(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117671199 A
(43)申请公布日 2024.03.08

(21)申请号 202211006456.X

(22)申请日 2022.08.22

(71)申请人 北京字跳网络技术有限公司
地址 100190 北京市海淀区紫金数码园4号楼2层0207

(72)发明人 李晨

(74)专利代理机构 北京开阳星知识产权代理有限公司 11710
专利代理师 吴崇

(51)Int.Cl.
*G06T 19/00*(2011.01)
*G06T 7/73*(2017.01)
*G06V 20/40*(2022.01)
*G06V 40/16*(2022.01)
*G10L 15/26*(2006.01)
*G10L 17/22*(2013.01)
*G10L 25/51*(2013.01)

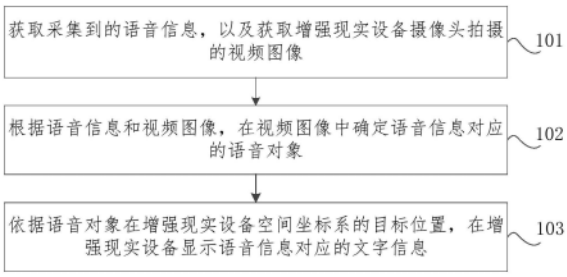
权利要求书2页 说明书10页 附图6页

(54)发明名称

信息显示方法、装置及电子设备

(57)摘要

本公开涉及一种信息显示方法、装置及电子设备，涉及人工智能技术领域，其中方法包括：首先获取采集到的语音信息；及获取所述增强现实设备摄像头拍摄的视频图像；再根据所述语音信息和所述视频图像，在所述视频图像中确定所述语音信息对应的语音对象；然后依据所述语音对象在所述增强现实设备空间坐标系的目标位置，在所述增强现实设备显示所述语音信息对应的文字信息。通过应用本公开的技术方案，可帮助听障用户快速准确地分辨出每一句话语所对应的说话人，提高了听障用户对文字的理解力，进而提升了对听障用户的辅助效果。




CN 117671199 A

1.一种信息显示方法,其特征在于,用于增强现实设备,包括:

获取采集到的语音信息;及,

获取所述增强现实设备摄像头拍摄的视频图像;

根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象;

依据所述语音对象在所述增强现实设备空间坐标系的目标位置,在所述增强现实设备显示所述语音信息对应的文字信息。

2.根据权利要求1所述的方法,其特征在于,所述根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象,包括:

从所述语音信息中确定语音信号的第一时间信息;及,

识别所述视频图像中的说话对象和相应说话时的第二时间信息;

将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象。

3.根据权利要求2所述的方法,其特征在于,所述将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象,具体包括:

获取所述语音信号开始的第一时间点;及,

获取说话对象开始说话时的第二时间点;

若所述第一时间点与所述第二时间点之间的时间差小于预设时长阈值,则将所述第二时间点对应的说话对象,确定为所述语音信息所对应的语音对象。

4.根据权利要求2所述的方法,其特征在于,所述将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象,具体包括:

获取所述语音信号的时间段;及,

获取说话对象的说话时间段;

若所述语音信号的时间段与所述说话时间段之间的相似度大于预设相似度阈值,则将所述说话时间段对应的说话对象,确定为所述语音信息所对应的语音对象。

5.根据权利要求2所述的方法,其特征在于,若存在多个同时说话的说话对象,则所述将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象,具体包括:

获取所述第一时间信息各自对应的声源方向信息;及,

获取所述第二时间信息各自对应的说话对象所处方向信息;

将所述第一时间信息和所述第二时间信息匹配的、且声源方向信息和说话对象所处方向信息匹配的说话对象,确定为语音信息所对应的语音对象。

6.根据权利要求2所述的方法,其特征在于,若存在多个同时说话的说话对象,则所述将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象,具体包括:

获取同时说话的说话对象各自的声纹特征;

将所述声纹特征与说话对象之前说话时的历史声纹特征进行匹配;

将所述第一时间信息和所述第二时间信息匹配的、且声纹特征与历史声纹特征匹配的说话对象,确定为语音信息所对应的语音对象。

7.根据权利要求2所述的方法,其特征在于,所述识别所述视频图像中的说话对象,具体包括:

通过人脸识别确定所述视频图像中的人物对象;

根据所述人物对象的口型变化,判断所述人物对象是否在说话;

将判定为在说话的人物对象,确定为所述说话对象。

8.根据权利要求7所述的方法,其特征在于,所述根据所述人物对象的口型变化,判断所述人物对象是否在说话,具体包括:

将所述人物对象的口型变化特征与样本对象说话时的口型变化特征进行匹配;

若匹配,则判定所述人物对象在说话。

9.根据权利要求7所述的方法,其特征在于,判断所述人物对象是否在说话的过程是通过机器学习模型计算得到的,所述机器学习模型是通过样本对象说话时的口型变化特征、和/或没有说话时的口型变化特征预先训练得到的。

10.根据权利要求1所述的方法,其特征在于,所述获取采集到的语音信息,具体包括:

采集在预设方向角度范围内的声源发出的语音信息,作为采集到的所述语音信息,其中,所述预设方向角度范围与摄像头拍摄所述视频图像时的方向角度范围相对应。

11.根据权利要求1所述的方法,其特征在于,所述依据所述语音对象在所述增强现实设备空间坐标系的目标位置,在所述增强现实设备显示所述语音信息对应的文字信息,具体包括:

将所述语音信息转换的文字信息,显示在所述语音对象所对应的所述目标位置的预设范围内。

12.根据权利要求11所述的方法,其特征在于,所述将所述语音信息转换的文字信息,显示在所述语音对象所对应的所述目标位置的预设范围内,具体包括:

从所述目标位置获取所述语音对象的人脸中心坐标;

基于所述人脸中心坐标,将所述文字信息显示在对应人脸旁的预设范围内。

13.根据权利要求1至12中任一项所述的方法,其特征在于,所述文字信息是所述语音信息在环境消噪处理后转换得到的。

14.一种信息显示装置,其特征在于,用于增强现实设备,包括:

获取模块,被配置为获取采集到的语音信息;及,获取所述增强现实设备摄像头拍摄的视频图像;

确定模块,被配置为根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象;

显示模块,被配置为依据所述语音对象在所述增强现实设备空间坐标系的目标位置,在所述增强现实设备显示所述语音信息对应的文字信息。

15.一种计算机可读存储介质,其上存储有计算机程序,其特征在于,所述计算机程序被处理器执行时实现权利要求1至13中任一项所述的方法。

16.一种电子设备,包括存储介质、处理器及存储在存储介质上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述计算机程序时实现权利要求1至13中任一项所述的方法。

# 信息显示方法、装置及电子设备

## 技术领域

[0001] 本公开涉及人工智能技术领域,尤其涉及一种信息显示方法、装置及电子设备。

## 背景技术

[0002] 近年来,随着科技的发展和社会的进步,听障人士受到关注程度也在不断地提高。

[0003] 目前,为了辅助听障人士明白他人讲话内容,可利用增强现实(Augmented Reality,AR)设备将所有人的话语全部无差别的显示在用户眼前。

[0004] 然而,这种方式会导致听障用户无法分辨出每一句话语所对应的说话人,对文字的理解力会变得很低,进而影响了对听障用户的辅助效果。

## 发明内容

[0005] 有鉴于此,本公开提供了一种信息显示方法、装置及电子设备,主要目的在于改善目前辅助听障用户的方式会导致听障用户无法分辨出每一句话语所对应的说话人,对文字的理解力会变得很低,进而影响了对听障用户的辅助效果的技术问题。

[0006] 第一方面,本公开提供了一种信息显示方法,用于增强现实设备,包括:

[0007] 获取采集到的语音信息;及,

[0008] 获取所述增强现实设备摄像头拍摄的视频图像;

[0009] 根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象;

[0010] 依据所述语音对象在所述增强现实设备空间坐标系的目标位置,在所述增强现实设备显示所述语音信息对应的文字信息。

[0011] 第二方面,本公开提供了一种信息显示装置,包括:

[0012] 获取模块,被配置为获取采集到的语音信息;及,获取所述增强现实设备摄像头拍摄的视频图像;

[0013] 确定模块,被配置为根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象;

[0014] 显示模块,被配置为依据所述语音对象在所述增强现实设备空间坐标系的目标位置,在所述增强现实设备显示所述语音信息对应的文字信息。

[0015] 第三方面,本公开提供了一种计算机可读存储介质,其上存储有计算机程序,所述计算机程序被处理器执行时实现第一方面所述的信息显示方法。

[0016] 第四方面,本公开提供了一种电子设备,包括存储介质、处理器及存储在存储介质上并可在处理器上运行的计算机程序,所述处理器执行所述计算机程序时实现第一方面所述的信息显示方法。

[0017] 借由上述技术方案,本公开提供的一种信息显示方法、装置及电子设备,与目前利用增强现实设备将所有人的话语全部无差别的显示在用户眼前的方式相比,本公开可实现将说话内容与相对应的说话对象进行关联显示,使得听障用户能够快速定位显示的话语是

哪个人说出的。具体的,在增强现实设备侧,首先获取采集到的语音信息,以及获取增强现实设备摄像头拍摄的视频图像;再根据语音信息和视频图像,在视频图像中确定语音信息对应的语音对象;然后依据语音对象在增强现实设备空间坐标系的目标位置,在增强现实设备显示语音信息对应的文字信息。通过应用本公开的技术方案,可帮助听障用户快速准确地分辨出每一句话语所对应的说话人,提高了听障用户对文字的理解力,进而提升了对听障用户的辅助效果。

[0018] 上述说明仅是本公开技术方案的概述,为了能够更清楚了解本公开的技术手段,而可依照说明书的内容予以实施,并且为了让本公开的上述和其它目的、特征和优点能够更明显易懂,以下特举本公开的具体实施方式。

## 附图说明

[0019] 此处的附图被并入说明书中并构成本说明书的一部分,示出了符合本公开的实施例,并与说明书一起用于解释本公开的原理。

[0020] 为了更清楚地说明本公开实施例或现有技术中的技术方案,下面将对实施例或现有技术描述中所需要使用的附图作简单地介绍,显而易见地,对于本领域普通技术人员而言,在不付出创造性劳动性的前提下,还可以根据这些附图获得其他的附图。

[0021] 图1示出了本公开实施例提供的一种信息显示方法的流程示意图;

[0022] 图2示出了本公开实施例提供的一种AR文字显示示例效果的示意图;

[0023] 图3示出了本公开实施例提供的另一种AR文字显示示例效果的示意图;

[0024] 图4示出了本公开实施例提供的又一种AR文字显示示例效果的示意图;

[0025] 图5示出了本公开实施例提供的再一种AR文字显示示例效果的示意图;

[0026] 图6示出了本公开实施例提供的另一种信息显示方法的流程示意图;

[0027] 图7示出了本公开实施例提供的增强现实设备示例的结构示意图;

[0028] 图8示出了本公开实施例提供的应用场景示例的流程示意图;

[0029] 图9示出了本公开实施例提供的应用场景示例的AR文字显示示例效果的示意图;

[0030] 图10示出了本公开实施例提供的一种信息显示装置的结构示意图。

## 具体实施方式

[0031] 下面将参照附图更详细地描述本公开的实施例。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。

[0032] 为了改善目前辅助听障用户的方式会导致听障用户无法分辨出每一句话语所对应的说话人,对文字的理解力会变得很低,进而影响了对听障用户的辅助效果的技术问题。本实施例提供了一种信息显示方法,如图1所示,可应用于增强现实设备(如AR眼镜)端侧,该方法包括:

[0033] 步骤101、获取采集到的语音信息,以及获取增强现实设备摄像头拍摄的视频图像。

[0034] 对于本实施例可通过麦克风阵列(Microphone Array)采集声源发出的语音信息。在采集语音信息的同时,本实施例还可通过摄像头(Camera)拍摄图像。其中,麦克风阵列和摄像头可为增强现实设备的内置装置或者与增强现实设备连接的外置设备等。

[0035] 步骤102、根据语音信息和视频图像,在视频图像中确定语音信息对应的语音对象。

[0036] 语音对象即说话对象,具体可以是人、动物、物品(如能够模仿人张嘴说话的假人玩具)等。例如,以说话对象为人作为示例,本实施例可基于人脸及口型识别,判断摄像头拍摄到的视频图像中是否存在说话的人,如果存在说话人,可结合当前采集到的语音信息,在视频图像中确定与该语音信息对应的说话人。如具体可基于说话人的位置和该语音信息的声源方向,确定与该语音信息对应的说话人。

[0037] 步骤103、依据语音对象在增强现实设备空间坐标系的目标位置,在增强现实设备显示语音信息对应的文字信息。

[0038] 例如,在增强现实空间中依据语音信息所对应的语音对象位置(目标位置),关联显示语音信息转换的文字信息。由于在实际当中,语音对象可能会存在边说话边移动、或者说话后迅速移动等情况,因此本实施例中的语音对象位置可以是动态的位置,即实时跟踪语音对象的位置,以便后续显示语音文字时能够做到准确地关联显示。

[0039] 对于语音转文字的过程,可基于自动语音识别(Automatic Speech Recognition,ASR)技术来实现。而对于本实施例中的文字与说话对象关联显示的方式可存在多种可选方式,目的是为了使得用户可直观了解到显示的文字信息是由哪个说话对象说出来的。

[0040] 例如,用户使用增强现实设备与其面前的其他用户进行交谈,在该用户能够观看到的增强现实空间中,可将面前说话人的语音信息转换为文字信息,并显示在该说话人脸部位置附近的特定区域内(如图2所示的AR图像展示效果);或者显示在该说话人头部上方位置的特定区域内(如图3所示的AR图像展示效果);或者显示在特定区域内并带有指示说话目标的效果(如图4所示的AR图像展示效果);或者显示在特定区域内并同时显示有说话人的标记,该标记也一并显示在AR画面当中,如图5所示,根据不同的人脸特征,对画面中的每个人进行标记,如出现两人,分别标记为人像a和人像b,然后在特定区域内将语音信息转换的文字信息前面加上对应说话人的标记。

[0041] 本实施例提供的信息显示方法,与目前利用增强现实设备将所有人的话语全部无差别的显示在用户眼前的方式相比,本实施例可实现将说话内容与相对应的说话对象进行关联显示,使得听障用户能够快速定位显示的话语是哪个人说出的。具体的,在增强现实设备侧,首先获取采集到的语音信息,以及获取增强现实设备摄像头拍摄的视频图像;再根据语音信息和视频图像,在视频图像中确定语音信息对应的语音对象;然后依据语音对象在增强现实设备空间坐标系的目标位置,在增强现实设备显示语音信息对应的文字信息。通过应用本实施例的技术方案,可帮助听障用户快速准确地分辨出每一句话语所对应的说话人,提高了听障用户对文字的理解力,进而提升了对听障用户的辅助效果。

[0042] 进一步的,作为上述实施例的细化和扩展,为了完整说明本实施例方法的具体实现过程,本实施例提供了如图6所示的具体方法,该方法包括:

[0043] 步骤201、增强现实设备接收图像处理的触发指令。

[0044] 图像处理的触发指令可用于触发执行本实施例的方法,进而实现开始辅助听障用户的过程。例如,增强现实设备被开启后自动输入图像处理的触发指令,或者用户通过点击预置功能键输入该指令等,使用该增强现实设备的用户,可观看到面前的说话人以及该说话人所说的文字内容。

[0045] 例如,如图7所示,本实施例中的增强现实设备可包括:麦克风阵列(MIC阵列)、显示模组、摄像头(Camera)以及系统级芯片(System on Chip,SOC)。其中,MIC阵列可负责采集人声;Camera可负责图像采集;显示模组,包括:左右显示模块(如双目AR眼镜设备),可负责显示处理好的文字信息;SOC可主要负责图像处理和音频信号处理、语音转文字、以及视频信息输出等。

[0046] 本增强现实设备可同时实现语音信息的处理(执行步骤202a至203a所示的过程)和拍摄图像的处理(执行步骤202b至203b所示的过程)。

[0047] 步骤202a、获取采集到的语音信息。

[0048] 为了保证采集到的语音信息后续能够准确处理得到本实施例方法所需的数据结果,可先做数据预处理,在数据预处理后再进行语音转文字的处理。而在增强现实设备的实际应用中,需要对佩戴者前方的说话人进行语音转换文字的辅助显示,因此可采集佩戴者前方的语音信息。并且为了保证语音信息转换为文字信息的准确性,可选的,文字信息可以是语音信息在环境消噪处理后转换得到的。

[0049] 进一步可选的,步骤202a具体可包括:采集在预设方向角度范围内的声源发出的语音信息,作为采集到的语音信息,其中,预设方向角度范围与摄像头拍摄视频图像时的方向角度范围相对应。预设方向角度范围可根据实际需求进行预先设置,如可取佩戴者正前方120度的范围,与佩戴者的视野范围相对应,即为佩戴者所观看到的人物提供语音转文字显示的辅助功能。而摄像头拍摄的图像也可是佩戴者前方的图像信息,对应佩戴者的视野范围。

[0050] 通过上述可选方式,可有效将用户视野所看到的说话人的说话内容转换为文字进行显示,降低了超出用户视野外的声源所造成的干扰。

[0051] 步骤203a、从语音信息中确定语音信号的第一时间信息。

[0052] 语音信号可为采集到对象说话时的语音信号,第一时间信息可为从语音识别角度确定的对象说话时的时间信息。

[0053] 例如,语音信息可在第一时间进行转换为文字,并记录该文字相应的时间信息,即第一时间信息,该时间信息具体可为时间戳或者时间段等。

[0054] 与步骤202a并列的步骤202b、获取增强现实设备摄像头拍摄到的视频图像。

[0055] 对于本实施例,摄像头具体可拍摄增强现实设备佩戴者前方的图像信息,并且为了保证拍摄到的图像信息后续能够准确处理得到本实施例方法所需的数据结果,可先做数据预处理,在数据预处理后再进行执行步骤203所示的过程,以提高说话对象位置的识别准确性。

[0056] 步骤203b、识别视频图像中的说话对象和相应说话时的第二时间信息。

[0057] 第二时间信息可为从图像识别角度确定的对象说话时的时间信息。

[0058] 以说话对象为人作为示例,可选的,识别视频图像中的说话对象,具体可包括:首先通过人脸识别确定视频图像中的人物对象;然后根据人物对象的口型变化,判断人物对象是否在说话;将判定为在说话的人物对象,确定为说话对象。

[0059] 例如,首先可提取视频图像信息中的图像特征,可包括人脸轮廓和口型轮廓的特征,记录不同口型轮廓特征的时间戳;然后根据图像特征识别出人脸,并根据该人脸的口型轮廓变化,识别判断该人脸所对应的人是否在说话;如果判定该人脸所对应的人在说话,则

依据该说话人的图像位置确定说话对象位置,并可依据记录的不同口型轮廓特征的时间戳,确定该说话人说话时的时间信息,即第二时间信息,该时间信息具体可为时间戳或者时间段等。通过这种基于人脸及口型识别的方式,可准确识别出图像信息中的说话对象位置和说话时的时间信息。并且口型变化不用于识别口语(浪费系统算力和响应时间),而是用于判断当前对象是否在说话,因此可保证处理效率。

[0060] 示例性的,为了说明具体如何根据人物对象的口型变化,判断人物对象是否在说话的过程,给出如下两种可选方式:

[0061] 作为一种可选方式,根据人物对象的口型变化,判断人物对象是否在说话,具体可包括:将人物对象的口型变化特征与样本对象说话时的口型变化特征进行匹配;若匹配,则判定人物对象在说话。

[0062] 通过这种可选方式,可直接根据样本对象说话时的口型变化特征进行判断,如果人物对象的口型变化特征与该样本特征匹配,即可判定人物对象在说话,可做到准确判别。

[0063] 作为另一种可选方式,判断人物对象是否在说话的过程可以是通过机器学习模型计算得到的,该机器学习模型可以是通过样本对象说话时的口型变化特征、和/或没有说话时的口型变化特征预先训练得到的。通过这种利用机器学习模型的判别方式,可做到快速准确地判别出目标对象是否在说话。

[0064] 步骤204、将与第一时间信息匹配的第二时间信息所对应的说话对象,确定为语音信息所对应的语音对象。

[0065] 本实施例中,通过时间信息匹配的方式,将两种事件(即采集到语音信息的事件1和识别出图像信息中存在说话对象的事件2)进行事件绑定,进而准确确定采集到的语音信息所对应的说话对象位置,从而辅助用户直观了解到显示的文字属于哪个人说的,提高了听障用户对文字的理解力。

[0066] 具体的时间信息匹配方式可有多种可选方式,示例性的,作为一种可选方式,步骤204具体可包括:获取语音信号开始的第一时间点(如人物说话语音的起始时间点);及,获取说话对象开始说话时的第二时间点(如通过图像识别确定的人物说话开始的时间点);若第一时间点与第二时间点之间的时间差小于预设时长阈值(根据实际需求进行预先设置),则将该第二时间点对应的说话对象,确定为语音信息所对应的语音对象。

[0067] 例如,采集到的一段语音信息m的语音起始时间点a(可通过时间戳表示),同时根据采集到的图像信息识别出说话对象n的说话开始时间点b(可通过时间戳表示),如果时间点a与时间点b之间的时间差小于一定时长阈值(考虑到两种事件处理在时间上很可能会存在一定的时间差,因此通过预设的时长阈值进行判别这两种事件是否相关联),那么可将说话对象n的图像位置,确定为语音信息m所对应的说话对象位置。通过这种利用事件发生时间点的判别方式,可准确确定语音信息所对应的说话对象。并且在说话对象说话的第一时间就可将语音信息与其进行关联,便于后续用户可观看到说话对象一边说一边显示语音文字的即时效果,进一步便于用户对显示文字的快速理解。

[0068] 作为另一种可选方式,步骤204具体可包括:获取语音信号的时间段;及,获取说话对象的说话时间段;若语音信号的时间段与说话时间段之间的相似度大于预设相似度阈值,则将说话时间段对应的说话对象,确定为语音信息所对应的语音对象。

[0069] 例如,采集到的一段语音信号x的发生时间段1,同时根据采集到的图像信息识别

出说话对象y的说话时间段2,如果时间段1与时间段2之间的相似度大于一定相似度阈值(考虑到两种事件处理在时间上很可能会存在一定的时间差,因此通过预设的时间段相似度阈值进行判别这两种事件是否相关联),那么可将说话对象y的图像位置,确定为语音信号x所对应的说话对象。通过这种利用事件发生时间段的判别方式,同样可准确确定语音信息所对应的说话对象,并且时间匹配更加精确,可做到精准关联。

[0070] 在实际的应用当中,视频图像信息中很可能会存在多个同时说话的说话对象,此时为了做到精准关联显示各自对应的语音文字内容,作为一种可选方式,若存在多个同时说话的说话对象,则步骤204具体可包括:首先获取第一时间信息各自对应的声源方向信息;及,获取第二时间信息各自对应的说话对象所处方向信息;然后根据第一时间信息和第二时间信息,并结合该声源方向信息和说话对象所处方向信息,确定语音信息各自所对应的语音对象。

[0071] 通过这种结合声源方向判别的方式,可在存在多个同时说话的说话对象的情况下,做到语音信息与各自对应说话对象的精准关联,进而可做到说话对象所说的语音的文字信息进行精确地关联显示,避免用户混淆说话对象的说话内容。

[0072] 示例性的,上述根据第一时间信息和第二时间信息,并结合声源方向信息和说话对象所处方向信息,确定语音信息各自所对应的语音对象,具体可包括:将第一时间信息和第二时间信息匹配的、且声源方向信息和说话对象所处方向信息匹配的说话对象,确定为语音信息所对应的语音对象。

[0073] 例如,在用户前方有两个声源发出的语音信息,分别为语音信息a和语音信息b,且拍摄的用户前方图像中有两个说话对象,分别为说话对象A和说话对象B。在通过时间信息匹配后,确定这两个说话对象同时说了话语。其中,说话对象A位于图像左侧的方向位置,而说话对象B位于图像右侧的方向位置。如果语音信息a的声源方向与说话对象A所处方向匹配、且语音信息b的声源方向与说话对象B所处方向匹配,可确定语音信息a所对应的说话对象为说话对象A,并且确定语音信息b所对应的说话对象为说话对象B。

[0074] 除了上述可选方式用于实现多人同时说话的语音精准关联以外,作为另一种可选方式,若存在多个同时说话的说话对象,则步骤204具体还可包括:首先获取同时说话的说话对象各自的声纹特征;然后根据第一时间信息和第二时间信息,并结合声纹特征,确定语音信息各自所对应的语音对象。

[0075] 由于用户声纹特征具备一定的唯一性,因此通过这种结合声纹特征判别的方式,可在存在多个同时说话的说话对象的情况下,也可做到语音信息与各自对应说话对象的精准关联,进而可做到说话对象所说的语音的文字信息进行精确地关联显示。

[0076] 示例性的,上述根据第一时间信息和第二时间信息,并结合声纹特征,确定语音信息各自所对应的语音说话对象,具体可包括:首先将声纹特征与说话对象之前说话时的历史声纹特征进行匹配;然后将第一时间信息和第二时间信息匹配的、且声纹特征与历史声纹特征匹配的说话对象,确定为语音信息所对应的语音对象。

[0077] 例如,在用户前方有两个声源发出的语音信息,分别为语音信息a和语音信息b,且拍摄的用户前方图像中有两个说话对象,分别为说话对象A和说话对象B。在通过时间信息匹配后,确定这两个说话对象同时说了话语。其中,说话对象A和/或说话对象B之前说过话,并记录有相应的声纹特征。如果语音信息a的声纹特征与说话对象A的历史声纹特征匹配、

和/或语音信息b的声纹特征与说话对象B的历史声纹特征匹配,可确定语音信息a所对应的说话对象为说话对象A,并且确定语音信息b所对应的说话对象为说话对象B。

[0078] 步骤205、依据语音对象在增强现实设备空间坐标系的目标位置,在增强现实设备显示语音信息对应的文字信息。

[0079] 可选的,步骤205具体可包括:将语音信息转换的文字信息,显示在语音对象所对应的目标位置的预设范围内。

[0080] 预设范围可根据实际需求进行预先设置,目的是为了使得用户可直观了解到显示的文字信息是由哪个说话对象说出来的。例如,用户使用增强现实设备的过程中,在能够观看到的增强现实空间中,可将面前说话人的语音信息转换为文字信息,并显示在该说话人位置的预设范围(适合显示文字信息的合适区域)内。

[0081] 示例性的,将语音信息转换的文字信息,显示在语音对象所对应的目标位置的预设范围内具体可包括:首先从目标位置获取语音对象的人脸中心坐标;然后基于人脸中心坐标,将文字信息显示在对应人脸旁的预设范围内。例如,可将语音对应的文字信息显示到人脸轮廓的旁边区域(如说话人人脸周围),不遮挡人脸,且具备信息指向性。通过这种可选方式,在方便听障人士理解文字信息的同时,也不影响用户面对沟通者。

[0082] 为了说明上述各实施例的具体实施过程,应用本实施例的方法给出如下应用示例,但不限于此:

[0083] 以AR眼镜为增强现实设备为例,目前,在AR眼镜的特殊应用中,尤其是针对中重度听障用户而言,借助AR眼镜,完成语音转文字,并进行近眼显示是一种非常实用的案例。然而目前市场上的相似产品在多人对话的场景仍然存在很多问题,例如把所有人的话语全部无差别的显示在用户眼前。这样会导致用户无法分辨每一句话语对应的说话人,对文字的理解力会变得很低。

[0084] 基于上述问题,采用本实施例方法,提出了一种基于人脸识别与口型识别的AR文字显示方案。例如,如图8所示,系统在采集语音信息,实施语音转文字的同时,会通过Camera采集人眼前的图像。首先识别出人脸个数,其次针对每个人脸,识别口型变化,口型变化不用于识别口语(浪费系统算力和响应时间),而是用于判断当前对象是否在说话。当系统完成语音采集和语音转文字后,系统同步时间信息。同时会通过口型变化判断图像中属于某个人脸在讲话。识别完成后,会将对应的文字信息显示到人脸轮廓的旁边,不遮挡人脸,且具备信息指向性。如图9所示,为听障人士佩戴AR眼镜后的实际示例效果,进而在方便听障人士理解文字信息的同时,也不影响用户面对沟通者。提高了听障用户对文字的理解力,从而提升了对听障用户的辅助效果。

[0085] 进一步的,作为图1和图6所示方法的具体实现,本实施例提供了一种信息显示装置,可应用于增强现实设备,如图10所示,该装置包括:获取模块31、确定模块32、显示模块33。

[0086] 获取模块31,被配置为获取采集到的语音信息;及,获取所述增强现实设备摄像头拍摄的视频图像;

[0087] 确定模块32,被配置为根据所述语音信息和所述视频图像,在所述视频图像中确定所述语音信息对应的语音对象;

[0088] 显示模块33,被配置为依据所述语音对象在所述增强现实设备空间坐标系的目标

位置,在所述增强现实设备显示所述语音信息对应的文字信息。

[0089] 在具体的应用场景中,确定模块32,具体被配置为从所述语音信息中确定语音信号的第一时间信息;及,识别所述视频图像中的说话对象和相应说话时的第二时间信息;将与所述第一时间信息匹配的所述第二时间信息所对应的说话对象,确定为所述语音信息所对应的语音对象。

[0090] 在具体的应用场景中,确定模块32,具体还被配置为获取所述语音信号开始的第一时间点;及,获取说话对象开始说话时的第二时间点;若所述第一时间点与所述第二时间点之间的时间差小于预设时长阈值,则将所述第二时间点对应的说话对象,确定为所述语音信息所对应的语音对象。

[0091] 在具体的应用场景中,确定模块32,具体还被配置为获取所述语音信号的时间段;及,获取说话对象的说话时间段;若所述语音信号的时间段与所述说话时间段之间的相似度大于预设相似度阈值,则将所述说话时间段对应的说话对象,确定为所述语音信息所对应的语音对象。

[0092] 在具体的应用场景中,确定模块32,具体还被配置为若存在多个同时说话的说话对象,则获取所述第一时间信息各自对应的声源方向信息;及,获取所述第二时间信息各自对应的说话对象所处方向信息;将所述第一时间信息和所述第二时间信息匹配的、且声源方向信息和说话对象所处方向信息匹配的说话对象,确定为语音信息所对应的语音对象。

[0093] 在具体的应用场景中,确定模块32,具体还被配置为若存在多个同时说话的说话对象,则获取同时说话的说话对象各自的声纹特征;将所述声纹特征与说话对象之前说话时的历史声纹特征进行匹配;将所述第一时间信息和所述第二时间信息匹配的、且声纹特征与历史声纹特征匹配的说话对象,确定为语音信息所对应的语音对象。

[0094] 在具体的应用场景中,确定模块32,具体被配置为通过人脸识别确定所述视频图像中的人物对象;根据所述人物对象的口型变化,判断所述人物对象是否在说话;将判定为在说话的人物对象,确定为所述说话对象。

[0095] 在具体的应用场景中,确定模块32,具体还被配置为将所述人物对象的口型变化特征与样本对象说话时的口型变化特征进行匹配;若匹配,则判定所述人物对象在说话。

[0096] 在具体的应用场景中,可选的,利用确定模块32判断所述人物对象是否在说话的过程是通过机器学习模型计算得到,所述机器学习模型是通过样本对象说话时的口型变化特征、和/或没有说话时的口型变化特征预先训练得到的。

[0097] 在具体的应用场景中,获取模块31,具体还被配置为采集在预设方向角度范围内的声源发出的语音信息,作为采集到的所述语音信息,其中,所述预设方向角度范围与摄像头拍摄所述视频图像时的方向角度范围相对应。

[0098] 在具体的应用场景中,显示模块33,具体被配置为将所述语音信息转换的文字信息,显示在所述语音对象所对应的所述目标位置的预设范围内。。

[0099] 在具体的应用场景中,显示模块33,具体还被配置为从所述目标位置获取所述语音对象的人脸中心坐标;基于所述人脸中心坐标,将所述文字信息显示在对应人脸旁的预设范围内。

[0100] 在具体的应用场景中,可选的,所述文字信息是所述语音信息在环境消噪处理后转换得到的。

[0101] 需要说明的是,本实施例提供的一种信息显示装置所涉及各功能单元的其它相应描述,可以参考图1和图6中的对应描述,在此不再赘述。

[0102] 基于上述如图1和图6所示方法,相应的,本实施例还提供了一种计算机可读存储介质,其上存储有计算机程序,该计算机程序被处理器执行时实现上述如图1和图6所示的信息显示方法。

[0103] 基于这样的理解,本公开的技术方案可以以软件产品的形式体现出来,该软件产品可以存储在一个非易失性存储介质(可以是CD-ROM,U盘,移动硬盘等)中,包括若干指令用以使得一台计算机设备(可以是个人计算机,服务器,或者网络设备等)执行本公开各个实施场景的方法。

[0104] 基于上述如图1和图6所示的方法,以及图10所示的虚拟装置实施例,为了实现上述目的,本公开实施例还提供了一种电子设备,具体可以为增强现实设备,如AR眼镜等,该设备包括存储介质和处理器;存储介质,用于存储计算机程序;处理器,用于执行计算机程序以实现上述如图1和图6所示的信息显示方法。

[0105] 可选的,上述实体设备还可以包括用户接口、网络接口、摄像头、射频(Radio Frequency,RF)电路,传感器、音频电路、WI-FI模块等等。用户接口可以包括显示屏(Display)、输入单元比如键盘(Keyboard)等,可选用户接口还可以包括USB接口、读卡器接口等。网络接口可选的可以包括标准的有线接口、无线接口(如WI-FI接口)等。

[0106] 本领域技术人员可以理解,本实施例提供的上述实体设备结构并不构成对该实体设备的限定,可以包括更多或更少的部件,或者组合某些部件,或者不同的部件布置。

[0107] 存储介质中还可以包括操作系统、网络通信模块。操作系统是管理上述实体设备硬件和软件资源的程序,支持信息处理程序以及其它软件和/或程序的运行。网络通信模块用于实现存储介质内部各组件之间的通信,以及与信息处理实体设备中其它硬件和软件之间通信。

[0108] 通过以上的实施方式的描述,本领域的技术人员可以清楚地了解到本公开可以借助软件加必要的通用硬件平台的方式来实现,也可以通过硬件实现。通过应用本实施例的方案,可实现将说话内容显示在相对应的说话对象位置附近,使得听障用户能够快速定位显示的话语是哪个人说出的。进而可帮助听障用户快速准确地分辨出每一句话语所对应的说话人,提高了听障用户对文字的理解力,从而提升了对听障用户的辅助效果。

[0109] 需要说明的是,在本文中,诸如“第一”和“第二”等之类的关系术语仅仅用来将一个实体或者操作与另一个实体或操作区分开来,而不一定要求或者暗示这些实体或操作之间存在任何这种实际的关系或者顺序。而且,术语“包括”、“包含”或者其任何其他变体意在涵盖非排他性的包含,从而使得包括一系列要素的过程、方法、物品或者设备不仅包括那些要素,而且还包括没有明确列出的其他要素,或者是还包括为这种过程、方法、物品或者设备所固有的要素。在没有更多限制的情况下,由语句“包括一个……”限定的要素,并不排除在包括所述要素的过程、方法、物品或者设备中还存在另外的相同要素。

[0110] 以上所述仅是本公开的具体实施方式,使本领域技术人员能够理解或实现本公开。对这些实施例的多种修改对本领域的技术人员来说将是显而易见的,本文中所定义的一般原理可以在不脱离本公开的精神或范围的情况下,在其它实施例中实现。因此,本公开将不会被限制于本文所述的这些实施例,而是要符合与本文所公开的原理和新颖特点相一

致的最宽的范围。

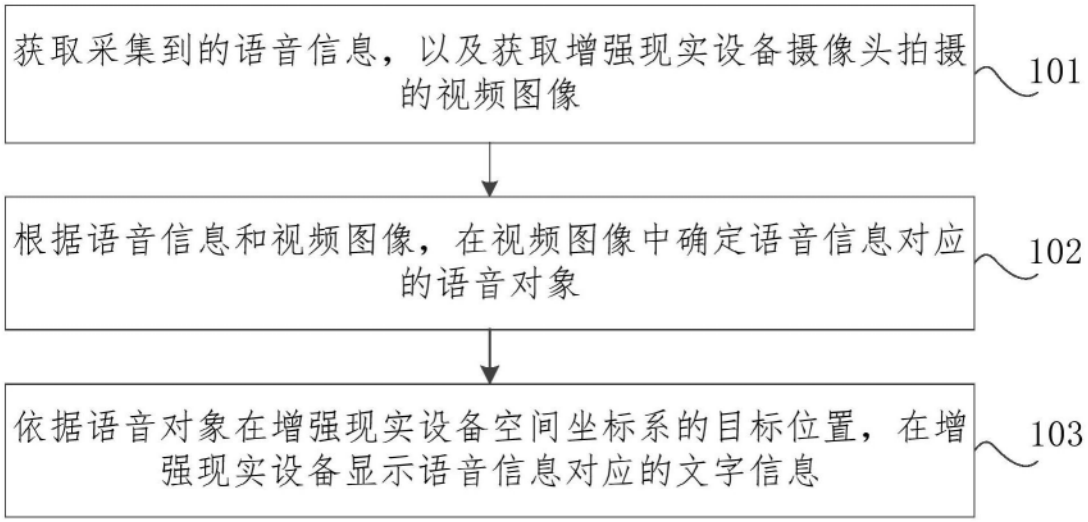

获取采集到的语音信息，以及获取增强现实设备摄像头拍摄的视频图像
101
根据语音信息和视频图像，在视频图像中确定语音信息对应的语音对象
102
依据语音对象在增强现实设备空间坐标系的目标位置，在增强现实设备显示语音信息对应的文字信息
103

图1

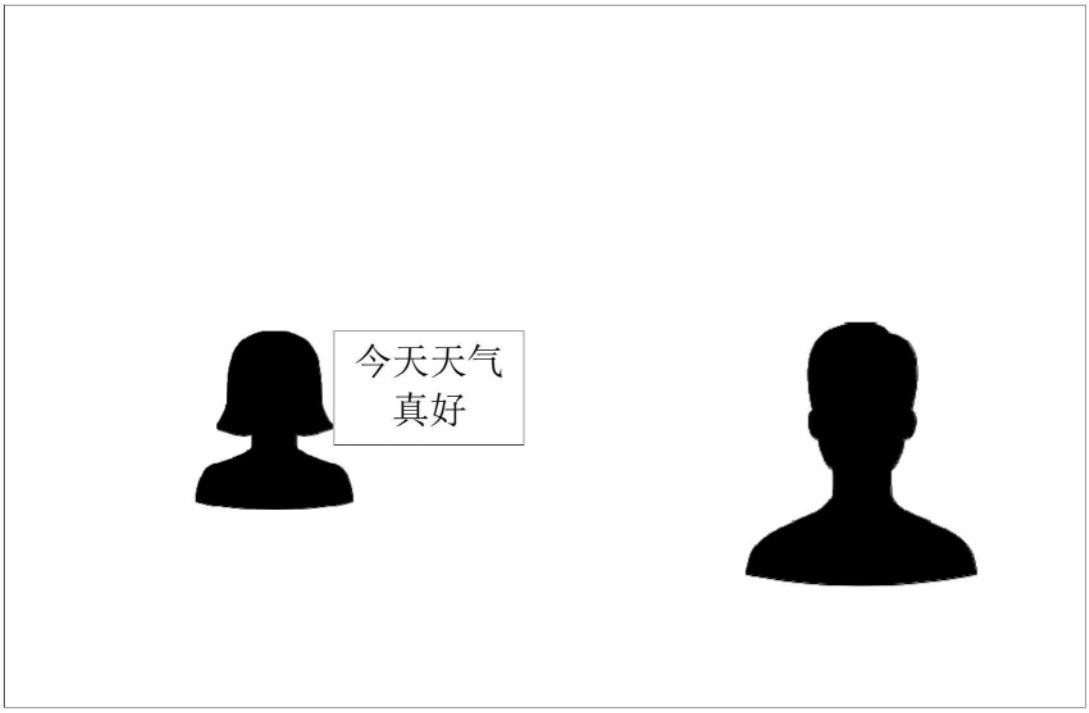

今天天气真好

图2

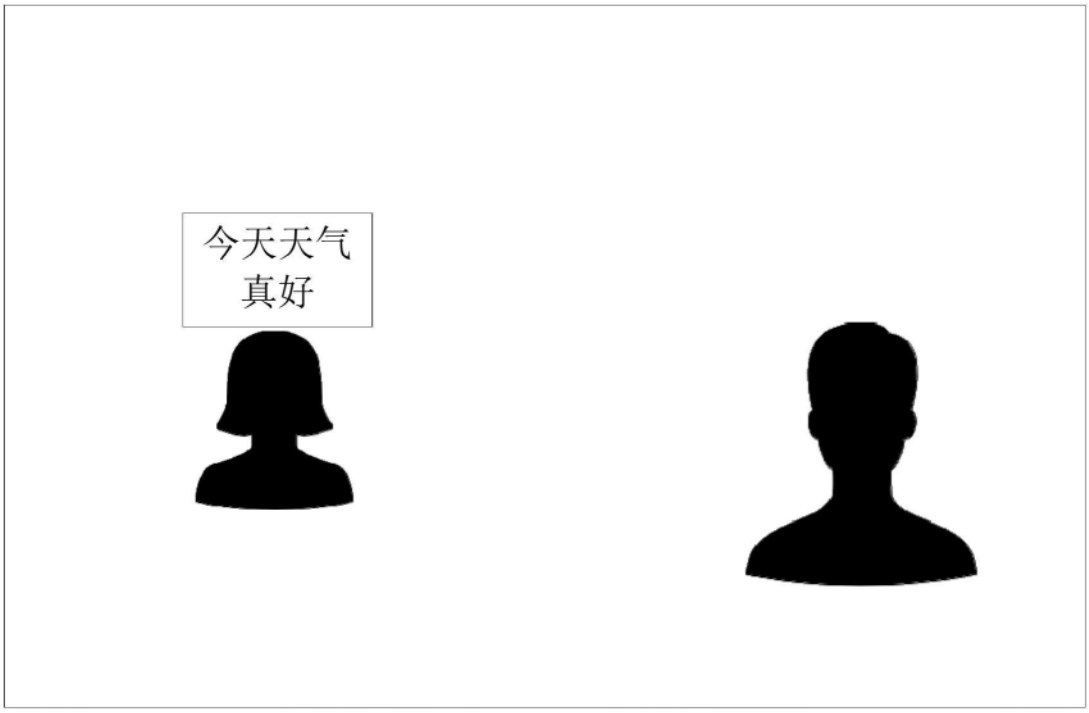



图3

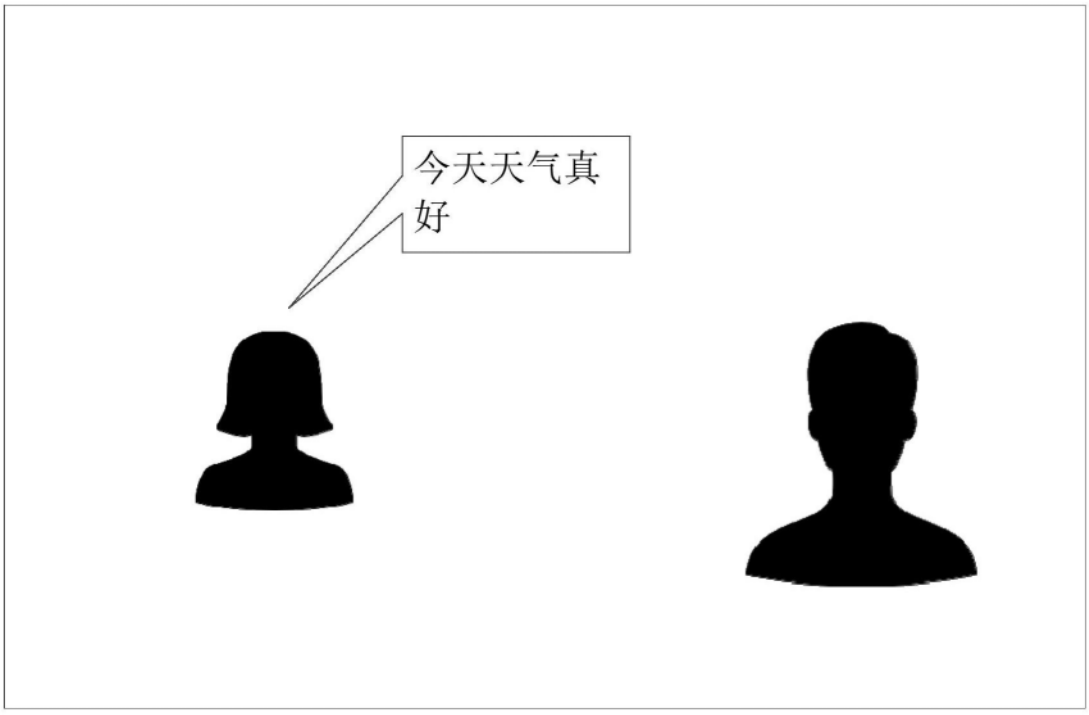



图4

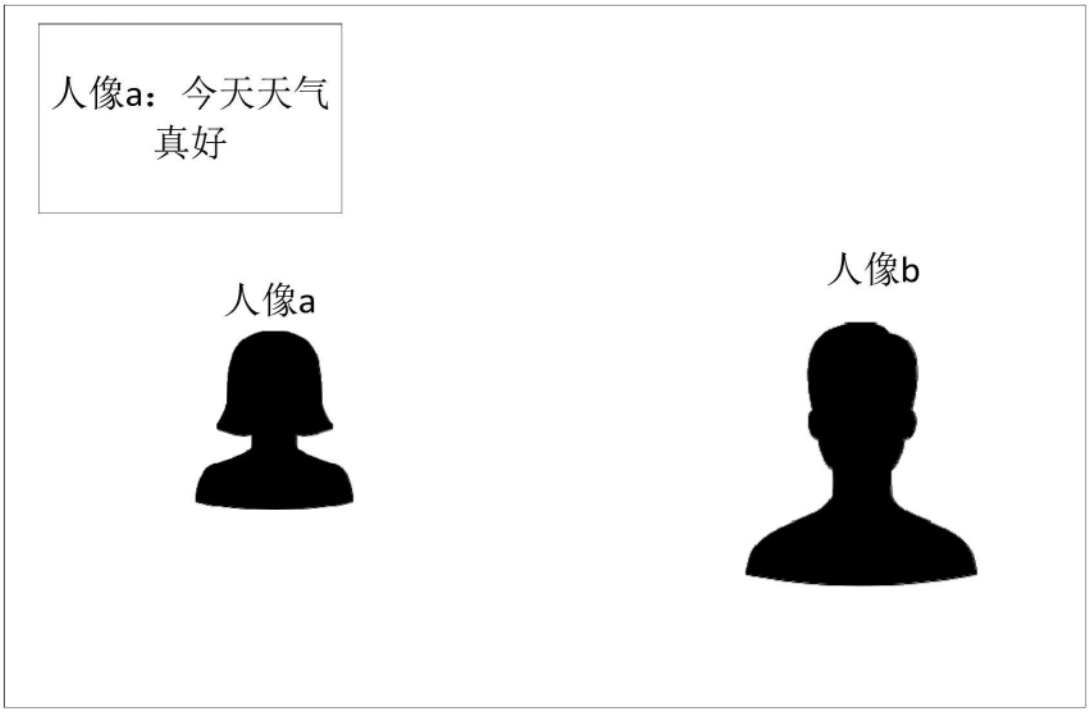

人像a：今天天气真好
人像a
人像b

图5

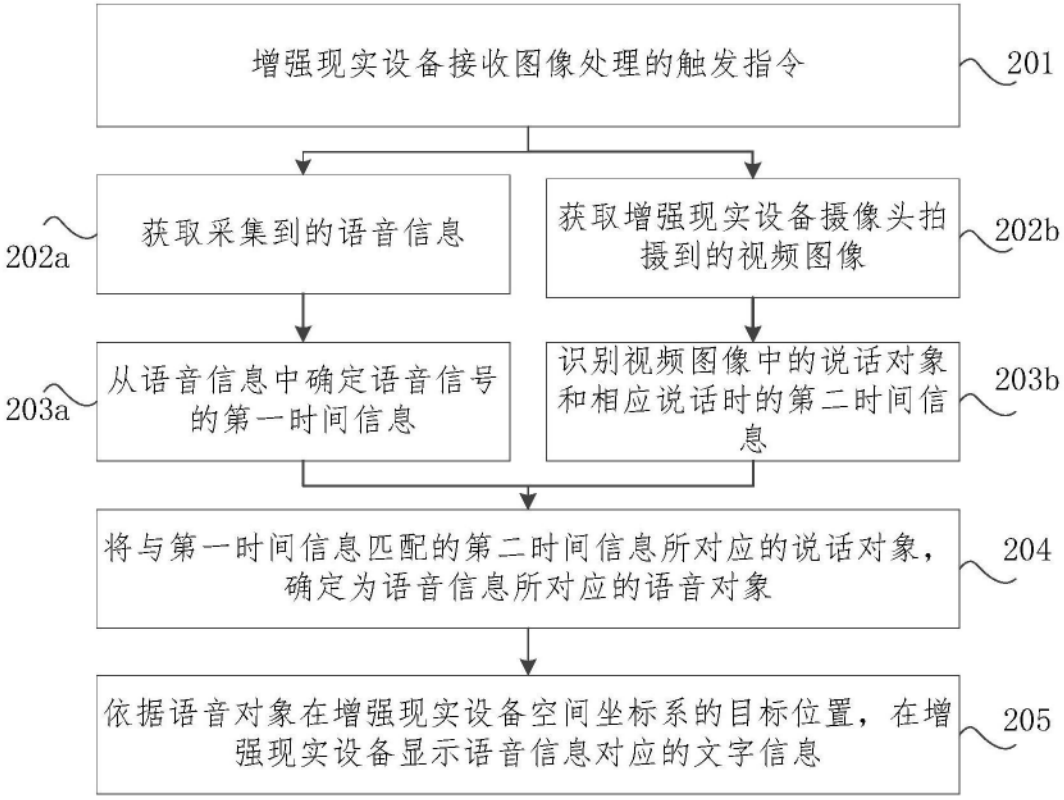

增强现实设备接收图像处理的触发指令
201
获取采集到的语音信息
202a
获取增强现实设备摄像头拍摄到的视频图像
202b
从语音信息中确定语音信号的第一时间信息
203a
识别视频图像中的说话对象和相应说话时的第二时间信息
203b
将与第一时间信息匹配的第二时间信息所对应的说话对象，确定为语音信息所对应的语音对象
204
依据语音对象在增强现实设备空间坐标系的目标位置，在增强现实设备显示语音信息对应的文字信息
205

图6

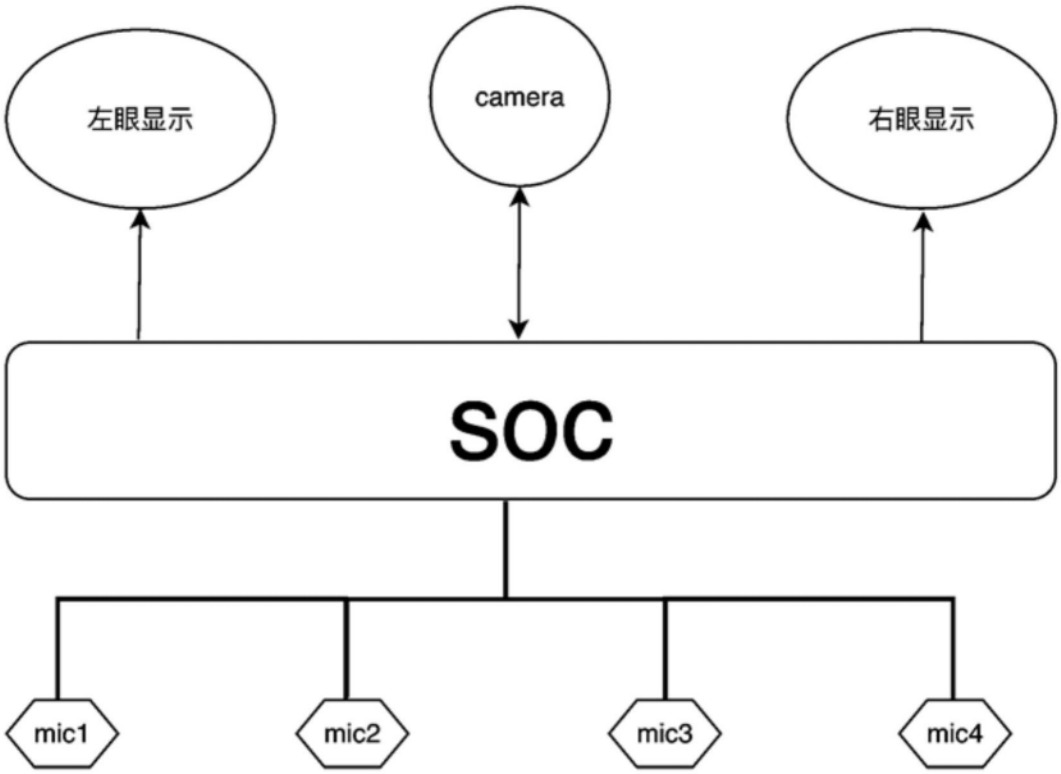



图7

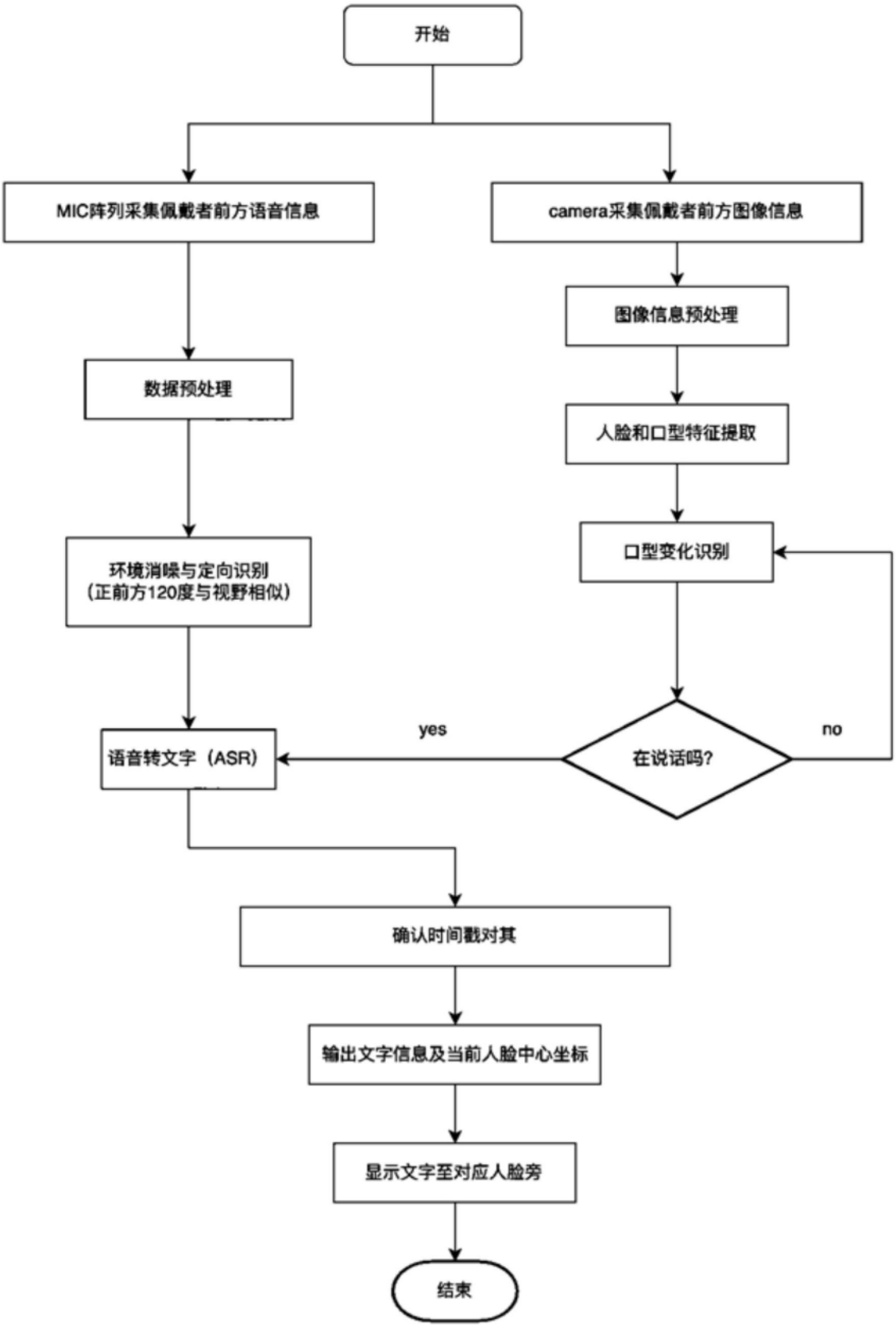

开始
MIC阵列采集佩戴者前方语音信息
camera采集佩戴者前方图像信息
图像信息预处理
数据预处理
人脸和口型特征提取
环境消噪与定向识别
(正前方120度与视野相似)
口型变化识别
yes
在说话吗?
no
语音转文字（ASR）
确认时间戳对其
输出文字信息及当前人脸中心坐标
显示文字至对应人脸旁
结束

图8

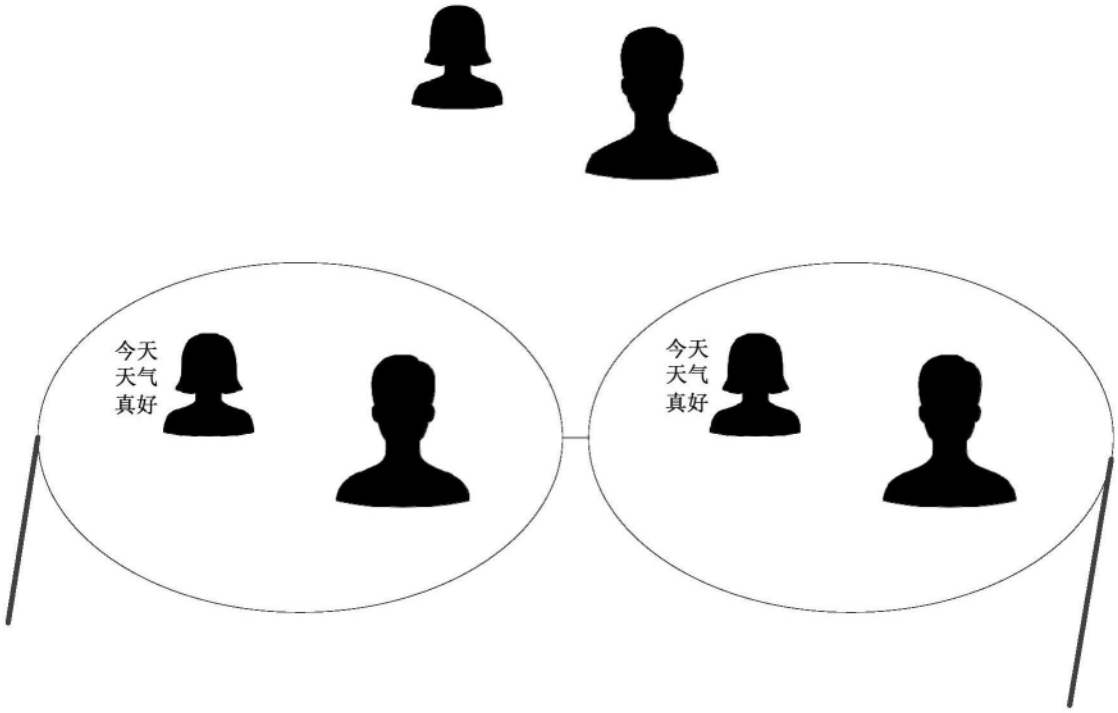



图9

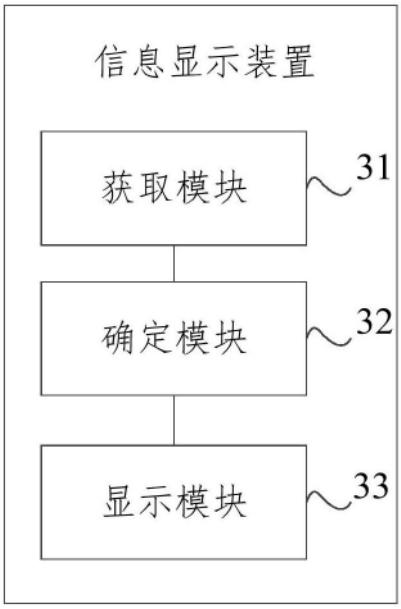



图10