(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115082602 A
(43)申请公布日 2022.09.20

(21)申请号 202210681368.3
(22)申请日 2022.06.15
(71)申请人 北京百度网讯科技有限公司
地址 100085 北京市海淀区上地十街10号百度大厦2层
(72)发明人 吴甜 李彦宏 肖欣延 刘昊 刘家辰 佘俏俏 吕雅娟
(74)专利代理机构 北京市汉坤律师事务所 11602
专利代理师 姜浩然 吴丽丽
(51)Int.Cl.
G06T 13/40(2011.01)
G06T 13/20(2011.01)
G06F 16/953(2019.01)
G06F 40/30(2020.01)
G06V 10/774(2022.01)
G06V 10/82(2022.01)

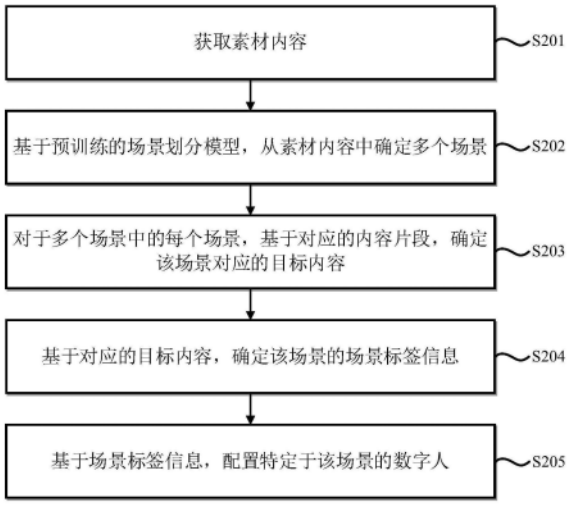
权利要求书5页 说明书17页 附图8页

(54)发明名称

生成数字人的方法、模型的训练方法、装置、设备和介质

(57)摘要

本公开提供了一种生成数字人的方法、模型的训练方法、装置、设备和介质，涉及人工智能领域，具体涉及自然语言处理、深度学习、计算机视觉、图像处理、增强现实和虚拟现实等技术领域，可应用于元宇宙等场景。实现方案为：获取素材内容；基于预训练的场景划分模型，从素材内容中确定多个场景，其中，多个场景中的每个场景分别对应于素材内容中的一个具有完整语义信息的内容片段；以及对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容；基于对应的目标内容，确定该场景的场景标签信息；以及基于场景标签信息，配置特定于该场景的数字人。



CN 115082602 A

1.一种生成数字人的方法,所述方法包括:

获取素材内容;

基于预训练的场景划分模型,从所述素材内容中确定多个场景,其中,所述多个场景中的每个场景分别对应于所述素材内容中的一个具有完整语义信息的内容片段;以及

对于所述多个场景中的每个场景,

基于对应的内容片段,确定该场景对应的目标内容;

基于所述对应的目标内容,确定该场景的场景标签信息;以及

基于所述场景标签信息,配置特定于该场景的数字人。

2.根据权利要求1所述的方法,其中,获取素材内容包括:

基于下列方式中的至少一者,获取所述素材内容:

基于网页地址,获取所述素材内容;或

基于搜索关键词,获得所述素材内容。

3.根据权利要求1或2所述的方法,其中,所述素材内容包括图像数据和视频数据中的至少一者以及文本数据。

4.根据权利要求1至3中任一项所述的方法,其中,基于预训练的场景划分模型,从所述素材内容中确定多个场景,包括:

通过对所述素材内容进行篇章结构分析和篇章语义分割,从所述素材内容中确定多个子主题,并且确定所述多个子主题之间的结构关系;以及

基于所述结构关系,将所述多个子主题划分为所述多个场景。

5.根据权利要求4所述的方法,其中,对于所述多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容,包括:

基于该场景与前一场景之间的结构关系,生成用于该场景的第一内容。

6.根据权利要求4或5所述的方法,其中,对于所述多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容,包括:

基于预训练的风格转换模型,将所述对应的内容片段转换为所述对应的目标内容,其中,所述风格转换模型是基于提示学习训练得到的。

7.根据权利要求6所述的方法,其中,对于所述多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容,还包括以下中的至少一项:

对所述对应的内容片段执行文本改写和文本压缩中的至少一种处理,以更新所述对应的内容片段;以及

对所述经转换的目标内容执行文本改写和文本压缩中的至少一种处理,以更新所述对应的目标内容。

8.根据权利要求1至7中任一项所述的方法,其中,所述场景标签信息包括语义标签,其中,对于所述多个场景中的每个场景,基于所述对应的目标内容,确定该场景的场景标签信息,包括:

对所述对应的目标内容进行情感分析,以获得所述语义标签。

9.根据权利要求8所述的方法,其中,所述语义标签用于标识所述对应的目标内容所表达的情感包括:积极、中性或消极。

10.根据权利要求8或9所述的方法,其中,对于所述多个场景中的每个场景,基于所述

标签信息,配置特定于该场景的数字人,包括:

基于所述语义标签,配置所述数字人的服饰、表情和动作中的至少一者。

11.根据权利要求10所述的方法,还包括:

将所述目标内容转换成语音,用于所述数字人播报。

12.根据权利要求11所述的方法,其中,对于所述多个场景中的每个场景,基于所述场景标签信息,配置特定于该场景的数字人,还包括:

基于所述语义标签,配置所述数字人语音的语气。

13.根据权利要求1至12中任一项所述的方法,还包括:

以全息图像的形式呈现所述数字人。

14.根据权利要求1至12中任一项所述的方法,还包括:

以视频的形式呈现所述数字人。

15.根据权利要求14所述的方法,还包括:

对于所述多个场景中的每个场景,

基于所述素材内容和该场景对应的目标内容,检索与该场景相关的视频素材;以及

将所述视频素材和所述数字人相结合。

16.根据权利要求15所述的方法,其中,对于所述多个场景中的每个场景,基于所述素材内容和该场景对应的目标内容,检索与该场景相关的视频素材,包括:

提取场景关键词;以及

基于所述场景关键词,检索与该场景相关的视频素材。

17.根据权利要求15或16所述的方法,其中,对于所述多个场景中的每个场景,基于所述素材内容和该场景对应的目标内容,检索与该场景相关的视频素材,包括:

提取句子级关键词;以及

基于所述句子级关键词,检索与该场景相关的视频素材。

18.根据权利要求17所述的方法,还包括:

基于所述句子级关键词,将检索到的视频素材和所述目标内容对齐。

19.根据权利要求15至18中任一项所述的方法,还包括:

响应于确定所述视频素材中包括特定素材,基于所述特定素材在所述视频素材中的显示位置,确定所述数字人的动作。

20.根据权利要求14至19中任一项所述的方法,还包括:

对于所述多个场景中的每个场景,

从该场景对应的目标内容中提取键-值形式的信息;以及

基于所述键-值形式的信息,生成用于所述视频的辅助素材。

21.根据权利要求15至20中任一项所述的方法,还包括:

确定所述视频素材相对应的场景所需的播放时长的占比;以及

基于所述占比,确定是否在相应场景中触发所述数字人。

22.一种场景划分模型的训练方法,包括:

获取样本素材内容和所述样本素材内容中的多个样本场景;

基于预设场景划分模型,从所述样本素材内容中确定多个预测场景;以及

基于所述多个样本场景和所述多个预测场景调整所述预设场景划分模型的参数,以得

到训练后的场景划分模型。

23.根据权利要求22所述的训练方法,其中,所述预设场景划分模型包括篇章语义分割模型和篇章结构分析模型,其中,基于预设场景划分模型,从所述样本素材内容中确定多个预测场景包括:

利用所述篇章语义分割模型和所述篇章结构分析模型对所述样本素材内容进行处理,以确定所述素材内容中多个预测子主题以及所述多个预测子主题之间的预测结构关系;以及

基于所述预测结构关系,将所述多个预测子主题划分为所述多个预测场景。

24.一种生成数字人的装置,所述装置包括:

第一获取单元,被配置为获取素材内容;

第一确定单元,被配置为基于预训练的场景划分模型,从所述素材内容中确定多个场景,其中,所述多个场景中的每个场景分别对应于所述素材内容中的一个具有完整语义信息的内容片段;

第二确定单元,被配置为对于所述多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容;

第三确定单元,被配置为基于所述对应的目标内容,确定该场景的场景标签信息;以及

数字人配置单元,被配置为基于所述场景标签信息,配置特定于该场景的数字人。

25.根据权利要求24所述的装置,其中,所述第一获取单元被进一步配置为基于下列方式中的至少一者,获取所述素材内容:

基于网页地址,获取所述素材内容;或

基于搜索关键词,获得所述素材内容。

26.根据权利要求24或25所述的装置,其中,所述素材内容包括图像数据和视频数据中的至少一者以及文本数据。

27.根据权利要求24至26中任一项所述的装置,其中,所述第一确定单元包括:

第一确定子单元,被配置为通过对所述素材内容进行篇章结构分析和篇章语义分割,从所述素材内容中确定多个子主题,并且确定所述多个子主题之间的结构关系;以及

第一划分子单元,被配置为基于所述结构关系,将所述多个子主题划分为所述多个场景。

28.根据权利要求27所述的装置,其中,所述第二确定单元包括:

生成子单元,被配置为基于该场景与前一场景之间的结构关系,生成用于该场景的第一内容。

29.根据权利要求27或28所述的装置,其中,所述第二确定单元包括:

转换子单元,被配置为基于预训练的风格转换模型,将所述对应的内容片段转换为所述对应的目标内容,其中,所述风格转换模型是基于提示学习训练得到的。

30.根据权利要求29所述的装置,其中,所述第二确定单元包括以下中的至少一项:

第一更新子单元,被配置为对所述对应的内容片段执行文本改写和文本压缩中的至少一种处理,以更新所述对应的内容片段;以及

第二更新子单元,被配置为对所述经转换的目标内容执行文本改写和文本压缩中的至少一种处理,以更新所述对应的目标内容。

31.根据权利要求24至30中任一项所述的装置,其中,所述第三确定单元包括:

情感分析子单元,被配置为对所述对应的目标内容进行情感分析,以获得所述语义标签。

32.根据权利要求31所述的装置,其中,所述语义标签用于标识所述对应的目标内容所表达的情感包括:积极、中性或消极。

33.根据权利要求31或32所述的装置,其中,所述数字人配置单元包括:

第一配置子单元,被配置为基于所述语义标签,配置所述数字人的服饰、表情和动作中的至少一者。

34.根据权利要求33所述的装置,还包括:

语音转换单元,被配置为将所述目标内容转换成语音,用于所述数字人播报。

35.根据权利要求34所述的装置,其中,所述数字人配置单元还包括:

第二配置子单元,被配置为基于所述语义标签,配置所述数字人语音的语气。

36.根据权利要求24至35中任一项所述的装置,还包括:

全息图像呈现单元,被配置为以全息图像的形式呈现所述数字人。

37.根据权利要求24至35中任一项所述的装置,还包括:

视频呈现单元,被配置为以视频的形式呈现所述数字人。

38.根据权利要求37所述的装置,还包括:

检索单元,被配置为对于所述多个场景中的每个场景,基于所述素材内容和该场景对应的目标内容,检索与该场景相关的视频素材;以及

结合单元,被配置为将所述视频素材和所述数字人相结合。

39.根据权利要求38所述的装置,其中,所述检索单元包括:

第一提取子单元,被配置为提取场景关键词;以及

第一检索子单元,被配置为基于所述场景关键词,检索与该场景相关的视频素材。

40.根据权利要求38或39所述的装置,其中,所述检索单元包括:

第二提取子单元,被配置为提取句子级关键词;以及

第二检索子单元,被配置为基于所述句子级关键词,检索与该场景相关的视频素材。

41.根据权利要求40所述的装置,还包括:

对齐单元,被配置为基于所述句子级关键词,将检索到的视频素材和所述目标内容对齐。

42.根据权利要求38至41中任一项所述的装置,还包括:

第四确定单元,被配置为响应于确定所述视频素材中包括特定素材,基于所述特定素材在所述视频素材中的显示位置,确定所述数字人的动作。

43.根据权利要求37至42中任一项所述的装置,还包括:

提取单元,被配置为对于所述多个场景中的每个场景,从该场景对应的目标内容中提取键-值形式的信息;以及

生成单元,被配置为基于所述键-值形式的信息,生成用于所述视频的辅助素材。

44.根据权利要求38至43中任一项所述的装置,还包括:

第五确定单元,被配置为确定所述视频素材相对应的场景所需的播放时长的占比;以及

第六确定单元，被配置为基于所述占比，确定是否在相应场景中触发所述数字人。

45.一种场景划分模型的训练装置，包括：

第二获取单元，被配置为获取样本素材内容和所述样本素材内容中的多个样本场景；

第七确定单元，被配置为基于预设场景划分模型，从所述样本素材内容中确定多个预测场景；以及

训练单元，被配置为基于所述多个样本场景和所述多个预测场景调整所述预设场景划分模型的参数，以得到训练后的场景划分模型。

46.根据权利要求45所述的训练装置，其中，所述预设场景划分模型包括篇章语义分割模型和篇章结构分析模型，其中，所述第七确定单元包括：

第二确定子单元，被配置为利用所述篇章语义分割模型和所述篇章结构分析模型对所述样本素材内容进行处理，以确定所述素材内容中多个预测子主题以及所述多个预测子主题之间的预测结构关系；以及

第二划分子单元，被配置为基于所述预测结构关系，将所述多个预测子主题划分为所述多个预测场景。

47.一种电子设备，包括：

至少一个处理器；以及

与所述至少一个处理器通信连接的存储器；其中

所述存储器存储有可被所述至少一个处理器执行的指令，所述指令被所述至少一个处理器执行，以使所述至少一个处理器能够执行权利要求1-23中任一项所述的方法。

48.一种存储有计算机指令的非瞬时计算机可读存储介质，其中，所述计算机指令用于使所述计算机执行根据权利要求1-23中任一项所述的方法。

49.一种计算机程序产品，包括计算机程序，其中，所述计算机程序在被处理器执行时实现权利要求1-23中任一项所述的方法。

生成数字人的方法、模型的训练方法、装置、设备和介质

技术领域

[0001] 本公开涉及人工智能领域，具体涉及自然语言处理、深度学习、计算机视觉、图像处理、增强现实和虚拟现实等技术领域，可应用于元宇宙等场景，特别涉及一种生成数字人的方法、一种神经网络的训练方法、一种视频生成装置、一种神经网络的训练装置、电子设备、计算机可读存储介质和计算机程序产品。

背景技术

[0002] 人工智能是研究使计算机来模拟人的某些思维过程和智能行为(如学习、推理、思考、规划等)的学科，既有硬件层面的技术也有软件层面的技术。人工智能硬件技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理等技术；人工智能软件技术主要包括计算机视觉技术、语音识别技术、自然语言处理技术以及机器学习/深度学习、大数据处理技术、知识图谱技术等几大方向。

[0003] 数字人是利用计算机技术对人体的形态和功能进行虚拟仿真的技术。数字人能够显著提升应用的交互性，增强智能信息服务的智能化水平。随着人工智能技术的不断突破，数字人的形象、表情、表达正在逐渐比拟真人，数字人的应用场景不断拓宽，数字人逐渐成为了数字世界的一种重要业务形态。

[0004] 在此部分中描述的方法不一定是之前已经设想到或采用的方法。除非另有指明，否则不应假定此部分中描述的任何方法仅因其包括在此部分中就被认为是现有技术。类似地，除非另有指明，否则此部分中提及的问题不应认为在任何现有技术中已被公认。

发明内容

[0005] 本公开提供了一种生成数字人的方法、一种神经网络的训练方法、一种视频生成装置、一种神经网络的训练装置、电子设备、计算机可读存储介质和计算机程序产品。

[0006] 根据本公开的一方面，提供了一种生成数字人的方法，包括：获取素材内容；基于预训练的场景划分模型，从素材内容中确定多个场景，其中，多个场景中的每个场景分别对应于素材内容中的一个具有完整语义信息的内容片段；以及对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容；基于对应的目标内容，确定该场景的场景标签信息；以及基于场景标签信息，配置特定于该场景的数字人。

[0007] 根据本公开的另一方面，提供了一种场景划分模型的训练方法，包括：获取样本素材内容和样本素材内容中的多个样本场景；基于预设场景划分模型，从样本素材内容中确定多个预测场景；以及基于多个样本场景和多个预测场景调整预设场景划分模型的参数，以得到训练后的场景划分模型。

[0008] 根据本公开的另一方面，提供了一种生成数字人的装置，装置包括：第一获取单元，被配置为获取素材内容；第一确定单元，被配置为基于预训练的场景划分模型，从素材内容中确定多个场景，其中，多个场景中的每个场景分别对应于素材内容中的一个具有完整语义信息的内容片段；第二确定单元，被配置为对于多个场景中的每个场景，基于对应的

内容片段，确定该场景对应的目标内容；第三确定单元，被配置为基于对应的目标内容，确定该场景的场景标签信息；以及数字人配置单元，被配置为基于场景标签信息，配置特定于该场景的数字人。

[0009] 根据本公开的另一方面，提供了一种场景划分模型的训练装置，包括：第三获取单元，被配置为获取样本素材内容和样本素材内容中的多个样本场景；第七确定单元，被配置为基于预设场景划分模型，从样本素材内容中确定多个预测场景；以及训练单元，被配置为基于多个样本场景和多个预测场景调整预设场景划分模型的参数，以得到训练后的场景划分模型。

[0010] 根据本公开的另一方面，提供了一种电子设备，包括：至少一个处理器；以及与至少一个处理器通信连接的存储器；其中存储器存储有可被至少一个处理器执行的指令，这些指令被至少一个处理器执行，以使至少一个处理器能够执行上述方法。

[0011] 根据本公开的另一方面，提供了一种存储有计算机指令的非瞬时计算机可读存储介质，其中，计算机指令用于使计算机执行上述方法。

[0012] 根据本公开的另一方面，提供了一种计算机程序产品，包括计算机程序，其中，计算机程序在被处理器执行时实现上述方法。

[0013] 根据本公开的一个或多个实施例，通过对素材内容进行场景切分，并以场景为粒度进行数字人的配置，从而确保了数字人与场景和目标内容的一致性，改善了素材内容和数字人之间的融合，提升了用户观看数字人的体验。

[0014] 应当理解，本部分所描述的内容并非旨在标识本公开的实施例的关键或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的说明书而变得容易理解。

附图说明

[0015] 附图示例性地示出了实施例并且构成说明书的一部分，与说明书的文字描述一起用于讲解实施例的示例性实施方式。所示出的实施例仅出于例示的目的，并不限制权利要求的范围。在所有附图中，相同的附图标记指代类似但不一定相同的要素。

[0016] 图1示出了根据本公开的实施例的可以在其中实施本文描述的各种方法的示例性系统的示意图；

[0017] 图2示出了根据本公开的实施例的生成数字人的方法的流程图；

[0018] 图3示出了根据本公开的实施例的从素材内容中确定多个场景的流程图；

[0019] 图4示出了根据本公开的实施例的确定每个场景对应的目标内容的流程图；

[0020] 图5示出了根据本公开的实施例的生成数字人的方法的流程图；

[0021] 图6示出了根据本公开的实施例的场景划分模型的训练方法的流程图；

[0022] 图7示出了根据本公开的实施例的生成数字人的装置的结构框图；

[0023] 图8示出了根据本公开的实施例的第一确定单元的结构框图；

[0024] 图9示出了根据本公开的实施例的第二确定单元的结构框图；

[0025] 图10示出了根据本公开的实施例的生成数字人的装置的结构框图；

[0026] 图11示出了根据本公开的实施例的场景划分模型的训练装置的结构框图；

[0027] 图12示出了能够用于实现本公开的实施例的示例性电子设备的结构框图。

具体实施方式

[0028] 以下结合附图对本公开的示范性实施例做出说明，其中包括本公开实施例的各种细节以助于理解，应当将它们认为仅仅是示范性的。因此，本领域普通技术人员应当认识到，可以对这里描述的实施例做出各种改变和修改，而不会背离本公开的范围。同样，为了清楚和简明，以下的描述中省略了对公知功能和结构的描述。

[0029] 在本公开中，除非另有说明，否则使用术语“第一”、“第二”等来描述各种要素不意图限定这些要素的位置关系、时序关系或重要性关系，这种术语只是用于将一个要素与另一要素区分开。在一些示例中，第一要素和第二要素可以指向该要素的同一实例，而在某些情况下，基于上下文的描述，它们也可以指代不同实例。

[0030] 在本公开中对各种所述示例的描述中所使用的术语只是为了描述特定示例的目的，而并非旨在进行限制。除非上下文另外明确地表明，如果不特意限定要素的数量，则该要素可以是一个也可以是多个。此外，本公开中所使用的术语“和/或”涵盖所列出的项目中的任何一个以及全部可能的组合方式。

[0031] 视频是数字世界最重要的信息载体之一，自然而然，数字人在视频生产中有着重要的应用空间。目前，数字人已经开始用于视频生产，比如通过数字人进行新闻播报，利用数字人形象进行宣传。然而，相关技术中，数字人在视频中的运用主要基于模板进行，比如固定数字人进行播报，数字人播报时可能会出现数字人与内容割裂，播报内容与数字人形象不匹配，用户体验差。另一些相关技术重点关注数字人偶像的精细构建，其目的以展示数字人形象为主。这种方式通常面向一些虚构、科幻的场景，难以用于真实信息的播报。此外，由于这种场景主要是在展示形象，数字人的各种属性通常与播报的内容无关。

[0032] 为解决上述问题，本公开通过对素材内容进行场景切分，并以场景为粒度进行数字人的配置，从而确保了数字人与场景和目标内容的一致性，改善了素材内容和数字人之间的融合，提升了用户观看数字人的体验。

[0033] 下面将结合附图详细描述本公开的实施例。

[0034] 图1示出了根据本公开的实施例可以将本文描述的各种方法和装置在其中实施的示例性系统100的示意图。参考图1，该系统100包括一个或多个客户端设备101、102、103、104、105和106、服务器120以及将一个或多个客户端设备耦接到服务器120的一个或多个通信网络110。客户端设备101、102、103、104、105和106可以被配置为执行一个或多个应用程序。

[0035] 在本公开的实施例中，服务器120可以运行使得能够执行生成数字人的方法和/或场景划分模型的训练方法的一个或多个服务或软件应用。

[0036] 在某些实施例中，服务器120还可以提供其他服务或软件应用，这些服务或软件应用可以包括非虚拟环境和虚拟环境。在某些实施例中，这些服务可以作为基于web的服务或云服务提供，例如在软件即服务(SaaS)模型下提供给客户端设备101、102、103、104、105和/或106的用户。

[0037] 在图1所示的配置中，服务器120可以包括实现由服务器120执行的功能的一个或多个组件。这些组件可以包括可由一个或多个处理器执行的软件组件、硬件组件或其组合。操作客户端设备101、102、103、104、105和/或106的用户可以依次利用一个或多个客户端应用程序来与服务器120进行交互以利用这些组件提供的服务。应当理解，各种不同的系统配

置是可能的，其可以与系统100不同。因此，图1是用于实施本文所描述的各种方法的系统的一个示例，并且不旨在进行限制。

[0038] 用户可以使用客户端设备101、102、103、104、105和/或106来输入与生成数字人相关的参数。客户端设备可以提供使客户端设备的用户能够与客户端设备进行交互的接口。客户端设备还可以经由该接口向用户输出信息，例如，向用户输出数字人生成结果。尽管图1仅描绘了六种客户端设备，但是本领域技术人员将能够理解，本公开可以支持任何数量的客户端设备。

[0039] 客户端设备101、102、103、104、105和/或106可以包括各种类型的计算机设备，例如便携式手持设备、通用计算机(诸如个人计算机和膝上型计算机)、工作站计算机、可穿戴设备、智能屏设备、自助服务终端设备、服务机器人、游戏系统、瘦客户端、各种消息收发设备、传感器或其他感测设备等。这些计算机设备可以运行各种类型和版本的软件应用程序和操作系统，例如MICROSOFT Windows、APPLE iOS、类UNIX操作系统、Linux或类Linux操作系统(例如GOOGLE Chrome OS)；或包括各种移动操作系统，例如MICROSOFT Windows Mobile OS、iOS、Windows Phone、Android。便携式手持设备可以包括蜂窝电话、智能电话、平板电脑、个人数字助理(PDA)等。可穿戴设备可以包括头戴式显示器(诸如智能眼镜)和其他设备。游戏系统可以包括各种手持式游戏设备、支持互联网的游戏设备等。客户端设备能够执行各种不同的应用程序，例如各种与Internet相关的应用程序、通信应用程序(例如电子邮件应用程序)、短消息服务(SMS)应用程序，并且可以使用各种通信协议。

[0040] 网络110可以是本领域技术人员熟知的任何类型的网络，其可以使用多种可用协议中的任何一种(包括但不限于TCP/IP、SNA、IPX等)来支持数据通信。仅作为示例，一个或多个网络110可以是局域网(LAN)、基于以太网的网络、令牌环、广域网(WAN)、因特网、虚拟网络、虚拟专用网络(VPN)、内部网、外部网、区块链网络、公共交换电话网(PSTN)、红外网络、无线网络(例如蓝牙、WIFI)和/或这些和/或其他网络的任意组合。

[0041] 服务器120可以包括一个或多个通用计算机、专用服务器计算机(例如PC(个人计算机)服务器、UNIX服务器、中端服务器)、刀片式服务器、大型计算机、服务器群集或任何其他适当的布置和/或组合。服务器120可以包括运行虚拟操作系统的一个或多个虚拟机，或者涉及虚拟化的其他计算架构(例如可以被虚拟化以维护服务器的虚拟存储设备的逻辑存储设备的一个或多个灵活池)。在各种实施例中，服务器120可以运行提供下文所描述的功能的一个或多个服务或软件应用。

[0042] 服务器120中的计算单元可以运行包括上述任何操作系统以及任何商业上可用的服务器操作系统的一个或多个操作系统。服务器120还可以运行各种附加服务器应用程序和/或中间层应用程序中的任何一个，包括HTTP服务器、FTP服务器、CGI服务器、JAVA服务器、数据库服务器等。

[0043] 在一些实施方式中，服务器120可以包括一个或多个应用程序，以分析和合并从客户端设备101、102、103、104、105和/或106的用户接收的数据馈送和/或事件更新。服务器120还可以包括一个或多个应用程序，以经由客户端设备101、102、103、104、105和/或106的一个或多个显示设备来显示数据馈送和/或实时事件。

[0044] 在一些实施方式中，服务器120可以为分布式系统的服务器，或者是结合了区块链的服务器。服务器120也可以是云服务器，或者是带人工智能技术的智能云计算服务器或智

能云主机。云服务器是云计算服务体系中的一项主机产品，以解决传统物理主机与虚拟专用服务器(VPS,Virtual Private Server)服务中存在的管理难度大、业务扩展性弱的缺陷。

[0045] 系统100还可以包括一个或多个数据库130。在某些实施例中，这些数据库可以用于存储数据和其他信息。例如，数据库130中的一个或多个可用于存储诸如音频文件和视频文件的信息。数据库130可以驻留在各种位置。例如，由服务器120使用的数据库可以在服务器120本地，或者可以远离服务器120且可以经由基于网络或专用的连接与服务器120通信。数据库130可以是不同的类型。在某些实施例中，由服务器120使用的数据库例如可以是关系数据库。这些数据库中的一个或多个可以响应于命令而存储、更新和检索到数据库以及来自数据库的数据。

[0046] 在某些实施例中，数据库130中的一个或多个还可以由应用程序使用来存储应用程序数据。由应用程序使用的数据库可以是不同类型的数据库，例如键值存储库，对象存储库或由文件系统支持的常规存储库。

[0047] 图1的系统100可以以各种方式配置和操作，以使得能够应用根据本公开所描述的各种方法和装置。

[0048] 根据本公开的一方面，提供了一种生成数字人的方法。如图2所示，该方法包括：步骤S201、获取素材内容；步骤S202、基于预训练的场景划分模型，从素材内容中确定多个场景，其中，多个场景中的每个场景分别对应于素材内容中的一个具有完整语义信息的内容片段；步骤S203、对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容；步骤S204、基于对应的目标内容，确定该场景的场景标签信息；以及步骤S205、基于场景标签信息，配置特定于该场景的数字人。

[0049] 由此，通过对素材内容进行场景切分，并以场景为粒度进行数字人的配置，从而确保了数字人与场景和目标内容的一致性，改善了素材内容和数字人之间的融合，提升了用户观看数字人的体验。

[0050] 根据一些实施例，在开始进行数字人生成之前，可以支持用户通过应用终端(例如，图1中的客户端101-106中的任一个)进行基本配置选项的设置。具体可以进行的设置的内容如下。

[0051] 本公开的方法可以应用于多种场景，例如播报场景、解说场景、主持场景等等。可以理解的是，本公开中将主要以播报场景作为示例对本公开的各种方法进行说明，但并不意图限定本公开的保护范围。

[0052] 在一些实施例中，可以支持用户对素材内容的类型的进行选择和配置。素材内容的类型和对应的文件、地址或内容可以包括：(1)文本文档，即具体包含文本内容或图文内容的文档；(2)文章URL，即希望用来生成数字人的图文内容对应的网址；(3)主题关键词与描述，描述希望生成数字人的主题，比如可以包括实体词、搜索关键词、搜索问题等形式。在一些示例性实施例中，素材内容可以包括图像数据和视频数据中的至少一者以及文本数据，从而能够丰富数字人播报、主持、或解说的内容。

[0053] 在一些实施例中，可以支持用户对语音播报(TTS,Text to Speech)功能进行配置，包括选择是否开启语音播报功能、语音播报的声音(例如，性别、口音等)、音色、音量以及语速等。

[0054] 在一些实施例中，可以支持用户对背景音乐进行配置，包括选择是否添加背景音乐、添加的背景音乐的类型等。

[0055] 在一些实施例中，可以支持用户对数字人资产进行设置，包括在预设的数字人资产中选择希望出现或使用的数字人的形象，或通过自定义的方式生成数字人的形象，以丰富数字人资产。

[0056] 在一些实施例中，可以支持用户对数字人背景进行配置，包括选择是否添加数字人背景、数字人背景的类型(例如，图像或视频)等。

[0057] 在一些实施例中，可以支持用户对作为最终呈现结果的视频的生成方式进行配置，包括选择全自动视频生成、人机交互辅助视频生成等。

[0058] 可以理解的是，除上述内容外，输入配置还可以根据情况为用户提供更多的系统控制，比如文案压缩的比例、最终呈现的结果中所使用的动态素材占比等，在此不作限定。

[0059] 根据一些实施例，步骤S201、获取素材内容可以包括：基于下列方式中的至少一者，获取素材内容：基于网页地址，获取素材内容；或基于搜索关键词，获得素材内容。针对上述几种不同类型的素材内容，具体可以使用如下方式进行获取：

[0060] 针对文本文档：直接读取本地或远程存储的文本文档中的内容。

[0061] 针对文章URL：主要是指互联网中已有的图文内容，比如新闻文章、论坛文章、问答类页面、公众号文章等内容的URL，基于已有的网页解析开源方案，获取URL对应的网页数据，并解析获取URL网页中的主体文本和图片内容，同时记录标题、正文、段落、加粗、图文位置关系、表格等重要原始信息。这些信息在后续的数字人生成过程中会被使用，例如，标题可以用作检索视觉素材的query，正文、段落、以及加粗内容可以用于生成播报内容，并且可以用于提取场景关键词和句子级关键词，图文位置关系可以提供原文中的图像素材和播报内容之间的对应关系，表格可以丰富数字人呈现结果中的内容表现形式，这些内容会在下文中进行详细介绍。

[0062] 针对主题关键词与描述：本系统也支持基于用户输入的主题描述来生成最终的数字人呈现结果。用户输入的主题关键词与描述可以是类似于百科词条的实体词，还可以是多个关键词，或者类似于事件描述或者问题描述的形式。根据主题关键词获取图文内容的方式是：(1)将主题关键词或者描述输入搜索引擎，(2)获取多条搜索返回结果，(3)从搜索结果中选择相关性排序较高，同时视觉素材更丰富的图文结果作为待生成视频的主体URL，(4)按URL内容提取方式提取URL内的图文内容等信息。

[0063] 在获取到素材内容之后，可以基于该素材内容生成用于数字人播报的文案。根据前面获取的素材内容，通过语义理解与文本生成等技术，结合数字人播报的需要进行场景划分、文字转化与生成，输出数字人视频/全息投影制作所需要的脚本文案，同时也提供场景的划分和语义分析信息。对素材内容的上述处理是决定数字人融合方式的关键基础。具体来说，可以通过以下方式生成数字人播报的目标内容。

[0064] 在一些实施例中，在步骤S202、基于预训练的场景划分模型，从素材内容中确定多个场景。多个场景中的每个场景可以分别对应于素材内容中的一个具有完整语义信息的内容片段。通过将素材内容划分为多个场景，以便于后续数字人播报时，对不同场景做出不同的处理。在本公开中，数字人融合是以场景为粒度进行融合的。

[0065] 根据一些实施例，如图3所示，步骤S202、基于预训练的场景划分模型，从素材内容

中确定多个场景可以包括:步骤S301、通过对素材内容进行篇章结构分析和篇章语义分割,从素材内容中确定多个子主题,并且确定多个子主题之间的结构关系;以及步骤S302、基于结构关系,将多个子主题划分为多个场景。由此,通过使用篇章结构分析和篇章语义分割的方法,能够准确地从素材内容中确定语义完整的多个子主题及其之间的结构关系,而利用子主题之间的结构关系(例如,总分结构、并列结构、递进结构等),能够将多个子主题进一步划分为适用于数字人播报任务的多个场景。

[0066] 在一些实施例中,场景划分模型是通过预先训练而得到的,并且可以进一步包括篇章结构分析模型、篇章语义分割模型、以及将子主题划分为场景的模型。这些模型可以是基于规则的模型,也可以是基于深度学习的模型,在此不做限定。在一个示例性实施例中,篇章语义分割模型能够输出语义完整的内容片段和相邻的内容片段之间的语义关系(例如,每组相邻的两个句子之间的语义相似度),篇章结构分析模型能够输出内容片段之间的关系(例如,总分、并列、递进等),两者结合即能够得到将素材内容划分为多个子主题时得到的分割边界的位置和子主题之间的关系。进一步地,上述方法仅基于文本(自然语言处理)对素材内容进行划分,得到的划分结果可能并不适合直接作为用于播报的场景。例如,部分子主题对应的内容片段很短,可能仅包括一句话,而如果将这样的子主题作为场景可能会导致频繁转场。因此,可以将多个子主题进一步划分为多个场景。在一些实施例中,可以使用子主题之间的结构关系来对多个子主题进一步划分,也可以使用其他的方式(例如基于规则的方法或基于序列标注的神经网络方法)对多个子主题进行进一步划分,在此不作限定。

[0067] 在得到多个场景后,可以确定每个场景的目标内容(例如,播报词、解说词、主持词)。在一些实施例中,可以将素材内容中的具有完整语义信息的内容片段直接作为目标内容。在另一些实施例中,由于素材内容的来源复杂、内容较多且通常是书面语,因此直接将每个场景的内容片段作为数字人的目标内容可能效果较差。通过对素材内容和/或内容片段进行如下处理,可以得到相应的目标内容。

[0068] 根据一些实施例,可以对素材内容和/或内容片段进行文本压缩、文本改写、风格转化等处理。在一些实施例中,如图4所示,步骤S203、对于多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容可以包括:步骤S401、对对应的内容片段执行文本改写和文本压缩中的至少一种处理,以更新对应的内容片段。可以根据用户设定的时长、内容的特点,根据素材内容整体和/或每个内容片段的内容,为每个场景输出简练且富有信息量的文字结果作为目标内容。在一些实施例中,可以采用是抽取式摘要算法确定目标内容,也可以使用生成式摘要算法确定目标内容,还可以使用其他方法确定目标内容,在此不做限定。

[0069] 根据一些实施例,如图4所示,步骤S203、对于多个场景中的每个场景,基于对应的内容片段,确定该场景对应的目标内容还可以包括:步骤S402、基于该场景与前一场景之间的结构关系,生成用于该场景的第一内容。由此,通过基于场景之间的结构关系生成相应的第一内容,可以使得场景之间的转换更连贯,过渡更自然。

[0070] 在一些实施例中,第一内容例如可以为转场词,也可以为过渡句,还可以为其他能够连接多个场景或指示场景之间的关系的内容,在此不作限定。

[0071] 根据一些实施例,如图4所示,步骤S203、对于多个场景中的每个场景,基于对应的

内容片段，确定该场景对应的目标内容还可以包括：步骤S403、基于预训练的风格转换模型，将对应的内容片段转换为对应的目标内容，其中，风格转换模型是基于提示学习训练得到的。由此，通过进行风格转换，可以将素材内容和/或内容片段转换为适合数字人播报的风格，而通过使用基于提示学习(prompt learning)的方法，使得预训练的风格转换模型能够根据数字人的各种特点和需求(例如，口语化要求、数字人特点(性别、口音设定等)、内容转场需求)，自动输出自然的目标内容。在一些实施例中，可以将上述各种特定和需求转换为属性控制，进而和将要进行风格转换的文本(以及可选地，文本的上下文)共同构造基于提示学习的风格转换模型的输入，并利用具有上述结构(属性控制与文本)的样本对模型进行训练，使得模型能够输出期望的风格转换结果。在一些实施例中，也可以将基于规则的方法(加入口语词、转场词等)和基于学习的进行结合，从而将每个场景的内容片段转换为符合数字人播报情景的目标内容。

[0072] 根据一些实施例，如图4所示，步骤S203、对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容还可以包括：步骤S404、对经转换的目标内容执行文本改写和文本压缩中的至少一种处理，以更新对应的目标内容。步骤S404的操作和步骤S401的操作可以一致。可以理解的是，步骤S401和步骤S404可以择一执行，也可以两者均执行，在此不做限定。

[0073] 在一个示例性实施例中，通过对一篇名称为“打造‘博物馆之城’!推进X市6处博物馆建设”执行上述步骤可以得到例如表1所示的多个场景和对应的目标内容。

[0074] 表1

[0075]

场景序号	场景文案
0	打造‘博物馆之城’！推进 X 市 6 处博物馆建设
1	3 月 2 日，X 市举行“推进建设‘博物馆之城’ 助力全国文化中心建设”议政会情况通报会。市政府相关负责人介绍，将推进 A 博物馆、B 自然博物馆等 6 处博物馆的建设。
2	市政府相关负责人介绍，截至 2021 年末，X 市共有 204 家各类备案博物馆，免费开放博物馆达 94 家。X 市博物馆藏品总数已达 1625 万件套，可移动文物数量和三级以上珍贵文物数量均居于全国首位，持续开放基本陈列 520 个，每年平均举办展览 600 多项，开展活动上千项，年均接待观众超过 5000 万人次。
3	市政府相关负责人介绍，将积极发挥市政府固定资产投资作用，强化资源整合与协同创新，加大博物馆建设支持力度。
4	目前正在建设的 X 市博物馆东馆已实现外立面亮相，2022 年底将基本完工；城墙遗址保护展示工程项目正在加快建设，2023 年将投

[0076]

	入使用；XX 博物馆改建工程、XX 遗迹保护工程、A 博物馆改造提升工程、B 自然博物馆项目正在加快推进前期工作。
5	这些博物馆的实施将进一步优化 X 市博物馆布局，推动形成布局合理、结构优化、特色鲜明、功能完备的博物馆事业发展新格局。

[0077] 在得到了每个场景对应的目标内容后，可以对每个场景和对应的目标内容进行进一步的语义分析，以得到用于视频素材检索与召回、用于视频素材与文案对齐、用于进行关键信息展示、和/或用于控制数字人的生成过程的丰富的场景相关信息。具体地，可以包括如下几种语义分析方式。

[0078] 在一些实施例中，可以进行场景关键词抽取，即自动抽取可以用于描述整个场景核心内容的关键词。如上例中，针对场景1，可以抽取针对该场景的关键词为“博物馆之城”、“X市”等。这些场景关键词将应用于构建素材检索query，如“X市博物馆之城”等，用于召回相关视频素材。

[0079] 在一些实施例中，可以进行句子级关键词抽取，即从句子级别出发，抽取更细粒度的关键词。如上例中，针对场景4中的句子“目前正在建设的X市博物馆东馆已实现外立面亮相，2022年底将基本完工”，可以自动抽取出句子级的关键词“X市博物馆东馆”，该句子级关键词除了被用于构建素材检索query外，还会被应用于做更加精准的视频素材与内容(文案)的对齐。

[0080] 在一些实施例中，可以进行信息抽取，以得到内容中所表达的键-值形式的结果(即，Key-Value对)。如上例中，针对场景2，可以自动抽取到多个Key-Value对，如“X市各类备案博物馆数量:204家”、“免费开放博物馆数量:94家”、“X市地区博物馆藏品总数:1625万件套”等。利用抽取到的这些关键信息，可以自动生成更加精准且丰富的数字人播报素材，例如数字人播报场景的背景中的展示板，从而能够显著提升数字人播报的整体质量和呈现效果。根据一些实施例，生成数字人的方法还可以包括：对于多个场景中的每个场景，从该场景对应的目标内容中提取键-值形式的信息；以及基于键-值形式的信息，生成用于最终的数字人呈现结果的辅助素材。由此，可以得到更加精准丰富的数字人视觉素材。

[0081] 在一些实施例中，可以进行情感分析，以用于对每个素材所表达的核心情感倾向进行输出。如上例中，针对场景5，整个场景在介绍X市博物馆事业蓬勃发展的情况，整体基调、情感和情绪是积极向上的。利用这类情感分析的结果，可以对数字人的表情、语气、动作进行更加丰富的控制。根据一些实施例，场景标签信息可以包括语义标签，步骤S204、基于对应的目标内容，确定该场景的场景标签信息可以包括：对对应的目标内容进行情感分析，以获得语义标签。由此，通过对目标内容进行情感分析，可以得到场景的基调与情感等信息。

[0082] 在一些实施例中，语义标签用于标识对应的目标内容所表达的情感可以包括：积极、中性或消极。可以理解的是，语义标签可以标识更丰富的情感，例如紧张、愉悦、愤怒等。在此不作限定。在一些实施例中，语义标签还可以包括其他内容，例如可以包括直接提取目标内容的文本语义特征、目标内容的类型(比如叙事型、评论型、抒情型等)、以及其他的能够体现场景和对应的目标内容的语义相关的信息的标签，在此不做限定。

[0083] 在一些实施例中，除上述方式外，还可以以其他方式对场景和对应的目标内容进

行情感分析，以得到相关信息。例如，可以直接提取目标内容的文本语义特征，用于作为后续数字人属性规划的输入信息。

[0084] 在得到了每个场景对应的目标内容后，还可以进行语音合成。语音合成的目的是生成用于数字人场景的声音，即将前面得到目标内容转化为语音，并且可选地，为数字人播报增加背景音乐。文本转换为语音可以通过调用TTS服务实现。在一些实施例中，如图5所示，生成数字人的方法还可以包括：步骤S505、将目标内容转换成语音，用于数字人播报。可以理解的是，图5中的步骤S501-步骤S504、以及步骤S507与图2中的步骤S201-步骤S205的操作类似，在此不作赘述。

[0085] 关于背景音乐，系统可以根据目标内容的类型(比如叙事型、评论型、抒情型等)，调用不同音调、音色的TTS能力和不同风格的背景音乐。此外，如上所述本公开的方法支持用户指定TTS、背景音乐等。系统可以提供多种音调、音色、语速的TTS和背景音乐让用户自主选择，可以支持用户主动定制专属的TTS音色。

[0086] 为了制作包含丰富视觉素材的数字人播报呈现结果，可以进行素材扩充，以为数字人播报补充视频、图像等素材。视频和图像素材的补充包括如下内容。

[0087] 根据一些实施例，如图5所示，生成数字人的方法还可以包括：步骤S506、对于多个场景中的每个场景，基于素材内容和该场景对应的目标内容，检索与该场景相关的视频素材；以及步骤S508、将视频素材和数字人相结合。由此，通过这种方式，可以检索到与场景以及对应的目标内容一致且密切相关的素材，从而丰富了数字人播报呈现结果中的视觉内容，提升了用户的观看体验。

[0088] 在一些实施例中，可以根据素材内容的标题以及前面描述的场景关键词、句子级关键词等，构建一个或者多个搜索关键词，进而通过在线全网图片/视频搜索以及离线图片/视频库获取内容相关的视频。随后可以通过例如视频拆条算法等方式将获得的视频内容进行拆分，以得到候选的视觉素材片段。可以理解的是，在实施本公开的方法时，可以使用各种方式进行视频检索和视频拆分，以得到候选视觉素材片段，在此不作限定。

[0089] 根据一些实施例，步骤S506、对于多个场景中的每个场景，基于素材内容和该场景对应的目标内容，检索与该场景相关的视频素材可以包括：提取场景关键词；以及基于场景关键词，检索与该场景相关的视频素材。由此，通过上述方式，能够得到与场景整体相关的视频素材，从而丰富了可使用的视频素材。

[0090] 根据一些实施例，步骤S506、对于多个场景中的每个场景，基于素材内容和该场景对应的目标内容，检索与该场景相关的视频素材可以包括：提取句子级关键词；以及基于句子级关键词，检索与该场景相关的视频素材。由此，通过上述方式，能够得到与目标内容中的句子相关的视频素材，从而丰富了可使用的视频素材。

[0091] 在一些实施例中，可以根据目标内容中的结构化数据生成动态报表，也可以基于深度学习模型进行文字生成图片和/或视频的方法对目标内容进行处理，以进一步提升了视频的素材丰富度。

[0092] 在一些实施例中，在得到目标内容、语音、以及与场景和目标内容对应的图像、视频素材后，可以将这些视觉素材和文字、语音进行对齐，以便在渲染生成阶段进行合成。具体实现上，主要是基于预训练模型的图文匹配，对文本和视觉素材进行相关性计算和排序，给每段文本找到对应的字幕、TTS语音时间段内相匹配的视频和图片内容，对齐语音-字幕-

视频时间轴。可以理解的是，用户可以在此过程中对时间轴进行调整，以实现手动对齐。

[0093] 根据一些实施例，生成数字人的方法还可以包括：基于句子级关键词，将检索到的视频素材和目标内容对齐。由于基于句子级关键词检索到的视频素材可能是和目标内容中的某个句子对应的，因此一段目标内容可能与多个基于句子级关键词检索到的视频素材对应。在这样的实施例中，可以利用目标内容中的多个句子级关键词将各自对应的视频素材和各自对应的句子进行对齐。

[0094] 在一个示例性实施例中，如表1所示的例子中，场景4中的多个句子的句子级关键词分别为“X市博物馆东馆”“城墙遗址”“XX博物馆;XX遗迹;A博物馆;B自然博物馆”，在基于每个句子对应的句子级关键词进行检索后，可以得到相应的图像或视频素材。进而，可以基于句子级关键词和多个句子之间的对应关系，以及句子级关键词和这些视频素材之间的对应关系，将视频素材和句子进行对齐，使得数字人在读每个句子的时候，场景中能够显示相应的视频素材，并且在句子之间的间隙完成视频素材的切换。

[0095] 可以理解的是，如果原始的素材内容中具有充足的素材，或者确定数字人生成结果不需要与视觉素材相结合，则可以不进行素材扩充。

[0096] 至此，数字人合成的前序工作全部执行完毕。数字人合成可以根据前面场景分析，以及可选地，素材补充的结果，为每个场景规划合适的数字人呈现方式，从而确保最终呈现的结果具有很好的交互性，给予用户良好的沉浸式体验。数字人合成具体包括如下内容。

[0097] 在一些实施例中，可以判断是否触发数字人，也即，是否为当前场景生成数字人。触发的核心考虑因素，主要是场景在视频中的位置以及场景对应的素材丰富度。这里的素材丰富度主要是指高相关的动态素材对整个场景播放时长的占比。除上述因素外，素材的清晰度、连贯性、素材与场景的相关度均可以作为数字人是否触发的因素。基于上述这些因素，系统可以接受用户定义规则，也支持基于机器学习方法自动判断是否触发。可以理解的是，本公开并不限定具体的触发逻辑，在实施本公开的方法时可以根据需求按照上述方式设置相应的触发逻辑，或使用满足相应的触发逻辑的样本对机器学习模型进行训练，在此不作限定。

[0098] 根据一些实施例，生成数字人的方法还包括：确定视频素材相对应的场景所需的播放时长的占比；以及基于占比，确定是否在相应场景中触发数字人。由此，实现了基于视频素材的丰富度判断是否触发数字人。

[0099] 在一些实施例中，在确定触发数字人后，可以进行数字人属性规划。数字人属性规划主要根据一系列内容特征确定数字人的服饰、姿态、动作、表情、背景等属性。内容特征可以包括场景的语义标签(例如，目标内容的基调、情感、语义特征、类型等等)、播报内容中的关键触发词、视觉素材的内容特点等。具体实现上，既可以使用基于规则的方法，根据上述内容特征确定数字人属性，同时也支持基于深度学习方法，以将内容特征作为输入以预测数字人属性配置。

[0100] 在一些实施例中，可以建立关键触发词和特定的数字人姿态、动作、或表情之间的映射关系，以使得在检测到关键触发词后，数字人做出或有一定概率做出相应的姿态、动作、或表情。可以理解的是，可以使用基于规则的方式使得在检测到关键触发词后，数字人一定做出特定的反应，也可以让模型学习大量的样本，以得到关键触发词和特定的数字人姿态、动作、或表情之间的关系，在此不做限定。

[0101] 在一些实施例中，视觉素材的内容特点，例如素材的清晰度、连贯性、素材与场景的相关度也可以作为数字人属性规划时所考虑的内容特征。此外，视觉素材中的特定内容可以触发数字人的特定姿态、动作、表情等。根据一些实施例，生成数字人的方法还可以包括：响应于确定视频素材中包括特定素材，基于特定素材在视频素材中的显示位置，确定数字人的动作。由此，通过对视频素材中的内容进行分析，并将视频素材的特定素材和数字人的动作进行结合，能够进一步提升视频素材、目标内容、数字人三者之间的一致性，并提升用户的观看体验。

[0102] 在一些实施例中，特定素材例如可以包括表格、图例、画中画等等。

[0103] 在一些实施例中，利用篇章分析的结果，能够进一步获得每个场景在素材内容(原始文本内容或图文内容)中的作用，比如主旨段落、总结段落等。根据这些信息我们可以对数字人的运镜、动作，以及数字人的具体展现形式(演播间、画中画等)进行更丰富的控制。

[0104] 在一些实施例中，通过情感分析而得到的场景的语义标签(例如，场景的基调和情感)可以作为数字人属性规划所考虑的内容特征。根据一些实施例，步骤S507、对于多个场景中的每个场景，基于场景标签信息，配置特定于该场景的数字人可以包括：基于语义标签，配置数字人的服饰、表情和动作中的至少一者。

[0105] 场景的语义表情还可以用于确定数字人的语气。根据一些实施例，步骤S507、对于多个场景中的每个场景，基于场景标签信息，配置特定于该场景的数字人可以包括：基于语义标签，配置数字人语音的语气。在一些实施例中，数字人语音的语气例如可以包括数字人语音的音量、音高、语调等等。此外，这些语义标签还可以在数字人配置过程中产生其他用途，例如为场景配置更加风格合适的演播室背景等等。

[0106] 在一些实施例中，数字人的属性可以包括数字人的服饰、姿态、动作、表情、背景等属性，具体到针对每个属性规划，可以首先人工为每类数字人属性设置选项。例如对于“服饰”属性，分别人工定义不同的选项“正装”、“休闲装”、“衬衣”等。基于这些人工给定的类别体系，我们使用机器学习中的分类算法，在人工标注的训练数据上，通过建模各类特征，如文本特征(词特征、短语特征、句子特征等)、数字人ID、数字人性别等，使得模型基于该特征拟合人工标注信号，至模型训练收敛。在预测阶段，在抽取的特征上进行模型预测，即可以得到对不同属性的规划。

[0107] 在一些实施例中，在得到数字人生成结果后，可以进行数字人生成结果的呈现。生成数字人的方法还可以包括：以全息图像的形式呈现数字人。由此，提供了一种能够保证数字人与目标内容一致的数字人生成结果。

[0108] 在一些实施例中，在得到视频素材和数字人生成结果后，可以执行步骤S508将视频素材和数字人相结合，进而进行视频渲染，以得到最终的视频。如图5所示，生成数字人的方法还可以包括：步骤S509、以视频的形式呈现数字人。由此，提供了另一种能够保证数字人与目标内容一致的数字人生成结果，并且提升了视频生成结果的生动性，提升了用户体验的沉浸感，同时能够充分发挥数字人的优势，利用数字人和视频素材的交互弥补相关素材不足的问题。此外，本公开的方法面向通用场景设计，能够兼容不同的内容类型，具备全领域运用的通用型。

[0109] 在一些实施例中，本公开支持端到端自动生成，同时也支持用户进行交互生成。也就是用户可以对生成的视频结果进行调整。在交互生成场景下，用户可以对视频的各个要

素进行调整，比如文字、语音、素材、虚拟人配置等等。交互生成的方式，一方面支持用户修改，从而生成质量更好的结果。同时用户交互生成的数据也将被记录下来，作为系统优化的反馈数据，从而指导各个步骤模型的学习，不断提升系统的效果。

[0110] 根据本公开的另一方面，提供了一种场景划分模型的训练方法。如图6所示，训练方法包括：步骤S601、获取样本素材内容和样本素材内容中的多个样本场景；步骤S602、基于预设场景划分模型，从样本素材内容中确定多个预测场景；以及步骤S603、基于多个样本场景和多个预测场景调整预设场景划分模型的参数，以得到训练后的场景划分模型。由此，通过上述方式能够实现对场景划分模型的训练，从而使得利用训练好的模型能够输出准确的场景划分结果，以提升利用该场景划分结果进行数字人生成而得到的最终呈现结果的用户观看体验。

[0111] 可以理解的是，样本素材内容和步骤S201中所获取的素材内容类似。样本素材内容中的多个样本场景可以是由人工或者使用基于模板或神经网络模型的方法对样本素材内容进行划分而得到的，其可以作为对样本素材内容进行划分的真实结果(ground truth)。通过利用预测结果和真实结果进行训练，可以使得训练后的场景划分模型具有对素材内容进行场景划分的能力。可以理解的是，在实施本公开的方法时，可以根据需求选择相应的神经网络作为场景划分模型，并使用相应的损失函数进行训练，在此不作限定。

[0112] 根据一些实施例，预设场景划分模型可以包括篇章语义分割模型和篇章结构分析模型。步骤S602、基于预设场景划分模型，从样本素材内容中确定多个预测场景包括：利用篇章语义分割模型和篇章结构分析模型对样本素材内容进行处理，以确定素材内容中多个预测子主题以及多个预测子主题之间的预测结构关系；以及基于预测结构关系，将多个预测子主题划分为多个预测场景。由此，通过对篇章结构分析模型和篇章语义分割模型进行训练，使得训练好的模型能够准确地从素材内容中确定语义完整的多个子主题及其之间的结构关系，而利用子主题之间的结构关系(例如，总分结构、并列结构、递进结构等)，能够将多个子主题进一步划分为适用于数字人播报任务的多个场景。

[0113] 可以理解的是，除了场景划分模型外，还可以对图1或图5中的方法所使用的其他模型进行训练，例如，对将子主题划分为场景的模型、用于生成目标内容的模型(用于文本改写、文本压缩、和/或风格转化的模型)、用于对场景进行情感分析的模型、场景关键词/句子级关键词抽取模型、数字人属性规划模型等模型进行训练。在一些实施例中，可以使用标注语料或用户交互数据对包括场景划分模型在内的这些模型进行训练。

[0114] 根据本公开的另一方面，提供了一种生成数字人的装置。如图7所示，生成数字人的装置700包括：第一获取单元702，被配置为获取素材内容；第一确定单元704，被配置为基于预训练的场景划分模型，从素材内容中确定多个场景，其中，多个场景中的每个场景分别对应于素材内容中的一个具有完整语义信息的内容片段；第二确定单元706，被配置为对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容；第三确定单元708，被配置为基于对应的目标内容，确定该场景的场景标签信息；以及数字人配置单元710，被配置为基于场景标签信息，配置特定于该场景的数字人。可以理解的是，装置700中的单元702-单元710的操作分别和图2中的步骤S201-步骤S205的操作类似，在此不作赘述。

[0115] 根据一些实施例，素材内容可以包括图像数据和视频数据中的至少一者以及文本数据。

[0116] 根据一些实施例，第一获取单元702可以被进一步配置为基于下列方式中的至少一者，获取素材内容：基于网页地址，获取素材内容；或基于搜索关键词，获得素材内容。

[0117] 根据一些实施例，如图8所示，第一确定单元800包括：第一确定子单元802，被配置为通过对素材内容进行篇章结构分析和篇章语义分割，从素材内容中确定多个子主题，并且确定多个子主题之间的结构关系；以及第一划分子单元804，被配置为基于结构关系，将多个子主题划分为多个场景。

[0118] 根据一些实施例，如图9所示，第二确定单元900可以包括以下中的至少一项：第一更新子单元902，被配置为对对应的内容片段执行文本改写和文本压缩中的至少一种处理，以更新对应的内容片段；以及第二更新子单元908，被配置为对经转换的内容执行文本改写和文本压缩中的至少一种处理，以更新对应的内容。

[0119] 根据一些实施例，如图9所示，第二确定单元900可以包括：生成子单元904，被配置为基于该场景与前一场景之间的结构关系，生成用于该场景的第一内容。

[0120] 根据一些实施例，如图9所示，第二确定单元900可以包括：转换子单元906，被配置为基于预训练的风格转换模型，将对应的内容片段转换为对应的目标内容，其中，风格转换模型是基于提示学习训练得到的。

[0121] 根据一些实施例，生成数字人的装置700还可以包括：提取单元，被配置为对于多个场景中的每个场景，从该场景对应的目标内容中提取键-值形式的信息；以及生成单元，被配置为基于键-值形式的信息，生成用于视频的辅助素材。

[0122] 根据一些实施例，第三确定单元708可以包括：情感分析子单元，被配置为对对应的目标内容进行情感分析，以获得语义标签。

[0123] 根据一些实施例，语义标签可以用于标识对应的目标内容所表达的情感包括：积极、中性或消极。

[0124] 根据一些实施例，如图10所示，生成数字人的装置1000可以包括：语音转换单元1010，被配置为将目标内容转换成语音，用于数字人播报。装置1000中的单元1002-单元1008、以及单元1014的操作分别和装置700中的单元702-单元710的操作类似，在此不作赘述。

[0125] 根据一些实施例，如图10所示，生成数字人的装置1000可以包括：检索单元1012，被配置为对于多个场景中的每个场景，基于素材内容和该场景对应的目标内容，检索与该场景相关的视频素材；以及结合单元1016，被配置为将视频素材和数字人相结合。

[0126] 根据一些实施例，检索单元1012可以包括：第一提取子单元，被配置为提取场景关键词；以及第一检索子单元，被配置为基于场景关键词，检索与该场景相关的视频素材。

[0127] 根据一些实施例，检索单元1012还可以包括：第二提取子单元，被配置为提取句子级关键词；以及第二检索子单元，被配置为基于句子级关键词，检索与该场景相关的视频素材。

[0128] 根据一些实施例，生成数字人的装置1000还可以包括：对齐单元，被配置为基于句子级关键词，将检索到的视频素材和目标内容对齐。

[0129] 根据一些实施例，生成数字人的装置1000还可以包括：第五确定单元，被配置为确定视频素材相对应的场景所需的播放时长的占比；以及第六确定单元，被配置为基于占比，确定是否在相应场景中触发数字人。

[0130] 根据一些实施例，生成数字人的装置1000还可以包括：第四确定单元，被配置为响应于确定视频素材中包括特定素材，基于特定素材在视频素材中的显示位置，确定数字人的动作。

[0131] 根据一些实施例，数字人配置单元1014可以包括：第一配置子单元，被配置为基于语义标签，配置数字人的服饰、表情和动作中的至少一者。

[0132] 根据一些实施例，数字人配置单元1014还可以包括：第二配置子单元，被配置为基于语义标签，配置数字人语音的语气。

[0133] 根据一些实施例，生成数字人的装置1000还可以包括：全息图像呈现单元，被配置为以全息图像的形式呈现数字人。

[0134] 根据一些实施例，生成数字人的装置1000还可以包括：视频呈现单元1018，被配置为以视频的形式呈现数字人。

[0135] 根据本公开的另一方面，提供了一种场景划分模型的训练装置。如图11所示，训练装置1100包括：第二获取单元1102，被配置为获取样本素材内容和样本素材内容中的多个样本场景；第七确定单元1104，被配置为基于预设场景划分模型，从样本素材内容中确定多个预测场景；以及训练单元1106，被配置为基于多个样本场景和多个预测场景调整预设场景划分模型的参数，以得到训练后的场景划分模型。

[0136] 根据一些实施例，预设场景划分模型可以包括篇章语义分割模型和篇章结构分析模型。第七确定单元1104可以包括：第二确定子单元，被配置为利用篇章语义分割模型和篇章结构分析模型对样本素材内容进行处理，以确定素材内容中多个预测子主题以及多个预测子主题之间的预测结构关系；以及第二划分子单元，被配置为基于预测结构关系，将多个预测子主题划分为多个预测场景。

[0137] 本公开的技术方案中，所涉及的用户个人信息的收集、存储、使用、加工、传输、提供和公开等处理，均符合相关法律法规的规定，且不违背公序良俗。

[0138] 根据本公开的实施例，还提供了一种电子设备、一种可读存储介质和一种计算机程序产品。

[0139] 参考图12，现将描述可以作为本公开的服务器或客户端的电子设备1200的结构框图，其是可以应用于本公开的各方面的硬件设备的示例。电子设备旨在表示各种形式的数字电子的计算机设备，诸如，膝上型计算机、台式计算机、工作台、个人数字助理、服务器、刀片式服务器、大型计算机、和其它适合的计算机。电子设备还可以表示各种形式的移动装置，诸如，个人数字处理、蜂窝电话、智能电话、可穿戴设备和其它类似的计算装置。本文所示的部件、它们的连接和关系、以及它们的功能仅仅作为示例，并且不意在限制本文中描述的和/或者要求的本公开的实现。

[0140] 如图12所示，电子设备1200包括计算单元1201，其可以根据存储在只读存储器(ROM)1202中的计算机程序或者从存储单元1208加载到随机访问存储器(RAM)1203中的计算机程序，来执行各种适当的动作和处理。在RAM 1203中，还可存储电子设备1200操作所需的各种程序和数据。计算单元1201、ROM 1202以及RAM 1203通过总线1204彼此相连。输入/输出(I/O)接口1205也连接至总线1204。

[0141] 电子设备1200中的多个部件连接至I/O接口1205，包括：输入单元1206、输出单元1207、存储单元1208以及通信单元1209。输入单元1206可以是能向电子设备1200输入信息

的任何类型的设备，输入单元1206可以接收输入的数字或字符信息，以及产生与电子设备的用户设置和/或功能控制有关的键信号输入，并且可以包括但不限于鼠标、键盘、触摸屏、轨迹板、轨迹球、操作杆、麦克风和/或遥控器。输出单元1207可以是能呈现信息的任何类型的设备，并且可以包括但不限于显示器、扬声器、视频/音频输出终端、振动器和/或打印机。存储单元1208可以包括但不限于磁盘、光盘。通信单元1209允许电子设备1200通过诸如因特网的计算机网络和/或各种电信网络与其他设备交换信息/数据，并且可以包括但不限于调制解调器、网卡、红外通信设备、无线通信收发机和/或芯片组，例如蓝牙TM设备、802.11设备、WiFi设备、WiMax设备、蜂窝通信设备和/或类似物。

[0142] 计算单元1201可以是各种具有处理和计算能力的通用和/或专用处理组件。计算单元1201的一些示例包括但不限于中央处理单元(CPU)、图形处理单元(GPU)、各种专用的人工智能(AI)计算芯片、各种运行机器学习模型算法的计算单元、数字信号处理器(DSP)、以及任何适当的处理器、控制器、微控制器等。计算单元1201执行上文所描述的各个方法和处理，例如生成数字人的方法和场景划分模型的训练方法。例如，在一些实施例中，生成数字人的方法和场景划分模型的训练方法可被实现为计算机软件程序，其被有形地包含于机器可读介质，例如存储单元1208。在一些实施例中，计算机程序的部分或者全部可以经由ROM 1202和/或通信单元1209而被载入和/或安装到电子设备1200上。当计算机程序加载到RAM 1203并由计算单元1201执行时，可以执行上文描述的生成数字人的方法和场景划分模型的训练方法的一个或多个步骤。备选地，在其他实施例中，计算单元1201可以通过其他任何适当的方式(例如，借助于固件)而被配置为执行生成数字人的方法和场景划分模型的训练方法。

[0143] 本文中以上描述的系统和技术的各种实施方式可以在数字电子电路系统、集成电路系统、场可编程门阵列(FPGA)、专用集成电路(ASIC)、专用标准产品(ASSP)、芯片上系统的系统(SOC)、复杂可编程逻辑设备(CPLD)、计算机硬件、固件、软件、和/或它们的组合中实现。这些各种实施方式可以包括：实施在一个或者多个计算机程序中，该一个或者多个计算机程序可在包括至少一个可编程处理器的可编程系统上执行和/或解释，该可编程处理器可以是专用或者通用可编程处理器，可以从存储系统、至少一个输入装置、和至少一个输出装置接收数据和指令，并且将数据和指令传输至该存储系统、该至少一个输入装置、和该至少一个输出装置。

[0144] 用于实施本公开的方法的程序代码可以采用一个或多个编程语言的任何组合来编写。这些程序代码可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理器或控制器，使得程序代码当由处理器或控制器执行时使流程图和/或框图中所规定的功能/操作被实施。程序代码可以完全在机器上执行、部分地在机器上执行，作为独立软件包部分地在机器上执行且部分地在远程机器上执行或完全在远程机器或服务器上执行。

[0145] 在本公开的上下文中，机器可读介质可以是有形的介质，其可以包含或存储以供指令执行系统、装置或设备使用或与指令执行系统、装置或设备结合地使用的程序。机器可读介质可以是机器可读信号介质或机器可读储存介质。机器可读介质可以包括但不限于电子的、磁性的、光学的、电磁的、红外的、或半导体系统、装置或设备，或者上述内容的任何合适组合。机器可读存储介质的更具体示例会包括基于一个或多个线的电气连接、便携式计算机盘、硬盘、随机存取存储器(RAM)、只读存储器(ROM)、可擦除可编程只读存储器(EPROM

或快闪存储器)、光纤、便捷式紧凑盘只读存储器(CD-ROM)、光学储存设备、磁储存设备、或上述内容的任何合适组合。

[0146] 为了提供与用户的交互,可以在计算机上实施此处描述的系统和技术,该计算机具有:用于向用户显示信息的显示装置(例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器);以及键盘和指向装置(例如,鼠标或者轨迹球),用户可以通过该键盘和该指向装置来将输入提供给计算机。其它种类的装置还可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈(例如,视觉反馈、听觉反馈、或者触觉反馈);并且可以用任何形式(包括声输入、语音输入或者、触觉输入)来接收来自用户的输入。

[0147] 可以将此处描述的系统和技术实施在包括后台部件的计算系统(例如,作为数据服务器)、或者包括中间件部件的计算系统(例如,应用服务器)、或者包括前端部件的计算系统(例如,具有图形用户界面或者网络浏览器的用户计算机,用户可以通过该图形用户界面或者该网络浏览器来与此处描述的系统和技术的实施方式交互)、或者包括这种后台部件、中间件部件、或者前端部件的任何组合的计算系统中。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括:局域网(LAN)、广域网(WAN)、互联网和区块链网络。

[0148] 计算机系统可以包括客户端和服务器。客户端和服务器一般远离彼此并且通常通过通信网络进行交互。通过在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序来产生客户端和服务器的关系。服务器可以是云服务器,也可以为分布式系统的服务器,或者是结合了区块链的服务器。

[0149] 应该理解,可以使用上面所示的各种形式的流程,重新排序、增加或删除步骤。例如,本公开中记载的各步骤可以并行地执行、也可以顺序地或以不同的次序执行,只要能够实现本公开公开的技术方案所期望的结果,本文在此不进行限制。

[0150] 虽然已经参照附图描述了本公开的实施例或示例,但应理解,上述的方法、系统和设备仅仅是示例性的实施例或示例,本发明的范围并不由这些实施例或示例限制,而是仅由授权后的权利要求书及其等同范围来限定。实施例或示例中的各种要素可以被省略或者可由其等同要素替代。此外,可以通过不同于本公开中描述的次序来执行各步骤。进一步地,可以以各种方式组合实施例或示例中的各种要素。重要的是随着技术的演进,在此描述的很多要素可以由本公开之后出现的等同要素进行替换。

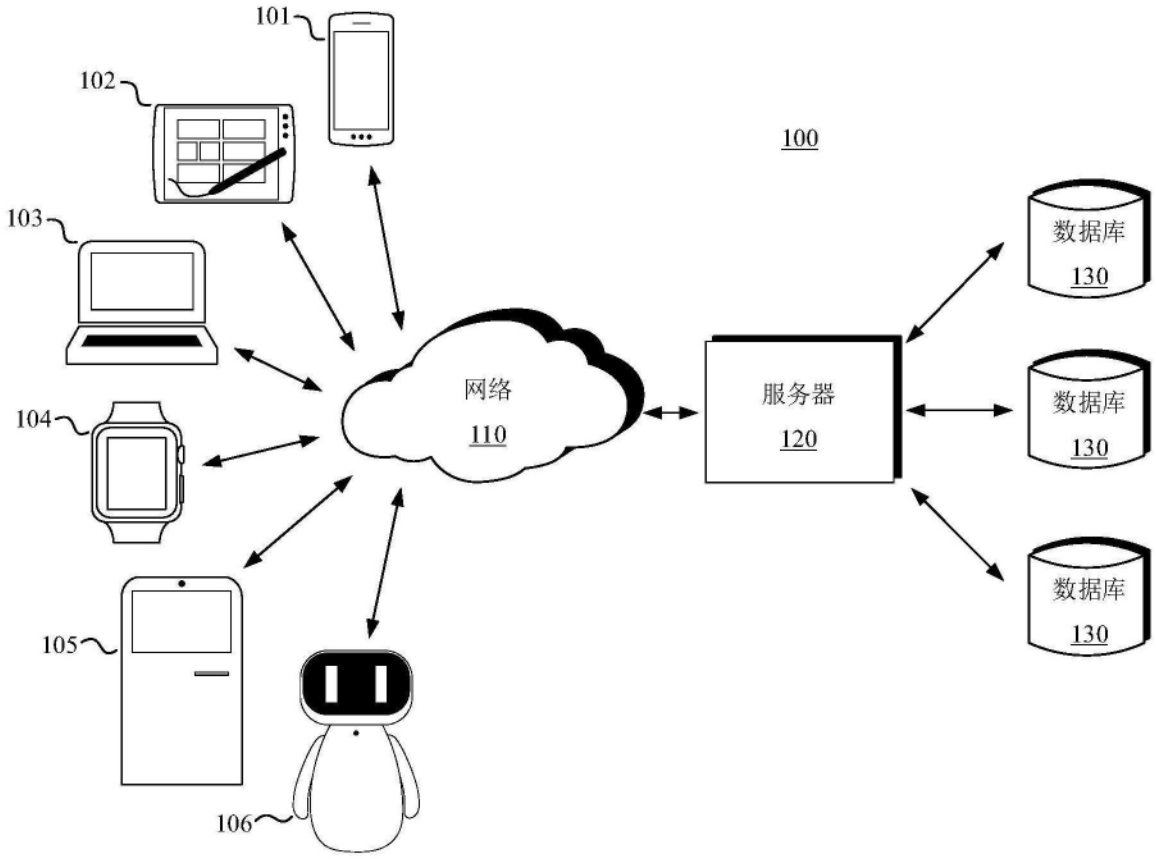

101
102
103
104
105
106
100
网络
110
服务器
120
数据库
130
数据库
130
数据库
130

图1

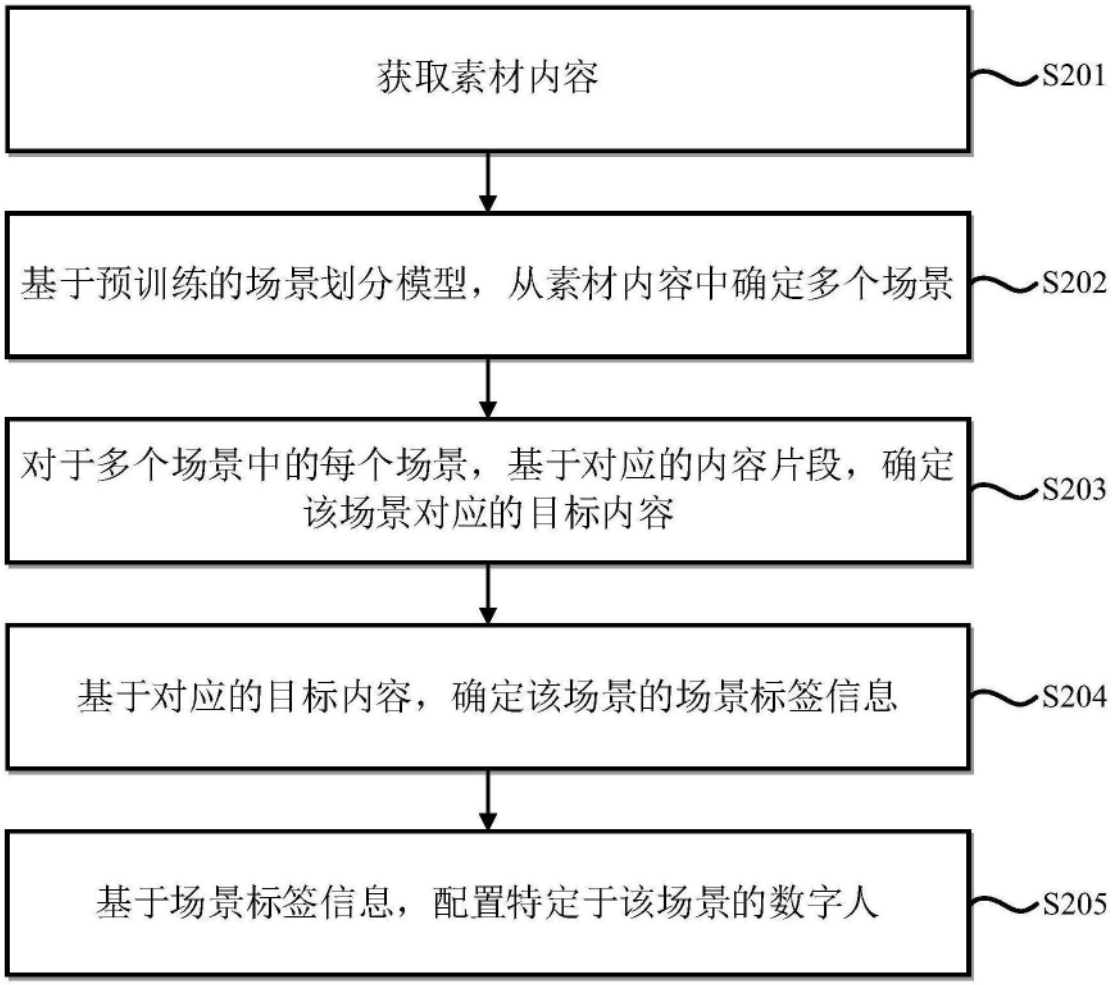

获取素材内容
S201
基于预训练的场景划分模型，从素材内容中确定多个场景
S202
对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容
S203
基于对应的目标内容，确定该场景的场景标签信息
S204
基于场景标签信息，配置特定于该场景的数字人
S205

图2

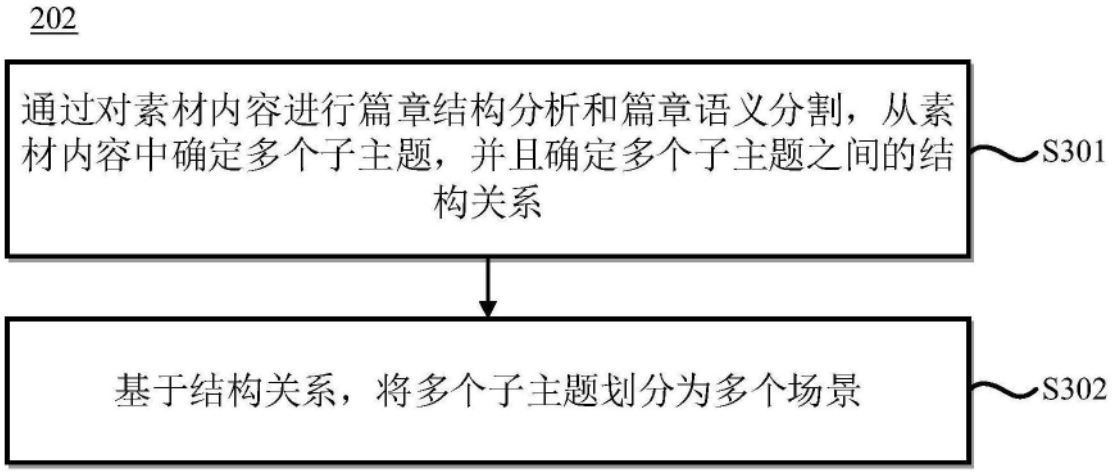

202
通过对素材内容进行篇章结构分析和篇章语义分割，从素材内容中确定多个子主题，并且确定多个子主题之间的结构关系
S301
基于结构关系，将多个子主题划分为多个场景
S302

图3

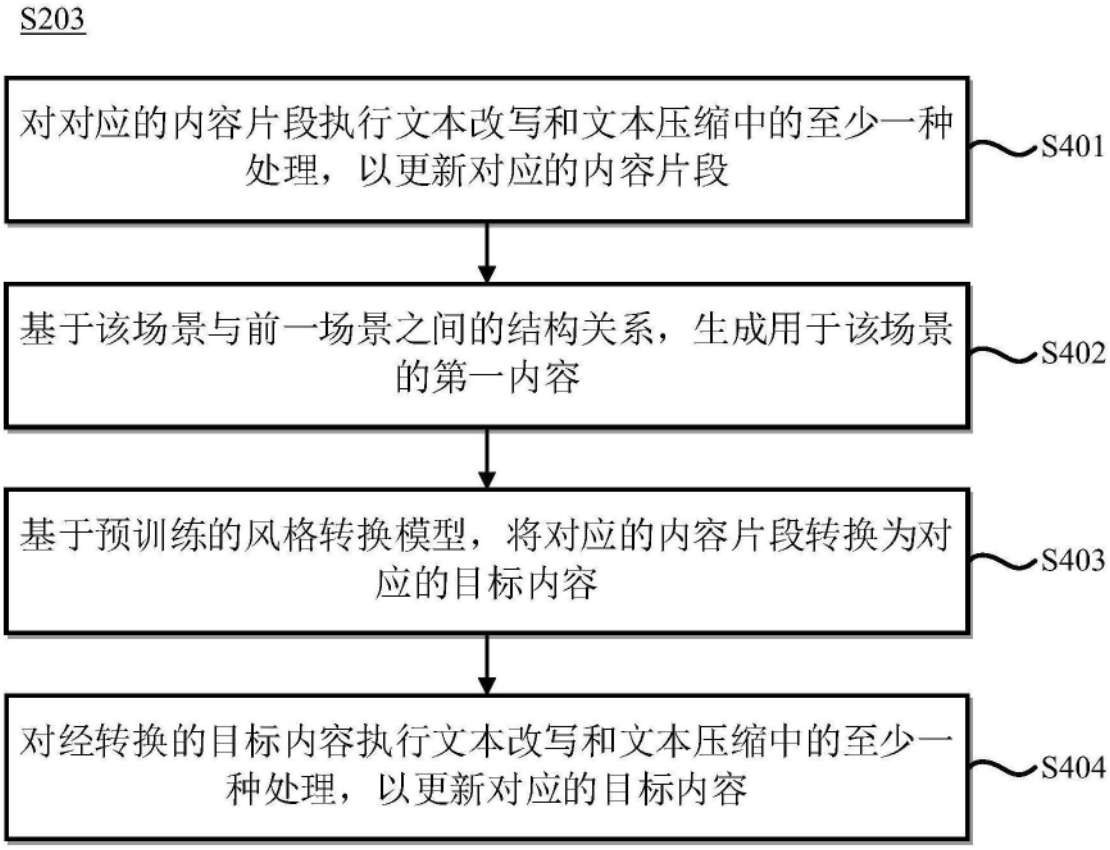

S203
对对应的内容片段执行文本改写和文本压缩中的至少一种处理，以更新对应的内容片段
S401
基于该场景与前一场景之间的结构关系，生成用于该场景的第一内容
S402
基于预训练的风格转换模型，将对应的内容片段转换为对应的目标内容
S403
对经转换的目标内容执行文本改写和文本压缩中的至少一种处理，以更新对应的目标内容
S404

图4

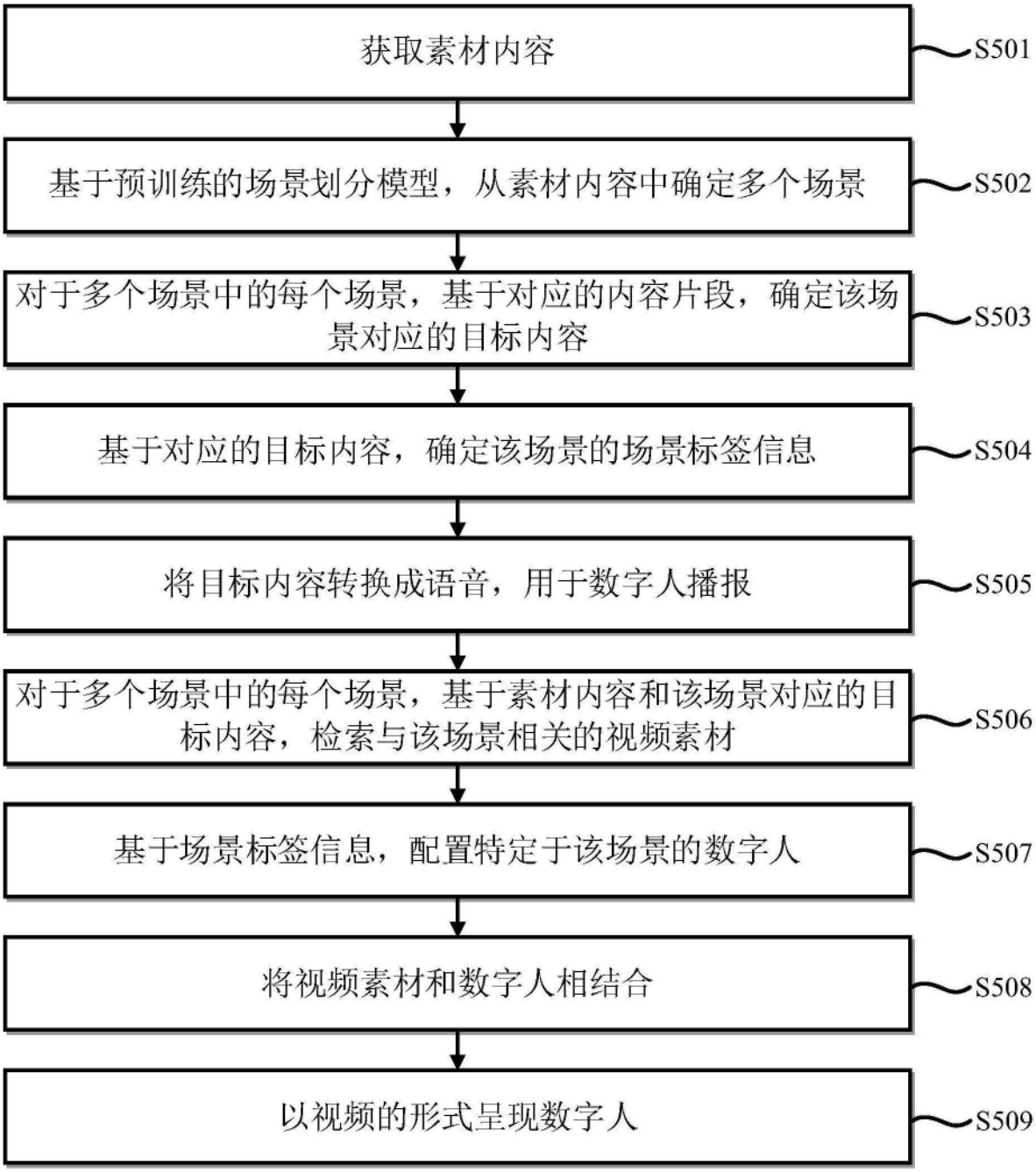

获取素材内容
S501
基于预训练的场景划分模型，从素材内容中确定多个场景
S502
对于多个场景中的每个场景，基于对应的内容片段，确定该场景对应的目标内容
S503
基于对应的目标内容，确定该场景的场景标签信息
S504
将目标内容转换成语音，用于数字人播报
S505
对于多个场景中的每个场景，基于素材内容和该场景对应的目标内容，检索与该场景相关的视频素材
S506
基于场景标签信息，配置特定于该场景的数字人
S507
将视频素材和数字人相结合
S508
以视频的形式呈现数字人
S509

图5

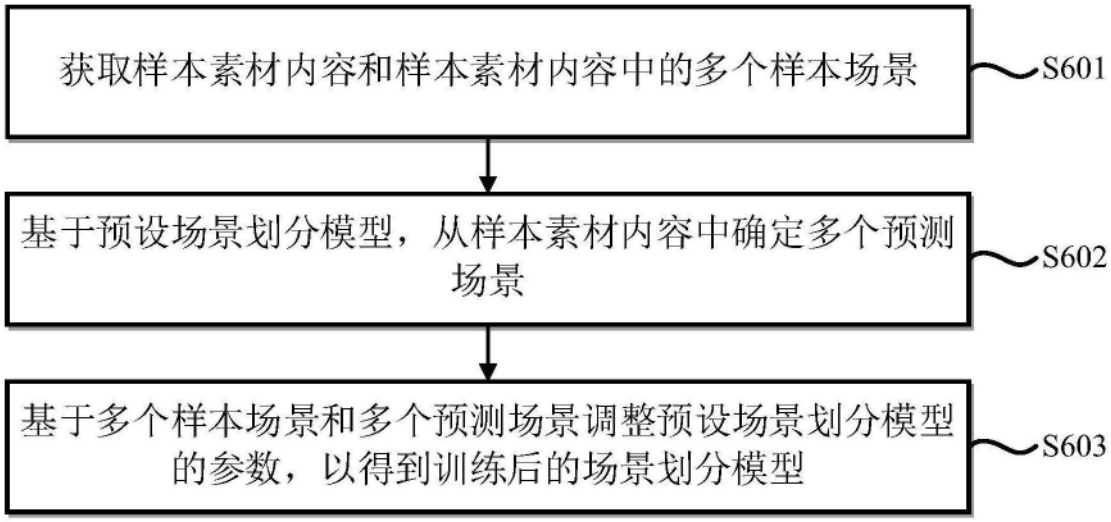

获取样本素材内容和样本素材内容中的多个样本场景
S601
基于预设场景划分模型，从样本素材内容中确定多个预测场景
S602
基于多个样本场景和多个预测场景调整预设场景划分模型的参数，以得到训练后的场景划分模型
S603

图6

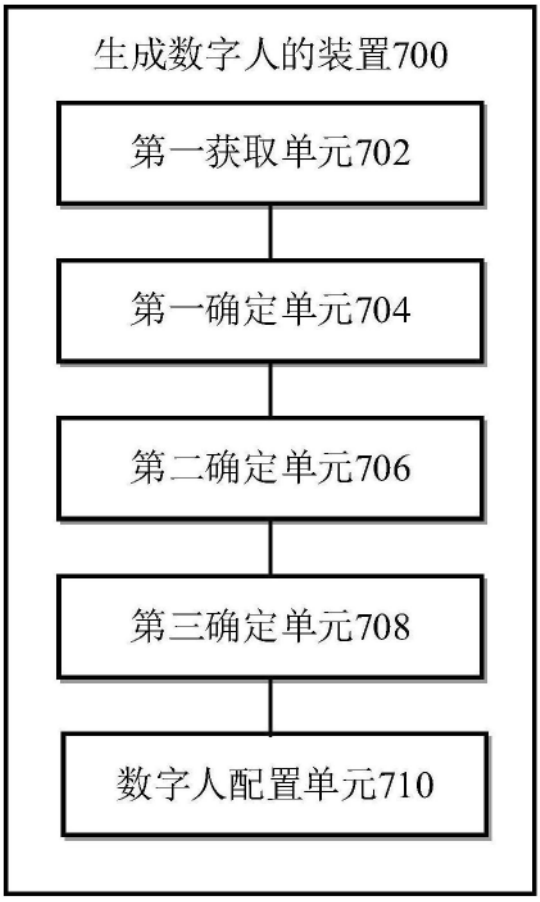

生成数字人的装置700
第一获取单元702
第一确定单元704
第二确定单元706
第三确定单元708
数字人配置单元710

图7

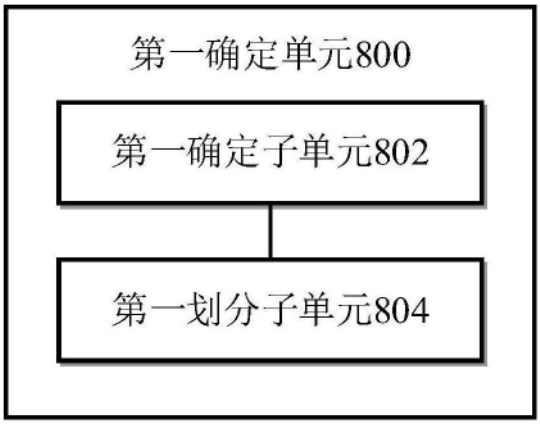

第一确定单元800
第一确定子单元802
第一划分子单元804

图8

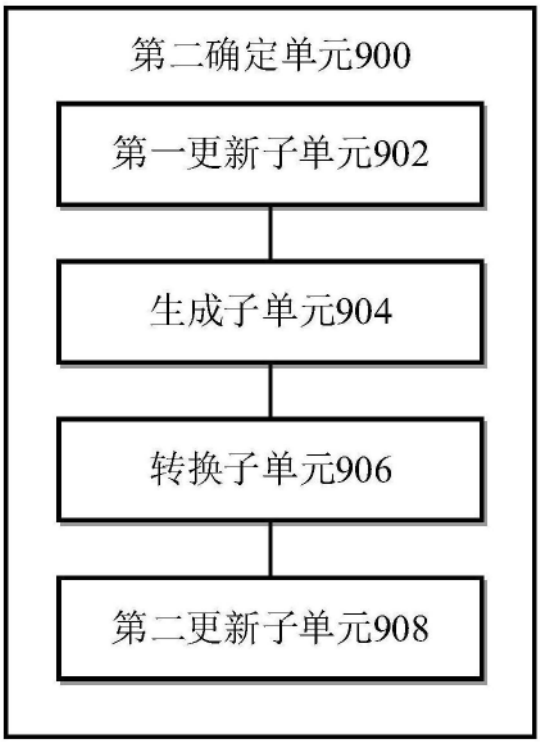

第二确定单元900
第一更新子单元902
生成子单元904
转换子单元906
第二更新子单元908

图9

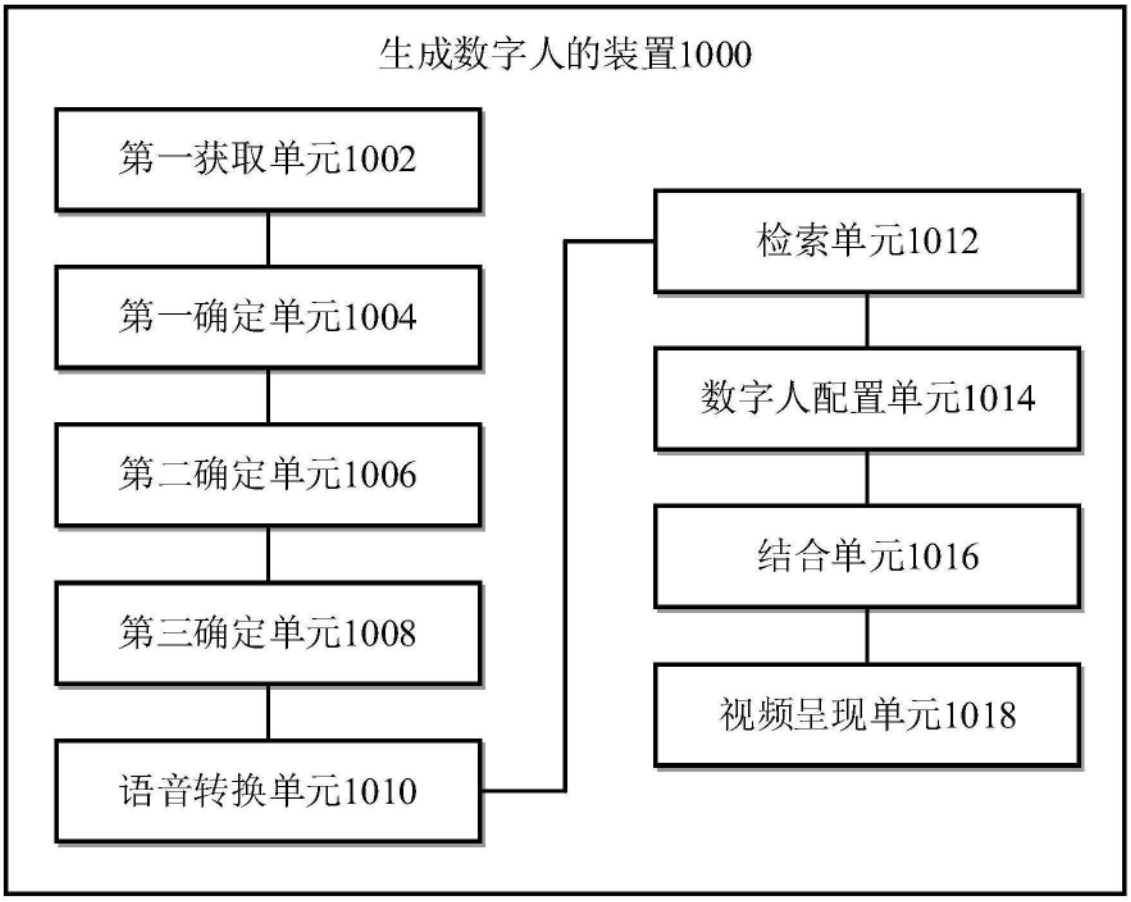

生成数字人的装置1000
第一获取单元1002
第一确定单元1004
第二确定单元1006
第三确定单元1008
语音转换单元1010
检索单元1012
数字人配置单元1014
结合单元1016
视频呈现单元1018

图10

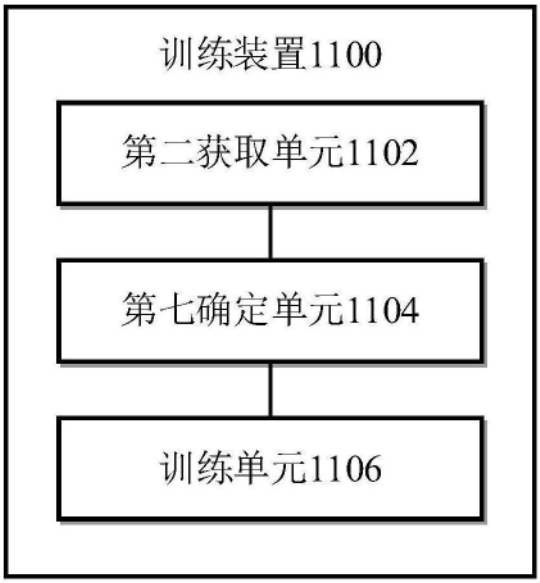

训练装置1100
第二获取单元1102
第七确定单元1104
训练单元1106

图11

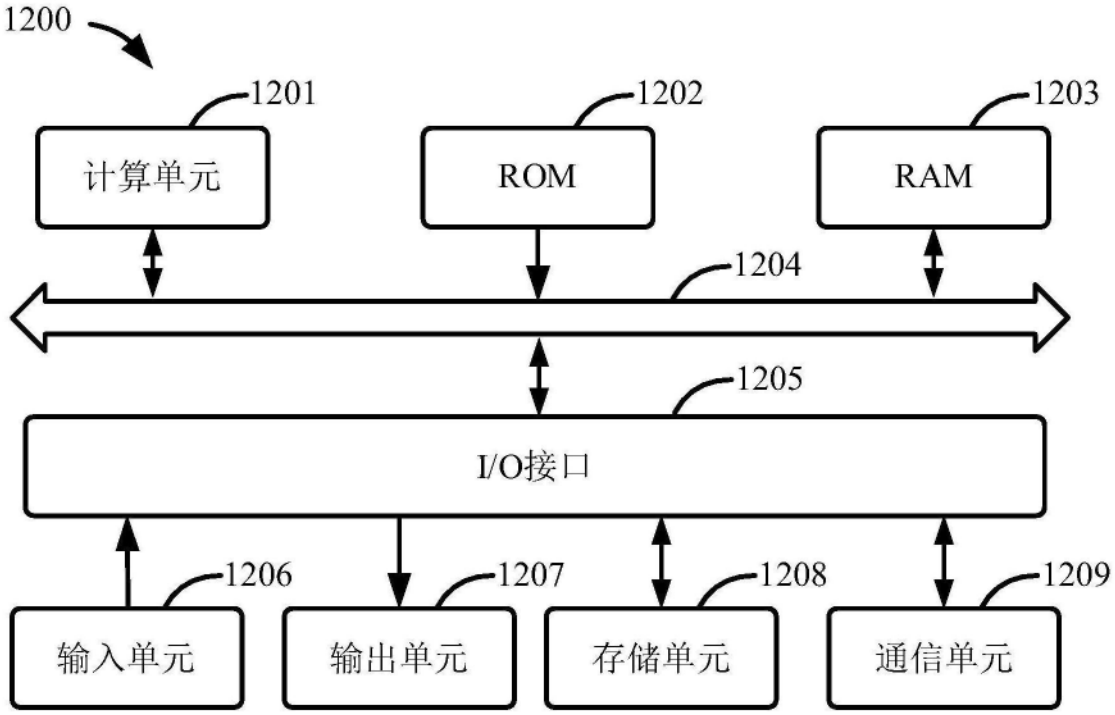

1200
1201
计算单元
1202
ROM
1203
RAM
1204
1205
I/O接口
1206
输入单元
1207
输出单元
1208
存储单元
1209
通信单元

图12