文章编号:1001 - 2486(2004)06 - 0049 - 05

辅助足球视频切分的音频自动分类与分段*

陈剑赟,李云浩,吴玲达,老松扬,白 亮
(国防科技大学人文与管理学院,湖南 长沙 410073)

摘 要:视频伴随音轨的自动分类与分段是辅助视频切分的一种有效手段。从足球视频的特征入手,归纳总结出足球视频中三类主要的音频类型,既而提出了基于 HMM 并且结合一定平滑策略的音频自动分类和分段的框架,在实现音频分类分段的同时完成了足球视频的切分。初步的实验结果验证了该方法的有效性和鲁棒性。

关键词:音频分类与分段;HMM;足球视频;视频切分
中图分类号:TN941 文献标识码:A

Automatic Audio Classification and Segmentation for Soccer Video Structuring

CHEN Jian-yun, LI Yun-hao, WU Ling-da, LAO Song-yang, BAI Liang
(College of Humanities and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract: Automatic classification and segmentation of the sound track is an effective approach for video structuring. We use this method to parse soccer video. Firstly, based on the characteristics of soccer video, the paper summarizes three audio classes for soccer video, namely game-audio, advertisement-audio and studio-audio. Then it proposes a framework of audio classification and segmentation using Hidden Markov Model and combining some smoothing rules. We develop a 26-coefficients feature stream for HMM model. And the experimental studies indicate that the proposed framework is effective and robust.

Key words: audio classification and segmentation; HMM; soccer video; video structuring

足球视频是一种具有广泛群众基础的视频类型。研究足球视频的内容分析具有重大的现实意义。先前的研究主要从句法分段和语义标注两条途径达到体育视频内容分析的目标[1~7]。但是,随着研究的深入,人们发现,如果仅仅考虑图像序列的视觉特征,例如颜色、纹理、运动矢量等,并不能全面地表征体育视频的内容,需要融合音频、字幕等其他特征来完成体育视频内容分析的目标。本文旨在讨论通过伴随音轨的自动分类与分段来辅助完成足球视频的切分。本文从足球视频的特征入手,总结出足球视频中常见的几种音频类型,提出了基于 HMM 的音频自动分类的算法;然后采用一定的平滑策略,在音频分类的基础上找到音频的分隔点,以达到最终的足球视频切分的目标。

1 辅助足球视频切分的音频自动分类与分段的框架图

图 1 是足球视频的一种典型的结构图。整段足球视频由相续的五个部分组成:比赛→广告→演

图 1 一种典型的足球视频结构

Fig. 1 Typical sequence of segments in soccer video

* 收稿日期:2004 - 06 - 28
基金项目:国家自然科学基金资助项目(60473117)
作者简介:陈剑赟(1977—),女,博士生。

播室→广告→比赛。事实上,这种结构对应于伴随音轨的结构:比赛音频→广告音频→演播室音频→广告音频→比赛音频。易见,不同的足球视频的伴随音轨都是由这三类音频组成的,所不同的只是这三种音频的组织结构。所以说,只要完成了伴随音轨的分类与分段,就可以辅助完成视频的切分。因此,本文抽取出足球视频中这三类主要的音频类型:比赛音频、广告音频和演播室音频。

HMM 是一种双随机过程的有限状态自动机,它具有刻画信号时序统计特性的能力[8],并且 HMM 算法成熟,运算效率较高,通用性和鲁棒性较强,所以本文采用 HMM 作为体育视频音频分类算法。系统框架如图 2 所示。从图 2 可见,整个系统主要包括训练阶段与决策阶段两大部分。在训练阶段,首先从足球视频中获得三类音频的训练数据,然后抽取训练数据的特征,采用 Baum-Welch 算法分别构建三类音频的 HMM 模型。在决策阶段,对于测试数据,按照同训练数据完全相同的方法和参数设置抽取音频特征流,然后以极大似然判别准则实现音频的分类,即对于测试的基本单元,分别计算对应每个 HMM 模型的概率值,并选取似然最大的模型作为该测试数据的类别,从而得到音频类型判别序列,最后对类型序列进行平滑,找到分割点,从而完成了音频的分段。以下详细讨论框架图(图 2)中的各个功能模块。

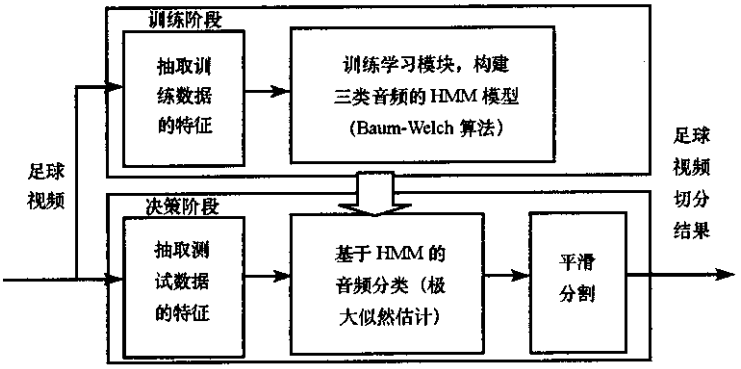

图 2 基于 HMM 的足球视频伴随音轨自动分类与分段的框架

Fig. 2 The HMM-based framework of automatic audio classification and segmentation for soccer video structuring

2 基于 HMM 的足球视频伴随音轨的自动分类与分段

2.1 音频特征提取

根据音频的短时特性[9],本文认为在极短的时间内,例如音频的 1 帧,讨论该帧属于什么类型的音频根本就没有意义,并且在极短的时间内,人也无法辨识该帧属于什么类型。所以,本文首先将音频分割成长度为 1s 的 clip,相邻 clip 之间有 0.5s 的重叠部分,把 clip 作为音频训练和测试用的基本单元。在每个 clip 中加 Hamming 窗形成帧,每个 clip 就是一个帧序列。对于每帧,抽取如下的音频特征[9]:

- 短时平均幅度;
- 短时过零率;
- MFCC 参数以及 MFCC 差分:MFCC(Mel-frequency Cepstral Coefficient,简称 MFCC)的分析着眼于人耳的听觉特征,因为人耳所听到的声音的高低与声音的频率并不成线性正比关系,而用 Mel 频率尺度则更加符合人耳的听觉特征,突出了有利于识别的信息;并且 MFCC 参数无任何前提假设,在各种情况下均可使用,通用性、鲁棒性强,适合作为音频分类分段的特征。本文采用 12 阶的 MFCC 系数。

所以,最终提取的特征参数为 26 维的特征矢量,包括 12 维 MFCC 参数、12 维 MFCC 差分参数、1 维的短时平均幅度和 1 维的短时过零率。

2.2 HMM 的训练

HMM 用三元组 $\lambda = (A, B, \pi)$ 来描述,其中 A 表示状态转移概率矩阵;B 表示观测信号概率密度;π 表示初始状态概率矢量。当观测信号取为连续矢量时,B 用椭圆对称或者对数凸概率密度函数的

有限混合来表示,一般采用高斯概率密度函数表示。此时的 HMM 称为连续的隐马尔可夫模型。

由于用于音频分类与分段的 26 维的特征矢量是连续变量,本文对三类音频类型构建连续隐马尔可夫模型,其中 HMM 的拓扑结构采用状态数 N 为 3 的遍历模型,观测信号概率密度函数采用高斯函数,混合项数 M 取 15。对于模型的训练,本文采用 Baum-Welch 算法,Baum-Welch 算法能从理论上证明经过有限次迭代后能够收敛,但是 Baum-Welch 算法和梯度算法一样都有可能陷入局部极值点,而不能得到全局最优的结果。

HMM 模型训练过程中的初值选择是一个非常重要的问题,好的初值选择可以保证达到收敛所需的迭代次数最少,即计算效率最高。本文采用经典的分段 K 平均算法(Segmental K-means Procedure)[8] 选取初值。对于训练过程中的下溢问题,本文的算法在迭代运算中乘以放大比例因子,使得每一步递推值维持在相似的水平。

2.3 测试音频的自动分类与分段

在训练得到三类音频的 HMM 模型之后,以极大似然准则判定测试音频的类型 给定一个 clip,按照与训练数据完全相同的方法和参数设置计算其特征序列,然后分别计算对应于三类 HMM 模型的概率值,并选取似然最大的模型作为该 clip 的类别[8]。经过分类之后,整段测试音频得到一个类型序列,然后对该类型序列进行平滑。

考虑到音频的短时特性,所以音频类型突然改变的可能性比较小。基于此假设,采用如下的平滑规则[10]

$$\text{Rule if}(c[1]\neq c[0]\ \&\&\ c[2]=c[0])\ \text{then}\ c[1]=c[0]$$

在此考虑三个连续 clip 的音频类型,$c[1]$,$c[0]$,$c[2]$ 分别表示当前 clip、前一个 clip 和下一个 clip 所属的音频类别。此平滑规则表明 如果当前 clip 的类别不同于前后 clip 的类别,并且前后 clip 的类别一致,则当前 clip 的类别判断一定是误判。

经过平滑处理后,得到一个分类精度得以提高的音频类型判别序列,易见,音频类型改变点就是音频的分割点,即视频的切分点。

3 实验结果及讨论

表 1 基于 HMM 的音频分类分段的训练与测试数据

Tab. 1 Experimental data of audio classification and segmentation

序号	名称	长度	来源
Soccer1	英超曼联队对阿斯维拉队比赛片段	24min38s	2003 – 3 – 15 采集自湖南卫视体育频道
Soccer2	英超曼联队对富尔汉姆队比赛片段	25min2s	2003 – 3 – 22 采集自湖南卫视体育频道
Soccer3	德甲拜仁慕尼黑队对勒沃库森队比赛片段	10min53s	2003 – 9 – 20 采集自湖南卫视体育频道

为了验证基于 HMM 的音频分类分段模型的有效性,选取若干段足球视频作为实验数据,如表 1 所示。每段足球视频的伴随音轨采用 16kHz 采样率,精度为 16bit。对于音频分类分段算法的评估,在此借鉴文献[2]的度量方法。选取一段足球视频作为训练数据,将剩余的足球视频作为测试数据,连带把本次的训练数据也作为测试数据进行了检验,这样交叉进行三次。以分类准确程度作为衡量音频分类好坏的指标,各个音频类别的分类精度 $ClaAccurary_i$ 定义如下:

$$ClaAccurary_i = \text{正确预测为 } i \text{ 类的 clip 数目} / \text{预测为 } i \text{ 类的 clip 数目}$$

得到基于 HMM 的足球视频伴随音轨的分类结果如表 2 所示。其中,表的最右边一列表示每行的平均分类精度,即测试数据一定时的平均分类精度 ;最下面一行表示每列的平均分类精度,即训练数据一定时的平均分类精度 ;而表的右下角的数值表示总的平均分类精度。从表中易见,实验结果比较令人满意。

表 2 基于 HMM 的足球视频伴随音轨的分类实验结果

Tab. 2 Experimental results of audio classification

测试数据		训练数据			平均精度
		Soccer1	Soccer2	Soccer3	
Soccer1	比赛音频	0.9196	0.8228	0.7998	0.8474
	广告音频	0.9310	0.8598	0.8964	0.8957
	演播室音频	0.8646	0.8917	0.88	0.8788
Soccer2	比赛音频	0.8420	0.9891	0.8758	0.9023
	广告音频	0.9400	0.9626	0.8598	0.9208
	演播室音频	0.8486	0.7810	0.8333	0.8210
Soccer3	比赛音频	0.8793	0.8470	0.9347	0.8870
	广告音频	0.8816	0.9075	0.9208	0.9033
	演播室音频	0.8114	0.8767	0.9066	0.8649
平均精度		0.8798	0.8820	0.8756	0.880

既然在分类的同时实现了音频的分割,有必要讨论一下音频分割以及最终视频切分的效果。图 3 是以 Soccer1 作为训练数据、Soccer2 作为测试数据得到的音频类型序列,其中横轴表示 clip 的编号,纵轴表示音频类型:1 代表演播室音频,2 代表广告音频,3 代表比赛音频。从音频类型序列易见,视频开始的时候是比赛;然后插播了广告;然后是一段演播室和比赛频繁切换的视频,这恰好与事实相符:因为当演播室主持人评述前一段比赛时,导演及时插入了当时比赛时的情形;接着是一段比较稳定的演播室视频;然后是广告视频;最后又进入了比赛。显然,图 3 中的阶跃点就是音频的分割点,也就是视频的切分点。所以,通过音频的分类分段,达到了辅助视频切分的目标,事实上也分析获得了该段视频的结构。

表 3 音频分割点偏差的分布

Tab. 3 Segmentation-point-offset distribution

分隔点偏差(Clip)	[0,5)	[6,10)	[11,15)	[16,20)	≥20
百分比(%)	73%	5%	8%	6%	8%

本文在对 clip 进行分类的同时,也同时考虑了该 clip 是否包含有分割点,这区别于传统的句法分段的方法。统计了表 2 中在不同的训练数据和测试数据情况下音频分割点偏差的分布情况,以分割点偏差的分布来衡量音频分割的优劣,偏差以 clip 来计算,结果如表 3 所示。从表中可见,分割点偏差在 0 ~ 5 个 clip 的百分比占到 70% 以上,结果还是比较令人满意的。

4 总结与展望

结合足球视频的基本特点,将足球视频的音频分为比赛、广告和演播室三类音频;然后提出了基于 HMM 的通用的音频自动分类模型;并且在分类基础上实行一定的平滑策略,同时实现音频的分段,辅助完成足球视频的切分。初步的试验结果表明该模型的有效性和鲁棒性,并且本文的思路也同样适用于其他的体育视频类型,例如网球、排球等,甚至可以用于其他的视频类型,以实现视频的切分和结构分析。

在接下来的工作中,将进一步提高算法的精度,一方面深入研究音频特征的提取,到底提取什么特征更加有效,更加能够表征不同音频类型的特征;另一方面需要充分考虑 HMM 模型的拓扑结构以及参数设置。目前还没有从理论上确定最优的 HMM 拓扑结构和参数设置,采用了一些经验值。虽然结果还比较令人满意,但是可以考虑 HMM 模型的多种拓扑结构以及多种参数设置,以获得更高的实验精度。此外,还可以考虑融合音频特征和图像序列特征完成视频内容分析的目标。例如,在图 3

中,第 1500 个 clip 到第 2000 个 clip 之间,演播室音频和比赛音频频繁地切换,可以结合考虑结合图像序列的颜色信息完成视频切分和结构分析,以获得更加完美的实验结果。

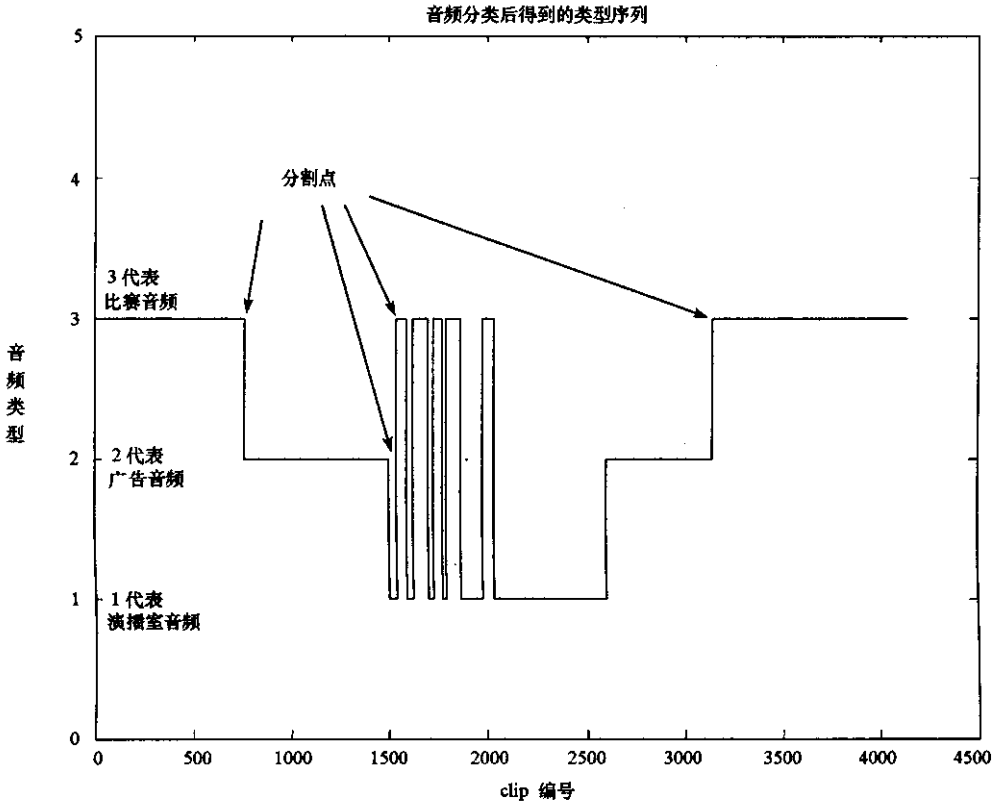

图 3 音频分类后得到的类型序列

Fig. 3 The final classification and segmentation sequence with soccer2 as testing data and soccer1 as training data

参 考 文 献:

[1] Zhong Z, Chang S F. Structure Analysis of Sports Video Domain Models [A]. In IEEE Conference on Multimedia and Expo., 2001 920 – 923.

[2] Xie L, Chang S F, Divakaran A, et al. Structure Analysis of Soccer Video with Hidden Markov Models [A]. In IEEE International Conference on Acoustics Speech and Signal Processing, Orlando, FL., 2002.

[3] Gong Y H, Sin L T, Chuan C H, et al. Automatic Parsing of TV Soccer Programs [A]. In IEEE International Conference on Multimedia Computing and Systems, Washington D. C., May, 1995.

[4] James A, Chang S F. Automatic Selection of Visual Features and Classifiers [A]. In SPIE Conference on Storage and Retrieval for Media Database, 2000, 3972 346 – 358.

[5] Rui Y, Gupta A, Acero A. Automatically Extracting Highlights for TV Baseball Programs [A]. In ACM International Conference on Multimedia, 2000 105 – 115.

[6] Sudhir G, Lee J C M, Jain A K. Automatic Classification of Tennis Video for High-level Content-based Retrieval [A]. In IEEE International Workshop on Content-based Access of Image and Video Database, Bombay, India, Jan., 1998.

[7] 熊华. 视频内容结构化技术的研究与实现[D]. 国防科技大学博士论文, 2001.

[8] Rabiner L, Juang B H. Theory and Implementation of Hidden Markov Models [M]. Book Chapter, Fundamentals of Speech Recognition, Prentice Hall, 1993.

[9] 杨行峻. 语音信号数字处理 [M]. 北京:电子工业出版社, 1995.

[10] Lu L, Zhang H J, Jiang H. Content Analysis for Audio Classification and Segmentation [J]. IEEE Transactions on Speech and Audio Processing, 2002, (7).