基于内容视频检索的关键技术研究

黄知义 周 宁
（武汉大学信息管理学院，武汉 430072）

〔摘 要〕 从分析基于内容的视频检索系统的基本结构和原理出发，文章重点探讨了基于内容视频检索的一些关键性技术，如：镜头分割、关键帧抽取、特征量提取、视频索引建立等的实现原理及其各自的优缺点，并对国内外视频检索的发展现状及发展趋势提出了一些看法。

〔关键词〕 视频检索；关键技术；镜头分割；关键帧抽取；特征提取

〔**Abstract**〕 This paper started from talking about the system structure and principle of content-based video retrieval. Then it mainly introduced its four key techniques: Scene Incision, Key Frame Extraction, Feature Extraction and Construction of Video Index and so on. At last, the author brought forward several viewpoints on the status quo and developing trend of video retrieval.

〔**Key words**〕 Video Retrieval; Key Techniques; Scene Incision; Key Frame Extraction; Feature Extraction
〔中图分类号〕TP391.3 〔文献标识码〕C 〔文章编号〕1008－0821（2005）10－0126－04

工作研究

1 引 言

近年来，随着多媒体编码、计算机多媒体处理和网络传输技术的飞速发展，人们已可通过互联网实时查询、欣赏和产生丰富多彩的视频信息。互联网正逐渐成为一个巨大的视频仓库。如何有效地组织和检索视频信息已成为数据库领域以及信息检索领域中研究的关键性问题。传统的基于文本的视频检索方法利用文本信息对视频内容进行注释，通过对关键字进行抽取来描述视频内容的语义特征。但由于目前的技术还不能对视频内容的语义特征进行自动描述，仍需使用手工的方法对视频进行解释和注释。这是一项耗时的工作，而且由于主观上的因素，可能造成注释的不准确性，因此，基于文本的检索方法已不能满足需求。

基于内容的视频检索（Content－Based Video Retrieval），根据视频的内容和上下文关系，对大规模视频数据库中的视频数据进行检索。它在没有人工参与的情况下，自动提取并描述视频的特征和内容。它以图像处理、模式识别、计算机视觉、图像理解等领域的知识为基础，从认知科学、人工智能、数据库管理系统及人机交互、信息检索等领域，引入新的媒体数据表示和数据模型，从而设计出可靠、有效的检索算法、系统结构以及友好的人机界面。目前，基于内容的视频检索研究，除了识别和描述图像的颜色、纹理、形状和空间关系外，主要的研究集中在视频分割、特征提取和描述（包括视觉特征、颜色、纹理和形状及运动信息和对象信息等）、关键帧提取和结构分析等方面。本文将就基于内容视频检索的基本原理及其一些关键技术进行分析阐述。

2 视频检索系统的基本原理

首先，视频作为一种表达信息的媒体，它有着自己独立的结构。一般来说，一段视频由一些描述独立故事单元的场景构成；一个场景由一些语义相关的镜头组成；而每个镜头是由一些连续的帧构成，它可由一个或多个关键帧表示。一段视频的典型结构如图 1 所示。

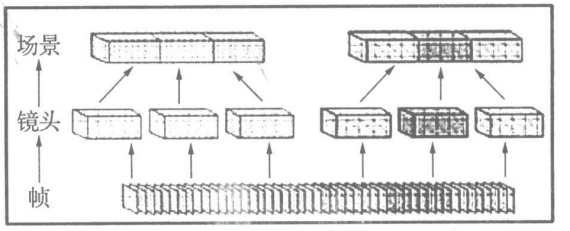

图 1 典型视频结构图

基于内容视频检索，首先要实现的是对视频数据的分析处理，基于内容的视频分析，即从所有的帧中提取主要内容，并从下至上对视频内容进行结构化描述。为了实现这个目标，我们需对视频进行如下处理：视频切分、关键帧抽取、特征提取和视频内容组织等。

图 2 描述了基于内容的视频处理的主要过程。视频首先被分割成各个镜头，并实现对各个镜头的特征提取，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。其中，特征提取包括关键帧中的视觉特征和镜头的运动特征的提取，所谓关键帧，即指从视频数据中抽取出来的、能概括镜头内容的一些静态图像，通过一定算法实现，对这些静态图像的视觉特征提取，主要从颜色、纹理、形状等几个角度来进行。镜头运动特性提取通过对镜头的运动分析（主要针对镜头运动的变化、运动目标的大小变化，视频目标的运动轨迹等）来进行，方法主要有基于光流方程的方法、基于块的方法、像素递归方法和贝叶斯方法等。然后，根据提取的关于镜头的动态特性和关键帧的一些静态特性，进行索引。最终，用户可以通过一种简单方便的方法浏览和检索视频。其整体模块图如图 3 所示。

收稿日期：2005－08－28
作者简介：黄知义（1982－），男，武汉大学信息管理学院情报学系 2004 级硕士研究生，研究方向为信息检索与信息可视化。
周 宁（1942－），男，武汉大学信息管理学院教授，博士生导师，主要研究方向为信息组织、信息检索、信息可视化。

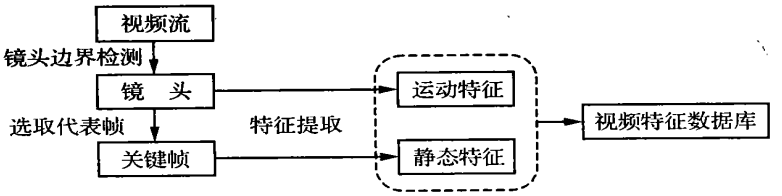

图 2 视频结构化处理过程

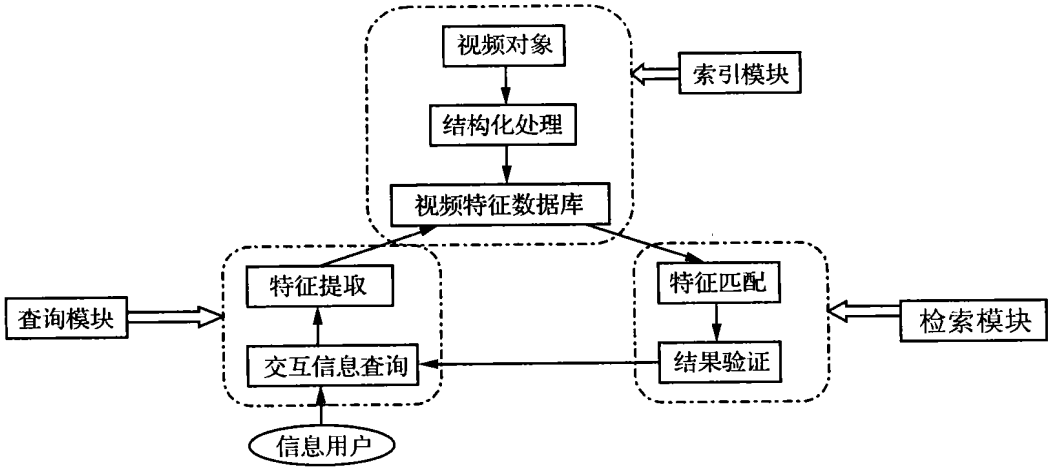

图 3 基于内容视频检索系统的模块描述图

由图3 可以看出，基于内容的视频信息检索系统大致由索引、查询、检索三大模块组成。索引模块中，系统运用镜头切割，关键帧抽取，特征提取等技术对视频信息进行预处理，并实现视频特征索引的建立；查询模块是实现人机交互的重要接口，通过该模块用户可以容易地实现交互信息查询，即用户可以较容易地向系统提交自己的各项特征要求（包括视频示例特征提取），并可根据对检索结果的进一步特征提取实现较精确检索；检索模块主要实现视频特征索引与用户提交的各项特征的相似度计算、特征匹配，并根据相关度进行结果排序提交。下面将就其一些关键技术进行分析介绍。

3 基于内容视频检索的关键技术

3.1 视频镜头切割

视频流由成千上万的图像帧组成，帧是组成视频的最小单位，如果对每一帧都进行单独处理，则索引和检索的效率会很低。幸运的是，视频通常是由大量的逻辑单位或分块组成，我们称这些分块为视频镜头。一个镜头是相邻帧的短序列，它描绘同一场景，表示一个摄像机动作，一个事件或连续的动作。任何视频都是由镜头衔接起来的，镜头是视频检索的基本单元。对视频数据进行索引处理，首先要把视频自动分割为镜头。

当视频情节内容发生变化时，会出现镜头切换。通常，镜头之间的切换方式主要可分为两大类：突变和渐变。突变是两个镜头之间没有过渡，直接从一个镜头跳跃至下一个镜头；渐变指一个镜头向另一个镜头渐渐过渡的过程，没有明显的镜头跳跃，包括淡入淡出、溶合和擦洗等。视频的镜头分割不仅要将镜头间的突变检测出来，还要将渐变分割出来。若手工对视频流进行镜头分割，相当耗费时间，不利于实用化。研究人员正致力于自动分割镜头技术。

目前视频镜头分割技术主要根据镜头在发生切换时其视频数据所反映的变化来进行。由于一个镜头内的相邻帧间的变化不会很大，它们之间的特征差值总会限定在某个阈值内。在镜头突变时，突变点前后两个相邻帧通常在内容上都显示着很大的量的变化，如果特征差值超过了给定的阈值，则意味着出现一个分割边界。通常检测的方法有：像素比较法，即比较相邻两帧对应的像素值的变化量来确定镜头边界；块比较法，把帧分为互不重叠的块，将这些块在下一帧进行块匹配，根据匹配产生的运动矢量和匹配值来确定镜头的边界，块匹配能够避免采用像素比较分割平扫镜头时产生的误分；直方图比较法，即根据相邻两帧的直方图差异来决定是否为镜头边界。

上述方法对镜头的突变切换有较好的效果，但对镜头的渐变切换检测效果不理想。对于镜头的渐变切换，由于相邻帧之间的特征差值很小，通常要比镜头阈值低得多，但却又高于镜头内的差值，因此，很难用单个的阈值来检测，更为复杂有效的分割方法必须被采用。双比较技术的提出，很好的解决了这个问题。

双比较技术要求使用两个差值阈值：阈值 T_b 用于检测正常的镜头切变；阈值 T_s 则较小，用于检测在渐变可能出现的地方、可能出现的帧。在镜头边界检测过程中，使用差值度量对相邻的帧进行比较。如果差值大于 T_b，则可宣布存在一个镜头边界，需要进行镜头切分。如果差值小于 T_b 但大于 T_s，则可把后一帧标记为一个潜在的变换帧，即可能渐变的开端（Fs），然后把该帧与后续帧比较，将相邻帧之间的差值加起来。在镜头发生渐变过程中“累计比较差”会逐渐增加，如果帧间差减到小于 T_s，而累计差超过 T_b，则标注为渐变的结束（Fe），而且可把相邻潜在转换帧当作一个特殊的块。（注：只有当相邻帧与帧之间的差值大于 T_s 时，才计算累计比较差。如果相邻帧差小于 T_s，而累计比较差小于 T_b，就放弃先前标注的渐变可能开始点（Fs），寻找下一个可能的渐变。这就是说，在一个渐变中，所有的帧与帧的差值，即当前帧与前一帧之间的差值应大

于 T_s 而小于 T_b。）

双比较技术的提出有效的解决了图像像素差法、颜色直方图差法、颜色直方图平方差法等一些基本的镜头分割技术中难以选取合适阈值的问题，很大程度上降低了镜头边界的误检率和漏检率。对渐变式镜头边界的检测，除了双比较技术法，还有统计分析法（基于帧间像素变化的分布进行统计，根据像素变化的数目分辨出种类渐变式镜头边界）、边界编辑模型法（根据视频的镜头边界编辑模型来检测镜头边界，实际上是视频编辑的反过程）等。

另外，由于数字视频文件通常以压缩形式存储，因此，镜头的检测要能够直接对压缩视频数据进行。对压缩视频镜头检测的方法与未经压缩的视频信息大体相似，主要是将相邻两帧的压缩数据进行比较。如，离散余弦变换（DCT）系数检测法，根据 DCT 系数定义，其中 DC 系数分量相当于帧的低分辨率图，对相邻的压缩帧对比 DC 系数，可使压缩视频镜头边界检测达到足够精度。再如，由于同一镜头内各帧之间的运动矢量通常是连续的，而且在 MPEG 数据流中，编码规则本身含有运动数据，因此，可以通过运动矢量的变化来检测镜头边界。

3.2 关键帧的提取

关键帧是表达镜头内容的代表帧，是从视频数据中抽取出来的、能概括镜头内容的静态图像。因此，关键帧的抽取是建立视频数据索引的关键，它不仅要解决如何从镜头的帧序列中抽取出能代表镜头内容的关键帧，还要从关键帧中析出其所含内容与所属类别。从目前的研究来看，已取得了一定的成果。

3.2.1 把镜头首帧和尾帧选为关键帧

通过对拍摄者或制作者的心理分析，研究者认为视频信息制造者总是希望镜头的开始就能抓住观众的眼球，镜头的结尾能够让观众回味无穷。因此，首帧往往决定了镜头的主题，尾帧通常表示一种特写。所以，选择首帧和尾帧作为镜头的关键帧不失为一种较为简便和有效的方法。

3.2.2 选择显著变化帧作为关键帧

一个镜头其帧序列的图像特征可能变化很大，即首帧和尾帧不能概括镜头全部内容，需要从中抽取新的帧。这种选取方法为：将镜头中的每一帧与首尾声帧颜色直方图进行比较，如果它与首尾帧均存在显著变化，将其作为关键帧，继续将未比较过的帧与这三个帧比较，若三个帧均有显著差异也作为关键帧。如此比较下去，直至比较完毕。

3.2.3 帧平均法选取关键帧

计算镜头中所有帧的某个位置像素的平均值，然后选择在该位置上像素值最接近平均值的那一帧作为关键帧。另一种方法是直方图平均法，即，选择与镜头平均直方图最接近的那一帧作为关键帧。

3.2.4 根据关键对象确定关键帧

把关键帧分解成更小的单元，从这些更小的单元中获取关键对象，这些关键对象可以从关键帧中抽取，也可从镜头甚至全局中抽取。如电视剧中的主要人物画像等。有了关键对象就可以根据适当的策略选取关键帧。如，关键帧的图像中必须有关键对象；具有多个关键对象的帧；根据关键对象的颜色、形状和运动的平均值等属性比较出关键帧。

当然关键帧的抽取有些方面还必须考虑到（许多研究者已注意到这个问题），如，关键帧数量的选取；如何将关键帧对应到相应的主题和类目；关键对象如何分割等。

3.3 特征提取

视频分割成镜头、关键帧被抽取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。特征提取可分为关键帧中的视觉特征和镜头的运动特征的提取，具体包括：颜色、纹理、形状和运动等几个方面。

3.3.1 颜色特征

颜色是图像最显著的特征，与其它特征相比，颜色特征计算简单、性质稳定，对于旋转、平移、尺度变化都不敏感。颜色特征包括颜色直方图、主要颜色、平均亮度等。其中利用主要颜色和平均亮度进行图像的相似匹配是很粗略的，但是它们可以作为层次检索方法的粗查，对粗查的结果再利用子块划分的颜色直方图匹配进行进一步的细查。为了能够在大规模图像数据集中进行快速的搜索，Smith 和 Change 等人提出了颜色集的概念：首先将 RBG 颜色空间转换为视觉上一致空间 HSV，然后量化为 m 个颜色条，颜色集就定义为量化后的颜色空间中颜色的一种选择。由于颜色集特征向量是二叉的，因而可以通过构造二叉树来进行快速的检索。

3.3.2 纹理特征

纹理是与物体表面材质有关的图像特征，具有照明不变性。纹理分析一直是计算机视觉的一个重要研究方向，其方法主要分为两类，即结构方法和统计方法。结构方法是假定图像有较小的纹理基元排列而成，它采用句法分析方法，但只适用于规则的结构纹理分析；统计方法是对图像的颜色强度的空间分布信息进行统计，又可进一步分为传统的基于模型的统计方法和基于频谱分析的方法，如马尔可夫随机场模型、Fourier 频谱特性等。20 世纪 70 年代初 Haralick 等人提出了纹理特征的共生矩阵表示法，即利用纹理在灰度级的空间相关性，先根据图像像素间的方向和距离构造一个共生矩阵，再从中提出有意义的统计数据作为纹理的特征表示。该方法的缺点是这些统计特征没有和人在视觉上对纹理特征的感知之间建立对应。于是不少人提出了其它的纹理特征度量方法，其中 Tamura 提出的纹理特征集可以很好地与人类视觉感知相对应，这些特征包括：粗糙度（Coarness）、对比度（Contrast）、方向性（Diractionality）、线像度（LineLikeness）、规则性（Regularity）、粗略度（Roughness）。其中最重要的特征是纹理粗糙度、对比度和方向性。另外，许多研究者也开始将小波变换应用于纹理特征表示。Manjunath 等人对三种小波变换方法（直角、树结构、Galbo）做了比较之后，发现 Galbo 小波变换最符合人类视觉特征的表达。目前也有不少研究者在研究利用神经网络进行纹理分割。

3.3.3 形状特征

形状分析首先需要采用合适的图像分割算法把不同对象从图像中分割出来，再用各种方法进行匹配测量。形状特征表示的一个重要准则是要求对位移、旋转、缩放的不变性，通常形状的表示可以分为基于边界和基于区域两类。它们分别采用傅里叶描述和矩不变量表述特征，另外新的研究方向有弹性变形模板和边界方向直方图。

3.3.4 运动特征

运动特征是视频镜头的重要特征，它反映了视频的时域变化，也是用视频例子进行检索的重要内容。运动分析的方法有基于光流方程的方法、基于块的方法、像素递归方法和贝叶斯方法等，但这些方法计算量大。于是，Tonomura 等人提出了 X 线断层分析的方法，将整个视频序列沿时间轴进行切片，从切片图像中分析运动情况。Patel 和 Sethi 提出利用 MPEG 中的 B 和 P 帧的运动向量来避免光流计

算和块匹配。该方法利用宏块的运动得到一个有九个分量的特征向量，再用这个特征向量判断镜头的运动。另外，用计算镜头内各帧平均亮度和主要颜色的均值和方差作为镜头运动量大小的度量，在新闻视频中也取得了较好的效果。

3.4 视频索引

实现对关键帧的视觉特征和镜头的运动特征的提取之后，即可进行视频索引的建立. 视频索引从不同的角度出发有不同的分类方式，从选取的索引内容出发，可以分成 3 类：基于特征的索引（Feature-based Indexing），基于注释的索引（Annotation-based Indexing）和基于特定领域的索引（Domain-specific Indexing）。

基于特征的索引：基于特征的索引是基于内容视频检索索引中重要的一块。目前，基于特征的索引技术的研究主要集中在图像特征索引和视频特征索引。图像特征索引面向一些不带时间延展性的特征如颜色、纹理、轮廓、形状等静态性的特征而建立；视频特征索引的建立则主要基于视频运动特征的提取而建立，在镜头层次表示视频数据的时间特征。上文中已就关键帧中的视觉特征和镜头的运动特征的提取作了详细阐述，在此特征索引的建立过程就不作累述。

3.4.1 基于注释的索引

基于特征匹配的视频检索的主要缺点是特征缺乏语义信息，使得用户在说明对视频数据的查询时感到不便。为此，研究人员提出了基于注释的检索。注释就是与特定视频段相关的语义属性集，可捕获视频的高级内容。目前基于注释的索引技术的研究主要集中在注释语言的选择、注释结构的设计，以及方便的人机交互式注释界面的设计 3 个方面。注释语言可以分为自然语句、关键词、源注释等。自然语句注释可以充分表达视频数据的语义内容，但随意性相当大。关键词注释可以预先给出供选择的关键词以减少随意性，但也存在信息表述不完全等不足，它的缺点在于忽视了多媒体与文本在数据组织方式、数据量等方面都有很大差别。而且关键词查询的方式在视频应用领域有两个致命的缺陷：首先，人工注解需要大量劳动力；其次，不同的人对同一视频内容有不同的理解，这种理解上的主观性和注释的不精确性会引起检索过程中匹配误差。源注释的基本思想是利用数字摄影机在视频数据流中加入相关信息作为视频注释的依据。

3.4.2 基于特定领域的索引

基于特定领域的索引是针对某个特定应用领域建立的索引，它们一般有固有的模式，如：新闻视频分主持人段和新闻内容段，篮球比赛分节等。对于特定领域的视频索引，可以首先根据它们的固有模式建立逻辑视频结构模型，在索引建立过程中，在对视频数据进行特征提取和分析的基础上将结果与模型匹配。一旦确定了匹配关系，就可以将模型对应的语义赋给相应的视频数据单元。

4 基于内容视频检索系统的发展现状

目前，国内外已研发出了多个基于内容的视频检索系统，主要有：

4.1 QBIC 系统

QBIC（Query By Image Content）是由 IBM Almaden 研究中心开发的，是“基于内容”检索系统的典型代表。QBIC 系统允许使用例子图像、用户构建的草图和图画及其选择的颜色和纹理模式、以及镜头和目标运动等图形信息，对大型图像和视频数据库进行查询。视频方面主要利用了颜色、纹理、形状、摄像机和对象运动来描述内容。

4.2 VisualSeek 系统

VisualSeek 是美国哥伦比亚大学电子工程系与电信研究中心图像和高级电视实验室共同研究的、一种在互联网上使用的“基于内容”的检索系统。它实现了互联网上的“基于内容”的图像/视频检索系统，提供了一套供人们在 Web 上搜索和检索图像及视频的工具。

4.3 VideoQ 系统

VideoQ 是哥伦比亚大学研究的一个项目，它扩充了传统的关键字和主题导航的查询方法，允许用户使用视觉特征和时空关系来检索视频。它有以下几个特征：集成文本和视觉搜索方法；自动的视频对象分割和追踪；丰富的视觉特征库，包括颜色、纹理、形状和运动；通过 WWW 互联网交互查询和浏览。

4.4 TV－FI 系统

TV－FI（Tsinghua Video Find It），是清华大学开发的视频节目管理系统。这个系统可以提供如下几个功能：视频数据入库、基于内容的浏览、检索等。TV－FI 提供多种模式访问视频数据，包括基于关键字的查询、基于示例的查询、按视频结构进行浏览、以及按用户自己预先定义的类别进行浏览。

5 结束语

基于内容的视频检索是当前信息检索的研究热点，它以图像处理、模式识别、计算机视觉、图像理解等领域的知识为基础，从认知科学、人工智能、数据库管理系统及人机交互、信息检索等领域，引入新的媒体数据表示和数据模型，实现对视频数据的有效检索。目前，在许多技术方面已取得了一定的进展，向基于内容的自动化视频管理迈进了一大步，但在实用性方面还存在一定的差距，如：视频关键帧的抽取，关键帧特征量的抽取等一些方面都还不能令人满意。为了取得基于高级特征和概念的自动化视频索引和检索，我们还需做更多的研究。

参 考 文 献

[1] 郑德俊，梅天宝. 基于 MPEG-7 视频信息检索探析[J]. 滁州学院学报，2004，(9)：75－78.

[2] 沈燕，任晓健. 基于内容的多媒体检索技术在数字档案馆中的应用[J]. 2004，(4)：91－93.

[3] 朱爱红，李连. 基于内容的视频检索中的镜头分割技术[J]. 情报检索，2004，(3)：66－68.

[4] 朱爱红，李连. 基于内容的视频检索关键技术研究[J]. 情报杂志，2004，(1)：45－47.

[5] 何立民，万跃华. 数字图书馆中基于内容的视频检索关键技术[J]. 中国图书馆学报，2003，(2)：52－56.

[6] 苏新宁. 视频信息索引技术研究进展[J]. 信息可视化与知识管理——2003 信息与信息资源管理学术研讨会论文选，21－29.

[7] Barrow H G. Parametric Correspondence and Chamfer Matching [C]. Proc. 5th Int. Joint Conf. Artificial Intelligence, 1977: 659－663.

[8] Tekalp A Murat. Digital Video Processing [M]. 北京：清华大学出版社，1998.

[9] Patel Nilesh V, et al. Video Shot Detection and Characterization for Video Databases [J]. Pattern Recognition, 1997, 30, (4)：583－592.