武打片中的动作场景检测方法

程文刚[1,2], 柳长安[1], 须 德[2]
(1. 华北电力大学计算机学院智能机器人研究所, 北京 102206; 2 北京交通大学计算机与信息技术学院计算机研究所, 北京 100044)

摘 要: 本文提出了一种简单有效的方法检测武打片中的动作场景: 首先根据动作场景的节奏特点, 从影片层次出发, 使用镜头长度和 MPEG-7运动活力描述符定义了镜头的步调函数来度量节奏, 由此定位快节奏区域, 找到动作场景的大体位置; 之后根据动作场景的内容发展特点, 从镜头层次出发, 分析快节奏区域及周边的镜头的内容, 根据视觉特征确定动作场景的边界点. 两个层次 (影片和镜头) 信息的充分利用使得方法简单易操作, 基于压缩视频的处理方法提高了运算速度, 实验结果表明了该检测方法的有效性.

关键词: 动作场景; 节奏; 步调函数; 压缩视频
中图分类号: TP391, TN941.1 文献标识码: A 文章编号: 0372-2112 (2006) 05-0915-06

An Approach to Action Scene Detection in Martial Arts Movies

CHENG Wen-gang[1,2], LIU Chang-an[1], XU De[2]
(1. *Institute of Intelligent Robot, School of Computer, North China Electric Power University, Beijing* 102206, *China*;
2. *Institute of Computer, School of Computer & Information Technology, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract This paper presents a simple and efficient approach to action scene detection in martial arts movies. Quick tempo is an important movie-level character for action scene. A shot pace function defined by shot length and MPEG-7 motion activity is used to measure the tempo. From the shot pace change curve, it is easy to locate the rough position of action scene. According to the character of action scene development, action scene boundary is detected at the shot level by analyzing the visual contents of shots within and around the above rough region. Two clues from the movie-level and shot-level make the method simple; working on compressed video directly makes the method very fast. Experimental results based on real-world movies verify its efficiency.

Key words action scene; tempo; pace function; compressed video

1 引言

作为一种重要的类型电影 (Genre Film), 武打动作片以其精心设计的故事内容、武术造型和打斗场面, 赢得了观众特别是年轻观众的喜爱. 与其他一些类型的视频一样, 其内容具有层次性的组织结构: 底层是一个个的镜头 (Shot), 一个镜头是摄像机一次摄像的开始和结束所决定的, 由一系列连续的帧 (Frame)组成; 中层是场景 (Scene), 也叫做故事单元, 它由时间相近, 语义相关的一组镜头构成, 组成场景的这些镜头一般发生在相同的时间和地点, 出现相同的人物或事件; 高层, 一系列的场景构成了整个视频. 在这种层次结构中, 镜头是视频的物理内容单元, 信息粒度比较小; 而场景则是视频的语义内容单元, 通常只有场景才能向观众传达相对完整的语义. 动作场景是描述追杀、打架、战斗或事故等剧烈动作事件的场景[1]. 很显然, 动作场景是一部武打片的标志性内容, 也是影片最吸引观众的部分. 动作场景的检测是对视频进行高层语义结构化的工作之一, 有利于视频内容的标识, 进而有助于实现基于语义的检索; 作为影片的精彩内容, 动作场景是制作电影宣传片 (Movie Trailer) 所必需的资料; 另外, 动作场景中经常包含暴力内容, 检测它们有助于电影的分级和在视频分发时实现内容的过滤. 动作场景存在于各种动作片 (如科幻、战争动作片), 甚至其他流派的影片 (如爱情片) 中, 但它在武打片中出现最多, 也最为典型, 因此本文以武打片为例研究动作场景的检测方法.

视频场景的分割是按照语义内容的连贯性把视频划分成一个个的场景, 动作场景是一类特定内容的场景, 因此其检测可以借鉴视频场景分割的方法. 目前的场景分割方法大体可以分为两类: 基于模型的方法和基于视频制作原理的方法. 在第一类方法中, 首先需要根据特定应用或领域建立一个先验模型. Swangberg等[2]提出了这类方法的一个理论框架, 并且在新闻和体育节目的分析中得到了实现和应用, 检测准确率比较高, 但这类方法的可扩展性较差, 很难应用到电影视频场景的分割中, 更不用说内容

收稿日期: 2004-08-06; 修回日期: 2006-01-23
基金项目: 国家自然科学基金 (No. 60305009), 北京交通大学校基金 (No. 2004SM013)

组成很不固定的动作场景. 以场景转移图 (STG) 为代表的一系列算法[3,4]属于第二类, 这类算法在一定的时间约束条件下, 将视觉内容相似的镜头聚类为镜头类, 时间上有交叉的几个镜头类组成场景. 这种方法适用于情景剧、电视访谈等类型的视频, 但由于动作场景中各个镜头的视觉内容变化太大, 镜头内容的重复性较低, 而且武打片中的动作场景一般持续时间较长, 导致过检测 (即把本属于同一场景的一组镜头错误地划分为多个场景) 非常严重. 此外, 还有一些方法综合运用视听信息进行场景的检测[5,6], 这类方法一般首先将音频流短时分段, 之后把每个分段分类为音乐、语音和静音和背景音等, 视听信息的同时变化通常标志着场景的切换, 但在时间较长的动作场景中, 音频类型没有一定的模式, 经常有这样的情况: 同一动作场景中, 有的部分是急促的音乐, 而有的部分是对话或打斗声; 同时, 由于视频的多功能性, 使得视频和对应音频之间有多种关系, 因此如何有效地综合视频和音频信息是一个困难的问题[6].

还有一些工作是研究如何从视频中抽取出特定语义内容的片段 (即语义片段的抽取), 动作场景的检测实质上属于这类工作. 目前语义片段的抽取主要集中在对话场景、射门事件等的检测上, 而这些工作大多采用了基于统计模型的方法. A latan等[7]使用隐马尔可夫模型 (HMM)模拟对话场景的结构, 根据视听特征 (人脸和音频类别) 确定镜头的类别标识, HMM 对镜头标识所形成的观测序列进行识别并确定对话场景的边界; 文献 [8] 假定射门和非射门事件都包含两个连续镜头, 这两个镜头的摄像机运动参数用来建立受控马尔可夫链 (Controlled M arkov Chain), 从而实现射门场景的检测; 类似地, 文献 [9] 使用四个 HMM 模型表示棒球比赛视频中的四种“有趣 ”的场景类型, 方法的依据是棒球比赛的精彩片段是由特定类型的镜头组成的. 以上工作都取得了良好的效果, 但基于统计模型的方法能够有效抽取视频语义片段的前提条件是: 镜头内容的类型比较固定且镜头之间的切换具有特定的方式[9]. 而动作场景的内容变化过于剧烈, 很难对其中的镜头按照内容分类并统计有意义的概率, 所以这类方法难以很好地解决动作场景检测的问题.

目前, 只有 Chen等[10] 在检测对话场景的同时涉及了一种内容组成比较固定的动作场景: 二人之间的简单打斗场景, 认为这种简单打斗场景和对话场景具有相同的内容进展模式. 利用演员在图像帧中的位置和摄像机方位, 首先把镜头分为四类: *A* (只有角色 *a*)、*B* (只有角色 *b*)、*C* (同时包含角色 *a* 和 *b*)和# (其他内容). 通过观测, 总结了在场景中各类镜头之间的转换模式, 建立有限状态机模型 (FSM)来检测场景. 对话场景和打斗场景使用了相同的模型, 因此它们是不可区分的, 鉴于动作场景中的镜头一般比较短, 最后使用平均镜头长度区分这两种场景. 然而, 实际的动作场景, 内容类型繁多, 不仅仅是这种二人之间的简单打斗场景, 镜头内容的组成也远比这种简单打斗场景更复杂; 并且, 如果动作剧烈或武打造型比较多, 即便是二人之间的打斗, 也不具有与对话类似的内容进展模式.

此外, Nam等[11]分别从视频流中提取小波系数、从音频流中提取能量特征, 并将这两种媒质特征融合在一起, 通过阈值判断去识别暴力血腥内容; 文献 [12] 使用分层的模型来识别视频中的爆炸片段. 这两个工作都取得了不错的效果, 但无论是暴力血腥内容还是爆炸片段, 都只是动作场景中的部分内容, 并不能与动作场景划等号: 动作场景的时间跨度更大, 内容也更丰富.

根据以上分析可知, 如何从一个长的视频序列中检测出动作场景是一个新的研究问题, 而现有的方法并不能很好地推广到这个问题的解决中. 事实上, 无论是视频场景的分割还是语义片段的抽取, 目前的方法实质上都采用了自底而上, 即从镜头构造场景或者语义片段的方法, 这种模式固然有利于确定边界, 但却没有充分应用视频的整体信息. 我们从文献 [13] 利用电影语法分析影片节奏的工作中得到启发, 利用节奏这一重要的整体信息辅助动作场景的检测, 提出了一种简单有效的动作场景检测方法: 分别从高层 (视频) 和底层 (镜头) 所提供的节奏和视觉内容信息进行分析, 根据动作场景的节奏特点和内容进展特点, 完成中层——动作场景的定位和检测.

2 检测方法

2.1 方法概述

节奏 (tempo) 是符合人感知的一个重要信息, 而动作场景是武打片中节奏最快的内容片段, 因此节奏为动作场景的检测提供了有益的线索. 步调 (pace) 是节奏的重要表现形式, 根据电影制作和剪辑原理, 定义了一个步调函数来描述内容进展的步调, 进而根据步调函数值的分布定位快节奏的区域, 从而找到动作场景所在的大体位置. 在武打片中, 一个动作场景通常描述一个动作事件, 事件会经历发生、发展和结束等阶段, 其中总有人物或地点的重复出现, 因此, 在大体位置确定的基础上, 通过分析快节奏区域及其周边镜头的内容, 来确定动作场景的最终边界.

2.2 节奏分析

“节奏在一部影视片中起着举足轻重的作用. 内容和片种的不同, 导致了结构和节奏的不同. 即使在同一影片中的不同内容段落, 也会在节奏的总谱中, 产生迥异的节奏 ”[14]. 在同一部武打片中, 描述对话、打斗、风景等不同内容的片段对应着不同的节奏. 为了不致使观众感到迷惑进而产生误解, 后期剪辑的时候一般不会把两个快节奏且表达不同内容的场景顺序组接. 这就使得动作场景与其周围的内容片段具有不同的特点, 使得节奏具有较强的辨识动作场景的能力. 但节奏的影响因素有很多, 且具有较大的主观性, 准确的度量比较困难. 文献 [14] 指出“速度是节奏的重要表现形式之一. 速度一方面表现在内容上, 高速奔

跑、追逐、打斗、抢险之类等情节, 本身就表现‘高速度’; 风光、抒情、恋爱、谈心的场面, 往往是低速度或中速度. 速度另一方面表现在镜头尺寸上, 即在屏幕上滞留的时间”, 其中速度指的是观众所体会到的内容进展步调的快慢程度. 因此, 可以通过度量步调来表征节奏. “节奏的构思阶段是从全片、场景进入单个镜头; 节奏的完成阶段又是从单个镜头、场景直到构成全片”[14]. 由此可知, 镜头是度量步调的基本元素. 步调的影响因素中, “高速奔跑”、“追逐”、“抒情”等内容可以使用运动信息描述其特点, 而镜头尺寸则直接对应了镜头长度, 因此可以定义一个镜头的步调函数, 使得这个函数能根据镜头的运动和长度反映该镜头步调的快慢程度. 根据人的主观感知, 镜头的运动越剧烈, 镜头长度越短, 步调应越快, 反之亦然.

MPEG-7运动活力 (Motion Activity)直观地描述了视频片段的“动作强度”或“动作步调”, 它提供了运动的强度、方向、空间分布、空间位置和时间分布等内容. 其中, 活力强度(Intensity of Activity)是一帧中运动矢量幅度的标准差, 它表明了运动向量幅度的一致性, 其突出的性能已经被实验验证[15], 在此使用帧的活力强度表示其运动活力. 定义镜头 $Shot_i$ 的运动活力 IAS_i 为:

$$IAS_i = \left(\sum_{j=1}^{N_i} IAF_j \right) \Big\backslash N_i \qquad (1)$$

式中 N_i 是镜头 $Shot_i$ 中的 P 帧总数, IAF_j 是第 j 个 P 帧的活力强度.

定义镜头 $Shot_i$ 的步调函数为:

$$Pace(Shot_i) = \frac{IAS_i / MaxIAS}{ShotLen_i / MaxShotLen} \qquad (2)$$

式中 $ShotLen_i$ 是 $Shot_i$ 的镜头长度, 即 $Shot_i$ 内的总帧数; $MaxIAS$, $MaxShotLen$分别表示属于同一视频的所有镜头的最大运动活力值和最大镜头长度, 用于对 IAS_i 和 $ShotLen_i$ 标准化. 由于本文实验中使用了较长的测试视频序列, 为了排除噪声的干扰, 首先对运动活力序列和镜头长度序列分别执行阿尔法裁减, 去掉其 5% 的最大值, $MaxIAS$ 和 $MaxShotLen$ 从剩余值中选取.

电影是一种历时性艺术, 因此影片中某个镜头体现出的步调总是受到“语境”, 即这个镜头周围的镜头所表达的内容的影响, 为了体现这个特点, 同时为了减少噪声, 我们对式 (2)计算出的值平滑处理, 把滤波后的值作为镜头实际的步调值. 实验中, 10个镜头宽度的中值滤波给出了较好的描述效果.

图 1中的 (a)、(b)、(c)分别是动作片《精武英雄》上集 722个镜头的镜头长度、运动活力、步调的变化曲线. 其中方框表示影片中各个动作场景对应上述三个值的分布位置示意. 从中可以看出: 属于动作场景的大部分镜头的运动活力值很大, 而镜头长度值很小, 致使这些镜头的步调值要远高于其他内容的镜头的步调值.

寻找步调值较高的一个连续片段可以定位动作场景的位置. 但是由于函数 $Pace(Shot_i)$ 理论上没有上界, 为方便检测, 把其变换为:

$$g(Pace(Shot_i)) = \exp(-Pace(Shot_i)) \qquad (3)$$

称式 (3)的值为镜头的步调变换值. 由于 $Pace(Shot_i) > 0$, 显然函数 $g(x)$有性质: $0 < g(x) < 1$; $g(x)$是减函数, 且随着 x 增大, $g(x)$递减的幅度变小, $g(x)$趋近于 0. 属于动作场景的大部分镜头的步调值要远高于表示其他内容的镜头的步调值, 使得这些镜头的步调变换值比较小, 接近零. 与图 1(c)对应的图 1(d)表示了这个特点. 因此, 只需设置一个合理的阈值 T_{pt} 便可以检测镜头步调变换值连续较低的区域 R. 为了减少动作场景中个别镜头的影响, 对相邻的任意两个区域 R_a 和 R_b, 如果相距很近 (不超过 W_{ab} 个镜头, W_{ab} 是经验参数, 实验中 $W_{ab} = 9$), 则将 R_a 和 R_b 合并连接为一个连续的区域 R_c. 经过区域检测和合并, 形成最终的区域 R, 确定了每个动作场景在整个视频序列中所处的大致位置.

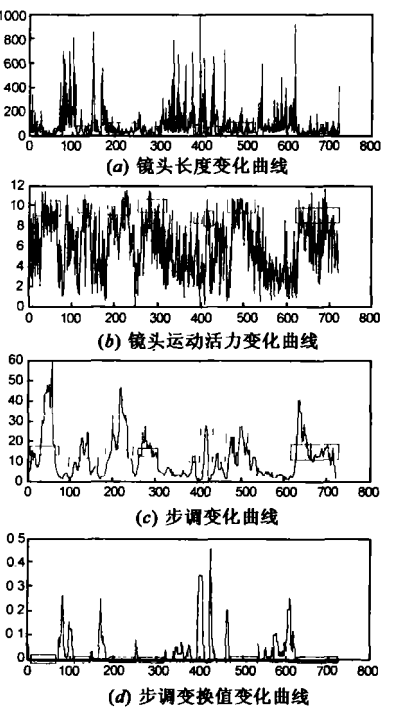


图 1　《精武英雄》上集 722个镜头的各项变化曲线

2.3　边界点检测

武打片中的动作场景通常只有一个动作剧烈的事件, 组成该场景的各个镜头反映了动作事件的不同侧面. 时序性是动作场景的一个重要特点: 事件有其发生、发展和结束的历程, 特殊的时候, 事件的发生或结束蕴涵在发展过程中. 图 2表示的是一个动作场景的内容进展过程, 其中, L_s 是开始阶段, L_p 是发展阶段, 而 L_d 则是动作事件的结束, 这三个阶段的界限并不一定非常明确. 作为发展和高潮部分的 L_p 动作剧烈; 阶段 L_s 用于交代场景中的人物和空间关系, 有时采用创建镜头 (Establishing Shot)或涵盖镜头 (Cover Shot), L_d 则是动作事件的结果, 首尾这两个阶段的动作一般不剧烈. 同时由于武打片中动作场景的设计是为了达到武术特技的展示、速度和运动的体现等

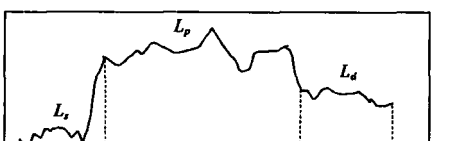


图 2　动作场景内容进展示意图

艺术效果, 因此阶段 L_p 持续时间最长, 而阶段 L_s 和 L_d 作为辅助理解的内容, 持续时间很短, 如影片《醉拳 II》结尾处的打斗场景近 20分钟, 而与阶段 L_s 和 L_d 对应的镜头数仅为 15和 2个.

根据 2.2节区域 R 中镜头的特点 (运动剧烈、镜头长度短)以及与实验结果的对应, 区域 R 大体相当于图 2中的阶段 L_p, 而动作场景发生和结束的过程却没有给出. 边界点的检测即分别确定阶段 L_s 和 L_d 的开始和结束位置. 为了保持语义的连贯型, 在动作事件的发生和结束阶段, 总会与其发展过程中出现相同的人物、地点以及连续的动作, 导致相似内容的镜头重复出现, 这样才能表明其中的各个镜头属于同一个场景, 使观众明白这个片段描述是同一语义内容. 由于相同的人物或者地点可以通过视觉相似性衡量; 又由于阶段 L_s 和 L_d 都比较短暂, 可以只需对 L_p 邻近的一些镜头进行检测即可, 因此, 我们将 2.2节检测出的区域向外扩展, 在两侧的镜头窗 (由于窗宽使用镜头数目度量, 称为镜头窗)W_l 和 W_r 中, 根据视觉内容的相似性确定场景的边界点, 如下图.

根据上述分析, 边界点的检测可以通过分析与阶段 L_p 内容相似的镜头在镜头窗 W_l 中出现的最早位置和在 W_r 中出现的最后位置来确定.

由于颜色特征具有较强的辨识能力, 它经常用来表示图像的内容. 类似于目前的大多数场景分割算法, 本文也使用颜色直方图描述帧的视觉内容. 定义两帧 f_i 和 f_j 之间的视觉相似度为:

$$SimFF(f_i, f_j) = \sum_{l=1}^{bins} \min(Hf_i(l), Hf_j(l)) \quad (4)$$

式中 $bins$ 表示直方图的格 (bin)总数; Hf_i 和 Hf_j 分别是两帧 f_i 和 f_j 的归一化直方图. 由于 HSV 颜色空间与人的视觉感知系统有较好的一致性, 本文选用 HSV 颜色空间来表示帧的颜色分量 (把 H 分成 16份, S 分为 4份, 由于 V 对光照敏感, 不使用 V).

为了减少冗余信息和计算量, 经常使用关键帧简单有效地表示镜头的内容. 根据动作场景中镜头长度比较短的特点, 同时考虑到本文实验中各个测试视频片段均包含大量的镜头, 我们简单地选取一个镜头的首尾两帧作为该镜头的关键帧, 从而使得这两帧在具有一定镜头内容概括能力的同时, 减少计算量. 定义两个镜头 $Shot_i$ 和 $Shot_j$ 之间的相似度为:

$$SimSS(Shot_i, Shot_j) = \max_{k \in KF_i, l \in KF_j} (SimFF(k, l)) \quad (5)$$

式中 KF_i 和 KF_j 分别是对应 $Shot_i$ 和 $Shot_j$ 的关键帧集合; k 和 l 是相应关键帧集合中的任意关键帧. 在一个视频片段中, 各个镜头的内容随时间变化, 经常会出现这样的情况: 两个镜头之间仅有部分相同的内容, 但实际上它们描述的是同一语义主题. 根据这个特点, 式 (5)使用 max函数来合理地度量镜头之间的内容相似性.

这样, 镜头窗 W_l 或 W_r 中的任意镜头 $Shot_u$ 与阶段 L_p 的相似度可以计算为:

$$SimSL(Shot_u, L_p) = \max_{Shot_v \in L_p} (SimSS(Shot_u, Shot_v)) \quad (6)$$

即 $Shot_u$ 与阶段 L_p 中最相似的镜头之间的相似度.

场景开始点的检测时, 对于镜头窗 W_l 中的镜头, 按照镜头出现的时间顺序逐个计算该镜头与片段 L_p 的相似度, 当足够相似时 (相似度超过阈值 T_{sl}时), 可以认为动作事件的参与人物或者发生地点出现了, 标志着一个动作场景的开始; 在特殊的时候, 由于动作事件没有明显的开始阶段 (如动作场景的开始就是打斗镜头), 此时镜头窗 W_l 中的所有镜头与片段 L_p 的相似度都比较低, 这种情况下, 我们把阶段 L_p 的开始点作为动作场景的开始位置. 同理, 类似的方法可以进行场景结束点的判断. 图 3中虚线箭头表示检测方向.

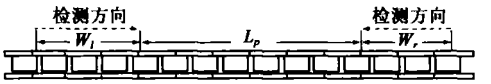


图 3 边界点检测镜头窗设置示意图

3 实验结果与分析

3.1 实验数据集

实验数据集如表 1所示, 所有片段均以 MPEG-1压缩视频流形式存在, 帧率 25 fps, 帧分辨率 352×240. 前 4个取自经典的香港武打片, 原因是: (1)目前的大部分武打片属于香港武打片, 这些片段具有一定的代表性; (2)这些片段具有较多的动作场景, 易于检测本文算法的有效性. 其中, 片段Ⅳ属于早期武打片 (70年代初), 拍摄镜头较长, 因此与其他三个武打片段相比, 镜头数目要少很多. 最后 2个片段分别取自恐怖片和战争动作片, 用于测试本文算法的适应性.

3.2 实验过程与参数设置

实验分为预处理 (包括镜头边界的检测、修正以及评

表 1 实验数据集

编号	名称	来源	类型	时间长度	帧数	镜头数
Ⅰ	Fist of Legend_1.mpg	《精武英雄》上集	武打片	00:46:14	69372	722
Ⅱ	Martial Arts of Shaolin_2.mpg	《南北少林》下集	武打片	00:39:23	56667	515
Ⅲ	Drunken master_2.mpg	《醉拳Ⅰ》下集	武打片	00:51:53	77830	684
Ⅳ	Fist of Fury (Bruce Lee)_1.mpg	《精武门》上集	武打片	00:49:12	73823	157
Ⅴ	MummyReturn_2.mpg	《木乃伊归来》下集	恐怖片	00:50:27	75686	368
Ⅵ	Braveheart_B.mpg	《勇敢的心》下集	战争动作片	01:20:00	120000	515

价标准的建立等)、检测(包括节奏分析和边界点确定)和结果分析(检测结果的绘制和统计分析)等三个阶段. 武打片中镜头之间的衔接绝大多数使用切变(cut), 而目前的大多数镜头边界检测算法对切变检测的准确率均很高, 本文使用文献[16]的算法检测镜头边界, 该算法利用 DCT 系数进行帧间比较, 具有计算量小, 速度快的优点. 为了不影响后续的动作场景检测, 我们对错误的镜头边界进行了修正. 统计一个镜头中的所有 P 帧的运动矢量可以快速的计算镜头的运动活力. 由于步调值经过滤波去噪, 且在动作场景的发展阶段 L_p 中时, 大多数镜头的步调变换值均很低, 使得 2.2节中的阈值 T_{pt} 很容易确定: 从几个视频片段的步调函数曲线中, 我们观测到动作场景区域的步调值均在 7以上, 根据步调变换函数(公式(4))可以计算出, 这些区域的步调变换值均小于 0.001, 对比测试了其他的一些值, T_{pt} = 0.001时的实验效果比较满意. 由于武打片中动作场景的开始和结束片段都比较短, 2.3节中镜头窗 W_l 和 W_r 均为 18个镜头的宽度时便可以很好的完成边界点的检测; 镜头相似性比较中的阈值 T_{sl} 是个经验值, 通过对不同数值进行测试的结果分析, T_{sl} = 0.83时结果比较好.

3.3 结果与性能分析

动作场景是一个语义级别的概念, 不同的浏览者可能理解不同, 特别是对于场景边界的确定, 主观性比较强. 为了减少这种主观性, 实验中通过四个人的共同商讨确定所有动作场景的实际位置和边界, 作为标准(Ground Truth). 在片段 V 和 VI中, 动作场景也包括战争和恐怖事件发生的情节. 图 4 分别给出了各个视频片段的步调变换值曲线, 检测出的动作场景片段使用"*"标识在曲线上.

表 2是检测结果的统计. 其中, 检测场景数、误检和漏检等三项不要求特别精确的场景边界, 只要是检测到的动作场景位于实际的动作场景边界中即可, 用于测试 2.2节算法的效果, 采用查全率(Recall)和精确率(Precision)来评价, 定义为:

$$\begin{aligned}&\text{查全率} = \text{正确检测数}/(\text{正确检测数} + \text{漏判数})\\&\text{精确率} = \text{正确检测数}/(\text{正确检测数} + \text{错判数})\end{aligned} \tag{7}$$

属于动作场景的镜头一般长度都很短, 并且动作场景包含的镜头数目通常比较多, 因此在实验中, 当检测到的场景边界与实际边界不超过 5个镜头时, 我们也认为其边界检测是正确的, 表中"边界确定准确的场景数"是对于所有正确检测到的动作场景, 其边界检测也准确的个数统计, 用于检测 2.3节算法的效果, 采用准确率来评价, 定义为:

$$\text{准确率} = \frac{\text{边界确定准确的场景数目}}{\text{实际场景数目}} \tag{8}$$

表 2　动作场景检测的统计结果表

视频编号	实际场景数	检测场景数	误检	漏检	边界确定准确的场景数	计算时间
Ⅰ	9	10	1	0	9	00:26:17
Ⅱ	5	6	1	0	4	00:20:39
Ⅲ	6	8	2	0	4	00:26:20
Ⅳ	4	3	0	1	2	00:19:18
总和(动作片)	24	27	4	1	19	——
Ⅴ	4	3	0	1	1	00:20:32
Ⅵ	6	6	0	0	4	00:34:20
总和(全部)	34	36	4	2	24	——

分析数据可见, 动作场景的检测方法是有效的. 误检和漏检的情况均比较少. 但由于个别不属于动作场景的片段具有类似于动作场景的特点, 如片段 I 中"陈真在霍元甲灵堂中"的场景, 采用了动作剧烈的几个短镜头, 以表达愤怒的心情, 这种情况下容易造成误检. 相对于误检, 漏检的情况很少出现. 表中"检测场景数"是利用节奏定位动作场景的结果, 达到 93.75% 的查全率和 88.23% 的精确率; 边界确定的准确率(武打片 79.17%, 全部 70.59%)也较高. 片段 V 和 VI的检测结果说明方法具有一定的适应性. 无论是镜头检测还是运动活力的提取, 都工作在 MPEG压缩视频流上, 充分利用了压缩域的信息, 大大减少了计算量; 在检测出动作场景所在大致区域的基础上进行场景边界的确定, 大大减少了镜头相似性比较的运算范围, 这都使得方法具有较高的效率. 表 2中"计算时间"一项统计了检测阶段的时间花费, 这些数据都是在 Founder微机(主频为 1.6GHz的 Intel Pentium4 CPU, 512MB的 DDR内存)上运行测得的, 操作系统为 Win2000 Server, 实验系统开发工具为 VC 6.0

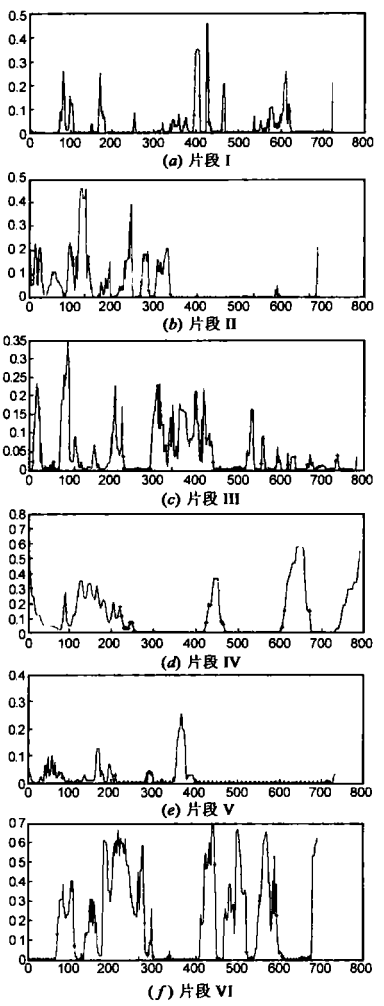


图 4　各个实验片段的步调变换值曲线与动作场景检测结果

4 结论

本文提出了一种不需要复杂模型, 简单有效的方法检测武打片中的动作场景: 首先从影片层次出发, 根据节奏的变化, 定位快节奏区域, 从而确定动作场景的大体位置;

之后,从镜头层次出发,对检测到的区域及其周围的镜头进行内容分析,从而确定动作场景的起止点.实验结果表明:使用镜头长度和运动活力定义的镜头步调函数能够很好地描述影片内容进展的节奏;边界确定算法是有效的;充分利用压缩域信息极大地减少了计算量,这些都保证了动作场景检测算法的性能.然而,该方法存在受限于阈值的缺陷,降低了自动化的程度,有待改进.对动作场景的内容进行更细粒度的划分是我们进一步的工作.

参考文献:

[1] Arijon D. Grammar of the film language[M]. Silman-James Press, 1976: 483.

[2] Swanberg D, Shu CF, Jain R. Knowledge guided parsing in video databases[A]. Proc of SPIE Conf on Storage and Retrieval for Image and Video Databases[C]. San Jose: SPIE Press, 1993. 173-187.

[3] Yeung MM, Yeo B L, Liu B. Segmentation of video by clustering and graph analysis[J]. Computer Vision and Image Understanding, 1998, 71(1): 94-109.

[4] Ngo CW, Ma YF, Zhang H J. Automatic video summarization by graph modeling[A]. Proc Of IEEE ICCV'03[C]. Washington: IEEE Computer Society Press, 2003. 104-109.

[5] Sundaram H, Chang SF. Computable scenes and structures in films[J]. IEEE Trans on Multimedia, 2002, 4(4): 482-491.

[6] Chen SC, Shyu ML, Liao W, et al. Scene change detection by audio and video clues[A]. Proc of IEEE ICME'02[C]. Lausanne: IEEE Computer Society Press, 2002. 365-368.

[7] Alatan AA, Akansu AN, Wolf W. Multimodal dialog scene detection using hidden markov models for content-based multimedia indexing[J]. Multimedia Tools and Applications, 2001, 14(2): 137-151.

[8] Xie L, Xu P, Chang SF, et al. Structure analysis of soccer video with domain knowledge and hidden markov models[J]. Pattern Recognition Letters, 2004, 25(7): 767-775.

[9] Chang P, Han M, Gong YH. Extract highlights from baseball game video with hidden markov models[A]. Proc of ICIP'02[C]. New York: IEEE Press, 2002. 609-612.

[10] Chen L, Ozsu M. Rule-based scene extraction from video[A]. Proc of IEEE ICIP'02[C]. New York: IEEE Press, 2002. 737-740.

[11] Nam J, Alghoniemy M, Tewfik A H. Audio-visual content-based violent scene characterization[A]. Proc of IEEE ICIP'98[C]. New York: IEEE Press, 1998. 353-357.

[12] 庄越挺,傅正钢,叶朝阳,吴飞.基于视听分层模型的实时爆炸场景识别[J].计算机辅助设计与图形学学报, 2004, 16(1): 90-97.

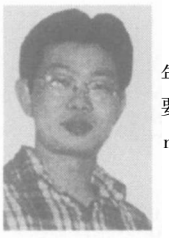
[13] Adams B, Dorai C, Venkatesh S. Novel approach to determining tempo and dramatic story sections in motion pictures[A]. Proc of IEEE ICIP'00[C]. New York: IEEE Press, 2000. 283-286.

[14] 傅正义.电影电视剪辑学[M].北京:北京广播学院出版社, 2002.

[15] Peker K, Divakaran A. Framework for measurement of the intensity of motion activity of video segments[J]. Journal of Visual Communications and Image Representation, 2003, 14(4): 265-284.

[16] Zhang H J, Low CY, Smoliar SW. Video parsing and browsing using compressed data[J]. Multimedia Tools and Applications, 1995, 1(1): 89-111.

作者简介:



程文刚 男,1977年出生于山东泰安,2005年获北京交通大学计算机应用技术博士学位,主要研究方向为视频内容分析和计算机视觉.E-mail:wengangcheng@sina.com

柳长安 男,1971年出生于黑龙江拜泉,2001年获哈尔滨工业大学计算机应用技术博士学位,主要研究方向为人工智能.

须 德 男,1944年出生于江苏常州,北京交通大学教授、博士生导师,目前主要研究方向为数据库系统及其应用和多媒体信息处理.