# 多模态体育视频语义分析*)

刘宇驰[1,2] 栾悉道[1] 戴端辉[3] 吴玲达[1]
(国防科学技术大学信息系统与管理学院 长沙 410073)[1] (空军雷达学院一系 武汉 430019)[2]
(陆军航空兵学院模拟训练中心 北京 101114)[3]

**摘 要** 以足球运动为例提出了一种体育视频语义结构,并提出相应的语义分析框架。视频被分解为纯视频流和音频流两种模态,每种模态均可依次提取和综合出低层内容和中层内容。视频流可根据低层(物理)内容分割为物理镜头,然后根据特定的中间层内容可以确定为语法镜头。音频也可以在物理特征的基础上形成有意义的中间层内容,如解说员兴奋时的声音。最后,根据视频流和音频流的中间层内容,按照足球比赛转播的规律,分析出比赛中的精彩事件,并选取相关的镜头作为反映此事件的序列组合。

**关键词** 体育视频,多模态,语义分析,语法镜头

## Multi-modal Analysis of Sports Video for Semantics

LIU Yu-Chi[1,2] LUAN Xi-Dao[1] DAI Duan-Hui[3] WU Ling-Da[1]
(National University of Defense and Technology, Changsha 410073)[1] (Air Force Radar Academy, Wuhan 430019)[2]
(Center of Simulation Training of Army Aviation Institute, Beijing 101114)[3]

**Abstract** A semantic structure of sports video, exampled with soccer, and corresponding framework for semantics analysis are proposed. Video is parsed into pure video stream and audio stream. Video is segmented into shots according to low/physical features, and then into syntactic shots with the help of specific middle level contents. Audio can be extracted meaningful middle contents, e.g. excited speech of commenter. According to rules of soccer broadcasting, semantics of highlights can be analyzed based on syntactic contents from video and audio streams.

**Keywords** Sports video, Multi-modal, Semantic analysis, Syntactic shot

## 1 引言

视频的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,因而研究得比较多。

为了获取语义,Ekin 等人[1]将足球视频中的镜头确定为预先指定类型,再根据视频编辑的一般规律,分析比赛中的重要语义"事件"。文[3,8]应用音频分析来探测棒球等比赛中的精彩镜头。还有对球员和足球进行探测和跟踪的[2]。统计方法也引入了语义分析中,在文[6,7]中,他们应用数据融合技术对足球视频进行语义索引。L. Xie 等人[4]应用 HMM 将视频流分为"比赛中"和"比赛中断"两种状态。这些方法取得了一定的成果。

本文以足球视频为例,提出了体育视频的一种语义结构,应用多模态的方法分析其语义。

## 2 体育视频的语义结构

在各种视频中,体育视频具有相对的结构性。这是因为体育场周围的摄像机数量有限,位置也相对固定,转播导演又会以特定的方式来组织由不同的摄像机所捕获的镜头。以足球视频为例,为了帮助观众理解和欣赏比赛,导演会适时地选用不同的镜头。如在进攻时,导演会切换到中距离的镜头表现比赛的关键情况(如传中),然后会有关键球员(如射门者)的特写,经常还会有观众等欢庆的场面,然后会用慢镜头重放比赛,接着再回到比赛直播中。同样的语义事件在转播中则表现为相似的镜头序列。

因此,体育视频的基本语义表达单位是镜头;一段视频是由各种"语义事件"来组成的,而每个"语义事件"则由数个特定的镜头序列构成。相对于直接根据低层特征区分出来的"物理镜头",这些镜头可称为"语法镜头"。我们确定了如下镜头类型:远景镜头、中景镜头、近景镜头、慢镜头、场内镜头和场外镜头。一个镜头可以赋予两种语法含义,如一个镜头可以同时赋予"远景镜头"和"场内镜头"。

## 3 多模态语义分析

视频往往包含两种模态,即纯视频流和音频流,这两种模态都是视频语义分析的信息来源。我们使用这两种模态进行以足球为代表的体育视频语义分析。

### 3.1 多模态语义分析框架

根据前面的讨论,我们提出了如图 1 所示的体育视频语义分析框架。

首先,视频被分流器分为视频流和音频流。然后,分别分析视频流和音频流。视频流先根据物理特征分割为"物理镜头",进而确定为如第 2 部分前述的"语法镜头"。音频流先提取出低层的"物理内容",再综合出"语法内容",具体见下一小节。最后,融合语法镜头和音频的语法内容,分析出特定的语义事件。

### 3.2 语法分析

我们应用文[5]中的算法分割出足球视频中的镜头,并选出相应的关键帧;此时它们并未赋予语义,称为"物理镜头"。

*)本文受到国家自然科学基金(编号:60473117)的资助。刘宇驰 博士研究生,主要研究方向:视频处理、视频语义分析;栾悉道 博生研究生,研究方向:视频处理、虚拟环境;戴端辉 讲师,主要研究方向:系统工程,作战模拟,媒体处理;吴玲达 教授,博士生导师,研究领域:多媒体与虚拟现实。

其后就要将各镜头确定"语法镜头"。

由于"语义鸿沟"的存在,无法根据方便提取的颜色等低层物理特征直接获得语义。为了填补这个鸿沟,就要引入特定的知识将物理特征"映射"到语义上去,在映射过程中,往往要引入"中间层内容"。中间层内容建立在物理特征的基础上,并直接用于语义的分析。为了区分各种镜头类型,我们根据关键帧合成了如下中间层内容。

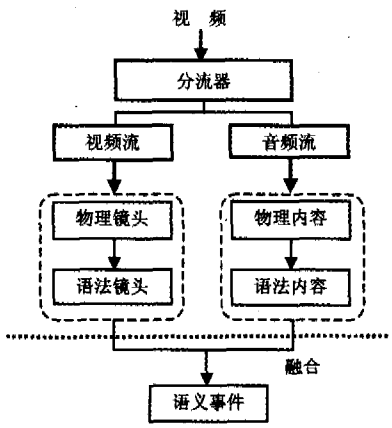


图 1 多模态语义分析框架

1)场地比率 $R(field)$:足球场是绿色的,有关场地的远、中、近景镜头所包含的场地颜色比率也有明显差异。该特征计算如下:

$$R(field)=A(field)/A(Image) \tag{1}$$

其中,$A(frame)$、$A(field)$分别表示关键帧中全部像素数和场地颜色像素数。

2)人脸比率 $R(face)$:关键帧中人脸所占的比例。用可伸缩的方框在关键帧中移动,当出现人脸肤色时,放大这个框,计算人脸肤色像素值在此框中的比例 $R(skin\text{-}color)$,在同一伸缩方框中选择 $R(skin\text{-}color)\geqslant 0.8$ 的最大方框作为候选人脸,然后计算 $R(face)$:

$$R(face)=A(Max(skin\text{-}color))/A(Image) \tag{2}$$

其中,$Max(skin\text{-}color)$表示最大的候选人脸方框。

3)边缘 $E(Image)$:用 Sobel 算子计算关键帧中边缘,这是因为远景观众中边缘要远远多于其它场景的镜头。

4)运动强度 $M(Image)$:它用于表现该镜头运动情况,在慢镜头中,其运动强度要比同样类型播放正常的镜头要低。我们所用的视频为 MPEG-1 编码,该特征源自于文[4],计算如下:

$$M(Image)=1/|\Phi|\sum_{\Phi}\sqrt{V_x^2+V_y^2} \tag{3}$$

其中,$\Phi$ 表示帧间编码宏块的数目,$V_x$、$V_y$ 则是每个宏块的运动向量。

各中间层内容的阈值的选取,根据手工标注的镜头类型计算出来的。

至于音频流,首先提出基本的物理特征,如:短时能量、过零率等,然后在此基础上计算语法内容。这里,我们计算一个重要的语法内容——"兴奋音"。由于在体育比赛中,当出现精彩场景时,解说员会因为兴奋而提高嗓音,因此探测兴奋音也可辅助分析精彩场景。

首先,将音频流中的语音部分提取出来,然后探测兴奋音。

我们用短时能量分离出语音,这是因为一般情况下,语音的短时能量比其他音频信号的要高。短时能量表征平均波形幅度,$m$ 个输入音频样本可用如下公式定义:

$$E_m=\frac{1}{N}\sum_{n=0}^{N-1}x(n)^2h(m-n) \tag{4}$$

这里,$x(n)$是输入样本,$h(m-n)$表示线性过滤器,我们应用适合探测语音端点的汉明窗过滤器。音频特征从一个固定长度为 10ms 的窗口提取。大小为 $N$ 的汉明窗计算如下:

$$h(p)=\begin{cases}0.54-0.46\cos(\frac{2\pi p}{N-1}), & 0\leqslant p\leqslant N-1\\ 0, & \text{其它}\end{cases} \tag{5}$$

下一步,探测其中的兴奋音部分。兴奋音在语音中表现为主要的频率,并且具有较高的能量值。我们应用自相关函数来进行兴奋音估计,自相关函数定义如下:

$$A(k)=\frac{1}{2N+1}\sum_{n=-N}^{N}x(n)x(n+k) \tag{6}$$

这里,$k$ 表示重叠的样本数。现在我们就可以计算不同的 $k$ 值下的峰值。当 $A(k)$超过一定的阈值时,指定窗口的兴奋音定义为最大的峰值(即 $\mathrm{Max}(A(k))$)所在的部分。

图 2 所示为 10 秒钟(1000 帧)音频流中兴奋音探测结果。图中(476,535)、(632,743)和(831,943)部分表示兴奋音所在的时间段,而(48, 356)表示一般语音的所在时间段,其余的则表示没有语音的部分。

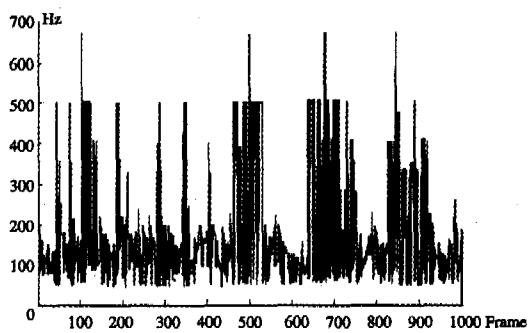


图 2 音频中兴奋音探测示例

### 3.3 语义分析

体育视频的语义以"事件"的形式表示,是在语法镜头和音频语法内容的基础上分析出来的。目前,我们只分析足球视频里的关键的语义事件,即进球和有威胁的射门,因为这是观众最需要的镜头,也是足球比赛剪辑节目要选取的。由于这两个事件都会出现慢镜头和兴奋音,因此我们提出了基于这两个语法内容的推理规则:

1)在有兴奋音的发生的相邻镜头里探测精彩事件;

2)在兴奋音刚发生以后的 6 个镜头里至少要有两个慢镜头;

3)如果满足上两个条件,选取兴奋音发生时前后各 6 个镜头作为表示事件的镜头序列。

## 4 实验情况

我们选取德甲、意甲和西甲三个联赛的足球视频作为实验对象。对每个联赛,各取一段时长 10 分钟左右的视频进行人工标注,确定有关语法镜头的四个中间层内容:场地比率、人脸比率、边缘和运动强度的阈值。然后,对三个联赛中的 5 段视频进行语义结构分析。这 5 段视频总共时长约 183 分钟,具体分析结果如表 1 所示。

表 1 中,"准确率"为探测正确数目与所探测出来的数目之比,"完全率"为探测正确数目与真实数目之比。大部分指标都在 90%以上,其余的也在 87%以上,这说明我们选择的特征合适,总体方法是有效的。

表1 实验结果

| | 实际数目 | 探测正确 | 探测错误 | 未探测到 | 准确率% | 完全率% |
|---|---|---|---|---|---|---|
| 远景 | 789 | 743 | 32 | 46 | 95.9 | 94.2 |
| 中景 | 358 | 330 | 49 | 28 | 87.1 | 92.2 |
| 近景 | 331 | 306 | 18 | 25 | 94.4 | 92.4 |
| 场内 | 1335 | 1285 | 49 | 50 | 96.3 | 96.2 |
| 场外 | 142 | 130 | 13 | 12 | 89.7 | 91.5 |
| 慢镜头 | 105 | 97 | 10 | 8 | 90.9 | 92.3 |
| 兴奋音 | 25 | 23 | 3 | 2 | 88.5 | 92.0 |
| 事件 | 11 | 10 | 1 | 1 | 90.9 | 90.9 |

**结论与展望** 本文中，我们分析了足球视频的语义结构，每个语义事件可视为特定的镜头序列，这些镜头因而可称为"语法镜头"。对已经分割出来的物理镜头，选择相应的中间层内容，分析出语法镜头；对音频流分析出解说员的"兴奋音"。最后综合视频流和音频流的语法内容，推理出精彩事件。实验表明效果良好，这种方法也可用于其他体育视频的分析，如篮球、棒球等。但是，我们分析出的语义事件还比较粗略。在未来，我们要确定镜头的语义，如裁判员镜头；还要细化"语义事件"，如犯规、任意球等。另外，还要将这种方法应用到其他体育视频的语义分析。这将要求我们选择更多的中间层内容，并要对视频中的文字进行识别。

## 参考文献


1 Ekin A, Tekalp A M. Generic event detection in sports video using cinematic features. In: Second IEEE Workshop on Event Mining: Detection and Recognition of Events in Video (EVENT 2003), Madison, Wisconsin, USA, June 2003. 17～24

2 Yu X, Xu C, Leong H, Tian Q, Wan K. Trajectory-Based Ball Detection and Tracking with Applications to Semantic Analysis of Broadcast Soccer Video. In: Proc. of ACM Multimedia, 2003. 11～20

3 Babaguchi N, Kawai Y, Kitashi T. Event based indexing of broadcasted sports video by intermodal collaboration. IEEE Trans. Multimedia, 2002, 4(1): 68～75

4 Xie L, Xu P, Chang S F, Divakaran A, Sun H. Structure analysis of soccer video with domain knowledge and hidden Markov models. PRL(25), May 2004(7): 767～775

5 Liu X M, Chen T. Shot Boundary Detection Using Temporal Statistics Modeling. In: IEEE Intl. Conf. on Acoustics, Speech and Signal Processing, ICASSP2002, Orlando, FL, U. S, May 2002

6 Lenardi R, Migliorati P, Prandini M. Semantic Indexing of Soccer Audio-Visual Sequence: A multimodal approach based on controlled Markov chains. IEEE Trans on Circuits&System for Video Technology, 2004, (5): 634～643

7 Barnard M, Odobez J M, Bengio S. Multi-modal audio-visual event recognition for football analysis. In: IEEE Workshop on Neural Networks for Signal Processing, NNSP, 2003. 469～478

8 Baillie M, Jose J M. Audio-based Event Detection for Sports Video, CIVR2003, July 2003. 300～310


(上接第 105 页)

和 $output(t)$ 进行 AND/OR 操作。操作后，染色体的因果关系矩阵并没有发生变化，改变的只是活动 $t$ 的路由模式 AND-OR/JOIN-SPLIT。

### 3.4 终止规则

当满足下面任一条件时，即可认为算法收敛，停止计算，输出最优解：(a)染色体的适应度达到给定的阈值(例如 1)；(b)经过 N 代进化，当 N 是所允许计算代(generation)的最大值；(c)染色体数组的第 0 位 chromo [0]连续 $q$ 代没有发生变化。

### 3.6 模拟退火操作

对于经过遗传算法选择复制、交叉、变异操作的群体作为模拟退火算法的初始群体，运用基于 Metropolis 判别准则的复制策略，产生下一代群体。即在染色体的邻域中随机产生新染色体 $i$ 和 $j$，竞争进入下一代群体的准则采用 Metropolis 判别准则：令 $\Delta F = Fit_i - Fit_j$，若 $\Delta F \leqslant 0$，则把染色体 $j$ 复制到下一代群体；否则产生[ 0, 1 ]之间的随机数 $r$。如果 $r < \exp(\Delta F/t_n)$，则同样把染色体 $j$ 复制到下一代群体，否则，把染色体 $i$ 复制到下一代群体。基于 Metropolis 判别准则的复制策略，在接受优质解的同时，有限度地接受劣质解，保证了群体的多样性，进一步避免了算法陷入局部最优解的可能性。

### 3.7 试验

基于本文提出的工作流重构算法，对包含活动数目为 8～25 的流程日志进行测试，测试日志中包含 5%～15%的噪声数据。初始群体规模 $popSize = 2N = 50$，算法中各系数确定如下：选择确定初温系数 $K=20$，退温操作系数 $\alpha=0.8$，交叉操作系数 $K_1=0.2$、$K_2=0.6$、$K_3=K_1+K_2=0.8$，变异操作系数 $K_4=0.01$、$K_5=0.1$、$K_6=K_4+K_5=0.11$。为防止解空间被破坏，当 $\varepsilon = 0.5Fit_{avg}$ 时，交叉率和变异率固定为 0.2和 0.01，算法终止规则中 $q=15$，$N=25$。试验结果表明：该算法能够在包含噪声数据的工作流日志中成功进行流程重构，并且由于采用全局策略和混合遗传算法，有效地避免了结果陷入局部最优解；并通过提高收敛速度，减少了流程模型重构的时间开销。

**结论** 随着知识抽取、数据挖掘等技术的兴起，基于日志的企业流程重构引起了学术界和工业界的高度重视。流程在结构上是由活动构成的，随着活动数目的增加，备选流程模式将会构成一个巨大的搜索空间。流程重构问题的实质就是在搜索空间中选出同流程实际行为最匹配的模式。本文针对目前重构算法采用本地策略以及无法有效处理噪声的情况，提出了一种新的工作流重构算法。该算法通过对流程活动依赖关系的度量，构建因果关系矩阵对流程实例(CASE)进行映射；然后通过因果关系矩阵来构建初始种群，结合遗传算法和模拟退火算法的思想，实现了对流程模型的有效挖掘。

## 参考文献


1 Cook J E, Wolf A L. Discovering Models of Software Processes from Event-Based Data. ACM Transactions on Software Engineering and Methodology, 1998, 7(3): 215～249

2 Cook J E, Wolf A L. Event-Based Detection of Concurrency. In: Proceedings of the Sixth International Symposium on the Foundations of Software Engineering (FSE-6), 1998. 35～45

3 Cook J E, Wolf A L. Software Process Validation: Quantitatively Measuring the Correspondence of a Process to a Model. ACM Transactions on Software Engineering and Methodology, 1999, 8 (2): 147～176

4 Agrawal R, Gunopulos D, Leymann F. Mining Process Models from Workflow Logs. In: Sixth International Conference on Extending Database Technology, 1998. 469～483

5 Maruster L, Weijters A J M M, van der Aalst W M P, et al. Process Mining: Discovering Direct Successors in Process Logs. In: Proceedings of the 5th International Conference on Discovery Science (Discovery Science 2002), Vol 2534 of Lecture Notes in Artificial Intelligence. Berlin: Springer-Verlag, 2002. 364～373

6 Kiepuszewski. Expressiveness and Suitability of Languages for Control FlowModeling in Workflows ( Submitted): [ Ph D Thesis ]. Queensland University of Technology, 2002

7 de Medeiros A K A, van Dongen B F, van der alst W M P, et al. Process Mining: Extending the α-algorithm to Mine Short Loops. BETA Working Paper Series, WP 113. Eindhoven University of Technology, Eindhoven, 2004

8 Hochbaum D. Approximation Algorithms for NP-hard Problems [M]. Berkeley, CA, PWS Publishing Company, 1997

9 周丽，孙树栋. 遗传算法原理及应用[M ]. 北京：国防工业出版社，2001

10 欧阳森，宋政湘，王建华，等. 一种快速收敛的遗传算法[ J ]. 计算机应用研究，2003, 20 (9): 50～52