(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107578004 A
(43)申请公布日 2018.01.12

(21)申请号 201710764473.2

(22)申请日 2017.08.30

(71)申请人 苏州清睿教育科技股份有限公司
地址 215000 江苏省苏州市工业园区星湖街328号创意产业园16-A301单元

(72)发明人 朱奇峰

(74)专利代理机构 苏州中合知识产权代理事务所(普通合伙) 32266
代理人 李中华

(51)Int.Cl.
G06K 9/00(2006.01)
G10L 13/08(2013.01)
G09B 5/06(2006.01)

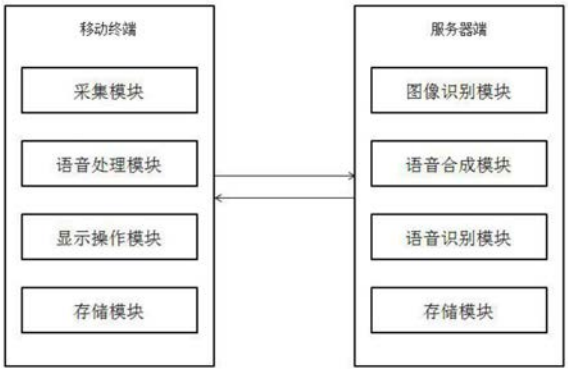
权利要求书2页 说明书4页 附图1页

(54)发明名称

基于图像识别和语音交互的学习方法及系统

(57)摘要

本发明公开了一种基于图像识别和语音交互的学习方法及系统，包括：用户通过移动终端拍照或在移动终端中选择一张图片上传至服务器端；服务器端的图像识别模块接收移动终端发来的图片，并将图片处理成至少一条文本推送至移动终端；用户选择其中一条文本或自己推荐一条文本，移动终端自动将该文本发送至服务器端的语音合成模块，语音合成模块接收移动终端发来的文本，并将文件处理合成为音频数据反馈至移动终端；移动终端接收音频数据供用户学习；用户根据音频数据进行跟读，并通过移动终端录音发送至服务器端的语音识别模块；所述语音识别模块接收移动终端发送来的录音，对录音进行分析并给出评价反馈至移动终端，达到随时随地进行拍照学习的目的。



CN 107578004 A

1.一种基于图像识别和语音交互的学习方法,其特征在于,包括:用户通过移动终端拍照或在移动终端中选择一张图片上传至服务器端;服务器端的图像识别模块接收移动终端发来的图片,并将图片处理成至少一条文本推送至移动终端;用户选择其中一条文本或自己推荐一条文本,移动终端自动将该文本发送至服务器端的语音合成模块,所述语音合成模块接收移动终端发来的文本,并将文件处理合成为音频数据反馈至移动终端;移动终端接收音频数据供用户学习;用户根据音频数据进行跟读,并通过移动终端录音发送至服务器端的语音识别模块;所述语音识别模块接收移动终端发送来的录音,对录音进行分析并给出评价反馈至移动终端。

2.根据权利要求1所述的基于图像识别和语音交互的学习方法,其特征在于,所述图像识别模块采用TensorFlow程序对图片进行处理,TensorFlow利用训练好的模型进行预测得到推荐文本。

3.根据权利要求1所述的基于图像识别和语音交互的学习方法,其特征在于,所述语音合成模块根据预先设置的语法知识库和语法字典对文本进行分析;将分析后的文本训练,生成有韵律的神经网络;结合预先设置的语音语料库生成音频数据。

4.根据权利要求1所述的基于图像识别和语音交互的学习方法,其特征在于,所述语音识别模块分析待识别的语音,得到语音参数,将所述语音参数与语音识别库中的语音模板进行一一比较,并采用判决的方法找出最接近该语音参数的模板,得出识别结果并评分。

5.根据权利要求4所述的基于图像识别和语音交互的学习方法,其特征在于,所述语音参数比较的标准是计量语音特征参数矢量之间的失真测度。

6.一种基于图像识别和语音交互的学习系统,其特征在于,包括:移动终端和服务器端,所述移动终端与所述服务器端通过网络进行连接,

所述移动终端,包括:采集模块、语音处理模块、显示操作模块和存储模块,所述采集模块,用于对物体进行图像采集,并将采集到的图像发送至服务器端;所述语音处理模块,用于接收服务器端生成的音频数据和为用户录音并将录音发送至服务器端;所述显示模块,用于显示服务器端反馈的文字信息以及对系统进行相应操作的按键;所述存储模块,用于存储采集到的图像、服务器生成的音频数据以及用户的录音;

所述服务器端,包括:图像识别模块、语音合成模块、语音识别模块和存储模块,所述图像识别模块,用于接收移动终端发来的图片,并根据图片内容将图片信息转化成推荐文本反馈至移动终端;所述语音合成模块,用于接收移动终端发送来的推荐文本,并根据所述推荐文本的内容生成相应的音频数据,将所述音频数据反馈到所述移动终端,所述语音识别模块,用于接收移动终端发来的录音,并对所述录音进行识别以及对所述英文语音信息做出评价,将评价内容反馈到移动终端供用户查看;所述存储模块,用于存储用户信息、音频数据以及用户的录音。

7.根据权利要求6所述的基于图像识别和语音交互的学习系统,其特征在于,所述图像识别模块实用GPU服务器,利用大量的模型学习图片,再使用集束算法进行筛选图片反馈结果。

8.根据权利要求6所述的基于图像识别和语音交互的学习系统,其特征在于,所述语音合成模块采用TTS内核,所述TTS内核的发声引擎小,不需要大量的声音文件支持。

9.根据权利要求6所述的基于图像识别和语音交互的学习系统,其特征在于,所述语音

识别模块的识别框架采用基于模式匹配的动态时间规整法和基于统计模型的隐马尔可夫模型法。

基于图像识别和语音交互的学习方法及系统

技术领域

[0001] 本发明涉及图像识别及语音交互领域，具体涉及一种基于图像识别和语音交互的学习系统及方法。

背景技术

[0002] 习主席说，建设“人人皆学、处处能学、时时可学”的学习型社会。坚持不懈推进教育信息化，努力以信息化为手段扩大优质教育资源覆盖面。我们将通过教育信息化，逐步缩小区域、城乡数字差距，大力促进教育公平，让亿万孩子同在蓝天下共享优质教育、通过知识改变命运。

[0003] 现有技术中，语音合成技术、在线录音技术、语音识别技术，都已经是相对成熟的技术，但是现有技术中还存在很多不足，例如：学生在学习中，学习内容都是教材规定好的，无法自动生成教学内容，对任意内容自动生成图文声音并茂的教学内容，并辅导使用者进行外语学习和练习的产品还没有。

发明内容

[0004] 为解决上述技术问题，本发明提出了一种基于图像识别和语音交互的学习方法及系统，以达到随时随地进行拍照学习的目的。

[0005] 为达到上述目的，本发明的技术方案如下：基于图像识别和语音交互的学习方法，包括：用户通过移动终端拍照或在移动终端中选择一张图片上传至服务器端；服务器端的图像识别模块接收移动终端发来的图片，并将图片处理成至少一条文本推送至移动终端；用户选择其中一条文本或自己推荐一条文本，移动终端自动将该文本发送至服务器端的语音合成模块，所述语音合成模块接收移动终端发来的文本，并将文件处理合成为音频数据反馈至移动终端；移动终端接收音频数据供用户学习；用户根据音频数据进行跟读，并通过移动终端录音发送至服务器端的语音识别模块；所述语音识别模块接收移动终端发送来的录音，对录音进行分析并给出评价反馈至移动终端。

[0006] 作为优选的，所述图像识别模块采用TensorFlow程序对图片进行处理，TensorFlow利用训练好的模型进行预测得到推荐文本。

[0007] 作为优选的，所述语音合成模块根据预先设置的语法知识库和语法字典对文本进行分析；将分析后的文本训练，生成有韵律的神经网络；结合预先设置的语音语料库生成音频数据。

[0008] 作为优选的，所述语音识别模块分析待识别的语音，得到语音参数，将所述语音参数与语音识别库中的语音模板进行一一比较，并采用判决的方法找出最接近该语音参数的模板，得出识别结果并评分。

[0009] 作为优选的，所述语音参数比较的标准是计量语音特征参数矢量之间的失真测度。

[0010] 基于图像识别和语音交互的学习系统，其特征在于，包括：移动终端和服务器端，

所述移动终端与所述服务器端通过网络进行连接，

[0011] 所述移动终端，包括：采集模块、语音处理模块、显示操作模块和存储模块，所述采集模块，用于对物体进行图像采集，并将采集到的图像发送至服务器端；所述语音处理模块，用于接收服务器端生成的音频数据和为用户录音并将录音发送至服务器端；所述显示模块，用于显示服务器端反馈的文字信息以及对系统进行相应操作的按键；所述存储模块，用于存储采集到的图像、服务器生成的音频数据以及用户的录音；

[0012] 所述服务器端，包括：图像识别模块、语音合成模块、语音识别模块和存储模块，所述图像识别模块，用于接收移动终端发来的图片，并根据图片内容将图片信息转化成推荐文本反馈至移动终端；所述语音合成模块，用于接收移动终端发送来的推荐文本，并根据所述推荐文本的内容生成相应的音频数据，将所述音频数据反馈到所述移动终端，所述语音识别模块，用于接收移动终端发来的录音，并对所述录音进行识别以及对所述英文语音信息做出评价，将评价内容反馈到移动终端供用户查看；所述存储模块，用于存储用户信息、音频数据以及用户的录音。

[0013] 作为优选的，所述图像识别模块实用GPU服务器，利用大量的模型学习图片，再使用集束算法进行筛选图片反馈结果。

[0014] 作为优选的，所述语音合成模块采用TTS内核，所述TTS内核的发声引擎小，不需要大量的声音文件支持。

[0015] 作为优选的，所述语音识别模块的识别框架采用基于模式匹配的动态时间规整法和基于统计模型的隐马尔可夫模型法。

[0016] 本发明具有如下优点：

[0017] (1).本发明利用移动终端进行图像采集，再通过服务器端生成文本、语音以及对用户录音的评分，达到随时随地进行拍照学习的目的。

[0018] (2).本发明利用语音合成模块将文本合成音频数据供用户学习，可以从听力的角度拓展学习。

[0019] (3).本发明利用语音识别模块对用户的录音进行评分，直观精确的让用户了解自身的学习情况。

附图说明

[0020] 为了更清楚地说明本发明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。

[0021] 图1为本发明实施例公开的基于图像识别和语音交互的学习系统功能模块图；

[0022] 图2为本发明实施例公开的语音合成流程图。

具体实施方式

[0023] 下面将结合本发明实施例中的附图，对本发明实施例中的技术方案进行清楚、完整地描述。

[0024] 本发明提供了一种基于图像识别和语音交互的学习方法及系统，其工作原理是通过移动终端进行图像采集，再通过服务器端生成文本、语音以及对用户录音的评分，达到随时随地进行拍照学习的目的。

[0025] 下面结合实施例和具体实施方式对本发明作进一步详细的说明。

[0026] 如图1和图2所示，基于图像识别和语音交互的学习方法，包括：用户通过移动终端拍照或在移动终端中选择一张图片上传至服务器端；服务器端的图像识别模块接收移动终端发来的图片，并将图片处理成多条英文文本推送至移动终端；用户选择其中一条英文文本或自己推荐一条英文文本，移动终端自动将该英文文本发送至服务器端的语音合成模块，所述语音合成模块接收移动终端发来的英文文本，并将英文文件处理合成为英文音频数据反馈至移动终端；移动终端接收英文音频数据供用户学习；用户根据英文音频数据进行跟读，并通过移动终端录音发送至服务器端的语音识别模块；所述语音识别模块接收移动终端发送来的录音，对录音进行分析并给出评价反馈至移动终端。

[0027] 其中，所述图像识别模块采用TensorFlow程序对图片进行处理，TensorFlow利用训练好的模型进行预测得到推荐文本，TensorFlow通过read_data_sets方法对引用数据进行封装，然后读取这些划分好的数据集，再通过next_batch来获取一小批的训练数据，在利用梯度下降算法时需要在所有的训练数据上计算梯度，随机选取一部分训练数据集，提供到神经网络的输入层，然后通过反向迭代方法去优化这个神经网络。

[0028] 其中，所述语音合成模块根据预先设置的语法知识库和语法字典对文本进行分析；将分析后的文本训练，生成有韵律的神经网络；结合预先设置的语音语料库生成音频数据。

[0029] 其中，所述语音识别模块分析待识别的语音，得到语音参数，将所述语音参数与语音识别库中的语音模板进行一一比较，并采用判决的方法找出最接近该语音参数的模板，得出识别结果并评分。

[0030] 其中，所述语音参数比较的标准是计量语音特征参数矢量之间的失真测度。

[0031] 基于图像识别和语音交互的学习系统，包括：移动终端和服务器端，所述移动终端与所述服务器端通过网络进行连接，

[0032] 所述移动终端，包括：采集模块、语音处理模块、显示操作模块和存储模块，所述采集模块，用于对物体进行图像采集，并将采集到的图像发送至服务器端；所述语音处理模块，用于接收服务器端生成的音频数据和为用户录音并将录音发送至服务器端；所述显示模块，用于显示服务器端反馈的文字信息以及对系统进行相应操作的按键；所述存储模块，用于存储采集到的图像、服务器生成的音频数据以及用户的录音；

[0033] 所述服务器端，包括：图像识别模块、语音合成模块、语音识别模块和存储模块，所述图像识别模块，用于接收移动终端发来的图片，并根据图片内容将图片信息转化成推荐文本反馈至移动终端；所述语音合成模块，用于接收移动终端发送来的推荐文本，并根据所述推荐文本的内容生成相应的音频数据，将所述音频数据反馈到所述移动终端，所述语音识别模块，用于接收移动终端发来的录音，并对所述录音进行识别以及对所述英文语音信息做出评价，将评价内容反馈到移动终端供用户查看；所述存储模块，用于存储用户信息、音频数据以及用户的录音。

[0034] 其中，所述图像识别模块实用GPU服务器，利用大量的模型学习图片，再使用集束算法进行筛选图片反馈结果。

[0035] 其中，所述语音合成模块采用TTS内核，所述TTS内核的发声引擎小，不需要大量的声音文件支持。

[0036] 其中，所述语音识别模块的识别框架采用基于模式匹配的动态时间规整法和基于统计模型的隐马尔可夫模型法。

[0037] 以上所述的仅是本发明所公开的基于图像识别和语音交互的学习系统及方法的优选实施方式，应当指出，对于本领域的普通技术人员来说，在不脱离本发明创造构思的前提下，还可以做出若干变形和改进，这些都属于本发明的保护范围。

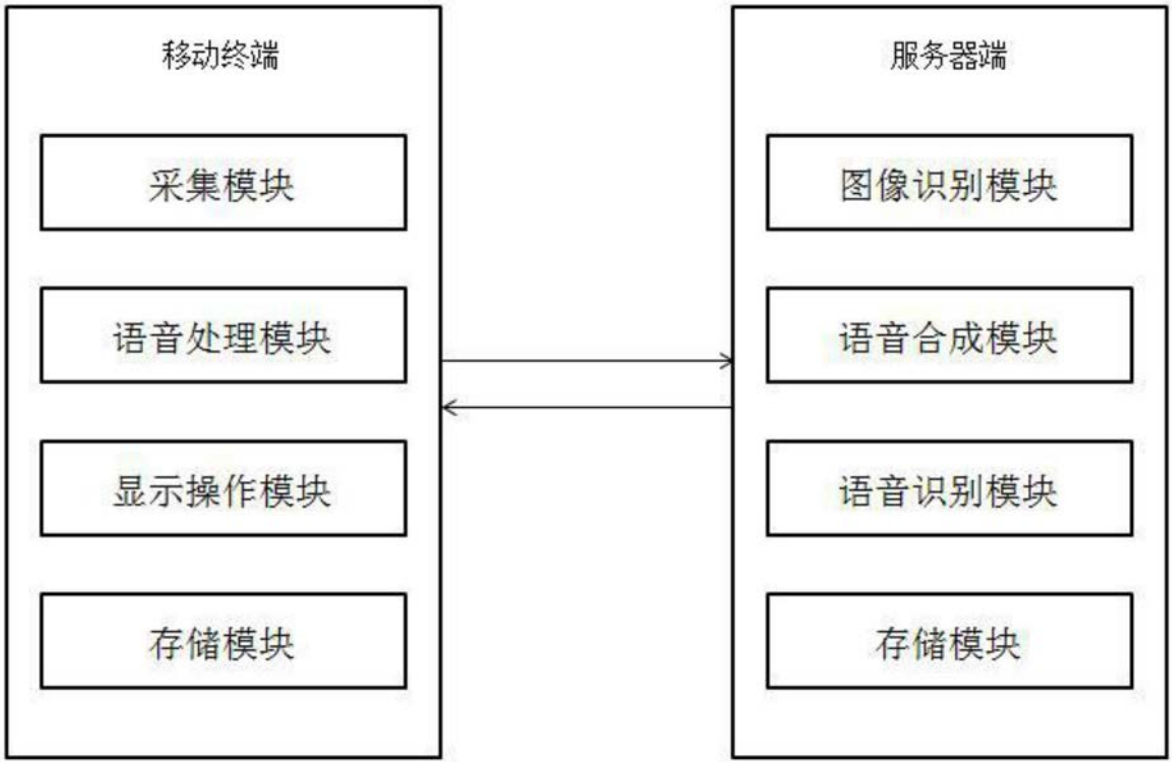

移动终端
采集模块
语音处理模块
显示操作模块
存储模块
服务器端
图像识别模块
语音合成模块
语音识别模块
存储模块

图1

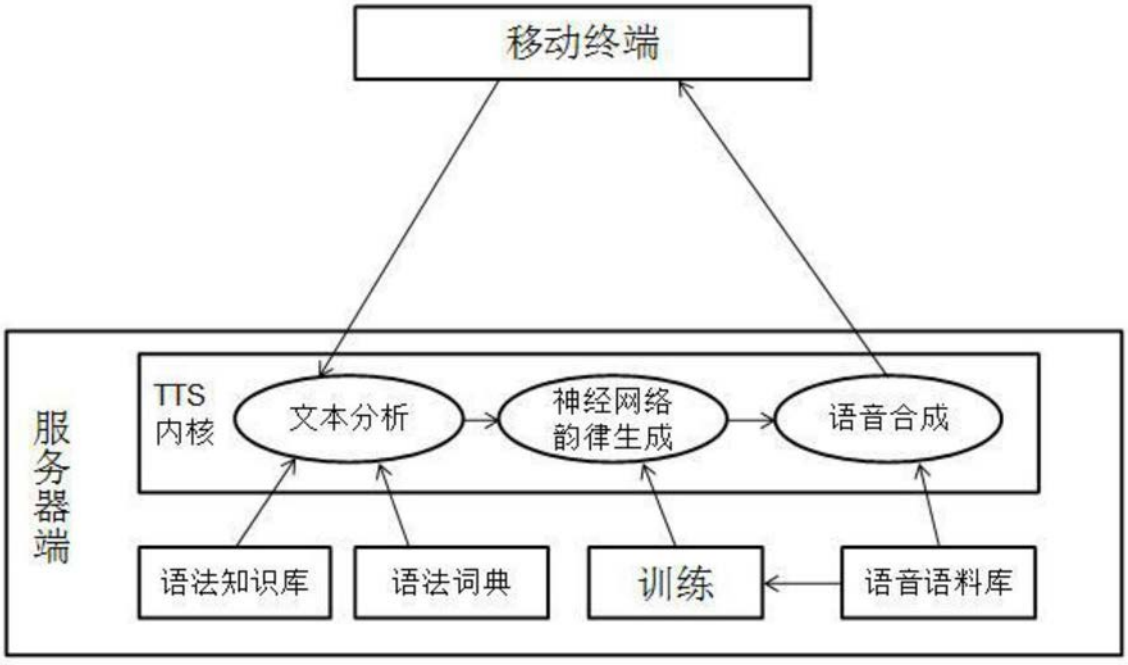

移动终端
服务器端
TTS
内核
文本分析
神经网络
韵律生成
语音合成
语法知识库
语法词典
训练
语音语料库

图2