基于内容检索的视频处理技术

金 红 周源华
(上海交通大学图象通信与信息处理研究所, 上海 200030)

摘 要 从分析视频数据的结构和特点出发,总结了基于内容检索的视频处理方法的一般步骤,即视频分割、关键帧选取、静态和动态特征提取以及视频聚类等,然后深入介绍了各个处理过程中的一些最新方法,并分析了各种方法和技术的优缺点;最后,对基于内容的视频检索提出一些值得进一步研究的问题.

关键词 视频数据库 基于内容的视频检索 镜头边界检测 特征提取
中图分类号: TN941.1 **文献标识码**: A **文章编号**: 1006-8961(2000)04-0276-08

Review of Video Parsing Techniques for Contented-Based Video Retrieval

JIN Hong, ZHOU Yuan-hua
(*Image Communication and Information Processing Insititute of SJTU, Shanhai* 200030)

Abstract Video contains the most affluent information but implies huge storage and complicated semantics. To search for required fragments among huge quantity of video is a tedious and time-consuming task for traditional manual indexing and sequential searching methods which certainly cannot meet the performance requirements of video databases. What the users want is to query by contents, that is, to get the desired fragments of video with just some given examples or feature descriptions. Because of the complicated structure and temporal variation of video data, it is very difficult to index video by content. Researchers have worked out various methods and techniques to solve the problem. The essential steps for content-based video indexing are video segmentation, key frame selection, static and dynamic feature extraction and video clustering. Starting from a brief description of the structures and characteristics of video, this paper generalizes the methods and techniques used in content-based video indexing, analyses in depth the newly proposed ones and their respective advantages and drawbacks. As a conclusion, the paper discusses some of the problems in content-based video indexing that are worth to be tackled in future researches.

Keywords Video database, Content-based video retrieval, Scene change detection, Feature extraction

0 引 言

随着多媒体技术的发展和信息高速公路的出现,数字视频的存储和传输技术都取得了重大的进展.人们可以坐在家中访问远端的多媒体数据库,如进行视频点播、电子购物和访问多媒体图书馆等.这些方面所具有的广阔的商业前景,使得视频检索技术的研究受到日益广泛的关注[1,2].

视频检索就是要从大量的视频数据中找到所需的视频片断.传统的视频检索只能通过快进和快退等顺序的方法人工查找,因而是一件非常繁琐耗时的工作,这显然已无法满足多媒体数据库的要求.用户往往希望只要给出例子或特征描述,系统就能自动地找到所需的视频片断点,即实现基于内容的视频检索.

视频数据比文本、图象包含更丰富的信息,但是却无法像文本那样直接地给出它的内容或者直接地

收稿日期: 1999-07-19; 改回日期: 1999-11-10

进行内容的比较. 要实现基于内容的视频检索, 首先必须对视频进行处理, 包括视频结构的分析和视频单元的自动索引. 视频结构的分析是指通过镜头边界的检测, 把视频分割成基本的组成单元—— 镜头; 视频单元的自动索引是指提取镜头的颜色、纹理和运动等各种特征, 形成描述镜头的特征空间. 然后依靠这个特征空间来进行镜头内容的比较.

1　视频数据

1.1　视频数据的结构

描述视频(包括描述它的元数据) 可从以下 3 个方面进行索引[3,4]:

(1) 文献数据　包括有关整个视频的信息(例如标题、摘要、主题、类型等) 以及制作视频的个人信息(例如制片人、导演、演员表等) . 传统的视频检索主要依靠这些元数据, 这些数据往往需要手工输入.

(2) 结构数据　视频数据从结构上自顶向下可分为电影、场景、镜头和帧(如图 1 所示) .

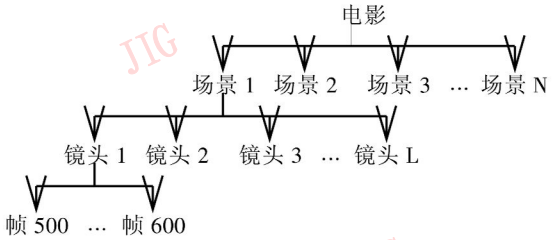


图 1　视频数据的结构

帧是视频数据的最小单元, 是一幅静止的画面. 镜头是视频数据的基本单位, 它是摄像头的一次连续的动作, 只能拍摄相邻地点连续发生的事情. 场景由内容相近的镜头组成, 从不同的角度描述同一个事件. 而电影则由许多场景组成, 叙述一个完整的故事.

结构层中每一个视频层次的数据都可以用一定的属性加以描述. 如:

电影的属性: 主要包括场景的个数和持续时间.

场景的属性: 如标题、持续时间、镜头数目、开始镜头、结束镜头等.

镜头的属性: 如持续时间、开始帧号、结束帧号、代表帧集合、特征空间等.

帧的属性: 帧有大量的属性, 如直方图、轮廓图、DC 及 AC 分量图等.

(3) 内容数据　表示视频的语义内容, 它包括音频数据、镜头内的一组代表帧或运动物体、由字幕得到的文本关键字以及从视频数据中提取的特征向量等.

1.2　镜头的切换

由于一个镜头只能拍摄相邻地点连续发生的事情, 它的描述能力有限, 所以大多数的视频都是由许多镜头通过编辑连接而成的. 有的视频切换频繁, 镜头的持续时间短, 如电视新闻节目、故事片等. 这些视频通过镜头的切换来反映不同地点或不同时间发生的事情. 也有的视频切换较少, 每个镜头的持续较长, 例如体育节目的转播. 而用于银行保安、交通监管的监控视频几乎没有镜头的切换, 对于这些视频人们关心的主要是镜头内物体的运动.

镜头的切换分为突变和渐变(abrupt change and gradual change) [5] 两类. 突变是一个镜头直接转换为下一个镜头, 中间没有时间上的延迟; 渐变则是加入了一些空间或时间上的编辑效果, 由前一个镜头慢慢地转换为下一个镜头. 渐变的方式有很多种, 而且不断有新的方式出现, 常用的有淡入/ 淡出(fade in/ out) 、慢转换(dissolve) 和扫转换(wipe) 等几类. 淡入是把画面逐渐加强, 淡出是把画面慢慢减弱直至消失; 慢转换是在上一个镜头画面逐渐减弱的同时, 下一个镜头的画面逐渐加强; 扫转换则是从画面的某一部分开始, 上一个镜头逐渐地被下一个镜头代替.

1.3　镜头内的运动

镜头内的运动包括由对象运动导致的局部运动和由摄像头运动导致的全局运动.

(1) 对象运动　对象的运动根据实际情况的不同千变万化, 但又是视频检索的一个重要方面, 特别是对于监控视频. 例如用户可能需要检索某个物体被移动的视频片段或汽车发动的视频片段. 针对这种情况, Courtney[6] 归纳了以下几种对象运动, 并进行了分析:

出现: 一个对象出现于镜头

消失: 一个对象从镜头中消失

进入: 一个运动的对象出现于镜头

退出: 一个运动的对象从镜头中离去

运动: 一个原本静止的对象开始运动

停止: 一个原本运动的对象停了下来

通过对以上对象运动的分析, 可实现对监控视频的基于内容的检索.

(2) 摄像头的运动　在视频的拍摄过程中, 摄像头可以按不同的方式运动, 以达到特定的拍摄效果. 摄像头的运动包括[7~9]:

摇镜头(tilt and pan) : 摄像头的位置不变, 而是

以云台为轴心,上下或左右转动拍摄方位.

转镜头(Z-rotation):以对象为中心,摄像头从不同的位置角度拍摄.

移动镜头(translation):摄像头的位置跟着拍摄对象移动,但不旋转角度.移动又可分为水平移动(horizontal translation)和垂直移动(vertical translation).

推拉镜头(zoom in and out):推镜头,即从远处开始逐渐推近到拍摄对象.拉镜头,即从近处开始逐渐拍成全景.

有时一个镜头内有几种摄像头运动,此时一般只分析主要的运动.

2 视频处理技术

基于内容的视频处理包括视频结构的分析、视频数据的自动索引和视频聚类.视频结构的分析是指通过镜头边界的检测,把视频分割成基本的组成单元——镜头;视频数据的自动索引包括代表帧的选取和静止特征与运动特征的提取;视频聚类就是根据这些特征进行的.视频处理的一般过程如图 2 所示.

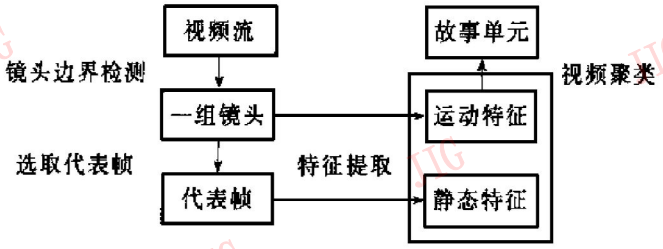


图2 视频数据的处理过程

2.1 镜头边界的检测

镜头是视频数据的基本单元.大部分视频是通过编辑由一个个镜头连接而成的,所以基于内容检索的视频处理,首先要把视频自动地分割为镜头,以作为基本的索引单元,这个过程就称为镜头边界的检测,也叫场景转换检测(Scene-Change-Detection, SCD),它是实现基于内容的视频检索的第一步.

镜头切换时,视频数据将发生一系列的变化,表现在颜色差异突然增大、新旧边缘的远离、对象形状的改变和运动的不连续性等各方面.镜头边界检测的目的就是寻找这些变化的规律.一般而言,同一个镜头内的各帧之间差异较小,而不同镜头的帧间差异较大.镜头边界检测方法可分为模板匹配法、直方图法、基于边缘的方法和基于模型的方法[1, 5, 10]等 4 种.

2.1.1 模板匹配法[5]

模板匹配法以两帧对应象素差的绝对值之和作为帧间差,其计算公式如下:

$$d(I_i, I_j) = \sum_{x=0, y=0}^{x<M, y<N} |I_i(x, y) - I_j(x, y)|$$

其中,I_i 表示第 i 帧视频,$d(I_i, I_j)$ 是 I_i 和 I_j 的帧间差,$I_i(x, y)$ 为第 i 帧 (x, y) 位置的象素值,M、N 为帧的宽度和高度.这种方法比较前后两帧对应象素之间的变化,如果变化超出一个阈值 t,则认为有镜头的切换.

模板匹配法的缺点是对噪声和镜头或物体运动非常敏感,因为它严格地局限于象素的位置.噪声和物体运动都会使帧间差增大,从而导致错误的场景转换检测.Zhang[1]等人提出了一种改进的方法,即把各帧划分为 8× 8 象素的小块,并对每个块取平均,再用这个平均值对前后帧的对应小块进行比较,这种方法可以去掉图象中的一些噪声,并对小的物体运动和镜头运动起到补偿作用.

2.1.2 直方图法[5, 11, 12]

直方图法是使用得最多的计算帧间差的方法,它不考虑象素的位置信息,而使用象素亮度和色彩的统计值,因而抗噪能力比模板匹配强.其基本原理是将颜色空间分为一个个离散的颜色小区间,然后计算落入每个小区间的象素数目.设颜色空间分为 n 个区间,H_{ik} 是第 i 帧中落入第 k 个颜色区间的象素数目.帧间差可用下面公式表示

$$d(I_i, I_j) = \sum_{k=1}^{n} |H_{ik} - H_{jk}|$$

颜色直方图法的缺点是,有时会漏掉场景变换,因为两幅图象可能有完全不同的结构,但其颜色直方图却很接近.与颜色直方图法相似的另一种计算帧间差的方法是 X^2 直方图法[5],据介绍这种方法用于镜头转换,检测效果要好于上述两种方法.两幅图象之差用下式求得

$$d(I_i, I_j) = \sum_{k=1}^{n} \frac{(H_{ik} - H_{jk})^2}{H_{jk}}$$

2.1.3 基于边缘的方法

这种镜头边界的检测方法是根据边缘特征[10, 11],它的基本思想是“在发生镜头转换时,新出现的边缘应远离旧边缘的位置,同样旧边缘消失的位置应远离新边缘的位置”.

首先提取前后两帧视频图象 I_i 和 I_{i+1} 的边缘图 E_i 和 E_{i+1},两帧视频图象之间的差异由下式计算:$diff = \max(din, dout)$,其中,din 是进入象素(新出

现的远离已有边缘的象素点) 所占的比例, $dout$ 是退出象素(新消失的远离新边缘的象素点) 所占的比例, 其中 $din = p_1/p_m$, p_1 为 E_{i+1} 中离 E_i 中最近边缘象素点的距离> r 的边缘象素点的总数, p_m 为 E_{i+1} 中的边缘象素点总数; $dout = p_2/p_n$, p_2 为 E_i 中离 E_{i+1} 中最近边缘象素点的距离> r 的边缘象素点的总数, p_n 为为 E_i 中的边缘象素点总数.

如果 $diff$ 大于某个设定的阈值 t, 则认为出现了镜头的切换.

上述 3 种检测方法都是通过计算帧间差来进行镜头边界的检测. 对于突变, 帧间差在镜头切换处会出现明显的峰值, 因而可以将帧间的差值与一个预先设定的阈值相比较, 当差值超过该阈值时, 则认为有镜头切换. 对于渐变切换, 由于两个镜头之间的切换是缓慢进行的, 帧间差虽然有所增大, 但没有一个明显的峰值, 而是会出现一个“高原”区. 为此, Zhang 等人[1]提出了一种双阈值比较技术, 它使用两个阈值 Tb 和 Ts, $Ts < Tb$. 如果前后两帧的帧间差 $d(I_i, I_{i+1})$ 满足 $Ts < d(I_i, I_{i+1}) < Tb$, 就认为它们是潜在的渐变切换的开始帧. 对每一个潜在的转换, 计算累积的差值, $Ac(i) += d(I_i, I_{i+1})$, 直到满足 $Ac(i) > Tb$, 就认为有镜头渐变, 当 $d(I_i, I_{i+1}) < Ts$ 时, 则认为渐变结束. 这种方法虽然能够成功地检测渐变转换, 但镜头的缓慢运动也具有上述特点, 从而导致误检测.

为了将渐变切换与镜头摇动和推拉镜头区分开来, Zhang 等人[1]还采用了光流计算方法. 其原理是镜头渐变切换时没有光流, 而镜头运动应适合某种特定的光流类型. 这种方法可以取得较好的检测效果, 但计算非常复杂, 且在镜头间的颜色直方图很接近或在光照变化很大的情况下, 会发生检测失败.

利用帧间差的镜头边界检测算法的一个重要问题就是要选择合适的阈值. 阈值过大, 会漏掉镜头切换; 阈值太小, 会引起误检测, 即把镜头内镜头或物体的运动(此时帧间差值增大) 误检测为镜头切换. 不同类型的视频应选择不同的阈值, 如体育比赛的镜头运动较多, 应选择较大的阈值, 而新闻节目主持人的镜头, 运动较少, 应选择较小的阈值. 为了使检测算法具有更强的适应性, 阈值应根据视频的内容自适应地选定.

2.1.4　基于模型的方法[11, 13~15]

上述方法都是利用帧间差自下而上来进行镜头边界的检测, 它对于突变检测可以取得较好的效果, 但是对于渐变检测则有一定的困难, 因为它在很大程度上忽略了渐变切换中帧之间结构上的相关性. 可是基于模型的方法是利用对镜头编辑的先验知识, 对各种镜头切换建立一定的数学模型, 自顶向下地进行镜头切换的检测, 因此这种方法对镜头渐变的检测往往能取得好的效果.

Hampapur 等人[13, 15]通过对视频制作过程的研究, 找到了一种可用于镜头边界检测的视频编辑模型(Video Edit Model). 例如, 一个典型的镜头渐变模型可表示为

$$f(x,y,t) = \alpha(t) g_1(x,y,t) + \beta(t) g_2(x,y,t)$$

其中, $g_1(x,y,t)$ 是即将逐渐消失的镜头; $g_2(x,y,t)$ 是即将逐渐出现的镜头, 如果镜头内没有运动或运动很小, 则可分别记为: $g_1(x,y,t) \cong g_1(x,y)$, $g_2(x,y,t) \cong g_2(x,y)$. $\alpha(t)$ 和 $\beta(t)$ 都是时间的线性函数; 假设渐变转换的持续时间为 0 到 T. 对于慢转换, 它们可以表示为

$$\alpha(t) = \begin{cases} 1, & t < 0 \\ 1 - t/T, & 0 \le t \le T \\ 0, & t > T \end{cases}$$

$$\beta(t) = 1 - \alpha(t)$$

对于淡出, 则 $g_2 = 0$; 对于淡入, 则 $g_1 = 0$. 在变化的过程中, 每幅图象上所有的象素都以线性规律变化. 可定义如下的常量图 CI(Constant Image)

$$CI(x,y,t) = \frac{\partial f(x,y,t)/\partial t}{f(x,y,t)}$$

假设镜头为无运动的线性淡出, 即: $\alpha(t) = 1 - t/T$, $\beta(t) = 0$, $g_1(x,y,t) \cong g_1(x,y)$, 则可以得到

$$\begin{aligned} CI(x,y,t) &= (\partial\alpha(t)/\partial t * g_1(x,y) + \alpha(t) * \\ &\quad \partial g_1(x,y)/\partial t)/(\alpha(t) g_1(x,y)) \\ &= -(1/T)\alpha(t) \end{aligned}$$

这样, 对于一定的时间 t, 我们得到所有象素均为常数的常量图 CI, 检测渐变只需检测模型的常量图. 对于给定的模型, 一旦检测到常量图, 则认为有一个渐变过程.

只要模型建立准确, 基于模型的方法对于渐变检测往往能得到较好的效果, 但是需要对每种切换类型建立模型, 而且建模过程比较复杂.

由于大量的视频数据都是以 JPEG、MPEG 等压缩形式存在, 因而有必要对压缩视频进行镜头切换的检测. 这种检测通常可以采用两种方法[5]:

(1) 先进行解压, 形成图象帧序列, 然后再使用

上面讨论的未压缩图象的检测方法. 这种方法的缺点是进行完全解压比较耗时.

(2) 不完全解压, 直接对压缩视频数据进行镜头转换的检测, 这种方法可以节省一些解码时间. 事实上, 上述各种帧间差的计算方法都用到压缩视频中的 DC 系数, 另外, 利用压缩视频中的运动向量还可以提取运动特征. Arman[16] 等人利用 M-JPEG 中的 DC 系数来计算帧间差; Patel 和 Sethi[12] 则使用 MPEG 压缩视频中 I 帧的 DC 系数来计算亮度直方图. Zhang[1] 等人还使用 MPEG 压缩视频中的 DCT 块和运动向量, 对非零的运动向量进行计数, 结果他们发现处于镜头转换处的 B 帧和 P 帧中有效运动向量的个数较少, 因此可对这些帧进行解压, 用非压缩视频的边界检测方法进行镜头边界的检测. 这类方法的缺点是, 目前的 MPEG 算法是面向数据压缩的, 而不是面向视频内容表示的. 例如, 镜头的开始帧不一定正好是 I 帧, 因而要取出单个的镜头, 还要依赖前面的镜头.

2.2 代表帧的选取[1, 12, 17]

代表帧是用于描述一个镜头的关键图象帧, 它反映一个镜头的主要内容. 代表帧的选取一方面必须能够反映镜头中的主要事件, 因而描述应尽可能地准确完全, 另一方面为便于管理, 数据量应尽量地小, 且计算不宜太复杂.

代表帧的选取方法很多, 比较经典的有帧平均法和直方图平均法[18]. 帧平均法是从镜头中取所有帧在某个位置上象素值的平均值, 然后将镜头中该点位置的象素值最接近平均值的帧作为代表帧; 直方图平均法则是将镜头中所有帧的统计直方图取平均, 然后选择与该平均直方图最接近的帧作为代表帧. 这些方法的优点是计算比较简单, 所选取的帧具有平均代表意义. 缺点是, 从一个镜头中选取一个代表帧, 无法描述有多个物体运动的镜头. 一般说来, 从镜头中选取固定数目的代表帧不是一种好的方法, 因为这种方法对于变化很少的镜头则选取的代表帧过多, 而对于运动较多的镜头, 用一两个代表帧又无法充分描述.

Yeung[4], Zhang[1] 等人依据帧间的显著变化来选择多个代表帧. 他们计算前一个代表帧与剩余帧之差, 如果差值大于某一个阈值, 则再选取一个代表帧. 这种方法可以根据镜头内容的变化程度选择相应数目的代表帧, 但是所选取的帧不一定具有代表意义, 而且在有镜头运动时, 容易选取过多的代表帧.

Wolf[17] 通过光流分析(Optical Flow Analysis)来计算镜头中的运动量, 在运动量取局部最小值处选取代表帧, 它反映了视频数据中的静止, 往往表示一种强调的实际情况. 这种方法首先用 Horn-Schunck 法计算光流, 对每个象素光流分量的模求和, 作为第 k 帧的运动量 $M(k)$, 即:

$$M(k)=\sum_i\sum_j |O_x(i,j,k)|+|O_y(i,j,k)|$$

其中 $O_x(i,j,k)$ 是帧 k 内象素 (i,j) 光流的 X 分量, $O_y(i,j,k)$ 是帧 k 内象素 (i,j) 光流的 Y 分量.

然后寻找 $M(k)$ 的局部最小值. 从 $k=0$ 开始, 扫描 $M(k)\sim k$ 曲线, 找到两个局部最大值 $M(k_1)$ 和 $M(k_2)$, $M(k_2)$ 的值与 $M(k_1)$ 的值至少相差 $p\%$(由经验设定), 如果 $M(k_3)=\min(M(k))$, $k_1<k<k_2$, 则把 k_3 选为代表帧. 然后把 k_2 作为当前的 k_1, 继续寻找下一个 k_2. Wolf 的这种基于运动的方法可以根据镜头的结构选择相应数目的代表帧. 如果先把图象中的运动对象从背景中取出, 再计算对象所在位置的光流, 可以取得更好的效果.

2.3 特征提取

镜头是视频检索的最小单位. 视频分割成镜头后, 就要对各个镜头进行特征提取, 得到一个尽可能充分反映镜头内容的特征空间, 这个特征空间将作为视频聚类和检索的依据. 视频数据的特征分为静态特征和动态特征.

2.3.1 静态特征提取

静态特征的提取主要针对代表帧, 可以采用通常的图象处理方法, 如提取颜色特征、纹理特征、形状和边缘特征等.

颜色(Color)特征提取 颜色是用于图象相似性比较的最常用的一个特征. 近年来提出了多种基于颜色的视频索引技术[2, 19~21]. 在 QBIC[21]、JACOB[2] 等许多系统中, 都采用了颜色直方图方法, 颜色直方图是给定一个离散的颜色空间, 把它分成 n 个区域, 每个区域取它的中心色作为代表, 计算落入每个区域内象素的个数, 就得到了颜色直方图——一个 n 维的特征空间. 在计算颜色直方图之前, 往往需要把 RGB 空间通过非线性变换, 形成其它的颜色空间(如 HSV, Munsell 空间). 因为在这些空间里, 颜色三元组之间的距离更符合人眼的视觉差异. Zhang *et al.*[1] 除了采用与上述方法相似的颜色直方图方法外, 还增加了关键帧的主要颜色和平均亮度两种特征.

纹理(Texture)特征提取 纹理是图象分类和识

别的另一个主要特征, 例如, 森林和沙滩具有不同的纹理特征. 纹理分析方法可以大致分为统计型和结构型两类[22,23], 其中统计方法是找出图象的数值特征, 例如 Fourier 频谱特性、共生矩阵(co-occurrence matrices)、Markov 随机场模型等; 结构方法则是首先假定纹理模式由纹理基元按一定的规则排列组成, 因而纹理分析就变为确定这些基元, 并定量分析它们的空间排列. 单纯的结构方法仅适用于非常规则的图案, 而实际图象很少具有这种规则性, 因此把统计方法和结构方法结合使用, 往往能取得较好的效果. 在各种纹理分析方法中, Zhang[1,10] 选择了 Tamura 特征(对比度, 方向性和粗糙度) 和同步自退化模型(Simultaneous Auto-Regressive (SAR) Model); 而 JACOB 系统[23] 中则采用了边缘密度(Edge Density) 的方法, 即边缘象素与总象素之比, 这种方法首先将代表帧分割为四个相等的区域, 然后对每个区域分别沿 0°、45°、90°和 135° 4 个方向计算边缘密度, 从而得到一个 16× 1 的基于纹理的特征向量.

形状 (Shape) 特征提取　代表帧中的主要对象往往反映了视频的重要内容, 这些对象可以通过其形状来表示. 形状分析首先要把对象从背景中分割出来, 再使用圆形度、矩形度、矩等各种方法进行形状的相似性比较. 由于形状的相似性比较仍是一个很困难的问题[1,10], 因而目前在视频处理领域使用得较少.

2.3.2　运动特征[6,18] 提取

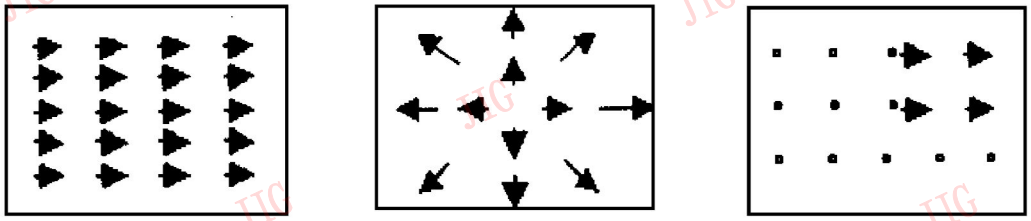
视频数据除了具有静态特征外, 还更具有运动特征, 它反映了视频数据的时域变化, 而且往往是用户检索时所能给出的主要内容, 例如用户可能要求检索有变焦的视频片段, 或者在监控系统中检索某个对象从画面上消失的视频帧. 因而对视频数据进行特征提取必须研究其运动特征.

摄像头的运动往往会给视频图象带来全局的影响, 例如水平移动镜头会使所有的象素点也水平地移动, 焦距拉长会使象素点从中心向四周发散, 焦距缩短会使象素点从四周向中心会聚. 在只有对象运动时, 大部分背景象素不变, 而只是运动对象和被摭挡的部分会发生变化.

由于运动特征无法从一幅静止的图象中获得, 所以必须对视频序列进行分析.

运动分析的方法有基于光流方程的方法、基于块的方法、象素递归方法和贝叶斯方法[22] 等, 但这些方法计算量都非常大. 为此, Tonomura 等人提出一种称为 X 线断层分析[12] 的方法. 它把一个镜头的视频序列看成一个整体, 通过对这个序列沿时间轴进行切片, 从而得到 x-t 切片图象和 y-t 切片图象. 然后分析切片图象, 即可以看出镜头的运动情况.

另一种可以避免耗时的光流和块匹配计算方法是利用 MPEG 视频流中 B 帧和 P 帧的运动向量[10,12]. 摇镜头时, 大多数的运动向量方向相同, 且大小相似; 推拉镜头时, 运动向量会由中心指向四周(或由四周指向中心); 对于对象运动, 大部分运动向量为零或很小, 仅是对应于对象运动的部分运动向量较大(如图 3 所示).



(a) 向左移动镜头　(b) 推镜头　(c) 对象运动

图 3　镜头运动的运动向量

Patel 和 Sethi[12] 计算宏块运动向量(指 B 帧和 P 帧) 在 8 个方向的分布, 即:

$$F_j = \frac{1}{K}\sum_{i=1}^{K}\theta(i,j)$$

其中 F_i 表示第 i 个方向的运动向量所占的比例; K 是帧中宏块的总数; 而 $\theta(i,j)$ 取值为 0 或 1, 表示第 i 个运动向量是否在方向 j 上. 加上没有运动的块所占的比例, 就得到一个有 9 个分量的特征向量. 利用这个特征向量可以将镜头分为静止、对象运动、镜头摇动、跟踪、变焦和综合运动 6 类.

Zhang[1,10] 等人用计算镜头内各帧平均亮度和主要颜色的均值和方差来作为镜头运动量大小的度量, 他们用这种方法把新闻节目视频段分为主持人和新闻内容, 并取得了良好的效果.

视频数据运动特征的研究, 多年来一直得到广泛的重视, 已取得了不少研究成果, 但到目前为止, 还有许多问题没有解决, 例如大物体(占据屏幕的大部分空间) 的运动与镜头的运动难以区分, 显露的被

遮挡背景与运动对象难以区别,特别在光照条件发生变化时,大多数的运动分析方法都会失效.

2.4 视频聚类

视频聚类是研究镜头间的关系,也就是如何把内容相近的镜头组合起来.根据聚类目的的不同,视频聚类可分为两类[3]:一类是把同属一个场景的镜头进行聚类,以形成层次型的视频结构——场景和电影.这种聚类不但要考虑镜头内容上的相似性,还要考虑其时间上的连续性,也就是说,虽然两个镜头内容很接近(特征向量之间的距离很小),但如果它们在时间上相距得很远,就不能认为它们属于同一个场景.把镜头聚类为故事单元后,其数量明显减少.例如对于一部典型的连续剧,半小时的节目中约有 300 个镜头,经过聚类后可形成约 20 个故事单元.

另一类聚类是对视频进行分类.它只考虑特征相似性,而不考虑时间连续性.根据镜头的重复程度,视频一般可分为对话型、动作型和其它类型 3 类.对话型视频是指一段实际的对话或者象对话一样由两个或多个镜头重复交替出现的视频.动作型视频则反应故事的展开,镜头不是固定在一个地点或跟随一个事件,因而很少发生镜头的重复.例如,一个有 13 个镜头的视频序列,各镜头分别标记为:

A B A B A B A B C D E F G

其中,前 8 个镜头可认为是对话型的,而后 5 个则是动作型的.根据镜头的持续时间可以对动作型视频进一步分类,例如一个有很多短镜头的视频可以认为是"快动作"型的.

通过视频聚类可以缩小检索的范围,提高检索的效率.

3 总结与讨论

视频数据处理是实现基于内容的视频检索的一项关键技术,它直接影响到视频特征匹配和检索的精度,其研究还处于起步阶段,各种理论和相关技术都不尽完善,需要继续做大量的研究探讨.我们认为视频处理技术可以从下面几方面进行研究:第一,镜头边界检测是基于内容检索的视频处理必不可少的第一步,能否准确地(漏检率和误检率都很低)检测出镜头边界,直接关系到以后的处理,而且镜头边界检测所用到的颜色、纹理和运动特征都可以用于镜头的索引,所以有必要对此进行重点的研究;第二,由于视频数据的特点在于其时变性和动态性,因而如何更好描述摄像头的各种运动和对象的运动也是一个研究重点;第三,基于内容的视频检索不应局限于镜头、场景和故事单元等这些基于帧的概念,应从视频对象(Video Object)的角度加以研究;第四,因为视频数据常常伴随着音频和字幕,利用这些信息或许可以为提取视频数据的高级语义找到一种可行的方法;第五,由于大量的视频数据是以压缩形式存储的,直接对压缩视频数据进行处理不但可以节省解压时间,还可以利用运动向量提取运动特征,因此是一种极有应用价值的方法[24,25].最后由于视频的数据量很大,对它进行各种处理往往要消耗大量的时间,因此寻找快速算法也是基于内容检索的视频处理必须研究的一个问题[26].

参 考 文 献

1 Zhang H J, Wu Jianhua, Zhong Di *et al*. An integrated system for content-based video retrieval and browsing. Pattern Recognition, 1997, 30(4): 643~ 657.

2 Edoardo Ardizzone, Marcola Cascia. Automatic video database and retrieval. Multimedia Tools and Applications, 1997, 4: 29~ 56.

3 Bolle R M, Yeo B L, Yeung M M. Video query: research directions. IBM Journal of Research and Development, 1998, 42(2): 233~ 252.

4 Yeung M M, Yeo B L. Video visualization for compact presentation and fast browsing of pictoriam content. IEEE Transactions on Circuits and Systems for Video Technology, 1997, 7(5): 771 ~ 785.

5 Haitao Jiang, AbdelSalam Helal. Scene change detection techniques for video databases systems. Multimedia Systems, 1998 (6): 186~ 195.

6 Courtney Jonathan D. Automatic video indexing via object motion analysis. Pattern Recognition, 1997, 30(4): 607~ 625.

7 Wei Xiong, John Chung-Mong Lee. Automatic dominant camera motion annotation for video retrieval. SPIE, 1997, 3312: 50~ 59.

8 Michael Hoetter. Differential estimation of the global motion parameters zoom and pan. Signal Processing, 1989, 16: 49~ 265.

9 Srinivasan M V, Venkatesh S, Hosie R. Qualitative estimation of camera motion parameters from video sequence. Pattern Recognition, 1997, 30(4): 593~ 606.

10 Zhang H J *et al*. Video parsing, retrieval and browsing: An integrated and content-based solution. In: Proc. of ACM Multimedia'95 San Francisco, 1995. 15~ 24.

11 Toller M S, Lewis, Nixon M S. Video segmentation using combined cues. Proc. SPIE, 1997, 3312: 414~ 425.

12 Patel Nilesh V, Sethi Ishwar K. Video shot detection and characterization for video databases. Pattern Recognition, 1997, 30 (4): 583~ 592.

13 Song S Moon-Ho, Kwon Tae-Hoon. On detection of gradual scene changes for parsing of video data. SPIE, 1997, 3312: 404~ 409.

14 Adnan M. Alattar. Wipe scene change detector for use with video compression algorithm and MPEG-7. IEEE Transactions on Consumer Electronics, 1998, 44(1): 43~ 51.

15 Hampapur A, Jain R, Weymouth T. Digital video segmentation. In: Proc. Second Annual ACM Multimedia Conference and Exposition ACM, New York, NY, USA, 1994, 357~ 364.

16 Arman F, Hsu A, Chiu M Y. Image processing on compressed video data for large video databases. ACM Multimedia, 1993, 267~ 272.

17 Wolf Wayne. Key frame selection by motion analysis. In: Proc. of IEEE Int. Conf. On Acoustics, Speech and Signal Processing, ICASSP, Atlanta, 1996, 7~ 10.

18 Bilge Gunsel, Tekalp A Murat. Content-based access to video objects: Temporal segmentation, visual summarization, and feature extraction. Signal Processing, 1998, 66: 261~ 280.

19 Michael J Swain. Color indexing. International Journal of Computer Vision, 1991, 7: 11~ 32.

20 Suresh K Choubey, Vijay V Raghavan. Generic and fully automatic content-based image retrieval using color. Pattern Recognition Letters, 1997, 18: 1233~ 1240.

21 Niblack W, Barber R. The QBIC project: querying images by content using color, texture, and shape. Proc. SPIE, 1993, 1908: 173~ 178.

22 Tekalp A Murat. Digital Video Processing. 北京: 清华大学出版社, 1998.

23 Jian Chang Mao, Anil K Jain. Texture classification and segmentation using multi-resolution simultaneous auto-regressive models. Pattern Recognition, 1992, 25(2): 173~ 188.

24 Shen Bo, Sethi Ishwar K. Direct feature extraction from compressed images. In: Proc SPIE 1996, 2670: 404~ 414.

25 Yasuyuki Nakajima, Kiyono Ujihara *et al*. Universal scene change detection on MPEG-coded data domain. In: Proc SPIE, 1999, 3024: 992~ 1003.

26 Adjeroh Donald A, Lee M C. Techniques for fast partitioning of compressed and uncompressed video. Multimedia Tools and Applications, 1997, 4: 225~ 243.

金　红　1968 年生, 57322 部队高级工程师. 现为上海交通大学图象通信与信息处理研究所博士生. 主要研究方向为图象通信与视频处理.

周源华　1940 年生, 上海交通大学图象通信与信息处理研究所教授, 博士生导师, 主要研究方向为目标识别、三维建模和多媒体与视频处理.