面向事件的影片摘要生成方法

王 辰[1] 刘桂清[2] 老松杨[1] 蒋 杰[1]

[1](国防科技大学信息系统与管理学院,长沙 410073) [2](国防科技大学电子科学与工程学院,长沙 410073)

摘 要 对视频摘要的研究已成为视频应用领域十分活跃的课题。为了获得更加实用的视频摘要,在介绍视频摘要概念和用途的基础上,依据目前的研究状况和影片类视频的特点,提出了一种适合于故事类影片的面向事件的影片摘要生成方法,并首先对该方法的细节进行了描述。该方法还使用了一种场景重要程度的评价算法,此算法的特点在于综合利用了影片中的多种媒体特征,其不仅考虑了运动特性,还考虑了感人内容对场景重要程度的影响。通过对摘要进行可用度测试和分析的结果,表明,此种方法生成的影片摘要具有较好的可用度。

关键词 影片摘要 视频事件探测 视频内容分析 音频内容分析 视频检索

中图法分类号: TP391.3 TN941.1 **文献标识码**: A **文章编号**: 1006-8961(2005)05-0642-08

Event-oriented Film Abstract

WANG Chen[1], LIU Gui-qing[2], LAO Song-yang[1], JIANG Jie[1]

[1](*School of Information System and Management, National University of Defense Technology, Changsha* 410073)

[2](*School of Electronic Since and Engineering , National University of Defense Technology, Changsha* 410073)

Abstract Now video abstract is an active research topic in the fields of video application. This paper first introduces the concepts and applications of video abstract. Schema for event-oriented film abstract described in this paper uses feature of films for reference. In this paper, the methodological detail is discussed, and a novel evaluation model of scene significance is presented. The evaluation integrates different video characters, including motion feature as well as impressive infection. The experimental result indicates that the film abstract generated using the event-oriented method is comparatively usable.

Keywords film abstraction, video event detection, video content analysis, audio content analysis, video retrieval

1 引 言

众所周知,对于一篇文章来说,摘要就是对文章的一个简单概述,它可以使人们快速了解文章的主要内容。文章的摘要已经被广泛地应用于对文献的检索中,一般情况下,用户完全可以只通过阅读文章的摘要,就能判断出该篇文章是否是自己感兴趣的。将这一思想应用于对视频的浏览和检索当中,就产生了视频摘要(video abstract)技术。

在对视频的查询与浏览过程中,面临的一个主要的问题是如何对视频信息加以表现。如果用户需要从查询结果中一段段地观看视频之后,才能从中找到自己想要的视频,这在视频查询中将是不可想象的。因此,对于视频的查询和浏览来说,特别是在 Internet 等带宽有限的条件下,如能用较少的数据量来代表视频流,以便使用户能够快速了解视频的内容,则显得尤为重要。

视频摘要就是以自动或半自动的方式,通过对视频的结构和内容进行分析来从原视频中提取有意义的部分,再将它们以某种方式合并而成的紧凑的、能充分表现视频语义内容的视频概要。视频摘要可以有多种媒体形式和表现形式,它可以是一段文字、一幅图像或图像组合,也可以本身就是一段视频或者由多种媒体组合成的多媒体文档形式。一般来说,视频摘要有以下几种类型:

基金项目:国家"863"高技术研究发展计划项目(2001AA115123)

收稿日期:2004-03-30;**改回日期**:2004-12-20

第一作者简介:王辰(1973 ~),男,讲师。2002 年获得国防科技大学系统工程专业博士学位。主要从事多媒体信息系统、数字视频分析处理及基于内容检索等方面的研究工作。已发表学术论文 20 多篇。获部委级科技进步三等奖 3 项。E-mail:wangchen@x263.net

(1) 文字描述(textual description)

这种方式是最紧凑的视频摘要形式,它非常便于用户理解和建立索引,但很难由计算机自动生成能准确概括视频内容的文字描述,其一般采用人工输入的方式,也可通过识别视频标题或视频中的其他注释文字来获得。

(2) 视频代表帧(video keyframe)

这是一种使用较多的视频表现形式,它是一幅从视频片段中抽取的图像,但是由于无法表现视频的时间和动态特性,因此这种方式在视频检索中多用于表示镜头或场景。由于对这种形式的视频摘要的研究起步较早,因此技术比较成熟,已有大量的文献介绍了这方面的内容。

(3) 情节串连图(filmstrip/storyboard)

这种摘要十分类似于电影海报,它是由一组从视频片段中抽取的图像,按照时间顺序组合而成,此类摘要有些用图像不同的大小来表示视频中相应内容的重要程度。此类摘要不仅可以向用户给出视频情节的总体描述,而且在浏览过程中可以方便地定位到视频中感兴趣的部分。

(4) 缩略视频(video skim)

这种摘要是由视频中的一些片段拼接而成,或者是由视频中的图像序列和声音片段合成得到。查询时,用户可以通过播放这些相对短小的视频片段来了解整个视频的内容,如电影预告片/宣传片(preview/trailer)就属于这一类。

(5) 多媒体视频摘要(multimedia video abstract)

其是由多种媒体形式组成的视频内容表现方式。它是将文字、图像、声音、视频等媒体综合集成在一起来表现视频的主要内容,例如,在一个 HTML(hyper text mark language)的页面中,可以包含文字形式的视频名称、简介,图像形式的演员照片、场景图,声音形式的精彩对白,视频形式的精彩片段等。这种多媒体视频摘要的生成可以基于其他形式的摘要生成技术,以便给用户以更加完整而丰富的视频内容表现,同时为用户提供多种浏览和检索视频的方式。

可以注意到,对于文章来说,摘要与原文是同一种媒体形式,即文字,但对于富含多媒体信息的视频来说,缩略视频和多媒体摘要则应该是最具表现力的摘要形式,因为,表现媒体的差异,往往会带来较多的信息丢失。缩略视频摘要的自动生成是视频摘要技术研究的热点和难点,也是本文重点关注的内容。由于影视节目在视频类媒体中占有很大的比例,同时也是应用十分广泛的视频类型,因此为影片自动生成视频摘要可以应用到很多领域,如

(1) 影视资料库

随着数字电影、电视的不断推广应用,电视台和电影公司将有大量的影视节目被数字化,并归档保存。而对于存档的数字视频来说,有效的检索方法是一个十分重要的问题,而视频摘要可以用于对数字化存储的视频材料进行索引和检索。

(2) 电影市场

电影海报和电影预告片是作为电影广告来介绍新影片的,可是制作这一类广告往往是费时、费力的,而视频摘要技术则不仅可以由计算机自动生成影片的广告素材,而且可以依据需要来生成满足不同要求的电影预告。

(3) 家庭娱乐

视频摘要可用于数字化的电视杂志,还可以取代文字形式的电视节目介绍,并能提供更加直观的预告。在播放连续剧的时候,视频摘要可以作为前一集内容的提示,让你快速地回顾那一集里都发生了什么。在视频点播系统中,视频摘要又可为用户选取节目提供方便。

2　影片摘要与相关研究

本文研究的影片摘要是一种缩略视频形式的视频摘要。所谓缩略视频指的是一种高度压缩源视频数据的视音频片段,它具有内容丰富、直观,表现力强的特点。

2.1　缩略视频

根据摘要内容和目的的不同,缩略视频摘要又分为总结摘要和亮点摘要两种,其中总结摘要提供给用户关于整个视频内容的介绍,例如用于影片介绍的电影剪辑,它会较为完整地介绍影片的主要内容;而亮点摘要则介绍最吸引人的部分,比如用于广告的电影宣传片,它呈现了影片中的许多精彩场面,而并不揭示故事的结局。

从目前的研究来看,主要有以下几种生成缩略视频的方法:

(1) 简单的生成方法

该方法是基于时间对视频进行采样,即每间隔一定的时间抽取一个代表帧或者一个片段,例如,可以抽取视频的前 10s 来获得缩略视频,这种生成方

法虽很容易实现,但由于完全没有基于视频的内容,因此效果很不可靠。

(2) 基于视觉信息的生成方法

该方法是根据视频中颜色、纹理、形状、运动方向和强度等视觉信息,同时基于模式识别的思想,应用各种视频和图像处理技术来进行镜头探测、关键帧提取、场景聚类、运动特征提取等一系列操作,最终生成具有代表性的关键帧序列或缩略视频。这种算法完全基于视觉特征,而忽略了音频、字幕等信息对表现视频的作用。

(3) 融合多特征的生成方法

该方法在基于视觉特征的方法的基础上融入其他媒体提供的信息,以便更准确地判断视频片段的重要程度,例如,采用人脸识别技术来探测新闻中重要人物的出现,采用音频处理技术来探测体育视频中的精彩片段等等。这种算法是目前研究的热点,且当前大多数视频摘要方法都是基于这种思想。

(4) 基于视频句法语义的生成方法

此方法不追求用模式识别的方法来获取视频中的数据语义,而是从视频的句法结构分析入手,通过探寻镜头与镜头之间、场景与场景之间的结构规则,试图从中分析出编导人员借此表现的情感和氛围[1],并以此为基础,将视频的句法语义尽可能完整地保存到摘要当中。这是一种新的思路,在视频模式识别技术还远远不够完善的今天,这种方法不失为生成视频摘要的一个新途径。

2.2 相关研究

1994 年,Carnegie Mellon 大学就开始开发 Informedia 视频数据库[2],是视频摘要研究领域的先驱;之后,Columbia 大学、Philips 研究院、微软研究院、AT&T 实验室、IBM Almaden 研究中心、德国曼海姆大学和加州大学 Berkeley 分校等大学和机构都展开了此方面的研究,并开发了多种形式的摘要和各式各样的生成算法。

近年来,基于多模态融合视频分析生成摘要的方法越来越被人们所重视,可以说目前正在进行的视频摘要研究大部分都是基于这类算法的。Columbia 大学、Philips 研究院、南加州大学、意大利 Napoli 大学、微软研究院都做过类似的系统,其大体思想都是先利用多种视、音频特征来对视频中的重要内容进行判断,其包括:(1)是否有文字、人脸等重要对象的出现;(2)是否有摄像机的运动;(3)伴音中是否出现关键词语等;然后,通过建立重要度评判模型将这些因素综合起来把视频摘要生产问题抽象成一个求解最优化的问题,并在此基础上抽取一些重要的视频帧和音频片段来生成摘要。当然,生成过程中还必须考虑一些其他因素的约束,比如客户端的处理能力、客户端的浏览模式、与其他媒体的同步性以及最终结果的可理解性等因素。

在德国曼海姆大学开发的 VAbstract 系统中[3],最精彩的电影片段被提取出来用于生成电影宣传片。具体来说,他们首先用图像的高对比度来检测含有重要物体/人物的帧;然后将动作场景用大的帧差来识别;接着由整个影片平均颜色组成相近的帧被包含到摘要中,并期望它们能反映影片的基调,此外,对话场景的识别由检测语言中的'a'字母来完成,这是由于在多数语言中,'a'字母出现得较频繁;最后,所有选中的场景(除了影片的结尾部分)以原来的时间序排列组成宣传片的内容。这种算法有很多有趣的观点,但是算法的有些部分还太简单,有待于改进,笔者也缺乏全面的实例来证明他们的结论。VAbstract 的一个改进版本——MoCA[3] 则利用了一些特殊的事件,比如主演的特写镜头、爆炸场面和枪声来帮助识别重要的场景。

基于视频句法语义的摘要生成方法,其典型的代表是 Columbia 大学的 Sundaram 等人提出的基于句法语义的效用模型[1]。Sundaram 等人认为,由于目前的模式识别技术还无法实现准确有效的从视频中抽取人们想要的文字、人脸等信息,所以基于这些模式识别技术的方法就可能有很大的不准确性;其次,由于模式库显然无法满足包含纷繁复杂的各种视频中出现的所有模式的要求,所以基于模式识别的方法势必产生很多遗漏。事实上,由于视频同文本类似,同样具有特定的句法结构,只要掌握了句法结构规则,即使是对于用人们不认识的文字写成的文章,也可以抽取出它的摘要,因此,Sundaram 等人致力于分析视频的句法结构,并以此句法结构为基础来生成摘要。

2.3 本文的方法

虽然在视频模式识别技术还远远不够完善的今天,基于句法语义的视频摘要生成方法可以说是一种新的解决问题的思路,但是它更适用于自动生成那些摘要长度已知,且为满足被动需求的视音频摘要。也就是说,该方法更适用于用户不知道原始视频的内容,并且对信息的需求是未知的情况。另外,这种方法需要对视频的编辑手法和技巧有深入的了解,因为这种方法生成的摘要都希望尽量保留视频

编辑时所融入的“语法”[4]。

对于影片摘要来说,用户的需求是明确的,即用户希望影片摘要包含影片的精彩内容。同时,由于电影制作过程十分复杂,因此试图从影片的编辑规律中找出明确的“语义”,并且合理地计算视频元素的效用是十分困难的。在众多的影视节目中,故事影片(feature)以讲述故事为特征,其影片内容由多个环环相扣的情节构成,而每段情节是通过一系列事件加以表现的,针对这一特点,本文提出了一种面向事件的影片摘要生成方法。

大家知道,影片摘要的生成是以影片内容分析为基础的,只有对影片的内容进行有效分析之后,才能以最紧凑的方式生成能够充分表达影片内容的摘要,也就是在影片分析的基础之上,通过选取具有代表意义的影片成分来最终合成为影片摘要。图 1 是面向事件的影片摘要和多媒体影片摘要的生成过程。

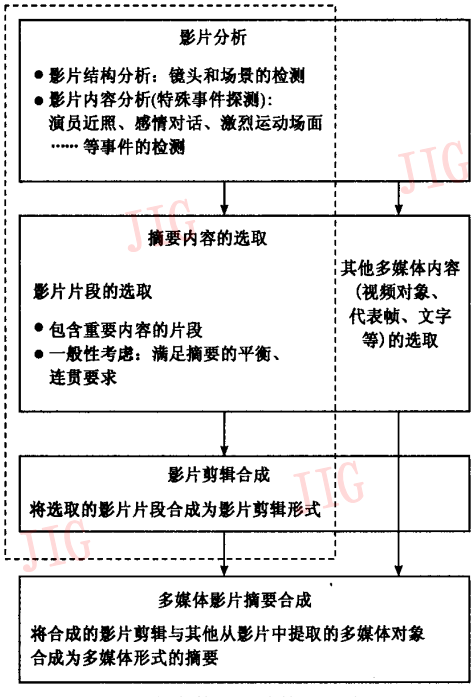

图 1　面向事件的影片摘要生成过程

Fig. 1　Process of event-oriented film abstract

3　面向事件的影片摘要生成

这一节将针对故事影片,介绍面向事件的影片摘要的生成过程。在下面的论述中,将多次出现“影片片段”一词,所谓影片片段 S_i 是由影片视频中的连续图像序列和其伴随音频组成,其长度最小可能是一个视频帧,最长可能是整段的影片。它可以用其在影片中跨越的时间间隔表示,即

$$S_i = \left[t_i^{\text{begin}}, t_i^{\text{end}} \right]$$

其中,$t_i^{\text{begin}}, t_i^{\text{end}} \in \mathbf{R}^+$,$t_i^{\text{begin}} \leqslant t_i^{\text{end}}$

对于帧频率固定的影片来说,影片片段也可以用它在视频中的起始和结束帧的序号来表示,即

$$S_i = \left[f_i^{\text{begin}}, f_i^{\text{end}} \right]$$

其中,$f_i^{\text{begin}}, f_i^{\text{end}} \in \mathbf{N}$,$f_i^{\text{begin}} \leqslant f_i^{\text{end}}$

3.1　影片分析过程

3.1.1　影片的分段

在进行影片分析之前,首先,将大段的影片分割成较小的单位,以便于进一步的分析,由于影片摘要关心更多的是影片故事的内容,因此,最理想的分割应该是将影片分割成一些完整的情节片段,但正如在本文前面所论述的,由于准确地理解影片语义内容是极其困难的,因此,可先将影片分割成具有一定语义内容的片段。虽然故事影片没有固定的编排模式,但由于场景的转换是故事影片叙事的主要方式,因此,可首先对影片进行场景分割(同时也进行了镜头的分割)。本文采用的场景探测方法[5]是融合了声像特征来进行场景检测,因为它能够获得比镜头更有意义的影片片段。

3.1.2　正面人脸的探测

在生成故事影片摘要的过程中,也不能忽视演员这一重要元素。由于人们需要了解故事影片中的主要演员是谁,因此需要能够在影片中探测人脸,并能够鉴别出一个镜头内以及不同镜头或场景中的同一演员的人脸,这也是影片内容的一个十分重要的线索。

Rowley, Baluja, 和 Kanade 提出的一种十分可靠的人脸探测方法[6]。由于这种方法对于探测正面垂直的人脸可以到达 90% 的探到率,而只有极小的误判率,因此,本文主要采用这种方法来探测视频帧中的正面人脸,同时也探测有轻微偏转的人脸(±30°),因为影片中人脸是运动的,故这样的探测是十分必要的,尽管这样探测的效果会有所降低。为了提高探测影片中人脸的效率,可首先对候选的人脸区域进行过滤,以使只有那些具有人脸皮肤颜色像素的区域和具有人脸结构特征(如:鼻、嘴、眼)的区域才被探测。这样可以减少大约 80% 的探测区域,而且在要处理的影片视频段中,每隔 3 帧进行探测。

尽管完成了视频中人脸的探测,但因为这些人脸

之间还完全没有关系,故下一步的任务就是按演员将这些探测到的人脸分组。由于只需要考虑重要的人脸,所以可只选择那些面积超过画面 30% 的人脸,并首先将同一镜头中的位置和大小相近的人脸分为一组,因为一般情况下,特别是在对话的场景中,在同一镜头内同一张脸的这些特征只有细微的变化。为减少错误的分组,可去除那些其中包含的人脸少于 3 个的分组;接下来,根据 CCV(color coherence vector)向量的相似程度,将时间上没有重叠的分组进行合并,以获得尽可能大的分组。这样的处理,在同一镜头内用于区分演员的人脸十分有效,而且计算简单,不用一些比较复杂的人脸识别算法。

同样,可利用 CCV 向量来合并不同镜头中的人脸分组。这种处理可以将同一演员在具有相似光照条件下的人脸合并在一起,但却不能保证将一个演员的所有人脸都合并在一起。这是因为演员人脸在影片中的变化很大的缘故,例如:同一演员在白天和晚上,其人脸 CCV 向量的差别可能比不同演员在同一场景中的人脸 CCV 向量的差别还大。这样的处理,获得的虽是多个演员的多个不同的人脸集合,但这种方法运算简单,这就避免了使用较为复杂的人脸识别算法。

3.1.3 对话的探测

对话是故事影片中出现最频繁的场面,对于影片摘要来说,全部记录这些对话是不可能的,但一些特殊的对话场面,如深情的倾诉、激烈的争吵……,由于其能给人很深的印象,也是表现故事情节的重要手段,因此,可探测出影片中的有感情对话,以作为影片摘要的候选内容。

在影片的对话场面中,镜头一般会在谈话者之间来回切换,利用这一特点可以初步地找出可能是对话的场面,通过判断伴随的声音是否是语音信号,则可以确定是否是对话场面,而对语音信号的分析则可以帮助判断某一场面是否是有感情的对话。

3.1.4 暴力场面的探测

同对话场面一样,暴力场面也是故事影片中经常出现的内容,由于几乎每一部电影或电视剧都会或多或少的出现暴力的场面(包括追逐、打斗),所以,可选取这类事件进行探测,一般可通过综合声音和视觉特征对视频进行分析来探测影片中的暴力场面,其中可视特征利用了镜头密度和镜头运动强度来选取候选片段,而音频处理则利用了音频分类和特殊声响探测的方法。详细的处理过程可参见文献[7],此处不再赘述。

3.1.5 文字的探测

故事影片中片名/剧名和演职人员的名字这两类重要都的信息是以文字形式出现的。这两种信息都应该被包含在影片的摘要当中,同时,这些也是多媒体影片摘要中用于影片检索的重要线索。

故事影片中出现这两类信息的地方,主要是片头和片尾,但一般情况下,由于在片尾出现的文字一般较小,不易探测和识别,而且片尾出现的演员表一般很长,包括全部的演职人员的名单和其他信息,而片头中出现的一般是主要演员的名字,因此,本文只在片头中进行文字的探测,一般是在影片开头的 5 ~ 10min 内进行探测。探测时是用影片中文字的大小来区分片名和演员名。其中,出现在画面中部,具有较大字体的文字,被认为是故事影片的片名,而且,具有这样大小的文字,在片头中一般不会很多。

本文采用文献[8]中的文字探测方法来探测影片中的文字,并利用比较成熟的文字识别方法或直接使用文字识别软件(如:清华紫光、尚书)来识别这些文字。

3.2 影片摘要内容的选取

3.2.1 影片摘要的一般特性

在选取摘要内容之前,应该首先明确人们对影片摘要有哪些基本的要求。一般来说,影片摘要具有以下一些特性:

(1) 重要的对象和人物

重要的对象和人物应当在影片摘要中出现,特别是明星演员的出现十分重要,因为对多数影迷来说,一部影视剧是否吸引他,更多地取决于影片中是否有明星的出现。对于由自己喜爱的演员出演的影片,他们可能会进行选择,而不太在乎故事本身的情节。

(2) 重要的事件

重要的事件是故事中的重要内容,也是推动故事情节的重要手段,特别是有特色的激烈运动场面和感人的对话场面,它们往往会成为某部影片的标志性内容,而且,这类事件也是故事影片中最吸引观众的部分。

(3) 故事的结局

一般情况下,影片摘要应当包含故事的结局内容,如果影片摘要被用于影片预告,那为了吸引观众,制作者一般会故意隐藏故事的结局,但是在用于检索的影片摘要中,故事的结局应该是必不可少的部分。

(4) 片头

片头部分包含影片的一个重要信息——片名。

由于主要演职人员的信息一般会出现在片头,而且片头一般会交代故事的起因和背景,主要演员一般也会在片头中首先出现,因此,故事影片的片头应该包含在影片的摘要当中。

(5) 摘要的平衡

影片摘要除了应该反映故事的主要内容以外,还应该符合人的观看习惯。摘要内容应该具有一定的连续性,这样才可以使得影片摘要能够反映故事情节的发展。

3.2.2　影片片段的选取

依据影片摘要的特性,可决定用于影片摘要的电影片段的选取方法。这里需要明确的问题是:(1)影片摘要的长度要求;(2)应该选择那些场景?(3)选择场景中的哪些影片片段包含在摘要中?

首先,必须确定影片摘要的长度。这一长度一般由人工来设定。一般情况下,故事影片摘要的长度在 20 ~ 60min 之间是比较合适的[9]。

接下来,就是确定选择影片中的哪些场景,以便从中抽取影片片段,用于组成影片摘要。选择场景时,首先依据特性(3)、(4)选择必须的场景,由于决定故事影片的片头和结尾应该被选取,因此可在影片开始和结束的部分各选择片长 2% 的内容加入摘要;然后,依据特性(2),选择包含特殊事件的场景,但这类场景在影片摘要中所占的比重,要由人来主观决定,本文选择的比例是 50%,如果这类场景的总长度超过这一比例的话,则将随机地从这类场景中选择一些符合长度要求的场景。对于以上的场景,一般会选取它们的完整片段,用于影片摘要的合成。

依据特性(5),影片摘要另外 50% 的部分,将从分布在整段影片中的其他场景中选取,如果随便选择一些没有吸引力的影片片段,则对于影片摘要来说是没有意义的,因此,场景的选择要依据一定的重要程度的规则进行取舍(见下一节)。本文是从选择出来的重要场景中,依照影片摘要长度的需要,选择若干个镜头作为合成影片摘要的片段。

3.3　场景重要程度评价

从影片观赏的角度来看,场景的重要程度应该体现在它包含吸引观众成分的多少,如对于故事影片来说,尽管能够吸引观众的成分很多,但这里只考虑其中比较重要的以下两类:紧张激烈的场面和感人的场面。

紧张激烈的场面惊险、刺激,引人入胜,其特点是:①镜头短;②镜头间重复内容少;③镜头内一般有较大的运动;④常伴随剧烈的声响(枪声、爆炸声、尖叫声、车轮的剧烈摩擦声等);⑤有时还伴有激烈的音乐背景,声场因而较为嘈杂。

感人场面的特点是:①镜头长;②常伴随清晰的抒情音乐,其声场一般比较安静;③镜头内部一般运动较少;④常出现人物的特写镜头。

因此,本文用紧张度 J 和感人度 G 这两个指标来衡量场景的重要程度。这两种指标的获得是依据两类故事场面的特点,通过综合多种媒体特征分析的结果得出的影片场景的重要程度的评价算法步骤如下:

(1)令当前场景 C 的紧张度 $J=0$,感人度 $G=0$;

(2)设场景 C 中镜头的平均运动强度 $m_{\text{intensity}} > \delta_{\text{intensity}}$,则 $J = J + w_1 m_{\text{intensity}}$;若 $m_{\text{intensity}} < 1 - \delta_{\text{intensity}}$,则 $G = G + k_1(1 - m_{\text{intensity}})$;其中 $\delta_{\text{intensity}}$ 为判断场景中运动强度的阈值,本文设定为 0.75;

(3)设场景 C 中包含剧烈声响的最大概率为 p_{noise},则 $J = J + w_2 p_{\text{noise}}$;

(4)设场景 C 中所有包含环境音的镜头的相对长度为 $l_{\text{surroundings}}$,则 $J = J + w_3 l_{\text{surroundings}}$;

(5)若场景 C 中的最大镜头长度 $L < \delta_{\text{short}}$,则 $J = J + w_4$;否则 $J = J$. 其中 δ_{short} 为判断场景中是否都是短促镜头的阈值;

(6)若场景 C 中的最大镜头长度 $L > \delta_{\text{long}}$,则 $G = G + k_2$,否则 $G = G$;其中 δ_{long} 为判断场景中是否有长镜头的阈值;

(7)设场景 C 中所有包含背景音乐的镜头的相对长度为 $l_{\text{background}}$,则 $G = G + k_3 l_{\text{background}}$;

(8)若场景 C 中包含正面人脸特写,则 $G = G + k_4$,否则 $G = G$;

(9)若 $J > 0$,则 $J = 0.5 + 0.5J$;若 $G > 0$,则 $G = 0.5 + 0.5G$。

算法中的 w_i 和 $k_i(i = 1, \cdots, 4)$ 为各种因素的经验权重,为简单起见,本文它们的取值都为 0.25。在评价场景重要程度的时,即可依据以上算法计算出场景的紧张度和感人度,笔者认为紧张度或感人度超过预先设定阈值的场景是比较重要的场景,其将被用于合成影片摘要。

3.4　影片摘要的合成

在选取了影片摘要的组成部分之后,接下来的工作就是将这些片段有效地组织起来,形成符合观看习惯的影片剪辑。影片摘要的合成阶段,需要考虑以下两个问题:(1)影片片段的合成顺序;(2)影片片段的衔接方式。

一般来说,合成影片摘要时,应按照影片片段的

时间顺序进行组合,这样才不会影响人对影片内容的理解,但是,当影片摘要被用于电影预告片时,制作者经常会打乱这些片段的顺序,以期待给观众更大的悬念。本文主要依据影片片段的时间顺序来拼接它们。

虽然采用由计算机直接合成影片摘要的方式未尝不可,但为了使影片摘要更加符合人的观看习惯、更具吸引力,可以利用人为的处理手段,如在影片片段的衔接方式上,可考虑硬切、溶化和扫描这 3 种编辑手段,但对这些编辑方式的使用,应该依据影片编辑的一般规则,这些规则来源于专业编辑人员的知识总结和相关的影片编辑模型。大家知道,这本身是一个十分复杂的过程。为简化起见,本文将运动剧烈的场景(包括暴力场面)与其他场景的衔接采用硬切的方式,而将柔和地衔接方式(溶化、扫描)用于较安静的场景(例如对话)。表 1 列出了不同情况下所使用的编辑方式,如果一种情况下有多种方式可以选择,则可以任选或由用户定制,一些更为成熟的编辑方式可以参见文献[10]。

表 1 影片片段的衔接规则

Tab.1 Film segment joining rules

	运动片段	安静片段	其他
运动片段	硬切	硬切	硬切
安静片段	硬切	溶化、扫描	硬切、溶化、扫描
其他	硬切	硬切、溶化、扫描	硬切、溶化、扫描

这里所说的影片片段包含其伴随音频,而对于伴随音频的衔接,可只采用溶化一种方式。

3.5 多媒体影片摘要的生成

由于多媒体形式的摘要包含对影片内容的多种形式的描述,因此更加有利于用户快速地了解影片内容,并可加速检索和浏览影片的过程。虽然多媒体形式的影片摘要中的内容比影片剪辑更加丰富,但它们大多可以在影片摘要的生成过程中获得,它们主要包括

(1) 影片的标题(片名);

(2) 重要事件、场景的代表帧;

(3) 影片缩略视频;

(4) 构成缩略视频的场景的代表帧。

这些内容应该被合理地组织,以便使多媒体形式的影片摘要发挥更大的作用。上面介绍的在 HTML 页面中组织多媒体信息的方式就是一个很好的例子。其中,影片标题可以实现快速的影片检索;影片摘要可以实现快速的浏览;代表帧可以标示影片中的重要内容。另外,通过点击代表帧,还可以定位到影片中的相应位置,以提高浏览影片的效率。

4 影片摘要可用度的测试与分析

根据故事影片摘要在不同领域的应用,对影片摘要一般有以下两方面的要求:一是要求影片摘要能够最大限度地反映影片的内容,使得在对影片摘要的浏览过程中,能够快速了解影片内容,这一要求主要用于对影片的检索和浏览、电影简介等领域,可称之为影片摘要的反映内容程度;另一个要求是,希望影片摘要能够更多地表现故事影片的精彩内容,以吸引人们进一步地观看,这一要求主要用于对影、视预告片的制作,可称之为影片摘要的表现精彩程度。

这两个要求有时是统一的,有时又是矛盾的,例如,一般来说,故事影片中的精彩部分是表现故事内容的主要元素,也是影片中比较重要的内容,但有时为了使影片摘要更加富有吸引力,常常需要在其中忽略一些影片中的重要内容,以增加悬念,而在影片摘要的生成过程中,有时又要兼顾这两种要求,因此在以这两种尺度评价影片摘要的可用程度时,应根据具体的应用要求做出判断。

本文将影片摘要表现精彩与反映内容的特性等同看待,因为二者的基础(即描述影片的内容)是相同的。同时,通过实验,还对本文提出的基于事件的影片摘要方法,其生成的影片摘要的可用度进行了测试。

由于对内容的感知是一个主观的过程,不同的人去评价一段影片摘要的优劣,会因测试者的理解能力和偏好而产生比较大的偏差,因此,本文的做法是:首先,让多个测试者对一段影片中他认为重要的内容进行标注,并综合他们的意见,人为地制作一段故事影片的简介片;然后以这一简介片作为标准,将用本文设计的方法生成的影片剪辑与简介片进行比较,并以此作为影片摘要可用度的评价。

由于处理过程十分复杂,致使自动生成影片摘要的过程很耗时(据测试,采用 Pentium Ⅲ 750MHz 处理器,128M 内存的计算机,对采用 MPEG1 标准压缩的一个小时的故事影片进行处理,大约需要 4 个小时),所以,本文用于测试的影片都不超过一个小时。实验中,本文选择了 5 段不同类型的故事电影(后两部为电视连续剧中的一集)进行测试。表 2 是测试影片与手工生成的简介片的情况。

对这 5 个测试影片,利用本文的方法分别生成了 2min 和 4min 的影片摘要。表 3 是影片摘要中的场景数。

表 2　测试影片与简介片的情况

Tab. 2　Testing film and trailer

测试影片	测试影片长度(min)	测试影片场景数	简介片长(s)	简介片场景数
最佳损友(上)	45	9 707	38	9
终极悍将(上)	51	10 874	120	45
吸血鬼女王(下)	48	10 326	64	21
反托拉斯行动(2)	46	9 965	48	21
钱王(3)	47	10 046	34	11

表 3　影片摘要中的场景数

Tab. 3　Scene number in film abstract

测试影片	120s 的影片摘要	240s 的影片摘要
最佳损友(上)	13	23
终极悍将(上)	9	22
吸血鬼女王(下)	10	18
反托拉斯行动(2)	6	12
钱王(3)	6	11

表 4 是自动生成的影片摘要与手工生成的简介片的比较结果。由该表可以看出,尽管存在一些吻合的场景,但二者之间的差别还是比较大的,笔者发现,在观看简介片和影片摘要时,二者给人的感觉效果却是差不多的。

表 4　影片摘要与简介片吻合场景数的比较

Tab. 4　Accordant scene in film abstract and trailer

测试影片	120s 的影片摘要	240s 的影片摘要
最佳损友(上)	1	3
终极悍将(上)	3	4
吸血鬼女王(下)	2	3
反托拉斯行动(2)	3	5
钱王(3)	0	1

通过分析,笔者发现造成这种情况有以下 3 种原因:一是该方法生成的影片摘要确实不能完全反映影片的内容,因为人类所能理解的影片内容极其丰富,而计算机能做到的还十分有限;二是人可以准确地确定重要内容的具体位置,而计算机却做不到;三是影片中有大量相似的场景,而且无论手工还是自动的方式,都会选择这一类场景,但它们选择的却不一定是同一个场景。

从实际的效果来看,造成上述情况应该主要是后两种原因,也就是说,手工方式与自动方式选择的影片片段虽然不同,但在内容上却是相似的。由于这种摘要生成方法还是能够反映影片主要内容的,因此,用本文的方法生成的影片摘要是具有一定可用性的。

5　结　论

本文首先介绍了影片摘要的概念和用途,并对影片摘要的形式进行了分类;然后,依据目前的研究状况和影片的特点,提出了一种适合于故事类影片的面向事件的影片摘要生成方法,同时对方法的细节进行了讲述,并提出了评价场景重要程度的算法,该场景重要程度的算法与现有的多数算法相比,其优势在于,它综合利用了影片中的多种媒体特征,即不仅考虑了运动特性,还考虑了感人特性对场景重要程度的影响;最后,对这种影片摘要方法的可用度进行了测试和分析。由于这种方法是建立在对影片内容的有效分析的基础之上的,因此可以预见,随着各种视频分析处理技术的不断完善,这种方法的效果会进一步提高。

参考文献(References)

1　Sundaram Hari, Xie Lexing, Chang Shih-Fu. A utility framework for the automatic generation of audio-video skims[A]. In: Proceedings of ACM Multimedia'02[C], Juan-les-Pins, France, 2002: 189 ~ 198.

2　Christel M G, Olligschlaeger A M. Interactive maps for a digital video library[A]. In: Proceedings of the IEEE International Conference on Multimedia Computing and Systems [C], Florence, Italy, 1999: 381 ~ 387.

3　Lienhart R, Pfei _ er S. Video abstracting[J]. Communications of the ACM, 1997, **40**(12): 54 ~ 62.

4　Ng T D, Christel M, Hauptmann A G, *et al*. Collages as dynamic summaries of mined video content for intelligent multimedia knowledge management[A]. In: AAAI Spring Symposium Series Intelligent Multimedia Knowledge Management[C], Palo Alto, CA, USA, 2003: 24 ~ 26.

5　Jiang H, Helal A, Elmagarmid A K, *et al*. Scene change detection techniques for video database system[J]. Multimedia Systems, 1998, **6**(3): 186 ~ 195.

6　Rowley H A, Baluja S, Kanade T. Human face detection in visual scenes[R]. Technical Report CMU-CS-95-158R, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, November 1995.

7　WANG Chen, LAO Song-yang, HU Xiao-feng. Audio-visual contend-based violet scene detection[J]. Mini-Micro Computer System, 2001, **22**(4): 456 ~ 458. [王辰,老松杨,胡晓峰. 基于声、像特征的视频暴力场面的探测[J]. 小型微型计算机系统, 2001, **22**(4): 456 ~ 458.]

8　WANG Chen, LAO Song-yang, HU Xiao-feng. Text detection in video frames, Mini-Micro Computer System, 2002, **23**(4): 478 ~ 481. [王辰,老松杨,胡晓峰. 视频中的文字探测[J]. 小型微型计算机系统, 2002, **23**(4): 478 ~ 481.]

9　Pfeiffer S, Lienhart R, Fischer S, *et al*. Abstracting digital movie automatically[A]. In: Proceedings of IEEE International Conference on Multimedia Computing and System [C], Tokyo, Japan, 1996: 509 ~ 516.

10　Yeo B, Yeung M M. Retrieving and visualizing video [J]. Communication of the ACM, 1997, **40**(12): 43 ~ 52.