Penalty Detection in Football Video on Audio and Shot

Yanliu Nie[1, a] and Jiande Fan[1, b]

[1]Information Engineering Department, Zhengzhou University of Industrial Technology, Xinzheng, 451100, China

[a]361045705@qq.com, [b]568498042@qq.com

Keywords: Audio; Goal; Excited audio; Bass; Penalty

Abstract. The present study on video detection is mianly based on image and video sequence. In the study of the video stream, goal shot play an important role, often indicates the highlight. In recent years, audiohas become more and more important with its rich information. Through analyzing the audio in order to find the excitement and bass, combining with the goal shot to determine whether the penalty shot or not. Experiments show that based on audio and video has higher recall and precision rate.

音频与球门融合的足球视频点球镜头检测

聂燕柳[1,a]，范建德[1,b]

1.郑州工业应用技术学院 、信息工程学院，中国 河南省 郑州市 451100

[a]361045705@qq.com, [b]568498042@qq.com

摘要：在足球镜头的检测中，图像和视频是其主要手段。在视频流的研究中，球门镜头起着十分重要的作用，往往预示着精彩镜事件的发生。近年来音频以其具有丰富的信息所起作用也越来越突出。本文通过对比赛音频分析判定兴奋音与低音，并且与球门镜头相结合，从而判定是否为点球镜头。实验表明，这种音视频融合的方法对点球事件的识别有较高的准确率和查全率。

关键词：音频；球门；兴奋音；低音；点球

1. 引言

足球做为世界第一大运动，各大联赛和世界杯等赛事吸引了亿万人的眼球，因而对足球视频语义内容的研究具有广阔的前景。

在视频分析与检索中，图像与声音越来越重要，文献[3]从视觉和听觉两个方面提取摄像机运动和大量音频特征从而检测出进球镜头，但误检率比较高。文献[4]对视频底层特征分析的基础了，提取音视频关键字作为中级特征，Chen[5-6]等人提取音量、能量和频谱等音频特征以及镜头类型特征，采用决策树来实现对进球事件的检测。Hanjalic[7-8]等人从视觉和听觉两个方面提取短时能量、镜头切换率和运动强度等特征，建立基于情感激励的精彩体育视频分析模型。辛宪阳[9]等人跟据兴奋音与有禁区白线的长距离镜头、特写镜头或场外镜头、含禁区的长距离镜头转换确定射门事件，但这种很容易误把角球镜头当作射门镜头。

本文通过在文献[9-10]的基础上利用兴奋音、低音和球门镜头检测，确定点球镜头。

2. 兴奋音检测

在足球比赛视频中，解说员激扬快速的解说音和现场观众的呼喊声，就往往预示着精彩事件的发生。本文通过计算过零率，短时能量，高过零率，低短时能量确定兴奋音。

2.1. 短时平均能量(Short time energy)

短时平均能量定义如公式 （1.1）：

$$E_n = \sum_{n=1}^{N} x(n)^2 \tag{1.1}$$

其中的 $x(n)$ 为音频帧的第 n 个采样点， N 表示一帧内采样点的个数。短时能量反映了在很短的时间内声音振幅能量的变换规律。在足球比赛当中，当有精彩镜头出现的时候，解说员激动的解说声和现场观众的呼喊声也会有明显的变化如图 1.1。

2.2. 短时过零率(Zero-Crossing Rate)

过零率可以比较准确地度量窄带信号的频率，也可以粗略反映宽带信号的频率特征。其计算方法如公式（1.2）：

$$Z_n = \frac{1}{2N}\sum_{n=1}^{N}[\mathrm{sgn}[x(n+1)] - \mathrm{sgn}[x(n)]] \tag{1.2}$$

其中 $x(n)$ 为离散采样信号， N 表示一帧内采样的个数。对于不同的音频说，不同的频率就有不同的过零率。足球比赛中精彩部分和非精彩部分的过零率也有差别如图 1.2。

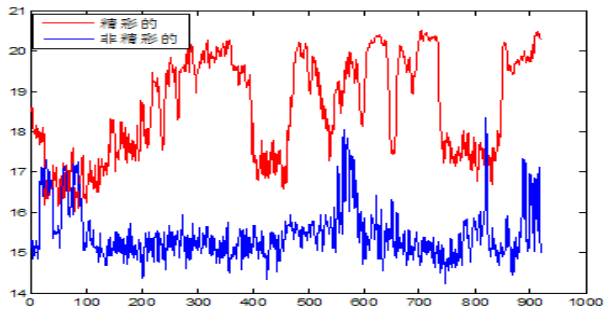

图 1.1 不同音频的短时能量

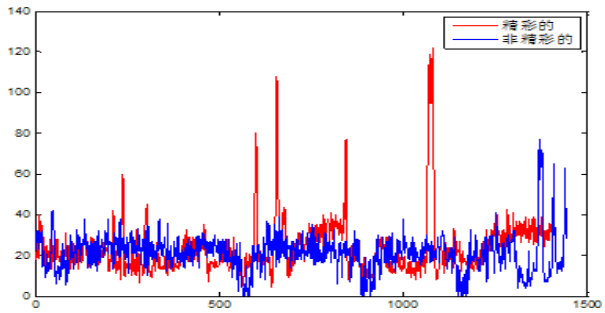

图 1.2 不同音频的短时过零率

2.3. 高过零率和低短时能量

2.3.1 高过零率

设定一个过零率的阈值，那么就可以计算出一个音频片段中过零率高于这个阈值的帧所占的比例，这个比例就是高过零率的比率。其计算公式（1.3）为：

$$HZCRR = \frac{1}{2N}\sum_{n=1}^{N}[\mathrm{sgn}(ZCR(n) - 1.5avZCR) + 1] \tag{1.3}$$

$$avZCR = \frac{1}{N}\sum_{n=1}^{N} ZCR(n) \tag{1.4}$$

2.3.2 低短时能量

如果设定一个能量的阈值，那么就可以计算出一个音频片段中能量低于这个阈值的帧所占的比例。这个比例就是低短时能量比例。其计算公式（1.5）为：

$$LSTER = \frac{1}{N}\sum_{n=1}^{N}[\mathrm{sgn}(0.5avSTE - E(n)) + 1] \tag{1.5}$$

$$avSTE = \frac{1}{N}\sum_{n=1}^{N} E(n) \tag{1.6}$$

其中 N 为一个片段中的帧数， $E(n)$ 是信号第 n 帧的能量。

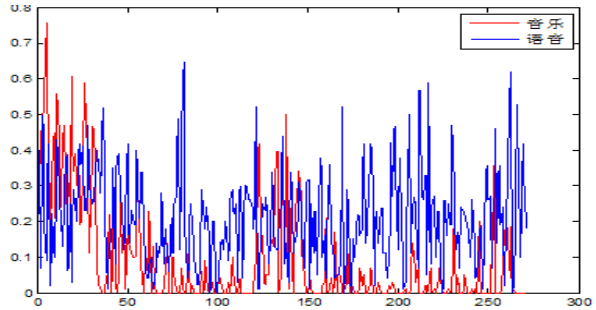

图 1.3 不同音频对应的高过零率

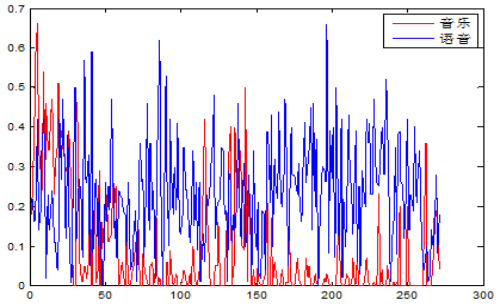

图 1.4 不同音频的低短时能量

如图 1.3 及 1.4 所示，语音与音乐的高过零率及低短时能量区别较大，由于足球视频中，解说声音占比很大，因此高过零率有很重要的参考价值。

2.3.3 音频分类

本文选用 SVM 作为分类器。我们主要把音频分为两种兴奋音和非兴奋音，其优化方程为：

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T w + C\sum_{i=1}^{l}\xi_i\right) \qquad (1.7)$$

$$s.t.\ \ y_i(w^T\varphi(x_i)+b)\geq 1-\xi_i,\xi_i\geq 0$$

本文的 SVM 分类器采用高斯核函数，即：

$$K(x,x_i)=\exp\{-\gamma\|x-x_i\|^2\}\ \ (\gamma>0) \qquad (1.8)$$

为检验本文方法的效果，采用 VC++和 DirectShow 开发实验平台。实验视频为 3 场足球比赛视频片断， 兴奋音的检测结果如表 1.1。

表 1.1 检测结果

	查全率	查准率
Test1	81.55%	77.78%.
Test2	77.78%	80.00%
Test3	68.67%	70.00%
平均	76.00%	75.92%

对三场比赛进行测试，分别是阿森那对涅茨克矿工队，切尔西对曼联，巴塞罗那对皇家社会，检测结果如表 1.1 对兴奋音的检测在查准率和查全率方面均有较好效果，产生的误检核漏检主要是因为现场的噪声以及解说员的情绪引起的；另一方面是 SVM 分类器分类产生误差，可以通过完善训练样本和分类器优化还进一步减小误差。

3. 球门检测

首先对视频流进行视频帧分割，然后将关键帧作如下处理：

（1）用下公式（2.1）将彩色帧图像转化为灰度图像；

$$Y = 0.299R + 0.587G + 0.114B \qquad (2.1)$$

其中 Y 表示灰度值，R，G，B 分别表示每个像素的红、绿、蓝分量。

（2）将灰度图像二值化处理，得到二值化图像 $I(x,y)$：如下公式（2.2）所示：

$$I(x,y)=\begin{cases}255, Y\geq T\\ 0, Y<T\end{cases} \qquad (2.2)$$

（3）对图像 $I(x,y)$ 做拉普拉斯锐化。

通过高斯滤波器对图像进行平滑。二维高斯滤波器的响应函数为：

$$G(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.3}$$

$I(x,y)$为图像函数，由线形系统中卷积和微分的可交换性得：

$$\nabla\{G(x,y)*I(x,y)\}=\{\nabla G(x,y)\}*I(x,y) \tag{2.4}$$

即对图像的高斯平滑滤波与拉普拉斯微分运算可结合成一个卷积算子如下公式（2.5）：

$$\nabla G(x,y)=\frac{1}{2\pi\sigma^4}\left(\frac{x^2+y^2}{\sigma^2}-2\right)e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{2.5}$$

$$=A^2\left(\frac{x^2}{\sigma^2}-1\right)e^{-\frac{x^2}{\sigma^2}}e^{-\frac{y^2}{\sigma^2}}+A^2\left(\frac{x^2}{\sigma^2}-1\right)e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{y^2}{\sigma^2}}$$

其中：$A=\frac{1}{\sqrt{2\pi\sigma^2}}$　　$K_1(x)=A\left(\frac{x^2}{\sigma^2}-1\right)e^{-\frac{x^2}{2\sigma^2}}$　　$K_2(x)=Ae^{-\frac{y^2}{2\sigma^2}}$

用上述算子卷积图像，得到经过拉普拉斯变换后的图像。将图像根据下公式 2.6 做垂直投影，其中b为每列黑色像素的个数，N为图像高度。

$$V(j)=\begin{cases}0, & 0<j<b\\ 255, & b\le j<N\end{cases} \tag{2.6}$$

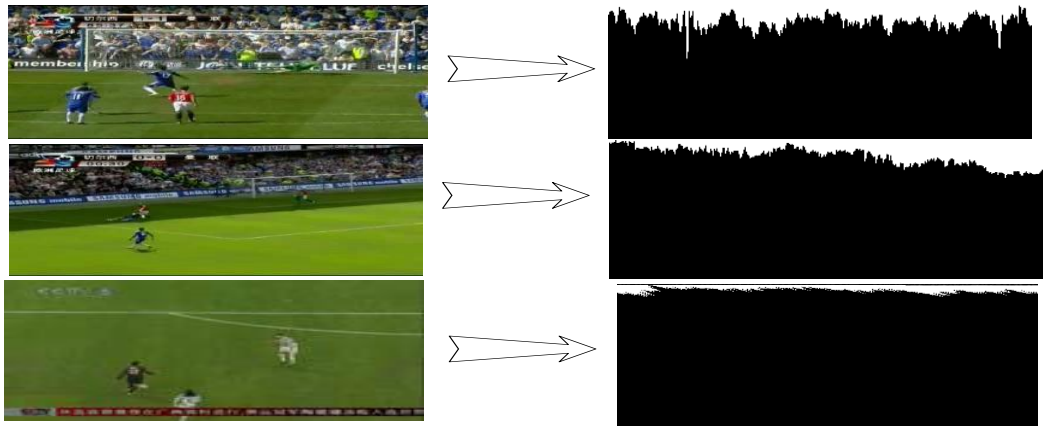

图 2.1 不图的镜头类型处理后的结果

对图像处于是后图像做自上而下，自左至右的扫描，统计每列中白色像素所占的比例$P_y=W/N$（W为白色像素所占个数，N为图像高度）。若 P_y>1/5，则我们认为检测到了球门镜头。

4. 点球镜头检测

传统的进球镜头运动进球和定位球进球，其中定位球进球包括角球进球和点球进球。运动进球会有大量的远镜头，会伴随着现场高昂的欢呼声和解说音。点球进球前有大量的球门镜头，并且有一段时间的低音如图 3.1。具体的检测规则如下：

检测是否出现大量球门镜头，音频处理区分兴奋音与非兴奋音。是否出现大量计算兴奋帧 a 计算短时平均能量E_a。计算非兴奋音频段b并计算其短时平均能量E_b。如果 $E_b<\frac{E_a}{2}$这说明出现

低音。如同时满足大量球门镜头和出现低音则出现点球镜头

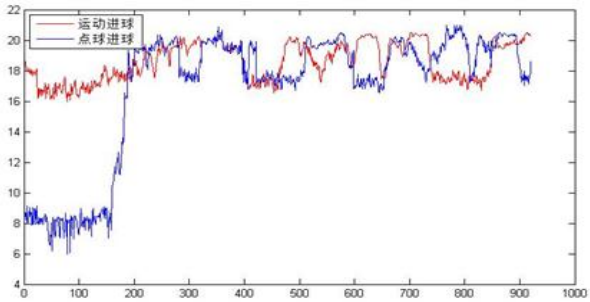

图 3.1 不同进球方式对应的音频类型

实验结果分析

表 3.1 检测结果

	点球事件实检数	误检数	漏检数	准确率
Test1	1	0	0	100%
Test2	2	0	1	50%
Test3	1	0	0	100%

表给出了对于用来测试的 3 段视频进行了音视频融合的点球事件检测的结果。从表中可以看出，本文对点球事件有较好的效果。其中视频 2 点球镜头的误检主要是主客场的观众情绪差异引起的。

5. 结论

本文基于音频兴奋音和球门镜头相融合检测，首先根据足球精彩事件的发生往往伴随着兴奋音这一特征，对音频内容进行兴奋音检测并利用 SVM 进行分类。在音频兴奋音检测的基础上加以球门镜头检测，最后实现点球镜头的检测。实验表明对于点球镜头有较高的查全率和准确率。如何去除现场噪音降低主客场影响以及如何判定点球镜头是否为进球镜头是未来的研究重点。

6. 致谢

河南省高等学校重点科研项目《大数据检索在图像标注及重构中的应用》（No. 15B520036）
河南省科技攻关计划项目《基于道路场景的车辆阴影检测和去除算法研究》（No. 162102210119）
感谢以上两个项目的支持。

参考文献

[1] 童晓峰,刘青山,卢汉清.体育视频分析[J]. 计算机学报.2008, 31(7): 1242-1251.

[2] 于俊清,何欢欢，何云峰，利用情感激励提取足球视频精彩镜头[J].计算机研究与发展.2010.47(10): 1823-1831

[3] Leonardi R. Migliorati P. Prandino M.Semantic indexing of soccer audio-visual sequences: A multimodal approach base on controlled Markov chains. IEEE Transactions on Ciruits and Systems for Video Technology. 2004. 14(5):634-643.

[4] Kang Yu-Lin, Lim Joo-Hwee, Kankanhalli M S, et al. Goal detection in soccer video using audio visual keywords, Proceedings of the International Conference on Image Processing. Singapore, 2004:1629-1632.

[5] Chen Shu-ching. Shyu Mei-Ling. Chen Min.et al. A decision tree-based multimodal data mining framework for soccer goal detection//Proceedings of the IEEE International Conference on Multimedia and Expo. Taipei, CHina. 2004:256-268

[6] Xie Zongxing. Shyu Mei-Ling.Chen Shu-ching, Viedo event detection with combined distance-based and rule-based data mining techniques//Proceedings of the IEEE International Conference on Multimedia and Expo. Beijing, CHina. 2007:2026-2029.

[7] Hanjalic A. Adaptive extraction of highlights from a sport viedo based on excitemeng modeling. IEEE Transactions on Multimedia. 2005, 7(6):1114-1122.

[8] Hanjalic A.Xu Li-Qun. Affective video content representation and modeling. IEEE Transactions on Multimedia. 2005,7(1):143-154.

[9] 辛宪阳. 基于多模态融合的足球视频语义分析[D]. 吉林大学, 2011.

[10] 聂燕柳,刘群,林彬.音视频融合的足球精彩镜头分类[J].计算机应用研究.2013, 30(1), 534-535.

Acknowledgement

This research was financially supported by the 2015 Research Project of Higher Education Reform in Henan Province (15B520036) and the 2016 Research Project of the Key Scientific and Technological in Henan Province (162102210119)

References

[1] Tong Xiao-Feng, Liu Qing-Shan, Lu Han-Qing. A survey on sports video analysis [J]. Chinese Journal of Computer. 2008, 31(7):1247-1251.

[2] Yu Jun-qing, He Huan-huan, He Yun-Feng. Highlights extraction for soccer video based on affection arousal.Journal of Computer Research and development. 2010.47(10):1823-1831

[3] Leonardi R. Migliorati P. Prandino M.Semantic indexing of soccer audio-visual sequences: A multimodal approach base on controlled Markov chains. IEEE Transactions on Ciruits and Systems for Video Technology. 2004. 14(5):634-643.

[4] Kang Yu-Lin, Lim Joo-Hwee, Kankanhalli M S, et al. Goal detection in soccer video using audio visual keywords, Proceedings of the International Conference on Image Processing. Singapore, 2004:1629-1632.

[5] Chen Shu-ching. Shyu Mei-Ling. Chen Min.et al. A decision tree-based multimodal data mining framework for soccer goal detection//Proceedings of the IEEE International Conference on Multimedia and Expo. Taipei, CHina. 2004:256-268

[6] Xie Zongxing. Shyu Mei-Ling.Chen Shu-ching, Viedo event detection with combined distance-based and rule-based data mining techniques//Proceedings of the IEEE International Conference on Multimedia and Expo. Beijing, CHina. 2007:2026-2029.

[7] Hanjalic A. Adaptive extraction of highlights from a sport viedo based on excitemeng modeling. IEEE Transactions on Multimedia. 2005, 7(6):1114-1122.

[8] Hanjalic A.Xu Li-Qun. Affective video content representation and modeling. IEEE Transactions on Multimedia. 2005, 7(1):143-154.

[9] Xin Xian-yang.Semantic Analysic for Soccer Video Based on Fusion of Multimodal Feature.

[10] Nie Yan-liu, Liu Qun, LinBin.Highlight deteciong 1n football video on video and audio [J].Appyication Research of computers.2013, 30(1), 534-535.

作者简介：第一作者聂燕柳（1987—），性别男，籍贯河南焦作，职称讲师，主要研究方向多媒体技术、数据挖掘，E-mail：361045705@qq.com.