

语义视频检索的现状和研究进展*

余卫宇, 谢胜利, 余英林, 潘晓舟
(华南理工大学 无线电与自动控制研究所, 广东 广州 510640)

摘 要: 概述了图像的可视化特征如颜色、纹理、形状和运动信息, 时空关系分析, 以及多特征目标提取和相似度量度; 分析了视频语义的提取, 语义查询、检索; 探讨了视频语义检索的性能评估, 存在的问题和发展方向。
关键词: 特征提取; 图像检索; 时空分析; 视频语义物体; 视频查询; 语义检索
中图法分类号: TP391. 4 **文献标识码:** A **文章编号:** 1001-3695(2005)05-0001-07

Actuality and Development of Video Retrieval Based Semantic

YU Wei-yu, XIE Sheng-li, YU Ying-lin, PAN Xiao-zhou
(Institute of Electronic Engineering & Control, South China University of Technology, Guangzhou Guangdong 510640, China)

Abstract: Many information systems require a large multimedia storage and retrieval, including browsing, querying, etc. Image retrieval technique has gradually matured. Video semantic retrieval is yet a research hotspot. In this paper, we first discuss some visual features such as color, shape, texture and motion information, spatial-temporal analysis, multi-feature object extraction and similarity measure. Second, we analysis video semantic object extraction, retrieval via semantic query. At last, we discuss the performance evaluation of video semantic retrieval, open issues and promising directions.
Key words: Feature Extraction; Image Retrieval; Spatial-Temporal Analysis; Video Semantic Object; Video Query; Semantic Retrieval

1 引言

Internet 的日益普及和数据量呈几何级数的增长, 数据已经不是单纯的文本信息了, 网络每天更新的图像或视频内容的数量已经达到了亿级。为了从浩如烟海的多媒体数据中 fastest 查找用户感兴趣的内容, 国内外已开发出不少商业图像检索系统, 如 IBM 的 QBIC, VisualSeek, Virage, MARS, Photobook, Web-Seek, Netra 等, 它们分别从不同的图像特征进行检索。对于视频的检索尚不成熟, 没有通用的数据库检索系统。视频与图像相比有着自身的特点。视频具有较大的层次性、结构性、复杂性, 需要对视频文件分层处理; 视频检索中查询系统设计要比图像检索复杂得多。一个精心设计的检索系统对于用户应该具有灵活性、模糊性、可操作性和记忆性。

视频检索系统(VRS) 目的是帮助用户在已有的数据库中检索所需的视频序列。主要有三类: ①用户想查询一个已知的视频, 且知道该序列存在于视频数据库中, 用户可用准确的方法找到所需的目标序列。这时使用“关键字”的文本检索可以满足用户的需要。如要查询体育比赛的节目。②当用户已知要查询的文件, 而不知道要检索的内容是否在数据库中, 要解决的问题是设计一个搜索工具使用户能够快速准确知道要查询的目标文件是否在数据库中, 避免浪费时间。③用户根据所知道的标题或一些相关的事件进行检索。在这种情况下, 搜索应该是分级的, 并要用户提供短小且完整的描述。可以用相关

反馈方法实现, 并要求查询检索系统具有模糊处理能力。如用户需查询足球比赛中的进球场面, 但是这种查询的困难之一是输入的标题或事件的标注是否与视频中的相匹配。

VRS 在实际应用中是非常重要的, 它的用户主要有: 新闻广播、商业广告、音乐视频、远程学习、视频档案及医学应用等。

2 视频时间特性分析

通常视频包含空间和时间两方面内容。视频应该能够捕捉到各个场景的空间和时间关系。为了达到这个目的, 首先应该对视频流进行处理, 视频的组成可以分成如图 1 所示的几部分^[1]。

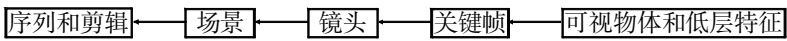


图 1 视频的结构模型

视频可以用一个视频结构模型描述。低层特征(Low Level Feature) 反映了图像物体的一些特征, 如颜色、纹理、形状等, 该层主要是分割出一些需要的物体; 帧层(Frame Level) 由于视频中每个帧是相互独立的, 在这层不包含时间分析; 镜头层是一系列连续的帧, 它是照相机的连续记录, 包含这组镜头各帧的时间关系及该组镜头中可能含有的视频语义; 场景层是一系列连续的镜头组成, 它具有相似的语义; 视频层是一个完整的视频, 可作为一个整体进行分析。

在上述各层中, 关键层是镜头层。区分镜头边界主要有三种类型^[2]: ①切换(Cut) 是指在不同镜头组的突变边界, 因为它的末帧与另一组镜头的首帧信息和运动状态有较大差别; ②溶解(Dissolve) 是指前一个镜头的最后一幅图像的内容和下一个镜头的头一幅图像相融合, 主要问题是融合效果与通过全局

运动所引起的改变的区别; ③擦除(Wipe)是指第二组镜头连续覆盖和消除第一个镜头的显示。镜头的切换检测是相对容易的, 主要原因是其突变性, 而融合和擦除对于检测是非常困难的。MPEG-4 文件压缩结构的一些有效的解决方法^[3~5]是基于全局运动估计和分割。场景的定义是对镜头的内容的深层理解。镜头的分割是困难的, 完全通过人工操作是不可行的, 通过半自动分割和标注工具达到目的^[6]。林通等人^[7]提出了镜头内容分析, 引入了两个内容描述符: 主色直方图和空间结构直方图的概念, 用主颜色直方图描述一组图像帧, 充分应用了时间信息, 并考虑了某些语义方面的信息。用空间结构直方图描述一幅图像的空间信息, 再利用镜头的高层聚类的自动场景标注。

如果把一个视频剪辑成若干个镜头组, 则需要提取每组镜头的关键帧。在单个镜头时, 如果场景的变化较大, 那么两帧可能是相似和相当不相似, 在这种情况下, 需要识别更多的关键帧。所以, 首先要确定场景的运动状况, 使这些镜头能进一步划分为小块。例如在相同镜头下不同的照相机运动的起始点可以作为关键帧。Dagtas^[9]提出了一种视频索引和检索的工具 PICTURESUE, 能够利用物体运动有效地处理用户的查询。该方法可以应用到视频数据库的许多方面, 从并行检索原始视频到时空数据模型。用可视化查询工具提供了一个友好接口满足用户查询, 同时分析了两种补充的运动模型有效的检索视频数据。

如何选择关键帧对于捕捉视频内容的变化是非常重要的。选择关键帧的理想方法是对于给定场景下, 比较各帧的相似度, 选择各帧中最小的差别作为关键帧, 但这种方法需要极大的计算量, 对于实时处理有较大的困难。一般来说, 相对于关键帧而言, 相同场景的其他帧具有较大的冗余度。我们可以很自然地先选择每个场景中的第一帧作为关键帧, 尽管该帧并不一定是场景中的最佳表达场景的内容。对于更一般的系统, 我们希望关键帧包含有最多的内容(如文本和没有遮挡的物体)。

3 特征提取

特征提取的目的是自动提取描述图像视频的特征。

3.1 彩色特征

颜色是视觉的特征之一。人观看外界事物时, 它是对人眼刺激最敏感的特征。许多学者对它进行了广泛的研究^[12]。Swain 和 Ballard^[2]最早提出了基于全局彩色分布的图像检索方法。图像的彩色分布是通过颜色直方图表示的, 通过离散化彩色空间计算有多少像素落入每个分散的颜色中。采用了互补颜色空间直方图来描述物体, 且利用直方图的相交及反投影算法完成物体的检索和定位。用 RGB 值表示颜色的方法虽然简单但不直观, 有许多其他彩色空间的表示方法, 如 HSV, YUV, HMMD 等。它们之间的转换是方便的, 如把 RGB 空间转换成 Y, Cr, Cb, 映射公式如下:

$$\begin{cases} Y = 0.299 \cdot R + 0.857 \cdot G + 0.114 \cdot B \\ Cb = -0.169 \cdot R - 0.331 \cdot G + 0.500 \cdot B \\ Cr = 0.500 \cdot R - 0.419 \cdot G - 0.081 \cdot B \end{cases} \quad (1)$$

还有些彩色空间如 YIQ 和 Lab 是根据经验亮度设计的, 它

用欧氏距离分别表示颜色感知距离^[13]。彩色图像检索的关键是彩色特征的提取, 目前主要方法有彩色直方图、累积直方图法、加权直方图法和颜色矩等的方法, 其主要优点在旋转, 变换和其他几何操作的不变性。一旦计算出彩色直方图, 可以使用直方图距离^[14]。

3.2 纹理特征

纹理是图像的重要特征之一, 目前虽然对纹理没有一个准确的定义, 但大部分人认为纹理是局部呈现不规则性, 甚至是杂乱无章的, 如岩石的表面、墙壁、水波等; 但从整体上却出现有规律的特性。纹理检测的常用方法是 Gabor 滤波器^[18], Wu^[15]提出了利用纹理描述子描述纹理, 其计算方法如下: 用 Gabor 滤波器组对图像进行滤波, 然后确定两个主纹理方向, 根据纹理方向确定其规则性和粗细度。

常用的纹理数据库有 Broadatz 纹理集、VisTEX、CURET 和 MeasTEX 框架。通过纹理能够有效地区分具有相似颜色(如天空和海洋、叶子和草地等)的内容。可以通过纹理特征进行图像检索时, 其主要方法有: ①基于空间性质的方法; ②基于统计特征矩阵的方法, 其主要有对比度、粗糙度、方向度、规则性, 周期性、方向性、随机性^[16]; ③基于随机模型的方法, 主要有 MRF 和 GRF 两种, 其思想是用概率模型, 采用最大似然估计, 通常首先对图像采用分块处理, 在各个块估计模型参数, 进行聚类, 构造与纹理相一致的数学模型参数。纹理是由不同的块组成的, 局部纹理的参数从而计算出。测量的主动轮廓能够被分割成纹理区域。扩展了传统的基于蛇形分割算法^[17]。

3.3 形状特征

颜色虽然在检索中应用非常广泛, 但它不能反映物体的具体特征。如要在图像数据库中检索“红色汽车”, 我们发现检索结果可能不仅有红色汽车, 还可能有日出, 国旗等红色特征明显的物体, 检索速度慢。而形状却弥补了这一缺陷。形状是物体的基本特征之一, 人类可以通过形状认识外界事物, 例如, 对于房屋、汽车、计算机、动物等。我们认识事物主要是通过其形状来区分不同物体的, 有些物体甚至是变形的, 或一幅图像中的物体是不完全的, 如对于有遮挡的汽车, 虽然我们没有看到它的全部, 但是由于我们具有各自形状的特征, 人可以通过特征判断出其属性, 认出该物体而不是其他的物体, 通过形状, 结合我们的感知认识可得到其语义信息。图 2(a)~图 2(d)分别是房屋、汽车、树和人, 有些通过形状判断会有困难, 如图 2(e)是 Tank, 我们要根据先验知识才能决定图中的物体是什么, 而有些图形我们要结合不同的图像特性, 如图 2(f)而通过形状是很难区分的, 但结合颜色信息我们可以很容易判断出黄色显示的是 IPI。而有些形状我们必须根据语义信息才能区分出来, 如图 2(g)单从前两幅图很难区分, 而把它们结合在一起, 就可判断出是花瓶的形状。所以在形状分析中借助语义信息是非常重要的。Biederman 指出自然物体通过它们的形状能够很容易识别。描述形状有许多方法, 其中主要有 Fourier 描述子、轮廓描述子和曲率描述子、区域描述子, 包括有 Zernike 矩和 Grid 描述子。在文献[24]中, Taoufik Gadi 提出了一种局部特征和测量索引的形状相似度检索的方法。首先根据形状的凹凸和曲率把形状标为记号, 记号区域有一个分布, 用两个不同的距离函数来处理符号相似度和形状相似度。通过把

形状符号排入 M 树的索引结构得到形状索引。

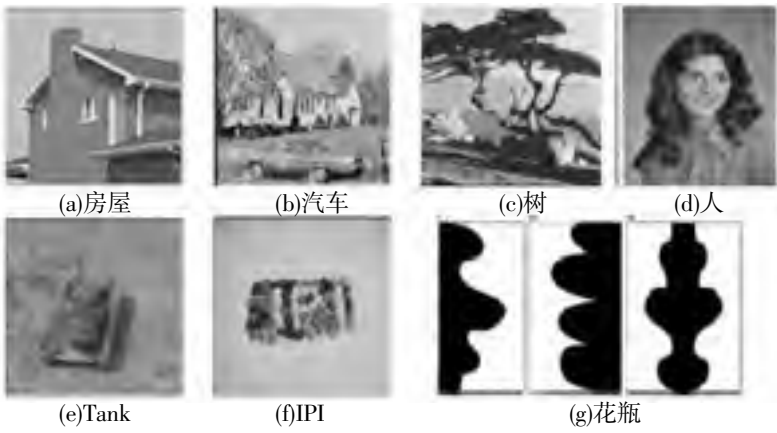


图 2 各种形状的图像

3.4 多特征空间

在视频检索系统中,典型的特征抽取方法主要有彩色直方图法、形状特征法、纹理特征法和轮廓特征法。其中,基于彩色的特征抽取能较好地表示图像的全局颜色信息,但对于局部处理的效果较差,且出现颜色特征相近而语义信息完全不相关的情况。形状特征抽取方法通用性差,而纹理特征抽取算法不能处理无纹理的图像,应用面较窄。在视频检索系统中已经使用了全局图像特征^[19,20]。Li^[21]提出了利用颜色、纹理和空间信息的综合检索的方法。在检索之前,先提取颜色直方图、小波提取纹理、无监督分割算法,然后利用了多特征滤波器组,彩色直方图滤波器、小波纹理滤波器、空间滤波器进行检索。

4 相似度量度

如何用数值来表示图像特征上的相似程度,是相似度量问题。相似度量直接影响检索效果。在模式识别技术中,特征相似度用特征向量的空间距离表示,常用的如马氏距离、欧氏距离等。直方图间的距离用欧氏距离表示为

$$M_E(Q,D)=\sqrt{\sum_{i=1}^L[H_Q(i)-H_D(i)]^2}$$

(2)

这里 $H_Q(i)$ 和 $H_D(i)$ 分别为查询图像 Q 和数据库图像 D 的特征的统计直方图。

在文献[22~26]中提到了视频的相似测量。笔者把视频序列的特征值表示为一串符号串。对于两个视频序列的匹配问题作为模式识别的问题。处理视频串距离作为视频序列匹配的新的相似度量。基于内容的图像检索文献[27~29]也论述了用该方法把视频序列与序列的关键帧相结合。但是,这种方法的缺陷是忽略了序列的时间信息,仅仅用于二维可视特征。这种方法要求先对视频进行分割,并准确分割出视频的关键帧。Ersboll^[30]中提出了任意给定形状视频物体相似度测量的一种新方法,比较视频物体的低层特征,并把它应用到检索视频物体上。实验表明使用相似度匹配技术优于通常的特征矢量方法,并且具有稳定性。Man-Kwan Shan^[31]提出考虑了时间顺序的序列帧的相似度测量及算法。相似度可以通过特征和语义描述来测量,视频矢量是独立于语义信息,通过视频分割和半自动分割提取出来。本文提出了七种相似度测量和响应算法。所有这些算法是基于动态程序方法和计算时间。Nguyen^[32]提出了基于运动相似度的合并区域的鲁棒性方法。Swanberg^[33]提出了一种广播视频流的分割方法。文献[34]提出了联合的视频检索方法,对不确定的形式和大容量的视频数

据通过相似度特征查询和标注。

5 语义提取

在多媒体信息检索中,语义是一个非常重要的概念,目前语义的定义还不完善。我们对语义的理解是人能够从图像或视频中所得到的信息,它包括视频中存在的重要的物体和它们之间的时间和空间关系以及它背后所隐含的内容。如物体 a 在物体 b 的上面,而物体 c 在物体 d 的下面,当它们的位置发生变换时,我们所得到的信息是不同的。例如在足球比赛中,当足球在两个门柱之间我们得到的信息是进球了,而检测发现足球在两个门柱外,则没有进球。另外时间信息对于视频理解也是非常重要的,如事件 X 发生在事件 Y 之前而不是之后,人们所得到的信息是不同的。人们希望能用语义检索代替内容检索。正是从该方面进行考虑的。最近视频检索正在从基于例子和关键词的检索到使用视频语义进行查询,用户能够通过具有语义信息的关键词进行查询,也必须根据多媒体特征表示这些语义结构特性^[35]。Harry 和 Marios^[36]详细分析了视频语义层的语义模型,如图 3 所示。文章从各方面分析了物体、事件和行为的语义模型。R. Naphade^[37]提出了视频检索的训练语义概念的三步骤:①标注集的界面;②基于标号样本的语义概念的表示框架;③基于关键词的查询系统。

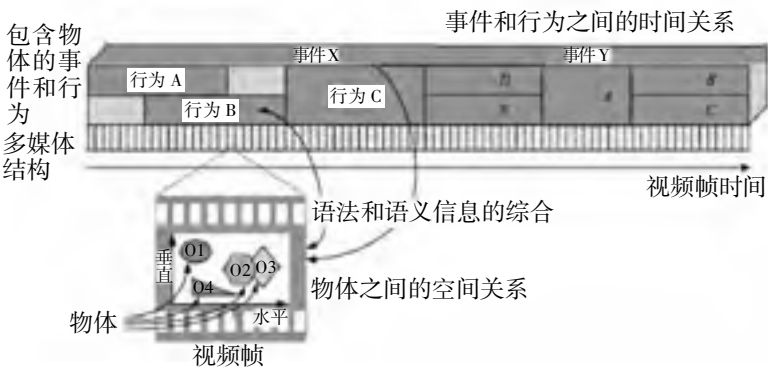


图 3 多媒体内容的语义表示

5.1 语义模型

在日常生活中,人们由于工作忙,上网的时间有限,希望在最短的时间内得到最多的信息,因此迫切需要有一个快速的视频检索,了解近几天的内容。而且每个人对视频的要求是非常不同的。所以“内容”本身也有不同的要求,如有些人比较关心新闻、体育、天气预报,而有些人对于特定的场景感兴趣,如要查看长城、白宫(White House)。在浏览新闻节目时,我们只是要了解新闻中有多少内容及观看感兴趣的片段。在上述所有应用中,传统的文本(Textual)方法的描述性能显然是不能令人满意的。因为关键字不能完全表达视频信息的时间结构,也不能给出视频内容的正确描述,具有较大的主观性。所以语义分析与提取至关重要。传统的检索方法^[38,39]不能有效地利用视频流中的语义信息。其实,语义是一种结构性的概念。Lee^[40]提出了一种新的视频语义模型,表达了视频的语义。根据视频的知识数据库划分了五层结构模型,包括原始数据层和语义数据层。可以通过不同的时间测量视频流的确定的语义表示。并设计了一个交互式的接口浏览和查询视频数据。文献[41]提出了基于在线事件结构的语义数据模型,可以充分表示视频文件的各种人的解释,并介绍了一种正规的查询语言实现用户的不同的查询。

5.2 语义物体提取

在实际中,需要面对广泛的数字媒体的传输、存储及使用等。为了达到该目的, MPEG-4 提供的一个基于内容的框架,提出了视频语义物体的概念(SVO)。但是 MPEG 组织没有制定具体的 SVO 的提取方法,而在大多数视频应用中 SVO 的提取却是必不可少的。它的应用主要有:视频的压缩和传输;多媒体应用如搜索、操作、合成;个人可视化通信服务等。

视频物体的提取方法已有很多文献报道,但是目前准确提取语义的视频物体的方法却是非常困难的,主要原因有:① SVO 的定义是多样的,任何有意义的物体都可以作为 SVO,如足球比赛中的球,行走中的人、照相机、运动的汽车等。可以说只要在视频流中出现的都可以当作 SVO 进行分析。② SVO 的提取是依赖于视频分割过程,而视频分割在图像处理和计算机视觉中仍是一个尚未完全解决的问题。③低层的分割标准(如颜色、纹理、灰度、运动等)不能进行语义物体分析。除了上述之外,还要考虑 SVO 的计算代价; SVO 的跟踪和 SVO 的后处理过程。文献[42, 43]研究了从低层特征映射到高层语义方法。基于与其他 SVO 不同的运动特性来提取 SVO。目前已有的大多数 SVO 自动提取方案利用了运动信息生成语义物体。运动信息的分割主要有时间分割^[44]、时空分割^[45, 46]、空间分割与时间跟踪^[47]三类。在语义物体提取中,我们不需要分割的精确性,只要分割出运动物体的轮廓,在不准确的 SVO 中仍包含有部分背景,且丢失分割的物体的部分内容。如上面提到的足球比赛场景,需要首先分割出几个静止物体,如场地、两个球门,而运动物体如人和足球。在分析射门球的场景时,我们要跟踪足球的运动路线及判断球的运动轨迹是否在两个门柱之间。在文献[48]中作者提出了一个新的视频语义提取方法。不单利用了视频序列的第一帧和所有帧进行时空分割,而且自适应完成时空分割和融合,并具有较低的计算复杂度。除此之外,作者提供了一个灵活的系统的转换设计,可以全自动提取和半自动提取,满足用户的不同需求。其算法如下:为了提取第 I 帧 SVO,系统给出了第 $I-1$ 帧和第 $I+1$ 帧。系统首先对全局运动进行估计与补偿,为了估计模型的参数,使用分层的块匹配和最小均方差值。然后计算帧 $I-1$ 和 k 帧的变化。最后根据两个准则提取 SVO。文献[49]提出了一种基于水线统计分析法,考虑了边缘标志提取算法结合颜色和边缘的生长算法。嵌入部分格型框架,设计了一种空时区域最大似然算法提取 SVO。

6 视频语义检索

有效的查询工具对于视频检索是必要的。对图像检索而言,通过例子的查询系统是相对直观和简单的。而对于视频文件来说,通过例子查询不是很直观而且不方便,它要求用户提供一个已知的视频内容。

6.1 运动检索

在视频中,运动是表示时间信息的唯一方法。基于运动查询是视频搜索引擎所特有的。运动查询的困难是怎样准确描述它。例如描述正在上升的气球是困难的。基于运动的查询相对而言不是很直观,用户需要用有效的方式表示运动。JACOB^[50, 51]是一个彩色和运动视频的搜索引擎,用户被要求选

择全局运动方向性和大小。VideoQ^[52]也是一种基于运动搜索引擎。在文献[53]中,作者提出了利用可变的自动场景分割和物体跟踪算法对各种语义层的物体运动计算。

6.2 文本检索

在 Web 中应用关键词查询文本内容是很普通的。在文献[55, 59]对视频文本分割和提取进行了分析。在视频的分析中,文本信息也是重要的语义信息之一,对于有些视频分析,使用运动信息分析或基本的特征分析会比较复杂,而用文本分析反而很直观性。Wernicke 和 Lienhart 提出了一种新的具有鲁棒性的多分辨率的视频文本的定位、分割和识别的方法。通过测试从 MPEG-1 HDTV(MPEG-2) 视频序列,文本的大小从 8 像素到半帧。但是文本查询有它内在的缺点,首先自动标注是非常困难的。因为它的工作量是相当巨大的,人工操作是不可能的,并且需要耗费大量的时间;其次是一幅图像的内在的语义信息提取对于不同的用户可能有不同的结果。Li^[54]针对视频关键帧中的文本信息,用混合小波和神经网络自动分割出图形和场景文本。总而言之,文本检索是非常重要的,因为它提供了用户较大的灵活性。但是它又存在较大的局限性。

6.3 可视化检索

使用可视化内容检索的系统应该能够提供一个查询方框图使用户很自然地指定有选择和非精确的查询;定义一个对于应用范围有意义的检索资源;定义一个使用户感觉满意的相似度度量。

可视化例子的查询是交互式的,它利用人类对图像分析和解释的本身的能力。它也要求能从图像自然地提取可视化特征并能够进行索引。

基于可视化例子的查询允许表达非精确性的和不完备性的查询输入,从图像中用户可以抽取其记忆的轮廓,所以不能期望达到较满意效果的结果。

7 语义视频检索的性能评价,存在的问题及发展方向

7.1 性能评价

如何评价处理后的性能,以及所用方法的优劣是图像处理后关键问题,制定一个评判标准是非常重要的。对于视频编码用 MSE 或 PSNR 判别算法的好坏。也有研究者提出了客观图像质量评价标准。为了评价实际应用系统性能,如查询的灵活性、查询的效率、检索的正确率等。迫切需要有一个客观的评价标准。MPEG-7 正在收集图像数据库作为测试的标准。文献[60]中提出了一种对图像检索的评价,对自动检索技术与它们的性能进行了比较。为了进行较好的评判,需要建立一个视频参考库,包括从视频镜头中提取的帧,尽管提取的每帧是不同的,但是它们具有相同的语义内容。检索的性能可以根据其效率和有效性评定。一般来说,效率可以根据对于一个查询需要多大的工作量评价。有效性可以用重检率(Recall) 和精度(Precision) 衡量。文献[61]提出了基于彩色检索的评价。这些分析是根据颜色相似度的人的评价,计算机的仿真结果是通过评价分数的方法进行分析的。其结果表明对于每一种情况没有一种最优的检索方法。文献[62]建立 NHANES2 的数字医学图像检索系统,通过使用形状参数作为检索的依据,为

了找到一种最佳的检索算法, 该组织分析了目前已有的各种形状检索算法, 制定了评价标准, 并给出了评价的结果和分析。

7.2 存在的问题及发展方向

多媒体信息中含大量的数据如图像、图形、视频剪辑和音频信号等。视频检索存在的问题主要有以下几点:

(1) 如何处理大数据量问题。Internet 上 Web 查询图像、视频信息比文本信息困难得多, 主要原因是网络上的数据不是固定不变的, 而是时刻改变的, 图像或视频内容每天及时更换。例如 Alta Vista^[63] 有约三百万幅图像。用目前的检索系统处理网上检索是不现实的。计算代价太大, 检索需要非常长的时间。

(2) 检索数据库的多样性。对于实际的系统, 由于用户的要求不同, 准确度不同, 检索的时间要求也不同。有些用户可能只需要快速查询相关的内容, 而有些用户需要精确的检索结果。目前的通用检索系统还是专业的, 不能像文本检索, 输入关键词就可以查到用户所需要的结果。如何解决用户的查询要求, 如何设计一种通用的查询语言, 如文本的关键字和 SQL 等。如何沟通用户的需求和计算机中的数据库相匹配, 将是研究的难点之一。

(3) 如何自动提取的视频数据库的特征。首先用户向数据库提交查询要求, 然后系统如何根据该要求自动提取出用户所需的结果, 交互式语义生成, 再根据视频语义的相关性查找满足用户所需求的视频剪辑。Chang^[64] 提出了一种视觉语义模板(Semantic Visual Template), 该方法需要用户提供一个颜色、纹理、形状特征的草图, 系统自动根据这些特征范围生成多个查询请求, 反馈给用户, 用户选出其中较好的检索结果。依次类推, 直到用户较为满意为止。但是这种方法还不算是完全的语义检索, 所以自动生成语义信息是研究的一个难点问题。

(4) 如何考虑人的因素。视频检索的一个最大障碍是如何考虑人在检索中的作用, 视频检索的最终是给人看的, 所以从视频内容分析人的感知是非常重要的。文献[65] 分析了图像数据库检索系统性能的心理研究。得出使用语义信息、早期输入的记忆和相对的图像相似度的判断可以达到最优的检索效果。

视频检索系统的发展方向也就是要解决上述存在的问题。

8 基于语义检索的应用

近几年来, 各国研究机构已经开发出一些实用系统, 主要有以下几个检索系统。

8.1 棒球的视频检索系统^[66]

系统首先通过图像相似度检测视频的特殊场景, 称为基本场景。在提取出这些场景后, 使用动态的过程匹配确定投掷和击球的准确位置。当击球手击球时, 确定击到球后的帧作为其最后帧, 然后对这些建立数据库索引, 形成视频数据库。该数据库可实现以下目的: ① 识别击球开始和结束; 击球员的挥棒及其结果; 击球员的改变; ② 运动员的辨别。该系统的前期工作主要是视频的分割; 视频场景文本的识别; 击球和接球的动作检索。在实验中, 系统使用了 2 小时约 20 000 帧的视频进行分析, 对于击球和接球的正确检索率分别为 96% 和 89%。

实验证明该系统具有较好的效果。

8.2 Gesture 的视频检索及其应用^[67]

该系统能够从视频流中提取特征参量, 执行自动查询, 检索相似结果, 用检索结果实时替代出现的物体。也对检索的准确度和运算代价进行了分析。本文使用姿势表示物体方向的运动, 照相机的运动, 物体的形状等, 并且设计了“交互式视频数据库”, 如图 4 所示。

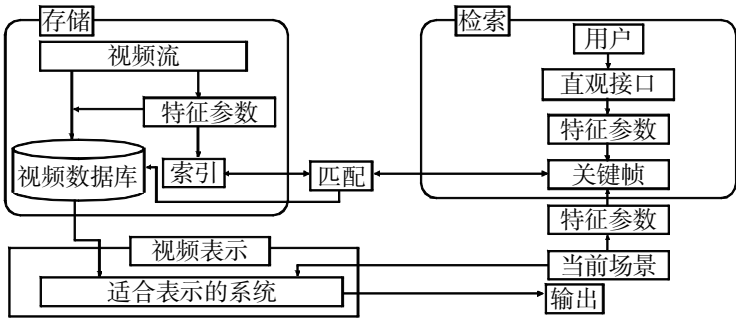


图 4 交互式的视频检索系统

(1) 存储视频和相关信息。作为电视新闻内容, 自动提取它是很困难的, 系统必须给出大量的人为给出索引。文献自动提取特征参量作为索引。为了在视频图像中输入上述参数, 设计了 GUI 工具用来输入彩色和形状。匹配信息的方法和关键: 例如, 如果用户要试图检索“红色跑车的场景”, 检索结果可能不是很好, 因为红色范围识别的范围太大。除此之外, 使用形状和运动作为关键和索引, 通过语言和数字系统的输入方法不仅是不直观而且很难, 检索出相应的图像。

(2) 关于 Gesture 的视频数据库的检索。主要是能客观地匹配视频中姿势的运动和物体的运动以便自动地进行检索。笔者描述了使用姿势的检索技术, 从用户的姿势提取特征参数; 从视频图像中提取特征参数; 从上述两者中实现完全匹配。

8.3 VideoQ 自动视频检索系统^[68]

VideoQ 是一个高级的基于内容的视频检索系统, 有下列主要特点: 视频序列物体自动分割和跟踪; 具有丰富的可视特征库, 如颜色、纹理、形状和运动信息; 多个物体的查询; 查询的空时限制。VideoQ 是基于视频搜索系统, 用户使用灵活的方法进行检索。场景切换检测和全局运动稳定性合并物体分割过程。

除上述系统外, ADVISOR^[69] 设计了高级的数字视频存储和在线检索系统。文献[70] 针对 MPEG-7 应用基于内容的视频检索数据库, 在线查询和滤波, 用可视特征和运动信息描述并对视频语义物体进行分类处理。

9 结论

图像检索技术已经有了长足的发展, 各国已经研制出许多基于内容的图像检索系统, 如 Illinois 大学的 MARS, UCSB Alexandria 数字图书馆的 Netra、MIT 的 Photobook 等。

在图像检索的早期, 人们利用图像的基本特征, 如颜色、纹理、形状及运动信息, 由于它们在检索中的局限性, 不能满足人们的实际需要, 从而综合各种视觉特征进行联合检索。但是由于视频中包含较大量的语义信息, 语义的视频检索仍是研究的重点和难点。近几年, 对于视频语义的研究也进行了很多工作, 仍未解决的问题如视频语义如何定义及理解, 如何建立通

用的视频数据库, 计算机如何自动提取视频的语义信息, 如何把人的主观感知加到视频检索中。视频检索的研究与其他领域有紧密联系, 如物体识别、计算机视觉, 并且它还与其他学科有着紧密的联系, 如模式识别、数据库建立、可视化查询、可视化数据模型、人视觉系统、心理学、感知认识学等。视频检索的研究还处在初级阶段, 检索的方法目前归根到底还是要结合图像的一些基本特征进行, 至今还没有突破性的进展。

参考文献:

[1] Rubin, G Davenport. Structured Content Modeling for Semantic Information[J] . SIGGHI Bull, 1989, 21(2) .

[2] Swain M J, Ballard D H. Color Indexing[J] . Journal of Computer Vision. 1991, 7(1) : 11-32.

[3] P Bouthemy, M Gelgon, F Ganansia. An Unified Approach to Shot Change Detection and Camera Motion Characterisation[R] . Rennes, France: Technical Report. IRISA, 1148, 1997.

[4] R Milanese, F Deguillaume, A Jacot-Descombes. Video Segmentation and Camera Motion Characterization Using Compressed Data [EB/OL] . <http://citeseer.ist.psu.edu/440740.html>.

[5] H H Yu, W Wolf. A Hierarchical Multi-resolution Video Shot Transition Detection Scheme[J] . Computer Vision and Image Understanding, 1999, 75(1/2) : 196-213.

[6] J Meng, *et al.* A Compressed Video Editing And Parsing System[C] . Boston, MA, USA: Proceedings of ACM Multimedia '96, 1996.

[7] 林通, 张宏江, 等. 镜头内容分析及其在视频检索中的应用[J] . 软件学报, 2002, 13(8) : 1577-1585.

[8] Yang Yongsheng, Lin Ming. A Survey on Content Based Video Retrieval[EB/OL] . <http://citeseer.ist.psu.edu/248151.html>.

[9] Serhan Dagtas, Wasfi AL-Khatib, Rangasamis L Kashyap. Models for Motion-based Video Indexing and Retrieval[J] . IEEE Transaction on Image Processing, 2000, 9(1) : 88-101.

[10] N Dimitrova, *et al.* Video Key Frame Extraction and Filtering: A Key Frame is Not Key Frame to Everyone[C] . International Conference on Information and Knowledge Management, 1997.

[11] H J Zhang, *et al.* Video Parsing, Retrieval and Browsing and Integrated and Content Based Solution[EB/OL] . www.cse.iitk.ac.in/users/cs726/Resources/Videoparsing.htm.

[12] Cotton. Color, Color Spaces and the Human Visual System[R] . Univ. of Birmingham, 1995.

[13] T Gevers, A W M Smeulders. A Comparative Study of Several Color Models for Color Image Invariants Retrieval[C] . Proceedings of the 1st International Workshop ID-MMS, 1996. 17-26.

[14] J Puzicha, *et al.* Empirical Evaluation of Dissimilarity Measures for Color and Texture[C] . Proceedings of the IEEE International Conference on Computer Vision, 1999. 1165-1172.

[15] Wu P, Manjunath B S, Newsam S, *et al.* Texture Descriptor for Browsing and Similarity Retrieval[J] . Signal Processing: Image Communication, 2000, 16(1, 2) : 33-43.

[16] K I Chang, K Bowyer, M Sivagurunath. Evaluation of Texture Segmentation Algorithms[C] . Proceedings of the Conference on Computer Vision and Pattern Recognition, 1999. 294-299.

[17] N Paragios, R Deriche. Geodesic Active Contours for Texture Segmentation [R] . Sophia-Antipolis, France; Technical Report 3340, ROBOVIS, INRIA, 1998.

[18] Dennis Dunn, William E Higgins. Optimal Gabor Filters for Texture

Segmentation[J] . IEEE. Trans. , 1995, 4(7) : 947-964.

[19] E Ardizzone, ML Cascia. Multifeature. Image and Video Content-based Storage and Retrieval[C] . Proc. of SPIE. 1996. 265-276.

[20] Y Deng, *et al.* Content-based Search of Video Using Color, Texture and Motion[C] . Proc. of ICIP, 1997. 534-537.

[21] Xiuqi Li, Shu-Ching Chen, Mei-ling Shyu, *et al.* Image Retrieval by Color, Texture, and Spatial Information[EB/OL] . <http://www.cs.fiu.edu/~chens/PDF/DMS02.pdf>.

[22] Glen Jeh, Jennifer Widom, Sim Rank. A Measure of Structural-Context Similarity[EB/OL] . <http://www-cs-students.stanford.edu/~glenj/simrank.pdf>.

[23] Eero Sormunen, Marjo Markkula, Kalervo Jarvelin. The Perceived Similarity of Photos-test Collection-based Evaluation Framework for the Content-based Image Retrieval Algorithms[EB/OL] . <http://citeseer.ist.psu.edu/cache/papers/cs/26686/http://zSzzSzwww.info.uta.fizSztutkimuszSzfirezSzarchivezSzPerceived.pdf/the-perceived-similarity-of.pdf>.

[24] Guo-Dong Guo, *et al.* Learning Similarity Measure for Natural Image Retrieval with Relevance Feedback[J] . IEEE Trans. on NN, 2002, 13(4) : 811-820.

[25] D A Adjero, M C Lee, I King. A Distance Measure for Video Sequence Similarity Matching[C] . Proc. Int. Workshop on Multimedia Database Management Systems, 1998. 72-79.

[26] D A Adjero, M C Lee, I King. A Distance Measure for Video Sequences[J] . Computer Vision and Image Understanding(Special Issue on Content-based Access for Image and Video Libraries) , 1999, 75(1, 2) : 25-45.

[27] Selim Aksoy, Robert M Haralick. Textural Features for Image Database Retrieval[C] . Santa Barbara, CA: Proceedings of IEEE Workshop on Content-based Access of Image and Video Libraries in Conjunction with CVPR '98, 1998. 45-49.

[28] Jian-kang Wu. Content-based Indexing of Multimedia Databases [J] . IEEE Trans. on Knowledge and Data Engineering, 1997, 9(6) : 978-989.

[29] David McG Squire, *et al.* Content-based Query of Image Databases: Inspirations from Text Retrieval[C] . Pattern Recognition Letters (Selected Papers from the 11th Scandinavian Conference on Image Analysis SCIA '99) , 2000. 1193-1198.

[30] B K Ersboll, *et al.* A Robust Distance Measure for the Retrieval of Video Objects[C] . Image Analysis and Interpretation, Proceedings 5th IEEE Southwest Symposium on, 2002. 40-44.

[31] Man-Kwan Shan, Suh-Yin Lee. Content-based Video Retrieval Based on Similarity of Frame Sequence[EB/OL] . <http://www.computer.org/proceedings/iw-mmdbms/8676/86760090abs.htm>.

[32] Hieu T Nguyen, *et al.* Detection of Moving Objects in Video Using A Robust Motion Similarity Measure[J] . IEEE Transaction on Image Processing, 2000, 9(1) : 137-141.

[33] Deborah Swanberg, Ramesh Jain. In Support of Similarity Measures. Department of Computer Science and Engineering[Z] . University of California at San Diego.

[34] Mi Hee, Yoon Yong Ik, Kio Chung Kim. Unified Video Retrieval System Supporting Similarity Retrieval[C] . 10th International Workshop on Database& Expert Systems Applications Sep, 1999. 1-3.

[35] Lew M S. Features Selection and Visual Learning. In: Principles of Visual Information Retrieval[Z] . Lew M S. ed Springer. 2001. Ch.

6: 145-162.

[36] Harry W AGIUS, Marios C Angelides. Modeling Content for Semantic-Level Querying of Multimedia[J] . Multimedia Tools and Applications, 2001, 15: 3-37.

[37] Milind R Naphade, Chang-Yung Lin, John R Smith. Learning Semantic Multimedia Representations from Small Set of Examples[R] . IBM Thomas J. Waston Research Center, NY.

[38] T Chua, L Ruan. A Video Retrieval and Sequencing System [EB/OL] . <http://citeseer.ist.psu.edu/491919.html>.

[39] M Yeung, B-L Yeo, B Liu. Extracting Story Units from Long Programs for Video Browsing and Navigation[C] . Multimedia Computing and Systems. Proceedings of the 3rd IEEE International Conference on, 1996. 296-305.

[40] Jia-Ling Koh, *et al.* Semantic Video Model for Content-based Retrieval[EB/OL] . National Taiwan Normal University.

[41] Lilac A E Al Safadi, Janusz R Getta. Semantic Modeling for Video Content-based Retrieval System[EB/OL] . <http://csdl.computer.org/comp/proceedings/acsc/2000/0518/00/05180002abs.htm>.

[42] Rong Zhao, W I Grosky. From Features to Semantics: Some Preliminary Results[EB/OL] . <http://citeseer.ist.psu.edu/484818.html>.

[43] A Mojsilovic, B Rogowitz. Capturing Image Semantics with Low-Level Descriptors[EB/OL] . IBM TJ Watson Research Center. <http://www.research.ibm.com/visualanalysis/papers/mojsilovic.pdf>.

[44] J Y A, *et al.* Representing Moving Images with Layers[J] . IEEE Trans. Image Processing, 1994, 3(5) : 625-638.

[45] R Mech, *et al.* A Noise Robust Method for 2D Shape Estimation of Moving Objects in Video Sequences Considering a Moving Camera [J] . Signal Processing, 1998, 66(2) : 203-217.

[46] Jinghong Mao, *et al.* Semantic Spatio-temporal Segmentation for Extracting Video Objects [C] . Multimedia Computing and Systems, IEEE International Conference on, 1999. 738-743.

[47] D Wang. Unsupervised Video Segmentation Based on Watersheds and Temporal Tracking[J] . IEEE Trans. on Circuits and System for Video Technology, 1998, 8(5) : 539-545.

[48] Fuhui Long, *et al.* Extracting Semantic Video Objects[J] . Computer Graphics and Applications, IEEE, 2001, 21(1) : 48-55.

[49] Daniel Gatica-perez, Chuang Gu, Ming-Ting Sun. Semantic Video Object Extraction Using Four-band Watershed and Partition Lattice Operators[J] . Circuits and Systems for Video Technology, IEEE Transactions on, 2001, 11(5) : 603-618.

[50] E ardizzone, M La Cascia. Automatic Video Database Indexing and Retrieval[J] . Multimedia Tools Applications, 1997, (4) : 29-56.

[51] Ardizzone E, *et al.* Motion and Color-based Video Indexing and Retrieval, Pattern Recognition[C] . Proceedings of the 13th International Conference on, 1996. 135-139.

[52] S-F Chang, *et al.* VideoQ-An Automatic Content-based Video Search System Using Visual Cues[EB/OL] . <http://www.acm.org/sigmm/MM97/papers/sundaram/acmMM97paper.html>.

[53] Pu-Jien Cheng, Wei-Pang Yang. Querying Video Contents by Motion Example[C] . Database Applications in Non-Traditional Environments (DANTE ’99) Proceedings International Symposium on, 1999. 287-293.

[54] Huiping Li, *et al.* Automatic Identification of Text in Digital Video Key Frames, Pattern Recognition[C] . Proceedings 14th International Conference on, 1998. 129-132.

[55] Jae-Chang Shim, Dorai C, Bolle R. Automatic Text Extraction From Video for Content-based Annotation and Retrieval, Pattern Recognition [C] . Proceedings 14th International Conference on, 1998. 618-620.

[56] Byung Tae Chun, Younglae Bae, Tai-Yun Kim. Automatic Text Extraction in Digital Videos Using FFT and Neural Network[C] . Fuzzy Systems Conference Proceedings, FUZZ-IEEE ’99 IEEE International, 1999. 1112-1115.

[57] Jain A K, *et al.* Automatic Text Location in Images and Video Frames, Pattern Recognition[J] . Proceedings 14th International Conference on, 1998. 1497-1499.

[58] Huiping Li, Doermann D. Automatic Text Tracking in Digital Videos [C] . Multimedia Signal Processing, ’98 IEEE 2nd Workshop on, 1998. 21-26.

[59] Yu Zhong, Hongjiang Zhang, *et al.* Automatic Caption Localization in Compressed Video[J] . Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2000, 22(4) : 385-392.

[60] Vincenzo Di Lecce, Andrea Guerriero. An Evaluation of the Effectiveness of Image Features for Image Retrieval[EB/OL] . <http://www.geradts.com/html/Documents/Articles/SPIE2001/drugs.pdf>.

[61] Petteri Kerminen. The Visual Goodness Evaluation of Color-based Retrieval Processes Finland[EB/OL] . <http://www.pori.tut.fi/~kermi/publications/CR1336.PDF>.

[62] Sameer Antani L, *et al.* Evaluation of Shape Indexing Methods for Content-based Retrieval of X-Ray Image[EB/OL] . <http://archive.nlm.nih.gov/pubs/antani/srmd03/srmd03.pdf>.

[63] AltaVista[EB/OL] . www.altavista.com

[64] Cheng S F, *et al.* Semantic Visual Templates: Linking Visual Features to Semantics[C] . Proceedings ’98 International Conference on, 1998. 531-535.

[65] T V Papathomas, *et al.* Psychophysical Studies of the Performance of an Image Database Retrieval System[C] . ISFJT/ \$PIE Symposium on Electronic Imaging: Science and Technology, Conference on Human Vision and Electronic Imaging III, 1998. 591-602.

[66] Kawashima T, *et al.* Indexing of Baseball Telecast for Content-based Video Retrieval[C] . Proceedings ’98 International Conference on, 1998. 871-874.

[67] Kuramoto M, Masaki T, Kitamura Y, *et al.* Video Database Retrieval Based on Gestures and Its Application[J] . IEEE Transactions on Multimedia, 2002, 4(4) : 500-508.

[68] Shih-Fu Chang, Chen W, Sundaram H. VideoQ. A Fully Automated Video Retrieval System Using Motion Sketches[C] . Applications of Computer Vision, WACV ’98. Proceedings 4th IEEE Workshop on, 1998. 270-271.

[69] Advanced Digital Video Storage and On-line Retrieval System[EB/OL] . DG IST-Project 10147. http://www.advisor-project.com/AD-ViSOR_description.pdf.

[70] Wensheng Zhou, *et al.* Video Analysis and Classification for MPEG-7 Applications[J] . Consumer Electronics, 2000, ICCEP: 344-345.

作者简介:

余卫宇(1972-) , 男, 硕士, 博士研究生, 主要从事图像恢复、图像/视频语义分析及多媒体信息检索方面的研究; 谢胜利(1957-) , 男, 教授, 博士生导师, 主要研究方向为智能信息处理、盲信号处理、非线性系统学习控制、图像视频处理、视频无线传输等; 余英林(1932-) , 男, 教授, 博士生导师, 博士, 主要研究方向有信号处理、神经网络、图像编码、图像恢复与重构、多媒体信息检索、视频语义分析等; 潘晓舟(1975-) , 男, 硕士研究生, 主要从事图像语义描述及图像视觉感知质量评价的研究。