(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 105913039 B
(45)授权公告日 2020.08.18

(21)申请号 201610265947.4

(22)申请日 2016.04.26

(65)同一申请的已公布的文献号
申请公布号 CN 105913039 A

(43)申请公布日 2016.08.31

(73)专利权人 北京光年无限科技有限公司
地址 100000 北京市石景山区石景山路3号玉泉大厦四层常青藤青年创业工作室193号

(72)发明人 徐振敬 陆羽皓

(74)专利代理机构 北京聿华联合知识产权代理有限公司 11611
代理人 张文娟 朱绘

(51)Int.Cl.
G06K 9/00(2006.01)
G06F 40/30(2020.01)
H04L 12/58(2006.01)
G10L 15/02(2006.01)
G10L 15/00(2013.01)

(56)对比文件
CN 104464733 A,2015.03.25
CN 103123619 A,2013.05.29
US 2016/0055885 A1,2016.02.25
CN 1797284 A,2006.07.05

审查员 魏旭阳

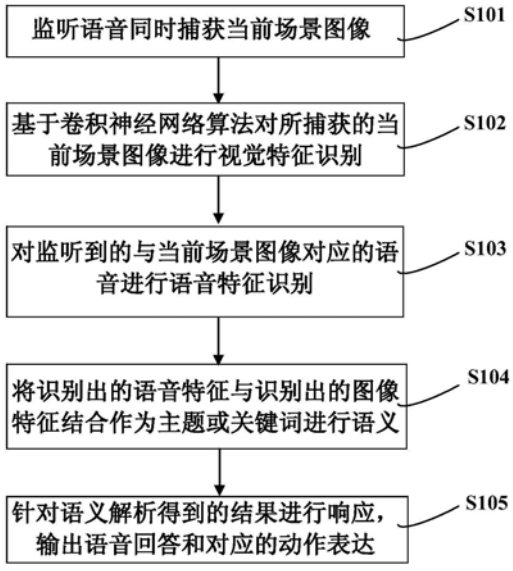
权利要求书2页 说明书6页 附图2页

(54)发明名称

基于视觉和语音的对话数据交互处理方法及装置

(57)摘要

本发明提供了一种基于视觉和语音的对话数据交互处理方法,其包括以下步骤:在监听语音的同时,捕获对应的当前场景图像;基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别;对监听到的与当前场景图像对应的语音进行语音特征识别;将识别出的语音特征与识别出的图像特征结合起来作为主题或关键词以进行语义解析;针对语义解析得到的结果进行响应,输出语音回答和对应的动作表达。由于采用了视觉识别与语音识别技术的结合得到对话话题的关键词,同时根据对话生成模型或者答案搜索库的方式提供答案,使得使用者同机器人的聊天可以更加符合用户的意图,对于用户的提问或者给出的话题可以给出适当的回复。



CN 105913039 B

1.一种基于视觉和语音的对话数据交互处理方法，其特征在于，所述处理方法包括以下步骤：

在监听语音的同时，捕获对应的当前场景图像，其中，所述当前场景图像包括人的图像、物体图像、动作序列以及背景图像；

基于卷积神经网络算法对所捕获的当前场景图像进行分类视觉特征识别，对所捕获的当前场景图像进行分类视觉识别包括对当前场景图像进行人体身份特征识别、主题特征识别以及发出动作的意图特征识别；其中，针对所捕获的当前场景图像进行进一步的分类，针对人体图像，需要识别面部特征、衣服颜色特征、样式特征、发型，而对于物体图像，需要识别出是属于哪一类物体，文具、玩具、宠物，对于背景图像，需要识别出环境特征，天气、地理位置，对于捕获的一系列动作，需要识别出人发出这些动作与语音配合的意图特征；

对监听到的与当前场景图像对应的语音进行语音特征识别；

将识别出的语音特征与分类识别出的图像特征结合起来作为主题或关键词以进行语义解析，其中，将所述识别出的语音特征和所述分类识别出的图像特征作为知识库匹配答案的参考项进行输出；

针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。

2.如权利要求1所述的基于视觉和语音的对话数据交互处理方法，其特征在于，在进行语义解析的步骤中，将所述识别出的语音特征和所述分类识别出的图像特征作为循环神经网络的对话生成模块的额外输入信息进行输出。

3.如权利要求1或2所述的基于视觉和语音的对话数据交互处理方法，其特征在于，在进行语义解析的步骤中，还需要发出命令机器人的其他运动系统做出某动作的信号。

4.一种基于视觉和语音的对话数据交互处理装置，其特征在于，所述处理装置包括：

监听与捕获模块，其用于在监听语音的同时，捕获对应的当前场景图像，其中，所述当前场景图像包括人的图像、物体图像、动作序列以及背景图像；

图像识别模块，其用于基于卷积神经网络算法对所捕获的当前场景图像进行分类视觉特征识别，在图像识别模块中，还包括对当前场景图像进行人体身份特征识别的身份识别单元、主题特征识别单元以及发出动作的意图特征识别单元；其中，针对所捕获的当前场景图像进行进一步的分类，针对人体图像，需要识别面部特征、衣服颜色特征、样式特征、发型，而对于物体图像，需要识别出是属于哪一类物体，文具、玩具、宠物，对于背景图像，需要识别出环境特征，天气、地理位置，对于捕获的一系列动作，需要识别出人发出这些动作与语音配合的意图特征；

语音识别模块，其用于对监听到的与当前场景图像对应的语音进行语音特征识别；

语义解析模块，其用于将识别出的语音特征与分类识别出的图像特征结合起来作为主题或关键词以进行语义解析，其中，在语义解析模块中还包括答案搜索接口单元，其用于将所述识别出的语音特征和所述分类识别出的图像特征作为知识库匹配答案的参考项进行输出；

对话输出模块，针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。

5.如权利要求4所述的基于视觉和语音的对话数据交互处理装置，其特征在于，在语义解析模块中还包括对话生成模块接口单元，其用于将所述识别出的语音特征和所述识别出的图像特征作为循环神经网络的对话生成模块的额外输入信息进行输出。

6.如权利要求4或5所述的基于视觉和语音的对话数据交互处理装置，其特征在于，所述处理装置还包括动作决策模块，其中所述动作决策模块与所述语义解析模块耦接，以接收所述语义解析模块发出的命令机器人的其他运动系统做出某动作的信号，并根据该信号作出相应动作的决策。

基于视觉和语音的对话数据交互处理方法及装置

技术领域

[0001] 本发明涉及智能机器人领域，具体地说，涉及一种基于视觉和语音的对话数据交互处理方法及装置。

背景技术

[0002] 在对话数据交互的技术领域中，需要提供一种能够让智能机器人根据当前聊天场景下的各种特征综合给出对话答案的交互数据处理方法或系统，从而提高用户的使用体验，满足用户的聊天需求。

发明内容

[0003] 为解决现有技术的上述问题，本发明提供了一种基于视觉和语音的对话数据交互处理方法，所述处理方法包括以下步骤：

[0004] 在监听语音的同时，捕获对应的当前场景图像；

[0005] 基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别；

[0006] 对监听到的与当前场景图像对应的语音进行语音特征识别；

[0007] 将识别出的语音特征与识别出的图像特征结合起来作为主题或关键词以进行语义解析；

[0008] 针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。

[0009] 根据本发明的一个实施例，在基于视觉和语音的对话数据交互处理方法中，对所捕获的当前场景图像进行视觉识别包括对当前场景图像进行人体身份特征识别、主题特征识别以及发出动作的意图特征识别。

[0010] 根据本发明的一个实施例，在进行语义解析的步骤中，将所述识别出的语音特征和所述识别出的图像特征作为知识库匹配答案的参考项进行输出。

[0011] 根据本发明的一个实施例，在进行语义解析的步骤中，将所述识别出的语音特征和所述识别出的图像特征作为循环神经网络的对话生成模块的额外输入信息进行输出。

[0012] 根据本发明的一个实施例，在进行语义解析的步骤中，还需要发出要指令机器人的其他运动系统做出某动作的信号。

[0013] 根据本发明的另一个方面，还提供了一种基于视觉和语音的对话数据交互处理装置，所述处理装置包括：

[0014] 监听与捕获模块，其用于在监听语音的同时，捕获对应的当前场景图像；

[0015] 图像识别模块，其用于基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别；

[0016] 语音识别模块，其用于对监听到的与当前场景图像对应的语音进行语音特征识别；

[0017] 语义解析模块，其用于将识别出的语音特征与识别出的图像特征结合起来作为主题或关键词以进行语义解析；

[0018] 对话输出模块，针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。

[0019] 根据本发明的一个实施例，在图像识别模块中，还包括对当前场景图像进行人体身份特征识别的身份识别单元、主题特征识别单元以及发出动作的意图特征识别单元。

[0020] 根据本发明的一个实施例，在语义解析模块中还包括答案搜索接口单元，其用于将所述识别出的语音特征和所述识别出的图像特征作为知识库匹配答案的参考项进行输出。

[0021] 根据本发明的一个实施例，在语义解析模块中还包括对话生成模块接口单元，其用于将所述识别出的语音特征和所述识别出的图像特征作为循环神经网络的对话生成模块的额外输入信息进行输出。

[0022] 根据本发明的一个实施例，所述处理装置还包括动作决策模块，其中所述动作决策模块与所述语义解析模块耦接，以接收所述语义解析模块发出的要指令机器人的其他运动系统做出某动作的信号，并根据该信号作出相应动作的决策。

[0023] 根据本发明的基于视觉和语音的对话交互数据处理方法或者实现该方法的装置，由于采用了视觉识别与语音识别技术的结合对对话场景进行分析，得到对话话题的关键词，同时根据对话生成模型或者搜索答案库的方式提供答案，使得使用者同机器人的聊天可以更加符合用户的意图，机器人对于用户的提问或者给出的话题可以给出适当的回复，而不会出现像现有技术中没有考虑应用场景的错误的语义识别的问题出现。

[0024] 本发明的其它特征和优点将在随后的说明书中阐述，并且，部分地从说明书中变得显而易见，或者通过实施本发明而了解。本发明的目的和其他优点可通过在说明书、权利要求书以及附图中所特别指出的结构来实现和获得。

附图说明

[0025] 附图用来提供对本发明的进一步理解，并且构成说明书的一部分，与本发明的实施例共同用于解释本发明，并不构成对本发明的限制。在附图中：

[0026] 图1是根据本发明的一个实施例的用于结合视觉识别和语音识别同时输出语音和对应动作的对话数据交互处理方法的流程图；

[0027] 图2是根据本发明的一个实施例对所捕获的当前场景图像中的各个特征进行分类的示意图；

[0028] 图3是根据本发明的一个实施例的对话数据交互处理装置的结构框图。

具体实施方式

[0029] 为使本发明的目的、技术方案和优点更加清楚，以下结合附图对本发明实施例作进一步地详细说明。

[0030] 本发明的方法是在智能机器人中实现的。如图1所示，其中显示了根据本发明的一个实施例的用于结合视觉识别和语音识别同时输出语音和对应动作的对话数据交互处理方法的流程图。

[0031] 在该图中，方法开始于步骤S101。在步骤S101中，机器人的听觉系统在监听外界语音的同时，还通过视觉感知系统捕获发出语音时刻对应的当前场景图像。例如，当监听到声

音“好累啊!”时,机器人同时捕获场景图像,例如用户正在球场拿着球的画面,或者用户在书桌上看书的画面。

[0032] 在该场景图像中,包括人的图像、物体图像、背景图像以及所捕获的一系列动作序列帧图像。通过对这些不同的图像进行分类视觉识别,从而判断出所发出语音的准确语义。

[0033] 为了提高视觉识别的准确性,本发明采用卷积神经网络算法进行视觉特征的提取和分析。因此,接下来,在步骤S102中,基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别。

[0034] 卷积网络最初是受视觉神经机制的启发而设计的,是为识别二维形状而设计的一个多层感知器。由于这种网络结构对平移、比例缩放、倾斜或者其它形式的变形具有高度不变性,因此,在图像识别技术领域,卷积网络得到广泛应用。

[0035] 而卷积神经网络是近年发展起来并引起广泛重视的一种高效的图像特征的识别方法。20世纪60年代,Hubel和Wiesel在研究猫脑皮层中用于局部敏感和方向选择的神经元时发现其独特的网络结构可以有效地降低反馈神经网络的复杂性,继而提出了卷积神经网络(Convolutional Neural Networks,简称CNN)。现在,CNN已经成为众多科学领域的研究热点之一。特别是在模式分类领域,由于该网络避免了对图像的复杂前期预处理,可以直接输入原始图像,因而得到了更为广泛的应用。K.Fukushima在1980年提出的新识别机是卷积神经网络的第一个实现网络。随后,更多的科研工作者对该网络进行了改进。其中,具有代表性的研究成果是Alexander和Taylor提出的“改进认知机”,该方法综合了各种改进方法的优点并避免了耗时的误差反向传播。

[0036] 一般地,CNN的基本结构包括两层。其一为特征提取层,每个神经元的输入与前一层的局部接受域相连,并提取该局部的特征。一旦该局部特征被提取后,它与其它特征间的位置关系也随之确定下来。其二是特征映射层,网络的每个计算层由多个特征映射组成,每个特征映射是一个平面,平面上所有神经元的权值相等。特征映射结构采用影响函数核小的sigmoid函数作为卷积网络的激活函数,使得特征映射具有位移不变性。

[0037] 此外,由于一个映射面上的神经元共享权值,因而减少了网络自由参数的个数。卷积神经网络中的每一个卷积层都紧跟着一个用来求局部平均与二次提取的计算层,这种特有的两次特征提取结构减小了特征分辨率。

[0038] CNN主要用来识别位移、缩放及其他形式扭曲不变性的二维图形。由于CNN的特征检测层通过训练数据进行学习,因此在使用CNN时,避免了显示的特征抽取,而隐式地从训练数据中进行学习。再者由于同一特征映射面上的神经元权值相同,所以网络可以并行学习,这也是卷积网络相对于神经元彼此相连网络的一大优势。卷积神经网络以其局部权值共享的特殊结构在语音识别和图像处理方面有着独特的优越性,其布局更接近于实际的生物神经网络,权值共享降低了网络的复杂性,特别是多维输入向量的图像可以直接输入网络这一特点避免了特征提取和分类过程中数据重建的复杂度。

[0039] 基于卷积神经网络的一个变形是神经认知机,其将一个视觉模式分解成许多子模式(特征),然后进入分层递阶式相连的特征平面进行处理。它试图将视觉系统模型化,使其能够在即使物体有位移或轻微变形的时候,也能完成识别。神经认知机能够利用位移恒定能力从激励模式中学习,并且可识别这些模式的变化形。在其后的应用研究中,Fukushima将神经认知机主要用于手写数字的识别。随后,国内外的研究人员提出多种卷积神经网络

形式，在邮政编码识别(Y.LeCun etc)、车牌识别和人脸识别等方面得到了广泛的应用。

[0040] 本发明利用了上述技术对所采集到的图像信息进行特征识别，例如识别出场景图像中针对人图像的身份特征、物体图像的主题特征、人发出一系列动作的意图特征以及背景图像中的其它特征等(如图2所示)，以便进行接下来的交互决策。

[0041] 继续参照图1，在步骤S103中，对监听到的与当前场景图像对应的语音进行语音特征识别。

[0042] 在步骤S104中，将上述识别出的语音特征和识别出的图像特征进行结合，并作为主题或关键词进行语义解析。

[0043] 例如，当机器人听到主人说“好累啊!”此时机器人捕捉画面，可能的画面是“主人拿着一个篮球”，也可能是“主人正在看书”。针对于这两种场景，图像识别模块分别识别到了“篮球”这个物体特征和“书”这个物体特征。背景也许分别是草场或者书房的特征。

[0044] 刚才提到说是主人，说明在此之前，机器人已经先进行了人体的身份特征识别，并识别出发出语音的对象就是主人，或者主人的朋友。对于身份特征识别，机器人需要进行精确的人面部特征的识别，对于画面中出现的三角形感兴趣区域采用卷积神经网络方法进行识别。因为在捕获场景图像时，人的图像尤其是人脸的特征因为角度的问题会发生变化。

[0045] 因此这些特征在机器人语义理解方面是非常有帮助的，虽然听到的是同一句话，但是假如没有这些视觉特征的话，机器人可能回答的答案是一样。但是我们知道，打篮球的累和看书学习的累显然有不同的处理方法，机器人要提供不同的回答。

[0046] 在步骤S104中，将识别出的语音特征与识别出的图像特征结合起来作为主题或关键词以进行语义解析。如上所述，机器人仅靠语音识别，而不考虑具体场景图像，很可能理解的意思是错误的或者说不恰当的，因此这样的交流没有意义。本发明在语义解析时，还基于视觉识别特征。通过将语音特征与图像识别技术获得的人体身份特征、背景图像中的主题特征以及发出动作的意图特征进行结合，这样获得的主题或关键词作为语义解析的基础。例如通过对话生成模块作为其额外输入信息，从而产生针对用户对话的准确的答案。或者，在语义解析后，可以将所识别出的语音特征和所识别出的图像特征作为知识库匹配答案的参考项进行输出。

[0047] 在步骤S105中，针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。

[0048] 语音回答的答案可例如通过上述的知识库匹配答案的方式获得，也可以通过采用循环神经网络的对话生成模块来获得。

[0049] 然而对应的动作表达需要动作决策模块来根据语义回答和情绪等作出相应的动作。因此，在语义解析之后，还需要发出命令机器人的其他运动系统作出某动作的信号。例如针对“你好”的语音发出自动作出伸出手来进行握手的动作。或者，根据当时的回答做出摇头的动作等等。

[0050] 针对所捕获的当前场景图像，如图2所示，其可以进行进一步的分类。例如针对人体图像，需要识别面部特征、表情特征、衣服颜色特征、样式特征、发型等等。而对于物体图像，需要识别出是属于哪一类物体如文具、玩具、宠物等等。对于背景图像，需要识别出环境特征，天气、地理位置等等。而对于捕获的一系列动作，需要识别出人发出这些动作与语音配合的意图特征等。当然，场景图像中还包括了其他特征，这里不一一赘述。

[0051] 由于本发明的方法描述的是在计算机系统中实现的。该计算机系统例如可以设置在机器人的控制核心处理器中。例如，本文所述的方法可以实现为能以控制逻辑来执行的软件，其由机器人控制系统中的CPU来执行。本文所述的功能可以实现为存储在非暂时性有形计算机可读介质中的程序指令集合。当以这种方式实现时，该计算机程序包括一组指令，当该组指令由计算机运行时其促使计算机执行能实施上述功能的方法。可编程逻辑可以暂时或永久地安装在非暂时性有形计算机可读介质中，例如只读存储器芯片、计算机存储器、磁盘或其他存储介质。除了以软件来实现之外，本文所述的逻辑可利用分立部件、集成电路、与可编程逻辑设备(诸如，现场可编程门阵列(FPGA)或微处理器)结合使用的可编程逻辑，或者包括它们任意组合的任何其他设备来体现。所有此类实施例旨在落入本发明的范围之内。

[0052] 因此，根据本发明的另一个方面，还提供了一种基于视觉和语音的对话数据交互处理装置600。如图3所示，该对话数据交互处理装置600包括：监听与捕获模块601、图像识别模块602、语音识别模块603、语义解析模块605、对话输出模块606。

[0053] 其中，监听与捕获模块601用于在监听语音的同时，捕获对应的当前场景图像。如图所示，其与交互层通信，通过交互层中的视觉系统和听觉系统来监听语音和捕获图像。

[0054] 图像识别模块602，其用于基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别。而语音识别模块603，其用于对监听到的与当前场景图像对应的语音进行语音特征识别。

[0055] 在图像识别模块602中，其还包括身份识别单元、主题识别单元、意图识别单元。通过这些单元，图像识别模块可以将捕获的场景图像中的各个要素进行逐一识别。例如，针对人的图像，通过识别面部特征、表情特征、衣服颜色特征、样式特征、发型等等。而对于物体图像，需要识别出是属于哪一类物体如文具、玩具、宠物等等。对于背景图像，需要识别出环境特征，天气、地理位置等等。而对于捕获的一系列动作，需要识别出人发出这些动作与语音配合的意图特征等。

[0056] 在语义解析模块605中，该模块用于将识别出的语音特征与识别出的图像特征结合起来作为主题或关键词以进行语义解析。在一个实施例中，为了将结合了语音特征和视觉特征得到的对话的主题或关键词进行对话的应用，可以通过设置答案搜索接口来把主题或关键词作为知识库的搜索输入项进行答案的搜索，或者通过对话生成接口将主题或关键词作为循环网络的对话生成模块的附加输入项来进行对话答案的生成。

[0057] 对话输出模块606针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达。输出语音回答就是通过音频处理系统将要输出的例如文本形式的回答转换成可以通过麦克风播放的语音。进行对应的动作表达，需要调用机器人的运动决策模块，通过该模块根据回答做出相应动作的决策，并通过执行机构运动相应部件。

[0058] 因此，所述处理装置600还包括动作决策模块，其中所述动作决策模块与所述语义解析模块耦接，以接收所述语义解析模块发出的命令机器人的其他运动系统做出某动作的信号，并根据该信号作出相应动作的决策。

[0059] 应该理解的是，本发明所公开的实施例不限于这里所公开的特定结构、处理步骤或材料，而应当延伸到相关领域的普通技术人员所理解的这些特征的等同替代。还应当理解的是，在此使用的术语仅用于描述特定实施例的目的，而并不意味着限制。

[0060] 说明书中提到的“一个实施例”或“实施例”意指结合实施例描述的特定特征、结构或特性包括在本发明的至少一个实施例中。因此，说明书通篇各个地方出现的短语“一个实施例”或“实施例”并不一定均指同一个实施例。

[0061] 虽然本发明所公开的实施方式如上，但所述的内容只是为了便于理解本发明而采用的实施方式，并非用以限定本发明。任何本发明所属技术领域内的技术人员，在不脱离本发明所公开的精神和范围的前提下，可以在实施的形式上及细节上作任何的修改与变化，但本发明的专利保护范围，仍须以所附的权利要求书所界定的范围为准。

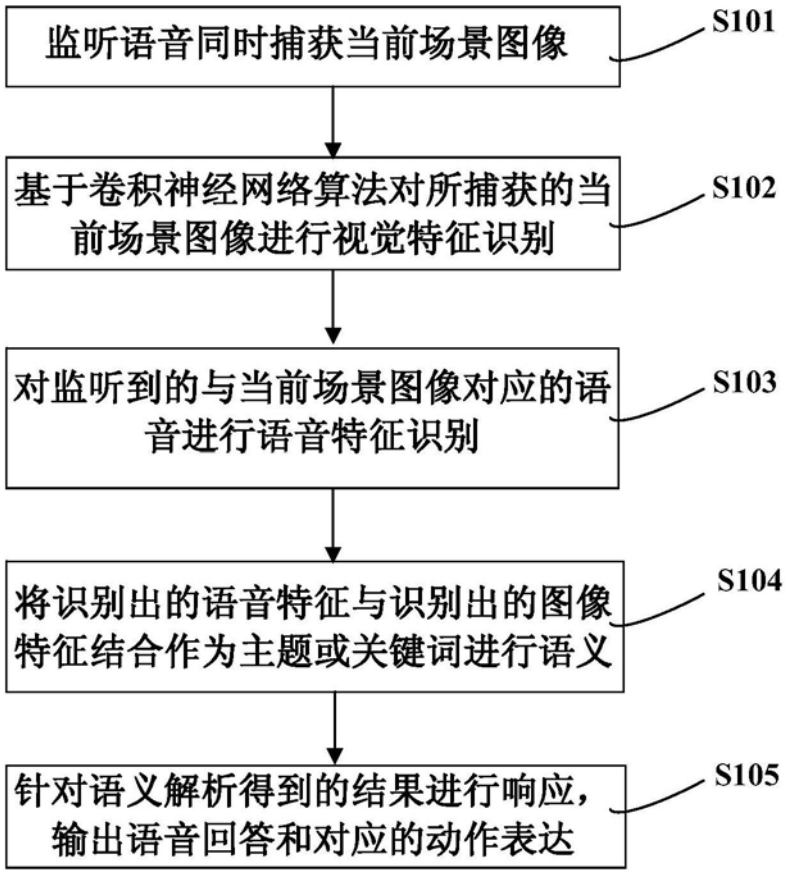

监听语音同时捕获当前场景图像
S101
基于卷积神经网络算法对所捕获的当前场景图像进行视觉特征识别
S102
对监听到的与当前场景图像对应的语音进行语音特征识别
S103
将识别出的语音特征与识别出的图像特征结合作为主题或关键词进行语义
S104
针对语义解析得到的结果进行响应，输出语音回答和对应的动作表达
S105

图1

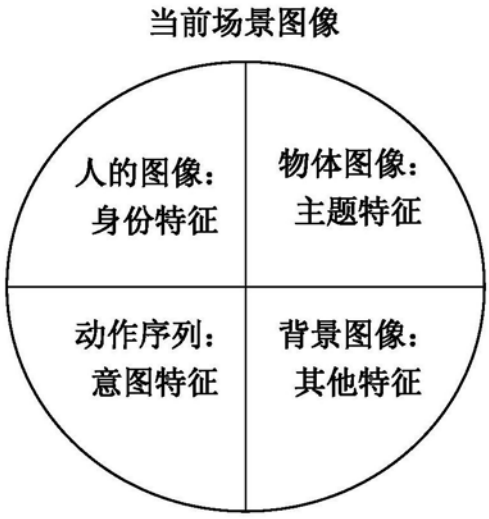

当前场景图像
人的图像：
身份特征
物体图像：
主题特征
动作序列：
意图特征
背景图像：
其他特征

图2

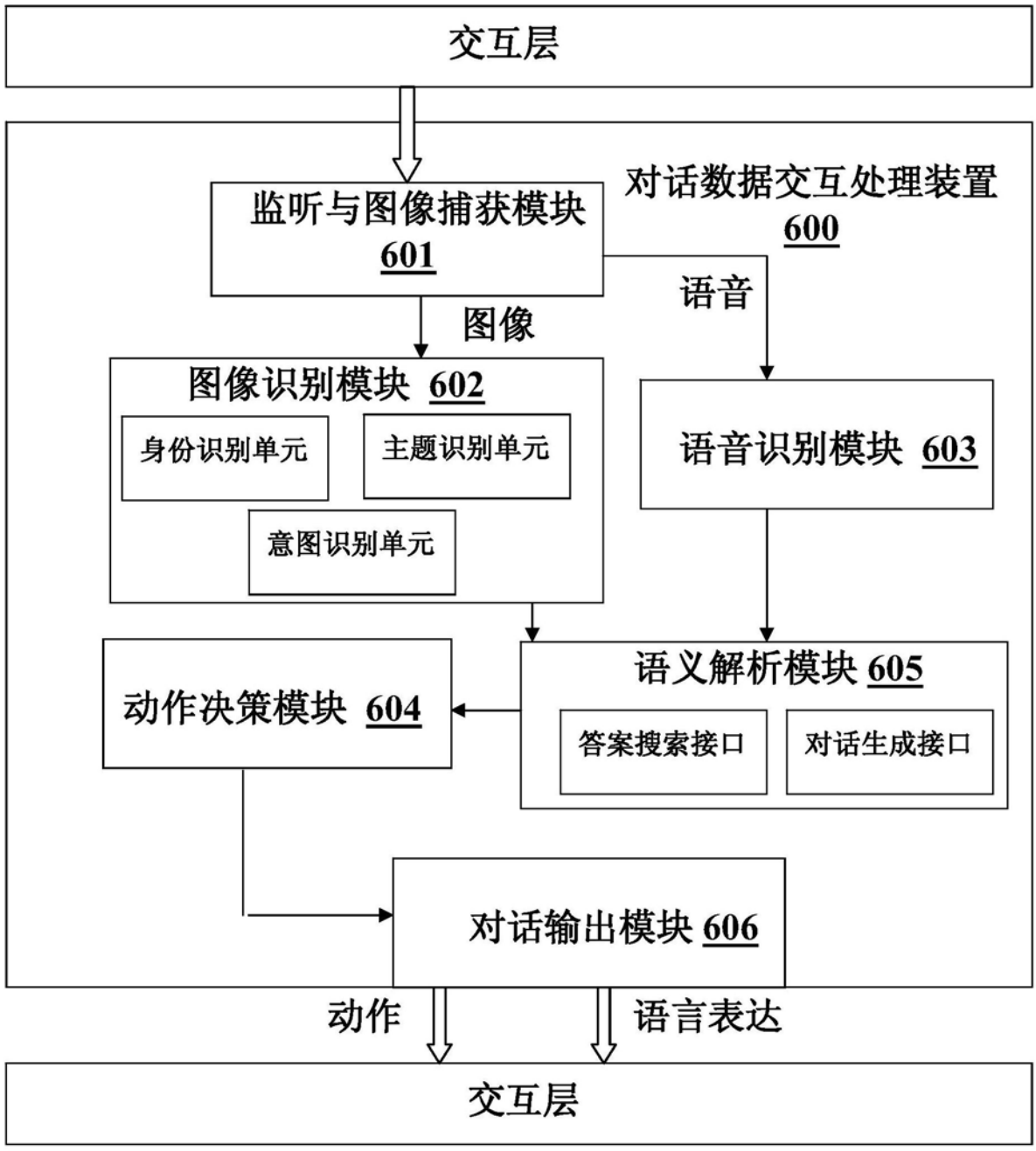

交互层
对话数据交互处理装置
600
监听与图像捕获模块
601
语音
图像
图像识别模块 602
身份识别单元
主题识别单元
意图识别单元
语音识别模块 603
动作决策模块 604
语义解析模块 605
答案搜索接口
对话生成接口
对话输出模块 606
动作
语言表达
交互层

图3