(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 104598541 A
(43)申请公布日 2015.05.06

(21)申请号 201410849018.9

(22)申请日 2014.12.29

(71)申请人 乐视网信息技术(北京)股份有限公司
地址 100089 北京市海淀区学院南路68号19号楼六层6184号房间

(72)发明人 王晓萌 谭傅伦 许泽军 王英杰 袁斌

(74)专利代理机构 北京恒都律师事务所 11395
代理人 李向东

(51)Int.Cl.
G06F 17/30(2006.01)

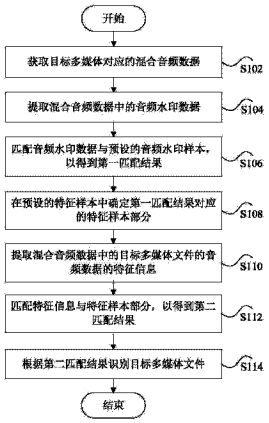
权利要求书3页 说明书19页 附图9页

(54)发明名称

多媒体文件的识别方法、装置

(57)摘要

本发明公开了一种多媒体文件的识别方法和装置。该多媒体文件的识别方法包括：获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据；提取混合音频数据中的音频水印数据；匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果；在预设的特征样本中确定第一匹配结果对应的特征样本部分；提取混合音频数据中的目标多媒体文件的音频数据的特征信息；匹配特征信息与特征样本部分，以得到第二匹配结果；根据第二匹配结果识别目标多媒体文件。通过本发明，能够提高视频识别的细度。



CN 104598541 A

1. 一种多媒体文件的识别方法，其特征在于，包括：

获取目标多媒体对应的混合音频数据，其中，所述混合音频数据包括所述目标多媒体文件的音频数据和音频水印数据；

提取所述混合音频数据中的音频水印数据；

匹配所述音频水印数据与预设的音频水印样本，以得到第一匹配结果；

在预设的特征样本中确定所述第一匹配结果对应的特征样本部分；

提取所述混合音频数据中的所述目标多媒体文件的音频数据的特征信息；

匹配所述特征信息与所述特征样本部分，以得到第二匹配结果；

根据所述第二匹配结果识别所述目标多媒体文件。

2. 根据权利要求 1 所述的多媒体文件的识别方法，其特征在于，所述混合音频数据还包括用户语音数据，所述方法还包括：

提取所述混合音频数据中的用户语音数据；

匹配所述用户语音数据与预设的语音样本，以得到第三匹配结果；以及

根据所述第三匹配结果从所述根据第二匹配结果识别得到的目标多媒体文件中选择一所述目标多媒体文件。

3. 根据权利要求 2 所述的多媒体文件的识别方法，其特征在于，

提取所述混合音频数据中的音频水印数据包括：提取所述混合音频数据中的高频部分的音频数据；

提取所述混合音频数据中的所述目标多媒体文件的音频数据的特征信息包括：提取所述混合音频数据中的低频部分的音频数据的特征信息；

提取所述混合音频数据中的用户语音数据包括：提取所述混合音频数据中的低频部分的音频数据；去除所述低频部分的音频数据中的所述目标多媒体文件的音频数据，以得到所述用户语音数据。

4. 根据权利要求 1 所述的多媒体文件的识别方法，其特征在于，提取所述混合音频数据中的所述目标多媒体文件的音频数据的特征信息包括：

提取所述混合音频数据中的低频部分的左声道数据和右声道数据；

采用以下公式合并所述左声道数据和所述右声道数据，以得到所述低频部分的立体声数据：$s = a*l+b*r$，其中，$a+b = 1$，s 为所述低频部分的立体声数据，l 为所述低频部分的左声道数据，r 为所述低频部分的右声道数据，a 和 b 为预设的参数；以及

提取所述立体声数据的时频特征数据得到所述目标多媒体文件的指纹信息，其中，所述指纹信息构成所述目标多媒体文件的音频数据的特征信息。

5. 根据权利要求 4 所述的多媒体文件的识别方法，其特征在于，若所述目标多媒体文件为第二多媒体文件的一个子多媒体文件，所述第一匹配结果为所述第二多媒体文件的标识信息，所述第二匹配结果为所述目标多媒体文件的标识信息，所述特征样本为预设的特征数据库中存储的至少一条多媒体记录，所述多媒体记录包括多媒体文件的指纹信息、与所述指纹信息对应的多媒体文件的标识信息，则：

在预设的特征样本中确定所述第一匹配结果对应的特征样本部分包括：在所述特征数据库中，定位到所述第二多媒体文件的标识信息对应的一条或多条多媒体记录；

匹配所述特征信息与所述特征样本部分，以得到第二匹配结果包括：匹配所述目标多

媒体的指纹信息与定位到的所述一条或多条多媒体记录，以确定所述目标多媒体的标识信息。

6. 根据权利要求 5 所述的多媒体文件的识别方法，其特征在于，所述低频部分的立体声数据为 N 个立体声数据，其中，所述 N 个立体声数据中的第 i 个立体声数据为 $s_i = a_i*l+b_i*r$，$a_i' + b_i' = 1$，$i = 1,2,3\cdots N$，则匹配所述目标多媒体文件的指纹信息与定位到的所述一条或多条多媒体记录，以确定所述目标多媒体文件的标识信息包括：

将每个立体声数据的时频特征数据分别与所述定位到的所述一条或多条多媒体记录进行匹配，得到所述立体声数据对应的多个匹配率；

根据所述多个匹配率中的最大值对应的一条所述多媒体记录确定所述目标多媒体文件的标识信息。

7. 一种多媒体文件的识别装置，其特征在于，包括：

获取模块，用于获取目标多媒体对应的混合音频数据，其中，所述混合音频数据包括所述目标多媒体文件的音频数据和音频水印数据；

第一提取模块，用于提取所述混合音频数据中的音频水印数据；

第一匹配模块，用于匹配所述音频水印数据与预设的音频水印样本，以得到第一匹配结果；

确定模块，用于在预设的特征样本中确定所述第一匹配结果对应的特征样本部分；

第二提取模块，用于提取所述混合音频数据中的所述目标多媒体文件的音频数据的特征信息；

第二匹配模块，用于匹配所述特征信息与所述特征样本部分，以得到第二匹配结果；

识别模块，用于根据所述第二匹配结果识别所述目标多媒体文件。

8. 根据权利要求 7 所述的多媒体文件的识别装置，其特征在于，所述混合音频数据还包括用户语音数据，所述装置还包括：

第三提取模块，用于提取所述混合音频数据中的用户语音数据；

第三匹配模块，用于匹配所述用户语音数据与预设的语音样本，以得到第三匹配结果；以及

验证模块，根据所述第三匹配结果从所述根据第二匹配结果识别得到的目标多媒体文件中选择一目标多媒体文件。

9. 根据权利要求 8 所述的多媒体文件的识别装置，其特征在于，

所述第一提取模块在提取音频水印数据时具体执行的步骤为：提取所述混合音频数据中的高频部分的音频数据；

所述第二提取模块在提取特征信息时具体执行的步骤为：提取所述混合音频数据中的低频部分的音频数据的特征信息；

所述第三提取模块在提取用户语音数据时具体执行的步骤为：提取所述混合音频数据中的低频部分的音频数据；去除所述低频部分的音频数据中的所述目标多媒体文件的音频数据，以得到所述用户语音数据。

10. 根据权利要求 7 所述的多媒体文件的识别装置，其特征在于，所述第二提取模块包括：

左右声道数据提取模块，用于提取所述混合音频数据中的低频部分的左声道数据和右

声道数据；

立体声数据合成模块，用于采用以下公式合并所述左声道数据和所述右声道数据，以得到所述低频部分的立体声数据：s = a*l+b*r，其中，a+b = 1，s 为所述低频部分的立体声数据，l 为所述低频部分的左声道数据，r 为所述低频部分的右声道数据，a 和 b 为预设的参数；以及

指纹信息提取模块，用于提取所述立体声数据的时频特征数据得到所述目标多媒体文件的指纹信息，其中，所述指纹信息构成所述目标多媒体文件的音频数据的特征信息。

11. 根据权利要求 10 所述的多媒体文件的识别装置，其特征在于，若所述目标多媒体文件为第二多媒体文件的一个子多媒体文件，所述第一匹配结果为所述第二多媒体文件的标识信息，所述第二匹配结果为所述目标多媒体文件的标识信息，所述特征样本为预设的特征数据库中存储的至少一条多媒体记录，所述多媒体记录包括多媒体文件的指纹信息、与所述指纹信息对应的多媒体文件的标识信息，则：

所述确定模块在确定所述特征样本部分时具体执行的步骤为：在所述特征数据库中，定位到所述第二多媒体文件的标识信息对应的一条或多条多媒体记录；

所述第二匹配模块在得到第二匹配结果时具体执行的步骤为：匹配所述目标多媒体的指纹信息与定位到的所述一条或多条多媒体记录，以确定所述目标多媒体的标识信息。

12. 根据权利要求 11 所述的多媒体文件的识别装置，其特征在于，所述低频部分的立体声数据为 N 个立体声数据，其中，所述 N 个立体声数据中的第 i 个立体声数据为 $s_i = a_i*l+b_i*r$，$a_i' + b_i' = 1$，$i = 1, 2, 3 \cdots N$，则所述第二匹配模块包括：

匹配率确定模块，用于将每个立体声数据的时频特征数据分别与所述定位到的所述一条或多条多媒体记录进行匹配，得到所述立体声数据对应的多个匹配率；

标识信息确定模块，用于根据所述多个匹配率中的最大值对应的一条所述多媒体记录确定所述目标多媒体文件的标识信息。

多媒体文件的识别方法、装置

技术领域

[0001] 本发明涉及多媒体文件识别技术领域，具体而言，特别涉及一种多媒体文件识别的方法、装置。

背景技术

[0002] 当前的视频搜索方式，通常使用的是视频的“关键字”搜索。这不但要求用户知晓该视频的相关信息，同时也要求搜索服务提供方能及时维护与视频一一对应的“关键字”数据库。而实际上，我们常常会遭遇到这样的尴尬：在大街小巷或者电视机前邂逅一段有趣的视频，但我们并不熟悉甚至不知道这段视频的信息，更别说通过“关键字”搜索到这段视频了。

[0003] 因而，基于声音识别视频便在这一实际需求的推动之下应运而生，它实现了由视频的声音识别视频本身。在基于声音识别视频的技术中，主要包括以下两种：基于音频水印的视频识别技术和基于音频指纹的视频识别技术。

[0004] 其中，在基于音频水印的视频识别技术中，常用的是基于声印码的视频识别技术，其原理在于：利用人耳对高频声音不敏感的特点，通过在音频数据的高频段中加入携带特定信息的声印码，识别终端在获取到这种携带了声印码的声音文件后，能从中提取它携带的声印码，将提取的声印码与数据库中的声印码样本匹配，从而实现了通过声音识别视频。其优点是识别速度快，一般为毫秒级。

[0005] 但是，该技术在区分视频时，仅依靠声印码数据来区分，因而无法区分添加相同声印码数据的视频，例如，当属于同一剧集的多集电视剧添加的声印码数据相同时，无法区分各集电视剧，从而在识别某集电视剧时，只能识别到该集电视剧属于某一剧集，而不能识别到该集电视剧具体为该剧集中的哪一集；当某电影添加的声印码数据相同时，无法区分该电影中的电影片段，从而在识别该电影中某一个片段时，只能识别到该电影片段属于某一电影，而不能识别到该电影片段具体为该电影中的哪一个片段，因此，这种基于声印码的视频识别技术的识别范围有限，识别细度低。

[0006] 针对现有技术中视频识别细度低的问题，目前尚未提出有效的解决方法。

发明内容

[0007] 本发明的主要目的在于提供一种多媒体文件识别的方法、装置，以解决现有技术中视频识别细度低的问题。

[0008] 依据本发明的一个方面，提供了一种多媒体文件的识别方法。

[0009] 根据本发明的多媒体文件的识别方法包括：获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据；提取混合音频数据中的音频水印数据；匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果；在预设的特征样本中确定第一匹配结果对应的特征样本部分；提取混合音频数据中的目标多媒体文件的音频数据的特征信息；匹配特征信息与特征样本部分，以得到第二匹配结果；

根据第二匹配结果识别目标多媒体文件。

[0010] 进一步地，混合音频数据还包括用户语音数据，该方法还包括：提取混合音频数据中的用户语音数据；匹配用户语音数据与预设的语音样本，以得到第三匹配结果；以及根据所述第三匹配结果从所述根据第二匹配结果识别得到的目标多媒体文件中选择一所述目标多媒体文件。

[0011] 进一步地，提取混合音频数据中的音频水印数据包括：提取混合音频数据中的高频部分的音频数据；提取混合音频数据中的目标多媒体文件的音频数据的特征信息包括：提取混合音频数据中的低频部分的音频数据的特征信息；提取混合音频数据中的用户语音数据包括：提取混合音频数据中的低频部分的音频数据；去除低频部分的音频数据中的目标多媒体文件的音频数据，以得到用户语音数据。

[0012] 进一步地，提取混合音频数据中的目标多媒体文件的音频数据的特征信息包括：提取混合音频数据中的低频部分的左声道数据和右声道数据；采用以下公式合并左声道数据和右声道数据，以得到低频部分的立体声数据：s = a*l+b*r，其中，a+b = 1，s 为低频部分的立体声数据，l 为低频部分的左声道数据，r 为低频部分的右声道数据，a 和 b 为预设的参数；以及提取立体声数据的时频特征数据得到目标多媒体文件的指纹信息，其中，指纹信息构成目标多媒体文件的音频数据的特征信息。

[0013] 进一步地，若目标多媒体文件为第二多媒体文件的一个子多媒体文件，第一匹配结果为第二多媒体文件的标识信息，第二匹配结果为目标多媒体文件的标识信息，特征样本为预设的特征数据库中存储的至少一条多媒体记录，多媒体记录包括多媒体文件的指纹信息、与指纹信息对应的多媒体文件的标识信息，则：在预设的特征样本中确定第一匹配结果对应的特征样本部分包括：在特征数据库中，定位到第二多媒体文件的标识信息对应的一条或多条多媒体记录；匹配特征信息与特征样本部分，以得到第二匹配结果包括：匹配目标多媒体的指纹信息与定位到的一条或多条多媒体记录，以确定目标多媒体的标识信息。

[0014] 进一步地，低频部分的立体声数据为 N 个立体声数据，其中，N 个立体声数据中的第 i 个立体声数据为 $s_i = a_i*l+b_i*r$，$a_i' + b_i' = 1$，$i = 1, 2, 3 \cdots N$，则匹配目标多媒体文件的指纹信息与定位到的一条或多条多媒体记录，以确定目标多媒体文件的标识信息包括：将每个立体声数据的时频特征数据分别与定位到的一条或多条多媒体记录进行匹配，得到立体声数据对应的多个匹配率；根据多个匹配率中的最大值对应的一条多媒体记录确定目标多媒体文件的标识信息。

[0015] 依据本发明的另一个方面，提供了一种多媒体文件的识别装置。

[0016] 根据本发明的多媒体文件的识别装置包括：获取模块，用于获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据；第一提取模块，用于提取混合音频数据中的音频水印数据；第一匹配模块，用于匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果；确定模块，用于在预设的特征样本中确定第一匹配结果对应的特征样本部分；第二提取模块，用于提取混合音频数据中的目标多媒体文件的音频数据的特征信息；第二匹配模块，用于匹配特征信息与特征样本部分，以得到第二匹配结果；识别模块，用于根据第二匹配结果识别目标多媒体文件。

[0017] 进一步地，混合音频数据还包括用户语音数据，该装置还包括：第三提取模块，用

于提取混合音频数据中的用户语音数据；第三匹配模块，用于匹配用户语音数据与预设的语音样本，以得到第三匹配结果；以及验证模块，用于根据所述第三匹配结果从所述根据第二匹配结果识别得到的目标多媒体文件中选择一所述目标多媒体文件。

[0018] 进一步地，第一提取模块在提取音频水印数据时具体执行的步骤为：提取混合音频数据中的高频部分的音频数据；第二提取模块在提取特征信息时具体执行的步骤为：提取混合音频数据中的低频部分的音频数据的特征信息；第三提取模块在提取用户语音数据时具体执行的步骤为：提取混合音频数据中的低频部分的音频数据；去除低频部分的音频数据中的目标多媒体文件的音频数据，以得到用户语音数据。

[0019] 进一步地，第二提取模块包括：左右声道数据提取模块，用于提取混合音频数据中的低频部分的左声道数据和右声道数据；立体声数据合成模块，用于采用以下公式合并左声道数据和右声道数据，以得到低频部分的立体声数据：s = a*l+b*r，其中，a+b = 1，s 为低频部分的立体声数据，l 为低频部分的左声道数据，r 为低频部分的右声道数据，a 和 b 为预设的参数；以及指纹信息提取模块，用于提取立体声数据的时频特征数据得到目标多媒体文件的指纹信息，其中，指纹信息构成目标多媒体文件的音频数据的特征信息。

[0020] 进一步地，若目标多媒体文件为第二多媒体文件的一个子多媒体文件，第一匹配结果为第二多媒体文件的标识信息，第二匹配结果为目标多媒体文件的标识信息，特征样本为预设的特征数据库中存储的至少一条多媒体记录，多媒体记录包括多媒体文件的指纹信息、与指纹信息对应的多媒体文件的标识信息，则：确定模块在确定特征样本部分时具体执行的步骤为：在特征数据库中，定位到第二多媒体文件的标识信息对应的一条或多条多媒体记录；第二匹配模块在得到第二匹配结果时具体执行的步骤为：匹配目标多媒体的指纹信息与定位到的一条或多条多媒体记录，以确定目标多媒体的标识信息。

[0021] 进一步地，低频部分的立体声数据为 N 个立体声数据，其中，N 个立体声数据中的第 i 个立体声数据为 $s_i = a_i*l+b_i*r$，$a_i' + b_i' = 1$，$i = 1, 2, 3\cdots N$，则第二匹配模块包括：匹配率确定模块，用于将每个立体声数据的时频特征数据分别与定位到的一条或多条多媒体记录进行匹配，得到立体声数据对应的多个匹配率；标识信息确定模块，用于根据多个匹配率中的最大值对应的一条多媒体记录确定目标多媒体文件的标识信息。

[0022] 通过本发明，在进行多媒体文件的识别时，将目标多媒体文件的音频水印数据与预设的音频水印样本进行匹配，得到第一匹配结果，在预设的特征样本中根据第一匹配结果进行筛选，筛选到第一匹配结果对应的特征样本部分，实现目标多媒体文件的初步识别，缩小识别的范围；以此为基础，再将该目标多媒体文件的音频数据的特征信息与上述特征样本部分进行匹配，能够得到第二匹配结果，也即在上述缩小的识别范围中进行进一步识别，最后根据第二匹配结果识别目标多媒体文件，若采用该方法识别视频时，能够解决现有技术中视频识别细度低的问题。

[0023] 上述说明仅是本发明技术方案的概述，为了能够更清楚了解本发明的技术手段，而可依照说明书的内容予以实施，并且为了让本发明的上述和其它目的、特征和优点能够更明显易懂，以下特举本发明的具体实施方式。

附图说明

[0024] 通过阅读下文优选实施方式的详细描述，各种其他的优点和益处对于本领域普通

技术人员将变得清楚明了。附图仅用于示出优选实施方式的目的，而并不认为是对本发明的限制。而且在整个附图中，用相同的参考符号表示相同的部件。在附图中：

[0025] 图 1 是根据本发明实施例一的方法流程图；

[0026] 图 2 是根据本发明实施例二的方法流程图；

[0027] 图 3 是根据本发明实施例三的方法流程图；

[0028] 图 4 是根据本发明实施例四的方法流程图；

[0029] 图 5 是根据本发明实施例五的方法流程图；

[0030] 图 6 是根据本发明实施例六的方法流程图；

[0031] 图 7 是根据本发明实施例七的终端录音模块框图；

[0032] 图 8 是根据本发明实施例七的终端音频识别模块框图；

[0033] 图 9 是根据本发明实施例七的服务器和数据库示意图；

[0034] 图 10 是根据本发明实施例七的方法流程图；

[0035] 图 11 是根据本发明实施例八的装置框图。

具体实施方式

[0036] 下面结合附图和具体实施方式对本发明做进一步说明。需要指出的是，在不冲突的情况下，本申请中的实施例及实施例中的特征可以相互组合。

[0037] 本发明实施例提供了多媒体文件识别的方法，在该方法中，首先利用目标多媒体文件的音频水印数据与预设的音频水印样本进行匹配，得到第一匹配结果；然后在预设的特征样本中得到第一匹配结果对应的特征样本部分；再将该目标多媒体文件的音频数据的特征信息与上述特征样本部分进行匹配，能够得到第二匹配结果，最后根据第二匹配结果识别目标多媒体文件。

[0038] 从中可以看出，在该方法中，首先利用基于音频水印的视频识别技术处理速度快的优点，基于目标多媒体文件的音频水印数据进行识别，能够快速的得到一个初步的识别结果；然后利用基于音频指纹的视频识别技术能够识别任何来源的音频数据的优点，在上述初步的识别结果基础上进一步基于目标多媒体文件的音频数据的特征信息进行识别，有效地结合了基于音频水印的视频识别技术和基于音频指纹的视频识别技术，提升了识别的细度。

[0039] 以下将对本发明所提供的多种实施例进行详细的描述。

[0040] 实施例一

[0041] 该实施例一提供了一种多媒体文件的识别方法的实施例，在该实施例的方法中，在目标多媒体文件的音频数据中预先添加音频水印数据，识别目标多媒体文件时，将会利用到预设的音频水印样本和预设的特征样本，具体地，如图 1 所示，该方法包括如下的步骤 S102 至步骤 S114。

[0042] 步骤 S102：获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据。

[0043] 该目标多媒体文件可以为视频（或音频），在该视频（或音频）播放的过程中，识别装置启动录音设备进行声音录制，从而得到该视频（或音频）的混合音频数据，由于该视频（或音频）中预先添加了音频水印数据，因而录制到的混合音频数据中既包括目标多媒

体文件的音频数据本身，也包括添加进去的音频水印数据。

[0044] 该识别装置可以为智能移动通信终端，例如手机、PAD；也可以为计算机；或者作为独立的识别单元嵌入到需要进行多媒体文件识别的装置中。

[0045] 步骤 S104：提取混合音频数据中的音频水印数据。

[0046] 根据音频水印数据的特点，将混合音频数据中的音频水印数据提取出来，例如，音频水印数据为声印码数据，由于声印码数据是在音频数据的高频段中加入的特定信息，因此，通过提取混合音频数据中高频部分的数据，即可得到声印码数据。

[0047] 步骤 S106：匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果。

[0048] 该预设的音频水印样本可以存储在位于识别装置本地的音频水印数据库中，从而识别装置完成音频水印数据与预设的音频水印样本的匹配，也可存储在识别装置远端的、位于音频水印识别服务器的音频水印数据库中，识别装置与音频水印识别服务器进行交互，将音频水印数据传输至音频水印识别服务器，由音频水印识别服务器完成音频水印数据与预设的音频水印样本的匹配，无论谁来执行该匹配步骤，均会得到第一匹配结果。

[0049] 该第一匹配结果可以为由多个多媒体文件构成的一个多媒体文件组，也即，通过该步骤，确定了目标多媒体文件在该多媒体文件组中。

[0050] 例如，预先将相同演员主演的所有的电影都添加相同的音频水印数据，并且预设的音频水印样本为存储在音频水印数据库中的多条音频水印记录，每条音频水印记录由音频水印数据、和与音频水印数据对应的电影主演名字构成。如果目标多媒体文件为某演员主演的电影，在匹配目标多媒体文件的音频水印数据与预设的音频水印样本时，必然能够根据目标多媒体文件的音频水印数据在该音频水印数据库中定位到一条音频水印记录，得到该电影的主演名字，也即，通过音频水印数据识别到该目标多媒体文件是某演员 A 主演的电影。

[0051] 步骤 S108：在预设的特征样本中确定第一匹配结果对应的特征样本部分。

[0052] 该预设的特征样本可以存储在位于识别装置本地的音频指纹数据库中，从而识别装置在本地完成特征样本部分的确定；也可存储在识别装置远端的、位于音频指纹识别服务器的音频指纹数据库中，识别装置与音频指纹识别服务器进行交互，将第一匹配结果传输至音频指纹识别服务器，由音频指纹识别服务器完成特征样本部分的确定，无论谁来执行该确定步骤，均会从预设的特征样本中筛选出一部分，该部分是第一匹配结果对应的特征样本部分。

[0053] 通过该步骤，在进行音频指纹匹配时，无需将目标多媒体文件的音频数据的特征信息与预设的特征样本整体进行匹配，只需将特征信息与第一匹配结果对应的特征样本部分进行匹配。

[0054] 例如，预设的特征样本为存储在音频指纹数据库中的多条多媒体记录，每条多媒体记录由多媒体文件的指纹信息、与指纹信息对应的电影名称和电影主演姓名构成。在通过步骤 S106 得到的第一匹配结果为：该目标多媒体是某演员 A 主演的电影，则在该步骤中，根据该第一匹配结果，在音频指纹数据库中，能够定位到该演员 A 对应的一条或多条多媒体记录。

[0055] 步骤 S110：提取混合音频数据中的目标多媒体文件的音频数据的特征信息。

[0056] 例如，音频水印数据为声印码数据时，由于声印码数据是在音频数据的高频段中

加入的特定信息，因此，通过提取混合音频数据中低频部分的数据，能够得到混合音频数据中的目标多媒体文件的音频数据。

[0057] 进一步地，提取该低频部分的数据的特征信息得到目标多媒体文件的音频数据的特征信息。具体可采用现有技术中任意的音频数据的特征提取方法，例如可提取音频的时域特征数据，具体如提取音频片段的幅值，也可提取音频的时频特征数据。

[0058] 步骤 S112 ：匹配特征信息与特征样本部分，以得到第二匹配结果。

[0059] 当预设的特征样本存储在位于识别装置本地的音频指纹数据库中时，由识别装置在本地完成特征信息与特征样本部分的匹配；当预设的特征样本存储在识别装置远端的、位于音频指纹识别服务器的音频指纹数据库中，由音频指纹识别服务器完成特征信息与特征样本部分的匹配，无论谁来执行该匹配步骤，均会得到与特征信息相匹配的结果。

[0060] 例如，在该音频指纹数据库中，特征样本部分为演员 A 对应的一条或多条多媒体记录，则将特征信息与该一条或多条多媒体记录的指纹信息分别进行匹配，以得到与该特征信息匹配成功的一个指纹信息，从而该指纹信息所在的多媒体记录中的电影名称即为目标多媒体文件的电影名称。

[0061] 步骤 S114 ：根据第二匹配结果识别目标多媒体文件。

[0062] 采用该实施例提供的多媒体文件的识别方法，首先基于目标多媒体文件的音频水印数据进行识别，利用了基于音频水印的视频识别技术处理速度快的优点，能够快速的得到一个初步的识别结果，该识别结果的细度可能较低，例如只是多媒体文件一个隶属范围，然而得到该范围的速度快，也即该方法首先快速的确定了目标多媒体文件所处的一个范围；在该范围中，进一步基于目标多媒体文件的音频数据的特征信息进行识别，利用了基于音频指纹的视频识别技术能够识别任何来源的音频数据的优点，能够识别到该目标多媒体文件。因而该方法有效地结合了基于音频水印的视频识别技术和基于音频指纹的视频识别技术，与单纯使用基于音频水印的视频识别技术的方法相比，提升了识别的细度和视频识别的应用范围；与单纯使用基于音频指纹的视频识别技术相比，缩短了识别的时间。

[0063] 实施例二

[0064] 该实施例二提供了一种多媒体文件的识别方法的实施例，该实施例为在实施例一的基础上的一个优选实施例。在该实施例的方法中，目标多媒体文件为目标视频；在该目标视频的音频数据中预先添加声印码数据；在获取目标视频的音频数据的同时，获取到了用户的语音数据；该识别方法能够在实施例一识别到目标多媒体文件的基础上，进一步根据用户的语音数据验证识别结果的准确性；在识别目标视频时，将会利用到预设的声印码样本和预设的特征样本；在根据用户的语音数据验证识别结果的准确性时，将会利用到预设的语音样本，具体地，如图 2 所示，该方法包括如下的步骤 S202 至步骤 S212。

[0065] 步骤 S202 ：获取目标视频对应的混合音频数据，其中，混合音频数据包括目标视频的音频数据、声印码数据和用户语音数据。

[0066] 用户在观看视频的过程中，可能对待识别视频中出现的场景、演员、甚至是物品等视频中的某具体细节内容熟悉，用户可在该视频播放的过程中通过语音输入其熟悉的内容。例如，在目标视频的播放过程中，识别装置启动录音设备进行声音录制，录制到当前环境中的所有声音信息，也即目标视频对应的混合音频数据，该混合音频数据中包括目标视频的音频数据本身，也包括添加进去的音频水印数据，还包括用户发出的用户语音数据。

[0067] 步骤 S204 :提取混合音频数据中的音频水印数据,并与预设的音频水印样本,以得到第一匹配结果。

[0068] 该步骤与实施例一中的步骤 S104 和步骤 S106 相同,此处不再赘述。

[0069] 步骤 S206 :在预设的特征样本中确定第一匹配结果对应的特征样本部分,提取混合音频数据中的目标视频的音频数据的特征信息,并将特征信息与特征样本部分进行匹配,以得到第二匹配结果,并根据第二匹配结果识别目标视频。

[0070] 该步骤与实施例一中的步骤 S108 至步骤 S114 相同,此处不再赘述。

[0071] 步骤 S208 :提取混合音频数据中的用户语音数据。

[0072] 根据声印码数据的特点,通过提取混合音频数据中低频部分的数据,可将声印码数据从混合音频数据中去除,再将去除声印码数据的混合音频数据中的目标视频的音频数据去除,也即将低频部分的数据中的目标视频的音频数据去除,即可得到用户语音数据。

[0073] 具体地,在去除低频部分的数据中的目标视频的音频数据时,需要获取到目标视频的音频数据。由于在步骤 S206 中已识别到目标视频,在该步骤中,利用步骤 S206 识别到的目标视频对应的音频数据来实现该步骤中用户语音数据提取,具体地,例如,第二匹配结果包括目标视频的 URL 信息,根据该 URL 信息可获取到目标视频的音频数据,然后用低频部分的数据减去获取到的目标视频的音频数据,即可得到用户语音数据。

[0074] 步骤 S210 :匹配用户语音数据与预设的语音样本,以得到第三匹配结果。

[0075] 该预设的语音样本可以存储在位于识别装置本地的语音数据库中,从而识别装置完成用户语音数据与预设的语音样本的匹配,也可存储在识别装置远端的、位于语音识别服务器的语音数据库中,识别装置与语音识别服务器进行交互,将用户语音数据传输至语音识别服务器,由语音识别服务器完成用户语音数据与预设的语音样本的匹配,无论谁来执行该匹配步骤,均会得到第三匹配结果。

[0076] 例如,预设的语音样本为存储在语音数据库中的多条语音记录,每条语音记录由语音特征信息、和与语音特征信息对应的关键字构成。在匹配用户语音数据与预设的语音样本时,首先根据用户语音数据提取到用户语音的语音特征信息,然后将该提取到用户语音的语音特征信息与语音数据库中的语音特征信息进行匹配,从而能够在该语音数据库中定位到一条或多条语音记录,进而得到该用户语音数据对应的关键字。具体如,得到用户语音数据对应的关键字为“拍摄地点”和“海南”。

[0077] 步骤 S212 :根据第三匹配结果和第二匹配结果识别得到的目标多媒体文件。由于通过第二匹配结果识别到了目标视频,但是,可能第二匹配结果给出的是多个目标视频,通过第三匹配结果识别到了用户输入的关键字,通过用户输入的关键字能够进一步确定识别的目标视频。

[0078] 具体如,根据第二匹配结果识别到了目标视频为某电影 B 和电影 C,第三匹配结果识别到的用户语音数据对应的关键字为“拍摄地点”和“海南”,而电影 B 拍摄地为“海南”,电影 C 的拍摄地为“北京”,则在该步骤中,可通过判断该电影 B 中的拍摄地点是否为海南来选择识别输出电影 B 还是电影 C。

[0079] 再如,根据第二匹配结果识别到了目标视频为某电视剧的第 7、8、10 集,第三匹配结果识别到的用户语音数据对应的关键字为“演员”和“刘若英”,而第 7、8 集中没有刘若英,则在该步骤中,可通过判断第三识别结果判断目标视频文件为电视剧的第 10 集。

[0080] 总而言之，可根据识别到的关键字来进一步识别目标视频，验证目标视频的识别结果的准确性，从而能够向用户提供高准确率的识别结果。

[0081] 采用该优选实施例提供的多媒体文件的识别方法，在上述实施例一的技术效果的基础上，能够结合用户语音数据来提高识别目标多媒体文件的准确性。

[0082] 实施例三

[0083] 该实施例三提供了一种多媒体文件的识别方法的实施例，该实施例为在实施例一的基础上的另一个优选实施例。在该实施例的方法中，在目标多媒体文件的音频数据中预先添加音频水印数据，识别目标多媒体文件时，将会利用到预设的音频水印样本和预设的特征样本，具体地，如图 3 所示，该方法包括如下的步骤 S302 至步骤 S320。

[0084] 步骤 S302 ：获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据。

[0085] 步骤 S304 ：提取混合音频数据中的音频水印数据。

[0086] 步骤 S306 ：匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果。

[0087] 上述的步骤 S302 至步骤 S306 分别与实施例一中的步骤 S102 至步骤 S106 分别一一对应，此处不再赘述。

[0088] 步骤 S308 ：在预设的特征数据库中定位到第一匹配结果对应的多条多媒体文件记录。

[0089] 该预设的特征数据库中存储有至少一条多媒体记录，该特征数据库中存储的多媒体记录构成预设的特征样本，每条多媒体记录包括多媒体文件的指纹信息、与指纹信息对应的多媒体文件的标识信息，其中，每条多媒体记录中的指纹信息由计算多媒体的音频数据的时频特征数据得到的多个指纹值构成。

[0090] 例如，预先将某电视频道播出的所有电视节目都添加相同的音频水印数据，并且预设的音频水印样本为存储在音频水印数据库中的多条音频水印记录，每条音频水印记录由音频水印数据、和与音频水印数据对应的电视频道名称构成。如果目标多媒体文件为某电视频道播出的电视节目，在匹配目标多媒体文件的音频水印数据与预设的音频水印样本时，必然能够根据目标多媒体文件的音频水印数据在该音频水印数据库中定位到一条音频水印记录，得到该电视节目的电视频道，也即，通过音频水印数据识别到该电视节目是电视频道 A 播出的节目。

[0091] 每条多媒体记录中与指纹信息相对应的多媒体文件标识信息可为多媒体文件的电视频道名称和电视节目名称，在第一匹配结果确定该目标多媒体文件是电视频道 A 播出的节目后，则通过该步骤，能够在该特征数据库中，定位到该电视频道 A 对应的多条多媒体记录，该多条多媒体记录为该电视频道 A 播出的多个电视节目。

[0092] 步骤 S310 ：提取混合音频数据中的低频部分的左声道数据和右声道数据。

[0093] 通过提取混合音频数据中低频部分的数据，能够得到混合音频数据中的目标多媒体文件的音频数据，该音频数据由左声道数据和右声道数据两部分数据组成。

[0094] 步骤 S312 ：合并左声道数据和右声道数据，以得到低频部分的 N 个立体声数据。

[0095] 具体地，采用以下公式进行合并 ：

[0096] $s_i = a_i * l + b_i * r$

[0097] 其中，$a_i' + b_i' = 1$，$i = 1, 2, 3 \cdots N$，s_1为第一个立体声数据，s_N为第 N 个立体声

数据，s_i为第 i 个立体声数据，a_i' 和 b_i' 为预设的权重参数，调节 a_i' 和 b_i' 的大小，可实现调节左右声道数据在立体声数据中所占的比重。

[0098] 步骤 S314：计算每个立体声数据的时频特征数据，得到每个立体声数据的多个指纹值。

[0099] 对各个立体声数据而言，每个立体声数据的多个指纹值构成自身的指纹信息，这 N 个立体声数据的指纹信息构成目标多媒体文件的指纹信息。

[0100] 具体地，对某个立体声数据而言，计算该立体声数据的时频特征数据得到该立体声数据的多个指纹值时，包括以下的步骤 S3142 至步骤 S3148：

[0101] 步骤 S3142：对该立体声数据进行短时傅里叶变换，以得到该立体声数据的时频分布图；

[0102] 步骤 S3144：获取时频分布图中的能量极大值点；

[0103] 步骤 S3146：根据两个不同时刻的极大值点 A[ta,fa,Va]、B[tb,fb,Vb] 构建一个指纹值为 fp[ta,fa,fb,tb-ta]，并转换为哈希码 fp[hashData，ta]，其中，ta 为极值大点 A 所处的时刻，fa 为极值大点 A 所处的频率，Va 为极值大点 A 的能量，tb 为极值大点 B 所处的时刻，fb 为极值大点 B 所处的频率，Vb 为极值大点 B 的能量，ta<tb，极大值点 A 和极值大点 B 为时频分布图中任意两个相邻的能量极大值点；

[0104] 步骤 S3148：将构建的所有指纹值按照时间顺序组合得到该立体声数据的多个指纹值。

[0105] 相应地，在特征数据库中，针对每条多媒体记录中的指纹信息，在计算多媒体的音频数据的时频特征数据得到的多个指纹值时，优选采用多媒体的立体声数据作为音频数据，优选采用上述的时频特征数据的方法计算指纹值，以保证目标多媒体文件的特征信息与特征数据库中的特征信息相一致，提高匹配准确率。

[0106] 步骤 S316：将每个立体声数据的多个指纹值分别与定位到的多条多媒体记录进行匹配，得到每个立体声数据对应的匹配率。

[0107] 例如：某第一立体声数据的指纹信息 fp(hashdata，t) 包括多个指纹值，依次为：[(10001,1),(10002,1),(20001,2) (30001,3)……]；

[0108] 某第二立体声数据的指纹信息 fp(hashdata，t) 包括多个指纹值，依次为：[(10002,11),(10004,11),(30001,14) (30005,16)……]；

[0109] 定位到第一条多媒体记录中的特征信息为：[(10003,10),(10002,20),(20001,21) (30001,31)……]；

[0110] 定位到第二条多媒体记录中的特征信息为：[(10002,11),(10004,11),(30001,14) (30005,16)……]。

[0111] 则第一立体声数据对应的匹配率包括：与第一条多媒体记录的匹配个数 3 和和第二条多媒体记录的匹配个数 2；第二立体声数据对应的匹配率包括：与第一条多媒体记录的匹配个数 2 和第二条多媒体记录的匹配个数 4。

[0112] 步骤 S318：根据多个匹配率中的最大值对应的一条多媒体记录确定目标多媒体文件对应的标识信息。

[0113] 例如，最大匹配率为第二立体声数据与第二条多媒体记录的匹配个数 4，因而，该步骤确定的目标多媒体文件对应的标识信息也即该第二条多媒体记录中的标识信息。

[0114] 步骤 S320 :根据目标多媒体文件对应的标识信息识别目标多媒体文件。

[0115] 例如,上述两条多媒体记录对应电视频道 A 播出的两个电视节目的记录,且第二条多媒体记录中的标识信息中的电视节目名称为《看见》,识别到的目标多媒体文件为电视频道 A 播出的《看见》。

[0116] 采用该实施例提供的多媒体文件的识别方法,在上述实施例一的技术效果的基础上,识别目标多媒体文件时,获取到的目标多媒体文件的音频数据是由左声道数据和右声道数据合并而成的立体声数据,相应地,预设的特征样本也为立体声数据的特征,使得目标多媒体文件的特征信息的源数据类型与特征样本的源数据类型相一致,均采用立体声数据,提高了识别的准确率 ;并且在合并左、右声道数据为立体声数据时,设置权重参数 a 和 b,以能够根据实际需要调整左右声道数据在立体声数据中所占的比重。

[0117] 进一步地,在构建目标多媒体文件的特征信息时,通过设置多组权重参数,将目标多媒体文件的左右声道数据转化为多组立体声数据,计算每组立体声数据对应的指纹值,从而目标多媒体文件的特征信息包括多组指纹值。在进行目标多媒体文件识别时,将每组指纹值与定位到的多条多媒体文件记录分别相匹配,根据最大匹配率对应的多媒体文件记录识别目标多媒体文件,进一步增加识别的准确性。

[0118] 实施例四

[0119] 该实施例四提供了一种多媒体文件的搜索方法的实施例,如图 4 所示,该方法包括如下的步骤 S402 至步骤 S406。

[0120] 步骤 S402 :接收搜索请求,其中,该搜索请求包括待搜索的目标多媒体文件的混合音频数据。

[0121] 步骤 S404 :根据搜索请求识别目标多媒体文件。

[0122] 步骤 S406 :根据识别结果搜索目标多媒体文件。

[0123] 在该实施例中,搜索目标多媒体文件时,首先需要识别到该多媒体文件,然后根据识别到的该多媒体文件的标识信息进一步搜索多媒体文件。其中,该处识别目标多媒体文件时,可采用上述的任一实施例。

[0124] 实施例五

[0125] 该实施例五提供了一种多媒体文件的搜索方法的实施例,该方法的执行主体可以为任意的终端,如图 5 所示,该方法包括如下的步骤 S502 至步骤 S512。

[0126] 步骤 S502 :获取目标多媒体对应的混合音频数据,其中,目标多媒体文件混合音频数据包括目标多媒体文件的音频数据和音频水印数据。

[0127] 步骤 S504 :提取目标多媒体文件混合音频数据中的音频水印数据。

[0128] 步骤 S506 :发送音频水印数据至音频水印识别服务器,以得到第一匹配结果,其中,目标多媒体文件第一匹配结果为音频水印识别服务器匹配音频水印数据与预设的音频水印样本得到的匹配结果。

[0129] 在该实施例中,音频水印识别服务器设置有音频水印数据库,该音频水印数据库用于存储预设的音频水印样本,其中,音频水印数据库中存储有多条音频水印记录,每条音频水印记录包括音频水印信息和与该音频水印信息相对应的多媒体文件标识信息。

[0130] 终端将目标多媒体文件的音频水印数据发送至音频水印识别服务器后,音频水印识别服务器在音频水印数据库定位到与目标多媒体文件的音频水印数据相匹配的音频水

印记录，从而得到第一匹配结果，也即从该音频水印记录中得到该目标多媒体文件对应的多媒体文件标识信息。

[0131] 该处的多媒体文件标识信息对于识别目标多媒体文件来讲，识别的细度相对较粗，也即不能唯一的确定目标多媒体文件，例如，通过该处的标识信息识别到目标多媒体文件属于某电视剧集，但是并不能确定具体是该电视剧集中的哪一集；又如，通过该处的标识信息识别到目标多媒体文件属于某电视频道的电视节目，但是并不能确定具体是该电视频道中的哪一个电视节目。

[0132] 步骤 S508：提取目标多媒体文件混合音频数据中的目标多媒体文件的音频数据的特征信息。

[0133] 步骤 S510：发送目标多媒体文件的音频数据的特征信息和第一匹配结果至音频指纹识别服务器，以得到第二匹配结果，其中，第二匹配结果为音频指纹识别服务器在预设的特征样本中确定第一匹配结果对应的特征样本部分后，匹配特征信息与特征样本部分得到的第二匹配结果。

[0134] 在该实施例中，音频指纹识别服务器设置有音频指纹数据库，该音频指纹数据库用于存储预设的特征样本，其中，音频指纹数据库中存储有多条多媒体记录，每条多媒体记录由多媒体的指纹信息、与指纹信息对应多媒体文件的标识信息。

[0135] 该处的多媒体文件标识信息对于识别目标多媒体文件来讲，识别的细度相对较细，通过该标识信息的内容，能够唯一的确定目标多媒体文件。该标识信息可包括上述音频水印数据库中的多媒体文件标识信息，还包括描述多媒体文件更细致的信息，例如多媒体文件的存储位置、多媒体文件的名称等。

[0136] 终端将音频数据的特征信息和第一匹配结果发送至音频指纹识别服务器后，音频指纹识别服务器首先在音频指纹数据库中定位到与第一匹配结果相对应的一条或多条多媒体记录，然后再将音频数据的特征信息与定位到的一条或多条多媒体记录进行匹配，从而得到第二匹配结果，以唯一识别到该目标多媒体文件。

[0137] 步骤 S512：发送第二匹配结果至多媒体管理服务器，以得到目标多媒体文件，其中，该目标多媒体文件为多媒体管理服务器根据第二匹配结果获取到的多媒体文件。

[0138] 例如通过第二匹配结果能够得到目标多媒体文件的 URL，终端将第二匹配结果发送至多媒体管理服务器，多媒体管理服务器根据第二匹配结果中的 URL 获取到目标多媒体文件后，将目标多媒体文件的相关数据返回至终端。该相关数据可以为目标多媒体文件的流媒体数据，终端接收到该流媒体数据直接播放目标多媒体文件；也可以为目标多媒体文件的下载地址，终端接收到该下载地址后，在相应地服务器上下载目标多媒体文件进行播放。

[0139] 在本发明提供的一个优选实施例中，目标多媒体文件混合音频数据还包括用户语音数据，在步骤 S512 之前，该方法还包括如下的步骤：

[0140] 步骤 S514：提取混合音频数据中的用户语音数据。

[0141] 步骤 S516：发送用户语音数据至语音识别服务器，以得到第三匹配结果，其中，第三匹配结果为语音识别服务器匹配用户语音数据与预设的语音样本得到的匹配结果。

[0142] 在该优选实施例中，语音识别服务器设置有语音数据库，该语音数据库用于存储预设的语音样本，其中，语音数据库中存储有多条语音记录，每条语音记录由语音特征信

息、和与语音特征信息对应的关键字构成。

[0143] 终端将用户语音数据发送至语音识别服务器后，语音识别服务器首先根据用户语音数据提取到用户语音的语音特征信息，然后将该提取到用户语音的语音特征信息与语音数据库中的语音特征信息进行匹配，从而能够在该语音数据库中定位到一条或多条语音记录，进而得到该用户语音数据对应的关键字。

[0144] 步骤 S518：根据第二匹配结果识别目标多媒体文件，并根据第三匹配结果验证识别目标多媒体文件得到的识别结果是否正确，其中，其中，当识别结果正确时，执行步骤 S512。

[0145] 在步骤 S510 中，得到第二匹配结果后，在该步骤中，根据第二匹配结果识别目标多媒体文件后，通过第三匹配结果验证识别结果的准确性，并且在识别结果准确时，再将第二匹配结果发送至多媒体管理服务器。

[0146] 例如，目标多媒体文件为某电影 Q，由第二匹配结果能够识别到该电影为电影 Q，并进一步得到该电影 Q 的描述信息中包括该电影 Q 的主演为王 XX 的信息，由第三匹配结果能够得到的关键字为“主演”和“王 XX”，则通过第三匹配结果验证识别结果是正确的，此时将第二匹配结果发送至多媒体管理服务器，以得到该电影 Q。

[0147] 在本发明提供的另一个优选实施例中，步骤 S504 提取目标多媒体文件混合音频数据中的目标多媒体文件的音频数据的特征信息时，可采用上述实施例三中描述的特征信息提取方式，此处不再赘述。

[0148] 实施例六

[0149] 该实施例六提供了另一种多媒体文件的搜索方法的实施例，在该实施例的方法中，目标视频的音频数据中预先添加有声印码数据；在获取目标视频的视频片段音频数据时，获取到了用户的语音数据；在识别目标视频时，将会利用到预设的音频水印数据库和预设的音频指纹数据库；在识别用户的语音数据时，将会利用到预设的语音数据库。具体地，如图 6 所示，该方法包括如下的步骤 S602 至步骤 S608。

[0150] 步骤 S602：开启录音模块，获取目标视频的视频片段的混合音频数据，其中，该混合音频数据包括目标视频的音频数据和用户的语音数据。

[0151] 录音模块开启后，实时录制当前环境中的声音信息以获取混合音频数据，在目标视频播放的过程中，如果有用户语音，则录制到的声音信息包括目标视频的视频片段的音频数据、用户的语音数据以及环境中的一些背景声音数据。

[0152] 在录音模块开启后，每当录音的时长到达时间 T2，则将长度为 T2 的声音数据封装，封装后的声音数据中包含一个视频片段的音频数据以及用户的语音数据和背景声音。

[0153] 步骤 S604：对混合音频数据的音频文件进行预处理。

[0154] 具体包括如下的步骤：

[0155] 1. 音频格式转换。

[0156] 调用第三方的软件（如：ffmpeg）将不同格式的音频文件统一转换为时间长度为 T2 的 PCM 编码的音频数据。

[0157] 2. 提取高频部分的音频数据。

[0158] 使用高通滤波器（滤波的频率范围与声印码占用的频率范围保持一致，假设为 H1Hz 到 H2Hz），获取时间长度为 T2，频率范围为 H1Hz 到 H2Hz 的音频数据 Music1。

[0159] 3. 提取低频部分的音频数据。

[0160] 使用低通滤波器，获取时间长度为 T2，频率范围为 L1Hz 到 L2Hz 的音频数据 Music2。

[0161] 步骤 S606 ：根据预处理后的混合音频数据的音频文件进行识别。

[0162] 具体地，需要识别的内容包括目标视频的视频片段和用户语音数据对应的关键字，包括如下的步骤 ：

[0163] 1. 接收预处理后的混合音频数据的音频文件，也即接收到两个音频片段，包括 Music1 和 Music2。

[0164] 2. 将获取到的低频部分的音频数据拼接，为语音提取准备数据。

[0165] 每接收到一个低频部分音频数据 Music2，则按时间顺序拼接成时间长度为 N*T2(N 为当前音频片段总数) 的音频数据 Music3。

[0166] 3. 利用声印码锁定目标视频。

[0167] 识别高频部分音频数据 Music1 中所携带的声印码信息，得到识别结果 Result1。

[0168] 例如，识别结果 Result1 为目标视频在音频指纹数据库中的 ID 号 (唯一标识视频指纹的 TrackID)。

[0169] Result1 ：{TrackID ：“……”}。

[0170] 4. 精确定位目标视频的视频片段

[0171] 提取 Music2 的指纹信息，将提取的指纹信息与 Result1 所指向的音频指纹数据库中的指纹信息进行匹配，得到匹配结果 Result2。

[0172] Result2 包含的信息为目标视频在音频指纹数据库中的索引信息 TrackID、存储位置信息 URL、以及视频片段在目标视频中的时间范围 timeStart 和 timeStop。

[0173] Result2 ：{TrackID ：“……”,URL:“http://……”,timeStart ：“……”,timeStop ：“……”} ；

[0174] 5. 提取目标视频的视频片段的音频数据，也即该视频片段的原声。

[0175] 读取 Result2，根据存储位置信息 URL 找到视频文件 vedio ；提取视频的文件 vedio 的音频数据 music ；根据 Result2 中的时间信息提取特定时间段 (也即 timeStart 至 timeStop) 的音频数据 music_clip，该 music_clip 即为视频片段的原声。

[0176] 6. 提取用户语音数据。

[0177] 对于拼接好的音频数据 Music3，实质上由如下三部分共同构成 ：

[0178] Music3 = a1* 视频片段的音频数据 +a2* 用户语音数据 +a3* 背景声音数据 (a1，a2，a3 为常量)

[0179] 假设录音条件足够好，即 ：a3 = 0，则 ：

[0180] 用户语音数据 = b1*Music3 - b2* 视频片段的音频数据 (b1，b2 为常量)

[0181] 因此，可提取用户语音数据 ：

[0182] 用户语音数据 word = b1*Music3 - b2*music_clip。

[0183] 7. 由用户语音数据解析语音指令。

[0184] 将用户语音数据 word 与语音数据库中的语音指令进行匹配，得到与 word 最为接近的语音指令 Command ：

[0185] Command ：{index ：{ “music，title， ……”}}

[0186] 步骤 S608 :根据识别结果返回检索结果。

[0187] 例如，Command 中的 index 信息包括音乐名称，演唱者姓名，则可由 Result2 得到描述该视频片段的所有信息，包括该视频片段中出现的物品信息、场景信息、背景音乐的音乐名称、该背景音乐的演唱者姓名等信息，在该所有信息中，判断 Command 中的 index 信息对应的内容是否在 Result2 得到描述该视频片段的所有信息中，若在，则可由 Result2 中的 URL 找到视频文件 vedio，将该视频文件 vedio 或该视频文件的链接地址等信息作为检索结果返回。

[0188] 实施例七

[0189] 该实施例七提供了另一种多媒体文件的搜索方法的实施例，在该实施例中，描述了实现该搜索方法的一个搜索系统。

[0190] 具体地，实现该方法的系统由终端和服务器及数据库两部分构成，分别说明如下。

[0191] 首先，终端包括录音模块、音频预处理模块和音频识别模块。其中，

[0192] 录音模块 :用于获取声音信息。输入的声音信息由两部分构成 :(1) 视频的声音数据（包含声印码）;(2) 用户的语音数据。用户可以在录音过程中的任意时间语音输入其语音信息。

[0193] 音频预处理模块 :如图 7 所示，音频格式转换单元将录音模块获取的音频数据进行数据转换，并分别由高频提取单元和低频提取单元进行音频提取，为下一步的视频识别和语音识别准备数据。

[0194] 音频识别模块 :如图 8 所示，该音频识别模块接收预处理后和音频数据，包括高频声音数据和低频声音数据，输出音频指纹检索结果，也即目标视频的检索结果，还输出用户的语音数据，各单元分别说明如下 :

[0195] 声印码识别单元 :用于与声印码识别服务器交互，向该声印码识别服务器上传高频音频信息，获取声印码识别服务器发来的声印码识别结果。

[0196] 指纹识别单元 :用于接收声印码识别结果，连同低频声音数据上传至音频指纹识别服务器，接收指纹识别服务器返回的音频指纹识别结果。

[0197] 音频拼接单元 :用于将低频声音数据的片段拼接成完整，为用户语音提取准备数据。

[0198] 语音识别单元 :一方面用于接收音频指纹识别结果，将该结果上传至视频管理服务器，并接收视频管理服务器发送的目标视频的音频片段。另一方面用于根据音频指纹识别结果识别获取原声音频数据，并根据该原声音频数据和音频拼接单元发送的低频声音数据，提取用户的语音数据。再一方面用于将语音数据上传至语音识别服务器，并接收语音识别服务器返回的语音识别结果。

[0199] 音频识别模块还可以包括 :识别结果验证单元，用于根据用户的语音数据判指纹识别单元的识别结果是否正确，并在正确时将音频指纹识别结果作为搜索意图发送至视频管理服务器，以及接收视频管理服务器的返回结果。

[0200] 终端还包括 :显示模块，用于将识别结果验证单元接收到的视频管理服务器返回的结果显示给用户，也用于根据返回信息的类型，调用多种多媒体文件资源，将结果显示给用户。

[0201] 上述内容描述了系统中的终端，参考图 9，以下将描述系统中服务器和数据库构

成。

[0202] 1. 语音识别服务器和语音数据库。

[0203] 语音识别服务器接收终端发送的语音数据，根据语音数据在语音数据库中进行识别，并返回与语音数据相对应的语音指令。

[0204] 音数据库中存储预先设定的语音指令：

[0205] Command：{“关键字 1”，“关键字 2”，“关键字 3”……}。

[0206] 指令可以用关键字描述，这些关键字可包括：视频类型（如：电视剧、电影、新闻），视频内容的信息（如：演员、物品、地点）。

[0207] 2. 音频水印识别服务器和音频水印数据库。

[0208] 接收终端发来的高频部分声音数据，从该高频部分声音数据中解析出其携带的声印码，将解析出的声印码与数据库中声印码数据进行匹配，获取声印码匹配度最高的匹配结果。将匹配结果返回给终端。

[0209] 在音频水印数据库中，存储的数据结构可采用如下的结构：

[0210] {“ID”，“url”，“声印码”，“TrackID”}

[0211] “ID”为该声印码在声印码数据库中的唯一标识。“url”为该声印码对应的视频文件的存储位置。“声印码”：为一段“01010101……”二进制的数列，每一个视频文件对应唯一的声印码。使用时，该声印码被加载在高频的音频数据中。“TrackID”为该声印码对应的视频数据所对应的指纹信息在音频指纹数据库中的标识。

[0212] 3. 音频指纹识别服务器和音频指纹数据库。

[0213] 指纹数据库中存储的指纹数据，其数据结构可采用如下的结果：

[0214] {TrackID：{}，fp：{}，“关键字 1”：{}，“关键字 2”：{}，……}

[0215] TrackID 为该指纹信息在音频指纹数据库中的唯一标识；fp 为视频文件对应的音频指纹数据，其结构为{“Hash1”，“time1”，“Hash2”，“time2”，“Hash3”，“time3”，……}；“关键字 N”：该关键字与语音服务器中的关键字一一对应，其结构{“内容 1”，“time1”，“内容 2”，“time2”，“内容 3”，“time3”，……}（例：“关键字＝演员”，则该关键字的内容标识视频在不同的播放时刻中，出现的演员的姓名，存储在“内容”中）。

[0216] 音频指纹识别服务器的功能说明如下：

[0217] 接收终端视频识别模块发来的低频部分声音数据，音频水印识别服务器发来的识别结果；从声音数据中提取目标音频的指纹信息；音频水印识别服务器发来的识别结果中提取 TrackID，根据 TrackID 确定指纹检索的范围；在确定的检索范围内，将目标指纹与 TrackID 指向的指纹信息进行匹配；获取当前播放视频的时间信息，将匹配到的音频片段返回给视频识别模块。

[0218] 4. 视频管理服务器和视频数据库。

[0219] 视频数据库用于储视频文件。

[0220] 视频管理服务器的功能说明如下

[0221] (1) 接收终端视频识别模块音频指纹识别单元发来的识别结果；提取结果获取数据库中视频索引信息 (URL) 和时间片段信息，检索到相应视频的片段；提取该视频片段的音频数据；将该音频数据返回给终端视频识别模块的语音识别单元。

[0222] (2) 接收终端识别结果验证单元发来的搜索意图，向终端返回搜索结果。

[0223] 采用上述的终端和服务器，参照图 10，实现本实施例的搜索方法的过程描述如下：

[0224] 步骤一：

[0225] 开启终端录音模块，获取目标视频的声音数据和用户的语音信息。录音模块一旦开启，终端就由其录音设备录制当前环境中的声音信息。直到接收到服务器发送来的视频检索结果或者是录音总时长大于预先设定的时间 T1 则停止录音。

[0226] 每当录音的时长到达时间 T2(T2<<T1)，则将长度为 T2 的声音数据封装，数据中包含视频的音频数据以及用户的语音指令，同时将该数据上传至音频数据预处理模块。

[0227] 步骤二：

[0228] 该步骤为下一步的视频识别准备数据。实现过程如下：

[0229] 1. 接收时间长度为 T2 的音频文件；

[0230] 2. 音频格式转换。调用第三方的软件（如：ffmpeg）将不同格式的音频文件统一转换为时间长度为 T2 的 PCM 编码的音频数据。

[0231] 3. 提取高频部分的音频数据。使用高通滤波器（滤波的频率范围与声印码占用的频率范围保持一致，假设为 H1 ～ H2Hz），获取时间长度为 T2，频率范围为 H1 ～ H2Hz 的音频数据 Music1。

[0232] 4. 提取低频部分的音频数据。使用低通滤波器，获取时间长度为 T2，频率范围为 L1 ～ L1Hz 的音频数据 Music2。

[0233] 5. 将 Music1 和 Music2 上传至视频识别模块。

[0234] 步骤三：

[0235] 步骤三分成两个阶段：阶段一提取语音信息和阶段二语音识别，其中，阶段一在终端进行，需要视频管理服务器提供目标视频的原声音频片段；阶段二在语音识别服务器进行，需要终端提供用户语音数据。在该步骤中，根据获取的音频数据，对目标视频进行检索，获取目标视频的检索结果和用户的语音信息，进而识别出用户的语音指令，具体地，两个阶段说明如下：

[0236] 阶段一

[0237] 1. 视频识别模块从音频预处理模块接收音频片段（包括 Music1 和 Music2）。

[0238] 2. 将获取到的低频部分音频数据拼接，为语音提取准备数据。

[0239] 每接收到一个低频部分音频数据 Music2，则按时间顺序拼接成时间长度为 N*T2(N 为当前音频片段总数）的低频音频数据 Music3，直到接收到视频检索结果为止。

[0240] 3. 利用声印码锁定目标视频。

[0241] 将高频部分音频数据 Music1 上传至声印码识别服务器，由声印码识别服务器提取并识别 Music1 中所携带的声印码信息。将识别结果 Result1 返回给终端的视频识别模块。识别结果 Result1 为目标视频在“视频指纹数据库”中的 ID 号（唯一标识视频指纹的 TrackID）。

[0242] Result1：{TrackID：“……”}；

[0243] 4. 精确定位播放视频的播放片段。

[0244] 将低频部分音频数据 Music2 连同 Result1，一同上传至“音频指纹识别服务器”。由指纹识别服务器提取 Music2 中的目标视频指纹信息，将目标视频指纹与 Result1 所指向

的指纹进行匹配，将匹配结果 Result2 返回给终端的视频识别模块。

[0245] Result2 包含的信息为检索结果在指纹库的索引信息 TrackID、在视频数据库中的索引信息 URL、以及时间范围 timeStart 和 timeStop。

[0246] Result2 :{TrackID :“……”,URL:“http://……”,timeStart :“……”,timeStop :“……”} ;

[0247] 5. 视频识别模块停止接收音频片段，同时向录音模块发送停止录音的信息。

[0248] 6. 提取目标视频的音频原声数据

[0249] 读取 Result2，将 Result2 上传至视频管理服务器 ;视频管理服务器根据索引信息找到视频文件 vedio ;提取视频的音频数据 music ;根据 Result2 中的时间信息提取特定时间段的音频数据 music_clip ;将 music_clip 返回给终端的视频信息推荐模块。

[0250] 7. 提取用户语音信息

[0251] 我们获取的，拼接好的音频数据 Music3，由如下三部分共同构成 :

[0252] Music3 = a1* 原声音频 +a2* 用户语音 +a3* 环境噪声 ;(a1，a2，a3 为常量)

[0253] 我们假定录音条件足够好的条件下 (即 :a3 ～= 0)，我们可以简单的通过如下方式获取 :

[0254] 用户语音 = b1*Music3 - b2* 原声音频 (b1，b2 为常量)

[0255] 其中原声音频由步骤 6 获得 :music_clip。

[0256] 用户语音 (word) = b1*Music3 - b2*music_clip ;

[0257] 阶段二

[0258] 8. 由用户语音解析用户指令

[0259] 将 word 上传至语音识别服务器，与语音数据库中的语音指令进行匹配。将与 word 最为接近的语音指令 Command 返回给识别结果验证单元。

[0260] Command :{index :{“music，title，belowing……”}}

[0261] 9. 根据用户指令，返回检索结果

[0262] 判断 Command 中的 index 信息对应的内容是否在 Result2 得到描述该视频片段的所有信息中，若在，则将 Result2 上传至视频管理服务器。由 Result2 中的 URL 定位匹配视频文件，并返回视频文件。

[0263] 在该实施例中，通过语音识别单元识别用户语音数据，通过识别结果进一步识别目标视频。使得搜索系统能够结合用户的语音数据将准确性高的检索结果返回至用户。

[0264] 以上是对本发明实施例提供的多媒体文件的识别方法和搜索方法进行的描述，以下将描述本发明实施例提供的多媒体文件的识别装置。

[0265] 实施例八

[0266] 该实施例八提供了一种多媒体文件的识别装置的实施例，如图 11 所示，该装置包括获取模块 810、第一提取模块 820、第一匹配模块 830、确定模块 840、第二提取模块 850、第二匹配模块 860 和识别模块 870。

[0267] 获取模块 810 用于获取目标多媒体对应的混合音频数据，其中，混合音频数据包括目标多媒体文件的音频数据和音频水印数据。该目标多媒体文件可以为视频 (或音频)，在该视频 (或音频) 播放的过程中，识别装置启动录音设备进行声音录制，从而得到该视频 (或音频) 的混合音频数据，由于该视频 (或音频) 中预先添加了音频水印数据，因而录制

到的混合音频数据中既包括目标多媒体文件的音频数据本身，也包括添加进去的音频水印数据。该识别装置可以为智能移动通信终端，例如手机、PAD；也可以为计算机；或者作为独立的识别单元嵌入到需要进行多媒体文件识别的装置中。

[0268] 第一提取模块 820 用于提取混合音频数据中的音频水印数据，根据音频水印数据的特点，将混合音频数据中的音频水印数据提取出来，例如，音频水印数据为声印码数据，由于声印码数据是在音频数据的高频段中加入的特定信息，因此，通过提取混合音频数据中高频部分的数据，即可得到声印码数据。

[0269] 第一匹配模块 830 用于匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果，该第一匹配结果可以为由多个多媒体文件构成的一个多媒体文件组，也即，通过该方法，确定了目标多媒体文件在该多媒体文件组中。

[0270] 确定模块 840 用于在预设的特征样本中确定第一匹配结果对应的特征样本部分，通过该模块，在进行音频指纹匹配时，无需将目标多媒体文件的音频数据的特征信息与预设的特征样本整体进行匹配，只需将特征信息与第一匹配结果对应的特征样本部分进行匹配。

[0271] 第二提取模块 850 用于提取混合音频数据中的目标多媒体文件的音频数据的特征信息，具体可采用现有技术中任意的音频数据的特征提取方法，例如可提取音频的时域特征数据，具体如提取音频片段的幅值，也可提取音频的时频特征数据。

[0272] 第二匹配模块 860 用于匹配特征信息与特征样本部分，以得到第二匹配结果。识别模块 870 用于根据第二匹配结果识别目标多媒体文件。

[0273] 采用该实施例提供的多媒体文件的识别装置，有效地结合了基于音频水印的视频识别技术和基于音频指纹的视频识别技术，与单纯使用基于音频水印的视频识别技术的方法相比，提升了识别的细度和视频识别的应用范围；与单纯使用基于音频指纹的视频识别技术相比，缩短了识别的时间。

[0274] 优选地，混合音频数据中还包括用户语音数据，该装置还包括第三提取模块、第三匹配模块以及验证模块。其中，第三提取模块用于提取混合音频数据中的用户语音数据，第三匹配模块用于匹配用户语音数据与预设的语音样本，以得到第三匹配结果；根据所述第三匹配结果从所述根据第二匹配结果识别得到的目标多媒体文件中选择一目标多媒体文件。

[0275] 采用该优选实施例，能够结合用户语音数据来提高识别目标多媒体文件的准确性。

[0276] 进一步优选地，第一提取模块在提取音频水印数据时具体执行的步骤为：提取混合音频数据中的高频部分的音频数据；第二提取模块在提取特征信息时具体执行的步骤为：提取混合音频数据中的低频部分的音频数据的特征信息；第三提取模块在提取用户语音数据时具体执行的步骤为：提取混合音频数据中的低频部分的音频数据；去除低频部分的音频数据中的目标多媒体文件的音频数据，以得到用户语音数据。

[0277] 优选地，第二提取模块包括左右声道数据提取模块、立体声数据合成模块和指纹信息提取模块。其中，左右声道数据提取模块用于提取混合音频数据中的低频部分的左声道数据和右声道数据；立体声数据合成模块用于采用以下公式合并左声道数据和右声道数据，以得到低频部分的立体声数据：$s = a*l+b*r$，其中，$a+b = 1$，s 为低频部分的立体声数

据，l 为低频部分的左声道数据，r 为低频部分的右声道数据，a 和 b 为预设的参数；以及指纹信息提取模块用于提取立体声数据的时频特征数据得到目标多媒体文件的指纹信息，其中，指纹信息构成目标多媒体文件的音频数据的特征信息。

[0278] 采用该优选实施例，识别目标多媒体文件时，获取到的目标多媒体文件的音频数据是由左声道数据和右声道数据合并而成的立体声数据，相应地，预设的特征样本也为立体声数据的特征，使得目标多媒体文件的特征信息的源数据类型与特征样本的源数据类型相一致，均采用立体声数据，提高了识别的准确率；并且在合并左、右声道数据为立体声数据时，设置权重参数 a 和 b，以能够根据实际需要调整左右声道数据在立体声数据中所占的比重。

[0279] 进一步优选地，若目标多媒体文件为第二多媒体文件的一个子多媒体文件，第一匹配结果为第二多媒体文件的标识信息，第二匹配结果为目标多媒体文件的标识信息，特征样本为预设的特征数据库中存储的至少一条多媒体记录，多媒体记录包括多媒体文件的指纹信息、与指纹信息对应的多媒体文件的标识信息，则确定模块在确定特征样本部分时具体执行的步骤为：在特征数据库中，定位到第二多媒体文件的标识信息对应的一条或多条多媒体记录；第二匹配模块在得到第二匹配结果时具体执行的步骤为：匹配目标多媒体的指纹信息与定位到的一条或多条多媒体记录，以确定目标多媒体的标识信息。

[0280] 进一步优选地，低频部分的立体声数据为 N 个立体声数据，其中，N 个立体声数据中的第 i 个立体声数据为 $s_i = a_i*l+b_i*r$，$a_i' + b_i' = 1$，$i = 1, 2, 3\cdots N$，则第二匹配模块包括匹配率确定模块和标识信息确定模块。其中，匹配率确定模块用于将每个立体声数据的时频特征数据分别与定位到的一条或多条多媒体记录进行匹配，得到立体声数据对应的多个匹配率；标识信息确定模块用于根据多个匹配率中的最大值对应的一条多媒体记录确定目标多媒体文件的标识信息。

[0281] 采用该优选实施例，在构建目标多媒体文件的特征信息时，通过设置多组权重参数，将目标多媒体文件的左右声道数据转化为多组立体声数据，计算每组立体声数据对应的特征信息，从而目标多媒体文件的特征信息包括多组特征西悉尼。在进行目标多媒体文件识别时，将每组特征信息与定位到的多条多媒体文件记录分别相匹配，根据最大匹配率对应的多媒体文件记录识别目标多媒体文件，进一步增加识别的准确性。

[0282] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉该技术的人在本发明所揭露的技术范围内，可轻易想到的变化或替换，都应涵盖在本发明的保护范围之内。因此，本发明的保护范围应该以权利要求的保护范围为准。


开始
获取目标多媒体对应的混合音频数据
S102
提取混合音频数据中的音频水印数据
S104
匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果
S106
在预设的特征样本中确定第一匹配结果对应的特征样本部分
S108
提取混合音频数据中的目标多媒体文件的音频数据的特征信息
S110
匹配特征信息与特征样本部分，以得到第二匹配结果
S112
根据第二匹配结果识别目标多媒体文件
S114
结束

图 1

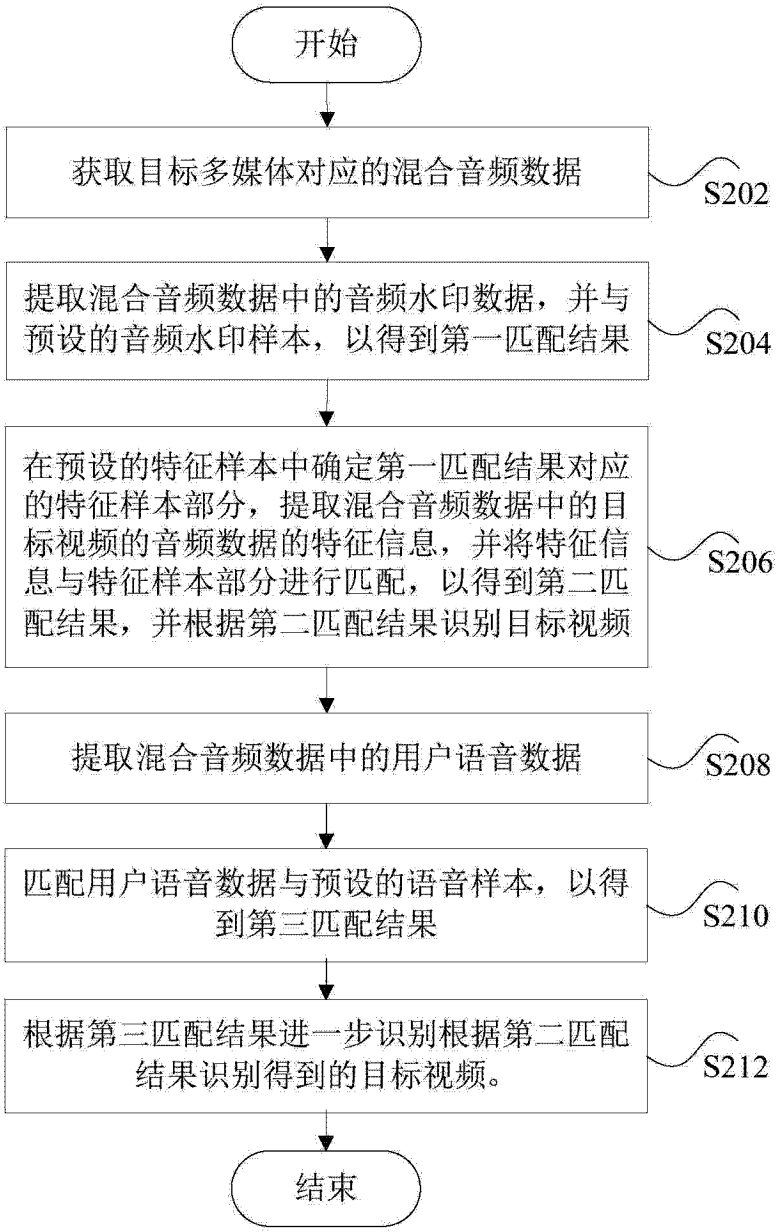

开始
获取目标多媒体对应的混合音频数据
S202
提取混合音频数据中的音频水印数据，并与预设的音频水印样本，以得到第一匹配结果
S204
在预设的特征样本中确定第一匹配结果对应的特征样本部分，提取混合音频数据中的目标视频的音频数据的特征信息，并将特征信息与特征样本部分进行匹配，以得到第二匹配结果，并根据第二匹配结果识别目标视频
S206
提取混合音频数据中的用户语音数据
S208
匹配用户语音数据与预设的语音样本，以得到第三匹配结果
S210
根据第三匹配结果进一步识别根据第二匹配结果识别得到的目标视频。
S212
结束

图 2


开始
获取目标多媒体对应的混合音频数据
S302
提取混合音频数据中的音频水印数据
S304
匹配音频水印数据与预设的音频水印样本，以得到第一匹配结果
S306
在预设的特征数据库中定位到第一匹配结果对应的多条多媒体文件记录
S308
提取混合音频数据中的低频部分的左声道数据和右声道数据
S310
合并左声道数据和右声道数据，以得到低频部分的N个立体声数据
S312
计算每个立体声数据的时频特征数据，得到每个立体声数据的多个指纹值
S314
将每个立体声数据的多个指纹值分别与定位到的多条多媒体记录进行匹配，得到每个立体声数据对应的匹配率
S316
根据多个匹配率中的最大值对应的一条多媒体记录确定目标多媒体文件对应的标识信息
S318
根据目标多媒体文件对应的标识信息识别目标多媒体文件
S320
结束

图 3

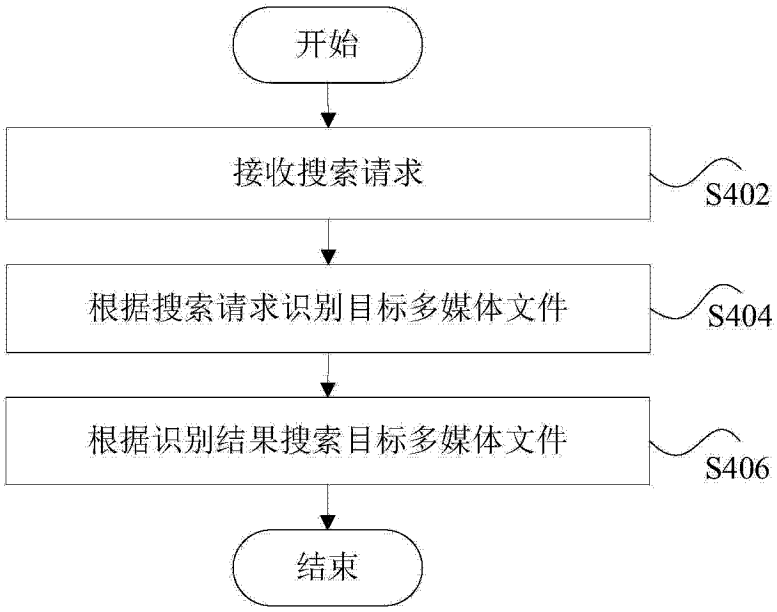

开始
接收搜索请求
S402
根据搜索请求识别目标多媒体文件
S404
根据识别结果搜索目标多媒体文件
S406
结束

图 4


开始
获取目标多媒体对应的混合音频数据
S502
提取混合音频数据中的音频水印数据
S504
发送音频水印数据至音频水印识别服务器，以得到第一匹配结果
S506
提取目标多媒体文件混合音频数据中的目标多媒体文件的音频数据的特征信息
S508
发送目标多媒体文件的音频数据的特征信息和第一匹配结果至音频指纹识别服务器，以得到第二匹配结果
S510
发送第二匹配结果至多媒体管理服务器，以得到目标多媒体文件
S512
结束

图 5

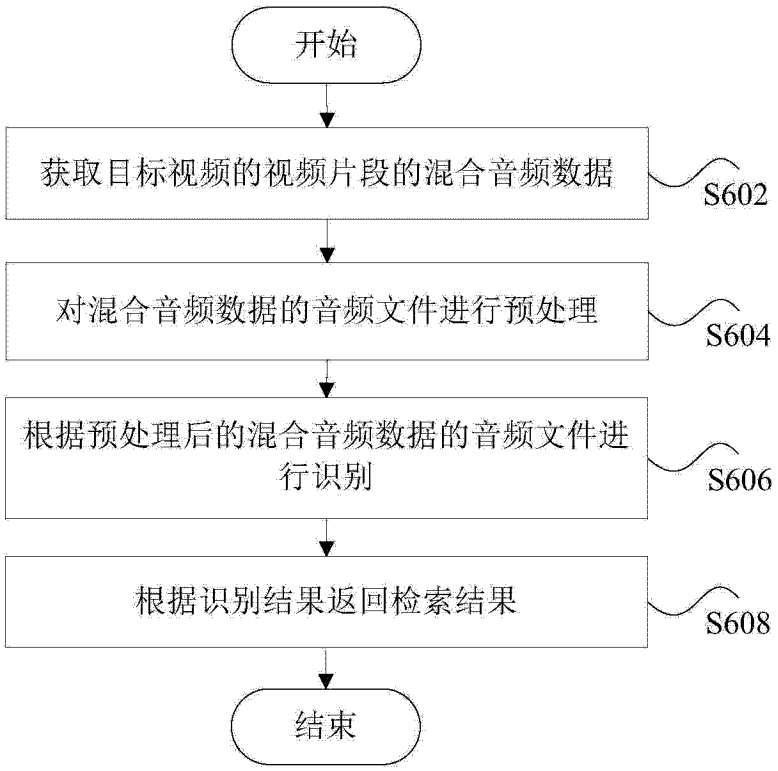

开始
获取目标视频的视频片段的混合音频数据
S602
对混合音频数据的音频文件进行预处理
S604
根据预处理后的混合音频数据的音频文件进行识别
S606
根据识别结果返回检索结果
S608
结束

图 6

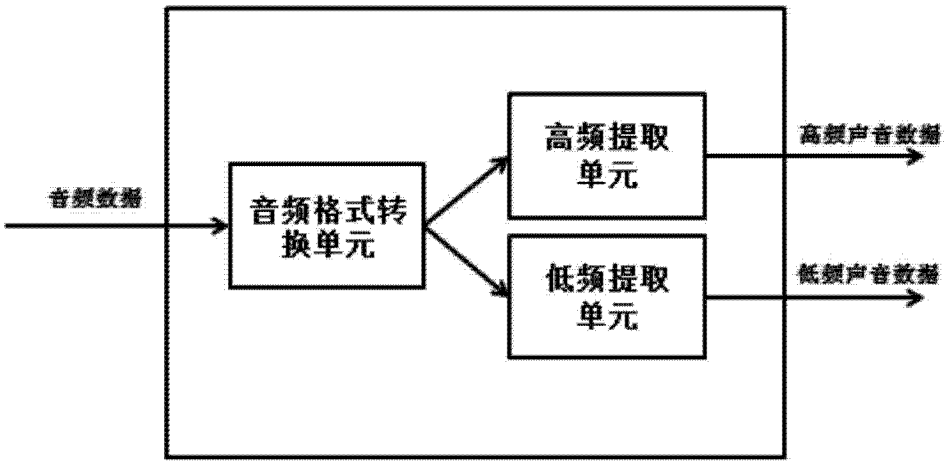

音频数据
音频格式转换单元
高频提取单元
低频提取单元
高频声音数据
低频声音数据

图 7

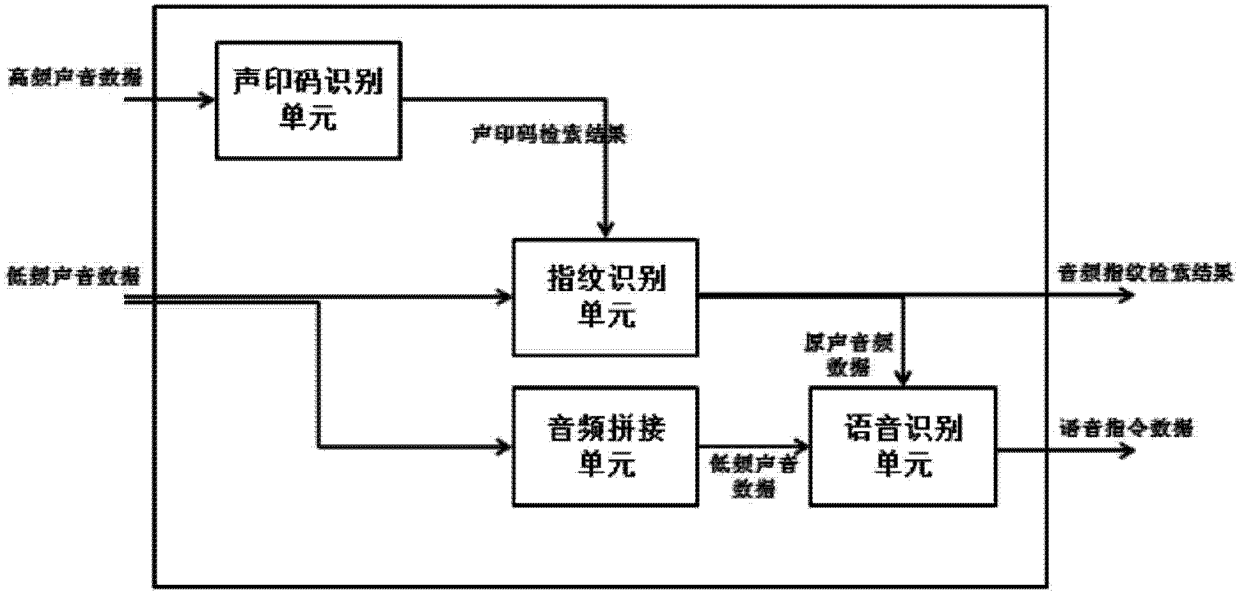

高频声音数据
声印码识别单元
声印码检索结果
低频声音数据
指纹识别单元
音频指纹检索结果
原声音频数据
音频拼接单元
低频声音数据
语音识别单元
语音指令数据

图 8

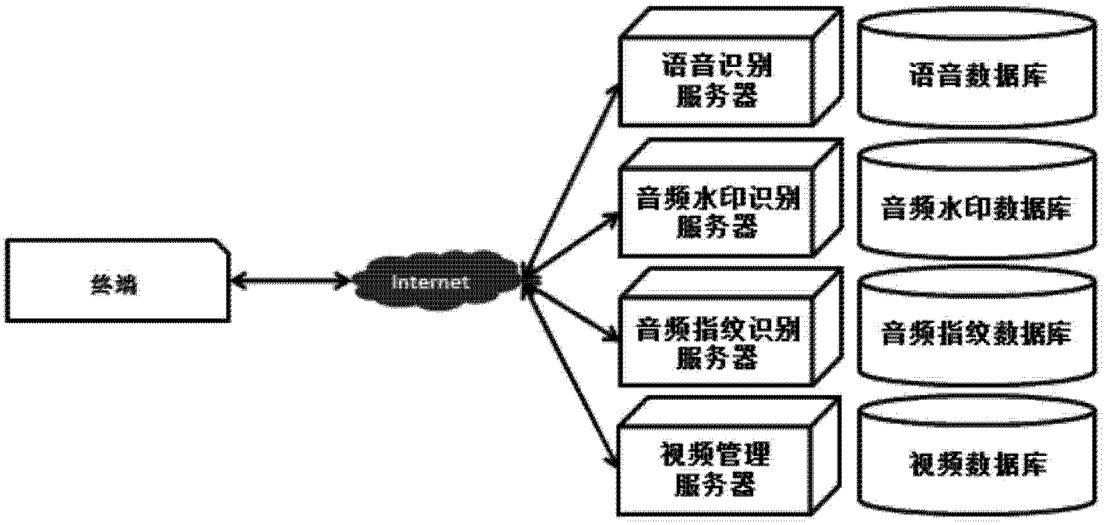

语音识别服务器
语音数据库
音频水印识别服务器
音频水印数据库
终端
Internet
音频指纹识别服务器
音频指纹数据库
视频管理服务器
视频数据库

图 9

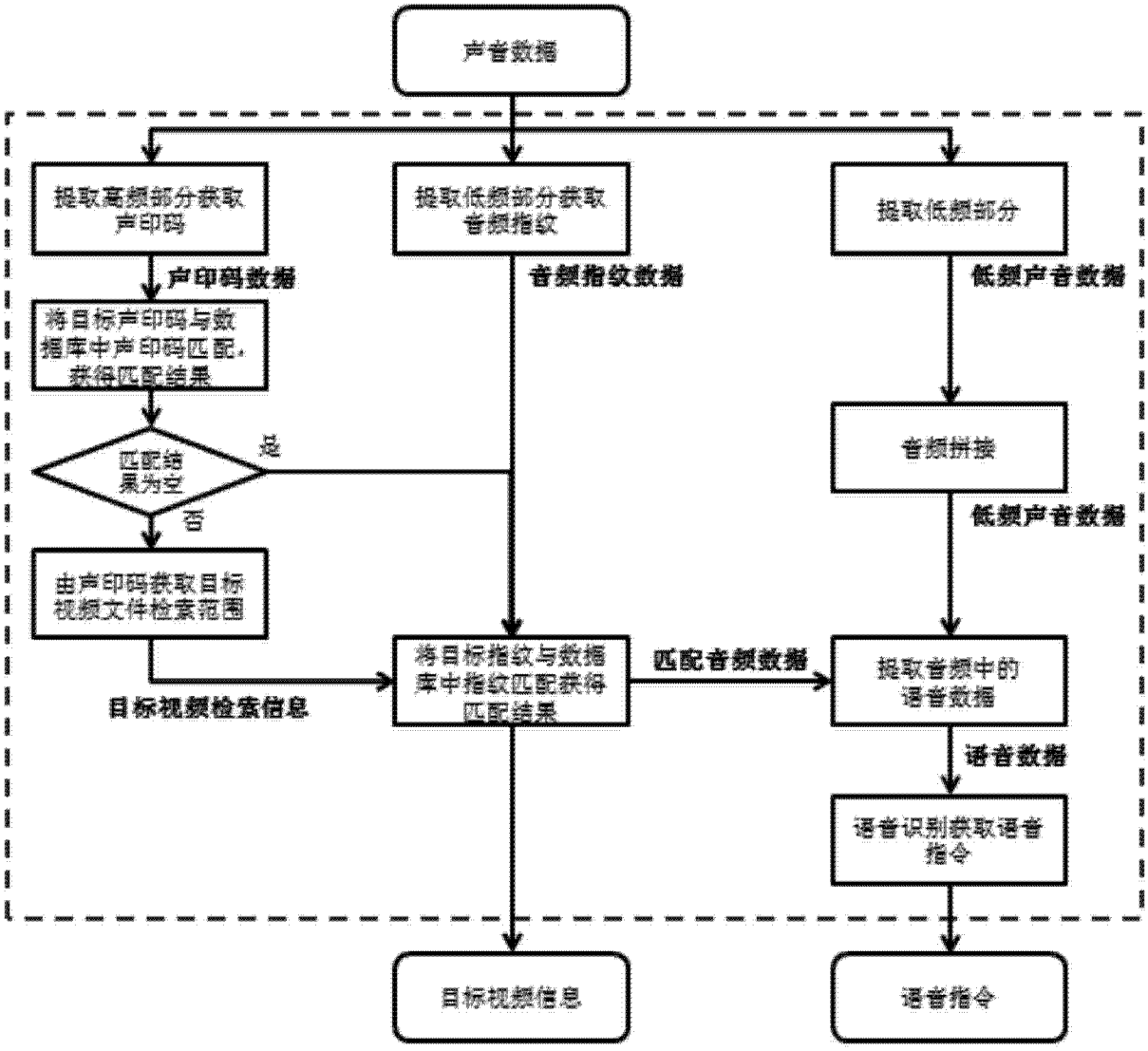

声音数据
提取高频部分获取声印码
提取低频部分获取音频指纹
提取低频部分
声印码数据
音频指纹数据
低频声音数据
将目标声印码与数据库中声印码匹配，获得匹配结果
匹配结果为空
是
否
音频拼接
低频声音数据
由声印码获取目标视频文件检索范围
目标视频检索信息
将目标指纹与数据库中指纹匹配获得匹配结果
匹配音频数据
提取音频中的语音数据
语音数据
语音识别获取语音指令
目标视频信息
语音指令

图 10

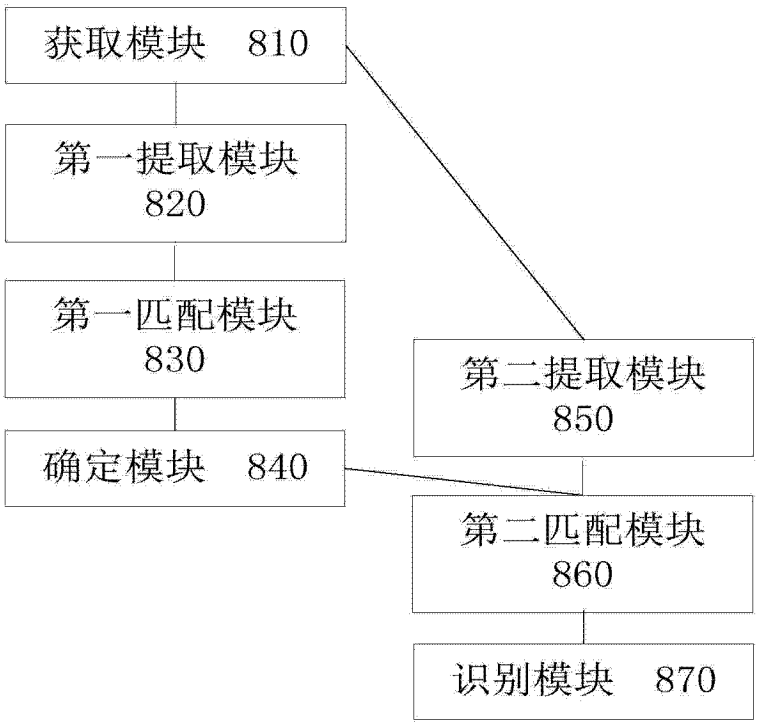

获取模块 810
第一提取模块 820
第一匹配模块 830
确定模块 840
第二提取模块 850
第二匹配模块 860
识别模块 870

图 11