(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 113894779 B
(45)授权公告日 2023.10.17

(21)申请号 202111066753.9

(22)申请日 2021.09.10

(65)同一申请的已公布的文献号
申请公布号 CN 113894779 A

(43)申请公布日 2022.01.07

(73)专利权人 人工智能与数字经济广东省实验室(广州)
地址 510308 广东省广州市海珠区新港东路2429号首层自编051房
专利权人 西安电子科技大学广州研究院

(72)发明人 石光明 张凡 李旭阳 谢雪梅

(74)专利代理机构 广州三环专利商标代理有限公司 44202
专利代理师 麦小婵 郝传鑫

(51)Int.Cl.
B25J 9/16(2006.01)
G06F 18/25(2023.01)
G06N 3/0464(2023.01)
G06N 3/084(2023.01)

(56)对比文件
CN 111055279 A,2020.04.24
CN 112580362 A,2021.03.30
CN 110232412 A,2019.09.13
CN 112388655 A,2021.02.23
CN 111204476 A,2020.05.29
US 2021125052 A1,2021.04.29

审查员 彭劲松

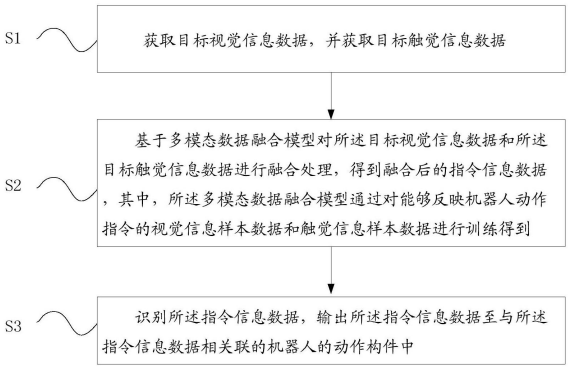
权利要求书2页 说明书10页 附图5页

(54)发明名称

一种应用于机器人交互的多模态数据处理方法

(57)摘要

本发明公开了一种应用于机器人交互的多模态数据处理方法,其中方法包括获取目标视觉信息数据,并获取目标触觉信息数据;基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理,得到融合后的指令信息数据,其中,所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到;识别所述指令信息数据,输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中。本发明实施例提供的应用于机器人交互的多模态数据处理方法,通过对视觉数据和触觉数据进行多模态的融合,提高了数据处理的精度,推进了机器人的智能化进程。



CN 113894779 B

1.一种应用于机器人交互的多模态数据处理方法,其特征在于,包括:

获取目标视觉信息数据,并获取目标触觉信息数据;

基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理,得到融合后的指令信息数据,其中,所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到;

识别所述指令信息数据,输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中;

所述多模态数据融合模型的构建,具体包括:

将所述视觉信息样本数据转换为高维视觉数据,并将所述触觉信息样本数据转换为高维触觉数据;

分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征;

基于CNN卷积神经网络技术,由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征;

对所述融合特征进行识别,根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

2.如权利要求1所述的应用于机器人交互的多模态数据处理方法,其特征在于,所述分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征,具体包括:

基于C3D三维卷积网络提取所述高维视觉数据的视觉特征,并基于TCN时间卷积网络提取所述高维触觉数据的触觉特征。

3.如权利要求1所述的应用于机器人交互的多模态数据处理方法,其特征在于,所述方法还包括:

基于CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。

4.一种应用于机器人交互的多模态数据处理装置,其特征在于,包括:

数据获取模块,用于获取目标视觉信息数据,并获取目标触觉信息数据;

模型处理模块,用于根据多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理,得到融合后的指令信息数据,其中,所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到;

指令输出模块,用于识别所述指令信息数据,输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中;

所述多模态数据融合模型,包括:

高维转换单元,用于将所述视觉信息样本数据转换为高维视觉数据,并将所述触觉信息样本数据转换为高维触觉数据;

特征提取单元,用于分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征;

特征融合单元,用于根据CNN卷积神经网络技术,由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征;

识别单元,用于对所述融合特征进行识别,根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

5.如权利要求4所述的应用于机器人交互的多模态数据处理装置,其特征在于,所述特

征提取单元，具体用于：

基于C3D三维卷积网络提取所述高维视觉数据的视觉特征，并基于TCN时间卷积网络提取所述高维触觉数据的触觉特征。

6.如权利要求4所述的应用于机器人交互的多模态数据处理装置，其特征在于，所述应用于机器人交互的多模态数据处理装置还包括：

模型训练模块，用于根据CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。

7.一种应用于机器人交互的多模态数据处理设备，其特征在于，包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如权利要求1至3中任意一项所述的应用于机器人交互的多模态数据处理方法。

8.一种计算机可读存储介质，其特征在于，所述计算机可读存储介质包括存储的计算机程序，其中，在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如权利要求1至3中任意一项所述的应用于机器人交互的多模态数据处理方法。

一种应用于机器人交互的多模态数据处理方法

技术领域

[0001] 本发明涉及机器人技术领域，尤其是涉及一种应用于机器人交互的多模态数据处理方法、装置、设备及存储介质。

背景技术

[0002] 随着人工智能技术的不断发展，智能家居机器人凭借着灵活的机械爪，高灵敏度的各种传感器，不仅可以识别三维物体，而且能够执行各种复杂的动作，逐渐受到了消费者的青睐。

[0003] 为了实现“人与物”之间的信息交互，进而达到家居机器人控制的智能化，如何对接收到的数据进行处理就显得尤为重要，在现有技术中，主要依靠图像识别和神经网络等相关技术，通过将摄像头采集到的图像信号输入至训练好的神经网络模型中，得到能够反映目标用户需求的数据，再将数据传递至机器人的控制器中以生成对应的动作信号，由机器人的动作构件执行动作。

[0004] 但是，当智能家居机器人在复杂的诸如遮挡物较多的环境下时，图像信号的获取不可避免地会受到影响，导致数据的处理精度降低，难以反映真实的用户需求，不利于机器人的智能化发展。

发明内容

[0005] 本发明提供了一种应用于机器人交互的多模态数据处理方法、装置、设备及存储介质，以解决现有的机器人仅依靠视觉数据，导致数据的处理精度较低的技术问题，通过对视觉数据和触觉数据进行多模态的融合，提高了数据处理的精度，推进了机器人的智能化进程。

[0006] 为了解决上述技术问题，本发明实施例提供了一种应用于机器人交互的多模态数据处理方法，包括：

[0007] 获取目标视觉信息数据，并获取目标触觉信息数据；

[0008] 基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0009] 识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中。

[0010] 作为其中一种优选方案，所述多模态数据融合模型的构建，具体包括：

[0011] 将所述视觉信息样本数据转换为高维视觉数据，并将所述触觉信息样本数据转换为高维触觉数据；

[0012] 分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征；

[0013] 基于CNN卷积神经网络技术，由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征；

[0014] 对所述融合特征进行识别，根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

[0015] 作为其中一种优选方案，所述分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征，具体包括：

[0016] 基于C3D三维卷积网络提取所述高维视觉数据的视觉特征，并基于TCN时间卷积网络提取所述高维触觉数据的触觉特征。

[0017] 作为其中一种优选方案，所述方法还包括：

[0018] 基于CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。

[0019] 本发明另一实施例提供了一种应用于机器人交互的多模态数据处理装置，包括：

[0020] 数据获取模块，用于获取目标视觉信息数据，并获取目标触觉信息数据；

[0021] 模型处理模块，用于根据多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0022] 指令输出模块，用于识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中。

[0023] 作为其中一种优选方案，所述多模态数据融合模型，包括：

[0024] 高维转换单元，用于将所述视觉信息样本数据转换为高维视觉数据，并将所述触觉信息样本数据转换为高维触觉数据；

[0025] 特征提取单元，用于分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征；

[0026] 特征融合单元，用于根据CNN卷积神经网络技术，由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征；

[0027] 识别单元，用于对所述融合特征进行识别，根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

[0028] 作为其中一种优选方案，所述特征提取单元，具体用于：

[0029] 基于C3D三维卷积网络提取所述高维视觉数据的视觉特征，并基于TCN时间卷积网络提取所述高维触觉数据的触觉特征。

[0030] 作为其中一种优选方案，所述应用于机器人交互的多模态数据处理装置还包括：

[0031] 模型训练模块，用于根据CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。

[0032] 本发明再一实施例提供了一种应用于机器人交互的多模态数据处理设备，包括处理器、存储器以及存储在所述存储器中且被配置为由所述处理器执行的计算机程序，所述处理器执行所述计算机程序时实现如上所述的应用于机器人交互的多模态数据处理方法。

[0033] 本发明再一实施例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包括存储的计算机程序，其中，在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如上所述的应用于机器人交互的多模态数据处理方法。

[0034] 相比于现有技术，本发明实施例的有益效果在于以下所述中的至少一点：首先获取目标视觉信息数据，并获取目标触觉信息数据，从而为后续的数据处理提供了两个模态

的数据类型;然后基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理,得到融合后的指令信息数据,其中,所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到;最后识别所述指令信息数据,输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中,由机器人的动作构件执行相应的诸如机械爪的开启或关闭动作。由此可见,整个过程针对于视觉和触觉两个模态的数据类型,通过构建的多模态数据融合模型进行数据处理,输出能够反映机器人动作的指令信息,不同于现有技术的仅依靠视觉数据的单一方式,当外界环境存在遮挡物时,机器人能够根据接收到的视觉数据和触觉数据进行同步分析处理,从而提高了复杂环境下机器人对数据处理的精度,实现了机器人的数据感知、判断和分析,以及决策控制,从而更好地保证机器人的工作效率,保障机器人的动作合理性,大幅度减少因数据处理方面而导致的决策失误,处理后的数据更能精准地反映用户的需求,完善后续对机器人的动作构件的控制流程,推进了机器人的智能化进程。

附图说明

[0035] 图1是本发明其中一种实施例中的应用于机器人交互的多模态数据处理方法的流程示意图;

[0036] 图2是本发明其中一种实施例中的应用于机器人交互的多模态数据处理方法的流程框图;

[0037] 图3是本发明其中一种实施例中的应用于机器人交互的多模态数据处理装置的结构示意图;

[0038] 图4是本发明其中一种实施例中的模型处理模块的结构示意图;

[0039] 图5是本发明其中一种实施例中的应用于机器人交互的多模态数据处理设备的结构框图;

[0040] 附图标记:

[0041] 其中,11、数据获取模块;12、模型处理模块;13、指令输出模块;14、模型训练模块;121、高维转换单元;122、特征提取单元;123、特征融合单元;124、识别单元;10处理器;20、存储器。

具体实施方式

[0042] 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,都属于本发明保护的范围。

[0043] 在本申请描述中,术语“第一”、“第二”、“第三”等仅用于描述目的,而不能理解为指示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”、“第三”等的特征可以明示或者隐含地包括一个或者更多个该特征。在本申请的描述中,除非另有说明,“多个”的含义是两个或两个以上。

[0044] 在本申请的描述中,需要说明的是,除非另有明确的规定和限定,术语“安装”、“相连”、“连接”应做广义理解,例如,可以是固定连接,也可以是可拆卸连接,或一体地连接;可

以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通。对于本领域的普通技术人员而言，可以具体情况理解上述术语在本申请中的具体含义。

[0045] 在本申请的描述中，需要说明的是，除非另有定义，本发明所使用的所有的技术和科学术语与属于本的技术领域的技术人员通常理解的含义相同。本发明中说明书中所使用的术语只是为了描述具体的实施例的目的，不是旨在于限制本发明，对于本领域的普通技术人员而言，可以具体情况理解上述术语在本申请中的具体含义。

[0046] 本发明一实施例提供了一种应用于机器人交互的多模态数据处理方法，具体的，请参见图1，图1示出为本发明其中一种实施例中的应用于机器人交互的多模态数据处理方法的流程示意图，其中包括步骤S1～S3：

[0047] S1、获取目标视觉信息数据，并获取目标触觉信息数据；

[0048] S2、基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0049] S3、识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中。

[0050] 应当说明的是，现有的机器人在进行数据处理的交互过程中，是依靠视觉数据来实现，但在数据的交互过程中经常会存在由于遮挡物挡住机器人的摄像头，导致图像不清晰的现象，即使在图像中添加大规模的注释数据，也难以概括到真实的图像，因此，机器人仅依靠图像技术，无法得到全面的用户需求，智能化效果较低。

[0051] 此外，需要说明的是，本实施例中的卷积神经网络的训练分析模型为预先训练好的模型，也可采用诸如深度神经网络的其他模型，由多个神经元连接在一起构成多层网络，多层网络包括卷积层、池化层和全连接层，卷积层用于提取特征向量，池化层对特征向量进一步筛选，全连接层进行最后的融合输出，各层之间通过激活函数连接在一起，并通过训练算法不断改进网络参数，从而提高模型的识别准确率，在训练过程中，首先需要提供海量的训练样本，样本可根据实际的数据类型进行参数调整，例如视觉数据，只提供海量的用户手部动作图片，通过对每一手部动作图片进行逐像素分析，获取每一手部动作图片对应的机器人操作动作，从而能够针对不同的图片，得到准确的用户需求，方便与其他数据类型的结果进行比对和融合，触觉数据亦可参阅上述步骤，在此不再赘述。

[0052] 本发明实施例提供的应用于机器人交互的多模态数据处理方法，多模态数据融合模型在其中扮演者数据融合的重要作用，数据融合的对象是视觉数据和触觉数据，正如背景技术所说，现有的智能家居机器人都配置有多种多样的高灵敏度传感器，故视觉数据和触觉数据都可以借助相应的传感器通过采样记录得到，例如，视觉数据可以有机器人顶部位置配置的多角度摄像头通过图像采集得到，而在机器人的动作部位诸如机械爪的指尖位置还设置有接触传感器或应力传感器，用于判断机器人(主要指四肢)是否接触到外界物体或测量被接触物体的触觉特征。摄像头和接触传感器的结构、型号与数量由实际的机器人的设计要求与场景限制决定，在此不再赘述。

[0053] 在机器人的运动过程中，其工作场景不断发生变化，考虑到工作环境诸如室内都放置有大体型的家具等物件，位于顶部的机器人摄像头难以兼顾各方面的视角，当用户处

于机器人的视野盲点下发指令时，仅凭借摄像头只能获取部分用户图像信息，这就会导致得到的数据不够准确，难以反映真实的用户需求。但触觉传感器不会存在上述技术缺陷，因此，本实施例除了获取目标视觉信息数据外，还会获取目标触觉信息数据，具体的，请参见图2，图2示出为本发明其中一种实施例中的应用于机器人交互的多模态数据处理方法的流程框图，当然，除了获取触觉信息据，还需要对获取到的两个类型的信息数据进行处理(即图中的融合步骤)，方能转化得到能够被机器人的控制器准确识别的数据，从而优化对数据的处理过程。

[0054] 进一步地，在上述实施例中，所述多模态数据融合模型的构建，具体包括：

[0055] 将所述视觉信息样本数据转换为高维视觉数据，并将所述触觉信息样本数据转换为高维触觉数据；

[0056] 分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征；

[0057] 基于CNN卷积神经网络技术，由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征；

[0058] 对所述融合特征进行识别，根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

[0059] 当然，在模型的构建过程中，样本数据的场景(例如需要机器人传递扳手螺丝刀的场景)、数量(例如视觉数据的序列设为若干个，触觉数据的序列设为若干个)、对象(可以分为用户类型和机械手类型，例如包括六个用户对机械爪C和机械爪D两个动作构件)、动作类型(机械爪传递物品和不传递物品)等参数指标均需要结合具体的训练要求进行设定，在此不再赘述。

[0060] 在本实施例中，首先设置构建用户和机器人之间的数据交互所用到的数据集的参数，例如视觉采集设备A和触觉采集设备B(一种高精度全向触觉传感器)，机械臂C和机械爪D、需要采集的六个人员(三男、三女)的手及相应的编号，传递12种物品的类别编号已经不同的场景S，$S=\{S_1,S_2,...,S_i,...,S_n\}$，其中i表示第i个场景，n表示场景总数，n可以为家庭场景、医疗场景与工作场景(如传递扳手、螺丝刀等的场景)。

[0061] 然后通过采集设备A采集在不同场景S_i传递过程中的原始视频，$V=\{V_1,V_2,...,V_i,...,V_n\}$，其中，$V_i$表示视频采集设备A对$S_i$进行采集所得到场景的原始视频，$V_i=\{V_{i1},V_{i2},...,V_{ij},...,V_{im}\}$，$V_{ij}$表示视频采集设备A对不同人员传递过程进行采集所得到的原始视频；

[0062] 将每个原始视频V_{ij}剪辑为1_{ij}个视频段，$V'_{ij}=\{V'^{1}_{ij},V'^{2}_{ij},...,V'^{k_{ij}}_{ij},...,V'^{l_{ij}}_{ij}\}$，$V'^{k_{ij}}_{ij}$表示一个人员与机器人传递交互不同物体的视频段；

[0063] 通过采集设备B采集在不同场景S_i传递过程中的原始触觉信号(一维数据)，$T=\{T_1,T_2,....,T_i,....T_n\}$，其中，$T_i$表示采集设备B对$S_i$进行采集所得到场景的原始触觉信号序列，$T_i=\{T_{i1},T_{i2},....,T_{ij},....T_{im}\}$，$T_{ij}$表示视频采集设备B对不同人员传递过程进行采集所得到的原始触觉信号序列；

[0064] 将每个原始触觉序列T_{ij}剪辑为1_{ij}个触觉序列，$T'_{ij}=\{T'^{1}_{i1},T'^{2}_{i2},....,T'^{k_{ij}}_{ij},....T'^{l_{ij}}_{im}\}$，$T'^{k_{ij}}_{ij}$表示一个人员与机器人传递交互不同物体的触觉序列。

[0065] 接着，建立集合M＝{M_1,M_2,...,M_i,...,M_n}，M_i表示某一物品在不同场景下的视频段、触觉序列和标签的集合。

$M_i=\{(V_{i1}^{'1},T_{i1}^{'1},X_{i1}^{'1}),(V_{i2}^{'2},T_{i2}^{'2},X_{i2}^{'2}),...,(V_{ij}^{'k_{ij}},T_{ij}^{'k_{ij}},X_{ij}^{'k_{ij}}),...,(V_{in}^{'l_{ij}},T_{in}^{'l_{ij}},X_{in}^{'l_{ij}})\}$，$(V_{in}^{'l_{ij}},T_{in}^{'l_{ij}},X_{in}^{'l_{ij}})$表示一个数据集的样本，$X_{in}^{'l_{ij}}$表示传递对应操作动作，X＝{$a_1$,$a_2$}，$a_1$表示机械爪的张开，$a_2$表示机械爪的闭合操作，其中视频段和触觉序列是对齐的。

[0066] 最后，在得到上述的数据集之后，就可以对网络进行训练，将数据集M＝{M_1,M_2,...,M_i,...,M_n}分为训练集和测试集，从训练集中拿到样本，样本包括正负样本$(V_{in}^{'l_{ij}},T_{in}^{'l_{ij}},X_{in}^{'l_{ij}})$，从样本中取出视频段$V_{ij}^{'k_{ij}}$、触觉序列$T_{ij}^{'k_{ij}}$和操作动作标签$X_{in}^{'l_{ij}}$。

[0067] 将样本输入到网络中，从视频段中提取出视频特征，这里使用神经网络(如3d卷积神经网络)来提取视频特征，从触觉序列中提取出触觉特征，这里使用神经网络(如因果卷积)来提取中间特征，将两个模态的特征进行融合，得到语义特征，将语义特征输入到传递网络中。

[0068] 将传递网络的输出与样本的动作标签进行拟合，通过反向传播的方式，对这个网络进行端到端的训练。

[0069] 将测试集数据输入到训练好的网络之中，来判断网络的好坏。

[0070] 将训练好的网络运用到实际的环境中，使用拍摄设备得到视频信息，通过机械爪上的高精度全向触觉传感器得到实时的触觉序列，将得到的数据输入到上述训练好的网络，就可以完成人机之间的数据处理交互。

[0071] 为便于理解，下面进行举例说明(相关参数的命名可能与上述不一致，以下述内容中的参数为准)，从训练样本集中取出视觉信息样本数据集合V和触觉信息样本数据集合T，转换为给定视觉序列(X_{v1},X_{v2},…,X_{vm})以及触觉序列(X_{t1},X_{t2},…,X_{tn})，其中m,n分别对应为视觉序列的长度和触觉序列的长度。通过视觉(E_v)和触觉(E_t)编码器函数提取视觉(F_v)和触觉(F_t)特征，之后基于F_v以及F_t构建一个融合特征($F_{v,t}$)。最后，$F_{v,t}$被输入分类函数$\mathbb{F}_c$以预测当前的传递语义y。这个过程可以用公示表示：

[0072] $$F_{v,t} = E_v(X_{v1},X_{v2},..,X_{vm}) \oplus E_t(X_{t1},X_{t2},..,X_{tn})$$

[0073] $$y = \mathbb{F}_c(F_{v,t}) \quad y \in 0,1$$

[0074] 上式中的0、1分别表示机械爪的动作类型，即相当于“不传递物品”与“传递物品”，优选地，上式中的E_v使用3D卷积神经网络实现，而E_t则使用时间卷积网络实现，同时上式中的$\mathbb{F}_c$则是由全连接层(FC)构成。给定当前的视觉和触觉序列，C3D-TCN的输出就是当前的场景中特征对应的动作类型。

[0075] 在本实施例中，优选地，基于C3D三维卷积网络提取所述高维视觉数据的视觉特征，可以使用5个的视觉图像序列作为输入，神经网络包含5层C3D和2层FC，每个卷积层的卷积核大小和步幅大小并不完全相同，从C3D层输出的特征被输入到2层FC层并变成4096维特征向量。

[0076] 在本实施例中，优选地，基于TCN时间卷积网络提取所述高维触觉数据的触觉特征，可以使用10个与图像序列对应时间的触觉序列作为输入，采用时间卷积网络(TCN)，最

终输出为5维特征向量。

[0077] 经过对视觉数据的特征提取和触觉数据的特征提取后，接下来对这两种模态的数据特征进行融合。多模态数据融合模型中采用的是合并(concat)操作，将两个不同模态提取到的特征进行concat特征融合，之后获取到综合了两个模态之后获得的语义特征，以神经网络为例，中间融合首先利用神经网络将原始数据转化成高维特征表达，然后获取不同模态数据在高维空间上的共性在进行融合，最终经过分类网络，在这里就是2层全连接层(FC)之后输出能够反映机器人动作指令的类别结果，也就是机械爪的动作类型，即相当于“不传递物品”与“传递物品”。

[0078] 在本发明另一实施例中，将频率为30帧的视频数据和采样率为1Mhz的触觉序列数据实时输入到训练好的基于神经网络的融合模型中；其中，3D卷积网络提取出视频数据的特征，时间卷积网络提取出触觉序列的特征；再将提取出来的两个特征进行融合得到新的特征；最后识别新的特征得到反映机械爪不同操作的信息类别。

[0079] 进一步地，所述应用于机器人交互的多模态数据处理方法还包括：

[0080] 基于CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。此外，训练好的多模态数据融合模型的好坏可以通过正确率来判定，例如将数据输入到训练好的网络模型中，看网络模型的输出结果和预期结果的差别，比如将视频和触觉输入到网络中后，机械爪执行了传递物品的操作，而数据集中进行的是不传递物品的操作，记录这样的错误的个数就可以得到正确率。当然，基于现有的神经网络技术中的其他方式诸如构建奖励函数均可以实现对训练好的多模态数据融合模型的好坏进行判定，根据不同的判定结果可以调整模型的样本集参数，以优化最终得到的多模态数据融合模型。

[0081] 在得到最终的多模态数据融合模型后，就可以将目标视觉信息数据和目标触觉信息数据输入至其中进行处理，以通过数据融合的方式得到准确的数据信息。

[0082] 本发明实施例提供的应用于机器人交互的多模态数据处理方法，整个过程针对于视觉和触觉两个模态的数据类型，通过构建的多模态数据融合模型进行数据处理，输出能够反映机器人动作的指令信息，不同于现有技术的仅依靠视觉数据的单一方式，当外界环境存在遮挡物时，机器人能够根据接收到的视觉数据和触觉数据进行同步分析处理，从而提高了复杂环境下机器人对数据处理的精度，实现了机器人的数据感知、判断和分析，以及决策控制，从而更好地保证机器人的工作效率，保障机器人的动作合理性，大幅度减少因数据处理方面而导致的决策失误，处理后的数据更能精准地反映用户的需求，完善后续对机器人的动作构件的控制流程，推进了机器人的智能化进程。

[0083] 本发明另一实施例提供了一种应用于机器人交互的多模态数据处理装置，具体的，请参见图3，图3示出为本发明其中一种实施例中的应用于机器人交互的多模态数据处理装置的结构示意图，其包括：

[0084] 数据获取模块11，用于获取目标视觉信息数据，并获取目标触觉信息数据；

[0085] 模型处理模块12，用于根据多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0086] 指令输出模块13，用于识别所述指令信息数据，输出所述指令信息数据至与所述

指令信息数据相关联的机器人的动作构件中。

[0087] 具体的，请参见图4，图4示出为本发明其中一种实施例中的模型处理模块的结构示意图，在上述实施例中，所述模型处理模块12包括：

[0088] 高维转换单元121，用于将所述目标视觉信息数据转换为高维视觉数据，并将所述目标触觉信息数据转换为高维触觉数据；

[0089] 特征提取单元122，用于分别提取所述高维视觉数据的视觉特征和所述高维触觉数据的触觉特征；

[0090] 特征融合单元123，用于根据CNN卷积神经网络技术，由卷积神经网络的全连接层输出对所述视觉特征和所述触觉特征进行特征共性融合后得到的融合特征；

[0091] 识别单元124，用于对所述融合特征进行识别，根据识别出的与所述融合特征相关联的机器人动作指令构建多模态数据融合模型。

[0092] 进一步地，在上述实施例中，所述特征提取单元122，具体用于：

[0093] 基于C3D三维卷积网络提取所述高维视觉数据的视觉特征，并基于TCN时间卷积网络提取所述高维触觉数据的触觉特征。

[0094] 进一步地，在上述实施例中，所述应用于机器人交互的多模态数据处理装置还包括：

[0095] 模型训练模块14，用于根据CNN卷积神经网络中的反向传播算法对所述多模态数据融合模型进行训练。

[0096] 参见图5，其是本发明实施例提供的应用于机器人交互的多模态数据处理设备的结构框图，本发明实施例提供的应用于机器人交互的多模态数据处理设备20，包括处理器21、存储器22以及存储在所述存储器22中且被配置为由所述处理器21执行的计算机程序，所述处理器21执行所述计算机程序时实现如上述应用于机器人交互的多模态数据处理方法实施例中的步骤，例如图1中所述的步骤S1～S3；或者，所述处理器21执行所述计算机程序时实现上述各装置实施例中各模块的功能，例如数据获取模块11。

[0097] 示例性的，所述计算机程序可以被分割成一个或多个模块，所述一个或者多个模块被存储在所述存储器22中，并由所述处理器21执行，以完成本发明。所述一个或多个模块可以是能够完成特定功能的一系列计算机程序指令段，该指令段用于描述所述计算机程序在所述应用于机器人交互的多模态数据处理设备20中的执行过程。例如，所述计算机程序可以被分割成数据获取模块11、模型处理模块12、指令输出模块13，各模块具体功能如下：

[0098] 数据获取模块11，用于获取目标视觉信息数据，并获取目标触觉信息数据；

[0099] 模型处理模块12，用于根据多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0100] 指令输出模块13，用于识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中。

[0101] 所述应用于机器人交互的多模态数据处理设备20可包括，但不仅限于，处理器21、存储器22。本领域技术人员可以理解，所述示意图仅仅是应用于机器人交互的多模态数据处理设备的示例，并不构成对应用于机器人交互的多模态数据处理设备20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,例如所述应用于机器人交互的多模态数据处理设备20还可以包括输入输出设备、网络接入设备、总线等。

[0102] 所述处理器21可以是中央处理单元(Central Processing Unit,CPU),还可以是其他通用处理器、数字信号处理器(Digital Signal Processor,DSP)、专用集成电路(Application Specific Integrated Circuit,ASIC)、现成可编程门阵列(Field-Programmable Gate Array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件等。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等,所述处理器21是所述应用于机器人交互的多模态数据处理设备20的控制中心,利用各种接口和线路连接整个应用于机器人交互的多模态数据处理设备20的各个部分。

[0103] 所述存储器22可用于存储所述计算机程序和/或模块,所述处理器21通过运行或执行存储在所述存储器22内的计算机程序和/或模块,以及调用存储在存储器22内的数据,实现所述应用于机器人交互的多模态数据处理设备20的各种功能。所述存储器22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手机的使用所创建的数据(比如音频数据、电话本等)等。此外,存储器22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,例如硬盘、内存、插接式硬盘,智能存储卡(Smart Media Card,SMC),安全数字(Secure Digital,SD)卡,闪存卡(Flash Card)、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。

[0104] 其中,所述应用于机器人交互的多模态数据处理设备20集成的模块如果以软件功能单元的形式实现并作为独立的产品销售或使用时,可以存储在一个计算机可读取存储介质中。基于这样的理解,本发明实现上述实施例方法中的全部或部分流程,也可以通过计算机程序来指令相关的硬件来完成,所述的计算机程序可存储于一计算机可读存储介质中,该计算机程序在被处理器执行时,可实现上述各个方法实施例的步骤。其中,所述计算机程序包括计算机程序代码,所述计算机程序代码可以为源代码形式、对象代码形式、可执行文件或某些中间形式等。所述计算机可读介质可以包括:能够携带所述计算机程序代码的任何实体或装置、记录介质、U盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器(ROM,Read-Only Memory)、随机存取存储器(RAM,Random Access Memory)、电载波信号、电信信号以及软件分发介质等。

[0105] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程,是可以通过计算机程序来指令相关的硬件来完成,所述的程序可存储于一计算机可读取存储介质中,该程序在执行时,可包括如上述各方法的实施例的流程。其中,所述的存储介质可为磁碟、光盘、只读存储记忆体(Read-Only Memory,ROM)或随机存储记忆体(Random Access Memory,RAM)等。

[0106] 相应地,本发明实施例提供一种计算机可读存储介质,所述计算机可读存储介质包括存储的计算机程序,其中,在所述计算机程序运行时控制所述计算机可读存储介质所在设备执行如上述实施例的应用于机器人交互的多模态数据处理方法中的步骤,例如图1中所述的步骤S1～S3。

[0107] 本发明实施例提供的应用于机器人交互的多模态数据处理方法、装置、设备及存储介质,有益效果在于以下所述中的至少一点:

[0108] 首先获取目标视觉信息数据，并获取目标触觉信息数据，从而为后续的数据处理提供了两个模态的数据类型；

[0109] 然后基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到；

[0110] 最后识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中，由机器人的动作构件执行相应的诸如机械爪的开启或关闭动作；

[0111] 整个过程针对于视觉和触觉两个模态的数据类型，通过构建的多模态数据融合模型进行数据处理，输出能够反映机器人动作的指令信息，不同于现有技术的仅依靠视觉数据的单一方式，当外界环境存在遮挡物时，机器人能够根据接收到的视觉数据和触觉数据进行同步分析处理，从而提高了复杂环境下机器人对数据处理的精度，实现了机器人的数据感知、判断和分析，以及决策控制，从而更好地保证机器人的工作效率，保障机器人的动作合理性，大幅度减少因数据处理方面而导致的决策失误，处理后的数据更能精准地反映用户的需求，完善后续对机器人的动作构件的控制流程，推进了机器人的智能化进程。

[0112] 以上所述是本发明的优选实施方式，应当指出，对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和润饰，这些改进和润饰也视为本发明的保护范围。

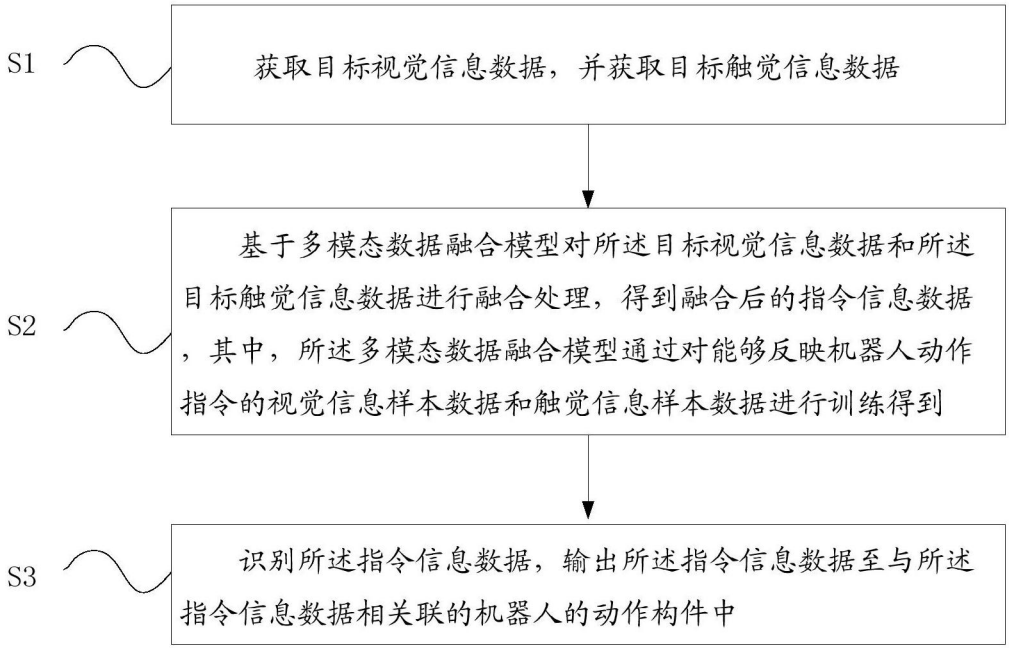

S1
获取目标视觉信息数据，并获取目标触觉信息数据
S2
基于多模态数据融合模型对所述目标视觉信息数据和所述目标触觉信息数据进行融合处理，得到融合后的指令信息数据，其中，所述多模态数据融合模型通过对能够反映机器人动作指令的视觉信息样本数据和触觉信息样本数据进行训练得到
S3
识别所述指令信息数据，输出所述指令信息数据至与所述指令信息数据相关联的机器人的动作构件中

图1

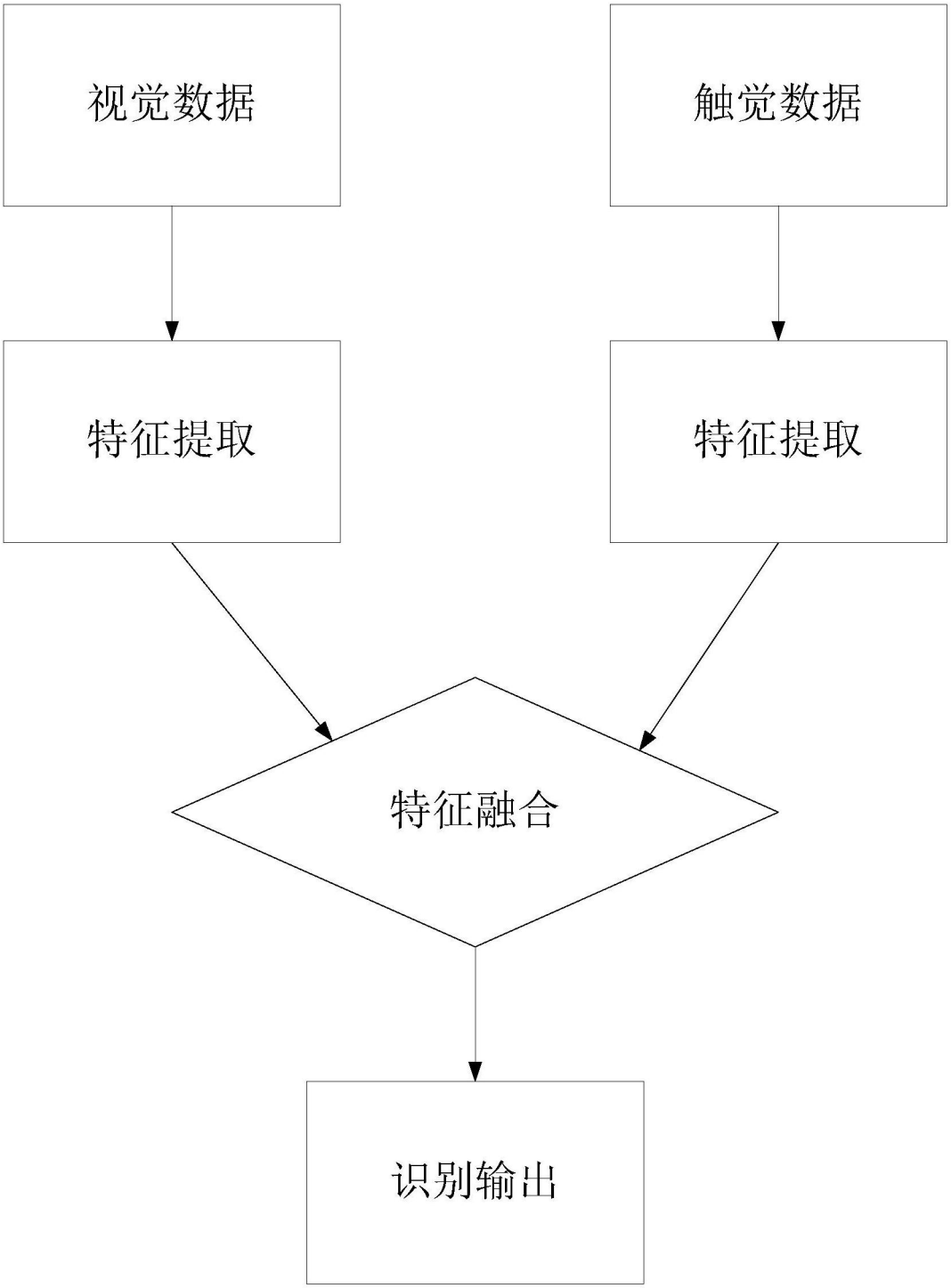


图2

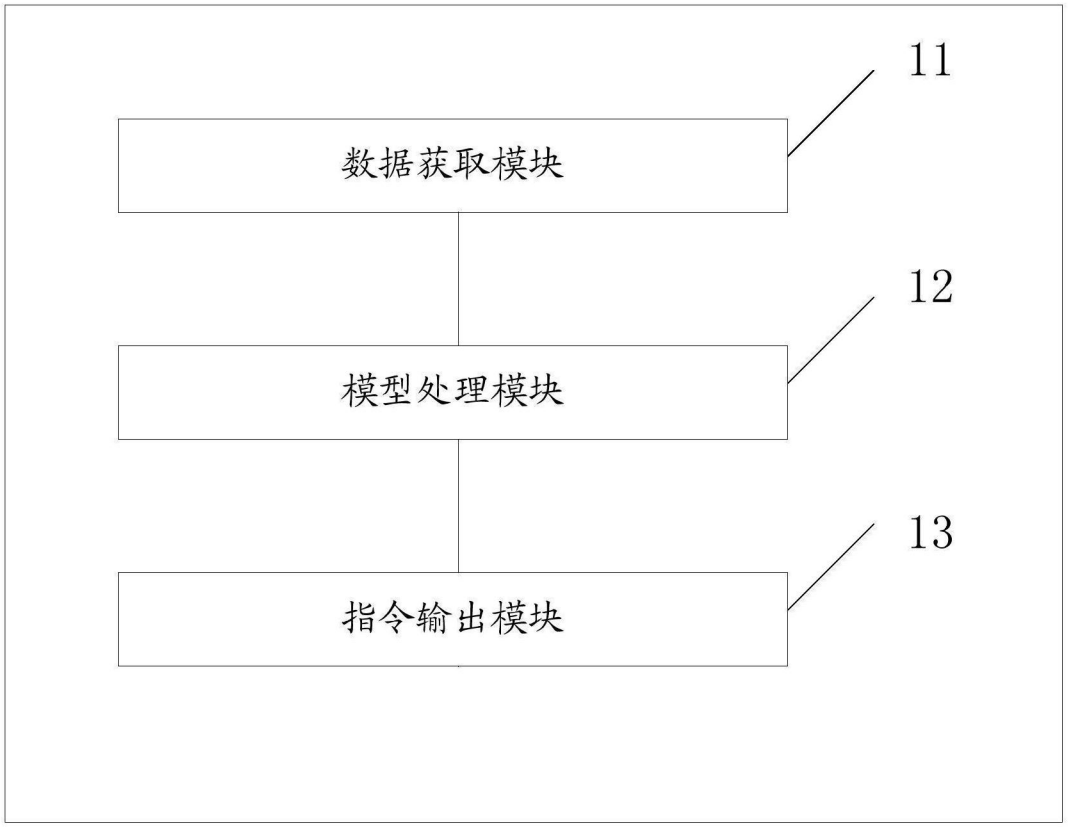


图3

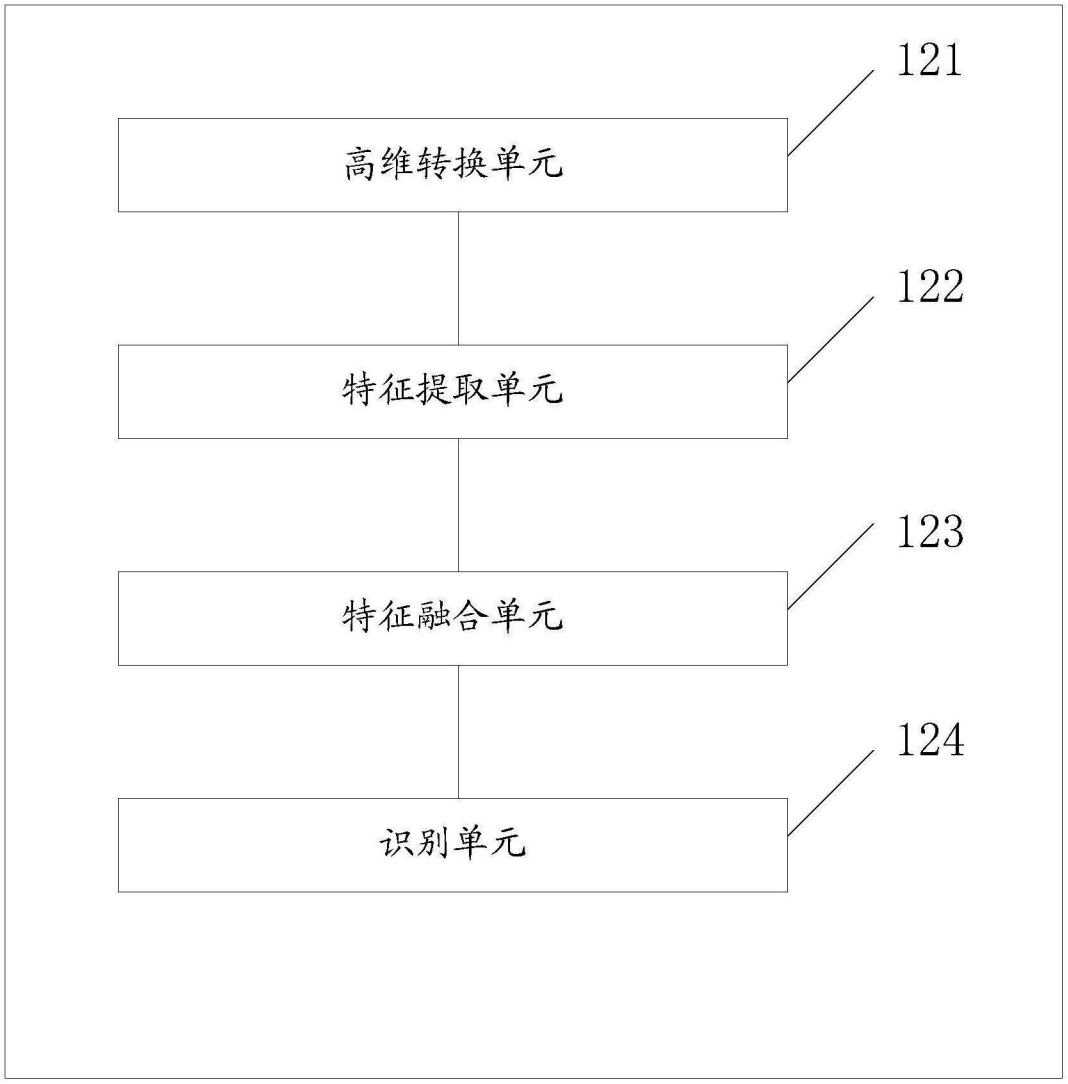


图4

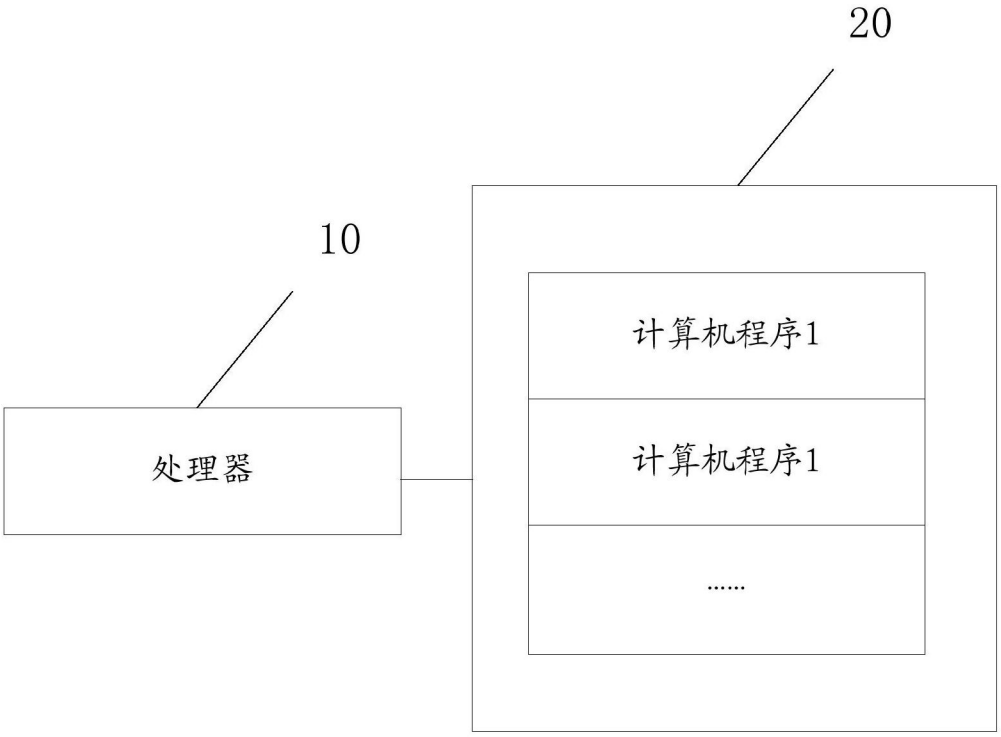


图5