(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 111681676 B
(45)授权公告日 2023.08.08

(21)申请号 202010517903.2

(22)申请日 2020.06.09

(65)同一申请的已公布的文献号
申请公布号 CN 111681676 A

(43)申请公布日 2020.09.18

(73)专利权人 杭州星合尚世影视传媒有限公司
地址 310000 浙江省杭州市江干区九华路1号8幢5楼507室

(72)发明人 薛媛 金若熙

(74)专利代理机构 杭州五洲普华专利代理事务所(特殊普通合伙) 33260
专利代理师 张瑜

(51)Int.Cl.
G10L 25/51(2013.01)
G10L 25/54(2013.01)
G10L 25/57(2013.01)
G06F 18/22(2023.01)
G06V 20/40(2022.01)
G06V 10/82(2022.01)
G06N 3/045(2023.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06N 3/088(2023.01)
G06N 20/20(2019.01)

(56)对比文件
CN 105279495 A,2016.01.27
US 2008159622 A1,2008.07.03
US 2017357720 A1,2017.12.14
US 2018336519 A1,2018.11.22
WO 02095508 A1,2002.11.28
CN 111199238 A,2020.05.26
CN 109819313 A,2019.05.28
CN 108197572 A,2018.06.22
CN 109919031 A,2019.06.21
CN 110046599 A,2019.07.23
US 2003053680 A1,2003.03.20

审查员 张辉

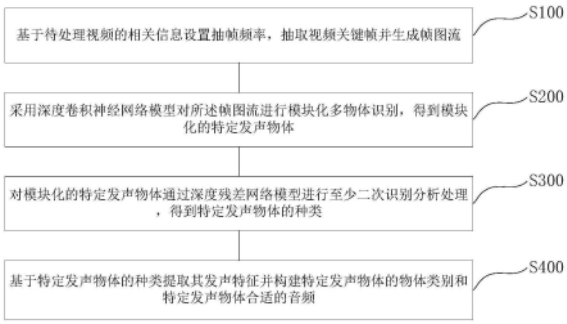
权利要求书3页 说明书13页 附图1页

(54)发明名称

视频物体识别构建音频方法、系统、装置及可读存储介质

(57)摘要

本发明公开一种视频物体识别构建音频方法，包括以下步骤：基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。通过采用深度卷积神经网络模型对视频进行模块化识别，再通过深度残差网络模型进行二次甚至更多次识别分析处理，能够等到更精确的特定发声物体的种类，进而能够构建出更合适的音频。



CN 111681676 B

1.一种视频物体识别构建音频方法,其特征在于,包括以下步骤:

基于待处理视频的相关信息设置抽帧频率,抽取视频关键帧并生成帧图流;

采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别,得到模块化的特定发声物体;

对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理,得到特定发声物体的种类;

基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频;

音频包括音频介绍和音频关键词,音频介绍为音频的介绍内容文本,音频关键词包括至少三个描述音频的词语,所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称;

所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频,具体步骤为:

基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数;

基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数,根据视频音频匹配分数得到特定发声物体至少一种合适的音频;

所述基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数,具体步骤如下:

对特定发声物体的物体类别和音频介绍进行分词处理得到单词;

分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例,得到第一比例和第二比例,将第一比例和第二比例进行加权平均处理,得到单词匹配分数,所述单词匹配分数=物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重,其中,音频介绍权重+音频关键词权重=1;

基于音频介绍的统计数据,得到物体类别TF-IDF向量,通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度,将第一余弦相似度作为TF-IDF匹配分数,所述TF-IDF匹配分数=cosine_similarity(物体类别TF-IDF向量,音频介绍TF-IDF向量);

将单词匹配分数和TF-IDF匹配分数进行加权平均处理,得到第一匹配分数,所述第一匹配分数=单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重,其中,单词权重+TF-IDF权重=1;

获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量,经过计算得到BERT向量的余弦相似度,将余弦相似度作为神经网络匹配分数。

2.根据权利要求1所述的视频物体识别构建音频方法,其特征在于,所述深度残差网络模型获得过程如下:

获取若干包含特定发声物体的图像,剔除不合格的特定发声物体的图像,得到合格特定发声物体的图像;

将合格特定发声物体的图像进行预处理,得到合格特定发声物体的图像数据集,并划分为训练集和验证集;

将训练集输入至初始深度残差网络模型中进行训练,再通过验证集对训练结果进行进

行验证，得到能够获取到特定发声物体的种类的深度残差网络模型。

3.根据权利要求1所述的视频物体识别构建音频方法，其特征在于，所述基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，具体为：

将第一匹配分数和神经网络匹配分数进行加权平均处理，得到视频音频匹配分数，所述视频音频匹配分数＝第一匹配分数*第一权重+神经网络匹配分数*神经网络权重，其中，第一权重+神经网络权重＝1。

4.根据权利要求1所述的视频物体识别构建音频方法，其特征在于，所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，还包括如下步骤：

根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配；

将所有音频进行混音处理，形成完整的音频文件，将音频文件添加进视频的音轨使得音频文件和视频同步。

5.一种视频物体识别构建音频系统，其特征在于，包括帧图流生成模块、第一处理模块、第二处理模块和提取构建模块；

所述帧图流生成模块被设置为：基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

所述第一处理模块，用于采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

所述第二处理模块，用于对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

所述提取构建模块被设置为：基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频；

音频包括音频介绍和音频关键词，音频介绍为音频的介绍内容文本，音频关键词包括至少三个描述音频的词语，所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称；

所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，具体步骤为：

基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数；

基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，根据视频音频匹配分数得到特定发声物体至少一种合适的音频；

所述基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数，具体步骤如下：

对特定发声物体的物体类别和音频介绍进行分词处理得到单词；

分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例，得到第一比例和第二比例，将第一比例和第二比例进行加权平均处理，得到单词匹配分数，所述单词匹配分数＝物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重，其中，音频介绍权重+音频关键词权重＝1；

基于音频介绍的统计数据，得到物体类别TF-IDF向量，通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度，将第一余弦相似度作为TF-IDF匹配分数，所述TF-IDF匹配分数＝cosine_similarity(物体类别TF-IDF向量，音频介绍TF-IDF向量)；

将单词匹配分数和TF-IDF匹配分数进行加权平均处理，得到第一匹配分数，所述第一匹配分数＝单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重，其中，单词权重+TF-IDF权重＝1；

获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，经过计算得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数。

6.一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现如权利要求1至4任意一项所述的方法。

7.一种视频物体识别构建音频装置，包括存储器、处理器以及存储在所述存储器中并可在所述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如权利要求1至4任意一项所述的方法。

视频物体识别构建音频方法、系统、装置及可读存储介质

技术领域

[0001] 本发明涉及计算机视觉视频检测技术领域，尤其涉及一种视频物体识别构建音频方法、系统、装置及可读存储介质。

背景技术

[0002] 在现有技术中，越来越多的神经网络模型被应用在各个技术领域中，比如安防、自动驾驶，图像识别等技术领域并且还在不断的追求识别的更高精度。现有技术中识别物体识别方法还有很多不足，比如视频中的图像不能很精确的归出特定发声物体的类型，类别识别的不够精细，因此在后续自动配音的过程中就会导致配音不够精确，难度很大。

发明内容

[0003] 本发明针对现有技术中的缺点，提供了一种视频物体识别构建音频方法、系统、装置及可读存储介质。

[0004] 为了解决上述技术问题，本发明通过下述技术方案得以解决：

[0005] 一种视频物体识别构建音频方法，包括以下步骤：

[0006] 基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0007] 采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0008] 对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0009] 基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0010] 作为一种可事实方式，所述深度残差网络模型获得过程如下：

[0011] 获取若干包含特定发声物体的图像，剔除不合格的特定发声物体的图像，得到合格特定发声物体的图像；

[0012] 将合格特定发声物体的图像进行预处理，得到合格特定发声物体的图像数据集，并划分为训练集和验证集；

[0013] 将训练集输入至初始深度残差网络模型中进行训练，再通过验证集对训练结果进行进行验证，得到能够获取到特定发声物体的种类的深度残差网络模型。

[0014] 作为一种可事实方式，音频包括音频介绍和音频关键词，音频介绍为音频的介绍内容文本，音频关键词包括至少三个描述音频的词语，所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称。

[0015] 作为一种可事实方式，所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，具体步骤为：

[0016] 基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数；

[0017] 基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，根据视频音频匹配分数得到特定发声物体至少一种合适的音频。

[0018] 作为一种可事实方式，所述基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数，具体步骤如下：

[0019] 对特定发声物体的物体类别和音频介绍进行分词处理得到单词；

[0020] 分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例，得到第一比例和第二比例，将第一比例和第二比例进行加权平均处理，得到单词匹配分数，所述单词匹配分数＝物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重，其中，音频介绍权重+音频关键词权重＝1；

[0021] 基于音频介绍的统计数据，得到物体类别TF-IDF向量，通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度，将第一余弦相似度作为TF-IDF匹配分数，所述TF-IDF匹配分数＝cosine_similarity(物体类别TF-IDF向量，音频介绍TF-IDF向量)；

[0022] 将单词匹配分数和TF-IDF匹配分数进行加权平均处理，得到第一匹配分数，所述第一匹配分数＝单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重，其中，单词权重+TF-IDF权重＝1；

[0023] 获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，经过计算得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数。

[0024] 作为一种可事实方式，所述基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，具体为：

[0025] 将第一匹配分数和神经网络匹配分数进行加权平均处理，得到视频音频匹配分数，所述视频音频匹配分数＝第一匹配分数*第一权重+神经网络匹配分数*神经网络权重，其中，第一权重+神经网络权重＝1。

[0026] 作为一种可事实方式，所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，还包括如下步骤：

[0027] 根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配；

[0028] 将所有音频进行混音处理，形成完整的音频文件，将音频文件添加进视频的音轨使得音频文件和视频同步。

[0029] 一种视频物体识别构建音频系统，包括帧图流生成模块、第一处理模块、第二处理模块和提取构建模块；

[0030] 所述帧图流生成模块被设置为:基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0031] 所述第一处理模块，用于采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0032] 所述第二处理模块，用于对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0033] 所述提取构建模块被设置为:基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0034] 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计

算机程序被处理器执行时实现如下的方法步骤：

[0035] 基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0036] 采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0037] 对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0038] 基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0039] 一种视频物体识别装置，包括存储器、处理器以及存储在所述存储器中并可在所述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如下的方法步骤：

[0040] 基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0041] 采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0042] 对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0043] 基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0044] 本发明由于采用了以上技术方案，具有显著的技术效果：

[0045] 基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。通过采用深度卷积神经网络模型对视频进行模块化识别，再通过深度残差网络模型进行二次甚至更多次识别分析处理，能够等到更精确的特定发声物体的种类，进而能够构建出更合适的音频。

附图说明

[0046] 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，显而易见地，下面描述中的附图仅仅是本发明的一些实施例，对于本领域普通技术人员来讲，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。

[0047] 图1是本发明的方法流程示意图；

[0048] 图2是本发明的系统结构示意图。

具体实施方式

[0049] 下面结合实施例对本发明做进一步的详细说明，以下实施例是对本发明的解释而本发明并不局限于以下实施例。

[0050] 一种视频物体识别构建音频方法，如图1所示，包括以下步骤：

[0051] S100、基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0052] S200、采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0053] S300、对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0054] S400、基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0055] 在本实施例中，待处理的视频是指用户提供的需要添加音效的视频片段，采取降频抽帧方式对待处理的视频提取视频关键帧，并且将抽帧频率设置为可调参数，抽帧频率的不设置下限，由视频本身的采码率(通常视频为25帧每秒)来决定抽帧频率的上限，对待处理的视频抽帧后可生成有时序的静态帧图即帧图流，Frame Image Stream，此帧图流用于下一步的特定发声物体识别。

[0056] 在实现的过程中，首先需要降频抽取视频关键帧：基于待处理的视频中出现的有配音价值的物体/人物需有一定的连续存在时长，一般不考虑在待处理的视频的一两帧之内出现消失的物体配音，因为这种从配音技术上来说，意义不大。在具体操作中，如果帧图流中的视频关键帧是这样的：识别的物体类别若2秒之前的帧种不含有该类别，则视为该物体从这一秒开始发声；若前2秒到该帧种已经存在次物体，则视为物体在继续发声，发声时间最小值设为5秒。实际操作中，还可以根据物体的发声规律，为不同物体设定不同的继续发声时间以及最小发声时间。通过降低视频关键帧的频率的方式抽取视频关键帧用以物体识别：比如采码率为25帧/秒的视频，通过降频后将采样关键帧的频率设置为1张/秒，也就是说从每25张关键帧图片中抽取一帧作为未来一秒视频内出现物体的识别输入样本，这样的话就能有效简单的降低读取次数，从而提升处理速度。同时，将抽帧频率设置为可调参数，抽帧频率的不设置下限，由视频本身的采码率(通常视频为25帧每秒)来决定抽帧频率的上限，这样使用者根据自己的视频样本特点决定合适的抽帧频率。

[0057] 再次，通过抽取视频关键帧生成的帧图流，并基于内嵌的深度卷积神经网络(Deep CNN)进行模块化的多物体识别。对于帧图流中的每张静态帧图，通过网络对图像个像素点RGB三色通道的像素值进行高度非线性运算，生成以各个可识别特定发声物体为中心的概率向量，深度卷积神经网络经由各概率向量中的最大值判断特定发声物体的类别，并根据在特定发声物体中心周围矩形区域内概率向量的数值分布特点裁定当前物体选择框的大小。所生成的选择框将用于截取每帧图像中某个具体特定发声物体的截图，以进行第二阶段更详细的特定发声物体识别。需要解释的是：在这个步骤中所有涉及到的神经网络均来自于python语言、TensorFlow深度学习框架里物体识别程序库里的预训练Fast-RCNN网络。

[0058] 本实施例得到模块化的特定发声物体，相应的是采用了模块化设计嵌入物体识别的各层深度卷积神经网络。使用的深度卷积神经网络可以任意切换各级物体识别中所需的一级深度神经网络以适应特殊的使用场景或特别的物体类别，例如对鞋与地面进行细化分类的识别网络便不基于任何预训练过的CNN模型。模块化设计可扩展为在各级识别中嵌入多个深度卷积神经网络，并利用集成学习(Ensemble Learning)的算法，提升总体物体识别的正确率、定位精度以及细化分类的识别准确度。

[0059] 例如：利用多个深度神经网络对同一张视频关键帧进行多物体识别，每个深度神经网络对于每个识别出的特定发声物体的选择框可能在大小位置上稍有不同，集成学习算

法可以利用每个深度神经网络对所识别的选择框的确信值(0到1之间,越接近1说明网络越确定选择框的正确性,确信值是模型对于物体识别是否正确的概率判断,可以理解为模型对于一次物体识别的信心,信心越高则这次物体识别的正确率越高。)将多个选择框进行加权平均,从而微调出一个更可靠的物体定位的选择框,以便产生更高质量的截图进行后续步骤的识别。

[0060] 得到模块化的特定发声物体还需要通过深度残差网络模型进行多级识别分析处理,得到特定发声物体的种类并提取其发声特征。具体可以参见以下方式:

[0061] 现有的深度神经网络还不能从一副自然图像中识别所有物体的细节,因此可以提出了多级物体识别网络的技术解决构架。在此实施例中,多级识别分析处理遵循的是“由粗到细”的设计理念:针对一副帧图流中的每张静态帧图,先利用一级的深度神经网络进行初步分析识别处理,得到大概的特定发声物体的种类(例如人物,鞋类,门窗),再针对每个物体所在位置的细部截图、利用新的神经网络进行物体细分种类的多级识别分析处理,得到特定发声物体的种类(例如鞋类是否是运动鞋、板鞋或皮鞋)。本实施例的多级识别分析处理能扩展为更多级(比如三级或者三级以上)的图像识别构架,一般情况下,由于实验中用到的抽帧图象清晰度受到限制,采用二级深度神经网络进行二级识别分析处理就可以实现目前所需的功能。

[0062] 在此,重点讲述通过二级深度神经网络进行二级识别分析处理的过程:初步识别分析处理是采用第一级深度识别网络,其来源于预训练Fast-RCNN网络;多级识别分析处理采用的是多级深度识别网络,在此,采用的是二级识别分析处理的二级深度识别网络,其针对第一级深度识别网络识别出的个别关键物件进行进一步细化识别,例如针对第一级深度识别网络在静态帧图中识别出的“鞋类”,二级深度识别网络会针对“鞋类”部分的截图再进行二次识别分析处理,以判断“鞋的种类”以及“地面种类”。更加具体地,本实施例可识别四种不同的细化鞋类(运动鞋,皮鞋,高跟鞋,其他),以及五种不同的细化地面(瓷砖地,木板地,水泥地,沙地,其他)。二级深度识别网络的具体网络架构是基于50层的深度残差网络(Resnet50)设计而成。参见以下深度残差网络模型获得过程如下:

[0063] S310、获取若干包含特定发声物体的图像,剔除不合格的特定发声物体的图像,得到合格特定发声物体的图像;

[0064] S320、将合格特定发声物体的图像进行预处理,得到合格特定发声物体的图像数据集,并划分为训练集和验证集;

[0065] S330、将训练集输入至初始深度残差网络模型中进行训练,再通过验证集对训练结果进行进行验证,得到能够获取到特定发声物体的种类的深度残差网络模型。

[0066] 在现有技术中是不存在针对鞋类或地面或者其他特定发声物体识别进行过预训练的深度残差网络,本实施例中使用的深度残差网络不基于任何预训练参数,其网络参数完全从随机数进行原始训练,训练所需的图像集均来自实际视频的截图,并针对出现的鞋子和地面种类进行人工标定。此图像训练集至少包含17000+张大小不一,宽高比不定,最高分辨率不超过480p的,并且主体为鞋和地面胡总恶化是其他特定发声物体的图片,在训练深度残差网络模型,需要剔除那些不合格的图像,比如很模糊的、图片中物体是残缺的这种图片,将剩余的合格的图像分为训练集和验证集。这些图片不同于公开的图像识别数据集,它们大多为形状非正方形的低分辨率图片,这是考虑到了实际使用场景中视频抽帧的图片

的截图形状并不规则、分辨率也可能由于视频压缩算法而降低，不规则与低分辨率可理解为图像集里内含的噪音，从而使在此数据集上训练出的网络拥有更强的抗噪能力，并且对于鞋类和地面有优化过的针对性。通过本实施例的深度残差网络得到对于地面的五种细化种类的识别正确率(计算于测试集上)达到73.4％，远高于随机选择(20％)和从众选择(35.2％)；对四种鞋类的识别精度也在同一量级；实际识别速度利用单一英伟达P100显卡可达100张图每秒。

[0067] 并额外将Resnet50固有的、网络末端的单层感知机(Multi-layer perceptron)加深为两层，配以随机失活设计(Dropout＝0.5)，以适应各种具体物件所需的识别类别的种类要求，这样就能在一定程度上避免因网络参数过多而造成的过拟合现象(对训练集的识别效果远优于测试集的情况)。

[0068] 本实施例中采用的深度残差网络(Resnet50)，本实施例是基于现有深度残差网络做了相应的训练，使得能够识别本实施例所需要的特定发声物体的种类，也就是说对单张图片的计算识别流程以及针对具体使用场景进行了相应的改动，深度残差网络(Resnet50)可读取像素值不低于224×224的方形RGB图像，对于形状为矩形以及长宽不为224像素的输入图像，本实施例采用常规的线性插值法将输入图像首先变形为规则的224×224×3(RGB三个颜色通道)的浮点数矩阵；矩阵输入网络后会经过一系列卷积区块进行卷积运算变形为抽象度越来越高，尺寸越来越小的特征图(feature maps)；卷积区块是卷积神经网络(CNN)常规设计的基础单元，Resnet50中使用的卷积区块由三个到四个二维卷积层(2D convolution)配合随机失活设计(dropout)、批量归一化层(batch normalization)以及线性整流层(rectified linear unit,ReLU)构成，同时与每个区块并行的还有一个残差通路(residual layer,只包含简单的一层二维卷积层或是对输入矩阵的简单复制)。前一区块输出的特征图通过残差通路和卷积区块通路分别计算后输出为两个新的，维度一致的矩阵，简单相加后构成下一区块的输入矩阵。深度残差网络(Resnet50)名称中的数字指代所有的卷积区块中一共包含50层二维卷积层。通过所有卷积区块后的深度残差网络输出为2048维的一阶向量，然后通过一层的感知机(Perceptron)输出为维度1000的向量，本实施例在此基础上增加了一层输出维度可调的感知机，用以符合实际物体细类识别的种类数量，也即用于鞋类识别的输出维度即为4，地面识别的输出维度即为5。深度残差网络最后的输出向量的各元素值代表了图像属于某一种类的概率值，最终的种类标定由最大的概率值决定。与Resnet50类似的常用深度残差网络还有Resnet34，Resnet101等；其他的常用图像识别网络还有Alexnet、VGGnet、InceptionNet等，在此，其实这些也可以应用在本实施例中，但是效果不佳，因此选取了深度残差网络(Resnet50)。

[0069] 另外，本实施例中二级识别网络构架即深度残差网络(Resnet50)同时支持反馈学习模式：在二级深度识别网络的识别准确率达不到场景需求的时候，可以通过第一级深度识别网络识别的物体选定框对帧图流进行截图，以截图作为新的数据集进行人工标定并微调二级深度识别网络即深度残差网络(Resnet50)。这样的话就可以在待处理的视频内容发生重大改变的时候利用已训练的模型和少量新数据迅速获得较高的识别准确率，并由此缩短适应新的应用场景的准备周期。第一级深度识别网络也可根据视频类型的变化或应用场景的变化进行阶段性的重新训练，以适应新的视频数据特点。

[0070] 进一步，二级深度识别网络中各级识别出的特定发声物体信息采用同样的格式合

并储存。针对每个物体储存信息有:物体大类(上级网络识别)、物体大类确信值、物体细类(二级深度识别网络识别)、物体细类确信值、物体定位选择框宽高以及中心(以帧图像素为测量单位),所有信息以json文件格式进行下一步处理。

[0071] 识别出特定发声物体的物体类型之后,为了使得特定发声物体的物体类别和特定发声物体音频产生关联,本实施例采用自然语言作为待处理的视频特定发声物体的物体类别与音频匹配的中间表达,自然语言作为匹配表达的方法使得表达利于人们理解和标注,以及音频库整理和维护。

[0072] 对于待处理的视频,从视频中识别出的物体类别作为自然语言表示(例如“猫”);对于音频,可以使用两类自然语言标注:音频介绍和音频关键词,即音频包括音频介绍和音频关键词,在此,音频介绍可以理解为:用一句话或短语来介绍音频的内容(例如“一个人在雪地里走路的声音”),音频关键词用三个关键单词来音频的内容(例如“鞋/雪地/脚步声”)。不同于音频介绍,音频关键词一定要包含了发声物体和发声声音类别,综上,音频关键词的引入连接了物体识别类别和声音介绍之间的不匹配。在此,可以将音频解析为音频介绍和音频关键词,其中,音频介绍为音频的介绍内容文本,音频关键词包括至少三个描述音频的词语,所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称。

[0073] 对于特定发声物体,直接使用物体识别的类别名作为其自然语言表示,因为计算机不能理解理解自然语言,因此进一步将自然语言表达映射为向量表达。具体来讲,本实施例引入两种自然语言的向量表达:TF-IDF(term frequency-inverse document frequency)与BERT(Bidirectional Encoder Representations from Transformers)。

[0074] 在具体的实施例中,通过音频介绍文本来计算TF-IDF向量,TF-IDF向量表示的是每个单词在一段文字中对一段文字整体的语义有多大的影响。具体为:先通过分词器“结巴分词”,对所有的音频的音频介绍进行中文分词:再计算每个单词在每个音频介绍中的词频TF,以及每个单词在所有音频介绍的集合中的词频DF;对于一个音频介绍,可以计算其中任何一个单词的TF-IDF:TF-IDF=TF*log(1/DF+1);在此一定要注意,此TF-IDF计算公式为一种归一化后的TF-IDF,目的是为了确保数值稳定性;最后,对于任意一段文字,都计算出这段文字的TF-IDF向量。先将所有文本库单词排序,按照该顺序,计算每个单词在该段文字的TF-IDF值,若该段文字不包含此单词,则视其TF-IDF值为0。最后,得到长度与文本库词汇量相同的向量,即为该段文字的TF-IDF向量表达。

[0075] 进一步地:计算BERT向量,在实施例中的BERT是一种Transformer神经网络结构,用大规模的无监督学习训练网络参数,所得的模型可以直接运用到下游自然语言理解问题中,直接对自然语言中的句子和短语进行向量映射。本实施例将两者(以及简单的单词匹配)结合的方法,这样结果更加精准。

[0076] 在一个实施例中,使用Pytorch库中的pytorch_pretrained_bert中的预训练的中文BERT模型计算句子的BERT向量表达。为了满足匹配的效率,采用最小的BERT模型“bert_base_chinese”。具体来讲,先将句子拆分陈逐个字符,在句子的最前和最后分别加入“[CLS]”和“[SEP]”字符,作为输入index_tokens,将和入index_tokens同长的全0列表作为输入segment_ids,将两个输入同时输入到预训练BERT模型中,取出最后一层神经网络对应第一个字符(“[CLS]”)的输出向量,作为该句的BERT向量。

[0077] 在一个实施例中,更具体地,步骤S400中所述基于特定发声物体的种类提取其发

声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，具体步骤为：

[0078] S410、基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数；

[0079] S420、基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，根据视频音频匹配分数得到特定发声物体至少一种合适的音频。

[0080] 音频和视频构建的过程即是视频中识别出的物体类别和音频介绍，音频关键词的匹配，通过计算出的匹配分数来选择合适的音频，匹配分数的计算是通过两种方式进行的，一种是传统方法，一种是神经网络方式，传统方法的优势在于当音频与视频的自然语言表达有相同单词时，能准确计算出分数；神经网络计算匹配分数的的优势在于当两者自然语言表达没有单词重叠时，可以做到对语言表达大意的匹配，本实施例同时使用两种方法的分数并将其合并，有助于做到两种方法的互补。

[0081] 在一个实施例中，在步骤S410中所述基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数，具体为：

[0082] S411、对特定发声物体的物体类别和音频介绍进行分词处理得到单词；

[0083] S412、分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例，得到第一比例和第二比例，将第一比例和第二比例进行加权平均处理，得到单词匹配分数，所述单词匹配分数＝物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重，其中，音频介绍权重+音频关键词权重＝1；

[0084] S413、基于音频介绍的统计数据，得到物体类别TF-IDF向量，通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度，将第一余弦相似度作为TF-IDF匹配分数，所述TF-IDF匹配分数＝cosine_similarity(物体类别TF-IDF向量，音频介绍TF-IDF向量)；

[0085] S414、将单词匹配分数和TF-IDF匹配分数进行加权平均处理，得到第一匹配分数，所述第一匹配分数＝单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重，其中，单词权重+TF-IDF权重＝1；

[0086] S415、获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，经过计算得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数。

[0087] 上述视频物体音效搜索匹配系统中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块可以硬件形式内嵌于或独立于计算机设备或者移动终端的处理器中，也可以以软件形式存储于计算机设备或者移动终端的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0088] 步骤S411-步骤S414是采用传统方法进行匹配的，首先用结巴分词器对特定发声物体的物体类别，声音介绍进行分词。然后计算特定发声物体的物体类别分别和声音介绍，声音关键词重合的单词比例，并将两个比例加权平均，作为单词匹配分数；根据声音介绍文本中的统计数据，得到特定发声物体的物体类别的TF-IDF向量表达。之后，计算物体TF-IDF向量与声音介绍TF-IDF向量的余弦相似度，作为TF-IDF匹配分数，将单词匹配分数和TF-IDF匹配分数加权平均，得到传统方法匹配分数也就是步骤中所说的第一匹配分数。当然，于其他实施例中，得到第一匹配分数的技术也可能是其他技术手段，在此不再赘述。

[0089] 而步骤S415中获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，进而得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数是通过神经网络匹

配的方法实现的。

[0090] 在一个实施例中，步骤S420中所述基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，具体为：将第一匹配分数和神经网络匹配分数进行加权平均处理，得到视频音频匹配分数，所述视频音频匹配分数＝第一匹配分数*第一权重+神经网络匹配分数*神经网络权重，其中，第一权重+神经网络权重＝1。

[0091] 实际操作中，可以根据需要调节加权平均的权重，如果希望特定发声物体的物体类别的名称准确出现在音频介绍或关键词中，可以增加传统匹配分数的权重来增加准确性；如果希望特定发声物体的物体类别的名称虽不在音频介绍或关键词中，但语义相同，可以增加神经网络匹配分数的权重来增加泛化性。

[0092] 具体地，可以根据最终匹配分数，对于每个识别出的物体，选择出10个最匹配的音频作为配音推荐，当然其他个数也是可以的。

[0093] 在一个实施例中，所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，还包括如下步骤：

[0094] S500、根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配；

[0095] S600、将所有音频进行混音处理，形成完整的音频文件，将音频文件添加进视频的音轨使得音频文件和视频同步。

[0096] 具体地，根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配，也就是现有技术中的给视频中的特定发声物体进行拟音，在此是通过视频音频匹配分数使得特定发声物体和音频进行匹配，这是单独的配音。后面还可以进行整体配音，就是将产生的音频进行混音，当找到配音所需的音频文件以及各个音频文件播放的起止时间后，就能读取所需的所有音频文件，并将每一个音频文件转化为统一的频域信号格式，以方便后续的剪辑。

[0097] 在本实施例中，可以读取任意常用格式的音频文件，包括wav和mp3等，提高了使用场景以及泛化至其它特定音频库的能力。

[0098] 将所有音频进行混音处理的具体过程为：每一段音频将被智能拉伸或压缩至配音所需要的时长，首先把音频开始和结束阶段的静音部分切除，这样可以使得配音与视频中触发配音的画面同时发生，使得配音效果最佳。再查看消除首尾静音后的音频时长是否比需要播放的时间更长，若是，则剪切音频至配音所需播放时长，并在末尾使用渐出效果，以消除音频突然暂停的突兀感。若不是，则循环播放该音频直至配音所需播放时长，在循环播放时，前后两段音频的首尾相连处将采用一定时长的重叠以及渐入渐出效果，以使循环播放处无缝衔接，让这段长音频听起来自然完整，使用户拥有最佳的听觉体验。渐入渐出的时长将与重叠的时长相等，时长将通过一个分段函数依据音频时长确定，若原音频时长小于20秒，则将重叠与渐入渐出时间设为音频时长的10％，这样能使得重叠部分时长适中，有利于平缓地过渡前后段音频，这样也有利于短视频更多的保留非重叠部分以播放给用户。若原音频时长大于20秒，则将重叠与渐入渐出时间设为2秒，这样可以避免长音频出现不必要的长过渡期，以尽可能地播放非重叠的音频。

[0099] 最后，按上述步骤处理过的各个音频合并到一起，并添加进视频的音轨，输出新的带有配音的视频文件，完成整个配音过程。

[0100] 实施例2：

[0101] 一种视频物体识别构建音频系统，如图2所示，包括帧图流生成模块100、第一处理模块200、第二处理模块300和提取构建模块400；

[0102] 所述帧图流生成模块100被设置为：基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0103] 所述第一处理模块200，用于采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0104] 所述第二处理模块300，用于对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0105] 所述提取构建模块400被设置为：基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0106] 在一个实施例中，所述第二处理模块300中深度残差网络模型获得过程如下：

[0107] 获取若干包含特定发声物体的图像，剔除不合格的特定发声物体的图像，得到合格特定发声物体的图像；

[0108] 将合格特定发声物体的图像进行预处理，得到合格特定发声物体的图像数据集，并划分为训练集和验证集；

[0109] 将训练集输入至初始深度残差网络模型中进行训练，再通过验证集对训练结果进行进行验证，得到能够获取到特定发声物体的种类的深度残差网络模型。

[0110] 在一个实施例中，所述提取构建模块400被设置为音频包括音频介绍和音频关键词，音频介绍为音频的介绍内容文本，音频关键词包括至少三个描述音频的词语，所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称。

[0111] 在一个实施例中，所述提取构建模块400被设置为：所述基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频，具体步骤为：

[0112] 基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数；

[0113] 基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，根据视频音频匹配分数得到特定发声物体至少一种合适的音频。

[0114] 在一个实施例中，所述提取构建模块400被设置为：

[0115] 对特定发声物体的物体类别和音频介绍进行分词处理得到单词；

[0116] 分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例，得到第一比例和第二比例，将第一比例和第二比例进行加权平均处理，得到单词匹配分数，所述单词匹配分数＝物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重，其中，音频介绍权重+音频关键词权重＝1；

[0117] 基于音频介绍的统计数据，得到物体类别TF-IDF向量，通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度，将第一余弦相似度作为TF-IDF匹配分数，所述TF-IDF匹配分数＝cosine_similarity(物体类别TF-IDF向量，音频介绍TF-IDF向量)；

[0118] 将单词匹配分数和TF-IDF匹配分数进行加权平均处理，得到第一匹配分数，所述第一匹配分数＝单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重，其中，单词权重+

TF-IDF权重＝1；

[0119] 获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，经过计算得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数。

[0120] 在一种可实施方式中，所述提取构建模块400被设置为：所述基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，具体为：

[0121] 将第一匹配分数和神经网络匹配分数进行加权平均处理，得到视频音频匹配分数，所述视频音频匹配分数＝第一匹配分数*第一权重+神经网络匹配分数*神经网络权重，其中，第一权重+神经网络权重＝1。

[0122] 在一个实施例中，还包括搜索匹配模块500和混音处理模块600；

[0123] 所述搜索匹配模块500，用于根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配；

[0124] 所述混音处理模块600，用于将所有音频进行混音处理，形成完整的音频文件，将音频文件添加进视频的音轨使得音频文件和视频同步。

[0125] 在混音处理模块中设置简单易用的函数接口，可以一键生成配音视频，极大地提升了使用者的工作效率。混音处理模块600虽然使用了很常见的音频工具，但是正如方法中的具体的混音步骤和参数是专门为电影、网剧、短视频而设计的，比如在方法实施例中提到的静音去除以及特效音频的压缩或延长方法可以针对性地解决上述特定类别视频的配音问题，即特效音频库中的音频长度很多时候并不满足视频配音需求的问题，这些特定的音频处理参数也是最适合本实施例的，别的技术或者音频处理参数是实现不了的。

[0126] 对于系统实施例而言，由于其与方法实施例基本相似，所以描述的比较简单，相关之处参见方法实施例的部分说明即可。

[0127] 实施例3：

[0128] 一种计算机可读存储介质，所述计算机可读存储介质存储有计算机程序，所述计算机程序被处理器执行时实现以下的方法步骤：

[0129] 基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流；

[0130] 采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0131] 对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类；

[0132] 基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频。

[0133] 在一个实施例中，处理器执行计算机程序时，实现所述对待处理视频进行识别处理，得到待处理视频中的特定发声物体的种类并提取其发声特征，具体为：

[0134] 将待处理的视频的相关信息降低抽帧频率，抽取视频关键帧；

[0135] 将抽取的视频关键帧生成帧图流；

[0136] 采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体；

[0137] 对模块化的特定发声物体通过深度残差网络模型进行多级识别分析处理，得到待处理视频中的特定发声物体的种类并提取其发声特征。

[0138] 在一个实施例中，处理器执行计算机程序时，实现音频介绍为音频的介绍内容文本，音频关键词包括至少三个描述音频的词语，所述描述音频的词语包括特定发声物体的类别名称和发声声音的类别名称。

[0139] 在一个实施例中，处理器执行计算机程序时，实现所述基于特定发声物体的物体类别、音频介绍以及音频关键词进行分数匹配处理分别得到第一匹配分数和神经网络匹配分数，具体为：

[0140] 对特定发声物体的物体类别和音频介绍进行分词处理得到单词；

[0141] 分别获取特定发声物体的物体类别与音频介绍、音频关键词重合的单词比例，得到第一比例和第二比例，将第一比例和第二比例进行加权平均处理，得到单词匹配分数，所述单词匹配分数＝物体类别和音频介绍的单词重合比例*音频介绍权重+物体类别和音频关键词单词重合比例*音频关键词权重，其中，音频介绍权重+音频关键词权重＝1；

[0142] 基于音频介绍的统计数据，得到物体类别TF-IDF向量，通过物体类别TF-IDF向量与音频介绍TF-IDF向量的第一余弦相似度，将第一余弦相似度作为TF-IDF匹配分数，所述TF-IDF匹配分数＝cosine_similarity(物体类别TF-IDF向量，音频介绍TF-IDF向量)；

[0143] 将单词匹配分数和TF-IDF匹配分数进行加权平均处理，得到第一匹配分数，所述第一匹配分数＝单词匹配分数*单词权重+TF-IDF匹配分数*TF-IDF权重，其中，单词权重+TF-IDF权重＝1；

[0144] 获取特定发声物体的物体类别的BERT向量和音频介绍的BERT向量，经过计算得到BERT向量的余弦相似度，将余弦相似度作为神经网络匹配分数。

[0145] 在一个实施例中，处理器执行计算机程序时，实现所述基于第一匹配分数和神经网络匹配分数得到视频音频匹配分数，具体为：

[0146] 将第一匹配分数和神经网络匹配分数进行加权平均处理，得到视频音频匹配分数，所述视频音频匹配分数＝第一匹配分数*第一权重+神经网络匹配分数*神经网络权重，其中，第一权重+神经网络权重＝1。

[0147] 在一个实施例中，处理器执行计算机程序时，实现所述根据视频音频匹配分数得到特定发声物体的一种或者几种合适音频步骤之后还包括：

[0148] 根据视频音频匹配分数将特定发声物体与选择的音频进行搜索匹配，使得音频介绍、音频关键词与特定发声物体的物体类别相互匹配；

[0149] 将所有音频进行混音处理，形成完整的音频文件，将音频文件添加进视频的音轨使得音频文件和视频同步。

[0150] 实施例4：

[0151] 在一个实施例中，提供了一种视频物体识别构建音频装置，该视频物体识别构建音频装置可以是服务器也可以是移动终端。该视频物体识别构建音频装置包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该视频物体识别构建音频装置的处理器用于提供计算和控制能力。该视频物体识别构建音频的装置的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该数据库用于存储视频物体识别构建音频的装置的所有数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现视频物体音效构建方法。

[0152] 本说明书中的各个实施例均采用递进的方式描述，每个实施例重点说明的都是与其他实施例的不同之处，各个实施例之间相同相似的部分互相参见即可。

[0153] 本领域内的技术人员应明白，本发明的实施例可提供为方法、装置、或计算机程序产品。因此，本发明可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本发明可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0154] 本发明是参照根据本发明的方法、终端设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理终端设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理终端设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0155] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理终端设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0156] 这些计算机程序指令也可装载到计算机或其他可编程数据处理终端设备上，使得在计算机或其他可编程终端设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程终端设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0157] 需要说明的是：

[0158] 说明书中提到的“一个实施例”或“实施例”意指结合实施例描述的特定特征、结构或特性包括在本发明的至少一个实施例中。因此，说明书通篇各个地方出现的短语“一个实施例”或“实施例”并不一定均指同一个实施例。

[0159] 尽管已描述了本发明的优选实施例，但本领域内的技术人员一旦得知了基本创造性概念，则可对这些实施例做出另外的变更和修改。所以，所附权利要求意欲解释为包括优选实施例以及落入本发明范围的所有变更和修改。

[0160] 此外，需要说明的是，本说明书中所描述的具体实施例，其零、部件的形状、所取名称等可以不同。凡依本发明专利构思所述的构造、特征及原理所做的等效或简单变化，均包括于本发明专利的保护范围内。本发明所属技术领域的技术人员可以对所描述的具体实施例做各种各样的修改或补充或采用类似的方式替代，只要不偏离本发明的结构或者超越本权利要求书所定义的范围，均应属于本发明的保护范围。

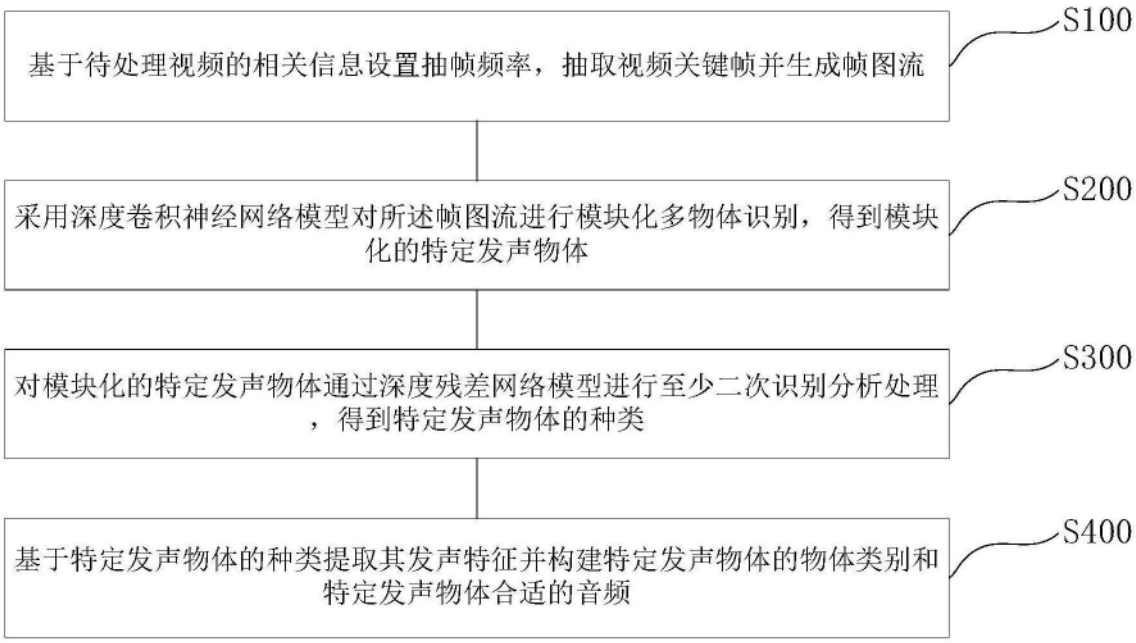

基于待处理视频的相关信息设置抽帧频率，抽取视频关键帧并生成帧图流
S100
采用深度卷积神经网络模型对所述帧图流进行模块化多物体识别，得到模块化的特定发声物体
S200
对模块化的特定发声物体通过深度残差网络模型进行至少二次识别分析处理，得到特定发声物体的种类
S300
基于特定发声物体的种类提取其发声特征并构建特定发声物体的物体类别和特定发声物体合适的音频
S400

图1

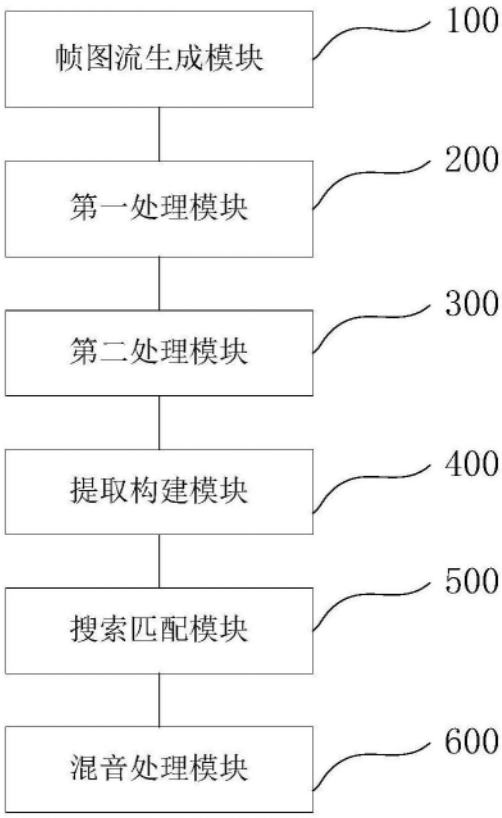

帧图流生成模块
100
第一处理模块
200
第二处理模块
300
提取构建模块
400
搜索匹配模块
500
混音处理模块
600

图2