第 37 卷 第 24 期
Vol.37 No.24

计 算 机 工 程
Computer Engineering

2011 年 12 月
December 2011

·开发研究与设计技术· 文章编号：1000—3428(2011)24—0281—03 文献标识码：A 中图分类号：TP391

# 面向音频指纹的帕尔森高斯核量化哈希方法

陈海浪[a,b]，欧阳建权[a,b]

(湘潭大学 a. 智能计算与信息处理教育部重点实验室；b. 信息工程学院，湖南 湘潭 411105)

**摘 要：**基于二进制哈希的音频指纹匹配方法鲁棒性较差。为此，提出一种帕尔森高斯核量化哈希方法，将 2 个音频中间哈希值的差异度通过概率密度函数量化编码到一个合适的整数范围内，以刻画失真的概率分布，实现音频指纹的提取。实验结果表明，与二值哈希法相比，该方法对多种失真具有更高的鲁棒性。

**关键词：**音频指纹；指纹提取；量化哈希；帕尔森高斯核；HFM 方法

# Parzen Gaussian Kernel Quantum Hash Method Orienting to Audio Fingerprint

CHEN Hai-lang[a,b], OUYANG Jian-quan[a,b]

(a. Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education;
b. College of Information Engineering, Xiangtan University, Xiangtan 411105, China)

**【Abstract】**Aiming at the poor robustness problem of the binary hash for the audio fingerprint matching technology, this paper proposes a Parzen Gaussian kernel quantum hash scheme, which encodes the intermediate hash difference between two audio contents into an integer at a suitable range according to its probability density function, and characters the probability distribution of distribution. Experimental results show that the method is more robust under various distortions than the binary hash.

**【Key words】**audio fingerprint; fingerprint extraction; quantum hash; Parzen Gaussian kernel; HFM method

**DOI:** 10.3969/j.issn.1000-3428.2011.24.094

## 1 概述

音频指纹是音频对象简短的摘要[1]，基于内容的音频指纹可以应用于版权保护、音频检索等，随着互联网上音频数据跨越式的发展，海量音频数据的检索和鉴别亟需高效的音频指纹匹配技术。目前，音频指纹匹配大多是基于二进制哈希的，即从音频内容提取出来的特征值通过域值判断直接编码成 0 或 1，并用海明距离表示两音频间的差异。为叙述方便，本文将从音频内容中提取出的特征实值称为中间哈希值。鲁棒性和可靠性是衡量一个音频指纹系统性能的重要指标[2]。系统的鲁棒性是指当音频经过一定的失真处理后仍然能够被识别的能力，可以从两方面来提高：(1)特征的选择。特征的选择直接决定了系统的鲁棒性[3]，但这种鲁棒性具有一定的局限性，即不能抵抗所有失真类型。如在文献[2]中，中间哈希值为子带能量差，该特征对多种失真类型如重采样、环境噪声、多种 MP3 有损压缩、带通过滤等具有较好的鲁棒性，但对线性速度或音调变化幅度超过 2%以及较严重的噪声如人的吵闹声等失真，鲁棒性不高。在文献[4]中，中间哈希值为一阶归一化频域子带矩，该特征对均衡化、噪声、电话通道过滤等具有较好的鲁棒性，但对回声及线性速度变化等鲁棒性不高。(2)指纹的匹配。如文献[3]使用一阶归一化频域子带矩[4]作为特征值，结合一种改进的 AdaBoost 算法来减小指纹值二值化过程中的失真，与传统的简单域值式差异二值法相比，系统的鲁棒性有所提高，但二进制编码将实值转换为 0、1 的过程所带来的实质性信息损失并没得到改善。这些音频指纹应用都基于二进制编码[1-4]。文献[5]首次将量化哈希应用于音频检索中，结合查询音频的中间哈希值，将数据库中候选音频二值指纹值为 0 或 1 的可能性编码为量化距离，在各种严重失真条件下，系统的鲁棒性相对二进制哈希法有明显的提高，但其量化过程是基于将两音频间特征值差异过程模拟为正态分布，而绝大多数失真类型所产生的特征值差异并不服从正态分布[5]，因此，不具一般性。鉴于此，本文采用文献[2]的能量差异值作为音频特征，通过更具一般性的非参数估计法——帕尔森高斯核概率密度函数估计法估算两音频间的距离并做出判断。

## 2 基于量化哈希的指纹提取

文献[2]的音频指纹模型是音频指纹领域的经典，也是许多商业应用的原型，其音频特征子带能量差具有区分性高(指纹值差异与歌曲差异呈正比)、粒度小(查询音频以正确识别数据库中音频记录的长度，通常要求 10 s 以下)等优点，并具有一定的鲁棒性，对失真类型如均衡化、回声、环境噪声及各种音频压缩等具有良好的鲁棒性，但在比较强的失真条件下如人的吵闹声及线性速度或音调变化等，鲁棒性不高。对此，在使用相同音频特征子带能量差的条件下，本文提出一种量化哈希方法，并结合修改后的音频指纹，以提高该系统在所有失真条件下的鲁棒性，使其更具实用性。

### 2.1 音频指纹

指纹提取过程如图 1 所示。通过将每帧的音频内容用快速傅里叶变换转换到频域，并选择与人的听觉感知最相关的频率范围(300 Hz~3 000 Hz)，将其按对数距离(巴克距离)分割

**基金项目：**国家科技支撑计划基金资助项目(2007BAH14B05)

**作者简介：**陈海浪(1986－)，女，硕士研究生，主研方向：多媒体技术；欧阳建权，教授

**收稿日期：**2011-06-01 **E-mail：**kissingman1@gmail.com

成 33 个子带，提取 32 个频域的能量差异值，用 $E_m[n]$表示第 $n$ 帧第 $m$ 个子带的能量，则第 $n$ 帧第 $m$ 个能量差异值 $d_m[n]$的定义如下：

$$d_m[n]=E_m[n]-E_{m+1}[n]-(E_m[n-1]-E_{m+1}[n-1]) \tag{1}$$

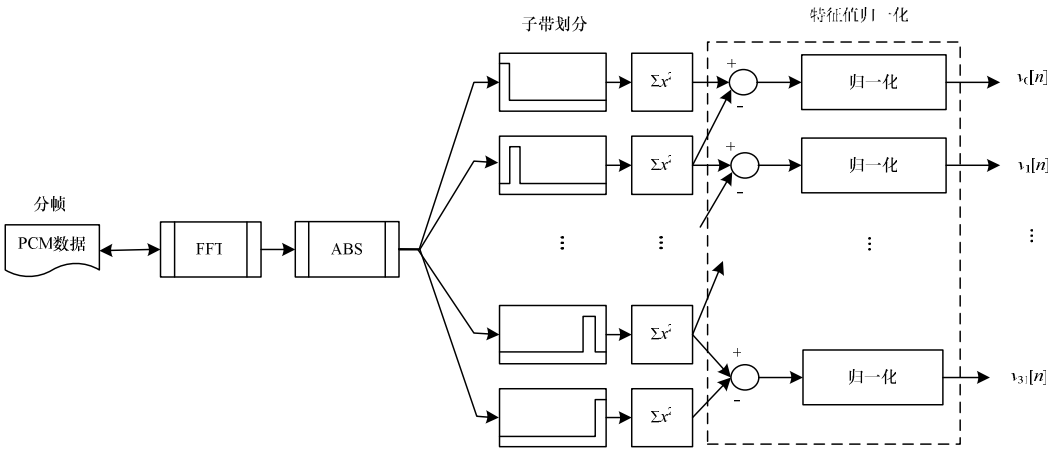


**图 1 音频指纹提取过程**

与文献[2]中直接进行二进制哈希编码不同，本文对该特征值块进行归一化处理，使特征值都处于一定的以 0 为中心的实数范围内，为方便叙述，本文称这种指纹提取方法为 HFM。

归一化过程如下：

$$v_m[n]=\frac{d_m[n]-\hat{\mu}_m}{\hat{\delta}_m} \tag{2}$$

其中：

$$\hat{\mu}_m=\frac{1}{N}\sum_{m=0}^{N-1}d_m[n] \tag{3}$$

$$\hat{\delta}_d=\sqrt{\frac{1}{N}\sum_{n=0}^{N-1}d_m^2[n]-\hat{\mu}_m^2} \tag{4}$$

本文将这样从 256 帧中提取出来的每帧 32 位的指纹序列称为一个中间哈希块，其表示为：

$$\boldsymbol{v}_N=[v[0],v[1],\cdots,v[N-1]]^{\mathrm{T}},N=256 \tag{5}$$

其中：

$$v[n]=[v_0[n],v_1[n],\cdots,v_{M-1}[n]],n=0,1,\cdots,N-1,M=32 \tag{6}$$

### 2.2 二进制哈希

许多传统的指纹提取系统都是基于二进制哈希[1-4]，并用海明距离来表示两音频间的差异程度。设 $\boldsymbol{c}[n]\in B^M$ 表示从 $v[n]\in\mathbf{R}^M$ 编码而来的二值向量，$B=\{0,1\}$。$\boldsymbol{c}_m[n]\in B$ 为 $\boldsymbol{c}[n]$的第 $m$ 个元素，$k$ 是一个预设的域值，实验中 $k$=0，$v_m[n]$编码为 $\boldsymbol{c}_m[n]$如下：

$$\boldsymbol{c}_m[n]=\begin{cases}1 & \text{if } v_m[n]>k\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

当两音频指纹块对应的指纹位相同时，距离为 1，否则为 0，那么，两音频间的海明距离即哈希位距离为 1 的个数。

### 2.3 量化哈希

由式(9)可看出，二进制哈希编码将一个实值转换为一个二进制值的过程会导致信息丢失，即忽略了音频失真程度对指纹值的影响，而量化哈希具体考量了这种影响程度，将两音频的中间哈希值差异即失真程度通过概率密度函数量化编码到一个合适的整数范围内，所得对应哈希位整数距离之和即为两音频间的量化距离，失真程度越大，量化距离越大，反之越小。本文使用与文献[2]相同的二进制哈希数据库，不同之处主要在于两音频间距离的计算。其指纹提取过程如图 2 所示，当输入一个查询音频片段，系统将通过 HFM 指纹提取模块得到与候选音频的中间哈希差异块(256×32)，并通过帕尔森窗口高斯核概率密度函数进行量化编码，得出与候选音频间的量化距离，当量化距离小于预设距离域值时，可接受两音频为匹配对象，反之，拒绝为不匹配对象。

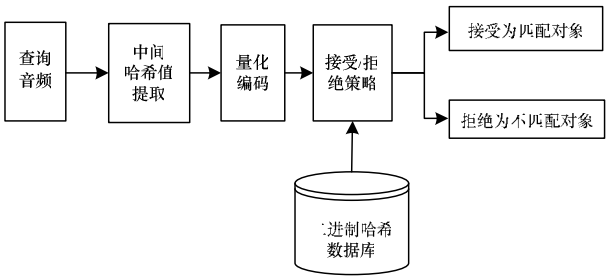


**图 2 量化哈希指纹提取过程**

#### 2.3.1 量化编码

查询音频经过量化编码后第 $n$ 帧的第 $m$ 个量化哈希位表示如下：

$$q_m[n]=\psi_m^-[n]|0\rangle+\psi_m^+[n]|1\rangle\in H^2 \tag{8}$$

其中，$|\psi_m^-[n]|^2$ 和 $|\psi_m^+[n]|^2$ 为候选音频指纹值为 0 或 1 的可能性。

通过查询音频与候选音频间中间哈希差异值的概率密度函数来表示，假设中间哈希差异值数据集合服从正态分布[5]，其概率密度函数如下：

$$p(x)=\frac{1}{\sqrt{2\pi\delta^2}}\exp(\frac{-(x-\mu)^2}{2\delta^2}) \tag{9}$$

其中，$\mu$ 为均值；$\delta$ 为均方差。

本文中的量化哈希位借助于量子物理学中量子位元标记法[6]，将一个 Hibert 空间 $H^2$ 中的量子位用数学表示如下：

$$|x\rangle=\psi_x^-|0\rangle+\psi_x^+|1\rangle\in H^2 \tag{10}$$

$$|\psi_x^-|^2+|\psi_x^+|^2=1 \tag{11}$$

其中，$|\psi_x^-|^2$ 和 $|\psi_x^+|^2$ 为 $|0\rangle$ 或 $|1\rangle$ 在 $|x\rangle$ 中的概率。

由于中间哈希差异值数据集合极少服从正态分布[5]，因此本文采用更具一般性的非参数估计法——帕尔森高斯核估计法来估算查询音频与候选音频间中间哈希差异的概率概率密度函数：

$$\overline{p}(x)=\frac{1}{K\sqrt{2\pi h^2}}\sum_{n=0}^{N-1}\exp(\frac{-(x-x_i)^2}{2h^2}) \tag{12}$$

其中，$i$=0, 1,$\cdots$, $N$−1；$h$ 为带宽。

2.3.2　距离计算

设 $\boldsymbol{c}[j]\in B^M$ 表示二进制哈希指纹数据库中的第 $j$ 个向量：

$$\boldsymbol{c}[j]=[\boldsymbol{c}_0[j],\boldsymbol{c}_1[j],\cdots,\boldsymbol{c}_{M-1}[j]]^{\mathrm{T}},j=0,1,\cdots,J-1 \tag{13}$$

其中，$J$ 是数据库中二值哈希向量的数目。

基于前述的量化哈希表示法，定义查询音频量化哈希位 $q_m[n]\in H^2$ 和数据库候选音频指纹二进制哈希位 $\boldsymbol{c}_m[j]\in B$ 的距离 $d_l$ 如下：

$$d_l(q_m[n],\boldsymbol{c}_m[j])=\begin{cases}|\psi_m^+[n]|^l & \text{if } \boldsymbol{c}_m[j]=0\\ |\psi_m^-[n]|^l & \text{otherwise}\end{cases} \tag{14}$$

其中，$l$ 为一个预设的参数，本实验中 $l$=0.8。

将距离 $d_l(q_m[n],\boldsymbol{c}_m[j])$ 编码到一个合适的整数范围内：

$$\widetilde{d}_l(q_m[n],\boldsymbol{c}_m[j])=\lfloor\beta d_l(q_m[n],\boldsymbol{c}_m[j])\rfloor \tag{15}$$

其中，$\beta$ 是一个预设的缩放比例参数，本文中 $\beta$=40；$\lfloor x\rfloor$ 表示不大于 $x$ 的最大整数。

设 $Q$ 为查询音频通过 HFM 提取出来的 32×256 中间哈希指纹块，$C$ 为数据库中相同大小的候选音频二进制哈希块，则 $Q$ 与 $C$ 间的距离可定义如下：

$$\Gamma_l(Q,C)=\beta^{-1}\sum_{i=0}^{MN-1}\widetilde{d}_{l,\beta}(Q[i],C[i]) \tag{16}$$

当 $\Gamma_l(Q,C)$ 小于预设的域值 $\tau$ 时，可判定两者是可接受为匹配的，反之，为不匹配。

## 3　实验分析比较

实验数据由 964 首不同风格的歌曲组成(立体声、44.1 KHz、16 位采样)，其中，经典 128 首；摇滚 114 首；爵士 145 首；乡村 135 首；hip-pop 108 首；R&B 178 首；蓝调 156 首，平均每首歌时长约 5 min，总时长约 70 h。实验中用 10 000 个匹配对和 1 000 000 个不匹配对来进行检验，每个匹配对约 3.33 s。计算机配置：Pentium D 3.0 GHz CPU、1 GB RAM，编程工具是 Matlab 7.0，所使用的音频处理工具为 Cool Edit Pro 2.1。

### 3.1　鲁棒性比较

为模拟实际应用环境，本文对音频作以下的失真处理：(1)D1：均衡化+96 Kb/s 有损 MP3 压缩；(2)D2：回声+96 Kb/s 有损 MP3 压缩；(3)D3：环境噪声+96 Kb/s 有损 MP3 压缩；(4)D4：人吵闹声+96 Kb/s 有损 MP3 压缩；(5)D5：32 Kb/s；(6)D6：均衡化+环境噪声+回声+96 Kb/s 有损 MP3 压缩；(7)D7：回声+升调 2%+92.5 ms 时延+96 Kb/s 有损 MP3 压缩；(8)D8：均衡化+高斯噪声+回声+升调 2%+92.5 ms 时延+96 Kb/s 有损 MP3 压缩。本文用误码率(BER)来衡量失真类型 D1~D8 的失真程度，BER 是指音频片段经过失真处理后用二值化哈希方法编码的指纹值与原始音频指纹值不一样的个数在指纹块中的比率(误码率=误码数/总码数×100%)，实验中的 BER 值是由数据库中各种风格歌曲 10 个共 70 个匹配对分别通过失真类型 D1~D8 处理后的 BER 平均值。

实验采用 EER(Equal Error Rate)作为评测标准，EER 定义为误检率和漏检率的平衡点。误检率为匹配对判定为不匹配的概率，漏检率为不匹配对判定为匹配的概率。误检率等于或者最接近漏检率时的概率即 EER。本文方法与文献[2]的经典二值化方法以及正态式量化哈希[5]的 EER 比较结果如表 1 所示。从表 1 中可以看出，HFM 对均衡化、回声、环境噪声及各种音频压缩都具有良好的鲁棒性[2]，因此，对于失真类型 D1、D2、D3、D5，本文方法性能的提高并不太明显，当查询音频受到比较强的失真处理，如 D4、D6、D7、D8，特别是盲检 D8 时，本文方法的性能明显优于二进制哈希方法，且对所有失真类型，本文方法的性能都优于正态量化哈希，说明本文方法更具实用性。

**表 1　3 种方法的 BER 比较结果**

| 失真类型 | BER | EER/(%) | | |
|---|---|---|---|---|
| | | 二进制 | 正态量化 | 本文方法 |
| D1 | 0.147 | 9.69E-05 | 9.17E-05 | 6.97E-05 |
| D2 | 0.203 | 5.97E-04 | 9.35E-05 | 1.27E-05 |
| D3 | 0.154 | 1.10E-05 | 6.39E-05 | 3.31E-06 |
| D4 | 0.385 | 9.40E-02 | 0.51E-02 | 2.28E-03 |
| D5 | 0.170 | 8.97E-05 | 1.74E-04 | 4.39E-05 |
| D6 | 0.431 | 1.31E+00 | 7.71E-01 | 6.53E-02 |
| D7 | 0.490 | 8.16E+01 | 5.33E+00 | 1.27E-01 |
| D8 | 0.591 | 1.31E+03 | 4.36E+02 | 7.10E+00 |

### 3.2　时间复杂度比较

本文方法与正态量化哈希方法[5]相比，时间复杂度的区别主要在于式(9)与式(12)的开销，都为 $O(N)$，2 种方法的复杂度差不多；与二进制哈希方法相比，本文方法所增加的时间开销集中于两音频哈希位间的量化距离 $|\psi_m^+[n]|^l$ 和 $|\psi_m^-[n]|^l$ 的计算：

(1)每个子带中间哈希向量的概率密度函数的计算，其时间复杂度为 $O(N)$。

(2)对概率密度函数进行积分运算以得到量化距离，其时间复杂度为 $O(N\log N)$。

相比二进制哈希法，两音频间量化哈希所增加的时间复杂度则为 $O(MN^2\log N)$，其中，$N$ 为帧数；$M$ 为每一帧的维度。因此，本文方法在音频实时检测方面还有待改善。

## 4　结束语

本文将量化哈希应用于音频检测中，以子带能量差作为音频特征，将查询音频的失真程度通过概率密度函数进行量化编码，并由此得出与数据库中候选音频间的量化距离。实验结果证明，在查询音频失真越严重的条件下，本文方法较传统二值哈希系统鲁棒性得到明显提高，后继工作将着力于对时间复杂度以及实时处理的改善，从而将其应用于广告监测应用。

## 参考文献

[1] 张　敏, 欧阳建权, 李泽洲, 等. 一种快速的特定音频指纹提取方法[J]. 计算机工程, 2010, 36(1): 211-213.

[2] Haitsma J, Kalker T. A Highly Robust Audio Fingerprinting System[C]//Proc. of International Conference on Music Information Retrieval. Paris, France: ISMIR Press, 2002: 107-115.

[3] Kim S, Yoo C D. Boosted Binary Audio Fingerprint Based on Spectral Subband Moment[C]//Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing. Honolulu, USA: [s. n.], 2007: 241-244.

[4] Jin Seo, Jin Minho, Lee Sunil, et al. Audio Fingerprinting Based on Normalized Spectral Subband Moments[J]. IEEE Signal Processing, 2006, 13(4): 209-212.

[5] Jin Minho, Yoo Chang Dong. Quantum Hashing for Multimedia[J]. IEEE Transactions on Information Forensics and Security, 2009, 4(4): 982-994.

[6] Bennett C H, Shor P W. Quantum Information Theory[J]. IEEE Transactions on Information Theory, 1998, 44(6): 2724-2742.

编辑　张　帆