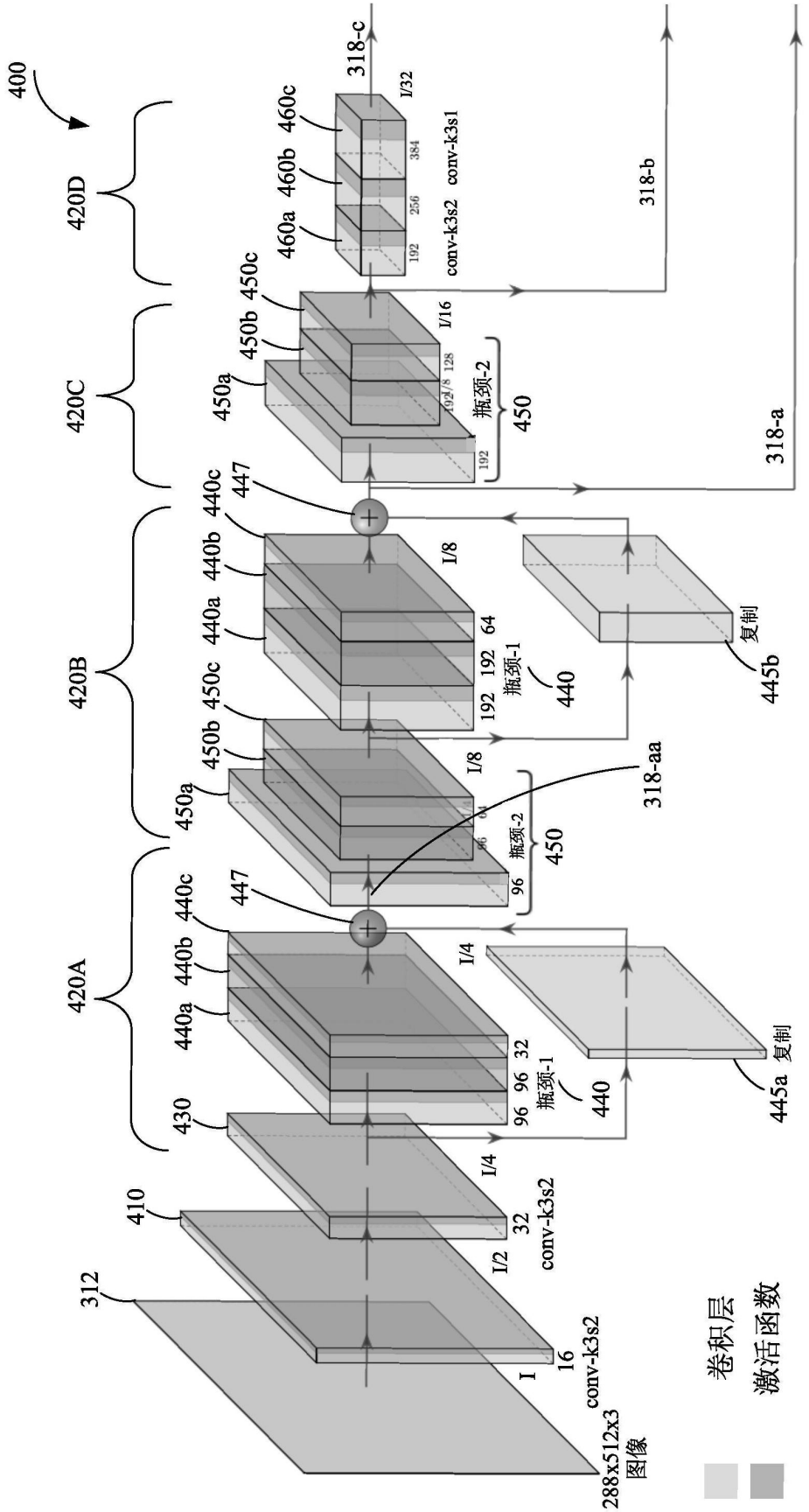

400
420D
420C
420B
420A
312
410
430
440
440a
440b
440c
447
445a
445b
450
450a
450b
450c
460a
460b
460c
318-a
318-b
318-c
318-aa
288x512x3
图像
I
16
conv-k3s2
I/2
32
conv-k3s2
I/4
96
96
32
瓶颈-1
复制
I/8
192
192
64
瓶颈-1
96
瓶颈-2
192/8
128
瓶颈-2
192
I/16
conv-k3s2
192
256
384
conv-k3s1
I/32
卷积层
激活函数

图4

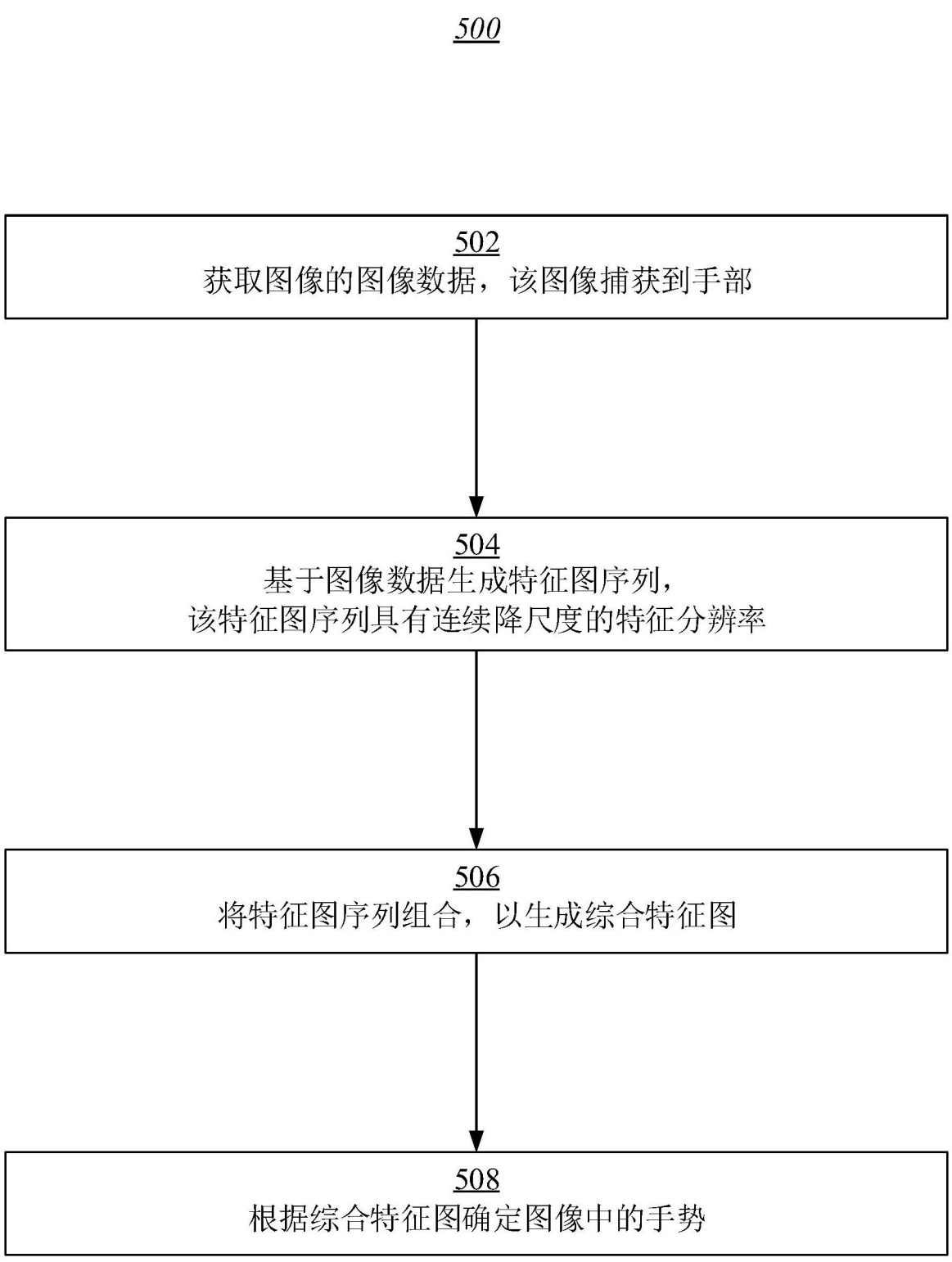

500
502
获取图像的图像数据，该图像捕获到手部
504
基于图像数据生成特征图序列，
该特征图序列具有连续降尺度的特征分辨率
506
将特征图序列组合，以生成综合特征图
508
根据综合特征图确定图像中的手势

图5