基于多模态融合的足球视频语义分析

张玉珍 魏带娣 王建宇 戴跃伟
(南京理工大学自动化学院 南京 210094)

摘 要 提出一种有效地融合多模态信息来检测足球视频精彩事件的系统算法和框架。首先从视频中抽取音频流,然后基于CHMM进行音频分类。接着根据时间对应关系在包含激昂解说音和欢呼声的相邻镜头里结合球门和慢镜头检测射门事件,其中慢镜头检测是基于徽标的。对射门事件进一步根据激昂解说音和欢呼声的长短、慢镜头的长短及比分字幕的出现检测进球事件。在哨音出现的相邻镜头中结合是否有慢镜头回放及回放长度来检测犯规事件。实验表明,提出的系统算法及框架是高效率的。

关键词 多模态融合,音频分类,徽标,慢镜头,球门
中图法分类号 TP391 **文献标识码** A

Semantic Analysis for Soccer Video Based on Fusion of Multimodal Features

ZHANG Yu-zhen WEI Dai-di WANG Jian-yu DAI Yue-wei
(School of Automation, Nanjing University of Science&Technology, Nanjing 210094, China)

Abstract This paper proposed a framework to fuse multimodal features to detect soccer highlights. First the audio stream was extracted from video and classified based on CHMM. Then according to time corresponding relationship, shoot event was detected based on the combination of goal and replay in the shots near to those including excited speech of commenter and cheer from audience, where replay was detected based on logos. For shoots scoring could be judged according to the length of excited speech and cheer and the one of replay and the caption appearance. In the shots close to those including whistles fouls could be detected based on the combination of replay appearance and the length of replay. Experiments prove the high efficiency of the proposed system.

Keywords Fusion of multimodal features, Audio classification, Logo, Slow-motion replay, Goal

1 引言

视频语义分析一直是视频研究的热点与难点。足球等体育视频往往由于场地和摄像机数量的限制,具有相对的结构性,而且广受欢迎,因此研究得比较多。体育视频中包含了多种模态的信息,如视觉特征(如颜色、形状、纹理)、文本特征(如比分字幕的出现)、运动特征(如慢镜头回放)、音频特征(如解说员的欢呼声、裁判的哨音)等,每种模态信息对视频检索都包含了丰富的语义。因此对于体育视频检索,应该有效地集成多模态信息,以便快速、准确地检索出所需的视频片段。

国内外对基于多模态融合的体育视频检索都做了一定的研究。文献[1,2]分别融合了多模态特征检索足球视频,但是都没有包含语义丰富的音频特征。文献[3]融合了多模态特征,但是没有包含文本特征。另外其音频中只提取出兴奋音,所以只实现了射门事件的检索,并且对于射门事件没有进一步地区分是一般射门事件还是进球事件。文献[4]对动态贝叶斯网络进行扩展,融合多模态特征,但是系统模型计算非常复杂。文献[5]基于视觉特征和足球视频的编辑管理,检索视频中的射门事件。文献[6]基于音频特征和视觉特征对足球视频进行分析。

足球视频有相对的结构性。如射门事件发生时,通常会有球门出现,而且会伴随解说员的激昂的解说音和观众的欢呼声,随后会有慢镜头回放,以便观众对整个事件过程看得更加清晰。若是射门进球事件,则激昂的解说音和观众的欢呼声及回放的镜头都会长些,同时还会有比分字幕出现。对于犯规事件,首先裁判会吹哨子,以示有人犯规,随后会有慢镜头回放,而且回放镜头持续时间较短。本文将针对足球视频的这些特点,融合视觉、文本、音频和运动等多模态特征来检索足球视频中的射门事件和犯规事件。对于射门事件,进一步判断是否为进球事件。本文对足球视频精彩事件检测的框架如图1所示,其中镜头分割可参见文献[10]。

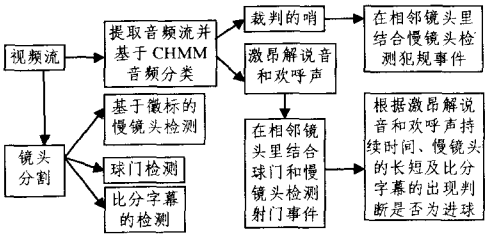

图1 基于多模态融合的足球视频精彩事件检测的流程图

到稿日期:2009-08-04 返修日期:2009-11-02 本文受南京理工大学科技发展基金(XKF09023)资助。
张玉珍(1973-),女,博士生,讲师,主要研究方向为基于语义的视频检索、图像处理和模式识别,E-mail:olindazh@163.com。

2 基于 CHMM 的音频分类

2.1 音频的预处理和特征提取

首先对分流出来的音频流进行预加重处理，以减少尖锐噪声影响，提升高频信号。然后再将音频序列分割为 1s 的音频片段，接着对每个音频片段加 40ms 的 hamming 窗构成音频帧，其中傅里叶变换长度 $Nfft=1024$，相邻帧之间重叠 1/2 帧。分帧后就可以在帧层次上提取音频特征。

对于每帧，根据足球音频的特点，抽取如下的音频特征：短时过零率、短时平均能量、12 阶的 MFCC 系数以及差分 MFCC，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 维的短时平均能量和 1 维的短时过零率。因此从一个长为 1s 的音频片段中就可以提取出一个观察矢量序列 $O=O_1O_2\cdots O_n$，其中 $O_i(i=1,2,\cdots,n)$表示从音频片段中第 i 个音频帧中提取出的 26 维特征矢量，n 表示一个音频片段经上述的音频分帧后被分为 n 个音频帧。

2.2 基于 CHMM 的音频分类

与其他分类器相比，隐马尔可夫模型基于随机数学模型输出概率，能够利用事件之间的概率相关性能提高事件检测的性能。因为声音表现为连续非平稳随机信号，而且声音识别特征参数是一个严格按照时间顺序变化的序列，所以本文采用无跨越由左向右型的 CHMM(连续隐马尔科夫模型)作为分类模型(如图 2 所示，状态数为 4)。在该模型中，对于每一个状态，都用若干个高斯概率密度函数(简称为 pdf)的线性组合表示，每个 pdf 都有各自的均值和协方差矩阵。

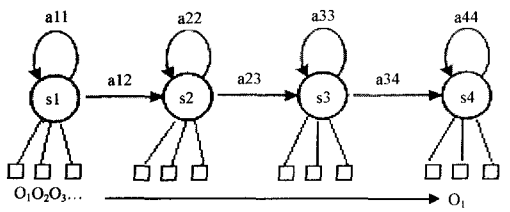

图 2 无跨越由左向右型的 CHMM 的结构图

本文基于 CHMM 分类器，将足球比赛音频片段分为解说员激昂解说、解说员平缓解说、欢呼声、哨音、背景噪音等 5 个类别。基于 CHMM 自动分类的过程是：a)训练过程中，首先对训练样本进行预处理和特征提取，然后对 α，β 和 ξ 分别进行动态标定处理，使用 K-means 聚类算法来初始化模型的参数，并运用 Baum-Welch 算法进行模型的训练，得到 5 个相应的 left-right CHMM 模型，记为 λ_1，λ_2，λ_3，λ_4 和 λ_5；b)在进行分类识别时，对测试样本进行预处理和特征提取，用前向算法计算观察序列对各个 CHMM 模型的输出概率 $P(O|\lambda_i)(i=1,2,3,4,5)$，并选取似然最大的模型来观察序列 O 的类别，即 $S=\arg\{\max_i P(O|\lambda_i)\}$。

为了验证该算法的有效性，本文选取了 CCTV-5 的一段长度为 2942s 包含 452 个镜头的足球比赛音频数据作为实验数据，每一个音频类别选择 10 个样本作为训练数据，其余音频片段作为测试片段。实验结果如表 1 所列。

表 1 足球比赛音频段类型检测实验结果

	解说员激昂解说	解说员平缓解说	欢呼声	哨音	背景噪音
应检数	336	1693	361	75	477
正确检测数	309	1518	335	70	439
误检数	25	159	28	18	41
漏检数	27	175	26	5	38
查全率	92.0%	90.0%	92.8%	93.3%	92.0%
查准率	92.5%	90.5%	92.3%	79.5%	91.5%

由表 1 可知，解说员激昂解说与欢呼声的查准率和查全率都很高，唯有哨音的查准率较低，误检较多，主要是由于背景噪音如突然的鼓声、观众口哨声以及其他尖锐背景噪音等容易误判为哨音。由于本文的音频分类是为了检测精彩事件，当犯规的时候，可以利用哨音查全率高的特点作为犯规事件的初选，后续再用慢镜头特征进一步定位犯规事件，这样就可以解决哨音查准率不高对犯规事件检测造成的影响。

3 基于徽标的慢镜头回放检测

足球比赛中，当出现精彩场面或观众感兴趣的片段之后，通常会出现从多个不同角度对精彩片段进行回放的慢镜头。因为慢镜头出现前后通常存在徽标，所以可以通过徽标检测慢镜头。

3.1 徽标镜头的检测

由于徽标镜头一般持续时间为 10～20 帧左右[7]，而在这个长度范围内的镜头有两种：徽标镜头和特写镜头，其中徽标镜头占多数。因此可将长度为 10～20 帧的镜头作为候选徽标镜头，并取镜头的中间帧作为其关键帧，图 3 为从一段视频中提取的部分候选徽标镜头的关键帧。通过对多种徽标镜头的观察(如图 4 所示)，发现它们都有一个共同的特点，就是镜头图像帧的中心位置都会出现徽标图案，而其它不含徽标的镜头大多都是特写镜头，其中心区域的颜色特征与徽标图案有着明显的区分。因此，可将候选徽标镜头的关键帧图像按横纵方向 1∶2∶1 分割为 9 个窗口，将对应中心区域的中心窗口的图案单独提取出来进行分析处理，以提高计算速度。由于 HSV 颜色空间与人的视觉感知系统有较好的一致性，因此本文选用 HSV 颜色空间模型来表征帧图像中心区域的图像特征。

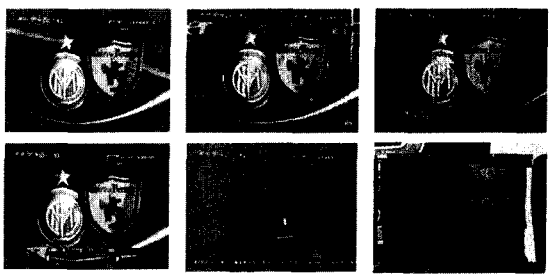

图 3 候选徽标镜头

图 4 各种各样的徽标镜头

在 HSV 颜色空间，求取候选徽标镜头图像 H,S,V 3 个

分量的各分量直方图，设定量化阶数 N 为 100，统计 3 个分量在 0～100 范围内的颜色直方图。处理结果如图 5 所示。

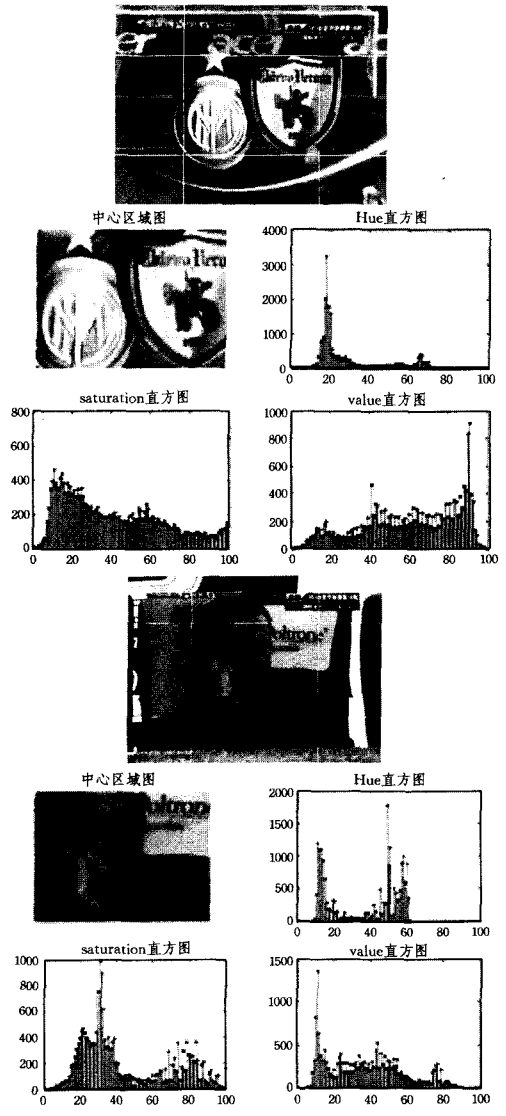

图 5　原图及其中心区域 HSV 颜色直方图

由图 5 可知，含徽标的图像的中心区域的 HSV 3 个分量的直方图与不含徽标图像的有着明显的区别。由于候选徽标镜头中，含徽标的图像占大多数，因此可以通过计算所有候选镜头关键帧图像的 HSV 3 个分量的直方图均值，并根据帧图像与均值的相似度来求得含徽标的镜头，求均值的公式如式(1)所示：

$$H_{mean}(i,j)=\frac{\sum_{n=1}^{L}H_n(i,j)}{L} \quad (1)$$

$$i=1,2,3, j=0,1,2,3,\cdots,N-1$$

式中，L 表示所有候选徽标镜头的个数，$H_n(i,j)$ 表示第 n 个候选徽标镜头的关键帧第 i 个颜色分量第 j 阶的直方图的值，$i=1,2,3$，分别代表 H,S,V 分量，N 代表颜色量化阶数，本文取值为 100。

接着计算每个候选徽标镜头关键帧 HSV 的直方图与均值直方图相似度，公式如下：

$$\mathrm{Sim}(n)=\sum_{i=1}^{3}\sum_{j=0}^{N-1}\min(H_n(i,j),H_{mean}(i,j)) \quad (2)$$

然后将相似度归一化到 0～1 之间，公式如下：

$$Uni\mathrm{Sim}(n)=\frac{\mathrm{Sim}(n)}{\max(\mathrm{Sim}(k))} \quad 1\leqslant k\leqslant L \quad (3)$$

若 $UniSim(n)$ 的值在[0.55,1]内，则认为该镜头是一个徽标镜头，否则不是，这样就可以得到一个徽标镜头集合 $LogoShotSet$。

3.2　根据徽标镜头定位慢镜头

设 $LogoShot$ 为徽标镜头集合 $LogoShotSet$ 中的任意一个镜头，若 $LogoShot$ 是奇次出现，则认为是慢镜头起始镜头之前的徽标镜头，并将该镜头的尾帧的下一帧作为慢镜头的起始帧，若 $LogoShot$ 是偶次出现，则认为是慢镜头结束后的徽标镜头，并将该镜头的起始帧的前一帧作为慢镜头的结束帧。这样就得到了慢镜头的终止帧。通过这种方式就可以找到所有的慢镜头。

如果徽标镜头存在漏检的话，从头到尾按顺序匹配两两徽标镜头，定位慢镜头，则会出现误匹配。为了解决这个问题，我们根据实验确定慢镜头最长不超过 10 个镜头，如果两个徽标镜头之间的镜头数大于 10，就认为徽标镜头存在漏检，则放弃前一个徽标镜头，让后一个徽标镜头与其后紧邻的徽标镜头继续匹配，这样就有效地解决了误匹配。

本算法的检测速度较快，因为仅需考虑视频中 10～20 帧之间的镜头帧图像，并且只选取每个关键帧的中心区域的图像特征作为区别性特征，因此时间代价非常小。经过上述方法就可以将慢镜头回放检测出来，将检测结果存入数据库，以便后续的精彩事件检测。

4　基于 Top-Hat 变换及规则的球门识别

Top-Hat 变换是形态学中一种很重要的变换。Top-Hat 变换的结果即为用原图减去原图被结构元素 S 的开运算。由于开运算能够除去原图灰度曲面中的“波峰”，通过求原图与开运算的差值便可以将原图中的“波峰”提取出来，因此 Top-Hat 变换又被称为波峰检测器。由于白色在 RGB 3 个通道的灰度值均很高，可以看作灰度曲面的“山峰”位置[8]，因此利用 Top-Hat 变换抽取灰度图中的山峰，从而达到增强白色的目的。足球视频中的球门始终是白色的。球门的两根球柱始终是相对水平面垂直的[5]。因此，可以基于 Top-Hat 变换、球门的白色特征及球柱的垂直特征提取球门的球柱，然后基于规则检测球门。球门检测算法具体可参考文献[5]。

5　字幕检测

当射门进球事件发生时，不仅有欢呼声和激昂解说音及慢镜头回放，而且在图像帧的下半部还会出现比分字幕。因此比分字幕的出现有助于进球事件的检测。字幕检测算法的主要思想是根据字幕区域具有比其他区域更高的空间频率，对图像帧下半部分首先进行两级小波分解并重构高频细节；然后将得到的图像分割为若干子块，提取二阶矩统计特征；最后用 K 均值聚类对所有子块聚成两类即字幕或非字幕；然后将这两类的聚类中心与事先从多个样本得到的聚类中心相比较。如果得到的两个聚类中心和已知非字幕聚类中心的距离都大于已知字幕聚类中心，说明当前图像的下半部分有字幕出现。算法的具体步骤可参考文献[9]。

6　基于多模态融合的精彩事件检测

6.1　音视频融合

音频流与视频流在时间上具有时间对应关系，通过该关系就可以将音频特征与视觉特征相融合。设视频流的播放速

度为 FrameRate，镜头的起始帧为 StartFrame，对应的音频段序号为 AudioStartID。镜头的终止帧为 EndFrame，对应的音频段序号为 AudioEndID。则有：

$$AudioStartID=\mathrm{int}\left[\frac{StartFrame}{FrameRate}\right]+1 \quad (4)$$

$$AudioEndID=\mathrm{int}\left[\frac{EndFrame}{FrameRate}\right]+1 \quad (5)$$

式中，int 表示取整操作。*StartFrame* 和 *EndFrame* 都是从整数 0 开始计数，依次加 1。根据式(4)和式(5)描述的时间对应关系，就可以将音频流与视频流中的镜头对应起来。

6.2 基于多模态融合的射门事件检测

足球视频中当出现精彩的射门事件时，一般都会有球门出现，随后会有慢镜头从不同角度对该射门事件的回放，同时还会伴有解说员激昂的解说音和观众的欢呼声，通过实验发现，这种激昂的解说音和观众的欢呼声一般会持续 4s 以上，而且对于射门进球事件，慢镜头回放的时间和解说员的激昂解说及观众的欢呼声会长些，并且还会出现比分字幕。

基于多模态融合的射门事件检测步骤如下。

Step1 对音频流基于 CHMM 进行分类，提取出欢呼声和解说员激昂解说的音频片段；

Step2 对视频流进行镜头分割；

Step3 检测视频流中的徽标镜头，并利用徽标镜头定位慢镜头，检测结果存入数据库；

Step4 利用时间对应关系将音频流与视频流对应起来；

Step5 提取出解说员激昂解说＋欢呼声的持续时间大于 4s 的镜头；

Step6 在满足上述条件的前面 2 个镜头及后面 8 个镜头中，进行球门检测，若不存在球门，则退出；若存在则进一步根据数据库中慢镜头数据检测是否存在慢镜头，若存在慢镜头回放则判断为射门事件。

Step7 针对射门事件，继续判断慢镜头持续的镜头个数是否大于 3 或者解说员激昂解说＋欢呼声的时间是否大于 10s，并且是否出现字幕，如果是，则认为是一次射门进球事件，否则为射门未进球事件。

经过上述步骤，即可将射门进球事件与射门未进球事件检测出来。

6.3 基于多模态融合的犯规事件检测

犯规事件也有其特点。首先裁判会吹哨子，以示有人犯规，随后会有慢镜头回放，而且该回放镜头持续时间较短。没有慢动作回放的犯规，通常不足以影响比赛结果，本文将忽略这种情况，因此裁判的哨音、短暂的慢镜头是犯规事件的两个特征。所以本文采用结合哨音与慢镜头及特写镜头的方法检测犯规事件。

犯规事件检测算法步骤如下。

Step1 基于 CHMM 对音频流进行分类，提取出含哨音的音频段；

Step2 对视频流进行镜头分割；

Step3 检测视频流中的徽标镜头，并利用徽标镜头定位慢镜头，检测结果存入数据库；

Step4 利用时间对应关系将音频流与视频流对应起来；

Step5 提取出含哨音的镜头，并根据数据库中慢镜头数据检测含哨音镜头的后续 6 个镜头序列中是否有慢镜头，若有，则进入下一步；

Step6 检测慢镜头持续时间是否小于 K 个镜头(本文根据经验设 $K=4$)，如果是，则认为是一个犯规事件。

通过上述步骤就能将犯规事件检测出来。

7 实验结果分析

为了验证算法的效果，本文选取 3 段足球比赛视频作为素材，表 2 是实验数据，表 3 是实验结果。对于射门事件(含进球事件)、进球事件及慢镜头的检测，由表 3 可以看出查全率和查准率都比较高。犯规事件的查准率稍微低点，主要是因为现场环境音中那些尖锐噪音也容易被误检为哨音，从而使得犯规事件检测存在一定的误检。因此后期工作中对犯规事件可以考虑检测裁判员镜头，以提高犯规事件的检测效果。

表 2 精彩事件实验数据

视频	射门	进球	犯规	慢镜头
1	22	5	15	40
2	15	2	10	27
3	18	3	12	32

表 3 精彩事件检测的实验结果(查全率和查准率单位均为%)

视	射门		进球		犯规		慢镜头	
频	查全率	查准率	查全率	查准率	查全率	查准率	查全率	查准率
1	86.4	90.5	80.0	100.0	86.7	81.2	87.5	92.1
2	86.7	92.9	100.0	100.0	80.0	80.0	88.9	92.3
3	88.9	94.1	100.0	100.0	83.3	76.9	90.6	93.5

结束语 提出了一种基于多模态融合的足球视频语义分析方法。首先是基于 CHMM 实现音频分类，然后根据时间对应关系融合视频流和音频流，最后在视频流的相应镜头中基于规则再次融合视觉特征(球门的出现)、运动特征(慢镜头)和文本特征实现精彩事件的检测。实验表明该系统及相应算法效果良好。而且本文方法也可以推广到其他体育视频如篮球、网球等。

参 考 文 献

[1] 金国英，陶霖密，徐光，等. 基于 HHMM 的多线索融合和事件推理方法[J]. 清华大学学报：自然科学版，2007，47(1)：112-115

[2] Chen J Y, Li Y H, Wu L D, et al. Semantic event detection in soccer video by integrating multi-features using Bayesian network[C]// Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speedh Procceding. Oct. 2004

[3] 刘宇驰，栾悉道，戴端辉，等. 多模态体育视频语义分析[J]. 计算机科学，2007，34(1)：109-111

[4] 王扉. 体育视频的内容分析技术研究[D]. 北京：中国科学院计算技术研究所，2005

[5] Yang Y, Lin S X, et al. Highlights extraction in soccer videos based on goal_mouth detection[C]// IEEE Proc. ISSPA 2007. 2007：1-4

[6] Barnard M, Odobez J M, Bengio S. Multi-modal audio-visual event recognition for football analysis[C]// IEEE Workshop on Neural Networks for Signal Processing, NNSP. 2003：469

[7] Kolekar M H, Palaniappan K, Sengupta S. A Novel Framework for Semantic Annotation of Soccer Sports Video Sequences[A]// Proceedings of 5th European Conference on Visual Media Production[C]. London, UK, 2008：1-9

[8] 刘国翌，杜威，李华. 足球场地标志线的自动提取[J]. 计算机辅助设计与图形学学报，2003，15(7)：870-874

[9] 王建宇，张峰，周献中，等. 利用小波变换和 K 均值聚类实现字幕区域分割[J]. 计算机辅助设计与图形学学报，2006，18(10)：1508-1512

[10] 张玉珍，王建宇，戴跃伟. 基于自适应双阈值和主色率的足球视频镜头的分割[J]. 南京理工大学学报：自然版，2009，4(8)：432-437

[11] 魏维，等. 音频高层语义分析[J]. 中国图象图形学报，2007，12(1)：141-147