

摘要

近几年来，基于内容的体育视频分析越来越受到人们的关注。研究者在结构分析和事件提取方面展开了大量的工作。但是这些工作有两个局限性，一是由于只研究有限种视频，例如足球、棒球，其通用性还有待提高；二是没有考虑诸如移动视频访问等新出现的应用中重要的可伸缩问题。本文在原有体育视频分析工作的基础上提出了针对周期性结构的得分型比赛（如网球、乒乓球）的通用分析框架。该框架的一个应用就是可伸缩视频精彩摘要，可以满足手机、掌上电脑等移动用户的需求。全文以网球和乒乓球运动为例子展开工作。

该分析框架是在充分分析现有的多模态信息融合方法并结合球拍运动周期性特点后提出的，是一个基于音视频中层特征、领域规则、采用时序分析方法形成精彩排序的通用体育视频内容分析方法，兼具了复杂度低、直观性强、通用性好、上下文相关和有感知性这五项优点。具体来说本文的工作可以分为以下几点：

首先，在体育视频中层特征提取上，本文采用了有监督的音频分类和无监督的场景聚类以适应通用性要求。体育视频中的声音鲁棒性比较好，有监督的音频分类可以做到针对一种比赛项目的通用性。当将这种方法扩展到其它比赛时，譬如跳水、棒球，采用有监督的方法也只需要少量的标注。对于视频来说，由于存在较大的场景差异，因此采用无监督的场景聚类，达到通用性的目的。本文提出了一种新的有效的场景聚类算法，无需先验知识，自动决定算法停止点。

其次，利用多模态信息融合提取了比赛结构事件。本文在详细分析了球拍运动的周期性特点的基础上，提出了一种适合于周期性结构得分型比赛的通用规则，即时域投票策略。该方法首先将标识的音频关键字按照时域对齐分散到各个聚类，然后通过投票获得每个聚类的语义以便提取结构事件。该方法充分利用了音频关键字的丰富语义和无监督的场景的可靠边界做到结构事件的准确提取。

再次，对于特定事件的提取，本文通过引入精彩排序技术解决其通用性问题。本文借助情感体验理论，设计了如下三个步骤。情感特征提取采用常用的音频、视频以及编辑手法三类特征；精彩等级的确定基于心理学实验生成事件的精彩程度事实，提出最优量化确定其数目；精彩模型的建立中提出了合理的主观感知评价标准，用于更真实地评价人的主观感知事实和计算精彩程度值之间的匹配程度，从而指导非线性精彩建模的建立。

最后，基于精彩排序技术，本文最终开发了一种可伸缩视频精彩摘要原型系统。无论是实验的结果还是用户的反馈，该框架和系统在球拍运动分析上得到了令人满意的效果。

关键词：体育视频分析，音频分类，场景聚类，多模态信息融合，时域投票策略，精彩排序，主观评价标准

Sports video analysis and highlight ranking based on audio/visual fusion

Xing Liyuan

Directed By Yu Hua and Huang Qingming

Content based sports video analysis has been paid increasing attention in recent years. Lots of work on structure analysis and events extraction has been done, but there are two limits on these methods. First, generality is not well supported as they focus on only finite kinds of video such as football and baseball. Second, they don't consider the scalability problem which is important for emerging application such as mobile video access. In this thesis, we propose a general framework of analyzing periodic structured scored game video (such as tennis, table tennis). A potential application of this proposed framework is flexible video summarization, which can satisfy the need of users of cell phone and Palm-PDA. The thesis takes tennis and table tennis for example in the following work.

This analysis framework is proposed based on adequate comparison of existing multi-model information fusion methods and the discovery of periodic characteristic of racquet sports video. It is a general sports video content analysis method based on audio/visual middle level features, domain rules, context, and highlights ranking. Its advantages include simplicity, intuitiveness, generality, context-sensitive and affectivity. The details are as follows:

Firstly, in order to extract audio/visual middle level features, we adopt supervised audio classification and unsupervised scene clustering to satisfy the requirement of generality. The audio is robust and supervised audio classification is general in the same type sports video. When we apply this method to other sports video, such as diving, baseball, only a little label work for audio is needed. Because scenes have much difference in appearance, we adopt unsupervised scene clustering for its universality, which groups video shots with similar visual content into same cluster. This thesis proposes a new effective scene clustering method which can automatically decide the stop point without prior knowledge.

Secondly, multi-model information fusion method is used to extract the structure events. By analyzing the periodicity characteristic of racquet sports, we propose a general fusion rule, temporal voting strategy, which is suitable for analyzing periodic structured scored game. It assigns labeled audio keywords to clusters according to time axis. Then we get the semantic meaning of each cluster by voting the audio keywords so that the structured events can be obtained. This method makes use of the semantic meaning of audio keywords and the confident boundary of unsupervised scene clusters, the structured events are extracted more accurately.

Thirdly, for extraction of specified events, highlight ranking is adopted in order to solve the generality problem. This thesis resorts to affective experience theory and designs three steps as follows. For highlights rank, it is determined by an optimal quantization process based on the ground truth of highlight degree. For highlight model, a reasonable subjective affective evaluation criterion is proposed, which is used for evaluating the match degree of human's subjective affective ground truth and computer's highlight rank, and used as a guide for the modeling of highlight.

Finally, based on the highlight ranking result, we developed a flexible video highlights summarization prototype system. Both the experimental results of our framework and the users' feedback of the system are encouraging in racquet sports video analysis.

Keywords: Sports video analysis, audio classification, scene clustering, multi model information fusion, temporal voting strategy, highlights ranking, subjective evaluation criterion.