一种通过视频片段进行视频检索的方法*

彭宇新[1,2+], Ngo Chong-Wah[3], 董庆杰[1,2], 郭宗明[1,2], 肖建国[1,2]

[1](北京大学 计算机科学技术研究所,北京 100871)

[2](北京大学 文字信息处理技术国家重点实验室,北京 100871)

[3](香港城市大学 计算机科学系,香港)

An Approach for Video Retrieval by Video Clip

PENG Yu-Xin[1,2+], NGO Chong-Wah[3], DONG Qing-Jie[1,2], GUO Zong-Ming[1,2], XIAO Jian-Guo[1,2]

[1](Institute of Computer Science and Technology, Peking University, Beijing 100871, China)

[2](National Key Laboratory of Text Processing Technology, Peking University, Beijing 100871, China)

[3](Department of Computer Science, City University of Hong Kong, Hong Kong, China)

+ Corresponding author: Phn: 86-10-62752426, Fax: 86-10-62981438, E-mail: peng_yuxin@icst.pku.edu.cn

http://www.icst.pku.edu.cn

Received 2002-11-25; Accepted 2003-03-20

Peng YX, Ngo CW, Dong QJ, Guo ZM, Xiao JG. An approach for video retrieval by video clip. *Journal of Software*, 2003,14(8):1409~1417.

http://www.jos.org.cn/1000-9825/14/1409.htm

Abstract: Video clip retrieval plays a critical role in the content-based video retrieval. Two major concerns in this issue are: (1) automatic segmentation and retrieval of similar video clips from video database; (2) similarity ranking of similar video clips. In this paper, motivated by the maximal matching and optimal matching in graph theory, a novel approach is proposed for video clip retrieval based on matching theory. To tackle the clip segmentation and retrieval, the retrieval process is divided into two phases: shot-based retrieval and clip-based retrieval. In shot-based retrieval, a shot is temporally partitioned into several sub-shots based on motion content. The similarity among shots is measured according to the color content of sub-shots. In clip-based retrieval, candidates of similar video clips are selected by modeling the continuity of similar shots. Maximal matching based on Hungarian algorithm is then adopted to obtain the final similar video clips. To rank the similarity of the selected video clips, four different factors: visual similarity, granularity, interference and temporal order of shots are taken into consideration. These factors are modeled by optimal matching based on Kuhn-Munkres algorithm and dynamic programming. Experimental results indicate that the proposed approach is effective and efficient in retrieving and ranking similar video clips.

Key words: content-based video retrieval; clip; similarity; maximal matching; optimal matching

摘 要: 视频片段检索是基于内容的视频检索的主要方式,它需要解决两个问题:(1) 从视频库里自动分割出与查询片段相似的多个片段;(2) 按照相似度从高到低排列这些相似片段.首次尝试运用图论的匹配理论来解决这两个

* **第一作者简介**: 彭宇新(1974－),男,贵州都匀人,博士生,主要研究领域为基于内容的视频检索.

问题.针对问题(1),把检索过程分为两个阶段:镜头检索和片段检索.在镜头检索阶段,利用相机运动信息,一个变化较大的镜头被划分为几个内容一致的子镜头,两个镜头的相似性通过对应子镜头的相似性计算得到;在片段检索阶段,通过考察相似镜头的连续性初步得到一个个相似片段,再运用最大匹配的 Hungarian 算法来确定真正的相似片段.针对问题(2),考虑了片段相似性判断的视觉、粒度、顺序和干扰因子,提出用最优匹配的 Kuhn-Munkres 算法和动态规划算法相结合,来解决片段相似度的度量问题.实验对比结果表明,所提出的方法在片段检索中可以取得更高的检索精度和更快的检索速度.

关键词: 基于内容的视频检索;片段;相似度;最大匹配;最优匹配

中图法分类号: TP391 **文献标识码**: A

随着电视台视频节目的积累,网上数字视频的增加,以及数字图书馆、视频点播、远程教学等大量多媒体的应用,如何在海量视频中快速检索出所需要的资料显得至关重要.传统的基于关键词描述的视频检索因为描述能力有限、主观性强、手工标注等原因,已经不能满足海量视频检索的需求.因此,从 20 世纪 90 年代开始,基于内容的视频分析和检索技术成为研究的热点问题.由于基于内容的图像检索的困难性和复杂性,大量的研究主要集中在视频内容的结构分析上,如镜头的分割、关键帧的提取、场景的构造等,视频检索方面的研究则相对较少,而这部分常常是应用的关键.视频检索一般分为镜头检索和片段检索.镜头一般是由摄像机一次摄像的开始和结束的所有帧构成,表示一个物理概念.而片段是由一连串语义相关的连续镜头构成,表示的是一个语义概念.目前视频检索的多数研究集中在镜头检索上[1~4],而片段检索方面的研究则刚刚开始[5~11].实际上,从用户的角度分析,他们对视频数据库的查询通常会是一个视频片段而很少会是单个的物理镜头.从信息量的角度分析,由几个镜头组成的视频片段比单个镜头有更多的语义,它可以表示用户感兴趣的事件,因此,查询的结果也比较有意义.例如,从新闻中检索出感兴趣的事件,从体育节目中检索出喜爱的体育运动,电视台检索某条广告是否播出等.基于这种考虑,本文提出了一种通过视频片段进行视频检索的方法,以满足用户通过视频片段来提交的查询需求.

视频片段检索需要解决两个问题:(1) 从视频库里自动分割出与查询片段相似的多个片段;(2) 按照相似度从高到低排列这些相似片段.目前已有的片段检索方法可以分为两类:(1) 把视频片段分为片段-帧两层考虑,片段的相似性利用组成它的帧的相似性来直接度量[5~7];(2) 把视频片段分为片段-镜头-帧三层考虑,片段的相似性通过组成它的镜头的相似性来度量,而镜头的相似性通过它的一个关键帧[8~10]或所有帧[11]的相似性来度量.方法(1)的缺点在于,限制相似的片段必须遵守同样的时间顺序,而实际的视频节目并不遵守这种约束,因为后期编辑的结果使得相似的片段完全可能具有不同的镜头顺序,如同一个广告的不同编辑.同时,这种基于每帧的比较,也使得检索速度比较慢.方法(2)的思想比较合理,但这种方法从已有的文献上看并没有很好解决片段检索的问题.文献[8~10]提出了影响视频相似度度量的顺序因子、速度因子、粒度因子、干扰因子,但它的片段是预先分割好的,并没有解决怎样在连续的视频节目里自动分割出多个相似片段的问题.与文献[8~10]相反,文献[11]完全忽略了镜头顺序、粒度、干扰因子的影响,两个片段的相似度仅仅取决于它们相似镜头的数量,因此,即使片段 Y 的所有镜头仅仅和片段 X 的一个镜头相似,Y 也会被认为与 X 相似;另外,镜头的相似性是根据两个镜头相似的最长帧序列来判断,这种基于每帧的比较和文献[5~7]类似,片段的检索速度也较慢.

针对上述问题,本文提出解决片段检索两个问题的一个新方法.为了分割出相似片段,本文采用了上述方法(2)的思想,把检索过程分为镜头检索和片段检索两个阶段:在镜头检索阶段,考虑了视频中的时间信息,把一个镜头内部随时间变化的内容,分解为几个内容一致的子镜头(sub-shots),这种基于子镜头的比较全面地反映了两个镜头是否相似;在片段检索阶段,通过考察相似镜头的连续性初步得到一个个相似片段,再运用最大匹配的 Hungarian 算法来确定真正的相似片段.为了排列相似片段,类似于文献[8~10],本文考虑了片段相似度度量的不同因子,不同于文献[8~10],提出用最优匹配的 Kuhn-Munkres 算法和动态规划算法相结合来度量这些因子的影响.本文首次尝试运用图论的匹配理论来解决视频检索问题,这是因为匹配的思想要求相似镜头必须一一对应(粒度),在这个条件下,求出的最大匹配和最优匹配可以客观而全面地反映两个片段相似的镜头数量和两个片段视觉相似的程度,从而避免了文献[11]中镜头计算的粒度问题.第 4 节的实验结果表明,与具有同样功能的文

献[11]相比,无论是检索的准确性,还是检索速度,本文提出的方法都取得了更好的效果.

本文第 1 节首先介绍了本文的理论基础——图论的最大匹配和最优匹配.第 2 节介绍怎样自动分割相似片段.第 3 节介绍视频片段的相似度模型.第 4 节给出了实验结果.第 5 节是总结.

1 最大匹配和最优匹配简介

最大匹配和最优匹配是图论中的两个经典问题,其中,最大匹配解决的一个典型应用问题如下:假设有 n 个工作人员 $x_1,x_2,\ldots,x_n$ 安排做 m 项工作任务 $y_1,y_2,\ldots,y_m$,如图 1 所示,其中边 $e_{ij}=(x_i,y_j)$表示 x_i 可以从事 y_j,如果每个人最多从事其中一项工作,且每项工作只能由一人承担.问:怎样才能让尽可能多的人安排上工作?

在上面的工作安排问题中,只着眼于每一工作人员都能安排一份工作,并没有考虑如何更好地发挥工作人员的专长问题.如图 2 所示,最优匹配解决的典型问题是:假设每个工作人员 x_i 从事 y_j 工作时的效率为 ω_{ij},问:怎样合理安排才能使总的工作效率最高?

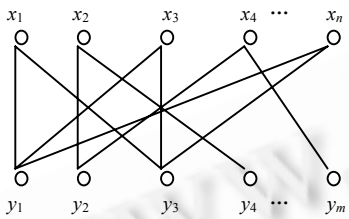

Fig.1 Maximal matching

图 1 最大匹配

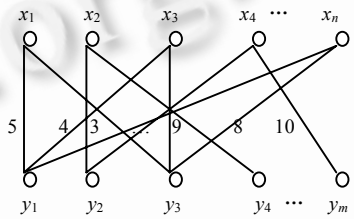

Fig.2 Optimal matching

图 2 最优匹配

考虑到最大匹配和最优匹配的不同性质,并且最大匹配的计算复杂度低于最优匹配,在第 2 节,本文首先运用最大匹配来得到相似片段,然后在第 3 节,运用最优匹配来计算两个相似片段的视觉因子.

2 自动分割相似片段

2.1 镜头检索

因为本文以镜头的相似性为基础来讨论片段的相似性,所以,无论是视频库,还是查询片段,首要任务都是先把它们分割为镜头.目前,很多镜头边界检测的算法被提了出来,本文使用了 Ngo 的时空切片算法(spatio-temporal slice)[12],因为它有这样一些优点:(1) 不仅能很好地检测出镜头突变(cut),还能检测出镜头缓变(gradual transition),如划像(wipe)、淡入淡出(dissolve);(2) 直接在 MPEG 流上分析 DC 图像实现,具有很快的检测速度.因为一个镜头是时间上的连续序列,它的内部经常会有内容上的变化,由于一个镜头内部的内容变化主要是由相机运动引起的,本文使用了 Ngo 的方法[13],根据相机的运动信息,把一个内容变化的镜头划分为几个内容一致的子镜头(sub-shots).如果一个镜头是静止(static)后变焦(zoom),那么该镜头就分为两个子镜头;如果是静止、扫描(pan)、静止,那么就分为 3 个子镜头.然后针对不同运动的子镜头,构造不同的关键帧表示,如静止子镜头可以用一个关键帧表示,扫描子镜头通过构造一个全景图(panoramic image)来表示,变焦子镜头可以用变焦之前和之后的两个关键帧来表示,这种基于运动检测的自适应关键帧构造,可以全面表示镜头的内容变化.

在此基础上,用关键帧 $\{r_{i1},r_{i2},\ldots,r_{ik}\}$ 表示镜头 s_i,把镜头 s_i 和 s_j 的相似度定义为

$$Sim(s_i,s_j)=\frac{1}{2}\left\{M(s_i,s_j)+\hat{M}(s_i,s_j)\right\},$$

其中,

$$M(s_i,s_j)=\max_{p=\{1,2,\ldots\}}\max_{q=\{1,2,\ldots\}}\left\{Inter\,sect(r_{ip},r_{jq})\right\},$$

$$\hat{M}(s_i,s_j)=\hat{\max}_{p=\{1,2,\ldots\}}\max_{q=\{1,2,\ldots\}}\left\{Inter\,sect(r_{ip},r_{jq})\right\},$$

$$Inter\sec t(r_i,r_j)=\frac{1}{A(r_i,r_j)}\sum_h\sum_s\sum_v\min\left\{H_i(h,s,v),H_j(h,s,v)\right\},$$

$$A(r_i,r_j)=\min\left\{\sum_h\sum_s\sum_v H_i(h,s,v),\sum_h\sum_s\sum_v H_j(h,s,v)\right\}.$$

$\hat{\max}$ 表示第二大的值.使用 $\hat{\max}$,是因为本文使用 HSV 颜色直方图来表示关键帧,它的缺点是如果两个关键帧有相似的颜色分布,即使它们的内容不一样,也会认为这两个关键帧相似,为了克服这种缺陷,使用 M 和 $\hat{M}$ 的平均值来加强算法的鲁棒性. $H_i(h,s,v)$ 是 HSV 颜色空间的直方图,本文用 H,S,V 分量在 $18\times3\times3$ 的三维空间中统计直方图,以归一化后的 162 个数值作为颜色特征值. $Inter\sec t(r_i,r_j)$ 表示两个直方图的交,本文用它来判断两个关键帧的相似性.

在计算出查询镜头 s_i 和每个镜头 s_j 的相似值之后,需要设定一个阈值 T_s 来决定哪些 s_j 与 s_i 相似,因为本文的目标是片段检索,所以希望能够保证相似镜头高的查全率(recall).根据实验结果, $T_s=0.5$ 能够取得较高的查全率.对这个低阈值带来的不相似镜头,在第 2.3 节利用最大匹配的长度约束(以镜头数表示)可以有效去掉,这是因为匹配要求这些误判的不相似镜头必须和查询片段的镜头一一对应相似,因此它们连成一个片段的概率非常小.与现有算法主要依赖于镜头检索的准确结果相比[8~11],本文提出的最大匹配的长度约束方法能够容忍镜头检索产生的适度错误.

2.2 初步得到可能的相似片段

对视频库 Y 而言,与查询片段 X 相似的镜头是少数,大量的镜头并不相似,因此首先将 Y 中与 X 相似的镜头 y_j 从小到大排序,然后考察这些 y_j 的连续性,如果 $|y_{j+1}-y_j|>2, j=1,2,\ldots,\lambda-1$,$\lambda$是查找的视频库 Y 的长度(以镜头数表示),则得到一个可能的相似片段 $Y_k=\{y_{i_2},y_{i+1},\ldots,y_j\}, i,j\in[1,\lambda]$.取 $|y_{j+1}-y_j|>2$,是考虑算法的鲁棒性,因为:(1) 后期编辑会插入无关镜头,如同一个广告的编辑,长广告会在短广告的基础上插入少量不相似的镜头;(2) 如果开始一个新片段,它们之间会有一段时间的间隔,这种间隔一般大于 2 个镜头.

2.3 最大匹配确认真正的相似片段

假设查询片段 $X=\{x_1,x_2,\ldots,x_n\}$,每个可能的相似片段 $Y_k=\{y_1,y_2,\ldots,y_m\}$,其中 x_j,y_j 表示镜头,那么,X 与 Y_k 的相似镜头对应图可以表示为图论中的二分图 $G_k=\{X,Y_k,E_k\}$,其中,顶点集 $V_k=X\cup Y_k$,边集 $E_k=\{e_{ij}\}$,e_{ij} 表示 x_i 与 y_j 相似.

经过第 2.2 节判断的可能相似片段,包含了不相似片段和真正的相似片段.通过大量的实验观察,可以归纳为如图 3~图 5 所示的 3 种典型情况,其中图 3 和图 4 是不相似片段的二分图,图 5 是相似片段的二分图.由于视频片段是由表示同一个语义的连续镜头组成,因此一个视频片段的内部镜头本身就会相似,本文称这个性质为视频片段的自相似性,由于这种自相似性的存在, X 和 Y_k 的二分图会出现普遍的一对多、多对一、多对多的情况.判断两个片段是否相似,可以从它们相似镜头的数量来判断.经过第 2.2 节的判断,可以知道,基本上每个 y_j 在 X 中都能找到相似镜头 x_i,但因为多对多相似的存在,未必每个 x_i 在 Y_k 中都能找到相似镜头 y_j.因此,考察 x_i 的相似情况,因为 Y_k 的长度可能会小于 X 的长度,考虑算法的鲁棒性,如果 X 中有一半镜头在 Y_k 中能找到相似镜头,就认为 Y_k 和 X 相似的镜头足够多,因此 Y_k 是 X 的相似片段.这个方法可以有效地辨别图 3 的情况.但在图 4 和图 5 中,查询片段 $X=\{x_1,x_2,\ldots,x_8\}$都有 6 个镜头找到相似镜头,如果用上述方法,它们都被判断为相似片段,但图 4 却是不相似片段的典型情况.

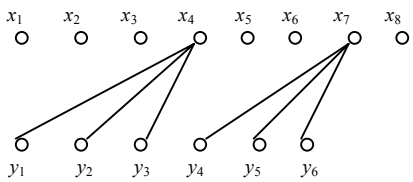

Fig.3 Two dissimilar video clips
图 3 两个不相似的视频片段

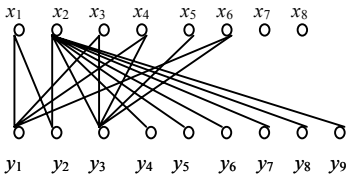

Fig.4 Two dissimilar video clips

图 4 两个不相似的视频片段

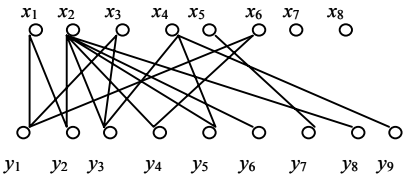

Fig.5 Two similar video clips

图 5 两个相似的视频片段

因此,进一步观察在 Y_k 和 X 一一对应而不是重复对应的情况下它们的相似情况.对图 4、图 5 使用求最大匹配的 Hungarian 算法,得到如图 6、图 7 所示情况,如果 $|M| \geq \left\lceil \frac{n}{2} \right\rceil$ (匹配 $M \subseteq E_k$,并且 M 中任意两条边都不相邻, n 是查询片段 X 的镜头数),就认为 Y_k 与 X 相似的镜头数足够多,因此它是真正的相似片段.

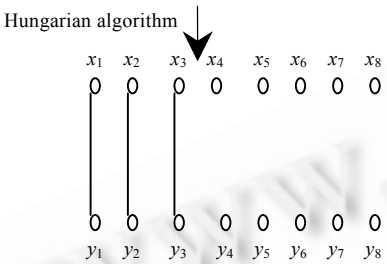

Fig.6 Result of Fig.4 after maximal matching

图 6 图 4 最大匹配后的结果

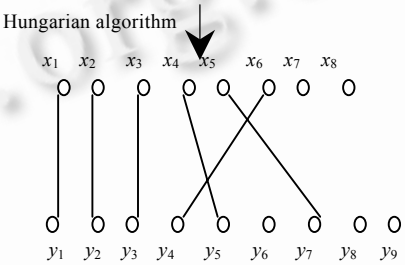

Fig.7 Result of Fig.5 after maximal matching

图 7 图 5 最大匹配后的结果

这样,从图 6、图 7 可以清楚地区分出不相似片段和相似片段,第 4 节的实验结果证明了这个方法的有效性.具体的 Hungarian 算法[14]如下(输入为二分图 $G_k=\{X,Y_k,E_k\}$;结点标记 0 表示尚未搜索,1 表示是饱和点,2 表示是无法扩大匹配的结点):

(1) 任给一初始匹配 M,给饱和点“1”标记;

(2) 判断 X 中的各结点是否都已有非零标记?

(2.1) 是, M 是最大匹配,结束;

(2.2) 否,找一“0”标记点 $x_0 \in X$,令 $A \leftarrow \{x_0\}, B \leftarrow \varnothing$, A , B 是两个集合;

(3) 判断集合 A 的邻接点集 $N(A)=B$ ($N(A) \subseteq Y_k$,是与 A 中结点邻接的结点集合)?

(3.1) 是, x_0 无法扩大匹配,给 x_0 标记“2”,转(2);

(3.2) 否,在 $N(A)-B$ 中找一点 y_i ,判断 y_i 是否标“1”?

(3.2.1) 是,则有边 $(y_i,z) \in M$.令 $A \leftarrow A \cup \{z\}, B \leftarrow B \cup \{y_i\}$,转(3);

(3.2.2) 否,存在从 x_0 至 y_i 的可增广路 P ,令 $M \leftarrow M \oplus P$ (M 与 P 进行环和),给 x_0 , y_i 标记“1”,转(2).

Hungarian 算法的时间复杂度是 $O(ne)$, n 是二分图 $G_k=\{X,Y_k,E_k\}$ 中 X 的结点数, e 是边数 $|E_k|$.

3 视频片段的相似度模型

经过第 2 节的计算,已经得到与查询片段视觉上相似的多个片段,接下来考虑按照相似度从高到低排列它们.类似于文献[8~10],本文考虑了片段相似度度量的下列因子:

(1) 视觉的相似度:两个片段在低层视觉特征(如颜色)上的相似程度;

(2) 粒度的相似度:由图 3~图 5 可知,相似片段的镜头一般会一对多、多对一、多对多地相似,这种情况会影响最终的相似度度量.例如与 X 多对一相似的 Y_k 应该被赋予较小的相似度;

(3) 时间顺序的相似度:与查询片段 X 颜色相似的多个片段,它们与 X 的时间顺序可能不一致.在这种情况下,与 X 时间顺序相似的片段应该被赋予更高的相似度;

(4) 干扰因子:查询片段 X 与相似片段 Y_k 中的一些镜头,可能不能找到它们对应的相似镜头,这种情况会影

响最终的相似度度量.

不同于文献[8~10],本文是基于图论的最优匹配来表示和建模上述的相似度模型,这样做的一个显著优点是,本文方法的有效性能够通过最优匹配来验证.另外,因为视觉是相似片段最重要的判断标准,本文不是像文献[8~10]那样采用上述因子的线形组合来判断两个片段是否相似,而是先利用最大匹配得到视觉上的相似片段后,再基于最优匹配来表示和建模相似度模型,这样视觉上相似的片段不会因为其他因子的影响而漏掉.另外,因为最大匹配的计算复杂度低于最优匹配,这样做也可以加快检索的速度.最优匹配和最大匹配一样,都是在粒度的前提下进行计算,下面具体计算其他 3 个因子.

3.1 最优匹配计算视觉因子

把每对相似镜头的相似值作为权值赋给 $G_k=\{X,Y_k,E_k\}$的每条边,这时的 G_k 就转化为一个带权的二分图,具体计算最优匹配的 Kuhn-Munkres 算法[15]如下:

(1) 给出初始标号 $l(x_i)=\max_j \omega_{ij}, l(y_j)=0, i,j=1,2,\ldots,t, t=\max(n,m)$;

(2) 求出边集 $E_l=\{(x_i,y_j)\mid l(x_i)+l(y_j)=\omega_{ij}\}$, $G_l=(X,Y_k,E_l)$ 及 G_l 中的一个匹配 M;

(3) 若 M 已饱和 X 的所有结点,则 M 即是 G 的最优匹配,计算结束,否则进行下一步;

(4) 在 X 中找 M 非饱和点 x_0,令 $A\leftarrow\{x_0\}, B\leftarrow\varnothing$, A,B 是两个集合;

(5) 若 $N_{G_l}(A)=B$,则转第(9)步,否则进行下一步,其中, $N_{G_l}(A)\subseteq Y_k$ 是与 A 中结点邻接的结点集合;

(6) 找一结点 $y\in N_{G_l}(A)-B$;

(7) 若 y 是 M 饱和点,则找出 y 的配对点 z,令 $A\leftarrow A\cup\{z\}, B\leftarrow B\cup\{y\}$,转第(5)步,否则进行下一步;

(8) 存在一条从 x_0 到 y 的可增广路 P,令 $M\leftarrow M\oplus E(P)$,转第(3)步;

(9) 按下式计算 a 值: $a=\min_{\substack{x_i\in A\\ y_j\notin N_{G_l}(A)}}\{l(x_i)+l(y_j)-\omega_{ij}\}$,修改标号:

$$l'(v)=\begin{cases} l(v)-a, & \text{若} v\in A\\ l(v)+a, & \text{若} v\in B\\ l(v), & \text{其他}\end{cases},$$

根据 l'求 $E_{l'}$及 $G_{l'}$;

(10) $l\leftarrow l', G_l\leftarrow G_{l'}$,转第(6)步.

Kuhn-Munkres 算法的时间复杂度是 $O(t^3), t=\max(n,m)$.求出最大权 ω 和取得 ω 的匹配 M 后,定义视觉相似度 $Vision=\frac{\omega}{n}$.为了确定 Y_k 与 X 相似的片段边界,取 X 关联 M 的所有 y,从小到大排序为 $\{y_\alpha,y_\beta,\ldots,y_\gamma\}$, $\alpha,\beta,\gamma\in[1,m]$.在这个集合中, y_α,y_β 可能并不连续,即 $y_\beta-y_\alpha>1$,根据视频片段连续性的定义,取 y_α 与 y_γ 之间的所有镜头构成相似片段 $Y'_k=\{y_\alpha,y_{\alpha+1},\ldots,y_\gamma\}$.

3.2 动态规划算法计算顺序因子

在 3.1 节计算的最优匹配 M 中,进一步考察 Y'_k和 X 按时间顺序对应的情况,即找到 Y'_k按时间顺序和 X 有边的最长镜头数目,以此来度量顺序因子.这个问题可以归结为最长公共子序列(LCS)问题:给定两个序列 $X=\{x_1,x_2,\ldots,x_n\}$和 $Y'_k=\{y_\alpha,y_{\alpha+1},\ldots,y_\gamma\}$,要求找出 X 和 Y'_k 的一个最长公共子序列,动态规划算法可以有效地解决这个问题.为了计算方便,把 $\{y_\alpha,y_{\alpha+1},\ldots,y_\gamma\}$表示为 $\{y_1,y_2,\ldots,y_l\}$, $l=\gamma-\alpha+1$,用 $c[i,j]$ 记录序列 X 和 Y'_k 的最长公共子序列的长度,建立递归关系如下:

$$c[i,j]=\begin{cases} 0, & \text{当} i=0\text{或} j=0\text{时}\\ c[i-1,j-1]+1, & \text{当} i,j>0\text{且}(x_i,y_j)\in M\\ \max(c[i,j-1],c[i-1,j]), & \text{当} i,j>0\text{且}(x_i,y_j)\notin M\end{cases}.$$

动态规划算法的时间复杂度为 $O(nl)$, n 是查询片段 X 的镜头数目, l 是相似片段 Y'_k 的镜头数目.定义顺序因子 $order=\frac{c[i,j]}{n}$.

3.3 计算干扰因子

在最优匹配 M 中, X 和 Y'_k 会有少量镜头没有边关联,这说明这些镜头不能找到对应的相似镜头,它们的存在体现了对应的不连续性,定义干扰因子 $Interference=\frac{2\times|M|}{n+l}$.这个等式表明在两个相似片段 X 和 Y'_k 的所有镜头中,能找到对应相似镜头的镜头比例.

3.4 计算总的相似度

根据前面的分析,本文用下列公式计算查询片段 X 和它的相似片段 Y'_k 的相似度:

$$Similarity(X,Y'_k)=\omega_1\cdot Vision+\omega_2\cdot Order+\omega_3\cdot Interference.$$

其中,$\omega_1,\omega_2,\omega_3$ 表明了人们对视觉、顺序、干扰因子的重视程度,不同的用户可以根据自己对这 3 个判断标准的喜好程度来调整它们.在第 4 节的实验中,分别取ω_1=0.4,ω_2=0.3,ω_3=0.3.实验结果表明,这种取法能够符合大多数人的相似性判断标准.

4 实验结果

实验数据是从电视录制的节目中获取的,这个视频数据库非常具有挑战性,总长为 3 小时 11 分钟,共 4 714 个镜头,286 936 帧图像,包括了广告、新闻、体育、电影各种类型的节目,这里面有重复的相同视频片段,如新闻的片头、广告等(精确的片段检索,见第 4.1 节);也有很多重复的相似视频片段,如体育节目中的不同网球比赛、不同时间长度和编辑的相同广告等(相似的片段检索,见第 4.2 节).为了验证提出方法的有效性,本文实现了文献[11]的方法并以此方法作为实验对比,主要有这样两个原因:(1) 文献[11]是目前所给出的实验数据最好的方法,也是最新的一种方法;(2) 与本文的方法功能一致,能够在视频库里自动分割出相似片段,然后按照相似度从高到低排列它们.在视频片段检索中,除了检索的准确性以外,检索速度也是非常重要的一个指标,基于这种考虑,本文也比较了两种方法的检索速度,使用的测试机器是 PIII Dual CPU 1G Hz,内存 256M.

图 8 是实验程序的用户界面:上面一行是查询的某条广告,显示的是它的关键帧,下面是检索的结果,按照相似度递减的顺序先后排列.检索出的第 1 行即是查询片段本身,它的相似度当然是最高的,其余的片段按照相似度递减的顺序先后排列.可以看到,排列的相似片段体现了第 3 节不同因子的作用,如前 3 个片段和查询片段在时间顺序上更为相似.具体的实验结果分别在第 4.1 节和第 4.2 节给出.

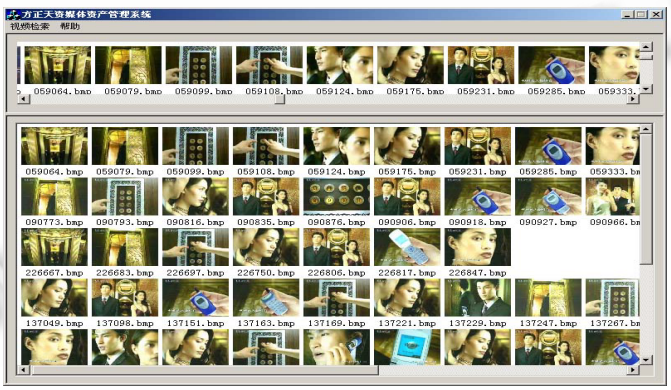

Fig.8 Retrieving and ranking similar video clips

图 8 检索和排列相似的视频片段

4.1 精确的片段检索

从表 1 可以看到,本文的方法和文献[11]都取得了 100%的查全率(recall),但在查准率上(precision),本文的方法优于文献[11],主要原因在于文献[11]仅仅计算了两个片段相似镜头的数量,而本文的方法考虑了相似镜头的对应关系(见第 2.3 节).在检索速度上,本文的方法快于文献[11],根据实验,总的检索时间基本上是等于相似镜头判断的时间,文献[11]采用按时间顺序逐帧比较的办法,而本文的方法只需比较每个镜头的关键帧,因此检索速

度快于文献[11].

Table 1 Result of exact video clip retrieval
表 1 视频片段精确检索的实验结果

Query clip	Frame numbers	The proposed approach			Match and tiling approach[11]		
		Precision (%)	Recall (%)	Speed (s)	Precision (%)	Recall (%)	Speed (s)
News logo	832	100	100	108	75	100	230
Football games in news	715	100	100	74	100	100	196
HuiYuan commercial	367	100	100	167	33.3	100	97
GuangMing commercial	374	100	100	89	100	100	101
FuLinMen commercial	432	100	100	99	100	100	116
Average	544	100	100	107	81.7	100	148

4.2 相似的片段检索

在表 2 中,无论是查全率,还是查准率,本文的方法都优于文献[11].查询片段 1 和 2 是两个难度很大的查询,在本文的视频库中,网球比赛共出现 4 次,我们漏掉了其中两个,原因是使用了蓝色网球场查询,而漏掉一个的网球场是绿色,另外一个主要是选手和观众镜头,反映蓝色球场的镜头很少,文献[11]也同样漏掉了这两个片段.与查询片段 1 类似,查询片段 2 也是一个语义很强而颜色特征很难利用的片段,综合整个片段反映这个语义的基本颜色特征,本文的方法也取得了不错的检索效果.这也说明了由几个镜头组成的视频片段查询,能够取得比单个镜头或图像查询更好的效果.在检索速度上,本文的方法同样快于文献[11],查询片段越长,本文方法的优势越明显.例如,在查询片段 2,本文方法的速度比文献[11]快了 5 倍多.此外,如图 8 所示,相比较文献[11]而言,本文方法的显著优势还表现在根据相似度从大到小排列相似片段上,因为除了视觉特征,本文还考虑了相似片段的不同因子,而文献[11]的相似度仅仅取决于相似镜头的数量.通过对几个人的测试结果表明,本文方法在相似片段的排序上,更加符合人的视觉特征和心理特征.

Table 2 Result of similar video clip retrieval
表 2 视频片段相似性检索的实验结果

Query clip	Frame numbers	The proposed approach			Match and tiling approach[11]		
		Precision (%)	Recall (%)	Speed (s)	Precision (%)	Recall (%)	Speed (s)
Tennis ball game	507	100	50	49	100	50	140
Doctors salvaging patient	1 806	60	85.7	93	50	50	507
TCL commercial	374	100	100	116	85.7	100	100
NaoBaiJin commercial	374	100	100	134	100	100	100
XiaXin commercial	374	100	100	108	100	50	99
Average	687	92	87.1	100	87.1	70	189

5 总 结

本文提出了一种新的基于图论最大匹配和最优匹配算法的视频片段检索方法.最大匹配用于从视频库里分割出与查询片段相似的多个片段,最优匹配用于表示和建模反映不同因子的相似度模型.实验结果表明,与具有同样功能的方法[11]相比,本文提出的方法可以取得更高的检索精度和更快的检索速度,同时在相似片段的排列顺序上,更加符合人的心理特征.

除了视频片段检索以外,本文提出的图论匹配方法在多媒体检索的相似性度量上有着广泛的应用.例如在镜头检索中,可以在下层帧序列的基础上按照这个方法计算镜头之间的相似性.我们下一步的工作,主要是继续运用该方法来解决这一类的复杂问题.

致谢 作者感谢北京大学计算机科学技术系的张立昂老师和耿素云老师在图论知识上给予的帮助.

References:

[1] Lin T, Ngo CW, Zhang HJ, Shi QY. Integrating color and spatial features for content-based video retrieval. In: Proceedings of the IEEE International Conference on Image Processing (ICIP 2001). 2001. 592~595. http://research.microsoft.com/asia/dload_files/group/mcomputing/ICIP01_lin-4th.pdf.

[2] Lin T, Zhang HJ, Feng JF, Shi QY. Shot content analysis for video retrieval applications. Journal of Software, 2002,13(8):1577~1585 (in Chinese with English abstract).

[3] Zhao L, Qi W, Li ZQ, Yang SQ, Zhang HJ. Content-Based retrieval of video shot using the improved neatest feature line method. Journal of Software, 2002,13(4):586~590 (in Chinese with English abstract).

[4] Ngo CW, Pong TC, Zhang HJ. On clustering and retrieval of video shots through temporal slices analysis. IEEE Transactions on Multimedia, 2002,4(4):446~459.

[5] Dimitrova N, Abdel-Mottaled M. Content-Based video retrieval by example video clip. In: Proceedings of IS&T and SPIE Storage and Retrieval of Image and Video Databases VI, Vol.3022. 1998. 184~196.

[6] Jain AK, Vailaya A, Wei X. Query by video clip. ACM Multimedia Systems, 1999,7(5):369~384.

[7] Tan YP, Kulkarni SR, Ramadge PJ. A framework for measuring video similarity and its application to video query by example. In: Proceedings of IEEE International Conference on Image Processing (ICIP 1999). 1999. 106~110. http://www.ee.princeton.edu/~ramadge/postscript/ICIP99_635.pdf.

[8] Liu XM, Zhuang YT, Pan YH. A new approach to retrieve video by example video clip. In: Proceedings of ACM Multimedia. 1999. 41~44. http://amp.ece.cmu.edu/Publication/Xiaoming/ACMMM99Poster.pdf.

[9] Wu Y, Zhuang YT, Pan YH. Content-Based video similarity model. In: Proceedings of the ACM Multimedia, 2000. http://www.acm.org/sigs/sigmm/MM2000/ep/wu/wu.pdf.

[10] Zhuang YT, Liu XM, Wu Y, Pan YH. A new approach to retrieve video by example video clip. Chinese Journal of Computers, 2000,23(3):300~305 (in Chinese with English abstract).

[11] Chen LP, Chua TS. A match and tiling approach to content-based video retrieval. In: Proceedings of IEEE International Conference on Multimedia and Expo (ICME 2001). 2001. 417~420. http://www.comp.nus.edu.sg/~chuats/papers/icme01.pdf.

[12] Ngo CW, Pong TC, Chin RT. Video partitioning by temporal slice coherency. IEEE Transactions on Circuits and Systems for Video Technology, 2001,11(8):941~953.

[13] Ngo CW, Pong TC, Zhang HJ. Motion-Based video representation for scene change detection. International Journal of Computer Vision, 2002,50(2):127~143.

[14] Dai YQ, Hu GZ, Chen W. Graph Theory and Algebra Structure. Beijing: Tsinghua University Press, 1995. 89~91 (in Chinese).

[15] Xiao WS. Graph Theory and Its Algorithms. Beijing: Aviation Industry Press, 1993. 134~142 (in Chinese).

附中文参考文献:

[2] 林通,张宏江,封举富,石青云.镜头内容分析及其在视频检索中的应用.软件学报,2002,13(8):1577~1585.

[3] 赵黎,祁卫,李子青,杨士强,张宏江.利用改进 NFL 算法对镜头进行基于内容的检索.软件学报,2002,13(4):586~590.

[10] 庄越挺,刘小明,吴翌,潘云鹤.通过例子视频进行视频检索的新方法.计算机学报,2000,23(3):300~305.

[14] 戴一奇,胡冠章,陈卫.图论与代数结构.北京:清华大学出版社,1995.89~91.

[15] 肖位枢.图论及其算法.北京:航空工业出版社,1993.134~142.