



(12) 发明专利

(10) 授权公告号 CN 111010589 B

(45) 授权公告日 2022. 02. 25

(21) 申请号 201911319864.9

H04N 21/235 (2011.01)

(22) 申请日 2019.12.19

H04L 65/80 (2022.01)

G06T 3/00 (2006.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 111010589 A

(56) 对比文件

CN 109637518 A, 2019.04.16

CN 109637518 A, 2019.04.16

CN 104618786 A, 2015.05.13

CN 109118562 A, 2019.01.01

CN 106653052 A, 2017.05.10

US 2006222071 A1, 2006.10.05

(43) 申请公布日 2020.04.14

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518000 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 朱绍明

审查员 周立秋

(74) 专利代理机构 北京派特恩知识产权代理有

限公司 11270

代理人 王姗姗 张颖玲

(51) Int. Cl.

H04N 21/233 (2011.01)

H04N 21/234 (2011.01)

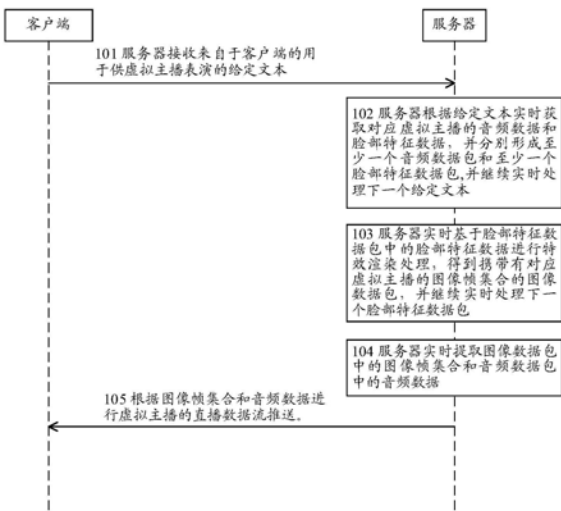
权利要求书3页 说明书22页 附图6页

(54) 发明名称

基于人工智能的直播方法、装置、设备及存储介质

(57) 摘要

本发明提供了一种基于人工智能的直播方法、装置、设备及存储介质;方法包括:接收用于供虚拟主播表演的给定文本,根据所述给定文本实时获取对应所述虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;实时基于所述脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包;实时提取所述图像数据包中的图像帧集合和所述音频数据包中的音频数据;根据所述图像帧集合和所述音频数据进行所述虚拟主播的直播数据流推送,通过本发明,能够有效提高直播的实时性和流畅性。



1. 一种基于人工智能的直播方法,其特征在于,所述方法包括:

接收用于供虚拟主播表演的给定文本,根据所述给定文本实时获取对应所述虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;

当形成针对所述给定文本的至少一个脸部特征数据包中的第一个脸部特征数据包时,实时基于所述第一个脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的第一个图像数据包;

当形成针对所述给定文本的至少一个音频数据包中的第一个音频数据包时,实时提取所述第一个音频数据包中的音频数据;

当形成携带有对应所述虚拟主播的图像帧集合的第一个图像数据包时,实时提取所述第一个图像数据包中的图像帧集合;

当从所述给定文本获取到所述第一个图像数据包所耗费的时间大于从给定文本获取到所述第一个音频数据包所耗费的时间,且提取到所述第一个音频数据包中的音频数据时,将所提取到的音频数据实时推送到直播客户端,并将后续提取到的音频数据包中的音频数据,实时推送至所述直播客户端,直至提取到所述第一个图像数据包中的图像帧集合,并将所提取到的图像帧集合实时推送至所述直播客户端;

当从所述给定文本获取到所述第一个图像数据包所耗费的时间小于从给定文本获取到所述第一个音频数据包所耗费的时间,且提取到所述第一个图像数据包中的图像帧集合时,将所提取到的图像帧集合时推送到直播客户端,并将后续提取到的图像数据包中的图像帧集合,实时推送至所述直播客户端,直至提取到所述第一个音频数据包中的音频数据,并将所提取到的音频数据实时推送至所述直播客户端。

2. 根据权利要求1所述的方法,其特征在于,所述根据所述给定文本实时获取对应所述虚拟主播的音频数据,包括:

当接收到所述给定文本时,将所述给定文本划分为至少两个语段,并实时将所述语段转化为对应所述语段的词向量;

对所述词向量进行编码处理以及解码处理,得到对应所述词向量的音频特征;

对所述音频特征进行合成处理,得到分别对应所述虚拟主播的各个语段的音频数据。

3. 根据权利要求2所述的方法,其特征在于,所述根据所述给定文本实时获取对应所述虚拟主播的脸部特征数据,包括:

预测所述虚拟主播的对应各个语段的嘴部关键点,并对所述嘴部关键点进行归一化处理,以使所述嘴部关键点适应于标准面部模板;

对经过归一化处理的嘴部关键点进行降维处理,得到对应所述虚拟主播的各个语段的口型特征数据;

对所述语段进行语义分析,得到所述语段所表征的语义;

根据所述语段所表征的语义,确定与所述语义匹配的脸部表情特征数据,并结合所述口型特征数据以及所述脸部表情特征数据形成对应所述虚拟主播的各个语段的脸部特征数据;

所述分别形成至少一个音频数据包和至少一个脸部特征数据包,包括:

基于对应于同一语段的脸部特征数据和音频数据,分别形成相互匹配的音频数据包和

脸部特征数据包。

4. 根据权利要求3所述的方法,其特征在于,在分别形成相互匹配的音频数据包和脸部特征数据包之前,所述方法还包括:

确定所述音频数据所表征的音素顺序;

基于所述音素顺序,确定携带有口型特征数据的脸部特征数据的顺序,以使所述脸部特征数据的顺序所表征的口型变化和所述音素顺序相匹配。

5. 根据权利要求3所述的方法,其特征在于,所述实时基于所述脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包,包括:

将所述音频数据的播放时长以及帧率参数的乘积确定为对应所述音频数据的图像帧数目;

实时提取所述脸部特征数据包中的脸部特征数据,并基于所述脸部特征数据进行特效渲染处理,得到对应所述脸部特征数据的脸部图像;

对所述脸部图像和背景图像进行合成处理,得到对应所述虚拟主播的图像帧,并将合成得到的多个所述图像帧组成为对应所述音频数据的图像帧集合;

其中,所述图像帧集合中图像帧的数目为所述图像帧数目;

对所述图像帧集合进行封包处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包。

6. 根据权利要求5所述的方法,其特征在于,

所述实时提取所述脸部特征数据包中的脸部特征数据,并基于所述脸部特征数据进行特效渲染处理,得到对应所述脸部特征数据的脸部图像,包括:

从所述脸部特征数据中,提取与所述图像帧数目相同数目的特征向量;

基于所述提取的特征向量进行特效渲染处理,得到与所述图像帧数目相同数目的脸部图像。

7. 根据权利要求6所述的方法,其特征在于,所述对所述脸部图像和背景图像进行合成处理,包括:

对所述脸部特征数据包进行解析处理,得到所述脸部特征数据包中所包含的脸部特征数据,并确定对应所述脸部特征数据的语段;

基于所述语段所表征的语义,提取对应所述语义的至少一组背景图像;

其中,每组背景图像中包括所述虚拟主播的模型、动作以及所处的直播场景;

将所述脸部图像与所述至少一组背景图像进行合成,得到所述虚拟主播的图像帧,其中,所述虚拟主播处于所述直播场景中、且具备所述动作以及所述脸部图像中相应的脸部表情和口型,所述动作的组数和所述背景图像的组数相同。

8. 根据权利要求1-7中任一项所述的方法,其特征在于,所述方法还包括:

响应于客户端发送的直播请求,为所述直播请求分配流媒体播放地址,并将所述流媒体播放地址返回至所述客户端;

所述根据所述图像帧集合和所述音频数据进行所述虚拟主播的直播数据流推送,包括:

将所述图像帧集合与所述音频数据依次实时推送至对应所述流媒体播放地址的流媒

体接口。

9. 一种基于人工智能的直播装置,其特征在于,所述装置包括:

文本转语音请求模块,用于接收用于供虚拟主播表演的给定文本,根据所述给定文本实时获取对应所述虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;

渲染模块,用于当形成针对所述给定文本的至少一个脸部特征数据包中的第一个脸部特征数据包时,实时基于所述第一个脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的第一个图像数据包;

视频推流模块,用于当形成针对所述给定文本的至少一个音频数据包中的第一个音频数据包时,实时提取所述第一个音频数据包中的音频数据;当形成携带有对应所述虚拟主播的图像帧集合的第一个图像数据包时,实时提取所述第一个图像数据包中的图像帧集合;当从所述给定文本获取到所述第一个图像数据包所耗费的时间大于从给定文本获取到所述第一个音频数据包所耗费的时间,且提取到所述第一个音频数据包中的音频数据时,将所提取到的音频数据实时推送到直播客户端,并将后续提取到的音频数据包中的音频数据,实时推送至所述直播客户端,直至提取到所述第一个图像数据包中的图像帧集合,并将所提取到的图像帧集合实时推送至所述直播客户端;当从所述给定文本获取到所述第一个图像数据包所耗费的时间小于从给定文本获取到所述第一个音频数据包所耗费的时间,且提取到所述第一个图像数据包中的图像帧集合时,将所提取到的图像帧集合时推送到直播客户端,并将后续提取到的图像数据包中的图像帧集合,实时推送至所述直播客户端,直至提取到所述第一个音频数据包中的音频数据,并将所提取到的音频数据实时推送至所述直播客户端。

10. 一种基于人工智能的直播设备,其特征在于,所述设备包括:

存储器,用于存储可执行指令;

处理器,用于执行所述存储器中存储的可执行指令时,实现权利要求1至8任一项所述的基于人工智能的直播方法。

11. 一种计算机可读存储介质,其特征在于,存储有可执行指令,用于被处理器执行时,实现权利要求1至8任一项所述的基于人工智能的直播方法。

基于人工智能的直播方法、装置、设备及存储介质

技术领域

[0001] 本发明涉及人工智能技术,尤其涉及一种基于人工智能的直播方法、装置、设备及存储介质。

背景技术

[0002] 人工智能(AI,Artificial Intelligence)是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法和技术及应用系统。

[0003] 随着通信技术的发展,网络通信带宽得到很大的提高,视频直播技术也日趋成熟,在各个方面得到应用,同时随着人工智能技术的发展,文本转语音技术与图像合成技术也成为研究热点,视频直播技术与人工智能技术的结合可以在许多地方发挥作用,比如代替真人进行新闻虚拟播报、代替游戏主播进行游戏虚拟解说等,具有广阔的应用前景,但是,在虚拟视频直播技术中,音频与图像的生成需要耗费大量的计算时间,难以保证虚拟视频直播的实时性。

发明内容

[0004] 本发明实施例提供一种基于人工智能的直播方法、装置、设备及存储介质,能够有效提高直播的实时性和流畅性。

[0005] 本发明实施例的技术方案是这样实现的:

[0006] 本发明实施例提供一种基于人工智能的直播方法,包括:

[0007] 接收用于供虚拟主播表演的给定文本,根据所述给定文本实时获取对应所述虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;

[0008] 实时基于所述脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包;

[0009] 实时提取所述图像数据包中的图像帧集合和所述音频数据包中的音频数据;

[0010] 根据所述图像帧集合和所述音频数据进行所述虚拟主播的直播数据流推送。

[0011] 本发明实施例提供一种基于人工智能的直播装置,包括:

[0012] 文本转语音请求模块,用于接收用于供虚拟主播表演的给定文本,根据所述给定文本实时获取对应所述虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;

[0013] 渲染模块,用于实时基于所述脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包;

[0014] 视频推流模块,用于实时提取所述图像数据包中的图像帧集合和所述音频数据包中的音频数据;根据所述图像帧集合和所述音频数据进行所述虚拟主播的直播数据流推送。

[0015] 在上述方案中,所述渲染模块,还用于:

[0016] 当形成针对所述给定文本的至少一个脸部特征数据包中的第一个脸部特征数据包时,实时基于所述第一个脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应所述虚拟主播的图像帧集合的第一个图像数据包;

[0017] 所述视频推流模块,还用于:

[0018] 当形成针对所述给定文本的至少一个音频数据包中的第一个音频数据包时,实时提取所述第一个音频数据包中的音频数据,以进行音频数据的推送;

[0019] 当形成携带有对应所述虚拟主播的图像帧集合的第一个图像数据包时,实时提取所述第一个图像数据包中的图像帧集合,以进行图像帧集合的推送。

[0020] 在上述方案中,当从所述给定文本获取到所述第一个图像数据包所耗费的时间大于从给定文本获取到所述第一个音频数据包所耗费的时间时,所述视频推流模块,还用于:

[0021] 当提取到所述第一个音频数据包中的音频数据时,将所提取到的音频数据实时推送到直播客户端,并将后续提取到的音频数据包中的音频数据,实时推送至所述直播客户端,直至

[0022] 提取到所述第一个图像数据包中的图像帧集合,并将所提取到的图像帧集合实时推送至所述直播客户端。

[0023] 在上述方案中,当从所述给定文本获取到所述第一个图像数据包所耗费的时间小于从给定文本获取到所述第一个音频数据包所耗费的时间时,所述视频推流模块,还用于:

[0024] 当提取到所述第一个图像数据包中的图像帧集合时,将所提取到的图像帧集合时推送到直播客户端,并将后续提取到的图像数据包中的图像帧集合,实时推送至所述直播客户端,直至

[0025] 提取到所述第一个音频数据包中的音频数据,并将所提取到的音频数据实时推送至所述直播客户端。

[0026] 在上述方案中,所述文本转语音请求模块,还用于:

[0027] 当接收到所述给定文本时,将所述给定文本划分为至少两个语段,并实时将所述语段转化为对应所述语段的词向量;

[0028] 对所述词向量进行编码处理以及解码处理,得到对应所述词向量的音频特征;

[0029] 对所述音频特征进行合成处理,得到分别对应所述虚拟主播的各个语段的音频数据。

[0030] 在上述方案中,所述文本转语音请求模块,还用于:

[0031] 预测所述虚拟主播的对应各个语段的嘴部关键点,并对所述嘴部关键点进行归一化处理,以使所述嘴部关键点适应于标准面部模板;

[0032] 对经过归一化处理的嘴部关键点进行降维处理,得到对应所述虚拟主播的各个语段的口型特征数据;

[0033] 对所述语段进行语义分析,得到所述语段所表征的语义;

[0034] 根据所述语段所表征的语义,确定与所述语义匹配的脸部表情特征数据,并结合所述口型特征数据以及所述脸部表情特征数据形成对应所述虚拟主播的各个语段的脸部特征数据;

[0035] 基于对应于同一语段的脸部特征数据和音频数据,分别形成所述音频数据包和所

述脸部特征数据包。

[0036] 在上述方案中,所述文本转语音请求模块,还用于:

[0037] 在对所述嘴部关键点进行归一化处理之前,

[0038] 向区块链网络发送查询交易,其中,所述查询交易指示了用于查询所述区块链网络中账本的智能合约,以使

[0039] 所述区块链网络中的共识节点通过执行所述智能合约的方式来查询所述账本,得到所述账本中存储的所述标准面部模板;或者

[0040] 根据所述标准面部模板的标识,从标准面部模板数据库中查询所述标识对应的所述标准面部模板,并确定所述查询到的标准面部模板的哈希值;

[0041] 从区块链网络中查询所述标识对应的哈希值,当所查询到的哈希值与所述确定的哈希值一致时,确定所述查询到的标准面部模板未经篡改。

[0042] 在上述方案中,所述文本转语音请求模块,还用于:

[0043] 在分别形成至少一个音频数据包和至少一个脸部特征数据包之前,

[0044] 将所述给定文本发送至区块链网络,以使

[0045] 所述区块链网络中的共识节点通过执行智能合约的方式来对所述给定文本进行合规性检查;

[0046] 当接收到超过预设数目的共识节点返回的合规性确认时,确定所述给定文本通过所述合规性检查。

[0047] 在上述方案中,所述文本转语音请求模块,还用于:

[0048] 在分别形成相互匹配的音频数据包和脸部特征数据包之前,

[0049] 确定所述音频数据所表征的音素顺序;

[0050] 基于所述音素顺序,确定携带有口型特征数据的脸部特征数据的顺序,以使所述脸部特征数据的顺序所表征的口型变化和所述音素顺序相匹配。

[0051] 在上述方案中,所述渲染模块,还用于:

[0052] 将所述音频数据的播放时长以及帧率参数的乘积确定为对应所述音频数据的图像帧数目;

[0053] 实时提取所述脸部特征数据包中的脸部特征数据,并基于所述脸部特征数据进行特效渲染处理,得到对应所述脸部特征数据的脸部图像;

[0054] 对所述脸部图像和背景图像进行合成处理,得到对应所述虚拟主播的图像帧,并将合成得到的多个所述图像帧组成为对应所述音频数据的图像帧集合;

[0055] 其中,所述图像帧集合中图像帧的数目为所述图像帧数目;

[0056] 对所述图像帧集合进行封包处理,得到携带有对应所述虚拟主播的图像帧集合的图像数据包。

[0057] 在上述方案中,所述渲染模块,还用于:

[0058] 从所述脸部特征数据中,提取与所述图像帧数目相同数目的特征向量;

[0059] 基于所述提取的特征向量进行特效渲染处理,得到与所述图像帧数目相同数目的脸部图像。

[0060] 在上述方案中,所述渲染模块,还用于:

[0061] 对所述脸部特征数据包进行解析处理,得到所述脸部特征数据包中所包含的脸部

特征数据,并确定对应所述脸部特征数据的语段;

[0062] 基于所述语段所表征的语义,提取对应所述语义的至少一组背景图像;

[0063] 其中,每组背景图像中包括所述虚拟主播的模型、动作以及所处的直播场景;

[0064] 将所述脸部图像与所述至少一组背景图像进行合成,得到所述虚拟主播的图像帧,其中,所述虚拟主播处于所述直播场景中、且具备所述动作以及所述脸部图像中相应的脸部表情和口型,所述动作的组数和所述背景图像的组数相同。

[0065] 在上述方案中,所述渲染模块,还用于:

[0066] 向区块链网络发送查询交易,其中,所述查询交易指示了用于查询所述区块链网络中账本的智能合约以及对应所述语义的查询参数,以使

[0067] 所述区块链网络中的共识节点通过执行所述智能合约的方式来查询所述账本,得到所述账本中符合所述查询参数的背景图像;或者

[0068] 根据所述语义从背景图像数据库中查询所述语义对应的背景图像,并确定所述查询到的背景图像的哈希值;

[0069] 从区块链网络中查询所述语义对应的哈希值,当所查询到的哈希值与所述确定的哈希值一致时,确定所述查询到的背景图像未经篡改。

[0070] 在上述方案中,所述装置还包括:

[0071] 流媒体服务模块,用于:

[0072] 响应于客户端发送的直播请求,为所述直播请求分配流媒体播放地址,并将所述流媒体播放地址返回至所述客户端;

[0073] 所述视频推流模块,还用于:

[0074] 将所述图像帧集合与所述音频数据依次实时推送至对应所述流媒体播放地址的流媒体接口。

[0075] 本发明实施例提供了一种基于人工智能的直播方法,所述方法包括:

[0076] 接收虚拟主播的直播选择指令;

[0077] 根据所述直播选择指令获取所述虚拟主播对应的图像帧集合和音频数据;

[0078] 其中,所述图像帧集合和所述音频数据对应于所述虚拟主播的给定文本;

[0079] 对获取的图像帧集合和音频数据进行同步处理,并实时呈现所述虚拟主播的图像帧,以及对应的音频数据。

[0080] 本发明实施例提供一种基于人工智能的直播装置,所述装置包括:

[0081] 指令接收模块,用于接收虚拟主播的直播选择指令;

[0082] 数据获取模块,用于根据所述直播选择指令获取所述虚拟主播对应的图像帧集合和音频数据;

[0083] 其中,所述图像帧集合和所述音频数据对应于所述虚拟主播的给定文本;

[0084] 呈现模块,用于对获取的图像帧集合和音频数据进行同步处理,并实时呈现所述虚拟主播的图像帧,以及对应的音频数据。

[0085] 本发明实施例提供一种基于人工智能的直播设备,包括:

[0086] 存储器,用于存储可执行指令;

[0087] 处理器,用于执行所述存储器中存储的可执行指令时,实现本发明实施例提供的基于人工智能的直播方法。

[0088] 本发明实施例提供一种存储介质,存储有可执行指令,用于引起处理器执行时,实现本发明实施例提供的基于人工智能的直播方法。

[0089] 本发明实施例具有以下有益效果:

[0090] 通过本发明实施例提供的基于人工智能的直播方法,通过特征提取、特效渲染以及推流并行处理的方式和音频数据和脸部特征数据分离的方式,增强虚拟视频直播的实时性,提高直播视频流畅性。

附图说明

[0091] 图1是本发明实施例提供的基于人工智能的直播系统100架构的一个可选的结构示意图;

[0092] 图2是本发明实施例提供的基于人工智能的直播设备的一个可选的结构示意图;

[0093] 图3A-3D是本发明实施例提供的基于人工智能的直播方法的一个可选的流程示意图;

[0094] 图4是本发明实施例提供的基于人工智能的直播系统架构的一个可选的结构示意图;

[0095] 图5是本发明实施例提供的基于人工智能的直播系统架构的一个可选的结构示意图;

[0096] 图6是本发明实施例提供的基于人工智能的直播方法的实现架构图;

[0097] 图7A-7B是本发明实施例提供的虚拟视频直播服务的整体框架示意图;

[0098] 图8是本发明实施例提供的虚拟视频推流服务的实现框架。

具体实施方式

[0099] 为了使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明作进一步地详细描述,所描述的实施例不应视为对本发明的限制,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0100] 在以下的描述中,涉及到“一些实施例”,其描述了所有可能实施例的子集,但是可以理解,“一些实施例”可以是所有可能实施例的相同子集或不同子集,并且可以在不冲突的情况下相互结合。

[0101] 除非另有定义,本文所使用的所有的技术和科学术语与属于本发明的技术领域的技术人员通常理解的含义相同。本文中所使用的术语只是为了描述本发明实施例的目的,不是旨在限制本发明。

[0102] 对本发明实施例进行进一步详细说明之前,对本发明实施例中涉及的名词和术语进行说明,本发明实施例中涉及的名词和术语适用于如下的解释。

[0103] 1) 文本转语音(Text To Speech):通过人工智能(例如,神经网络),把文本智能地转化为自然语音流。

[0104] 2) 虚拟视频直播:由虚拟角色所合成的视频直播,由虚拟角色担任虚拟主播的角色,在新闻、游戏等场景下进行直播。

[0105] 3) 交易(Transaction),等同于计算机术语“事务”,交易包括了需要提交到区块链网络执行的操作,并非单指商业语境中的交易,鉴于在区块链技术中约定俗成地使用了“交

易”这一术语,本发明实施例遵循了这一习惯。

[0106] 例如,部署 (Deploy) 交易用于向区块链网络中的节点安装指定的智能合约并准备好被调用;调用 (Invoke) 交易用于通过调用智能合约在区块链中追加交易的记录,并对区块链的状态数据库进行操作,包括更新操作 (包括增加、删除和修改状态数据库中的键值对) 和查询操作 (即查询状态数据库中的键值对)。

[0107] 4) 区块链 (Blockchain), 是由区块 (Block) 形成的加密的、链式的交易的存储结构。

[0108] 5) 区块链网络 (Blockchain Network), 通过共识的方式将新区块纳入区块链的一系列的节点的集合。

[0109] 6) 账本 (Ledger), 是区块链 (也称为账本数据) 和与区块链同步的状态数据库的统称。其中,区块链是以文件系统中的文件的形式来记录交易;状态数据库是以不同类型的键 (Key) 值 (Value) 对的形式来记录区块链中的交易,用于支持对区块链中交易的快速查询。

[0110] 7) 智能合约 (Smart Contracts), 也称为链码 (Chaincode) 或应用代码,部署在区块链网络的节点中的程序,节点执行接收的交易中所调用的智能合约,来对状态数据库的键值对数据进行更新或查询的操作。

[0111] 8) 共识 (Consensus), 是区块链网络中的一个过程,用于在涉及的多个节点之间对区块中的交易达成一致,达成一致的区块将被追加到区块链的尾部,实现共识的机制包括工作量证明 (PoW, Proof of Work)、权益证明 (PoS, Proof of Stake)、股份授权证明 (DPoS, Delegated Proof-of-Stake)、消逝时间量证明 (PoET, Proof of Elapsed Time) 等。

[0112] 9) 人工智能 (AI, Artificial Intelligence) 是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法和技术及应用系统。

[0113] 随着通信技术的发展,网络通信带宽得到很大的提高,视频直播技术也日趋成熟,在各个方面得到应用。同时随着人工智能技术的发展,文本转语音技术与图像合成技术也成为了人们的研究热点,发展迅速。视频直播技术与人工智能技术的结合可以在许多地方发挥作用,比如代替真人进行新闻虚拟播报、代替游戏主播进行游戏虚拟解说等,具有广阔的应用前景。在虚拟视频直播技术中,音频与图像的生成需要耗费大量的计算时间,为保证虚拟视频直播的实时性,虚拟视频实时推流的实现成为影响最终直播视频质量的重要因素,相关技术中的直播方法大都是针对已有稳定的音频、图像数据输入 (如本地视频推流),或者是能够快速获取到音频、图像数据 (如摄像头获取数据) 等应用场景,无法在虚拟视频直播中得到很好应用,具体而言,上述相关技术中的技术方案在实际应用中存在以下缺陷:过分依赖稳定的数据源,且在需要较大计算力的场景下,相关技术中的技术方案无法适用,基于此,本发明实施例提出了一种基于人工智能的直播方法,通过并行处理以及数据分离的方式,尽可能让图像帧数据与音频数据均匀地到达视频推流模块,可以理解为视频由两条时间线组成,一条是音频时间线,一条是图像时间线,由客户端对接收到的两条时间线的数据进行同步处理后,直播视频才会播放,因此,在渲染模块渲染图像期间,视频推流模块可以等待文本转语音请求模块获取音频数据包,以此减少两条时间线的互相等待,解决在需要耗费较大计算力用于获取音视频数据的前提下,虚拟视频实时推流的问题,有效避免不稳定数据源对实时推流的干扰,增强虚拟视频直播的实时性,提高直播视频流畅性。

[0114] 本发明实施例提供一种基于人工智能的直播方法、装置、设备和存储介质,能够很好地应对数据源不稳定的问题,有效提高虚拟视频推流的实时性,下面说明本发明实施例提供的基于人工智能的直播设备的示例性应用,本发明实施例提供的设备可以实施为服务器。下面,将说明设备实施为服务器时示例性应用。

[0115] 参见图1,图1是本发明实施例提供的基于人工智能的直播系统100的一个可选的架构示意图,终端400(示例性示出了终端400-1和终端400-2)通过网络300连接服务器200,网络300可以是广域网或者局域网,又或者是二者的组合。终端400-1和终端400-2上分别设置有直播客户端410-1和410-2,这里的直播客户端具备提供直播内容的功能以及对直播视频进行呈现的功能,服务器200中包括文本转语音请求模块2551、渲染模块2552和视频推流模块2553,文本转语音请求模块2551基于终端400-1和400-2返回的给定文本,流式获取到对应的脸部特征数据与音频数据,且将脸部特征数据和音频数据分离成两个包,将每次获取到的脸部特征数据包推送给渲染模块2552,将每次获取到的音频数据包推送给视频推流模块2553,渲染模块2552对每次获取到的脸部特征数据通过三维渲染处理,获取虚拟角色的表情图像,并推送给视频推流模块2553,视频推流模块2553异步接收到音频数据包与表情图像数据包,进行视频合成,并推流给客户端,文本转语音请求模块2551、渲染模块2552和视频推流模块2553相互独立并行协助,将音频数据、脸部特征数据相互分离并独立分流处理的方式,这里的脸部特征数据中可以包括口型特征数据,从而将音频数据和口型特征数据相互分离并独立分流处理,能够将文本数据自动实时合成为视频推送到直播客户端410-1和410-2,在渲染模块2552对每次获取到的脸部特征数据进行三维渲染处理的过程中,还可以调用基于人工智能的直播系统100中的背景图像数据库500中的形象资源,以实现更加丰富的视频直播呈现。

[0116] 参见图2,图2是本发明实施例提供的基于人工智能的直播服务器200的结构示意图,图2所示的服务器200包括:至少一个处理器210、存储器250、至少一个网络接口220和用户接口230。服务器200中的各个组件通过总线系统240耦合在一起。可理解,总线系统240用于实现这些组件之间的连接通信。总线系统240除包括数据总线之外,还包括电源总线、控制总线和状态信号总线。但是为了清楚说明起见,在图2中将各种总线都标为总线系统240。

[0117] 处理器210可以是一种集成电路芯片,具有信号的处理能力,例如通用处理器、数字信号处理器(DSP, Digital Signal Processor),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,其中,通用处理器可以是微处理器或者任何常规的处理器等。

[0118] 用户接口230包括使得能够呈现媒体内容的一个或多个输出装置231,包括一个或多个扬声器和/或一个或多个视觉显示屏。用户接口230还包括一个或多个输入装置232,包括有助于用户输入的用户接口部件,比如键盘、鼠标、麦克风、触屏显示屏、摄像头、其他输入按钮和控件。

[0119] 存储器250可以是可移除的,不可移除的或其组合。示例性的硬件设备包括固态存储器,硬盘驱动器,光盘驱动器等。存储器250可选地包括在物理位置上远离处理器210的一个或多个存储设备。

[0120] 存储器250包括易失性存储器或非易失性存储器,也可包括易失性和非易失性存储器两者。非易失性存储器可以是只读存储器(ROM, Read Only Memory),易失性存储器可

以是随机存取存储器 (RAM, Random Access Memory)。本发明实施例描述的存储器250旨在包括任意适合类型的存储器。

[0121] 在一些实施例中,存储器250能够存储数据以支持各种操作,这些数据的示例包括程序、模块和数据结构或者其子集或超集,下面示例性说明。

[0122] 操作系统251,包括用于处理各种基本系统服务和执行硬件相关任务的系统程序,例如框架层、核心库层、驱动层等,用于实现各种基础业务以及处理基于硬件的任务;

[0123] 网络通信模块252,用于经由一个或多个(有线或无线)网络接口220到达其他计算设备,示例性的网络接口220包括:蓝牙、无线相容性认证(WiFi)、和通用串行总线(USB, Universal Serial Bus)等;

[0124] 呈现模块253,用于经由一个或多个与用户接口230相关联的输出装置231(例如,显示屏、扬声器等)使得能够呈现信息(例如,用于操作外围设备和显示内容和信息的用户接口);

[0125] 输入处理模块254,用于对一个或多个来自一个或多个输入装置232之一的一个或多个用户输入或互动进行检测以及翻译所检测的输入或互动。

[0126] 在一些实施例中,本发明实施例提供的装置可以采用软件方式实现,图2示出了存储在存储器250中的基于人工智能的直播装置255,其可以是程序和插件等形式的软件,包括以下软件模块:文本转语音请求模块2551、渲染模块2552、视频推流模块2553和流媒体服务模块2554,这些模块是逻辑上的,因此根据所实现的功能可以进行任意的组合或进一步拆分,将在下文中说明各个模块的功能。

[0127] 在另一些实施例中,本发明实施例提供的基于人工智能的直播装置255可以采用硬件方式实现,作为示例,本发明实施例提供的基于人工智能的直播装置255可以是采用硬件译码处理器形式的处理器,其被编程以执行本发明实施例提供的基于人工智能的直播方法,例如,硬件译码处理器形式的处理器可以采用一个或多个应用专用集成电路(ASIC, Application Specific Integrated Circuit)、DSP、可编程逻辑器件(PLD, Programmable Logic Device)、复杂可编程逻辑器件(CPLD, Complex Programmable Logic Device)、现场可编程门阵列(FPGA, Field-Programmable Gate Array)或其他电子元件。

[0128] 下面将结合本发明实施例提供的终端的示例性应用和实施,说明本发明实施例提供的基于人工智能的直播方法。

[0129] 参见图3A,图3A是本发明实施例提供的基于人工智能的直播方法的一个可选的流程示意图,将结合图3A示出的步骤101-105进行说明。

[0130] 在步骤101中,服务器接收来自于客户端的用于供虚拟主播表演的给定文本。

[0131] 在一些实施例中,客户端将用于供虚拟主播表演的给定文本发送至服务器,这里的给定文本可以是用户输入的,还可以是通过网络爬去的,这里的给定文本来源于直播内容提供商,这里的给定文本还可以来源于观看直播的用户,虚拟主播可以实时基于观看直播的用户通过客户端输入的文本进行表演,另外在直播过程中,观看直播的用户可以和虚拟主播进行互动,在虚拟主播进行直播表演的过程中,用户输入互动文本后,客户端可以根据互动文本查询对应的话术,生成对应话术的给定文本,或者将互动文本发送至服务器,由服务器根据互动文本查询对应的话术,生成对应话术的给定文本,使得虚拟主播根据对应话术的给定文本进行直播表演,实现了观看用户和虚拟主播之间的互动。

[0132] 在步骤102中,服务器根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包,并继续实时处理下一个给定文本。

[0133] 在一些实施例中,在步骤102中,服务器根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包,并继续实时处理下一个给定文本,可以由文本转语音请求模块请求文本转语音服务来获取音频数据包和脸部特征数据包,文本转语音服务用于根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,以分别形成至少一个音频数据包和至少一个脸部特征数据包并继续实时处理下一个给定文本,或者也可以由文本转语音模块直接根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,以分别形成至少一个音频数据包和至少一个脸部特征数据包并继续实时处理下一个给定文本,这里实时处理下一个给定文本意味着,文本转语音请求模块不会等待完成整个推流过程之后才处理下一个给定文本,即只要处理完一个给定文本之后,就开始处理下一个给定文本,下一个给定文本可以是来自于同一个客户端的,也可以是来自于不同客户端的,给定文本可以是针对同一场直播的,接收到的不同给定文本也可以是针对不同直播的,对应形成的音频数据包和脸部特征数据包中会携带有直播标识,不同直播对应应有不同直播标识,进一步的,文本转语音请求模块基于处理器所配置的处理能力,该模块自身也会具有对应于处理器配置的处理能的并行处理能力,即同时处理多个给定文本,这多个给定文本可以是来自于不同客户端或者是针对不同直播的给定文本。在处理过程中,还可以对给定文本进行划分,在划分之后再生成对应各个语段的音频数据包和脸部特征数据包,每处理完一个语段以生成一个音频数据包和一个脸部特征数据包之后,就紧接着处理下一个语段以产生下一个音频数据包和下一个脸部特征数据包,或者基于处理器所配置的并行处理能力来同时对多个语段产生多个音频数据包和脸部特征数据包。

[0134] 参见图3B,图3B是本发明实施例提供的基于人工智能的直播方法的一个可选流程示意图,步骤102中服务器根据给定文本实时获取对应虚拟主播的音频数据,可以具体通过步骤1021-1023实现。

[0135] 在步骤1021中,当接收到给定文本时,服务器将给定文本划分为至少两个语段,并实时将语段转化为对应语段的词向量。

[0136] 在步骤1022中,服务器对词向量进行编码处理以及解码处理,得到对应词向量的音频特征。

[0137] 在步骤1023中,服务器对音频特征进行合成处理,得到分别对应虚拟主播的各个语段的音频数据。

[0138] 在一些实施例中,可以构建一个端到端的深度学习文本转语音模型,直接用深度学习的方法训练出一个文本转语音模型,模型训练完成后,针对给定文本,模型就能生成对应的音频,服务器中的文本转语音请求模块可以将给定文本划分为至少两个语段,并将语段实时转化为对应语段的词向量,对于文本转语音模型而言,需要先把文本转化为对应的词向量,这里的给定文本可以是一句话,将给定文本进行语段划分,对语段进行分词,以确定对应各个词的词向量,可以通过查询字典的方式确定词向量,使用字典下标作为字典中每一个字对应的标识,然后每个语段就可以通过遍历字典转化成其对应的词向量;对词向

量进行编码处理得到中间语义表示,再通过解码处理得到对应词向量的音频特征,基于各音频合成算法,对最终得到的音频特征进行合成,得到对应虚拟主播的各个语段的音频数据。

[0139] 在一些实施例中,步骤102中服务器根据给定文本实时获取对应虚拟主播的脸部特征数据,可以根据以下技术方案实现,预测虚拟主播的对应各个语段的嘴部关键点,并对嘴部关键点进行归一化处理,以使嘴部关键点适应于标准面部模板;对经过归一化处理的嘴部关键点进行降维处理,得到对应虚拟主播的对应各个语段的口型特征数据;对语段进行语义分析,得到语段所表征的语义;根据语段所表征的语义,确定与语义匹配的脸部表情特征数据,并结合口型特征数据以及脸部表情特征数据形成对应虚拟主播的各个语段的脸部特征数据;在执行上述技术方案之后,服务器基于对应于同一语段的脸部特征数据和音频数据,分别形成相互匹配的音频数据包和脸部特征数据包。

[0140] 在一些实施例中,根据给定文本的各个语段,预测口型的表达,这里使用谱特征来表示音频,为了计算口型表示,需要从面部提取嘴部关键点,并做归一化处理从而不受图像大小、面部位置、面部旋转、面部大小的影响,归一化在此过程中非常重要,因为,它能使生成的嘴部关键点兼容于任何图像,从而归一化处理后的嘴部关键点能够适应于标准面部模板,这里的标准面部模板是虚拟主播的标准面部模板,不同的虚拟主播角色对应于不同的标准面部模板,虚拟主播可以为动物,则标准面部模板可以是动物头像,虚拟主播也可以是卡通人物,则标准面部模板可以是卡通人物头像,对归一化处理过的嘴部关键点进行降维处理,对嘴部关键点特征去相关,只使用最重要部分作为口型的表征,对于脸部特征数据而言,除了口型特征数据还有脸部表情特征数据,口型特征数据是和音频数据相关,而脸部表情特征数据则是和给定文本的语段相关,对给定文本的语段进行语义分析,对于给定文本的语段所表征的语义而言,虚拟主播会呈现适应于语义的脸部表情特征数据,结合口型特征数据和脸部表情特征数据形成虚拟主播的脸部特征数据。

[0141] 在一些实施例中,将给定文本划分为至少两个语段,例如,针对给定文本“大家好,我是主播小A”而言,可以将其划分为两个语段“大家好”和“我是主播小A”,或者三个语段“大家好”“我是”和“主播小A”,对应于不同的语段,基于语段,生成“大家好,我是主播小A”中各个语段的音频数据,即“大家好”的音频数据和“我是主播小A”的音频数据,同样是基于划分后得到的语段,得到对应“大家好”的脸部特征数据和对应“我是主播小A”的脸部特征数据,“大家好”的脸部特征数据中包括对应“大家好”的口型特征数据和脸部表情特征数据,“我是主播小A”的脸部特征数据中包括对应“我是主播小A”的口型特征数据和脸部表情特征数据。

[0142] 在一些实施例中,在对嘴部关键点进行归一化处理之前,可以通过以下技术方案获取标准面部模板,向区块链网络发送查询交易,其中,查询交易指示了用于查询区块链网络中账本的智能合约,以使区块链网络中的共识节点通过执行智能合约的方式来查询账本,得到账本中存储的标准面部模板;或者根据标准面部模板的标识,从标准面部模板数据库中查询标识对应的标准面部模板,并确定查询到的标准面部模板的哈希值;从区块链网络中查询标识对应的哈希值,当所查询到的哈希值与确定的哈希值一致时,确定查询到的标准面部模板未经篡改。

[0143] 在一些实施例中,任意一个用户都可以通过直播客户端向区块链网络或者标准面

部模板数据库提交标准面部模板,向区块链网络提交标准面部模板可以通过查询交易的方式实现,查询交易通过执行对应查询的智能合约来查询共识节点中的账本,得到账本中存储的标准面部模板,参见图4,图4是本发明实施例提供的基于人工智能的直播系统100的一个可选的架构示意图,终端400向区块链网络上传标准面部模板,标准面部模板被共识节点执行智能合约进行共识通过后,记录到各个记账节点的账本中,服务器200可以从区块链网络获取标准面部模板。终端400上链存储标准面部模板、服务器200上链存储标准面部模板、查询标准面部模板和更新标准面部模板都是通过提交交易的方式实现的,交易中携带实现数据操作的智能合约的调用信息,通过共识节点调用智能合约来执行,共识节点执行智能合约的结果需要获得其他节点的共识才能有效,从而保证了数据可靠性和一致性。终端400和服务器200到认证中心注册获取证书(其中包括业务主体的公钥、以及认证中心对业务主体的公钥和身份信息签署的数字签名),用来与终端400或服务器200针对交易的数字签名一起附加到交易中,并被发送到区块链网络,以供区块链网络从交易中取出数字证书和签名,验证消息的可靠性(未经篡改)和发送消息的业务主体的身份信息,区块链网络会根据身份进行验证,例如是否具有发起交易的权限。共识功能和排序功能是共识节点能够实现的功能,区块链和状态数据库是账本的两种维护数据的方式,上链标准面部模板的交易和查询标准面部模板的交易在区块链中有完整的记录;在状态数据库中仅记录部分数据。

[0144] 在一些实施例中,还可以从标准面部模板数据库中查询标准面部模板,查询依据是标准面部模板的标识,标识和标准面部模板一一对应,根据唯一标识查询到标准面部模板,并确定查询到的标准面部模板的哈希值,以对查询到的标准面部模板进行哈希验证,当查询到的哈希值和预置的确定的哈希值一致时,则查询到的标准面部模板未经篡改,从而,基于此标准面部模板对嘴部关键点进行归一化处理。

[0145] 在一些实施例中,用户上传的一些标准面部模板会存在暴力等不符合社会公序良俗的元素,因此,共识节点将调用对应合规性检查的智能合约来对标准面部模板进行合规性检查,智能合约可以采用人工智能模型来进行合规性检查,根据合规来训练人工智能模型,通过执行智能合约,以使用人工智能模型对输入的标准面部模板进行识别,以识别出不符合合规的标准面部模板,对于不符合合规的标准面部模板,将不会存储到区块链网络上。

[0146] 在一些实施例中,在形成至少一个音频数据包和至少一个脸部特征数据包之前,还可以执行以下技术方案,将给定文本发送至区块链网络,以使区块链网络中的共识节点通过执行智能合约的方式来对给定文本进行合规性检查;当接收到超过预设数目的共识节点返回的合规性确认时,确定给定文本通过合规性检查。

[0147] 在一些实施例中,除了对标准面部模板进行合规性检查之外,还可以对给定文本进行合规性检查,在形成至少一个音频数据包和至少一个脸部特征数据包之前,将给定文本发送到区块链网络进行合规性检查,通过共识节点对给定文本进行共识验证,该共识验证基于合规来进行,以实现合规性检查,当接收到超出数目阈值个共识节点返回的共识验证确认通过时,才确定给定文本通过合规性检查,可以进行后续处理,为了避免合规检查的延迟,可以在接收到一个共识节点的确认时就进行后续处理,当在后续的等待时间内没有接收到足够数量的共识节点返回的确认,则终止处理并返回原因,这里终止处理的时机可以出现在形成至少一个音频数据包和至少一个脸部特征数据包的时候,也可以是在进行特

效渲染处理的时候,以及进行视频推流的时候,原因可以是对给定文本的合规性检查不通过,要求客户端重新提交给定文本。

[0148] 在一些实施例中,在分别形成相互匹配的音频数据包和脸部特征数据包之前,可以确定脸部特征数据包中的脸部特征数据的顺序,通过以下技术方案实现,基于音素顺序,确定携带有口型特征数据的脸部特征数据的顺序,以使脸部特征数据的顺序所表征的口型变化和音素顺序相匹配。

[0149] 在一些实施例中,脸部特征数据包中的脸部特征数据本身是不具备顺序属性的,但是图像帧的呈现是具备顺序的,且图像帧是基于脸部特征数据得到的,因而脸部特征数据是具备顺序的,从口型的角度而言,口型变化和音频的输出应该保持一致,否则将会出现口型和音频不一致的情形,因而,确定音频数据中的音素顺序,基于音素顺序确定口型特征数据的顺序,由于口型特征数据、脸部特征数据和图像帧三者对应,因而,可以确定包含口型特征数据的脸部特征数据的顺序,对每个脸部特征数据进行顺序编号,使基于按顺序排列的脸部特征数据分别得到的图像帧所表征的口型变化和音素顺序相匹配,于是,在输出音频时,按照脸部特征数据编号的顺序来呈现图像帧。

[0150] 在步骤103中,服务器实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,并继续实时处理下一个脸部特征数据包。

[0151] 在一些实施例中,在步骤103中,服务器实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,并继续实时处理下一个脸部特征数据包,通过渲染模块实现,这里的实时处理下一个脸部特征数据包意味着,渲染模块不会等待完成整个推流过程之后才处理下一个脸部特征数据包,即只要处理完一个脸部特征数据包之后,就开始处理下一个脸部特征数据包,脸部特征数据包可以是针对同一场直播的,接收到的不同脸部特征数据包也可以是针对不同直播的,对应形成的图像数据包中会携带有直播标识,不同直播对应应有不同直播标识,进一步的,渲染模块基于处理器所配置的处理能力,该模块自身也会具有对应于处理器配置的处理能力的并行处理能力,即同时处理多个脸部特征数据包,这多个脸部特征数据包可以是针对不同直播的脸部特征数据包,除此之外,渲染模块的并行还可以体现在模块内部的流程处理,即将渲染模块所执行的生成脸部图像的过程和将脸部图像与背景图像合成的过程分成两个子模块处理,子模块之间的处理是独立并行的,每生成出一个脸部图像,合成子模块就将其与背景图像合成,在合成时,脸部图像生成子模块开始生成下一个脸部图像。

[0152] 在一些实施例中,除了将口型变化考虑到脸部特征数据的排序,即渲染阶段生成的图像帧的呈现顺序中,还可以在特效渲染阶段将背景图像中的动作也考虑到图像帧的排序中,对于同一个脸部特征数据包的背景图像而言,同一个脸部特征数据包的背景图像可以表征同一组动作,而一组动作的呈现也可以是动态的过程,将一组动作的背景图像按照包含口型特征数据的脸部图像的顺序进行排序,使得口型变化和动作的呈现均能够符合后续音频数据的输出,使得后续在进行视频推送时,图像帧和音频数据匹配且同步。

[0153] 参见图3C,基于图3A,图3C是本发明实施例提供的基于人工智能的直播方法的一个可选流程示意图,步骤103中服务器实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,可以具体通过步骤

1031-1034实现。

[0154] 在步骤1031中,服务器将音频数据的播放时长以及帧率参数的乘积,确定为对应音频数据的图像帧数目。

[0155] 在步骤1032中,服务器提取脸部特征数据包中的脸部特征数据,并基于脸部特征数据进行特效渲染处理,得到对应脸部特征数据的脸部图像。

[0156] 在步骤1033中,服务器对脸部图像和背景图像进行合成处理,得到对应虚拟主播的图像帧,并将合成得到的多个图像帧组成为对应音频数据的图像帧集合,其中,图像帧集合中图像帧的数目为图像帧数目。

[0157] 在步骤1034中,服务器对图像帧集合进行封包处理,得到携带有对应虚拟主播的图像帧集合的图像数据包。

[0158] 在一些实施例中,服务器中的渲染模块将音频数据的播放时长以及帧率参数的乘积确定为对应音频数据的图像帧数目;这里的播放时长是通过文本转语音请求模块所确定得到的,将音频数据包的播放时长随同脸部特征数据包发送至渲染模块,由渲染模块据此得到对应音频数据的图像帧数目,提取脸部特征数据包中的脸部特征数据,并基于脸部特征数据进行特效渲染处理,得到对应脸部特征数据的脸部图像;对脸部图像和背景图像进行合成处理,得到对应虚拟主播的图像帧,并将合成得到的多个图像帧组成为对应音频数据的图像帧集合,其中,图像帧集合中图像帧的数目为图像帧数目;对图像帧集合进行封包处理,得到携带有对应虚拟主播的图像帧集合的图像数据包。

[0159] 在一些实施例中,渲染模块负责基于脸部特征数据生成图像帧集合,图像帧集合中的图像帧的数目和对应的音频数据的长度相关,例如,音频数据的长度为1秒,帧率是每秒25帧,则对应该音频数据所对应的脸部特征数据包中的图像帧集合中的图像帧的数目为25帧。图像帧的生成过程主要是基于脸部特征数据进行特效渲染处理,使得到的脸部图像上能够表征出对应的脸部特征数据,这里的脸部特征数据包括口型特征数据和脸部表情特征数据,例如,对于“大家好”这个语段所对应的脸部特征数据包,其所得到的脸部图像上的口型是对应于“大家好”的口型,其所得到的脸部图像上的表情是对应于“大家好”的表情,且这里的口型特征数据可以是适应于任一标准面部模板的,标准面部模板包括真实人物面部模板、卡通人物面部模板、动物面部模板或者图形模板,在虚拟直播中,可以让一个普通的长方体开口说话进行直播,将脸部图像和背景图像进行合成,则得到了虚拟主播的图像帧,在基于脸部特征数据进行特效渲染处理的过程中会生成多个脸部图像,从而和多个背景图像合成形成多个图像帧,多个图像帧组成了图像帧集合,这里的脸部图像的数目、背景图像的数目以及图像帧的数目一致。

[0160] 在一些实施例中,上述提取脸部特征数据包中的脸部特征数据,并基于脸部特征数据进行特效渲染处理,得到对应脸部特征数据的脸部图像的过程可以通过以下技术方案实现,从对应脸部特征数据的特征向量中,提取与图像帧数目相同数目的特征向量;基于提取的特征向量进行特效渲染处理,得到与图像帧数目相同数目的脸部图像。

[0161] 在一些实施例中,图像帧集合中图像帧的数目是根据音频数据的播放长度所确定的,音频数据的长度为1秒,帧率是每秒25帧,则对应同一语段的脸部特征数据所生成得到的图像帧集合中的图像帧的数目为25帧,即意味着需要合成25帧图像帧,即意味着需要得到25个脸部图像,这里,通过特征向量来对脸部特征数据进行数学上的表达,将1秒音频数

据所对应的脸部特征数据包中的脸部特征数据以特征向量进行呈现,不同的特征向量可以表征不同的口型,因为口型变化是一个动态的过程,即便只是1秒的音频数据,其口型变化也是一个动态的过程,其中,需要25帧图像来呈现该动态过程,因而,从特征向量中提取与图像帧数目相同数目的特征向量,基于提取的特征向量进行特效渲染处理,得到与图像帧数目相同数目的脸部图像。

[0162] 在一些实施例中,上述步骤中对脸部图像和背景图像进行合成处理的过程可以通过以下技术方案实现,对脸部特征数据包进行解析处理,得到脸部特征数据包中所包含的脸部特征数据,并确定对应脸部特征数据的语段;基于语段所表征的语义,提取对应语义的至少一组背景图像;其中,每组背景图像中包括虚拟主播的模型、动作以及所处的直播场景;将脸部图像与至少一组背景图像进行合成,得到虚拟主播的图像帧,其中,虚拟主播处于直播场景中、且具备动作以及脸部图像中相应的脸部表情和口型,动作的组数和背景图像的组数相同。

[0163] 在一些实施例中,在进行合成的过程中,需要确定脸部图像所适应的背景图像,背景图像中包括虚拟主播的模型、动作以及所处的直播场景,虚拟主播的模型可以是三维模型,不同的虚拟主播模型对应了不同的体态,以及不同的整体形象,背景图像中的动作是虚拟主播会呈现的动作,例如摆手或者跳舞等动作,直播场景是虚拟主播所处的环境,例如可以是在野外、森林或者是具有特定环境的房间等等,背景图像的确定也和语义相关,脸部特征数据包中的脸部特征数据是与语段一一对应的,每个语段会有其表征的语义,基于语义来选择适应于该语义的背景图像,将脸部图像和背景图像合成之后,得到的虚拟主播的图像帧中可以呈现以下信息,虚拟主播处于直播场景中、且虚拟主播具备动作以及脸部图像中相应的脸部表情和口型。

[0164] 在一些实施例中,上述提取对应语义的背景图像的过程,可以通过以下技术方案实现,向区块链网络发送查询交易,其中,查询交易指示了用于查询区块链网络中账本的智能合约以及对应语义的查询参数,以使区块链网络中的共识节点通过执行智能合约的方式来查询账本,得到账本中符合查询参数的背景图像;或者根据语义从背景图像数据库中查询语义对应的背景图像,并确定查询到的背景图像的哈希值;从区块链网络中查询语义对应的哈希值,当所查询到的哈希值与确定的哈希值一致时,确定查询到的背景图像未经篡改。

[0165] 在一些实施例中,参见图5,图5是本发明实施例提供的基于人工智能的直播系统100的一个可选的架构示意图,终端400向区块链网络上传背景图像,背景图像被共识节点执行智能合约进行共识通过后,记录到各个记账节点的账本中,服务器200可以从区块链网络获取背景图像。终端400上链存储背景图像、服务器200上链存储背景图像、查询背景图像和更新背景图像都是通过提交交易的方式实现的,交易中携带实现数据操作的智能合约的调用信息,通过共识节点调用智能合约来执行,共识节点执行智能合约的结果需要获得其他节点的共识才能有效,从而保证了数据可靠性和一致性。终端400和服务器200到认证中心注册获取证书(其中包括业务主体的公钥、以及认证中心对业务主体的公钥和身份信息签署的数字签名),用来与终端400或服务器200针对交易的数字签名一起附加到交易中,并被发送到区块链网络,以供区块链网络从交易中取出数字证书和签名,验证消息的可靠性(未经篡改)和发送消息的业务主体的身份信息,区块链网络会根据身份进行验证,例如是

否具有发起交易的权限。共识功能和排序功能是共识节点能够实现的功能,区块链和状态数据库是账本的两种维护数据的方式,上链背景图像的交易和查询背景图像的交易在区块链中有完整的记录;在状态数据库中仅记录部分数据。

[0166] 在一些实施例中,背景图像可以由专门的美术内容提供商提供,或者由直播发起用户提供,可以将背景图像上传到区块链网络中,使得背景图像不易被篡改,防止在进行直播生成的过程中,所合成出的图像帧中出现有违社会工序良俗的内容,或者故意损害直播发起用户的内容,因而,获取背景图像的过程可以通过向区块链网络发起查询交易来实现,查询交易中指示了用于查询区块链网络中账本的智能合约以及对应语义的查询参数,查询参数可以是背景图像的标识,或者是背景图像的上传者信息等,通过执行对应于查询交易的智能合约可以得到账本中符合查询参数的背景图像,在另一种实施方式中,根据语义从背景图像数据库中查询背景图像以及背景图像的哈希值,对查询到的哈希值进行哈希验证,将背景图像的哈希值和区块链网络中存储的语义的哈希值和背景图像的哈希值进行比对,比对结果一致,则表征查询到的背景图像未被篡改。

[0167] 在步骤104中,服务器实时提取图像数据包中的图像帧集合和音频数据包中的音频数据。

[0168] 在一些实施例中,步骤103中服务器实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,可以通过以下技术方案实现,当形成针对给定文本的至少一个脸部特征数据包中的第一个脸部特征数据包时,实时基于第一个脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的第一个图像数据包;步骤104中服务器实时提取图像数据包中的图像帧集合和音频数据包中的音频数据,可以通过以下技术方案实现,当形成针对给定文本的至少一个音频数据包中的第一个音频数据包时,实时提取第一个音频数据包中的音频数据,以进行音频数据的推送;当形成携带有对应虚拟主播的图像帧集合的第一个图像数据包时,实时提取第一个图像数据包中的图像帧集合,以进行图像帧集合的推送。

[0169] 在一些实施例中,服务器对图像的处理过程和对音频的处理过程是分成两条线进行的,对于图像处理而言,一旦形成了针对给定文本的至少一个脸部特征数据包中的一个脸部特征数据包时,则实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,一旦形成携带有对应虚拟主播的图像帧集合的图像数据包时,实时提取图像数据包中的图像帧集合,以进行图像帧集合的推送,上述图像处理过程和音频处理无关,对于音频处理而言,一旦形成了音频数据包,则实时提取音频数据包中的音频数据,以进行音频数据的推送,从给定文本生成对应给定文本的音频数据包,以及提取音频数据包中的音频数据并进行推送的过程是独立于图像处理进行的,图像处理的过程为从给定文本生成对应给定文本的脸部特征数据包,进行特效渲染生成图像数据包,以及提取图像数据包中的图像帧集合,以进行图像帧集合的推送的过程是独立于上述音频数据过程的。

[0170] 在步骤105中,服务器根据图像帧集合和音频数据进行虚拟主播的直播数据流推送。

[0171] 在一些实施例中,当从给定文本获取到第一个图像数据包所耗费的时间大于从给定文本获取到第一个音频数据包所耗费的时间时,步骤105中根据图像帧集合和音频数据

进行虚拟主播的直播数据流推送,可以通过以下技术方案实现,当提取到第一个音频数据包中的音频数据时,将所提取到的音频数据实时推送到直播客户端,并将后续提取到的音频数据包中的音频数据,实时推送至直播客户端,直至提取到第一个图像数据包中的图像帧集合,并将所提取到的图像帧集合实时推送至直播客户端

[0172] 在一些实施例中,当从给定文本获取到第一个图像数据包所耗费的时间小于从给定文本获取到第一个音频数据包所耗费的时间时,步骤105中根据图像帧集合和音频数据进行虚拟主播的直播数据流推送,可以通过以下技术方案实现,当提取到第一个图像数据包中的图像帧集合时,将所提取到的图像帧集合时推送到直播客户端,并将后续提取到的图像数据包中的图像帧集合,实时推送至直播客户端,直至提取到第一个音频数据包中的音频数据,并将所提取到的音频数据实时推送至直播客户端。

[0173] 在一些实施例中,图像数据包和音频数据包是异步生成的,即图像数据包和音频数据包异步抵达视频推流模块,由视频推流模块对生成的图像数据包和音频数据包进行数据提取以及推送,当最先抵达视频推流模块的是图像数据包时,则对图像数据包中的图像帧集合进行提取以及推送,当最先抵达视频推流模块的是音频数据包时,则对音频数据包中的音频数据进行提取以及推送,即最先得到的的是音频数据包,则实时推送音频数据包中的音频数据,最先得到的的是图像数据包,则实时推送图像数据包中的图像帧集合,并不会互相等待,对于后续的音频数据包和图像数据包而言,仍然采取先生成先推送的策略,将音频数据和图像帧数据推送到客户端之后,由客户端对音频数据和图像帧数据进行同步处理,并进行视频播放,为了避免客户端对音频数据和图像帧数据进行同步处理时耗费等待时间,即尽量保证音频数据和图像帧数据的均匀到达,需要控制生成对应同一语段的图像数据包和音频数据包的耗时的差值小于差值阈值,最佳情形是生成对应同一语段的图像数据包和音频数据包的耗时完全相同,即两者同时到达视频推流数据包,同时推送到客户端进行同步处理以及播放。

[0174] 在一些实施例中,服务器中的视频推流模块可以直接将提取到的图像帧集合和音频数据推送至客户端进行直播,还可以响应于客户端发送的直播请求,为直播请求分配流媒体播放地址,并将流媒体播放地址返回至客户端。

[0175] 在一些实施例中,在客户端向服务器发送包括给定文本的直播选择指令时,可以由服务器为客户端分配流媒体服务器地址,使得客户端不断从该地址拉取直播视频,服务器选择合适的流媒体服务器地址发送给直播客户端,流媒体服务器地址的选择方式可以是固定的,也可以是预分配选择范围等方式。

[0176] 在一些实施例中,步骤105中服务器根据图像帧集合和音频数据进行虚拟主播的直播数据流推送,可以通过以下技术方案实现,将图像帧集合与音频数据依次实时推送至对应流媒体播放地址的流媒体接口。

[0177] 参见图3D,图3D是本发明实施例提供的基于人工智能的直播方法的一个可选的流程示意图,将结合图3D示出的步骤201-203进行说明。

[0178] 在步骤201中,客户端接收虚拟主播的直播选择指令。

[0179] 在步骤202中,客户端根据直播选择指令获取虚拟主播对应的图像帧集合和音频数据;其中,图像帧集合和音频数据对应于虚拟主播的给定文本。

[0180] 在步骤203中,客户端对获取的图像帧集合和音频数据进行同步处理,并实时呈现

虚拟主播的图像帧,以及对应的音频数据。

[0181] 在一些实施例中,直播请求是用于创建直播的指令,即该直播请求是针对于直播发起方,直播请求中携带有虚拟主播的给定文本,直播选择指令是用于选择观看直播的指令,即该直播选择指令是针对于观看直播的用户,步骤201 中的直播选择指令是针对于观看直播用户的,即客户端将接收到的来自于观看用户的直播选择指令,该直播选择指令会指向虚拟主播,虚拟主播会对应有直播发起方提供的给定文本,在客户端接收到直播选择指令后,客户端根据直播选择指令获取虚拟主播对应的图像帧集合和音频数据;其中,图像帧集合和音频数据对应于虚拟主播的给定文本,这里图像帧集合和音频数据的生成过程可以通过前述方案实现,在获取到图像帧集合和音频数据后,需要对两者进行同步处理,并实时呈现虚拟主播的图像帧,以及对应的音频数据,这里的音频数据和图像帧数据需要对应呈现,即音频数据和图像帧是对应于同一语段相互匹配的影音数据。

[0182] 在一些实施例中,在直播的过程中,客户端可以接收到来自于直播观看用户的互动文本,这里的互动文本是指在直播过程中和虚拟主播的互动,例如,对于美食直播而言,互动文本的内容可以是询问餐厅的地址或者菜品,这时,客户端或者服务器查询与互动文本匹配的形象资源和话术资源,这里的话术资源可以是预先配置好的,可以用于回答观看用户提问的话术,以使得直播观看用户可以获得及时的响应,同时,这里的话术可以设定快捷回复话术,快捷回复话术可以实时响应互动文本,在通过快捷回复话术响应互动文本的同时,客户端可以获取直播内容提供商上传的具体用于回答互动文本的具备实质性内容的给定文本,或者由服务器直接根据互动文本在网上爬取具体用于回答互动文本的具备实质性内容的给定文本,对话术资源进行语音合成处理,并结合形象资源,确定用于响应互动文本的音频数据以及脸部特征数据,后续的处理过程和本发明实施例提供的基于人工智能的直播方法相同。

[0183] 在一些实施例中,还可以对虚拟主播的话术资源和形象资源进行配置,接收美术资源提供方上传的虚拟主播的形象资源,并创建虚拟主播的话术资源;其中,形象资源包括以下至少之一:场景资源,模型资源和动作资源;当接收到新虚拟主播的形象资源时,为接收的形象资源分配新版本标识,并生成对应新版本的配置信息,配置信息包括以下至少之一:场景资源配置项目,模型资源配置项目,动作资源配置项目和话术资源配置项目;当接收到虚拟主播的形象资源是已有虚拟主播的更新资源时,更新已有虚拟主播的对应版本的形象资源和配置信息。

[0184] 参见图6,图6是本发明实施例提供的基于人工智能的直播方法的实现架构图,文本转语音请求模块2551基于给定文本获取到脸部特征数据与音频数据(对应于图6中的步骤101-102,在步骤101中,服务器接收来自于客户端的用于供虚拟主播表演的给定文本,在步骤102中,服务器根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包,并继续实时处理下一个给定文本),渲染模块2552 对每次获取到的脸部特征数据通过三维渲染处理(对应于图6中的步骤103,在步骤103中,服务器实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包,并继续实时处理下一个脸部特征数据包),获取虚拟角色的脸部图像,视频推流模块2553将每次获取到的脸部图像与音频数据实时合成为虚拟视频推送给客户端(对应于图6中的步骤104和105,在步骤104中,服务器实

时提取图像数据包中的图像帧集合和音频数据包中的音频数据,在步骤105中,服务器根据图像帧集合和音频数据进行虚拟主播的直播数据流推送)。文本转语音请求模块2551、渲染模块2552以及视频推流模块2553相互独立并行协助的方式,能够将文本数据自动实时合成为视频推送到客户端。这里三个模块是相互独立并行的,且音频数据和脸部特征数据是分开的数据流,相互独立并行表征文本转语音请求模块2551对第一个给定文本处理完毕后,将得到的第一个脸部特征数据包发送至渲染模块2552,将得到的第一个音频数据包发送至视频推流模块2553,且文本转语音请求模块2551继续对第二个给定文本进行处理以得到第二个脸部特征数据包以及第二个音频数据包,将第二个音频数据包直接推送至视频推流模块2553,渲染模块2552对接收到的第一个脸部特征数据包进行处理生成第一个图像数据包,此时渲染模块2552会继续对接收到的第二个脸部特征数据包进行处理以生成第二个图像数据包,即文本转语音请求模块2551不会等待第一个给定文本完成全部推流过程之后再继续进行下一步处理,而是继续处理接收到的给定文本,且脸部特征数据包和音频数据包并不是绑定在一起进行处理的,而是分成两条处理线进行的,即,文本转语音请求模块2551对第一个给定文本处理生成第一个音频数据包和第一个脸部特征数据包的过程中,生成第一个音频数据包和生成第一个脸部特征数据包所耗费的时间是不一样的,一旦第一个音频数据包产生之后,即会被发送至视频推流模块2553,这个过程和脸部特征数据包的生成进度是没有关系的,一旦第一个脸部特征数据包产生之后,即会被发送至渲染模块2552,这个过程和音频数据包的生成进度是没有关系的,渲染模块2552也不会受其他模块处理进度的影响,而会继续处理其所接收到的下一个脸部特征数据包。

[0185] 下面,将说明本发明实施例提供的基于人工智能的直播方法在一个实际的应用场景中的示例性应用。

[0186] 本发明实施例提供的基于人工智能的直播方法能应用到包括虚拟新闻直播、虚拟游戏解说等多个项目和产品应用中,从而有效提高直播的实时性。

[0187] 在虚拟直播的场景中,通过本发明实施例提供的基于人工智能的直播方法可以根据输入文本(给定文本),并行独立地获取到音频数据与图像数据,并实时推流给客户端以实现实时直播。

[0188] 图7A-7B是本发明实施例提供的虚拟视频直播服务的整体框架示意图,参见图7A-7B,虚拟视频直播服务的整体框架包括直播客户端、文本转语音服务和虚拟视频推流服务,其中虚拟视频推流服务的实现方法为本发明的主要创新所在。图7A中,直播客户端向虚拟视频推流服务发送直播请求,直播请求中携带有给定文本,虚拟视频推流服务向文本转语音服务请求获取音频数据和脸部特征数据,文本转语音服务返回音频数据和脸部特征数据作为响应,虚拟视频推流服务根据获取到的响应继续执行本发明实施例提供的直播方法,将图像帧集合和音频数据推送至直播客户端。图7B中,直播客户端向虚拟视频推流服务发送直播请求,直播请求中携带有给定文本,虚拟视频推流服务将流媒体服务地址作为响应返回给客户端,且向文本转语音服务请求获取音频数据和脸部特征数据,文本转语音服务返回音频数据和脸部特征数据作为响应,虚拟视频推流服务根据获取到的响应继续执行本发明实施例提供的直播方法,将图像帧集合和音频数据推送至流媒体服务,直播客户端从对应流媒体服务的流媒体服务地址拉取视频。

[0189] 图8是本发明实施例提供的虚拟视频推流服务的实现框架,参见图8,虚拟视频推

流服务中包括文本转语音请求模块、渲染模块、视频推流模块。直播客户端向虚拟视频直播服务端发送文本请求,客户端发送的文本请求中的给定文本是直播视频的虚拟角色将会说出的文字;文本转语音请求模块向文本转语音服务发起请求,流式获取给定文本相对应的音频数据与脸部特征数据(音频数据包1-3和脸部特征数据包1-3),文本转语音请求模块每从文本转语音服务获取到一个音频数据包和脸部特征数据包,就将脸部特征数据包推送给三维渲染模块,将音频数据包推送给视频推流模块,当文本转语音请求模块接收到文本转语音服务的结束包时,发送结束包给三维渲染模块和视频推流模块;三维渲染模块每获取到一个脸部特征数据包,提取其中的脸部特征数据进行三维渲染获取一组对应脸部表情图像,同时将一张脸部表情图像与一张背景图像合成为一张完整的图像,得到一组完整的直播视频图像帧作为图形数据包,并推送给视频推流模块。当接收到一个结束包时,发送结束包给视频推流模块;视频推流模块每获取到一个文本转语音请求模块推送来的音频数据包或三维渲染模块推送过来的图像数据包1-3,提取其中的音频数据或图像帧数据。通过快进动态图像专家组(FFMPEG, Fast Forward Moving Picture Expert Group)工具将音频或图像帧数据异步推送给客户端,由客户端完成同步,直到接收到文本转语音请求模块推送的结束包与三维渲染模块推送的结束包后,结束该视频的推送。

[0190] 这里,文本转语音请求模块、三维渲染模块和视频推流模块是独立并行协作实现虚拟视频实时推流,文本转语音请求模块在获取文本转语音服务的流式数据前,需要先请求文本转语音服务以获取客户端提供的给定文本最终生成的音频的时长,用于估算最终生成的直播视频时长,从文本转语音服务流式获取音频与脸部特征数据,能够快速获取到音频和图像数据,实现实时直播,将脸部特征数据与音频数据分离,并分别推送给三维渲染模块与视频推送模块,能够有效缓解三维渲染模块处理耗时对实时推流的时延影响。

[0191] 三维渲染模块进行渲染的过程可以由开放图形库工具与卷积神经网络进行实现,背景图像包括人物的肢体动作以及背景环境信息,本发明实施例提供的基于人工智能的直播方法根据给定文本确定视频长度,在渲染的过程中,从预先保存的通用背景图像中选择n组合适的背景图像与脸部表情进行匹配合成,每一组背景图像都是一个完整的动作,n组背景图像正好能够在视频结束时完成n个动作。

[0192] 视频推流模块主要使用了FFMPEG工具进行视频推流,由于音频数据与图像数据是相互独立且异步到达视频推流模块的,在推流过程中能够有效利用三维渲染的耗时用于接收文本转语音请求模块获取到的音频数据,避免大量等待音频数据包的时间,从而提高推流实时性,当收到第一个数据包时,进行推流初始化,并推送音频或者图像数据,当接收到结束包时,结束推流,此时完成一个完整的视频推流过程。

[0193] 上述文本转语音请求模块、三维渲染模块、视频推流模块所进行的处理都是十分耗时的过程,若将这些处理用串行的方式进行实现、那么数据处理时长必将大于生成视频的时长,无法实现虚拟视频的直播,因此,文本转语音请求模块、三维渲染模块和视频推流模块相互独立并行协作,只要各个模块的处理时间小于视频时长,同时客户端仅需要在发送请求后,等待一个固定延时后就能够实现虚拟视频实时直播,该固定延时等于文本转语音请求模块获取的第一个数据传递到视频推流模块并推流成功的耗时。

[0194] 本发明在文本转语音请求模块、三维渲染模块、视频推流模块相互独立并行协作的基础上,对文本转语音请求模块获取的音频数据与脸部特征数据进行分离分流处理的方

式,能够有效避免音频数据源与脸部特征数据源不稳定的情况(例如,每次只接收到音频数据或脸部特征数据,且音频数据与脸部特征数据的获取并在时间上不是均匀分布的)对实时推流的影响,在三个模块相互独立协作的过程中,在三维渲染模块渲染图像期间,视频推流模块在等待文本转语音请求模块获取音频数据,有效利用了文本转语音服务的音频处理耗时敏感与三维渲染模块的图像处理耗时敏感的特点,提高实时推流的实时性。

[0195] 下面继续说明本发明实施例提供的基于人工智能的直播装置255的实施为软件模块的示例性结构,在一些实施例中,如图2所示,存储在存储器250的基于人工智能的直播装置255中的软件模块可以包括:文本转语音请求模块251,用于接收用于供虚拟主播表演的给定文本,根据给定文本实时获取对应虚拟主播的音频数据和脸部特征数据,并分别形成至少一个音频数据包和至少一个脸部特征数据包;渲染模块252,用于实时基于脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的图像数据包;视频推流模块253,用于实时提取图像数据包中的图像帧集合和音频数据包中的音频数据;根据图像帧集合和音频数据进行虚拟主播的直播数据流推送。

[0196] 在一些实施例中,渲染模块252,还用于:当形成针对给定文本的至少一个脸部特征数据包中的第一个脸部特征数据包时,实时基于第一个脸部特征数据包中的脸部特征数据进行特效渲染处理,得到携带有对应虚拟主播的图像帧集合的第一个图像数据包;视频推流模块253,还用于:当形成针对给定文本的至少一个音频数据包中的第一个音频数据包时,实时提取第一个音频数据包中的音频数据,以进行音频数据的推送;当形成携带有对应虚拟主播的图像帧集合的第一个图像数据包时,实时提取第一个图像数据包中的图像帧集合,以进行图像帧集合的推送。

[0197] 在一些实施例中,当从给定文本获取到第一个图像数据包所耗费的时间大于从给定文本获取到第一个音频数据包所耗费的时间时,视频推流模块253,还用于:当提取到第一个音频数据包中的音频数据时,将所提取到的音频数据实时推送到直播客户端,并将后续提取到的音频数据包中的音频数据,实时推送至直播客户端,直至提取到第一个图像数据包中的图像帧集合,并将所提取到的图像帧集合实时推送至直播客户端。

[0198] 在一些实施例中,当从给定文本获取到第一个图像数据包所耗费的时间小于从给定文本获取到第一个音频数据包所耗费的时间时,视频推流模块253,还用于:当提取到第一个图像数据包中的图像帧集合时,将所提取到的图像帧集合时推送到直播客户端,并将后续提取到的图像数据包中的图像帧集合,实时推送至直播客户端,直至提取到第一个音频数据包中的音频数据,并将所提取到的音频数据实时推送至直播客户端。

[0199] 在一些实施例中,文本转语音请求模块251,还用于:当接收到给定文本时,将给定文本划分为至少两个语段,并实时将语段转化为对应语段的词向量;对词向量进行编码处理以及解码处理,得到对应词向量的音频特征;对音频特征进行合成处理,得到分别对应虚拟主播的各个语段的音频数据。

[0200] 在一些实施例中,文本转语音请求模块251,还用于:预测虚拟主播的对应各个语段的嘴部关键点,并对嘴部关键点进行归一化处理,以使嘴部关键点适应于标准面部模板;对经过归一化处理的嘴部关键点进行降维处理,得到对应虚拟主播的对应各个语段的口型特征数据;对语段进行语义分析,得到语段所表征的语义;根据语段所表征的语义,确定与语义匹配的脸部表情特征数据,并结合口型特征数据以及脸部表情特征数据形成对应虚拟

主播的各个语段的脸部特征数据;基于对应于同一语段的脸部特征数据和音频数据,分别形成音频数据包和脸部特征数据包。

[0201] 在一些实施例中,文本转语音请求模块2551,还用于:在对嘴部关键点进行归一化处理之前,向区块链网络发送查询交易,其中,查询交易指示了用于查询区块链网络中账本的智能合约,以使区块链网络中的共识节点通过执行智能合约的方式来查询账本,得到账本中存储的标准面部模板;或者根据标准面部模板的标识,从标准面部模板数据库中查询标识对应的标准面部模板,并确定查询到的标准面部模板的哈希值;从区块链网络中查询标识对应的哈希值,当所查询到的哈希值与确定的哈希值一致时,确定查询到的标准面部模板未经篡改。

[0202] 在一些实施例中,文本转语音请求模块2551,还用于:在分别形成至少一个音频数据包和至少一个脸部特征数据包之前,将给定文本发送至区块链网络,以使区块链网络中的共识节点通过执行智能合约的方式来对给定文本进行合规性检查;当接收到超过预设数目的共识节点返回的合规性确认时,确定给定文本通过合规性检查。

[0203] 在一些实施例中,文本转语音请求模块2551,还用于:在分别形成相互匹配的音频数据包和脸部特征数据包之前,确定音频数据所表征的音素顺序;基于音素顺序,确定携带有口型特征数据的脸部特征数据的顺序,以使脸部特征数据的顺序所表征的口型变化和音素顺序相匹配。

[0204] 在一些实施例中,渲染模块2552,还用于:将音频数据的播放时长以及帧率参数的乘积确定为对应音频数据的图像帧数目;实时提取脸部特征数据包中的脸部特征数据,并基于脸部特征进行特效渲染处理,得到对应脸部特征数据的脸部图像;对脸部图像和背景图像进行合成处理,得到对应虚拟主播的图像帧,并将合成得到的多个图像帧组成为对应音频数据的图像帧集合;其中,图像帧集合中图像帧的数目为图像帧数目;对图像帧集合进行封包处理,得到携带有对应虚拟主播的图像帧集合的图像数据包。

[0205] 在一些实施例中,渲染模块2552,还用于:从脸部特征数据中,提取与图像帧数目相同数目的特征向量;基于提取的特征向量进行特效渲染处理,得到与图像帧数目相同数目的脸部图像。

[0206] 在一些实施例中,渲染模块2552,还用于:对脸部特征数据包进行解析处理,得到脸部特征数据包中所包含的脸部特征数据,并确定对应脸部特征数据的语段;基于语段所表征的语义,提取对应语义的至少一组背景图像;其中,每组背景图像中包括虚拟主播的模型、动作以及所处的直播场景;将脸部图像与至少一组背景图像进行合成,得到虚拟主播的图像帧,其中,虚拟主播处于直播场景中、且具备动作以及脸部图像中相应的脸部表情和口型,动作的组数和背景图像的组数相同。

[0207] 在一些实施例中,渲染模块2552,还用于:向区块链网络发送查询交易,其中,查询交易指示了用于查询区块链网络中账本的智能合约以及对应语义的查询参数,以使区块链网络中的共识节点通过执行智能合约的方式来查询账本,得到账本中符合查询参数的背景图像;或者根据语义从背景图像数据库中查询语义对应的背景图像,并确定查询到的背景图像的哈希值;从区块链网络中查询语义对应的哈希值,当所查询到的哈希值与确定的哈希值一致时,确定查询到的背景图像未经篡改。

[0208] 在一些实施例中,装置还包括:流媒体服务模块2554,用于:响应于客户端发送的

直播请求,为直播请求分配流媒体播放地址,并将流媒体播放地址返回至客户端;视频推流模块2553,还用于:将图像帧集合与音频数据依次实时推送至对应流媒体播放地址的流媒体接口。

[0209] 本发明实施例提供一种基于人工智能的直播装置,装置包括:

[0210] 指令接收模块,用于接收虚拟主播的直播选择指令;

[0211] 数据获取模块,用于根据直播选择指令获取虚拟主播对应的图像帧集合和音频数据;

[0212] 其中,图像帧集合和音频数据对应于虚拟主播的给定文本;

[0213] 呈现模块,用于对获取的图像帧集合和音频数据进行同步处理,并实时呈现虚拟主播的图像帧,以及对应的音频数据。

[0214] 本发明实施例提供一种存储有可执行指令的存储介质,其中存储有可执行指令,当可执行指令被处理器执行时,将引起处理器执行本发明实施例提供的方法,例如,如图3A-3D示出的基于人工智能的直播方法。

[0215] 在一些实施例中,存储介质可以是FRAM、ROM、PROM、EPROM、EE PROM、闪存、磁表面存储器、光盘、或CD-ROM等存储器;也可以是包括上述存储器之一或任意组合的各种设备。

[0216] 在一些实施例中,可执行指令可以采用程序、软件、软件模块、脚本或代码的形式,按任意形式的编程语言(包括编译或解释语言,或者声明性或过程性语言)来编写,并且其可按任意形式部署,包括被部署为独立的程序或者被部署为模块、组件、子例程或者适合在计算环境中使用的其它单元。

[0217] 作为示例,可执行指令可以但不一定对应于文件系统中的文件,可以可被存储在保存其它程序或数据的文件的一部分,例如,存储在超文本标记语言(H TML,Hyper Text Markup Language)文档中的一个或多个脚本中,存储在专用于所讨论的程序的单个文件中,或者,存储在多个协同文件(例如,存储一个或多个模块、子程序或代码部分的文件)中。

[0218] 作为示例,可执行指令可被部署为在一个计算设备上执行,或者在位于一个地点的多个计算设备上执行,又或者,在分布在多个地点且通过通信网络互连的多个计算设备上执行。

[0219] 综上所述,通过本发明实施例提供的基于人工智能的直播方法,通过并行处理和数据分离的方式,解决在需要耗费较大计算力用于获取音视频数据的前提下,虚拟视频实时直播的问题,有效增强虚拟视频直播的实时性,提高直播视频流畅性且促进了人工智能在直播服务中的发展,解放了真人直播,降低了劳动力成本。

[0220] 以上所述,仅为本发明的实施例而已,并非用于限定本发明的保护范围。凡在本发明的精神和范围之内所作的任何修改、等同替换和改进等,均包含在本发明的保护范围之内。

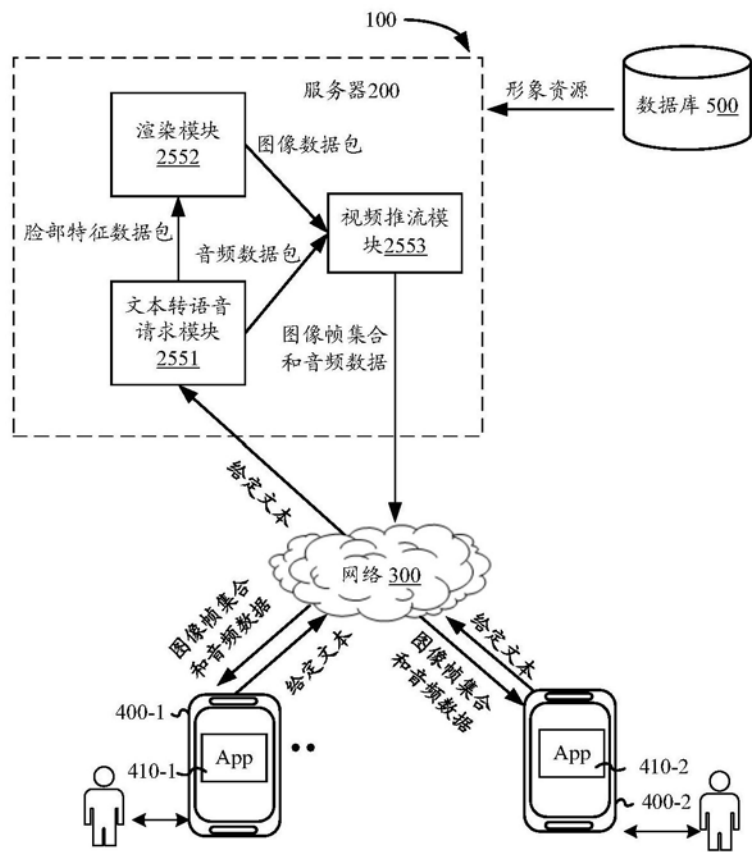


图1

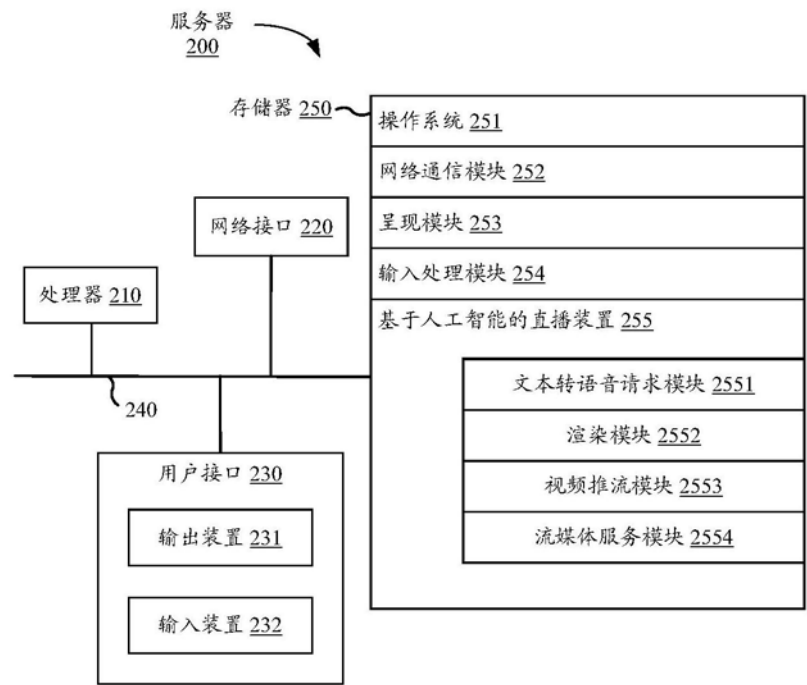


图2

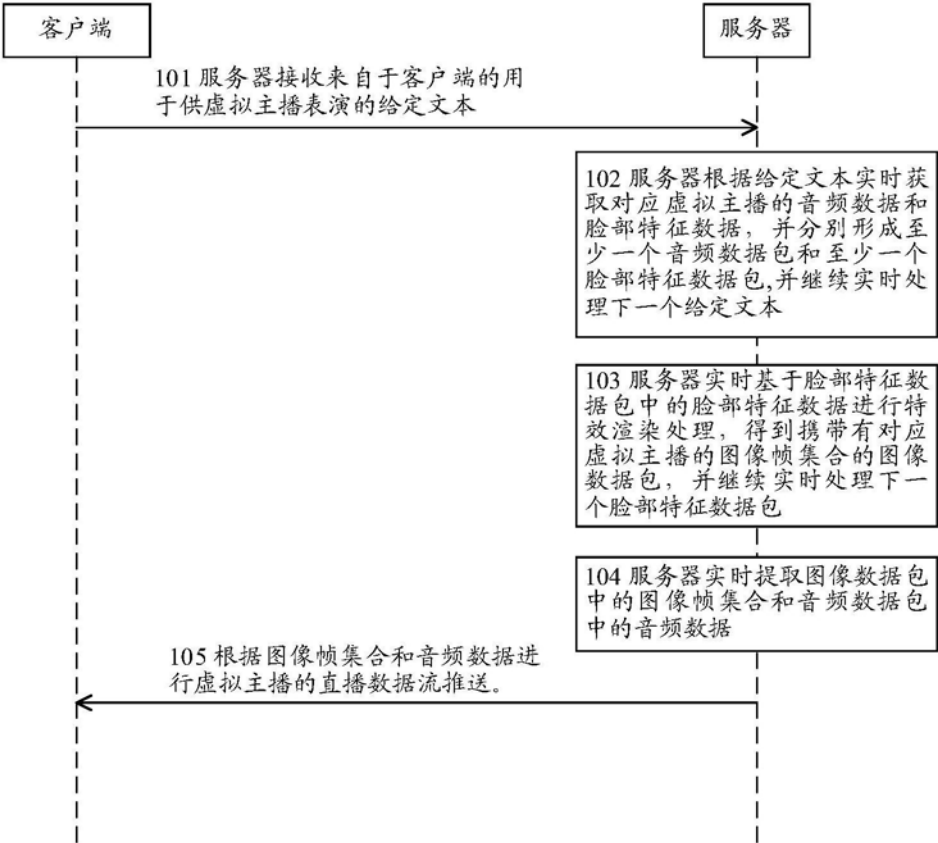


图3A

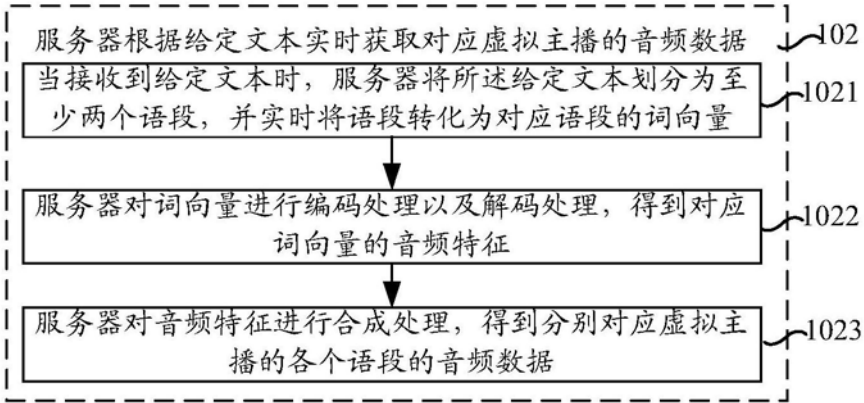


图3B

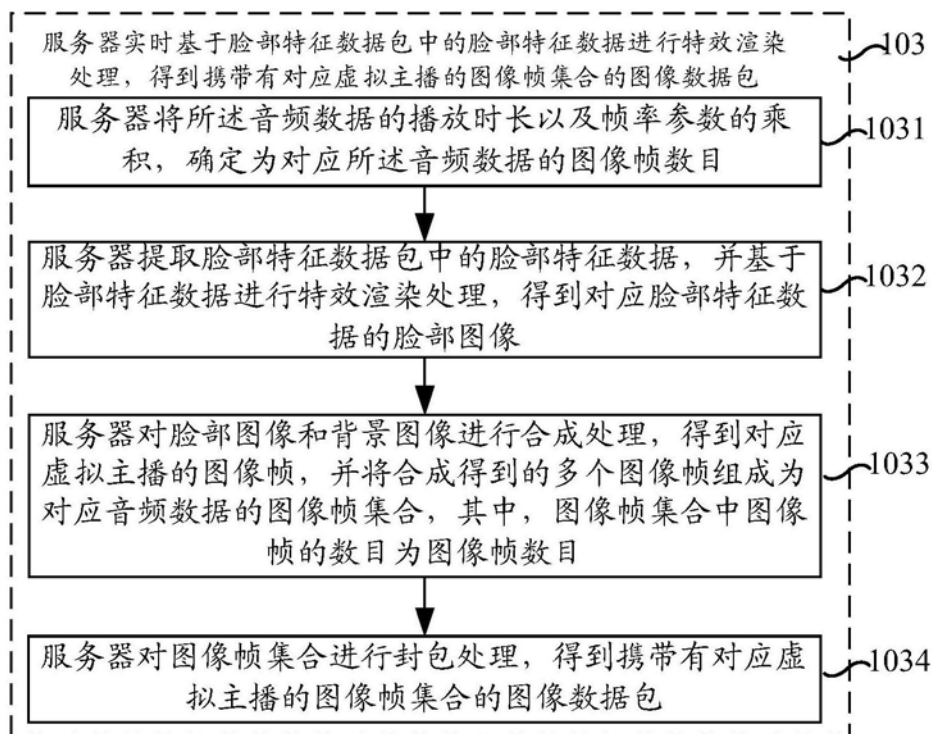


图3C

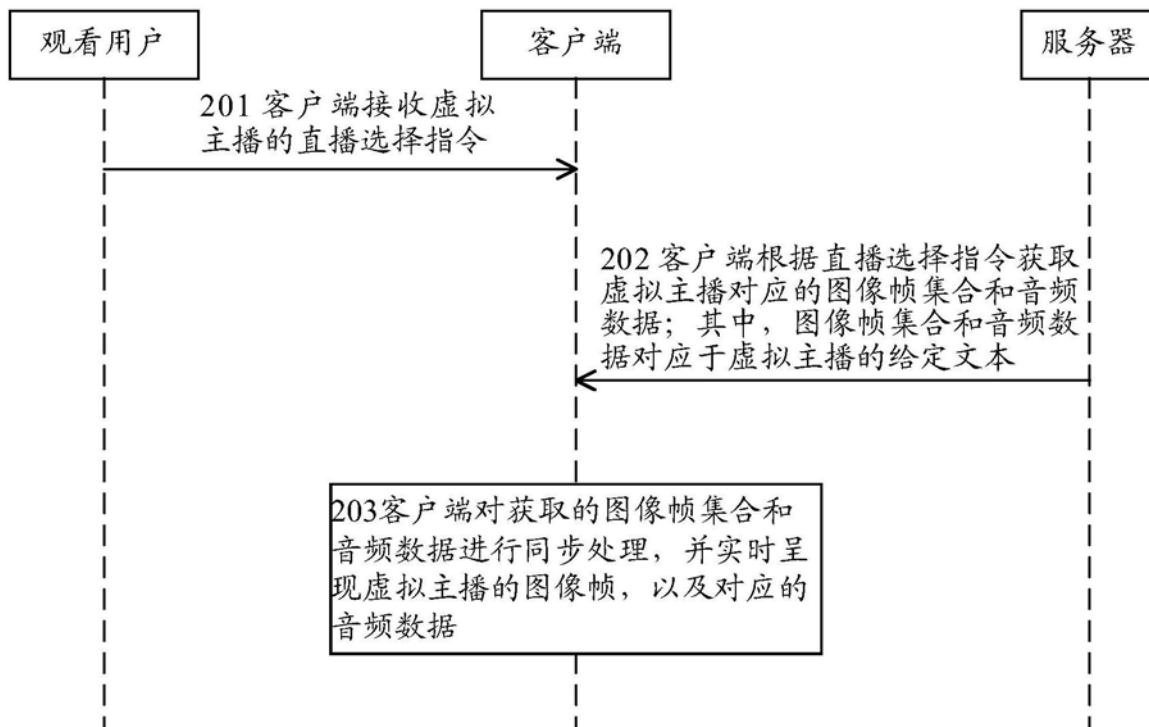


图3D

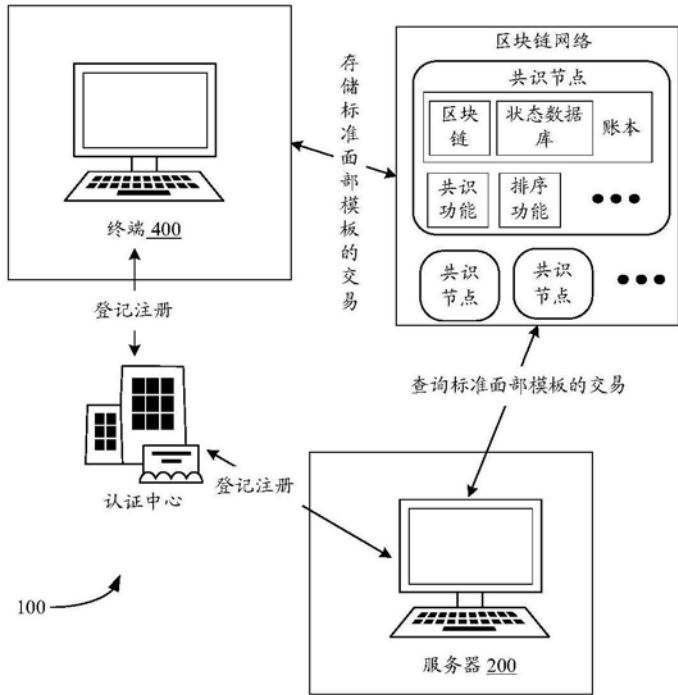


图4

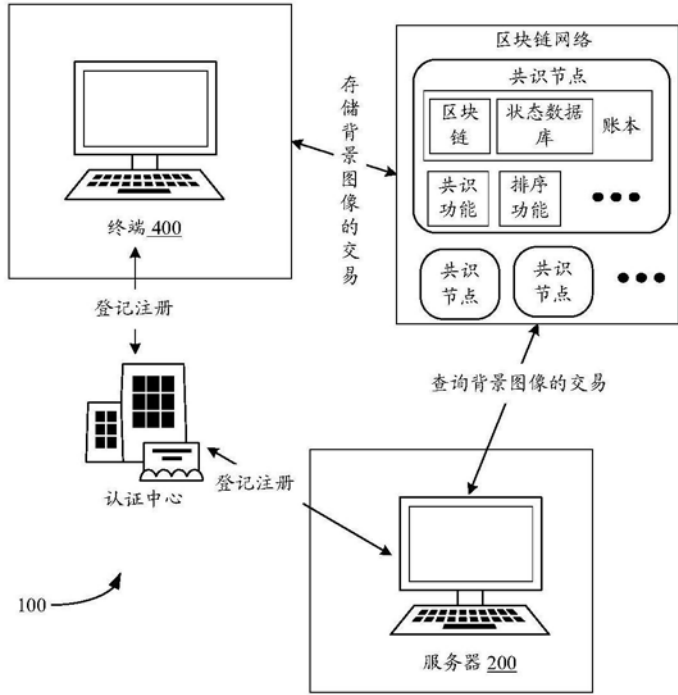


图5

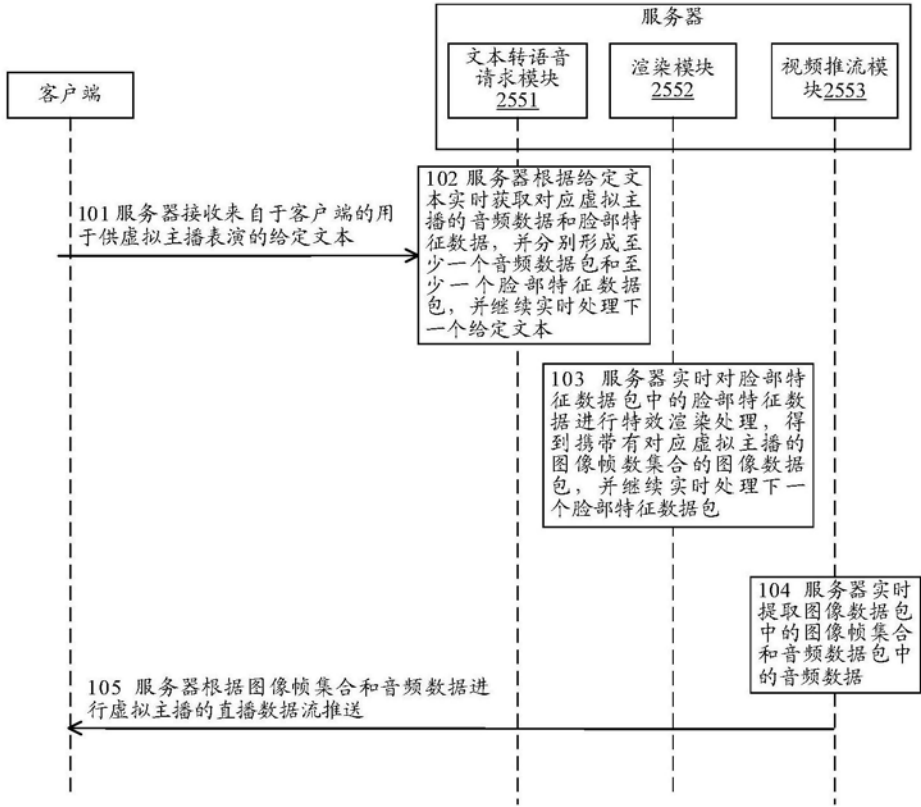


图6

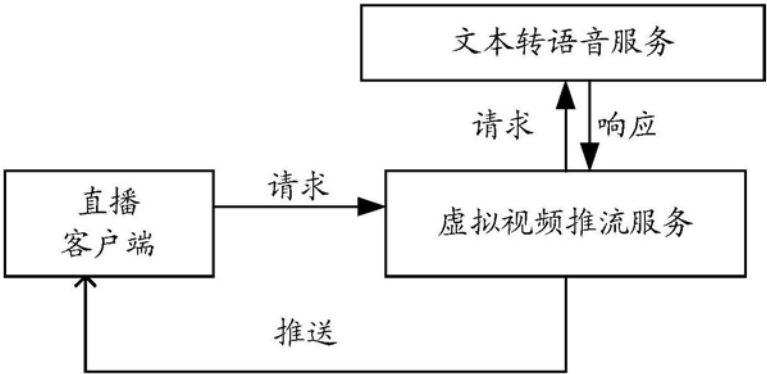


图7A

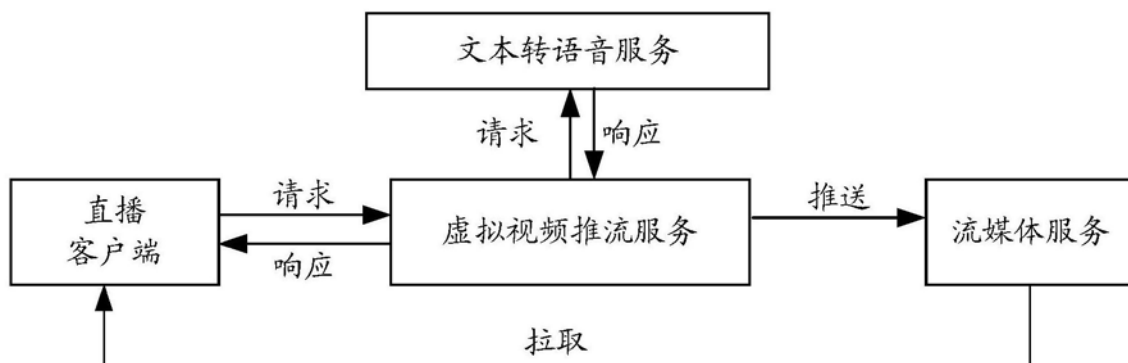


图7B

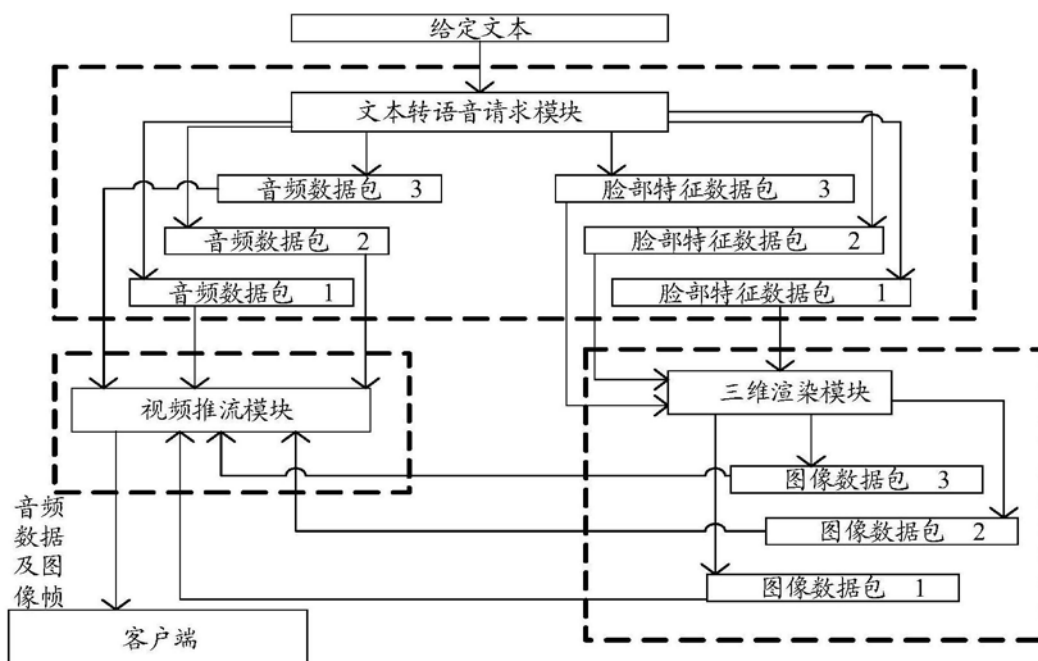


图8