(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 108447501 B
(45)授权公告日 2020.08.18

(21)申请号 201810258659.5

(22)申请日 2018.03.27

(65)同一申请的已公布的文献号
申请公布号 CN 108447501 A

(43)申请公布日 2018.08.24

(73)专利权人 中南大学
地址 410083 湖南省长沙市岳麓区麓山南路932号

(72)发明人 段桂华 滕明英 王琰 张振宇

(74)专利代理机构 长沙市融智专利事务所(普通合伙) 43114
代理人 龚燕妮

(51)Int.Cl.
G10L 25/24(2013.01)
G10L 25/45(2013.01)
H04N 21/233(2011.01)
H04N 21/81(2011.01)
H04N 21/854(2011.01)
G06K 9/62(2006.01)

(56)对比文件
CN 102024033 A,2011.04.20
CN 1835580 A,2006.09.20
CN 1920947 A,2007.02.28
CN 103403710 A,2013.11.20
CN 106340310 A,2017.01.18
CN 103198838 A,2013.07.10
CN 104936022 A,2015.09.23
CN 107293307 A,2017.10.24
CN 106162321 A,2016.11.23
WO 02073593 A1,2002.09.19
US 2005027766 A1,2005.02.03
EP 1760693 A1,2007.03.07
EP 2263335 A1,2010.12.22
EP 3142107 A1,2017.03.15
B. Srinivas 等.Movie Piracy Detection Based on Audio Features Using Mel-Frequency Cepstral Coefficients and Vector Quantization.《International Journal of Soft Computing and Engineering》.2012,第2卷(第4期),

审查员 蒋栗

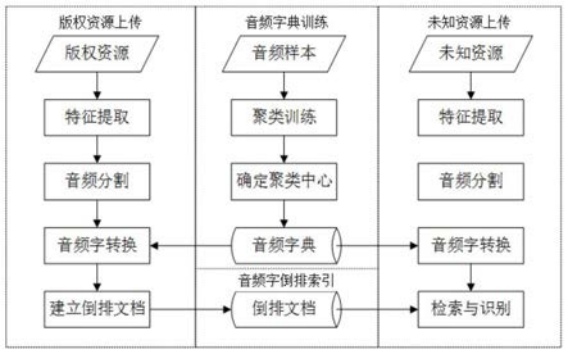
权利要求书4页 说明书11页 附图2页

(54)发明名称

一种云存储环境下基于音频字的盗版视频检测方法与系统

(57)摘要

本发明公开了一种云存储环境下基于音频字的盗版检测方法和系统,通过多维聚类构建的音频字典,对版权资源分割后的短时音频段进行特征提取,音频段转换为可以表征音频内容的音频字序列后,建立倒排索引。检索时,将用户提交的查询转换为音频字后直接定位候选段落,并根据候选段落与查询的内容相似度与阈值的关系确定视频是否为盗版。充分利用了音频特征在传统音视频媒体文件中的稳定性,以及静默片段等特征,检测结果高效而准确,以极低的本地计算成本和较少的网络带宽,较容易的为正版资源提供了一种版权鉴定保护方案。



CN 108447501 B

1.一种云存储环境下基于音频字的盗版视频检测方法,其特征在于,包括以下步骤:

步骤1:提取音效,并进行音频字标记;

提取各种视频中的音轨,从音轨中提取各种常见音效,对所提取的不同音效依次进行标号,获得每种音效的音频字;

依据音频字典中的音频字对每个音频片段标记时,寻找音频字典中音频字的超向量与音频片段中所有帧音频的超向量均值距离最小的音频字作为对应音频片段的标记音频字;

步骤2:提取各音频帧超向量;

先提取每帧音频的多维特征,并对多维特征分别进行归一化处理,构建音频帧的超向量;

所述每帧音频的多维特征包括宏观声学特征、时域特征、频域特征以及倒谱特征,所述音效超向量是指对音频多维特征分别进行归一化处理后得到的特征形成的一维向量;

其中,宏观声学特征包括音调、音高、带宽,时域特征包括短时能量、短时平均幅度、短时平均过零率、短时自相关系数,频域特征包括谱能量、子带能量比、谱质心、谱带宽、谱密度,倒谱特征包括Delta倒谱系数、LPC倒谱系数、梅尔倒谱系数;

步骤3:构建音频字典;

依次对每一种音效收集100个样本,提取每个样本中所有音频帧的超向量均值,将每一种音效的所有样本的超向量均值使用k-均值聚类算法聚成3个类,每一个聚类中心作为一个新的音频字w_i,每一个新的音频字均进行唯一标记w_i,利用新的音频字生成音频字典为$W=\{w_1,w_2,\cdots,w_k\}$,$k=1200$;

步骤4:音频分割:

采用3阶段的自顶向下多层分割方法,将步骤1中提取出的音轨D_i依据声学特征分割成音频片段;

步骤5:音频字转换;

计算每个音频片段中所有帧音频的超向量均值,并依据音频字典中的音频字对每个音频片段标记,得到每一个音轨对应的音频字序列$ID_i=\{id_j^i\}$,$1\leqslant j\leqslant N_i$,$id_{N_i}^i$表示第i个音轨中的第j个音频片段对应的音频字;$N_i$表示第i个音轨中包含的音频片段总数;

步骤6:构建音频字在音轨中的时刻位置索引表;

对所有上传的正版视频资源依次进行音轨提取、音频帧超向量提取、音频分割以及音频字转换,并将所有正版视频资源中音频字出现在音轨中的时刻位置进行记录,获得各正版视频中所有音频字出现在各音轨中的时刻位置倒排索引表;

步骤7:音频检索;

将上传的待检测的视频进行音轨提取、音频帧超向量提取、音频分割以及音频字转换得到对应的音频字序列,将待检测视频的音频字序列中包含的音频字按照顺序与所述时刻位置索引表中音频字进行匹配,若待检测的音频字序列中存在至少N个连续的音频字与某一正版视频中连续的音频字一一匹配,则选取对应正版视频中连续匹配的音频字的起始位置所在的候选音频段落C,计算各正版视频候选段落C与待检测视频的音频字序列的匹配度,若匹配度超过设定的匹配度阈值,则判定当前待检测的视频属于对应正版视频的盗版资源;

所述匹配度是指音频字的出现时间的吻合程度。

2.根据权利要求1所述的方法，其特征在于，所述归一化处理是指进行规整向量计算；

规整向量f'_d计算公式为：

$$f_d^{'}=\frac{f_d-\mu_d}{\sigma_d},\ d=1,\cdots,D$$

其中，D为特征总维数，f_d为原始特征，μ_d和σ_d分别为同一类音效特征的均值和标准差；通过该公式将各维特征规整到均值为0，方差为1的分布中。

3.根据权利要求1所述的方法，其特征在于，所述采用3阶段的自顶向下多层分割方法，将步骤1中提取出的音轨D_i依据声学特征分割成音频片段的具体过程如下：

第1阶段：静音分割；

以静音作为分割点对音轨进行粗略分割，得到各粗音频段；

其中，所述静音的短时帧能量至少连续2s均小于能量门限E_{th}；

$$E_{th}=\lambda_s\cdot\min\left(\frac{1}{2}(E_{max}-E_{min}),\ E_{mean}-E_{min}\right)+E_{min}$$

其中，E_{max}、E_{min}和E_{mean}分别代表当前音轨文档中短时帧能量的最大值、最小值和均值，λ_s为静音因子，$\lambda_s\in[0,1]$；

第2阶段：距离分割；

距离分割将经过静音分割后得到的各粗音频段，依据Hotelling's T^2距离再分割成无明显音频波动的音频片段；

利用逐渐增长的第一分析窗依次对各粗音频段进行扫描，并在分析窗中每隔0.2s设置一个测试点，若第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离超过预设第一门限时，对应的测试点所在位置当作音频类型改变点，以音频类型改变点对粗音频段进行分割；

第一分析窗初始长度为3s，如果窗内未发现音频类型改变点，则第一分析窗窗长增加1s，再次对粗音频段进行扫描；如果第一分析窗内找到音频类型改变点，则将第一分析窗长度重置为初始长度，并以得到的新的音频类型改变点作为起点继续搜索下一音频类型改变点直至搜索至粗音频段尾端；

第3阶段：声学特征分割；

根据音频特征的均值和方差，对无明显音频波动的音频片段进行分割；

利用第二分析窗对各无明显音频波动的音频片段进行扫描，以第二分析窗的中点对第二分析窗内的音频片段进行分割得到左侧数据窗和右侧数据窗，计算中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差，其中，μ_1和μ_2分别是第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值：

当欧式距离或者方差超过预设第二门限时，认为第二分析窗内部存在较大的数据变化，则当前中点为音效改变点，以音效改变点对对应的无明显音频波动的音频片段进行分割；

否则，将左侧数据窗向后增加5帧，右侧数据窗向后平移5帧，继续计算左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差直到找到新的音效改变点或者搜索至当前无明显音频波动的音频片段的数据尾端；

第二分析窗的长度初始为30帧。

4.根据权利要求3所述的方法，其特征在于，所述静音因子λ_s设置为0.1。

5.根据权利要求3所述的方法，其特征在于，所述第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离采用以下公式计算：

$$T^2 = \frac{b(N-b)}{N}(S_1 - S_2)^T \sum^{-1} (S_1 - S_2)$$

其中，N为第一分析窗总长度，Σ为协方差矩阵符号，b和S_1分别为第一分析窗测试点左侧数据窗长度和所包含的所有音频帧的超向量均值，S_2为右侧数据窗所包含的所有音频帧的超向量均值。

6.根据权利要求3所述的方法，其特征在于，所述第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$采用以下公式计算获得：

$$\mathrm{dis}(\mu_1,\mu_2) = \sqrt{\sum_{d=1}^{D} |\mu_1(d) - \mu_2(d)|^2}$$

其中，$\mu_1(d)$为左数据窗中所有帧音频的超向量中第d维的特征均值，$\mu_2(d)$为右数据窗中所有帧音频的超向量中第d维的特征均值，D为超向量中的特征维数。

7.根据权利要求1-6任一项所述的方法，其特征在于，所述匹配度按照以下公式计算：

$$R(Q,C) = \frac{\sum_{n=1}^{N_q} hit(Q_n, C_n)}{N_q}$$

其中，R(Q,C)表示待检测视频的音频字序列Q与选中的候选音频段落C的匹配度，$hit(Q_n,C_n)$表示待检测视频的音频字序列的第n个音频字与候选音频段落中的第n个音频字相同或者不同，$hit(Q_n,C_n) = \begin{cases} 1, & Q_n = C_n \\ 0, & Q_n \neq C_n \end{cases}$。

8.根据权利要求7所述的方法，其特征在于，在对上传的待检测的视频进行音频检索前，先提取所上传的视频文件的MD5值，并将该值与所有上传的正版视频资源的MD5值进行比对，若与已上传的正版视频资源的MD5值相同，则判定当前上传的待检测视频属于盗版视频，结束当前上传的待检测的视频的检测流程。

9.一种云存储环境下基于音频字的盗版视频检测系统，其特征在于，采用权利要求1-8任一项所述的一种云存储环境下基于音频字的盗版视频检测方法，包括：

正版资源上传模块，用于上传正版视频资源；

正版资源音频字文档倒排模块，获取正版视频资源，利用音频特征提取模块提取各正版视频资源中包含的音频特征，再依据音频字字典将音频特征转换为音频字，记录所有正版视频资源中音频字出现在各音轨中的时刻位置，形成各正版视频中所有音频字出现在各音轨中的时刻位置倒排索引表，得到音频字倒排索引表；

版权数据库，用于存储各正版资源的音频字倒排索引表；

音频特征提取模块，用于从音轨中提取各帧音频的超向量，所述超向量包括宏观声学特征、时域特征、频域特征以及倒谱特征；

音频字字典构建模块，利用音频特征提取模块对常见音效样本集进行超向量提取，对提取的超向量进行聚类，获取常见音效对应的音频字，构建音频字字典；

待检测资源上传模块，通过云存储提供商的客户端上传待检测视频资源至云服务端；

检测模块，在云服务端将上传的待检测的视频资源利用音频特征提取模块提取各正版视频资源中包含的音频特征，再依据音频字字典将音频特征转换为音频字得到音频字序列，将待检测视频资源的音频字序列中包含的音频字按照顺序与所述版权数据库中的音频字倒排索引表中音频字进行匹配，若待检测的音频字序列中存在至少N个连续的音频字与某一正版视频中连续的音频字一一匹配，则选取对应正版视频中连续匹配的音频字的起始位置所在的候选音频段落C，计算各正版视频候选段落C与待检测视频的音频字序列的匹配度，若匹配度超过设定的匹配度阈值，则判定当前待检测的视频属于对应正版视频的盗版资源。

一种云存储环境下基于音频字的盗版视频检测方法与系统

技术领域

[0001] 本发明属于版权检测领域，特别涉及一种云存储环境下基于音频字的盗版视频检测方法与系统。

背景技术

[0002] 云计算时代，当云存储和各类UGC(User Generated Content)视频网站服务商的兴起，多媒体信息呈现爆炸式增长。数字音视频具有传播速度快、覆盖范围广、盗版成本低等特点，这都对数字版权形成了巨大的挑战。我们常常可以看到，一部影片，在不同视频网站的版本各不相同，甚至在同一网站，能搜索到一部影片的多个版本，通过而这些所谓的“山寨”版本，大多是由个人或团体从各种渠道获得的正版或盗版片源，经过翻录等手段获得盗版的视频副本，再使用私人账号将其上传到视频网站而来。这些盗版视频，严重地侵犯了视频制作方和发行方的合法权利与利益，对多媒体产业的发展以及社会价值取向的形成具有不良影响。

[0003] 面对这种情况，目前主流网站多在用户上传视频后、视频发布前，预先通过机器检测或人工审核等方式对视频内容进行预先审核，在视频发布后接受用户举报，查实后删除违规视频并视情节对账号进行封号处理。网站平台的数据流量大且时效性高，目前的机器检测大多采用图像识别技术，这项技术漏判、误判的情况时有发生。而且对海量视频帧进行画面识别导致处理效率相对低下，如果发布者通过降低清晰度、加快或放慢视频播放速度等方式，对图像识别进行干扰，进一步降低图像识别的准确度。人工审核的准确度很高，但耗费的审核时间相当长。举报封删的方法是建立在观众举报的基础之上的，观众可能出于各种原因并不举报违规视频，同时审核举报信息也会耗费一定时间。举报封删是一种事后补救措施，即便删除违规视频能阻止进一步扩散，但其已造成的负面影响是难以消除的。

[0004] 由于目前盗版检测的方式还存在上诉缺点，因此有必要设计一种新的云存储环境下的盗版检测方案，能够满足以下要求：(i)准确性。不仅可以检测出翻录的视频，对经过噪声和变速处理的视频也应该起到较为准确的检测作用。(ii)低成本。更少的数据存储空间和I/O开销。(iii)实时性。检测速度应该满足网站的高实时性要求。

发明内容

[0005] 本发明提供了一种云存储环境下基于音频字的盗版视频检测方法与系统，其目的在于，克服现有技术中盗版资源人工审核周期长或图像识别准确度较低的问题。

[0006] 一种云存储环境下基于音频字的盗版视频检测方法，包括以下步骤：

[0007] 步骤1:提取音效，并进行音频字标记；

[0008] 提取各种视频中的音轨，从音轨中提取各种常见音效，对所提取的不同音效依次进行标号，获得每种音效的音频字；

[0009] 步骤2:提取各音频帧超向量；

[0010] 先提取每帧音频的多维特征，并对多维特征分别进行归一化处理，构建音频帧的

超向量；

[0011] 所述每帧音频的多维特征包括宏观声学特征、时域特征、频域特征以及倒谱特征，所述音效超向量是指对音频多维特征分别进行归一化处理后得到的特征形成的一维向量；

[0012] 其中，宏观声学特征包括音调、音高、带宽，时域特征包括短时能量、短时平均幅度、短时平均过零率、短时自相关系数，频域特征包括谱能量、子带能量比、谱质心、谱带宽、谱密度，倒谱特征包括Delta倒谱系数、LPC倒谱系数、梅尔倒谱系数；

[0013] 步骤3：构建音频字典；

[0014] 依次对每一种音效收集100个样本，提取每个样本中所有音频帧的超向量均值，将每一种音效的所有样本的超向量均值使用k-均值聚类算法聚成3个类，每一个聚类中心作为一个新的音频字w_i，每一个新的音频字均进行唯一标记w_i，利用新的音频字生成音频字典为$W=\{w_1,w_2,\cdots,w_k\}$，$k=1200$；

[0015] 步骤4：音频分割：

[0016] 采用3阶段的自顶向下多层分割方法，将步骤1中提取出的音轨D_i依据声学特征分割成音频片段；

[0017] 音频片段中的声学特征变化程度较小；

[0018] 步骤5：音频字转换；

[0019] 计算每个音频片段中所有帧音频的超向量均值，并依据音频字典中的音频字对每个音频片段标记，得到每一个音轨对应的音频字序列$ID_i=\{id_j^i\}$，$1\leqslant j\leqslant N_i$，$id_{N_i}^i$表示第i个音轨中的第j个音频片段对应的音频字；$N_i$表示第i个音轨中包含的音频片段总数；

[0020] 步骤6：构建音频字在音轨中的时刻位置索引表；

[0021] 对所有上传的正版视频资源依次进行音轨提取、音频帧超向量提取、音频分割以及音频字转换，并将所有正版视频资源中音频字出现在音轨中的时刻位置进行记录，获得各正版视频中所有音频字出现在各音轨中的时刻位置倒排索引表；

[0022] 倒排文件记录的是音频字出现在第几个音轨的第几个位置上，使用倒排索引可以加速检索速度。

[0023] 步骤7：音频检索；

[0024] 将上传的待检测的视频进行音轨提取、音频帧超向量提取、音频分割以及音频字转换得到对应的音频字序列，将待检测视频的音频字序列中包含的音频字按照顺序与所述时刻位置索引表中音频字进行匹配，若待检测的音频字序列中存在至少N个连续的音频字与某一正版视频中连续的音频字一一匹配，则选取对应正版视频中连续匹配的音频字的起始位置所在的候选音频段落C，计算各正版视频候选段落C与待检测视频的音频字序列的匹配度，若匹配度超过设定的匹配度阈值，则判定当前待检测的视频属于对应正版视频的盗版资源；

[0025] 所述匹配度是指音频字的出现时间的吻合程度。

[0026] 进一步地，所述归一化处理是指进行规整向量计算；

[0027] 规整向量f'_d计算公式为：

[0028] $$f_d^{'}=\frac{f_d-\mu_d}{\sigma_d},\ d=1,\cdots,D$$

[0029] 其中，D为特征总维数，f_d为原始特征，μ_d和σ_d分别为同一类音效特征的均值和标准差；

[0030] 通过该公式将各维特征规整到均值为0，方差为1的分布中。

[0031] 进一步地，所述采用3阶段的自顶向下多层分割方法，将步骤1中提取出的音轨D_i依据声学特征分割成音频片段的具体过程如下：

[0032] 第1阶段：静音分割；

[0033] 以静音作为分割点对音轨进行粗略分割，得到各粗音频段；

[0034] 其中，所述静音的短时帧能量至少连续2s均小于能量门限E_{th}；

[0035] $$E_{th} = \lambda_s \cdot \min\left(\frac{1}{2}(E_{max} - E_{min}),\ E_{mean} - E_{min}\right) + E_{min}$$

[0036] 其中，E_{max}、E_{min}和E_{mean}分别代表当前音轨文档中短时帧能量的最大值、最小值和均值，λ_s为静音因子，$\lambda_s \in [0,1]$；

[0037] E_{range}表示能量的浮动范围，能量门限应当在E_{min}和$E_{min}+E_{range}$之间；

[0038] 第2阶段：距离分割；

[0039] 距离分割将经过静音分割后得到的各粗音频段，依据Hotelling's T^2距离再分割成无明显音频波动的音频片段；

[0040] 利用逐渐增长的第一分析窗依次对各粗音频段进行扫描，并在分析窗中每隔0.2s设置一个测试点，若第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离超过预设第一门限时，对应的测试点所在位置当作音频类型改变点，以音频类型改变点对粗音频段进行分割；

[0041] 第一分析窗初始长度为3s，如果窗内未发现音频类型改变点，则第一分析窗窗长增加1s，再次对粗音频段进行扫描；如果第一分析窗内找到音频类型改变点，则将第一分析窗长度重置为初始长度，并以得到的新的音频类型改变点作为起点继续搜索下一音频类型改变点直至搜索至粗音频段尾端；

[0042] 第3阶段：声学特征分割；

[0043] 根据音频特征的均值和方差，对无明显音频波动的音频片段进行分割；

[0044] 利用第二分析窗对各无明显音频波动的音频片段进行扫描，以第二分析窗的中点对第二分析窗内的音频片段进行分割得到左侧数据窗和右侧数据窗，计算中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差，其中，μ_1和μ_2分别是第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值；

[0045] 当欧式距离或者方差超过预设第二门限时，认为第二分析窗内部存在较大的数据变化，则当前中点为音效改变点，以音效改变点对对应的无明显音频波动的音频片段进行分割；

[0046] 否则，将左侧数据窗向后增加5帧，右侧数据窗向后平移5帧，继续计算左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差直到找到新的音效改变点或者搜索至当前无明显音频波动的音频片段的数据尾端；

[0047] 第二分析窗的长度初始为30帧。

[0048] 利用声学特征一阶、二阶统计量保证每个短时音频段内的声学特征数值处于较小的变化范围内；

[0049] 上述三个阶段是依次递进，由粗到细的过程，后面步骤的输入为上一步骤的输出；

[0050] 进一步地，所述静音因子λ_s设置为0.1。

[0051] 音频总量为210h时有最好的分割效果。

[0052] 进一步地，所述第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离采用以下公式计算：

[0053] $$T^2 = \frac{b(N-b)}{N}(S_1 - S_2)^T \sum^{-1} (S_1 - S_2)$$

[0054] 其中，N为第一分析窗总长度，Σ为协方差矩阵符号，b和S_1分别为第一分析窗测试点左侧数据窗长度和所包含的所有音频帧的超向量均值，S_2为右侧数据窗所包含的所有音频帧的超向量均值。

[0055] 进一步地，所述第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$采用以下公式计算获得：

[0056] $$\mathrm{dis}(\mu_1,\mu_2) = \sqrt{\sum_{d=1}^{D} |\mu_1(d) - \mu_2(d)|^2}$$

[0057] 其中，$\mu_1(d)$为左数据窗中所有帧音频的超向量中第d维的特征均值，$\mu_2(d)$为右数据窗中所有帧音频的超向量中第d维的特征均值，D为超向量中的特征维数。

[0058] 进一步地，所述依据音频字典中的音频字对每个音频片段标记时，寻找音频字典中音频字的超向量与音频片段中所有帧音频的超向量均值距离最小的音频字作为对应音频片段的标记音频字。

[0059] 进一步地，所述匹配度按照以下公式计算：

[0060] $$R(Q,C) = \frac{\sum_{n=1}^{N_q} hit(Q_n, C_n)}{N_q}$$

[0061] 其中，R(Q,C)表示待检测视频的音频字序列Q与选中的候选音频段落C的匹配度，$hit(Q_n,C_n)$表示待检测视频的音频字序列的第n个音频字与候选音频段落中的第n个音频字相同或者不同，$hit(Q_n,C_n) = \begin{cases} 1, & Q_n = C_n \\ 0, & Q_n \neq C_n \end{cases}$。

[0062] 进一步地，在对上传的待检测的视频进行音频检索前，先提取所上传的视频文件的MD5值，并将该值与所有上传的正版视频资源的MD5值进行比对，若与已上传的正版视频资源的MD5值相同，则判定当前上传的待检测视频属于盗版视频，结束当前上传的待检测的视频的检测流程。

[0063] 利用文件的MD5值进行判断，可加速检测效率。

[0064] 一种云存储环境下基于音频字的盗版视频检测系统，采用上述的一种云存储环境下基于音频字的盗版视频检测方法，包括：

[0065] 正版资源上传模块，用于上传正版视频资源；

[0066] 正版资源音频字文档倒排模块，获取正版视频资源，利用音频特征提取模块提取各正版视频资源中包含的音频特征，再依据音频字字典将音频特征转换为音频字，记录所

有正版视频资源中音频字出现在各音轨中的时刻位置，形成各正版视频中所有音频字出现在各音轨中的时刻位置倒排索引表，得到音频字倒排索引表；

[0067] 版权数据库，用于存储各正版资源的音频字倒排索引表；

[0068] 音频特征提取模块，用于从音轨中提取各帧音频的超向量，所述超向量包括宏观声学特征、时域特征、频域特征以及倒谱特征；

[0069] 音频字字典构建模块，利用音频特征提取模块对常见音效样本集进行超向量提取，对提取的超向量进行聚类，获取常见音效对应的音频字，构建音频字字典；

[0070] 待检测资源上传模块，通过云存储提供商的客户端上传待检测视频资源至云服务端；

[0071] 检测模块，在云服务端将上传的待检测的视频资源利用音频特征提取模块提取各正版视频资源中包含的音频特征，再依据音频字字典将音频特征转换为音频字得到音频字序列，将待检测视频资源的音频字序列中包含的音频字按照顺序与所述版权数据库中的音频字倒排索引表中音频字进行匹配，若待检测的音频字序列中存在至少N个连续的音频字与某一正版视频中连续的音频字一一匹配，则选取对应正版视频中连续匹配的音频字的起始位置所在的候选音频段落C，计算各正版视频候选段落C与待检测视频的音频字序列的匹配度，若匹配度超过设定的匹配度阈值，则判定当前待检测的视频属于对应正版视频的盗版资源。

[0072] 有益效果

[0073] 本发明提供了一种云存储环境下基于音频字的盗版检测方法和系统，过多维聚类构建的音频字典，对版权资源分割后的短时音频段进行特征提取，音频段转换为可以表征音频内容的音频字序列后，建立倒排索引。检索时，将用户提交的查询转换为音频字后直接定位候选段落，并根据候选段落与查询的内容相似度与阈值的关系确定视频是否为盗版。具有以下三个特性：

[0074] (1)充分利用了音频特征在传统音视频媒体文件中的稳定性，以及静默片段等特征，检测结果高效而准确，大幅提高本方案的商用价值；

[0075] (2)音频特征描述相较于视频描述能更有效的减少资源音频字典数据库的存储空间和I/O开销；

[0076] (3)基于倒排索引的音频匹配算法可以在音频字典中以常数级时间完成检索，满足实时性的要求。

[0077] 本发明通过上述三种特性以极低的本地计算成本和较少的网络带宽，加快了检索速度，提高了检索准确率，较容易的为正版资源提供了一种版权鉴定保护方案，在上传阶段对视频进行检测，从源头遏制了盗版的传播，在保证用户无感知的前提下，具有较好的鲁棒性。

附图说明

[0078] 图1为本发明的流程图；

[0079] 图2为实验系统架构图；

[0080] 图3为算法设计图；

[0081] 图4为音频字序列示意图；

[0082] 图5为应用本发明对各种音频进行检索的准确率关系示意图。

具体实施方式

[0083] 下面将结合附图和实施例对本发明做进一步的说明。

[0084] 图1为发行方和盗版方在上传资源时应用本发明所述方法的流程图，过程如下：

[0085] 发行方是由版权平台授权的企业或个人，盗版方是未经认证的团体或个人；

[0086] 首先，发行方请求上传版权资源，得到批准后，按照本发明所述方法提取对应的音频字序列，并向版权数据库写入版权资源的音频字序列信息；

[0087] 其次，盗版方发出资源上传请求，使用云存储提供商的客户端上传资源；

[0088] 接着，客户端计算资源的MD5值并发送给版权数据库，如果该MD5值已经存在于版权数据库中，则返回上传失败，否则，在客户端按照本发明所述方法计算上传资源的音频字序列，并与版权资源的音频字序列采用倒排索引式查找与匹配，如果匹配成功，说明该资源存在版权，上传失败，并且把此盗版资源的MD5值写入到版权数据库中，否则上传成功。

[0089] 图2为应用本发明所述系统的整体架构示意图，包含四大主体：Issuer、Copyright Cloud、Client以及Pirate。

[0090] Issuer即为发行方，是由Copyright Cloud(版权平台)授权的企业或个人。Issuer可以向Copyright Cloud写入有关data(所发行资源的音频字序列)信息。

[0091] Copyright Cloud即为版权平台，是公正的第三方平台。存储有版权资源的音频字序列，以及盗版资源的MD5等数据，并且在接收到Client发送的info(MD5或音频字序列)后进行matching(匹配)，并把result返回给Client。

[0092] Client即为云存储服务提供商的客户端，或者称之为云盘客户端，是所有用户上传资源到云盘的唯一途径。Pirate请求上传资源后，在Client本地计算资源的MD5和音频字序列后，把info(MD5或音频字序列)发送给Copyright Cloud，并接收Copyright Cloud返回的result(匹配结果)。

[0093] Pirate即为未经认证的团体或个人，Pirate可以向Client发送request(资源上传请求)，Client会根据所上传的资源返回result(上传结果)。

[0094] Issuer在本地对资源进行利用本发明所述方法进行特征提取和音频分割，将音频数据切分成音频段，并且根据音频字典将其转换为Audio word(音频字序列)，并将其发送到Copyright Cloud。

[0095] Copyright Cloud接收到Issuer发来的信息之后，为该资源建立基于音频字序列的倒排文档，并将其插入到当前版权平台的倒排索引表中。

[0096] Pirate在Client对资源进行处理，由于针对某个资源的，所以可以计算Resource MD5(资源的MD5)，如果当前MD5存在于版权平台的数据库中，则可以直接认为当前资源为盗版资源。否则利用利用本发明所述方法进行特征提取和音频分割，将音频数据切分成音频段，并且根据音频字典将其转换为Audio word(音频字序列)，并将其发送到Copyright Cloud进行进一步匹配。

[0097] Copyright Cloud接收到Client发来的消息之后，检索系统对所有的音频字进行“命中”检测，进而识别是否为盗版资源，若为盗版资源则将该资源的Resource MD5写入到版权平台的数据库中，完成版权平台的更新。

[0098] 本发明所述检测方法的具体步骤如图3所示，具体如下：

[0099] 一种云存储环境下基于音频字的盗版视频检测方法，包括以下步骤：

[0100] 步骤1：提取音效，并进行音频字标记；

[0101] 提取各种视频中的音轨，从音轨中提取各种常见音效，对所提取的不同音效依次进行标号，获得每种音效的音频字；

[0102] 音轨就是视频的声音信息，与视频时长相同，有许多公开提取方法，本方案中使用的是FFmpeg开源程序提供的工具包来提取音轨。

[0103] 常见音效至少包括400种，比如语音、枪声、爆炸声、欢呼声、笑声、哽咽、小提琴声、汽笛声等；音效的区分依据为宏观声学特征，包括音调、音高、带宽；

[0104] 步骤2：提取各音频帧超向量；

[0105] 先提取每帧音频的多维特征，并对多维特征分别进行归一化处理，构建音频帧的超向量；

[0106] 所述每帧音频的多维特征包括宏观声学特征、时域特征、频域特征以及倒谱特征，所述音效超向量是指对音频多维特征分别进行归一化处理后得到的特征形成的一维向量；

[0107] 其中，宏观声学特征包括音调、音高、带宽，时域特征包括短时能量、短时平均幅度、短时平均过零率、短时自相关系数，频域特征包括谱能量、子带能量比、谱质心、谱带宽、谱密度，倒谱特征包括Delta倒谱系数、LPC倒谱系数、梅尔倒谱系数；

[0108] 所述归一化处理是指进行规整向量计算；

[0109] 规整向量f'_d计算公式为：

[0110] $$f_d^{'} = \frac{f_d - \mu_d}{\sigma_d},\ d = 1, \cdots, D$$

[0111] 其中，D为特征总维数，f_d为原始特征，μ_d和σ_d分别为同一类音效特征的均值和标准差；

[0112] 通过该公式将各维特征规整到均值为0，方差为1的分布中。

[0113] 步骤3：构建音频字典；

[0114] 依次对每一种音效收集100个样本，提取每个样本中所有音频帧的超向量均值，将每一种音效的所有样本的超向量均值使用k-均值聚类算法聚成3个类，每一个聚类中心作为一个新的音频字w_i，每一个新的音频字均进行唯一标记w_i，利用新的音频字生成音频字典为$W=\{w_1, w_2, \cdots, w_k\}$，$k=1200$；

[0115] 步骤4：音频分割：

[0116] 采用3阶段的自顶向下多层分割方法，将步骤1中提取出的音轨D_i依据声学特征分割成音频片段；

[0117] 音频片段中的声学特征变化程度较小；

[0118] 具体过程如下：

[0119] 第1阶段：静音分割；

[0120] 以静音作为分割点对音轨进行粗略分割，得到各粗音频段；

[0121] 其中，所述静音的短时帧能量至少连续2s均小于能量门限E_{th}；

[0122] $$E_{th} = \lambda_s \cdot \min\left(\frac{1}{2}(E_{max} - E_{min}),\ E_{mean} - E_{min}\right) + E_{min}$$

[0123] 其中，E_{max}、E_{min}和E_{mean}分别代表当前音轨文档中短时帧能量的最大值、最小值和均值，λ_s为静音因子，$\lambda_s \in [0,1]$，静音因子λ_s设置为0.1，音频总量为210h时有最好的分割效果。

[0124] E_{range}表示能量的浮动范围，能量门限应当在E_{min}和$E_{min}+E_{range}$之间；

[0125] 第2阶段：距离分割；

[0126] 距离分割将经过静音分割后得到的各粗音频段，依据Hotelling's T^2距离再分割成无明显音频波动的音频片段；

[0127] 利用逐渐增长的第一分析窗依次对各粗音频段进行扫描，并在分析窗中每隔0.2s设置一个测试点，若第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离超过预设第一门限时，对应的测试点所在位置当作音频类型改变点，以音频类型改变点对粗音频段进行分割；

[0128] 第一分析窗初始长度为3s，如果窗内未发现音频类型改变点，则第一分析窗窗长增加1s，再次对粗音频段进行扫描；如果第一分析窗内找到音频类型改变点，则将第一分析窗长度重置为初始长度，并以得到的新的音频类型改变点作为起点继续搜索下一音频类型改变点直至搜索至粗音频段尾端；

[0129] 所述第一分析窗内部测试点左右两边数据窗之间的Hotelling's T^2距离采用以下公式计算：

[0130] $$T^2 = \frac{b(N-b)}{N}(S_1 - S_2)^T \sum^{-1} (S_1 - S_2)$$

[0131] 其中，N为第一分析窗总长度，Σ为协方差矩阵符号，b和S_1分别为第一分析窗测试点左侧数据窗长度和所包含的所有音频帧的超向量均值，S_2为右侧数据窗所包含的所有音频帧的超向量均值。

[0132] 第3阶段：声学特征分割；

[0133] 根据音频特征的均值和方差，对无明显音频波动的音频片段进行分割；

[0134] 利用第二分析窗对各无明显音频波动的音频片段进行扫描，以第二分析窗的中点对第二分析窗内的音频片段进行分割得到左侧数据窗和右侧数据窗，计算中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差，其中，μ_1和μ_2分别是第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值；

[0135] 当欧式距离或者方差超过预设第二门限时，认为第二分析窗内部存在较大的数据变化，则当前中点为音效改变点，以音效改变点对对应的无明显音频波动的音频片段进行分割；

[0136] 否则，将左侧数据窗向后增加5帧，右侧数据窗向后平移5帧，继续计算左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$和方差直到找到新的音效改变点或者搜索至当前无明显音频波动的音频片段的数据尾端；

[0137] 第二分析窗的长度初始为30帧。

[0138] 所述第二分析窗的中点左右两侧数据窗内音频片段中每一帧音频的超向量均值之间的欧氏距离$dis(\mu_1,\mu_2)$采用以下公式计算获得：

[0139] $$\mathrm{dis}(\mu_1,\mu_2)=\sqrt{\sum_{d=1}^{D}|\mu_1(d)-\mu_2(d)|^2}$$

[0140] 其中，$\mu_1(d)$为左数据窗中所有帧音频的超向量中第d维的特征均值，$\mu_2(d)$为右数据窗中所有帧音频的超向量中第d维的特征均值，D为超向量中的特征维数。

[0141] 利用声学特征一阶、二阶统计量保证每个短时音频段内的声学特征数值处于较小的变化范围内；

[0142] 上述三个阶段是依次递进，由粗到细的过程，后面步骤的输入为上一步骤的输出；

[0143] 步骤5：音频字转换；

[0144] 计算每个音频片段中所有帧音频的超向量均值，并依据音频字典中的音频字对每个音频片段标记，得到每一个音轨对应的音频字序列$ID_i=\{id_j^i\}$，$1\leqslant j\leqslant N_i$，$id_{N_i}^i$表示第i个音轨中的第j个音频片段对应的音频字；$N_i$表示第i个音轨中包含的音频片段总数；

[0145] 所述依据音频字典中的音频字对每个音频片段标记时，寻找音频字典中音频字的超向量与音频片段中所有帧音频的超向量均值距离最小的音频字作为对应音频片段的标记音频字。

[0146] 步骤6：构建音频字在音轨中的时刻位置索引表；

[0147] 对所有上传的正版视频资源依次进行音轨提取、音频帧超向量提取、音频分割以及音频字转换，并将所有正版视频资源中音频字出现在音轨中的时刻位置进行记录，获得各正版视频中所有音频字出现在各音轨中的时刻位置倒排索引表；

[0148] 倒排文件记录的是音频字出现在第几个音轨的第几个位置上，使用倒排索引可以加速检索速度。

[0149] 图4中(i，j)表示音频文件i的第j个位置，w_k表示某音频字，阴影表示该音频字出现在该位置。

[0150] 表1为音频倒排索引表，对应的音频字序列示意图如图4所示。倒排索引表由两部分构成，即索引项列表和每个索引项自身的事件表，事件表中的每一项均是一个指针，指向了含有该索引项的内容在音频文件中的具体位置，每一个位置(a，b)中a代表文档编号，b代表该索引项在文档中的具体位置。

[0151] 表1为音频倒排索引表

[0152]

音频字	事件表
*w*1	(1,3),(1,8)
*w*2	(1,6),(1,10)
*w*3	(1,6),(1,10)
*w*4	(1,1),(1,7),(1,9)
*w*5	(1,2),(1,5)

[0153] 步骤7：音频检索；

[0154] 在对上传的待检测的视频进行音频检索前，先提取所上传的视频文件的MD5值，并将该值与所有上传的正版视频资源的MD5值进行比对，若与已上传的正版视频资源的MD5值相同，则判定当前上传的待检测视频属于盗版视频，结束当前上传的待检测的视频的检索

流程。

[0155] 利用文件的MD5值进行判断，可加速检测效率。

[0156] 将上传的待检测的视频进行音轨提取、音频帧超向量提取、音频分割以及音频字转换得到对应的音频字序列，将待检测视频的音频字序列中包含的音频字按照顺序与所述时刻位置索引表中音频字进行匹配，若待检测的音频字序列中存在至少N个连续的音频字与某一正版视频中连续的音频字一一匹配，则选取对应正版视频中连续匹配的音频字的起始位置所在的候选音频段落C，计算各正版视频候选段落C与待检测视频的音频字序列的匹配度，若匹配度超过设定的匹配度阈值，则判定当前待检测的视频属于对应正版视频的盗版资源；

[0157] 所述匹配度是指音频字的出现时间的吻合程度。

[0158] 所述匹配度按照以下公式计算：

[0159] $$R(Q,C)=\frac{\sum_{n=1}^{N_q} hit(Q_n,C_n)}{N_q}$$

[0160] 其中，R(Q,C)表示待检测视频的音频字序列Q与选中的候选音频段落C的匹配度，$hit(Q_n,C_n)$表示待检测视频的音频字序列的第n个音频字与候选音频段落中的第n个音频字相同或者不同，$hit(Q_n,C_n)=\begin{cases}1, & Q_n=C_n\\ 0, & Q_n\neq C_n\end{cases}$。

[0161] 表2中的实验数据来自互联网的137部电影及网剧，总时长200h，依据IMDb分类标准。

[0162] 表2为实验的各类视频数量分布

[0163]

	战争片	动作片	灾难片	纪录片	音乐剧	犯罪片
数量（部）	15	19	12	38	23	80
时长（h）	40.6	33.9	27.2	35.6	32.9	29.8

[0164] 表3分别以5min,10min,15min,30min作为视频长度进行实验，以验证不同长度的视频(同一类别)对平均音频字数目、平均音频字计算时长、平均检索用时的影响。从数据库音频中随机截取不同长度的音频段落作为查询，记录下该查询音频在数据库中的音频文档编号作为标签，该标签用来评价检索算法的性能。平均音频字数目是指每1s的音频生成的音频字序列数目。平均音频字计算时长是指每1min的音频生成的音频字序列所需的时间。平均检索用时是指每1min的音频检索所需的时间。可以看到平均音频字序列长度、平均检索用时、音频字计算时长三个系统性能指标与视频类别没有明显的关系。

[0165] 表3为视频时长与系统性能关系表

[0166]

时长（min）	平均音频字数目	平均音频字计算时长（s）	平均检索用时（s）
5	51	0.071	0.011
10	53	0.077	0.013
15	62	0.068	0.010
30	58	0.070	0.011

[0167] 表4使用时长为30分钟的不同类别(战争片、动作片、灾难片、记录片、音乐剧、犯罪片)的视频来检测视频类别对方案性能的影响。可以看到平均音频字序列长度、平均检索用

时、音频字计算时长三个系统性能指标与视频长度没有明显的关系。

[0168] 表4为视频类别与系统性能关系表

[0169]

视频类别	平均音频字数目	平均音频字计算时长（s）	平均检索用时（s）
战争片	59	0.071	0.011
动作片	65	0.073	0.011
灾难片	61	0.069	0.012
记录片	52	0.071	0.009
音乐剧	50	0.072	0.011
犯罪片	58	0.070	0.011

[0170] 图5为应用本发明对各种音频进行检索的准确率关系示意图，分别选取同一个视频的5s、10s、15s、20s、25s、30s音频片段，分别进行翻录、加入噪声、变速等操作，并分别测试检索准确率。应用本发明所述的检测系统返回与处理后的音频片段相似度超过0.91的音频字文档编号，如果本发明所述的检测系统返回的编号与原音频片段的标签一致，则认为检索成功，否则认为检索失败。系统可以准确的检索出音频与翻录音频，对于噪声音频和变速音频，当样本时间较长时也能获得较好的检索效果。如果能够分布采样计算音频字序列，最终的系统检索准确率可以稳定在95％左右。

[0171] 本文中所描述的具体实施例仅仅是对本发明精神作举例说明。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，但并不会偏离本发明的精神或者超越所附权利要求书所定义的范围。

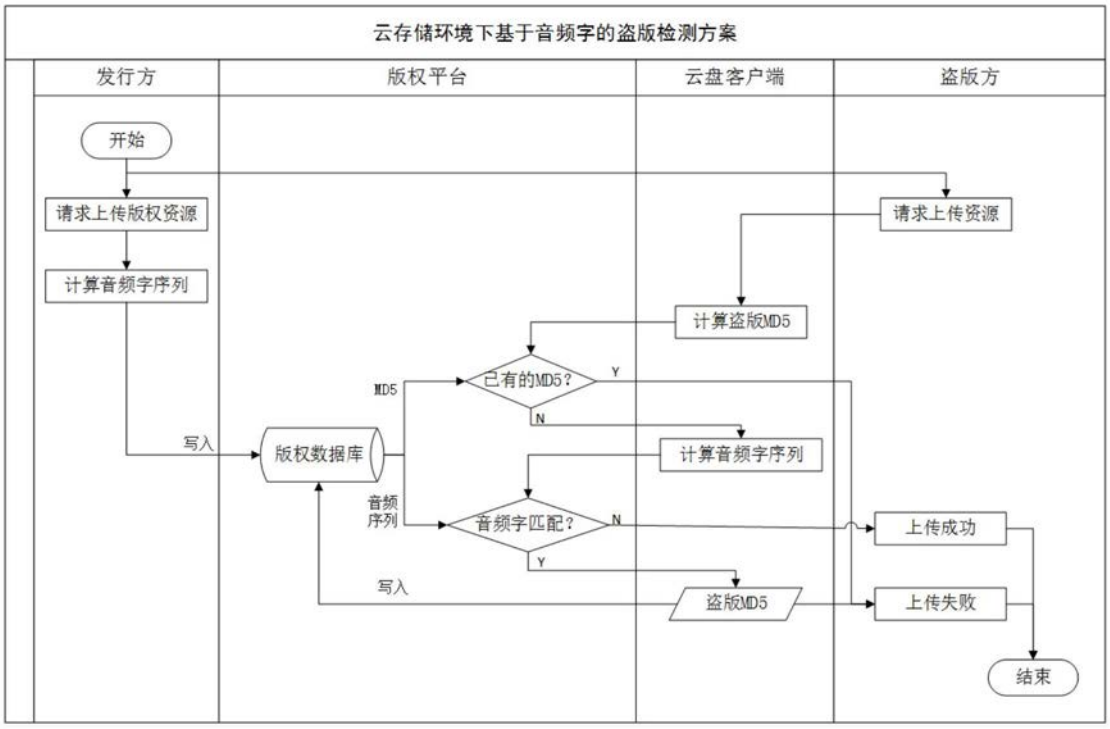

云存储环境下基于音频字的盗版检测方案
发行方
版权平台
云盘客户端
盗版方
开始
请求上传版权资源
计算音频字序列
写入
版权数据库
请求上传资源
计算盗版MD5
MD5
已有的MD5?
Y
N
计算音频字序列
音频序列
音频字匹配?
N
Y
上传成功
写入
盗版MD5
上传失败
结束

图1

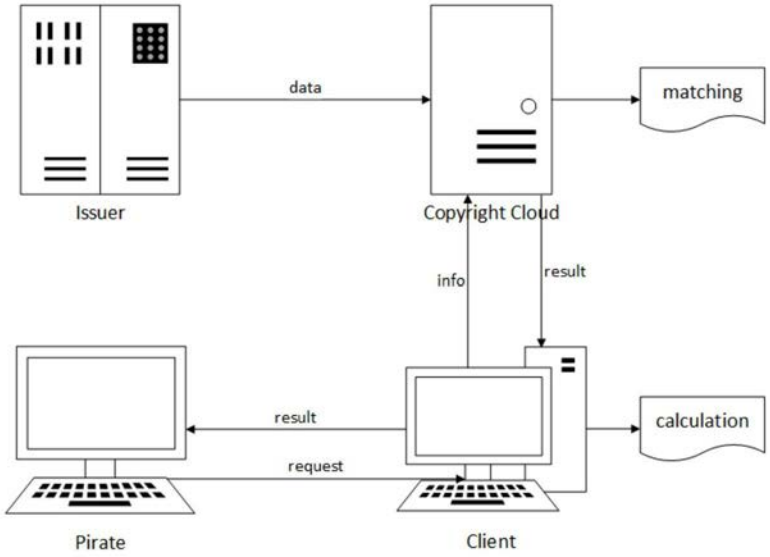

data
matching
Issuer
Copyright Cloud
info
result
result
calculation
request
Pirate
Client

图2

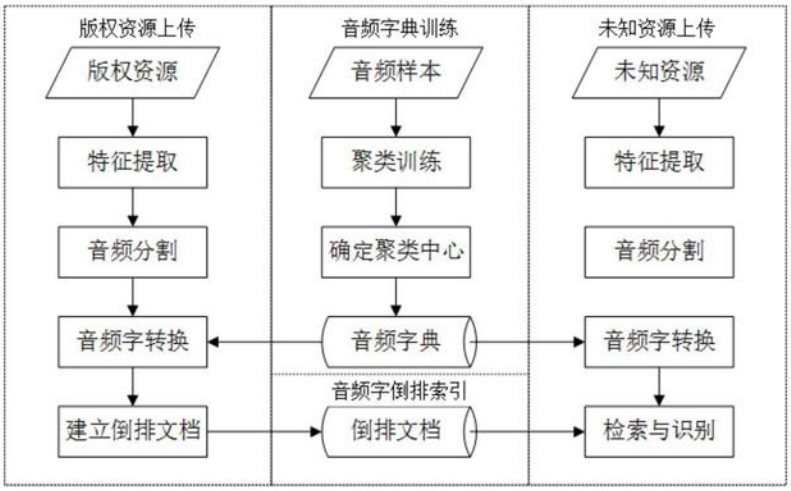

版权资源上传
音频字典训练
未知资源上传
版权资源
音频样本
未知资源
特征提取
聚类训练
特征提取
音频分割
确定聚类中心
音频分割
音频字转换
音频字典
音频字转换
音频字倒排索引
建立倒排文档
倒排文档
检索与识别

图3

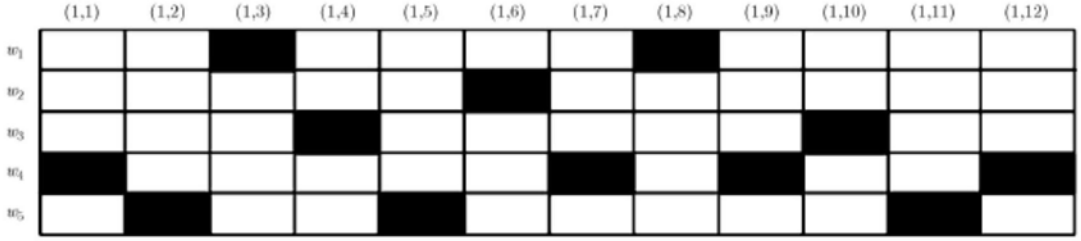

(1,1)
(1,2)
(1,3)
(1,4)
(1,5)
(1,6)
(1,7)
(1,8)
(1,9)
(1,10)
(1,11)
(1,12)

图4

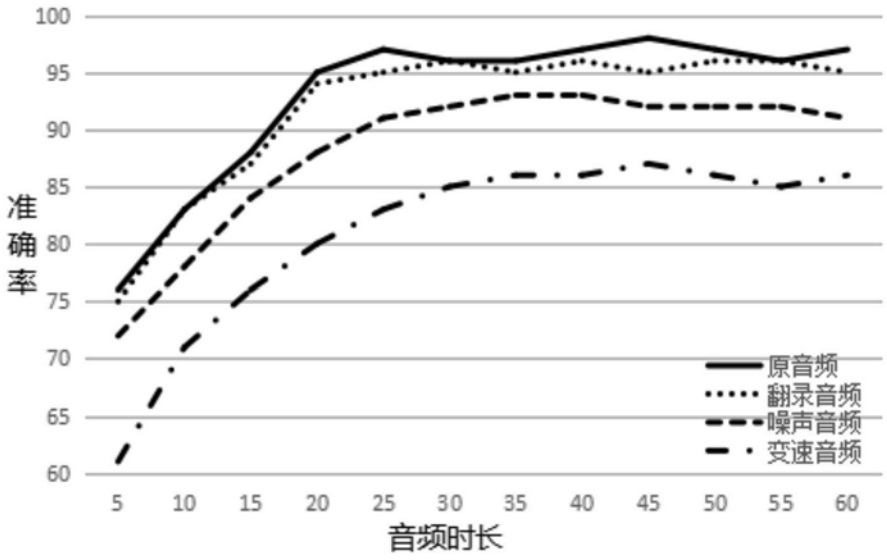

100
95
90
85
80
75
70
65
60
准确率
5
10
15
20
25
30
35
40
45
50
55
60
音频时长
原音频
翻录音频
噪声音频
变速音频

图5