一种从 MPEG 压缩视频流中提取关键帧的方法

杨 胜 钟玉琢
(第三军医大学计算机教研室,重庆 400038) (清华大学计算机系,北京 100084)

摘 要 在基于视频内容检索的多媒体系统中,由于需要进行镜头分割和提取关键帧,还需要用静态图象来表示视频内容以及对该图象的特性进行分析,因此根据视频序列中相邻画面一般具有相似性和连续性这一镜头分割和关键帧提取的共同理论依据,构造了关键帧提取系统,它能直接提取关键帧,而不用先进行镜头分割,且只需要 I 帧信息及其频域直流分量的信息,即能达到最小程度的解码.在关键帧的判定方面,通过分析当前镜头分割技术的特点及其发展方向,提出了质点等价法和基于宏块互异的方法.

关键词 镜头分割 关键帧 基于内容检索 MPEG JPEG DCT
中图法分类号:TP941.1 **文献标识码**:A **文章编号**:1006-8961(2001)03-0254-05

A Unified Approach to Extraction of Keyframes from MPEG Compressed Video

YANG Sheng
(*Teaching Group of Computer Science, Third Military Medical University, ChongQing* 400038)
ZHONG Yu-zhuo
(*Department of Computer Science, TsingHua University, Beijing* 100084)

Abstract Keyframes are still images, which best represent the content of the video sequence in an abstracted manner, and may be extracted from original compressed data. Keyframes are frequently used to supplement the text of a video log, but there has been little progress in identifying them automatically. The challenge is that the extraction of keyframes needs to be automatic and content based so that they maintain the important content of the video while removing all redundancy. Up to now, more and more video materials are stored and transmitted in the compression data, so it is practical to study a unified approach to extraction of keyframes based on compressed video data. In this paper, we present a system for extraction of keyframes, which is based on different formulae comparing discrete cosine transform (DCT) direct current (DC) coefficients over the I-frames in MPEG video stream, and for which only minimal decoding is needed. In theory semantic primitives of the video, such as interesting objects, actions and events should be used. However, because such general semantic analysis is not currently feasible, we explore two methods instead, in which one is from the idea of dot-in-mass, and the other is based on the number of unequal macro-blocks.

Keywords Shot segmentation, Keyframes, Content-based retrieval, MPEG, JPEG, DCT

0 引 言

当前,虽然基于视频内容的多媒体查询与检索系统方面的研究越来越多,其应用前景也非常广阔,但广泛使用的却很少,究其根源,主要与视频内容的获取不能完全自动化有关,如镜头分割、关键帧提取、画面注解、物体分割以及物体的运动估计等由于时常需要人工干预,因此,研究全自动的视频内容提取(如关键帧的提取)技术,是很有价值的.

如今基于视频内容的查询与检索系统[1~4],都运用了关键帧提取技术,而且通过其实现方法的分

收稿日期:1999-06-11 改回日期:2000-03-02

析，还发现了如下一些共同点：①提取的源数据都是压缩的数据流，如 MPEG 流，并且对压缩数据进行最小程度的解码；②提取技术都利用了视频节目相邻帧都具有连续性与相似性这一特征；③一般均先进行镜头分割，然后从镜头中取一帧或多帧作为关键帧. 如今选取关键帧的方法很多，并且各有优缺点，这说明关键帧的提取仍有待进一步研究.

1　关键帧提取的原理和过程

1.1　关键帧提取的原理

关键帧提取的目的，是希望用关键帧来代表视频节目的某个(或某些)特征. 如果将所有视频图象重叠起来(即在图象坐标系下考虑问题)，则可将第一幅图象的每个特征与后续图象中的匹配特征连接起来，这样即得到一条视频流图象特征的轨迹，由此可见，关键帧的提取，就是对某特征的轨迹进行“跟踪”的过程，也是对轨迹上关键的特征值(及其对应的帧)进行记录的过程.

1.2　关键帧提取系统

关键帧提取的步骤如下：①寻找图象中某特征的量化参数；②判断该特征量化的参数是否为关键的特征值，由此，可构造出从 MPEG 压缩视频流中提取关键帧的系统(如图 1)，该系统可不经过镜头分割，即能直接从 *I* 图中提取关键帧. 由图 1 可见，提取关键帧的关键是特征参数的选取和确定关键帧的判定规则.

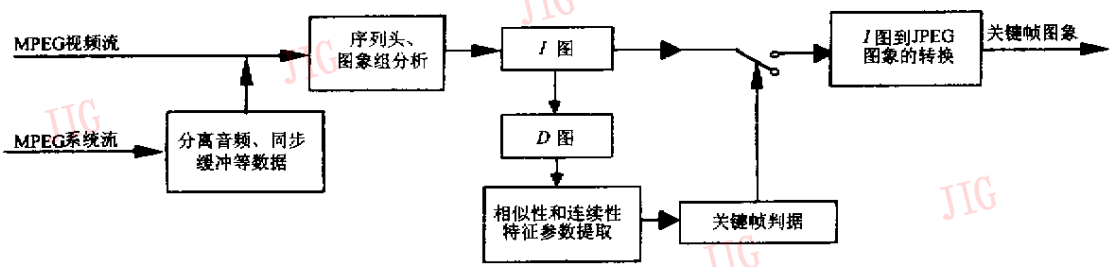

图 1　关键帧提取过程框图

对该系统的设计作如下说明：

(1) 根据 MPEG-Ⅰ国际标准(系统部分)，分析 MPEG-Ⅰ文件，如果该文件为系统流，则转换并生成为视频流文件；否则，直接生成视频流文件. 所谓系统流一般是由视频流、音频流、填充流等复合而成，并包含各个流的播放同步以及缓冲区管理等数据. 由于关键帧的提取只与视频流有关，故采取“MPEG-Ⅰ文件⇒视频流文件”的措施.

(2) 根据 MPEG-Ⅰ国际标准(视频部分)，分析视频流文件，并通过删除 *P* 帧和 *B* 帧，生成仅由 *I* 帧组成的视频流文件. 视频流主要由画面组组成，而画面组则由 *I* 帧、*P* 帧和 *B* 帧(或 *D* 帧)等 3 种类型的帧画面组成，且 MPEG 视频编码要求约每 13 帧中有一 *I* 帧. 由不被打断的连续画面组成的镜头，其播放时间以秒为单位才有意义，故按画面频率24 帧/s 计算，则每个镜头内必然包括 *I* 帧，现已得到实验证实. 由此可见，建立在镜头基础上的关键帧，完全可以仅从 *I* 帧中提取，故生成视频流文件可采取‘视频流⇒*I* 帧组成的视频流’的措施.

(3) 首先分析相邻 *I* 帧的连续性和相似性，并用适当的参数表示，然后找到由 *I* 帧组成的视频流的间断点(与镜头分割点有一定区别). 大家知道，视频流中相邻画面具有很强的相似性，实验也表明，仅由 *I* 帧组成的视频流，其相邻 *I* 帧画面也具有很强的相似性.

(4) 把与每一间断点对应的 *I* 帧画面，转换成 JPEG 静止图象. 由于 MPEG 和 JPEG 的压缩与编码算法非常相似，故以 JPEG 格式来表示关键帧最简单[5,6].

2　质点等价法

2.1　镜头分割与关键帧提取的关系

在视频节目的拍摄过程中，将摄像机一次拍摄过程所获得的节目称为一个镜头；而在节目编辑时，则把镜头拼接起来，来表达连贯的节目语义. 这种拼接有如下两种方式：一种是直接把前一镜头的最后一帧与后一镜头的第一帧进行连接，但播放时其有明显的切变过程；另一种是把前一镜头的最后一帧进行艺术加工，使其逐渐隐去，而后一镜头的第一帧逐渐凸显出来，这样播放时即有渐变的过程. 这种互相连接的镜头以及连接方式的选择，应该由节目要

表达的语义来决定,即内容决定形式.由于拼接的两个镜头毕竟是摄像头的两次操作,因此拼接起来后在某些特性上会有所变化,特别是切变拼接时.

当然,编码时可以把镜头拼接点作为图象组的开始点;更进一步,可以通过在压缩编码标准中,定义一个标志位,来表示镜头的拼接.尽管目前 MPEG 压缩标准不支持,但随着数字视频编辑的应用,会有这种发展的趋势.

镜头分割,就是把已经拼接好的节目再分割成一个个的镜头,其手段就是通过某特性的变化来寻找镜头的边界点,然后即可实现镜头分割.在基于视频内容检索的多媒体系统中,只是把镜头作为最小的用户操作单位,而不是语义单位,同时用户使用镜头时,主要是通过播放器进行播放,因此把镜头的边界定义成一个点,价值显然不大,而且播放时快进与后退已是非常方便,也不费时;而用户可能对最小的语义单元更感兴趣.

大家知道,镜头分割与关键帧的提取是紧密相连的,如果镜头变化时的变化特性与关键帧代表的特性相同,则镜头分割主要关心特性的变化点,而关键帧则关心特性的一致性,而且前者希望得到一个点是非常困难的,也很容易误判;而后者则是希望得到一致的线段,然后再归纳线段的共同点,这显然要容易得多,鲁棒性也好得多.

2.2 镜头分割算法

镜头分割算法,归纳起来有如下两大类:其中一类是建立在相邻图象的差图基础上的算法,如象素差图、D 图的差图等,其特点是对切变镜头的分割效果好,但对大物体的快速运动则容易误判,渐变也不容易检测到;另一类是建立在图象的组成元素基础上的算法,如直方图统计、纹理分析等,其特点是对运动不敏感,但指标过于简单,容易误判[7].

由于同一镜头中相邻图象的相似性较好,主要表现在图象的组成元素相同或近似;而其连续性则表现在相同元素在画面中的位置不变,或者变化很小,因此,镜头分割的算法应该既要考虑图象组成元素的近似,又要考虑元素的位置关系.基于象素差图的方法,由于考虑元素太细,因此对位置的变化敏感;而基于直方图的方法,则根本没有考虑元素的位置关系.

当然,理想的算法应该把图象划分成合适的组成单元,然后在后续图象中相应位置附近进行匹配.这个过程类似于 MPEG 标准中的运动估计,但以宏块为组成单元则有一定的局限性.一种改进的方法就是在图象中找到有代表性的物体或特征(如主颜色或物体),并跟踪其出现和消失的过程,但这样的物体或特征难以寻找.

2.3 质点等价法

运用物理学上的等效方法,可以把整幅图象等效为一个象素点(相当于质点的质量),其位置即为矩心(相对于质心的坐标)

$$M_0 = \sum_{x=1}^{W}\sum_{y=1}^{H} f(x,y)$$

$$M_x = \sum_{x=1}^{W}\sum_{y=1}^{H} xf(x,y)$$

$$M_y = \sum_{x=1}^{W}\sum_{y=1}^{H} yf(x,y)$$

$$x_0 = \frac{M_x}{M_0}$$

$$y_0 = \frac{M_y}{M_0}$$

其中,M_0 表示质点的质量;$f(x,y)$ 表示 (x,y) 位置的象素值;W 和 H 分别表示水平和垂直方向的象素个数;(x_0,y_0) 表示质心[8].由此,即可构造三维跟踪特征质点 (M_0,x_0,y_0).

这种方法抽象地考虑了图象组成,又兼顾了位置关系.

判定关键帧的规则可确定为

$$\Delta T = \lambda_1(1 - M_0^C/M_0^P)^2 + \lambda_2(1 - x_0^C/x_0^P)^2 + \lambda_3(1 - y_0^C/y_0^P)^2 \quad (1)$$

其中,λ_1、λ_2 和 λ_3 表示加权因子;M_0^C 和 M_0^P 分别表示当前帧(C)与前一帧(P)的质点质量;(x_0^C,y_0^C) 和 (x_0^P,y_0^P) 分别表示当前帧(C)与前一帧(P)的质点质心;ΔT 为判定的依据.

3 宏块互异数目的算法

3.1 宏块互异数目能反映视频流的相似性和连续性

一般表征相邻画面的相似性,可以用组成画面宏块的相似性来反映,且在相似的前提下,表征相邻画面的连续性,可以从相同宏块的位置来反映.由于视频节目可能有物体的运动或者摄像机的推、拉和摇等动作,而相邻画面内的相同组成单元可能有位置的小变化,因此,画面的相似性与连续性,应采用小范围图块的均值来反映,而不能用具体的象素值;同时,差异应该是小图块互异的数目,而不是小图块的差异值.

在 MPEG 标准中,图象压缩是基于 DCT 变换

的,并且该变换是以小图块(8 × 8 象素块)为变换的基本单元.在压缩视频流中,DCT 系数的直流分量很容易获得,它是宏块均值的 8 倍,而且由于宏块是 3 个颜色分量的最小单位,因此,相邻画面的宏块互异数目,能够很好地反映视频节目的连续性,其构造公式如下:

$$S_C = \sum_{i=1}^{W}\sum_{j=1}^{H} Diff(MB_{ij}^{C} - MB_{ij}^{P}) \qquad (2)$$

其中,

$$Diff(MB_{ij}^{C} - MB_{ij}^{P}) = \begin{cases} 1 & \sum_{k=1}^{6} \left| B_k^C - B_k^P \right| > T \\ 0 & \text{其他} \end{cases}$$

在式(2)中,MB 为实块,S_C 表示相邻 I 图间宏块的差异数目,W 表示水平方向宏块数,H 表示垂直方向宏块数,B_k^C 和 B_k^P 分别表示当前帧与前一帧中组成对应宏块的块,T 表示宏块互异的阈值(一般取 T 为 6). $Diff$ 是通过比较两宏块中块的差异值来确定宏块是否是互异的函数.

3.2　关键帧的判定标准

提取关键帧的目的有如下两个方面:其一是希望用它来静态表示视频节目的主题和部分内容,而不是动态的细节;其二,则是希望从关键帧中提取颜色、纹理和形状等特征,以作为多媒体数据库的数据源,而不需要对每一画面都重复提取.由此看出,关键帧应具有代表性,即对前者,应代表主题方面的特征;对后者,则视提取特征的不同而不同.

当前,一般采用保守原则来提取关键帧,即关键帧的提取"宁愿错,勿能少",同时,在代表特征不具体的情况下,一般以去掉重复(或冗余)画面为原则,据此,设计以下判断规则:

(1) 如果 $\lambda S_C > 0.75 MB_s$,则得到关键帧;

(2) 如果 $\lambda S_C > S_P$ && $\lambda S_C < S_N$ && $\lambda S_C > 0.5 MB_s$,则得到关键帧;

(3) 如果得到关键帧,则 $\lambda = 1$;否则,$\lambda = \lambda + \Delta\lambda$.

其中,S_C、S_P 和 S_N 分别为当前、前一和后一宏块互异数目;MB_s 表示画面宏块总数;λ 表示特征因子,$\Delta\lambda$ 表示特征增强因子;符号"&&"表示"与"的逻辑关系.

上述规则(1)表示相邻画面变化较大;规则(2)表示相邻画面变化趋势有改变,且变化较大;规则(3)表示特征的增强(与遗忘因子相反),以避免丢失关键帧.

关键帧,一般取与上述规则中所得到的关键帧 S_C 所对应的前一帧,因为当前帧变化剧烈,而且没有成型,前一帧则比较成熟.按照本规则,虽然最后一帧不能判断,但可根据其画面复杂程度来判断是否作为关键帧,如可通过比较相邻宏块的差异数目来进行判断.

4　实验及其讨论

本文选择《大公鸡》和《两只老虎》等两部 VCD 儿歌,以及《魂断蓝桥》等 3 部黑白电影片头,还有佳能打印机等 5 部广告片作为测试数据源.利用公式(2),其判定标准中的增强因子 $\Delta\lambda = 0.07$,跟踪视频流中图片宏块的差异特征,其实验说明和统计数据见表 1.如果采用质点特征跟踪,则式(1)中的加权因子 λ_1、λ_2 和 λ_3 都选为 1,其实验结果与表 1 类似,但冗余的关键帧稍多.

表 1　宏块的差异特征实验说明与统计数据(* 表示统计数据)

视频源	视频源特点(测试帧数)	画面特征(宏块数)	总帧数*	I 帧数*	关键帧数*	实验结论
儿歌 VCD	变化快、物体运动多(多)	彩色 352 × 288(396)	2 960	210	67	镜头中关键帧没有遗漏,但有少量冗余;Pentium(r) Ⅱ 333 机器提取与播放时间接近
电影片头	变化慢、渐变多(中)	黑白 352 × 288(396)	1 500	120	23	
广告片	变化很快、切变多(少)	彩色 352 × 288(396)	850	65	18	

本文选择佳能打印机的广告片作为视频源来进行实验,图 2 为质点等价法的实验数据,该数据表明,本文方法对相似性与连续性反映准确,同时其鲁棒性表现在大物体的快速运动上(图 2 中 22#、23# I 帧所示),而在质点特征(M_0, x_0, y_0)中,由于仅表现为重心特征(x_0, y_0)的变化连续,因此,判定关键帧的规则可以比较简单.

若采用同样的视频源进行实验,图 3 为宏块互异法的实验数据,其中,3# I 帧反映色彩的变化,23# I 帧反映大物体的运动;由图 3 数据表明,由于本方法对相似性与连续性反映复杂,因此判定关键帧的规则也比较复杂.

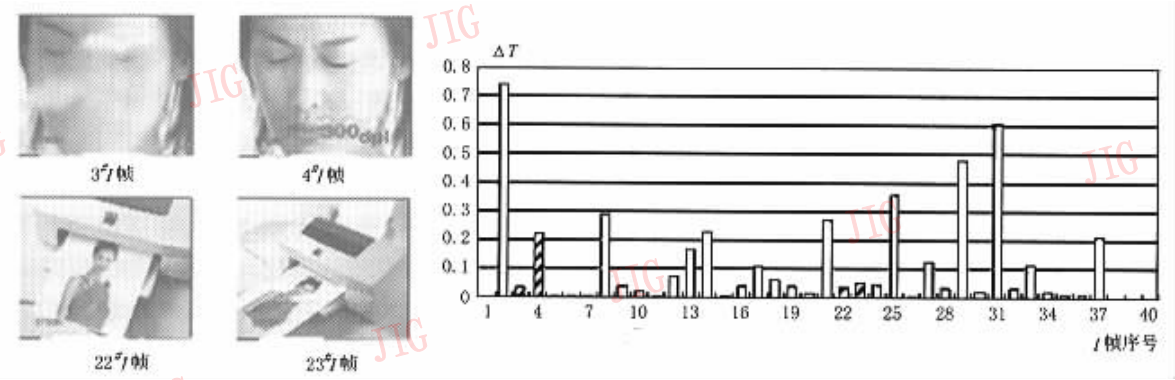

图 2 质点等价法的实验数据

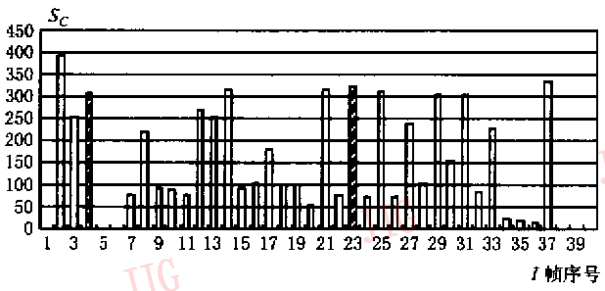

图 3 宏块互异法的实验数据

5 结 语

虽然关键帧能够代表视频节目的内容,并且可进一步提取特征,但寻找图象序列中的特征以及对其量化是困难的,若能避开镜头分割,而直接利用由 *I* 帧的宏块象素均值等技术构成的关键帧提取系统来进行关键帧提取,则实现方法比较简单.同时,运用抽象的质点表示图象的组成,并用质心代表位置,来对质点的特征进行跟踪,进而提取关键帧的方法,也很有启发意义,如可用图象的 4 个角表示质量,再利用宏块互异数目的方法,来进行特征提取是比较鲁棒的.

当然,具体实现中,某些阈值还需要经过多次测试,其理论上的指导作用仍有待于进一步探索.

参 考 文 献

1 Zhang H J, Low C Y, Gong Y *et al*.. Video parsing using compressed data. In: SPIE Image and Video Processing Ⅱ, 1994, 2182: 142149.

2 Minerva Yeung, Yeo Boon-Lock, Liu Bede. Segmentation of video by clustering and graph analysis. Computer vision and image understanding, 1998, 71(1): 94109.

3 Bolle R M, Yeo B L, Yeung M M. Video query: Research directions. IBM J. RES. DEVELOP, 1998, 42(2): 233251.

4 Nevenka Dimitrova, Thomas McGee, Herman Elebaas *et al*. Video content management in consumer devices. IEEE Transactions on knowledge and data enginerring, 1998, 10(6): 988995.

5 Gregory K Wallace. The JPEG still picture compression standard. IEEE Transactions on consumer electronics, 1992, 38(1): 18-33.

6 祁卫,钟玉琢. 基于 MPEG 国际标准压缩视频流的镜头切分算法. 北京: 清华大学学报, 1997, (9): 5054.

7 徐建华. 图象处理与分析. 北京: 科学出版社. 1992: 177182.

杨 胜 1970 年生,1995 年 7 月获重庆大学自动控制理论与应用专业硕士学位,现为第三军医大学计算机教研室讲师.主要从事医学图像处理与多媒体数据库的研究工作.

钟玉琢 1963 年毕业于清华大学自动控制系,现为清华大学计算机科学系教授、博士生导师.长期从事机器人视觉、智能计算机声文图一体化接口以及多媒体计算机方面的教学和科研工作.