文章编号: 1001- 2486(2003) 05- 0052- 06

开放架构的数字视频管理系统 iView 研究与实现*

王 炜[1], 吕荣聪[2], 武德峰[1], 张 军[1], 汤大权[1], 李志强[1]
(1. 国防科技大学人文与管理学院, 湖南 长沙 410073; 2. 香港中文大学计算机系, 香港 沙田)

摘 要: iView 是我们研制的基于多模态元数据提取, 具有开放架构且支持无线移动存取的数字视频全内容管理系统。本文讨论了系统的需求, 体系框架设计实现, 描述了其中涉及的关键技术思想, 并对存在的问题和研究方向做了简单探讨。

关键词: 视频数据库; 元数据; 多模态; 全内容检索
中图分类号: TP31 文献标识码: A

The Research and Implementation of an Open Digital Video Management System: iView

WANG Wei[1], Michael R. Lyu[2], WU De-feng[1], ZHANG jun[1], TANG Da-quan[1], LI Zhi-qiang[1]
(1. College of Humanities and Management, National Univ. of Defense Technology, Changsha 410073, China;
2. Department of Computer Science and Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong)

Abstract: Our open digital video management system named iView is presented, which supports full content retrieval based on multimodal metadata extraction and fusion, and supports kinds of wireless access mode. Requirements of iView are discussed first, then the design of framework. We also describe the key ideas and technology involved. Finally, its future research trend is pointed out.

Key words: Video database; metadata; multimodal; full content retrieval

随着多媒体和 Internet 应用的普及, 传统信息管理模式已无法满足对海量数字视频进行有效管理的客观需求。近年来基于内容的媒体处理技术为视频有效管理指出了方向, 即基于内容特征提取结构化内容描述信息用于辅助多媒体信息的管理[1,2]。视频包含其它类型媒体, 内容丰富, 在如何有效使用和管理上最具挑战性, 对视频数据库进行有效内容管理的需求也更迫切。虽然视频内部存在丰富的未开发的内容和知识, 但这些未结构化的数据很难管理, 无法直接使用关键字检索, 如果手工注释, 工作量大且具有相当的主观随意性。有效的视频管理需要以某种自动方式提取视频中蕴涵的未开发的内容和知识, 提供一个类似卡片索引目录的工具来完成视频归档并借此寻找所需内容。即视频必须伴随一个结构化内容索引, 通过创建视频内容的丰富索引, 释放视频库中丰富的知识资源, 把视频转换成基于索引对内容进行精细颗粒度存取和控制的容易管理的有用信息[3,4]。传统的视频管理也没有利用到 Internet 及 Web 使用模式。Internet 网络的发展, 尤其是无线移动网络的迅速发展同时也要求能够在任意时间、任意地点检索存取到用户需要的视频信息。而无线视频则以短的、个性化的视频信息片段交换和娱乐内容为其主要特征。

1 系统目标及体系结构

iView(Video over Internet and Wireless)的目标是创建以视频数据为主的分布开放式信息仓储中心, 能够数字化、存储、管理和发布海量的各种格式的数字视频数据内容, 提供综合的公共视频信息服务, 并使得有关用户可以通过不同网络(包括移动无线网络)和不同平台来方便快捷地访问、存取这些内容和服务而不受时空限制。香港文化上的特殊性也要求 iView 系统要能适应英文、普通话及粤语等多语言环境。

* 收稿日期: 2003- 02- 28
基金项目: 香港研究基金委员会资助项目(CUHK4222/01E); ITF 创新与技术基金资助项目(ITS/29/00)
作者简介: 王炜(1973—), 男, 副教授, 博士。

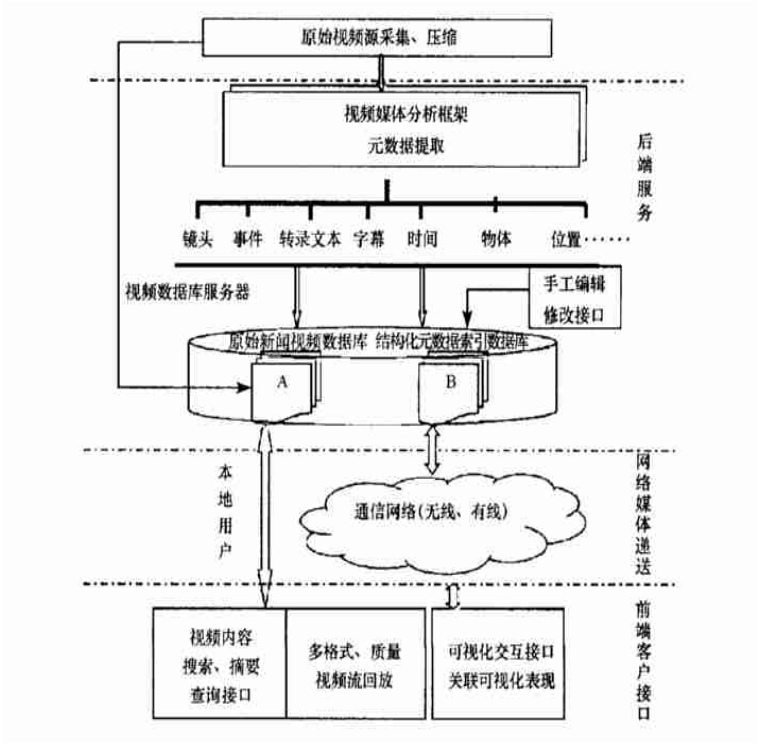

图 1　iView 系统体系结构图
Fig. 1　Architecture of iView

索引的构造是整个系统的核心, 其基础是视频元数据。元数据是用于描述原始视频资源属性和内容且机器可理解的智能化信息, 或者说是关于视频数据资源的特征信息。正是自动/半自动抽取的元数据中独立的众多数据值以及对这些数据值的应用, 使得视频数据成为真正有用的可搜索的信息资源。元数据提取所依赖的算法大部分基于信号分析, 特定元数据面向特定应用, 无法适应普遍应用背景, 因此 iView 主要针对香港 TVB 电视台授权采集的 2000 年以来的新闻节目构成的视频资产库的管理和检索来设计实现。其开放客户/服务器逻辑体系结构如图 1 所示。整个功能框架可分为后端和前端部分。后端涉及数字新闻视频的压缩、采集、归档, 多模态元数据的提取、索引、融合以及人工修改、补充、完善和入库。前端涉及本地的或经由各种不同带宽条件的固定、无线移动网络的视频主题、片段或摘要的检索, 基于多模态元数据融合关联可视化接口以及相应视频媒体的递送组成部分。

视频内容与元数据相分离是系统设计的关键点, 是系统开放性得以保证的基础。这种分离使得 iView 能够集成很大范围内伸缩的数字视频格式和技术。无论最终视频存储模式是在线(磁盘阵列) 、近似在线(磁带遥控设备) 或是脱线(物理磁带) , 也无论视频是 MPEG1, MPEG2, MPEG4 或其它任何操作系统支持的格式, 分离的元数据都能有效定位、跟踪和操纵视频的一个或多个再现。可扩展的模块化体系也是系统设计的开放性考虑之一。因为解决广泛意义上的图像理解是非常困难的问题, 典型的视觉分析一般限制在一个狭窄的问题领域。如何针对不同的视频应用领域构造、集成适用的视频元数据及索引对有效发挥系统的功效至关重要[5, 6]。iView 系统虽然目前主要针对新闻视频管理领域, 但考虑到未来可能被集成到广泛的各种视频应用环境中, 不同的应用领域需要不同类型的元数据分析, 甚至在新闻视频管理领域, 也会随着技术发展不断产生或更新更有效的元数据描述及算法[7, 8], 因此, 后端以系统软总线模式组织, 提供一个开放、可扩展的支持多模态融合的视频媒体分析框架, 方便加入新的特征抽取方法, 以便紧密集成和灵活配置各种现有以及未来可能的元数据分析捕获算法和第三方媒体管理程序, 最终实现针对特定应用的定制。

各种媒体元数据分析插件遵从功能接口标准和数据交换标准, 不仅可以直接处理原始数据(视频帧、音轨) , 也可以存取任何其它元数据分析插件生成的元数据。每种不同的元数据选择被定义为不同的模态。多种模态之间的协作可以完成对原始视频内容更全面的刻画。即成组的插件可以有效合作和

交换信息,为视频索引提供完整框架。此外,符合接口和数据交换标准的元数据手工编辑模块也可以认为是一种人工干预的广义分析插件。

2 多模态元数据及相关处理

iView 系统的核心是开放的视频媒体分析框架,其功能在于高度自动化地建立一套丰富的多模态融合的结构化元数据索引。所有元数据就像解锁图书馆中信息价值的分类卡片一样,充当对原始视频内容的引用,并对其增值。在此基础上完成视频搜索、导航、预览,并迅速定位特定视频节段用于回放,在此过程中并不修改原始视频数据,也不关心原始视频数据物理存储。

作为索引素材的视频元数据大致分为 3 类:

(1) 外部环境中包含的关于视频数据的各种不同形式的客观信息,包括创建时间、长短、格式、时序安排、Closed Caption 文本或 TELE 文本以及其它的关联间接信息等等。

(2) 通过用户编辑接口手工标注的片段标记和手工评注等。

(3) 通过各种音视频信号分析算法自动化抽取的元数据。包括关键帧,字幕文本,语音识别文本,说话人和人脸定位、识别,户内、户外检测,对象(例如主播人头像、人脸、Logo)进入、退出屏幕的检测等。

2.1 视频分割

最终有效的基本检索单位是视频段,依赖于基于内容的视频分割技术。视频分割是执行任何数字视频内容管理的前提环节且已得到充分的研究,本文立足在现有视频分割研究基础上,为提高针对渐变镜头的鲁棒性,同时有效检测突变和渐变镜头,基于颜色、形状边缘模糊统计直方图以及双阈值检测等多种手段综合进行视频有效分割。

2.2 不同模态的元数据集

多模态元数据提取如图 2 所示。篇幅所限,本文不讨论具体算法细节。iView 中接受的元数据模态包括:

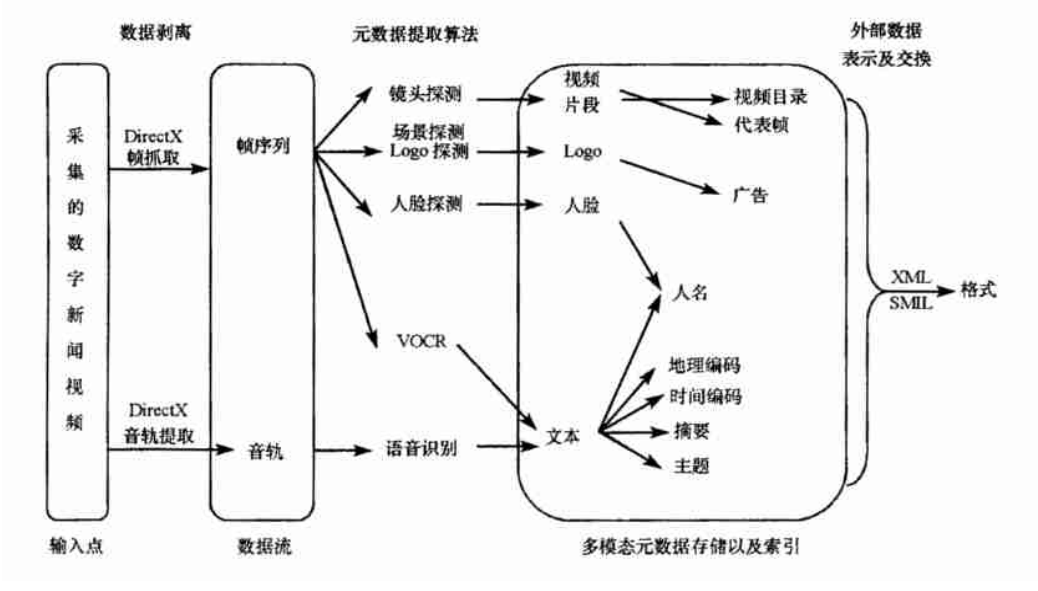

图 2 多模态视频元数据提取示意图

Fig. 2 Multimodal video metadata abstraction

关键帧:从整个镜头序列中挑选出来的从特定测度而言最具代表性的,用于表征整个镜头内容的特殊视频帧。iView 为每个镜头产生一个关键帧,并提取关键帧显著颜色、颜色直方图、形状等数据描述每个节段的不同视觉内容。

Closed Captioned 文本:标准新闻节目中在 CC 文本中包含新闻解说词,提取并剖析后可用于索引目的。欧洲新闻节目中也包括类似的 Teletext 文本。这些文本中都有特殊标记指示说话人的改变或是话

题、故事的改变, 可用于自动化的故事分割。

Logo 标记: 香港与欧洲类似, 在节目中包含电视台 Logo 徽记, 但广告中不包含。通过探测和识别是否出现 Logo 以及类别, 可有效地区分节目和广告视频。Logo 的探测主要基于 Susan 边缘特征的动态局部阈值提取、边缘图像增强和聚类匹配方法实现。

人脸标识: 人脸出现与否是视频内容的一个重要线索。iView 探测和定位视频中出现人脸的时间和位置, 并基于文本和数字坐标描述。人脸的探测算法基于神经网络构造。

字幕文本定位识别(VOCR): 视频帧中出现的文字, 尤其是后期编辑叠加的字幕, 包含了与局部视音频语义密切相关的重要信息, 例如新闻中的主题、日期、人名等。iView 通过灰度变化、亮度边缘检测、动态局部阈值、候选区域增强、由粗到精的定位分割, 并辅之以字幕文本区域预测、长宽比、前背景亮度对比等先验知识来完成字幕文本的自动化探测定位[9]。新闻视频通常用屏幕下三分之一区域显示场所位置、标题、说话者名字, 通过区域预测, 可以优化执行性能, 缩短计算时间。文本识别因为背景复杂和解析度低而无法直接运用传统 OCR 得到良好效果, iView 通过多帧增强、动态阈值二值化、字符分割以及基于 2 维 Gabor 小波特征的模式匹配实现文本识别。

音频分割与分类: 基于内容分割音轨, 找出音频数据流中的所有边界, 将其划分到预定义的类别中, 例如语音、音乐、噪音、静默、室内、户外等, 使得音频中每段时间都赋予一个分类。该分割与视频分割不同, 可能存在相互重叠或覆盖。

语音分割与识别: 当音频信号中包含语音时, 采用两阶段方法检测说话者分界。首先检测语音和非语音边界, 然后定位真正说话者语音阶段, 通过分类判别是英语还是普通话或粤语音频流, 而后通过对 IBM 提供的 Viavoice 语音识别引擎的参数适应性调整, 以实时且与说话人无关的模式将连续语音流转换为对应文本。识别引擎支持多种语言且与领域无关的语音识别。iView 主要针对三种口语, 使用的识别词汇超过 65 000 个, 并且可对新闻节目扩展定制。虽然语音识别的精度依赖于说话人口音、清晰度、语速、周围环境噪音等诸多因素, 但即使是不完美的识别(不到 70%) 仍具有很大参考价值。一般特定单词的内容重要性和出现的频率成反比, 例如名词、专有名词、人名等, 携带搜索所需的大部分信息。因为携带重要内容的单词的识别精度在知识辅助前提下大大高于所有语音词汇的全体识别率, 在执行元数据上关键词类型的搜索时, 语音识别文本的效用仍然很显著。此外我们发现, 在特定视频节段中, 反映主题且检索概率较高的用户感兴趣语音词汇往往在视频局部多次重复出现, 即使引擎不能每次都正确识别, 但只要识别一两个实例, 需要的视频片段也会被迅速定位。通过设计为新闻领域定制的附加关键词增强和过滤机制, 并使用关键字测点定位算法用于提供与说话人无关的关键字音频定位, 可有效提高语音识别元数据的作用。

基于知识的命名实体标识: 基于 VOCR、语音识别以及任何文本为主的多模态元数据中的文本流, 借助自定义领域相关的命名实体词典, 可抽取包含地理(国家、地区、城市)、时间、组织、人物以及其它专有名词等称谓信息, 进一步构成实体标识元数据。结合知识库中的空间坐标、时间关系等知识, 就可以支持可能涉及特定时间、地点(包括附近地点) 以及人物的检索。

节段结构描述: 系列用户定义的视频节段, 每个节段具有指定的标注信息, 包含文本、日期、数字等结构描述。定义的节段描述可以在原始视频中覆盖重叠, 或保持多对一关系。

人工评注文本: 类似 CC 文本的手工文本元数据, 用户可用来记载任意的间接描述信息。

此外, 系统允许任意符合接口标准的自定义元数据模态扩展及对以上元数据的人工编辑修改。

2.3　基于参考时间轴的多模态元数据融合

在上述全方位多模态元数据基础上融合并建立索引, 就可为用户提供全内容视频检索。多模态融合基础的元数据和传统非时基系统中数据类型的区别在于共享的参考时间轴。在所有 iView 元数据表示中, 时间编码是一个共有的关键成分。所有的元数据元素或者打上一个时间戳; 或者跨越一段时间, 由一个进入时间戳和一个退出时间戳表示。iView 的时间模型基于时间编码方法的工业标准 SMPTE, 记为 HH:MM:SS:FF, FF 表示每秒中的视频帧数目。如图 3 所示, 不同模态的元数据元素参考同一时间轴实现关联融合, 形成索引中另一层可以查询的信息。后继搜索和浏览算法可以基于此来推断数据元素

间的同步关系, 并向用户从全方位揭示视频片段内容的视听概览。

3 系统前端

前端分为三个功能接口: 检索, 回放以及支持交互的全内容关联的信息可视化接口。

检索主要通过多语言文本关键字和 QBE 形式实现, 并辅之以基于交互式地图的地理主题检索和基于可视化图形接口的多主题检索。交互式地图检索通过在可缩放矢量地图上圈定范围和选择时间范围来逐步缩小、定位关心的新闻内容; 可视化图形交互检索接口是指将系统聚类产生的若干个主题以图标形式呈现在二维或三维空间中, 用户通过选择并移动主题图标的相对位置向服务器询问感兴趣主题, 从而逐步缩小范围, 找到需要信息。

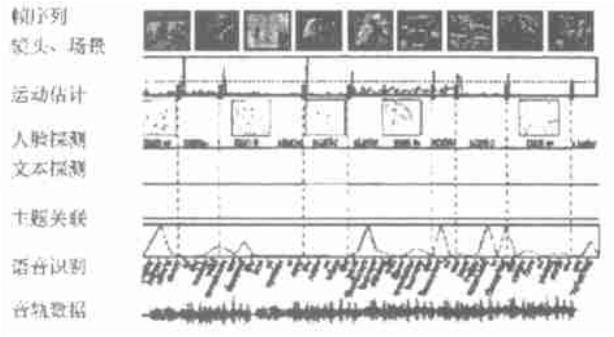

图 3 多模态索引基于时间线的融合示意图
Fig. 3 Multimodal index fusion based on timeline

全内容检索通过服务器将查询转换成为对基于时间关联的多模态元数据索引的查询操作来实现。其结果采用标准的 XML 格式与前台交换。视频媒体分析框架产生的元数据与存储库的交换也采用与平台和应用无关的 XML 标准。从数据结构交换的角度也体现了系统的开放性设计, 不仅通过开放的视频媒体分析框架可以集成各种不同数字视频处理功能, 也可进一步采用支持 XML 的通用框架向各种不同应用平台提供视频片段检索结果并回放。基于 XML 的方案显著提高了元数据的可重用性。元数据索引可以因此和广泛的系统互操作, 从前端 Web 浏览器到后台 DBMS 等等。

为辅助用户迅速把握主体内容, iView 继承并扩展 informedia 系统中若干表现手法[10,11]。所有符合检索条件的视频片段结果以关键帧为代表, 构成图像列表向用户呈现, 伴随每个片段的状态条给出这组视频片段对查询的不同匹配程度。用户可翻页寻找需要的内容并点击选择。被选中的视频展开相应的由若干离散代表帧构成的静态视频故事板摘要并起到导航作用, 辅助用户在进一步搜索和浏览过程中快速定位。因为所有模态元数据自动根据时间信息关联, 在播放原始视频片段的同时, 可以激活包括基于交互地图的地理位置、音轨识别文本、VOCR 主题摘要、关键帧等不同元数据表现窗口, 自动地以时间同步模式按需回放, 即随着时间推移, 以时间同步模式自动高亮显示不同元数据可视化窗口中涉及的关键元素, 以辅助用户全面、迅速理解视频内容。同时, iView 也支持用户在不同窗口中的相互激活、跳转和关联。

最终用户选定的视频将通过调用通用视频播放插件实现流媒体回放, 目前支持 Microsoft Media, Real, QuickTime 等格式。因为并未修改原始视频, 在本地或宽带条件下, 特定视频片段播放结束后用户也可选择连续播放后继视频浏览。

随着各种无线移动网络的迅速普及, 除了通过固定 Internet 网络检索存取视频以外, 以笔记本或 PDA, 移动电话等移动手持设备在任意时间、任意地点检索存取用户需要的视频信息也是重要应用方向。由于无线网的带宽资源和手持设备的计算能力问题, 无线视频以短的、个性化的信息片段交换为主要特征, 尤其是处于 PDA、移动电话等环境时。系统在前端功能全集的基础上裁减, 以适应不同设备能力。变动涉及视频片段质量的调整, 包括降低分辨率以及抽取少量帧来反映关键的视频内容轮廓并按照 SMIL 标准与音轨同步向用户提供信息。尽可能在不影响用户综合理解的前提下, 在有限网络条件和视频内容精细度之间动态平衡。目前在 iPAQ PocketPC 中可通过 802.11 或蓝牙提供用户满意的低分辨率视频片段回放; 在 Nokia 7650 移动电话上可支持连续语音和连环画模式的多代表帧视频轮廓[12] 的递送。不同环境下前台客户应用程序界面如图 4 所示。

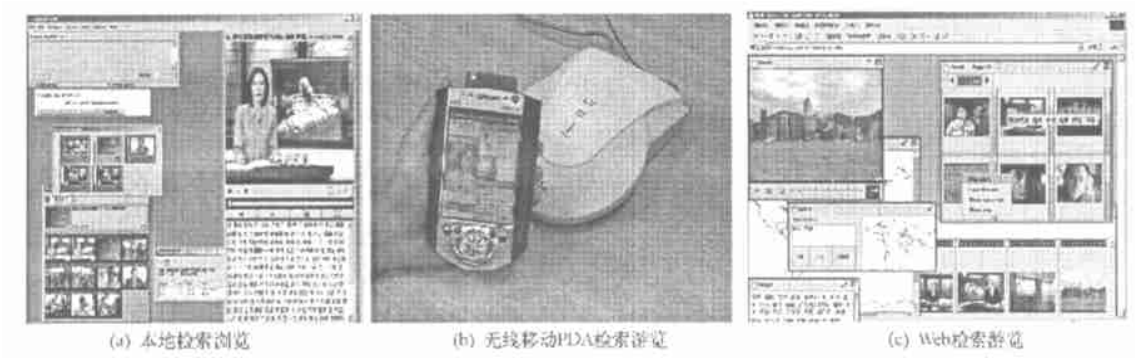

图 4　iView 不同应用示意图
Fig. 4　Different terminals of iView application

4　结论及展望

iView 用于有效管理、分布或动态发布新闻类型的流媒体视频。它将模拟或数字新闻视频通过多模态元数据抽取改造成可充分索引的数据类型,并用于视频归档和友好人机交互模式的全内容检索目的: 首先用视频采集设备完成视频捕捉,然后执行自动视觉场景变化检测,文本字幕检测、人脸检测、logo 等有意义对象的检测,执行自动语音识别及说话人关联等等,抽取的元数据进一步通过手工标注或远程处理增强;而后执行基于时间轴的交叉索引,并管理多个针对不同环境的视频编码程序,最终基于丰富的视频索引为视频发布者和浏览者提供针对视频内容的不同精细颗粒度控制。其开放的体系结构设计支持对数字视频环境中典型设备、处理技术以及应用的集成,并在对 TVB 新闻视频的管理中显示了良好的效果。

基于 XML 交换的元数据不仅是 iView 系统的核心,类似形式的元数据发掘和运用也将迅速成为所有媒体装配和处理过程的核心功能。为了改善不同元数据集间的互操作以及延续性, MPEG7 多媒体内容描述标准针对内容描述的不同方面,为各种各样的多媒体元数据给出了一个统一的描述规范。对 iView 而言,下一步的发展是基于 MPEG7 标准的元数据交换,以进一步提高开放性,实现和相关系统基于标准的协同。

参 考 文 献:

[1] Sutcliffe A, et al. Empirical Studies in Multimedia Information Retrieval[M]. Intelligent Multimedia Information Retrieval, AAAI Press, Menlo Park, CA, 1997.

[2] Vellaikal A, Kuo C C J. Hierarchical Clustering Techniques for Image Database Organization and Summarization[C]. Multimedia Storage and Archiving Systems III, Proc SPIE 3527, 1998: 68- 79.

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

[4] Brunelli R, Mich O, Modena C M. A Survey on the Automatic Indexing of Video Data[J]. Journal of Visual Communication and Image Representation. 1999, 10(2): 78- 112.

[5] Wactlar H D, Christel M G, Gong Y, Hauptmann A G. Lessons Learned from Building a Terabyte Digital Video Library[J]. IEEE Computer, 1999, 42(2): 66- 73.

[6] Wactlar H D, et al. Intelligent Access to Digital Video: the Informedia Project[J]. IEEE Computer 1996, 29(5): 46- 52.

[7] Brown M, Foote J, Jones G, Sparck-Jones K, Young S. Automatic Content-based Retrieval of Broadcast News[C]. ACM Multimedia, 1995, San-Francisco, USA.

[8] Bertini M, Bimbo A D, Pala P. Content-based Indexing and Retrieval of TV News[J]. Pattern Recognition Letters, 2001, 22(5): 503- 516.

[9] Cai M, Song J Q, Lyu M R. A New Approach for Video Text Detection[C]. International Conference On Image Processing, 2002, Rochester, New York, USA.

[10] Wactlar H D. New Directions in VideoInformation Extraction and summarization[C]. The 10th DELOS Workshop, 1999, Sanorini, Greece.

[11] Wactlar H D. Multi-Document Summarization and Visualization in the Informedia digital Video Library[C]. New Information Technology 2001 Conference, 2001, Tsinghua University, Beijing.

[12] 王炜等. 针对无线移动环境的音频同步视频连环画的自动生成[J]. 国防科技大学学报, 2003, 25(3).