DOI:10.16751/j.cnki.hbkj.2007.06.019
第27卷第6期 咸宁学院学报 Vol.27, No.6
2007年12月 Journal of Xianning College Dec. 2007
文章编号:1006-5342(2007)06-0066-03

基于内容的音频检索关键技术研究与设计*

吴春辉,陈洪生

(咸宁学院 计算机系,湖北 咸宁 437100)

摘 要:基于内容的音频检索是多媒体检索技术中一个重要的组成部分.本文在综述了国内外现行的音频信息检索方法的基础上,通过一种利用声学特征的相似性检索音频文件的方法来分析基于内容的音频检索的关键技术,并在综合运用该关键技术的基础上设计了一个简单的系统对基于内容的音频检索方法进行了测试.

关键词:基于内容的音频检索;关键技术;多媒体;声学特征;音频
中图分类号:TN912.3 文献标识码:A

0 引 言

随着互联网的发展,越来越多的人能够更加方便、快捷、经济地接触到数字媒体,多媒体数据也已经成为互联网信息高速公路上所传送数据的主要部分.这时人们面临的问题不再是缺少多媒体内容,而是如何在浩如烟海的多媒体世界中找到自己所需要的信息.为了方便人们快速的检索到自己所需的多媒体信息,近年来,国内外在多媒体数据库技术的研究中出现了一个新热点—基于内容的检索 CBR(Content Based Retrieval)技术.

1 基于内容的音频检索

基于内容的音频检索技术突破了基于关键词匹配的传统检索技术的限制,它根据音频本身所固有的特征而不是人工标注的外部属性或者关键词对音频进行检索.他的核心思想是通过一定的计算机处理,分析音频的结构和语义,建立它们的结构化的组织和索引,使得"无序"的音频变得"有序",从而有利于用户的检索和浏览.

基于内容的音频检索技术主要分为三大部分:音频内容的获取、音频内容的描述(音频特征提取)、特征相似度匹配.而其中音频内容描述是整个基于内容的音频检索技术的核心技术,它是在音频内容获取的基础之上进行的,同时是进一步进行音频特征相似度匹配的必要前提.

特征提取指的是寻找原始音频信号表达形式,提取能代表原始信号的数据.要抽取特征和属性,通常要对数据库中的多媒体数据项进行预处理.因为在检索过程中,其实是对这些特征和属性而不是对信息项本身进行搜索和比较,所以特征抽取的质量决定着检索效果.

2 基于内容的音频检索关键技术研究与设计

2.1 基于内容的音频检索关键技术研究

在基于内容的音频检索关键技术研究中,我们主要用到的方法是基于模糊聚类的音频例子检索算法,即首先对音频数据库中的每个音频例子自动提取其音频特征,然后对特征向量进行模糊聚类,用形成的聚类质心去表示数据库中的每个音频例子.对于用户提交"哼"出来的音频例子,在得到其聚类质心后,将提交音频例子的聚类质心与数据库中所有音频例子聚类质心进行快速匹配,取最相似的几个音频例子反馈给用户.

假设在实验中我们从每个音频例子的短时音频帧中提取了基音频率、倒谱系数、相对信噪比以及功率谱特征四个特征,每个短时帧的这四个特征组成的向量就构成了这个音频例子的特征.

(1)音频例子聚类质心提取

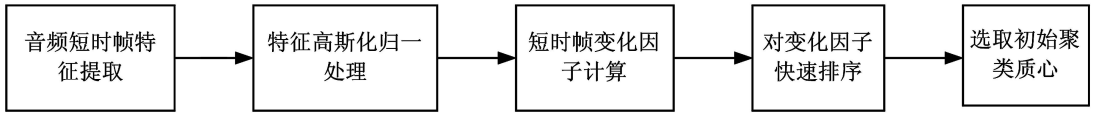

图1 初始聚类质心选取

* 收稿日期:2007-10-29

通过上述步骤得到初始聚类质心后,用如下公式计算每个短时帧对每个质心的隶属度 μ_{ij}:

$$\mu_{ij} = [1/d(\chi_j, V_i)]^{(q-1)-1} / \sum_{k=1}^{K} 1/d(\chi_j, V_k)]^{(q-1)-1} \quad (1)$$

根据计算出来的隶属度,形成新的聚类质心 V_i^*:

$$V_i^* = \sum_{j=1}^{n} (\mu_{ij})^{ij} \chi_j \sum_{j=1}^{N} (\mu_{ij})^{ij} \quad (2)$$

用新的聚类质心 V_i^* 更换 V_i,重新计算隶属度 μ_{ij},得到新的隶属度 μ_{ij}^*,如果

$\max_{ij}[|\mu_{ij} - \mu_{ij}^*|] < \varepsilon$ 成立,则聚类停止.

每个音频例子通过以上方式得到 K个质心,为每个音频例子建立了索引用 K个聚类质心来表征每一个音频例子.

(2)音频例子的相似度比较

既然每个音频都可以用 K个质心表征,那么两个音频之间的相似度也可以通过这 K个质心来计算.假设 $V=\{v_1, v_2, \cdots, v_i, \cdots, v_k\}$ 表示用户提交检索的例子音频 request所形成的模糊聚类质心,$W=\{w_1, w_2, \cdots, w_i, \cdots, w_k\}$ 表示音频检索数据库中与 V进行相似比较的某个音频 clip的聚类质心,则用如下的方法计算 request和 clip的相似度:

1)对于 V中的每个 $v_i(1 \leqslant i \leqslant K)$,在 W中找到与其最相似的 $w_j(1 \leqslant j \leqslant K)$,记为 $g(v_i, W) = \arg\min_{w \in W} d(v_i, w_j)$,其中 d表示余弦相似度.同理,对于 W中的每个 V中的每个 $w_j(1 \leqslant j \leqslant K)$,在 V中找到与其最相似的 $v_i(1 \leqslant i \leqslant K)$,记为 $g(w_j, V) = \arg\min_{v \in V} d(w_j, v_i)$.

2) request和 clip之间的相似度

$$Dis(V, W) = (|V| + |W|)^{-1} [\sum_{\omega \in V} d(v, g(v, W)) + \sum_{\omega \in W} d(w, g(w, W))] \quad (3)$$

这样,由 Dis可以求出音频数据库中所有音频与 request的相似程度,然后选择若干最相似的音频返回给用户,完成检索.

2.2　基于内容的音频检索关键技术设计

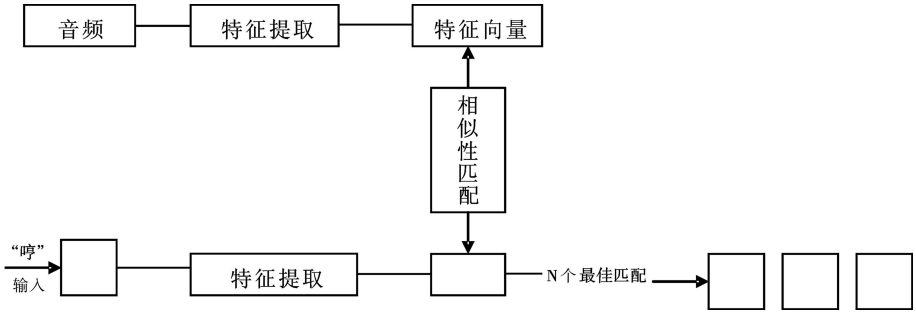

图 2　基于内容的音频检索流程图

音频数据库建立后,首先对多媒体数据库中的音频数据进行特征提取,并通过特征对数据聚类.音频检索主要采用基于“哼”的音频检索方式(Query by Humming),用户通过“哼”的检索界面提交一个用嘴通过麦克风哼出来的语音例子,系统对用户通过哼提交的语音提取特征,并对特征矢量进行模糊聚类,然后检索引擎对特征矢量与聚类参数集匹配,按相关性排序后通过查询接口返回给用户.

图 3为基于哼的检索界面,用户进入该界面后,可以先录一段自己的语音,然后保存为. wav文件,检索时可以检索出来,并且按照匹配的结论把结果按照从小到大的顺序反馈给用户,表示可能是三个人中的一个在说话.并且排在第一位的说话人的概率最大.在实验中,说话者李伟先通过录音录了一段自己的语音,然后检索,可以看出,该实验中检索出说话者最大可能性是李伟.

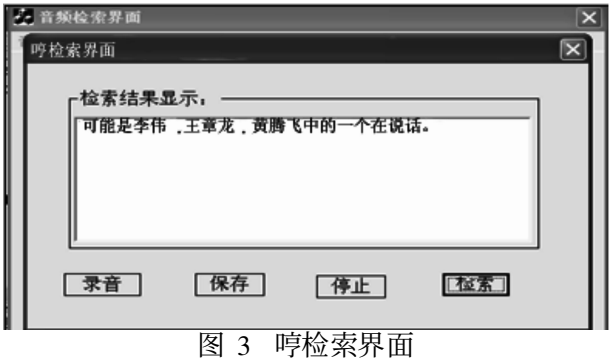

图 3　哼检索界面

在“哼”检索过程中,用到的基本的类以及类之间的相互关联如图 4.

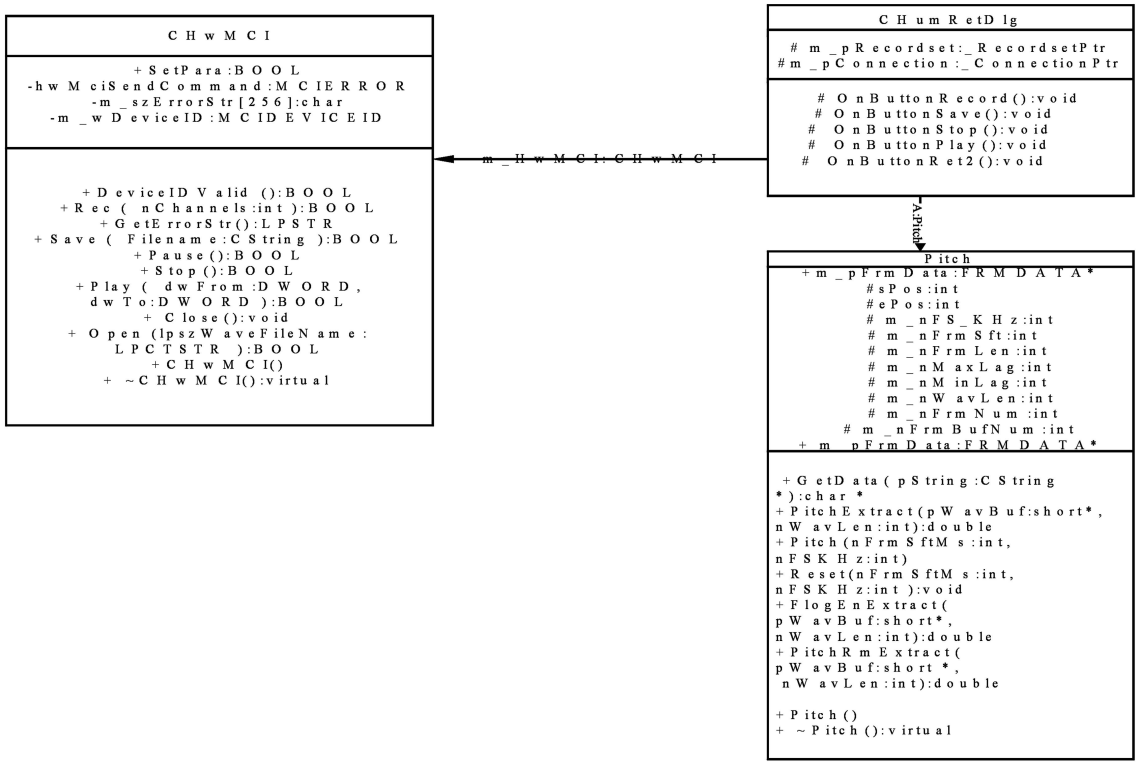

图 4 类与类间关系图

3 结束语

基于内容的音频检索是一个新兴的研究领域,在国内外仍处于研究、探索阶段.当今时代,随着现代信息技术的发展,多媒体信息可以说是无处不在,但是由于多媒体类型丰富,数据量大等特点,使得如何能高速的检索就显得尤其重要.在本论文中介绍了一种基于内容得音频检索方法,并提供了一个简单的测试系统,该系统能很好的用于检索测试,但用于实际的应用还需要进一步加强和改进.

参考文献:

[1]李恒峰,李国辉.基于内容的音频检索与分类[J].计算机工程与应用,2000,54~56.

[2]李国辉,李恒峰.基于内容的音频检索:概念和方法[J].小型微型计算机系统,2000,1173~1177.

[3] Jonathan T. Foote In C.-C. J. Kuo et al. Content-Based Retrieval of Music and Audio[J]. Multimedia Storage and Archiving Systems II, Proc. of SPIE, 1997, 3229: 138~147.

[4]黄志军,曾斌.多媒体数据库技术[M].北京:国防工业出版社,2005.

[5]李国辉等.基于内容的检索[N].计算机世界专题,1998—05—26.