

人机交互方法、装置、系统、电子设备以及计算机介质

相关申请的交叉引用

本申请要求于 2021 年 2 月 9 日提交的申请号为 202110174149.1、发明名称为“人机交互方法、装置、系统、电子设备以及计算机介质”的中国专利申请的优先权，其全部内容通过引用结合在本申请中。

技术领域

本公开涉及人工智能技术领域，具体涉及计算机视觉、深度学习等技术领域，尤其涉及人机交互方法、装置、电子设备、计算机可读介质以及计算机程序产品。

背景技术

传统的虚拟数字人客服系统仅仅能完成简单的人机交互，可以理解成为是一个没有感情的机器人，只是做到简单的语音识别与语义理解，在较为复杂的柜台客服系统中，仅通过简单的语音识别和语义理解无法针对各种不同情绪的用户做出情绪反应，用户交互体验较差。

发明内容

本公开的实施例提出了人机交互方法、装置、电子设备、计算机可读介质以及计算机程序产品。

第一方面，本公开的实施例提供了一种人机交互方法，该方法包括：接收用户的至少一种模态的信息；基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；基于意图信息，确定对用户的答复信息；基于用户情绪特征，选定向用户反馈的人物情绪特征；基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。

在一些实施例中，上述至少一种模态的信息包括用户的图像数据以及音频数据，上述基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征，包括：基于用户的图像数据，识别用户的表情特征；由音频数据，得到文本信息；基于文本信息，提取用户的意图信息；基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征，还包括：用户情绪特征还由文本信息得到。

在一些实施例中，上述基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征，包括：将音频数据输入已训练完成的语音情绪识别模型，得到语音情绪识别模型输出的语音情绪特征；将表情特征输入已训练完成的表情情绪识别模型，得到表情情绪识别模型输出的表情情绪特征；对语音情绪特征、表情情绪特征加权求和，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述至少一种模态的信息包括用户的图像数据以及文本数据，上述基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征，包括：基于用户的图像数据，识别用户的表情特征；基于文本数据，提取用户的意图信息；基于文本数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频，包括：基于答复信息、人物情绪特征，生成答复音频；基于答复音频、人物情绪特征以及预先建立的动画人物形象模型，得到与人物情绪特征对应的动画人物形象的播报视频。

在一些实施例中，上述基于答复音频、人物情绪特征以及预先建立的动画人物形象模型，得到与人物情绪特征对应的动画人物形象的播报视频，包括：将答复音频、人物情绪特征输入已训练完成的口型驱动模型，得到口型驱动模型输出的口型数据；将答复音频、人物情绪特征输入已训练完成的表情驱动模型，得到表情驱动模型输出的表情数据；基于口型数据、表情数据对动画人物形象模型进行驱动，得到三维模型动作序列，对三维模型动作序列进行渲染，得到视频帧图片序列；合成视频帧图片序列，得到与人物情绪特征对应的动画人物形象的播报视频，其中，口型驱动模型、表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到。

第二方面，本公开的实施例提供了一种人机交互装置，该装置包括：接收单元，被配置成接收用户的至少一种模态的信息；识别单元，被配置成基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；确定单元，被配置成基于意图信息，确定对用户的答复信息；选定单元，被配置成基于用户情绪特征，选定向用户反馈的人物情绪特征；播报

单元，被配置成基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。

在一些实施例中，上述至少一种模态的信息包括用户的图像数据以及音频数据，上述识别单元包括：识别子单元，被配置成基于用户的图像数据，识别用户的表情特征；文本得到子单元，被配置成由音频数据，得到文本信息；提取子单元，被配置成基于文本信息，提取用户的意图信息；特征得到子单元，被配置成基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述识别单元中的用户情绪特征进一步地还由文本信息得到。

在一些实施例中，上述特征得到子单元包括：语音得到模块，被配置成将音频数据输入已训练完成的语音情绪识别模型，得到语音情绪识别模型输出的语音情绪特征；表情得到模块，被配置成将表情特征输入已训练完成的表情情绪识别模型，得到表情情绪识别模型输出的表情情绪特征；求和模块，被配置成对语音情绪特征、表情情绪特征加权求和，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述至少一种模态的信息包括：用户的图像数据以及文本数据；上述识别单元包括：识别模块，被配置成基于用户的图像数据，识别用户的表情特征；提取模块，被配置成基于文本数据，提取用户的意图信息；特征得到模块，被配置成基于文本数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述播报单元包括：生成子单元，被配置成播报单元；视频得到子单元，被配置成基于答复音频、人物情绪特征以及预先建立的动画人物形象模型，得到与人物情绪特征对应的动画人物形象的播报视频。

在一些实施例中，上述视频得到子单元包括：口型驱动模块，被配置成将答复音频、人物情绪特征输入已训练完成的口型驱动模型，得到口型驱动模型输出的口型数据；表情驱动模块，被配置成将答复音频、人物情绪特征输入已训练完成的表情驱动模型，得到表情驱动模型输出的表情数据；模型驱动模块，被配置成基于口型数据、表情数据对动画人物形象模型进行驱动，得到三维模型动作序列；图片得到模块，被配置成对三维模型动作序列进行渲染，得到视频帧图片序列；视频得到模块，被配置成合成视频帧图片序列，

得到与人物情绪特征对应的动画人物形象的播报视频。口型驱动模型、表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到。

第三方面，本公开的实施例提供了一种人机交互系统，该系统包括：采集设备、显示设备以及分别与采集设备、显示设备连接的交互平台；采集设备用于采集用户的至少一种模态的信息；交互平台用于接收用户的至少一种模态的信息；基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；基于意图信息，确定对用户的答复信息；基于用户情绪特征，选定向用户反馈的人物情绪特征；基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频；显示设备用于接收并播放播报视频。

第四方面，本公开的实施例提供了一种电子设备，该电子设备包括：一个或多个处理器；存储装置，其上存储有一个或多个程序；当一个或多个程序被一个或多个处理器执行时，使得一个或多个处理器实现如第一方面中任一实现方式描述的方法。

第五方面，本公开的实施例提供了一种计算机可读介质，其上存储有计算机程序，该程序被处理器执行时实现如第一方面中任一实现方式描述的方法。

第六方面，本公开的实施例提供了一种计算机程序产品，包括计算机程序，计算机程序在被处理器执行时实现如第一方面任一实现方式描述的方法。

根据本公开的实施例提供的人机交互方法和装置：首先，接收用户的至少一种模态的信息；其次，基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；再次，基于意图信息，确定对用户的答复信息；从次，基于用户情绪特征，选定向用户反馈的人物情绪特征；最后，基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。由此，通过对用户的至少一种模态的信息进行分析确定反馈用户的人物情绪特征，为不同情绪的用户提供了有效的情绪反馈，保证了人机交互过程中的感情交流。

附图说明

通过阅读参照以下附图所作的对非限制性实施例所作的详细描述，本公

开的其它特征、目的和优点将会变得更明显：

图 1 是本公开的一个实施例可以应用于其中的示例性系统架构图；

图 2 是根据本公开的人机交互方法的一个实施例的流程图；

图 3 是本公开的识别用户的意图信息以及用户情绪特征的一个实施例的流程图；

图 4 是根据本公开的人机交互装置的实施例的结构示意图；

图 5 是根据本公开的人机交互系统的实施例的结构示意图；

图 6 是适于用来实现本公开的实施例的电子设备的结构示意图。

具体实施方式

下面结合附图和实施例对本公开作进一步的详细说明。可以理解的是，此处所描述的具体实施例仅仅用于解释相关发明，而非对该发明的限定。另外还需要说明的是，为了便于描述，附图中仅示出了与有关发明相关的部分。

需要说明的是，在不冲突的情况下，本公开中的实施例及实施例中的特征可以相互组合。下面将参考附图并结合实施例来详细说明本公开。

图 1 示出了可以应用本公开的人机交互方法的示例性系统架构 100。

如图 1 所示，系统架构 100 可以包括终端设备 101、102、自动柜员机 103，网络 104 和服务器 105。网络 104 用以在终端设备 101、102、自动柜员机 103 和服务器 105 之间提供通信链路的介质。网络 104 可以包括各种连接类型，通常可以包括无线通信链路等等。

终端设备 101、102、自动柜员机 103 通过网络 104 与服务器 105 交互，以接收或发送消息等。终端设备 101、102、自动柜员机 103 上可以安装有各种通讯客户端应用，例如即时通信工具、邮箱客户端等。

终端设备 101、102 可以是硬件，也可以是软件；当终端设备 101、102 为硬件时，可以是具有通信和控制功能的用户设备，上述用户设备可与服务器 105 进行通信。当终端设备 101、102 为软件时，可以安装在上述用户设备中；终端设备 101、102 可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

服务器 105 可以是提供各种服务的服务器，例如为终端设备 101、102、自动柜员机 103 上客户问答系统提供支持的后台服务器。后台服务器可以对

终端设备 101、102、自动柜员机 103 上采集的相关用户的至少一种模态的信息进行分析处理，并将处理结果（如动画人物形象的播报视频）反馈给终端设备或自动柜员机。

需要说明的是，服务器可以是硬件，也可以是软件。当服务器为硬件时，可以实现成多个服务器组成的分布式服务器集群，也可以实现成单个服务器。当服务器为软件时，可以实现成多个软件或软件模块（例如用来提供分布式服务的软件或软件模块），也可以实现成单个软件或软件模块。在此不做具体限定。

需要说明的是，本公开的实施例所提供的人机交互方法一般由服务器 105 执行。

应该理解，图 1 中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要，可以具有任意数目的终端设备、网络和服务器。

如图 2，示出了根据本公开的人机交互方法的一个实施例的流程 200，该人机交互方法包括以下步骤：

步骤 201，接收用户的至少一种模态的信息。

本实施例中，人机交互方法运行其上的执行主体可以在同一时段接收用户的不同来源的信息。不同来源的信息即为不同模态的信息，不同来源的信息为多个时称为至少一种模态的信息。具体地，至少一种模态的信息可以包括：图像数据、音频数据、文本数据中的一种或多种。

本实施例中，用户的至少一种模态的信息是由用户发出的信息或/和与用户有关的信息。比如，图像数据是对用户的人脸、用户的肢体、用户头发进行拍摄得到的图像数据等，音频数据是对用户发出的语音进行录制后得到的音频数据，文本数据是用户向执行主体输入的文字、符号、数字等数据。通过用户的至少一种模态的信息，可以对用户的意图进行分析，确定用户的问题、目的以及用户发问或者进行信息输入时的情绪状态等。

实践中，不同模态的信息可以是不同传感器采集的对同一事物的描述信息。比如，视频检索时，不同模态的信息包括同一时段采集的同一用户的音频数据和图像数据，其中的音频数据和图像数据在同一时刻相互对应。再如任务型对话交流过程，用户通过用户终端向执行主体发送的同一时段同一用户的图像数据、文本数据等。

本实施例中，人机交互方法的执行主体（如图 1 所示的服务器 105）可以通过多种手段接收用户的至少一种模态的信息。比如，实时从用户终端（如图 1 所示的终端设备 101、102、自动柜员机 103）采集待处理数据集，并从待处理数据集中提取至少一种模态的信息。或者，从本地内存获取包含多种模态的信息的待处理数据集，并从待处理数据集中提取至少一种模态的信息。可选地，上述至少一种模态的信息还可以是终端实时发送的信息。

步骤 202，基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征。

本实施例中，用户的意图信息是表征用户的问题、目的、寒暄等内容的信息。执行主体在得到用户的意图信息之后，可以基于意图信息的内容的不同，做出不同的反馈。

用户情绪特征是用户在发出或展示出不同模态的信息时个人的情绪状态，具体地，情绪状态包括：愤怒、悲哀、高兴、生气、厌恶等。

进一步地，基于用户的不同模态的信息可以有不同的识别用户的意图信息以及用户情绪特征的方式。

在本公开的一些可选实现方式中，至少一种模态的信息包括用户的图像数据以及音频数据，上述基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征，包括：基于用户的图像数据，识别用户的表情特征；由音频数据，得到文本信息；基于文本信息，提取用户的意图信息；基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征。

本可选实现方式中，在用户的至少一种模态的信息包括用户的图像数据以及音频数据时，基于用户的图像数据，识别用户的表情特征；基于音频数据，得到文本信息；基于文本信息，提取意图信息；基于音频数据以及表情特征，得到用户情绪特征。由此，基于用户的面部表情（表情特征）以及声音（音频数据）两方面所包含的情绪，综合确定用户的情绪，一定程度上提高了分析用户情绪特征的可靠性。

在本公开的一些可选实现方式中，至少一种模态的信息包括：用户的图像数据以及文本数据，上述基于至少一种模态的信息，识别用户的意图信息以及用户情绪特征的方法包括以下步骤：基于用户的图像数据，识别用户的表情特征；基于文本数据，提取用户的意图信息；基于文本数据以及表情特征，得到与意图信息对应的用户情绪特征。

本可选实现方式提供的识别用户的意图信息以及用户情绪特征的方法，当用户的模态信息包括图像数据以及文本数据时：基于图像数据，识别用户的表情特征；基于文本数据，提取意图信息；进一步基于文本数据以及表情特征，得到用户情绪特征。由此，基于用户的面部表情（表情特征）以及语言（文本信息）两方面所包含的情绪，综合确定用户的情绪，为聋哑人士的意图信息和情绪的提取提供了可靠地情绪分析方式。

可选地，至少一种模态的信息包括：用户的图像数据、文本数据以及音频数据。上述基于至少一种模态的信息，识别用户的意图信息以及用户情绪特征的方法包括以下步骤：基于用户的图像数据，识别用户的表情特征；基于文本数据以及音频数据提取用户的意图信息；基于文本数据、表情特征、音频数据，得到与意图信息对应的用户情绪特征。

本可选实现方式中，在至少一种模态的信息包括用户的图像数据、文本数据以及音频数据三者时，可以通过用户的面部表情（表情特征）、声音（音频数据）以及语言（文本信息）三方面所包含的情绪，综合确定用户的情绪，提高了用户情绪分析的可靠性。

本实施例中含有的文本信息以及文本数据均是文本的不同表现形式，采用文本信息和文本数据仅是用于区分文本的来源或者处理方式不同。

进一步地，由于用户的语言、文字以及表情均可以反映出用户的情绪，所以能够得到用户情绪特征。在本实施例的一些可选实现方式中，上述基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征，包括：

将音频数据输入已训练完成的语音情绪识别模型，得到语音情绪识别模型输出的语音情绪特征；将表情特征输入已训练完成的表情情绪识别模型，得到表情情绪识别模型输出的表情情绪特征；对语音情绪特征、表情情绪特征加权求和，得到与意图信息对应的用户情绪特征。

本可选实现方式中，通过训练完成的表情情绪识别模型、语音情绪识别模型，分别识别表情情绪特征、语音情绪特征，从而从用户的至少一种模态的信息快速得到了用户的实时的情绪状态，为实现有感情的人物形象提供了可靠的基础。

可选地，上述基于文本数据、表情特征、音频数据，得到与意图信息对应的用户情绪特征还可以包括：将文本数据输入已训练完成的文本情绪识别模型，得到文本情绪识别模型输出的文本情绪特征；将音频数据输入已训练

完成的语音情绪识别模型，得到语音情绪识别模型输出的语音情绪特征；将表情特征输入已训练完成的表情情绪识别模型，得到表情情绪识别模型输出的表情情绪特征；对文本情绪特征、语音情绪特征、表情情绪特征加权求和，得到与意图信息对应的用户情绪特征。

本实施例中，上述语音情绪识别模型用于识别用户的音频数据中的情绪特征，以确定用户在发出语音时的情绪状态；上述表情情绪识别模型用于识别用户的表情特征中与情绪相关的表情特征，以确定用户在表达某种表情时的情绪状态；上述文本情绪识别模型用于识别用户的文本数据中的情绪特征，以确定用户输出的文本所表达的情绪状态。

上述表情情绪识别模型、语音情绪识别模型、文本情绪识别模型可以是在给定同一个用户的大量的、标注完成的文本数据、表情特征、音频数据基础上，训练出来的模型，而得到的语音情绪特征、表情情绪特征以及文本情绪特征均是用于表征用户的情绪状态（喜、怒、哀、惧）。需要说明的是，本可选实现方式中的语音情绪识别模型、表情情绪识别模型也可以适用于其他实施例。

步骤 203，基于意图信息，确定对用户的答复信息。

本实施例中，用户的答复信息是与用户的意图信息相对应的信息，而答复信息也是需要动画人物形象需要播报的音频内容。例如，用户意图信息是一个问题：李四有多高？而答复信息就是一个答案：李四身高 1.8 米。

在得到用户的意图信息之后，执行主体可以通过多种途径确定答复信息，比如，通过查询知识库、搜索知识图谱等。

步骤 204，基于用户情绪特征，选定向用户反馈的人物情绪特征。

本实施例中，人物情绪特征表征动画人物形象的情绪状态的特征，其中，人物情绪状态可以是与用户情绪特征所表征的情绪状态相同，也可以是与用户情绪特征所表征的情绪状态不同。例如，用户情绪特征为发怒时，人物情绪特征可以表现为安抚；用户情绪特征为高兴时，人物情绪特征也可以同样表现为高兴。

人机交互方法运行于其上的执行主体，在得到用户情绪特征之后，可以基于用户情绪特征，从预设的情绪特征库中选取一个或多个情绪特征作为人物情绪特征。该人物情绪特征应用于动画人物形象中，实现动画人物形象的情绪特征的体现。

步骤 205, 基于人物情绪特征与答复信息, 生成与人物情绪特征对应的动画人物形象的播报视频。

本实施例中, 动画人物形象的播报视频是虚拟的动画人物播报信息的视频, 人物情绪特征与答复信息均是动画人物形象需要表现出来的信息。为了生动、直观地表现答复信息, 可以将答复信息转换为答复音频。通过动画人物形象的播报视频中动画人物虚拟的张口动作体现播报答复音频。通过动画人物虚拟的表情变化体现人物情绪特征。

在动画人物形象与用户进行沟通的过程中, 根据人物情绪特征, 可以使动画人物形象的语音合成的音频中带有人物情绪信息, 比如安抚情绪。同时, 还可以选择与人物情绪特征对应的面部表情呈现在动画人物形象的面部, 提高了动画人物形象的表情的丰富度。

为了使答复音频更加生动, 在本实施例的一些可选实现方式中, 基于人物情绪特征与答复信息, 生成与人物情绪特征对应的动画人物形象的播报视频, 包括: 基于答复信息、人物情绪特征, 生成答复音频; 基于答复音频、人物情绪特征以及预先建立的动画人物形象模型, 得到与人物情绪特征对应的动画人物形象的播报视频。

本可选实现方式中, 动画人物形象模型可以是通过三维形象建模得到三维模型, 其中, 三维形象建模是利用三维制作软件通过虚拟三维空间构建出具有三维数据的模型的过程。进一步地, 还可以针对动画人物形象的各个部位进行建模(比如, 脸部轮廓建模、嘴部独立建模、头发独立建模、躯干独立建模、骨骼独立建模、面部表情建模等), 组合选取的各个部位的模型得到动画人物形象模型。

本可选实现方式中, 基于答复信息、人物情绪特征生成答复音频包含的预先分析的人物情绪因素, 使得到的动画人物形象的播报视频中的音频更加富含感情, 从而感染用户; 基于人物情绪特征得到的动画人物形象的播报视频中的动画人物动作更加富含感情, 具有情感感染力。

在本实施例的一些可选实现方式中, 上述基于答复音频、人物情绪特征以及预先建立的动画人物形象模型, 得到与人物情绪特征对应的动画人物形象的播报视频, 包括: 将答复音频、人物情绪特征输入已训练完成的口型驱动模型, 得到口型驱动模型输出的口型数据; 将答复音频、人物情绪特征输入已训练完成的表情驱动模型, 得到表情驱动模型输出的表情数据; 基于口

型数据、表情数据对动画人物形象模型进行驱动，得到三维模型动作序列；对三维模型动作序列进行渲染，得到视频帧图片序列；合成视频帧图片序列，得到与人物情绪特征对应的动画人物形象的播报视频。口型驱动模型、表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到。

本可选实现方式中，口型驱动模型是用于识别动画人物的嘴唇在三维空间中的运行轨迹的模型，并且口型驱动模型还可以与口型库结合，得到动画人物形象的在不同时刻的口型数据，口型数据也是动画人物形象口型变化的数据。

本可选实现方式中，表情驱动模型是用于识别动画人物的面部特征点在三维空间中的运行轨迹的模型，并且表情驱动模型还可以与表情库结合，得到动画人物形象的在不同时刻的表情数据，表情数据也是动画人物形象表情变化的数据。

本可选实现方式中，由于口型驱动模型、表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到，从而使得到的动画人物形象的嘴型和声音更加贴合、统一、无违和感，使播报视频中的动画人物更加生动、形象。

可选地，还可以采用语音动画合成（STA，Speech-to-Animation）模型，直接实现与人物情绪对应的动画人物形象的播报视频。语音动画合成模型可以是多种不同类型的模型（虚拟形象模型、语音合成模型等）统一训练得到，其结合人工智能与计算机图形学，能实时解算语音对应的发音口型，并精细驱动动画人物形象面部表情，实现动画的音画同步呈现。

语音动画合成模型训练中涉及的数据主要包括形象数据、声音数据与文本数据。三种数据存在一定的交集，即，用于训练形象的视频数据中的音频、用于训练语音识别的音频数据与用于训练语音合成的音频数据是一致的。用于训练语音识别的音频数据对应的文本数据与用于训练形象的音频对应的文本数据是一致的。这些一致性是为了提升语音动画合成模型训练过程中的准确性，除此以外还需要有人工标注的数据：形象的表情、情绪特征。

语音动画合成模型包括：虚拟形象模型、语音合成模型。虚拟形象的模型建模除形象基本的面部及面部轮廓、五官、躯干等基本静态模型外，还有针对形象的口型、表情、动作等动态模型。语音合成模型除了最基本的音色

模型外，还融入了人物情绪特征。

根据本公开的实施例提供的人机交互方法：首先，接收用户的至少一种模态的信息；其次，基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；再次，基于意图信息，确定对用户的答复信息；从次，基于用户情绪特征，选定向用户反馈的人物情绪特征；最后，基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。由此，通过对用户的至少一种模态的信息进行分析确定动画人物形象的人物情绪特征，为不同情绪的用户提供了有效的情绪反馈，保证了人机交互过程中的感情交流。

在本公开的另一实施例中，至少一种模态的信息包括用户的图像数据以及音频数据。如图 3，示出了本公开的识别用户的意图信息以及用户情绪特征的方法的一个实施例的流程 300，该方法包括以下步骤：

步骤 301，基于用户的图像数据，识别用户的表情特征。

本实施例中，表情特征识别是指对人脸的器官特征、纹理区域和预定义的特征点进行定位和提取。表情特征识别还是人脸表情识别中的核心步骤，也是人脸识别的关键，它决定着最终的人脸识别结果，直接影响识别率的高低。

本可选实现方式中，人脸的表情也属于一种肢体语言，通过人脸的表情可以反映用户的情绪，每个用户情绪特征均具有与其相对应的表情。

用户的图像数据包括人脸图像数据，通过对人脸图像数据进行分析，确定用户的表情特征。

可选地，用户的图像数据还可以包括用户的肢体图像数据，通过对肢体图像数据进行分析，还可以更加明确用户的表情特征。

步骤 302，由音频数据，得到文本信息。

本实施例中，可以通过成熟的音频识别模型，得到文本信息。例如采用 ASR（Automatic Speech Recognition，语音识别）模型，ASR 模型可以将声音转化为文字。将音频数据输入 ASR 模型，可以得到 ASR 模型输出的文字，从而达到识别文本信息的目的。

步骤 303，基于文本信息，提取用户的意图信息。

本可选实现方式中，文本信息是将用户的音频数据转换为文本后的信息。

通过成熟的意图识别模型得到意图信息，例如，采用 NLU（Natural Language Understanding，自然语言理解）模型对文本信息进行句子检测、分词、词性标注、句法分析、文本分类/聚类、信息抽取等处理对文本信息进行语义分析，确定用户的意图信息。

步骤 304，基于音频数据、文本信息以及表情特征，得到与意图信息对应的用户情绪特征。

本可选实现方式中，在判断用户情绪特征时，可以从用户的音频数据（语气）以及用户的表情特征结合音频模型识别出的文本信息，协同判断出用户情绪特征。这比仅根据用户表情或者仅根据用户声音信息判断用户表情更加准确，从而便于选出更加适合的答复信息和人物情绪特征应用于动画人物形象，并通过动画人物形象与用户进行沟通。

本实施例提供的识别用户的意图信息以及用户情绪特征的方法，在用户的模态信息包括图像数据以及音频数据时：基于图像数据，识别用户的表情特征；基于音频数据，得到文本信息；基于文本信息，提取意图信息；进一步基于音频数据、文本信息以及表情特征，得到用户情绪特征。由此，基于用户的面部表情（表情特征）、声音（音频数据）以及语言（文本信息）三方面所包含的情绪，综合确定用户的情绪，提高了分析用户情绪特征的可靠性。

进一步参考图 4，作为对上述各图所示方法的实现，本公开提供了人机交互装置的一个实施例，该装置实施例与图 2 所示的方法实施例相对应，该装置具体可以应用于各种电子设备中。

如图 4 所示，本公开的实施例提供了一种人机交互装置 400，该装置 400 包括：接收单元 401、识别单元 402、确定单元 403、选定单元 404、播报单元 405。其中，接收单元 401，可以被配置成接收用户的至少一种模态的信息。识别单元 402，可以被配置成基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征。确定单元 403，可以被配置成基于意图信息，确定对用户的答复信息。选定单元 404，可以被配置成基于用户情绪特征，选定向用户反馈的人物情绪特征；播报单元 405，可以被配置成基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。

在本实施例中，人机交互装置 400 中，接收单元 401、识别单元 402、确

定单元 403、选定单元 404、播报单元 405 的具体处理及其所带来的技术效果可分别参考图 2 对应实施例中的步骤 201、步骤 202、步骤 203、步骤 204、步骤 205。

在一些实施例中，上述至少一种模态的信息包括用户的图像数据以及音频数据。上述识别单元 402 包括：识别子单元（图中未示出）、文本得到子单元（图中未示出）、提取子单元（图中未示出）、特征得到子单元（图中未示出）。其中，识别子单元，可以被配置成基于用户的图像数据，识别用户的表情特征。文本得到子单元，可以被配置成由音频数据，得到文本信息。提取子单元，可以被配置成基于文本信息，提取用户的意图信息。特征得到子单元，可以被配置成基于音频数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述识别单元中的用户情绪特征进一步地还由文本信息得到。

在一些实施例中，上述特征得到子单元包括：语音得到模块（图中未示）、表情得到模块（图中未示）、求和模块（图中未示）。其中，语音得到模块，可以被配置成将音频数据输入已训练完成的语音情绪识别模型，得到语音情绪识别模型输出的语音情绪特征。表情得到模块，可以被配置成将表情特征输入已训练完成的表情情绪识别模型，得到表情情绪识别模型输出的表情情绪特征。求和模块，可以被配置成对语音情绪特征、表情情绪特征加权求和，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述至少一种模态的信息包括用户的图像数据以及文本数据，上述识别单元 402 包括：识别模块（图中未示）、提取模块（图中未示）、特征得到模块（图中未示）。其中，识别模块，可以被配置成基于用户的图像数据，识别用户的表情特征。提取模块，可以被配置成基于文本数据，提取用户的意图信息。特征得到模块，可以被配置成基于文本数据以及表情特征，得到与意图信息对应的用户情绪特征。

在一些实施例中，上述播报单元 405 包括：生成子单元（图中未示）、视频得到子单元（图中未示）。其中，生成子单元，可以被配置成播报单元。视频得到子单元，可以被配置成基于答复音频、人物情绪特征以及预先建立的动画人物形象模型，得到与人物情绪特征对应的动画人物形象的播报视频。

在一些实施例中，上述视频得到子单元包括：口型驱动模块（图中未示）、

表情驱动模块（图中未示）、模型驱动模块（图中未示）、图片得到模块（图中未示）、视频得到模块（图中未示）。其中，上述视频得到子单元包括：口型驱动模块，被配置成将答复音频、人物情绪特征输入已训练完成的口型驱动模型，得到口型驱动模型输出的口型数据；表情驱动模块，被配置成将答复音频、人物情绪特征输入已训练完成的表情驱动模型，得到表情驱动模型输出的表情数据；模型驱动模块，被配置成基于口型数据、表情数据对动画人物形象模型进行驱动，得到三维模型动作序列；图片得到模块，被配置成对三维模型动作序列进行渲染，得到视频帧图片序列；视频得到模块，被配置成合成视频帧图片序列，得到与人物情绪特征对应的动画人物形象的播报视频。口型驱动模型、表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到。

根据本公开的实施例提供的人机交互装置：首先，接收单元 401 接收用户的至少一种模态的信息；其次，识别单元 402 基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；再次，确定单元 403 基于意图信息，确定对用户的答复信息；从次，选定单元 404 基于用户情绪特征，选定向用户反馈的人物情绪特征；最后，播报单元 405 基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。由此，通过对用户的至少一种模态的信息进行分析确定动画人物形象的人物情绪特征，为不同情绪的用户提供了有效的情绪反馈，保证了人机交互过程中的感情交流。

进一步参考图 5，作为对上述各图所示方法的实现，本公开提供了人机交互系统的一个实施例，该系统实施例与图 2 所示的方法实施例相对应。

如图 5 所示，本公开的实施例提供了一种人机交互系统 500，该系统 500 包括：采集设备 501、显示设备 502 以及分别与采集设备 501、显示设备 502 连接的交互平台 503。采集设备 501 用于采集用户的至少一种模态的信息。交互平台 503 用于接收用户的至少一种模态的信息；基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；基于意图信息，确定对用户的答复信息；基于用户情绪特征，选定向用户反馈的人物情绪特征；基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。显示设备 502 用于接收并播放播报视频。

本实施例中，采集设备为采集用户的至少一种模态的信息的设备，基于不同模态的信息，采集设备的种类不同。比如，至少一种模态的信息包括用户的图像数据以及音频数据，相应地，采集设备可以包括摄像头、扬声器。进一步，至少一种模态的信息包括用户的文本数据，则采集设备还可以包括键盘、鼠标等输入装置。

本实施例中，采集设备 501、显示设备 502 以及交互平台 503 三者可以分离设置，也可以是集成在一起形成一体化机（如图 1 的自动柜员机、终端设备）。

下面参考图 6，其示出了适于用来实现本公开的实施例的电子设备 600 的结构示意图。

如图 6 所示，电子设备 600 可以包括处理装置（例如中央处理器、图形处理器等）601，其可以根据存储在只读存储器（ROM）602 中的程序或者从存储装置 608 加载到随机访问存储器（RAM）603 中的程序而执行各种适当的动作和处理。在 RAM 603 中，还存储有电子设备 600 操作所需的各种程序和数据。处理装置 601、ROM 602 以及 RAM 603 通过总线 604 彼此相连。输入/输出（I/O）接口 605 也连接至总线 604。

通常，以下装置可以连接至 I/O 接口 605：包括例如触摸屏、触摸板、键盘、鼠标等的输入装置 606；包括例如液晶显示器（LCD，Liquid Crystal Display）、扬声器、振动器等输出装置 607；包括例如磁带、硬盘等的存储装置 608；以及通信装置 609。通信装置 609 可以允许电子设备 600 与其他设备进行无线或有线通信以交换数据。虽然图 6 示出了具有各种装置的电子设备 600，但是应理解的是，并不要求实施或具备所有示出的装置。可以替代地实施或具备更多或更少的装置。图 6 中示出的每个方框可以代表一个装置，也可以根据需要代表多个装置。

特别地，根据本公开的实施例，上文参考流程图描述的过程可以被实现为计算机软件程序。例如，本公开的实施例包括一种计算机程序产品，其包括承载在计算机可读介质上的计算机程序，该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中，该计算机程序可以通过通信装置 609 从网络上被下载和安装，或者从存储装置 608 被安装，或者从 ROM 602 被安装。在该计算机程序被处理装置 601 执行时，执行本公开的实施例

的方法中限定的上述功能。

需要说明的是，本公开的实施例的计算机可读介质可以是计算机可读信号介质或者计算机可读存储介质或者是上述两者的任意组合。计算机可读存储介质例如可以是——但不限于——电、磁、光、电磁、红外线、或半导体的系统、装置或器件，或者任意以上的组合。计算机可读存储介质的更具体的例子可以包括但不限于：具有一个或多个导线的电连接、便携式计算机磁盘、硬盘、随机访问存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM 或闪存）、光纤、便携式紧凑磁盘只读存储器（CD-ROM）、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开的实施例中，计算机可读存储介质可以是任何包含或存储程序的有形介质，该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。而在本公开的实施例中，计算机可读信号介质可以包括在基带中或者作为载波一部分传播的数据信号，其中承载了计算机可读的程序代码。这种传播的数据信号可以采用多种形式，包括但不限于电磁信号、光信号或上述的任意合适的组合。计算机可读信号介质还可以是计算机可读存储介质以外的任何计算机可读介质，该计算机可读信号介质可以发送、传播或者传输用于由指令执行系统、装置或者器件使用或者与其结合使用的程序。计算机可读介质上包含的程序代码可以用任何适当的介质传输，包括但不限于：电线、光缆、RF（Radio Frequency，射频）等等，或者上述的任意合适的组合。

上述计算机可读介质可以是上述服务器中所包含的；也可以是单独存在，而未装配入该服务器中。上述计算机可读介质承载有一个或者多个程序，当上述一个或者多个程序被该服务器执行时，使得该服务器：接收用户的至少一种模态的信息；基于至少一种模态的信息，识别用户的意图信息以及与意图信息对应的用户情绪特征；基于意图信息，确定对用户的答复信息；基于用户情绪特征，选定向用户反馈的人物情绪特征；基于人物情绪特征与答复信息，生成与人物情绪特征对应的动画人物形象的播报视频。

可以以一种或多种程序设计语言或其组合来编写用于执行本公开的实施例的操作的计算机程序代码，程序设计语言包括面向对象的程序设计语言——诸如 Java、Smalltalk、C++，还包括常规的过程式程序设计语言——诸如“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机

上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网(LAN)或广域网(WAN)—连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。

附图中的流程图和框图，图示了按照本公开的各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分，该模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意，在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个接连地表示的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或操作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

描述于本公开的实施例中所涉及到的单元可以通过软件的方式实现，也可以通过硬件的方式来实现。所描述的单元也可以设置在处理器中，例如，可以描述为：一种处理器，包括接收单元、识别单元、确定单元、选定单元、播报单元。其中，这些单元的名称在某种情况下并不构成对该单元本身的限定，例如，接收单元还可以被描述为“配置成接收用户的至少一种模态的信息”的单元。

以上描述仅为本公开的较佳实施例以及对所运用技术原理的说明。本领域技术人员应当理解，本公开的实施例中所涉及的发明范围，并不限于上述技术特征的特定组合而成的技术方案，同时也应涵盖在不脱离上述发明构思的情况下，由上述技术特征或其等同特征进行任意组合而形成的其它技术方案。例如上述特征与本公开的实施例中公开的(但不限于)具有类似功能的技术特征进行互相替换而形成的技术方案。

权利要求书

1、一种人机交互方法，所述方法包括：

接收用户的至少一种模态的信息；

基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征；

基于所述意图信息，确定对所述用户的答复信息；

基于所述用户情绪特征，选定向所述用户反馈的人物情绪特征；

基于所述人物情绪特征与所述答复信息，生成与所述人物情绪特征对应的动画人物形象的播报视频。

2、根据权利要求1所述的方法，其中，

所述至少一种模态的信息包括所述用户的图像数据以及音频数据，

所述基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征，包括：

基于所述用户的图像数据，识别所述用户的表情特征；

由所述音频数据，得到文本信息；

基于所述文本信息，提取所述用户的意图信息；

基于所述音频数据以及所述表情特征，得到与所述意图信息对应的用户情绪特征。

3、根据权利要求2所述的方法，其中，所述基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征，还包括：

所述用户情绪特征还由所述文本信息得到。

4、根据权利要求2所述的方法，其中，所述基于所述音频数据以及所述表情特征，得到与所述意图信息对应的用户情绪特征，包括：

将所述音频数据输入已训练完成的语音情绪识别模型，得到所述语音情绪识别模型输出的语音情绪特征；

将所述表情特征输入已训练完成的表情情绪识别模型，得到所述表情情绪识别模型输出的表情情绪特征；

对所述语音情绪特征、所述表情情绪特征加权求和，得到与所述意图信息对应的用户情绪特征。

5、根据权利要求1所述的方法，其中，所述至少一种模态的信息包括所述用户的图像数据以及文本数据；

所述基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征，包括：

基于所述用户的图像数据，识别所述用户的表情特征；

基于所述文本数据，提取所述用户的意图信息；

基于所述文本数据以及所述表情特征，得到与所述意图信息对应的用户情绪特征。

6、根据权利要求1-5之一所述的方法，其中，所述基于所述人物情绪特征与所述答复信息，生成与所述人物情绪特征对应的动画人物形象的播报视频，包括：

基于所述答复信息、所述人物情绪特征，生成答复音频；

基于所述答复音频、所述人物情绪特征以及预先建立的动画人物形象模型，得到与所述人物情绪特征对应的动画人物形象的播报视频。

7、根据权利要求6所述的方法，其中，所述基于所述答复音频、所述人物情绪特征以及预先建立的动画人物形象模型，得到与所述人物情绪特征对应的动画人物形象的播报视频，包括：

将所述答复音频、所述人物情绪特征输入已训练完成的口型驱动模型，得到所述口型驱动模型输出的口型数据；

将所述答复音频、所述人物情绪特征输入已训练完成的表情驱动模型，得到所述表情驱动模型输出的表情数据；

基于所述口型数据、所述表情数据对所述动画人物形象模型进行驱动，得到三维模型动作序列；

对所述三维模型动作序列进行渲染，得到视频帧图片序列；

合成所述视频帧图片序列，得到与所述人物情绪特征对应的动画人物形象的播报视频，

其中，所述口型驱动模型、所述表情驱动模型基于预标注的同一人的音频以及由该音频得到的音频情绪信息训练得到。

8、一种人机交互装置，所述装置包括：

接收单元，被配置成接收用户的至少一种模态的信息；

识别单元，被配置成基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征；

确定单元，被配置成基于所述意图信息，确定对所述用户的答复信息；

选定单元，被配置成基于所述用户情绪特征，选定向所述用户反馈的人物情绪特征；

播报单元，被配置成基于所述人物情绪特征与所述答复信息，生成与所述人物情绪特征对应的动画人物形象的播报视频。

9、一种人机交互系统，所述系统包括：采集设备、显示设备以及分别与所述采集设备、所述显示设备连接的交互平台；

所述采集设备用于采集用户的至少一种模态的信息；

所述交互平台用于接收所述用户的至少一种模态的信息；基于所述至少一种模态的信息，识别所述用户的意图信息以及与所述意图信息对应的用户情绪特征；基于所述意图信息，确定对所述用户的答复信息；基于所述用户情绪特征，选定向所述用户反馈的人物情绪特征；基于所述人物情绪特征与所述答复信息，生成与所述人物情绪特征对应的动画人物形象的播报视频；

所述显示设备用于接收并播放所述播报视频。

10、一种电子设备，包括：

一个或多个处理器；

存储装置，其上存储有一个或多个程序；

当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现如权利要求 1-7 中任一所述的方法。

11、一种计算机可读介质，其上存储有计算机程序，其中，该程序被处理器执行时实现如权利要求 1-7 中任一所述的方法。

12、一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现如权利要求 1-7 中任一项所述的方法。