余一清　汪宏斌　周洞汝
(武汉大学计算机学院　430072)

基于内容的视频检索研究

Research of Content-based Video Retrieval

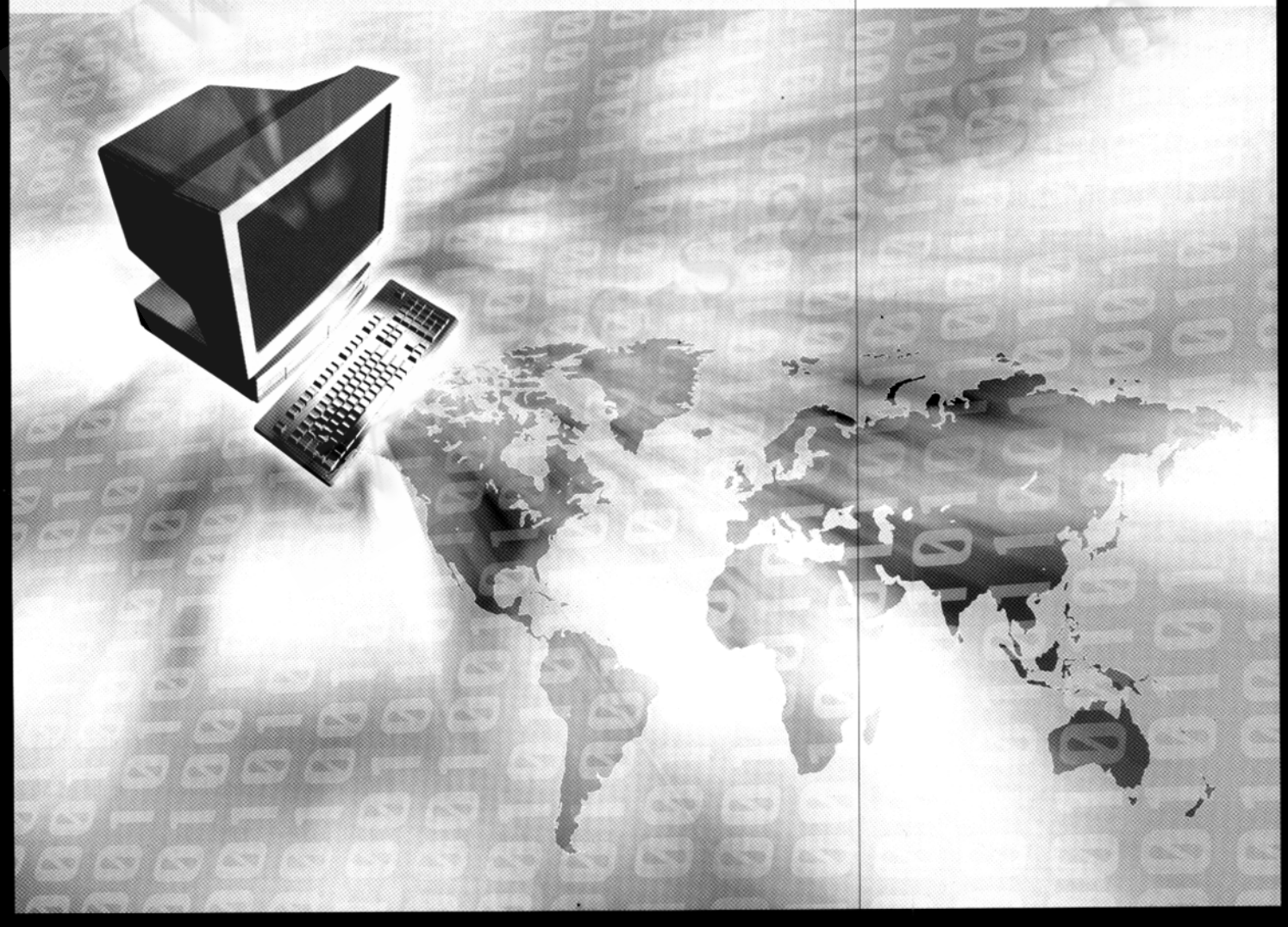
摘要：基于内容的检索能使用户根据媒体特征对媒体内容进行检索和查询。由于多媒体数据中含有丰富的视频数据，并且是随时间动态变化的，其特征很难用一般的静态特征来描述，为了取得视频数据的特征，对视频数据的处理非常重要。本文将介绍基于内容的视频检索中相似索引的处理技术和方法。

关键词：基于内容的索引　视频　相似性

1 引言

现在，多媒体数据已经广泛的应用于Internet、企业和事业信息系统中。不同于常规数据，多媒体数据本身包含大量的视听信息线索，用常规的关键词和属性方法难以对多媒体的内容进行有效的描述和检索，因此需要新的多媒体信息检索方法。基于内容的（Content-based）多媒体信息检索就是一种有效的多媒体信息检索方法。

视频图像作为多媒体信息里一种视觉媒体，与静态图像有着非常大的区别，最大的区别就是时间特性，由此使得视频具有独特的时间逻辑结构和运动特性。基于内容的视频信息检索即使根据用户的视觉、运动或语义属性方面的查询要求，从结构化视频中，搜索出与用户查询要求相似的所有帧、镜头、场景或视频节目。基于内容



的视频浏览是一种随机存取的非线性浏览方式。通过对视频的基于内容的抽象，包括机构分析，获得视频的各种形式的逻辑结构、总结和概要，从而支持视频的基于内容的检索。

2 视频检索的主要特征

为了对视频进行分类、索引和检索，我们先对视频的主要特征进行描述。

2.1 视频序列的组成

视频数据可用幕、场景、镜头、帧等描述。帧是一幅静态的图象，是组成视频的最小单位；镜头是又一序列帧组成的一段视频，它描绘同一场景，表示的是一个摄像机动作，一个事件或连续的动作；场景包含有多个镜头，针对同一批对象，但拍摄的角度不同、表达的含义不同；幕是由一序列相关的镜头组成的一段视频，包含一个完整的事件或故事情节，任何视频都是由一个个镜头衔接起来的，因此镜头是视频检索的基本单元，往下就是镜头中对象的运动或图象帧，往上是场景。为了检索需要，必须将视频分成一幕幕或一个个场景以及一个个镜头，因此对视频中镜头划分是视频处理中最基本的内容。

2.2 视频的单元特征

视频单元可以是镜头或场景，我们以镜头作为研究对象，对于低层特征来说，从时间和空间二个方面考虑，可以从一个镜头中提取三种类型的特征：

(1) 关键帧特征。关键帧特征是从空间里提取的特征，使用的是与图像检索中相同的特征，即从镜头的关键帧中提取的主要是颜色、纹理和形状等。除底层特征外，也可以包括相应的高层语义和属性。

它是视频镜头的基本和独特的性质。

(2) 关键对象特征。通常情况下，希望对特定的目标进行查询，这就涉及到视频中的对象和子区域的特征问题，但关键帧特征要有提供足够的信息用于基于视频对象的查询，因为视频对象不是静态的，他们有时基特性。关键对象的属性可以是运动特征、形状、生命周期、运动对象的轨迹，以及对象颜色、纹理等图像属性，甚至是语义，关键对象的生命周期是对象在镜头中出现到对象消失之间的那段时间。

(3) 运动特征。它有面向对象的运动特征和一般性的运动特征：面向镜头的运动特征。运动特征可以提供有关镜头操作、行为、运动复杂性和分布方面的信息，这些信息在查询中有很大的价值。

3 视频的检索

图像检索是比较两幅图像之间的相似性，而对于视频检索来说，涉及到时间和空间两个维度的比较，即如何结合多特征来定义两个镜头序列的内容相似性，这是一个关键的问题，困难在于可能要涉及到多种特征，不同的权重，内容的比较可以基于关键帧特征、镜头的运动特征、对象特征，或以上特征的组合。本文将侧重介绍相似索引中的语义索引技术，语义检索就是按主题检索，一般通过对视频数据的分析，利用关键字表示视频的语义内容，例如基于文本的注释。

3.1 关键帧相似性计算

我们用关键帧的内容特征来进行内容比较，先定义基于两个关键帧集合之间的相似性来定义镜头的相似性。如果两个相比较的镜头用和表示，而它们的关键帧集合分别用 Ki= {fim, m=1,…,M},Kj= {fjm,n=1,…, N} 表示，那么两个镜头的相似性可以定义为：

sim(STi,STj)=max [S(f,fj1),S(fi1,fi2),…, S(fi1,fiN),…S(fiM,fj2),…, S(fiM,fjN)]

这里S是两个图像之间的相似性度量，可以用任何一种或多种图像特征的组合进行度量，按上述公式，两个镜头之间存在MXN个相似性度量值，而选择其中的最大值作为镜头的相似性度量，定义是假设两个镜头的相似性可以由一对最相似的关键帧来确定，或如果在两个镜头中有一对相似的关键帧，那么它们就认为是相似的。

3.2 基于内容检索的表达机制

基于内容检索是基于相似度的检索，即给定检索要求，在视频数据库中查找有关镜头和场景等媒体对象。由于在实际应用中，被检索的媒体对象库一般都是巨容库(容量都在几十万以致上百万以上)，因此不可能采用对库中对象逐个计算相似度函数再进行比较的方法，而需要快速索引手段的支持。由于媒体对象的"内容"是通过特征来体现和构成的，特征空间中的接近意味着内容的相似，因此实现中一种较为有效的技术是基于特征空间中k-最近邻搜索（k-Nearest Neighbor Search）的相似索引（Similarity Indexing）技术。在这种方法中，通过在特征空间建立一种距离度量，而把任何两点间的相似度与其间的距离对应起来，使得相似度与距离成单调递减关系。根据这个距离度量，可对库中所有对象的特征值建立近邻索引结构。利用此索引结构，给定检索特征，可在很小的查找范围快速得到内容特征与检索特征距离最近的（按照相似检索的定义也就是最相似的）一组对象，从而大大加快检索的速度。

在实际应用中，检索要求的形式不仅限于此，许多应用中与检索要求最相似的媒体对象可能并非仅存在于某一特征值处。

从广义的角度来看，我们可认为每个对象的内容都被其内容描述所充分表示，即具有同样内容描述特征的对象内容相同。这样，每个检索请求就都可对应于内容描述空间上的一个相似度函数分布，内容描述空间上任一点的相似度函数值表明了内容描述为该点坐标值的对象对此检索请求的满意程度——即相似度。这样，单特征最近邻搜索对应的相似度函数在对应特征空间是以被检索特征值为中心的以距离为变元的球对称单

峰单调递减函数，在整个特征空间为一“岭”状函数。Virage系统的多特征检索对应的相似度函数在相关特征空间是以被检索特征值为中心的（椭球对称）单峰函数，在整个特征空间也为一“岭”状函数。而在以相似索引中，这样的函数显然难以用一个单峰函数来近似。

通过上述分析可知，检索表达机制应支持具有任意形状相似度分布函数的检索请求。

3.3 相似检索技术

根据特征空间中距离近的对象内容相似的思想，在特征空间建立一种距离度量，并将任何两点间的相似度与其间距离对应起来。据此距离度量，可对库中所有对象根据特征值建立近邻索引结构。利用索引结构，给定检索特征，可只在很小的查找范围查找，从而快速得到与检索特征距离最近的一组对象（即最近邻）。

相似度与距离的对应关系可采用多种函数形式来描述，一般说来相似度作为距离的函数sf(d)只要满足以下条件即可：

· sf (d) 在距离为零(d=0)处取值为1，即sf (0)=1。

· sf(d)是距离的单调递减函数。

· sf(d) 在无穷远处取值为零，即。

常采用的函数有指数函数$sf(d)e^{-d}$、函数$sf(d)=2/(1+e^{d})$等，也有一些系统采用分段线性函数（设定足够大的最大距离d_{max}，sf(d)在$[0,d_{max}]$区间从1线性下降至0，在$[d_{max},+\infty]$区间上取值为0）。在本论文中，我们将采用$sf(d)=e^{-d}$形式的指数型相似度 - 距离对应函数。

从上面的分析可知，对于这种基于距离索引的相似索引，其一次查找的相似性分布函数在对应特征空间中为一以检索特征为球心的球对称单峰函数，峰值在检索特征处，取值为1，无穷远处取值为零，相似性函数值随着离球心距离的增大（以指数形式）单调下降。在一维情况下，对于数据的查找可用折半查找来进行，在高维空间中，这种算法不再有效，但其思想却是类似的。目前，所有相似索引/最近邻搜索的基本思想都是根据特征数据集在特征空间中的分布特性，将数据集切分为子数据集，并对每个子数据集建立描述。这样，检索时通过对各子数据集描述的比较，可以抛弃不满足条件的子数据集而仅对合格者进行检索和计算，从而大大降低搜索量。

根据这个思想，相似索引结构一般都由下面两部分组成：

（1）数据集描述：用来对描述数据集的各种特性，通常包括如下描述：

· 对象个数：用于说明当前数据集(及其子集)中含有的对象个数。

· 包络(envelop) 包络为高维空间中的一个封闭曲面，用于描述数据集的空间位置与分布，使得数据集中的所有点都位于此包络的内部。

· 代表对象标识 指定一个可代表此数据集的对象的标识。

（2）子集指针：指向子数据集结点的指针(对应于索引树的内部结点)或指向对象及其描述的指针（对应于索引树的叶结点）。

采用不同的特征空间距离定义、包络类型和数据集切分方法，可以得到不同类型的索引结构，但其索引结构的组成和查找算法的思想却是类似的。

3.4 相似索引结构的生成算法

输入：数据集，数据集切分准则与方法。

输出：索引结构（索引树）。

算法如下：

Step1 对数据集生成索引结构并建立包络描述（如对象个数、包络和代表对象标识）。

Step2 判断数据集是否需要进一步切分。

Step3 若数据集需要进一步切分，则用指定切分方法将数据集切分为子数据集，对每个子数据集调用此过程并将返回指针填入索引结构的子集指针中，返回索引结构。

Step4 若不需进一步切分，则填入对象及其描述，生成叶节点。

4 小结

多媒体信息检索的主要工作是设计快速的检索方法，对于多媒体数据库（数据集合）进行搜集，精确或近似地匹配查询的对象。这些对象可以是二维图像、医学图像、一维时间序列、数字语音或音乐、视频等。在基于内容的多媒体检索研究中，许多方法是通过自动提取的特征来检索的，如图像的颜色和纹理。

本文介绍了一种视频索引技术，在未来的工作研究中，将关注其他类型的媒体检索。因为我们生活的信息环境是全方位的，多媒体信息还包含典型的音频媒体以及图形、动画等媒体。随着信息化进程的深入，这些媒体数据类型将会越来越多，就会面临检索问题，即需要对数字音频、语音和音乐进行基于内容的检索，对合成媒体，如动画、VRML数据进行检索等。在研究这些多媒体检索时，将更注重多媒体的互相关联和互补关系，以提高检索算法的效率。