# 利用改进 NFL 算法对镜头进行基于内容的检索*

赵 黎[1], 祁 卫[2], 李子青[2], 杨士强[1], 张宏江[2]

[1](清华大学 计算机科学与技术系,北京 100084);
[2](微软亚洲研究院,北京 100080)
E-mail: li.zhao@media.cs.tsinghua.edu.cn; lizhao99@mails.tsinghua.edu.cn
http://imlab.cs.tsinghua.edu.cn/~zhaoli

摘要: 基于镜头的分类和检索对于视频库的管理和查询非常重要.将“最近特征线”法(nearest feature line,简称 NFL)用于镜头的分类和检索.将镜头中的代表帧看做是某个特征空间中的点,通过这些点间的连线表征该镜头的总体特征信息,然后计算查询图像和特征线的距离,以决定镜头与查询图像的相似度.为了更适于视频数据,对原来的 NFL 方法进行了改进,基于镜头内部内容活动程度对特征线进行限制.实验结果表明,改进的 NFL 方法比传统的 NFL 方法以及常用的聚类方法,如最近邻法(nearest neighbor,简称 NN)和最近中心法(nearest center,简称 NC),在性能上有所提高.

关 键 词: 基于内容检索;最近特征线(NKL);视频检索;视频分类;视频镜头
中图法分类号: TP391 文献标识码: A

视频信息表现力强,但其数据的海量性和内容的复杂性使得管理和使用都很困难.对于一些多媒体应用,如数字化的视频点播系统、数字图书馆等,要求实现对视频信息进行基于内容的管理和检索,能够使用户方便地获取其需要的信息.因此,人们提出了很多方法对视频进行基于内容的分析,从而实现基于内容的检索和访问[1~5].

一般的思路是首先进行镜头边界检测,以独立的镜头作为视频序列的基本结构单元和检索单元;然后在每个镜头的内部提取“关键帧”来代表该镜头的内容.目前已经有很多成熟的技术来进行镜头检测和关键帧的提取[5,6].

对于用户来讲,最直观和最方便的检索方法就是,用户向检索系统提交一幅查询图像,然后检索系统根据内容上的相似性按顺序向用户返回视频库中的一组镜头,作为检索结果.

最“直接”的搜索匹配方法是,将每个镜头都看做是图像帧的序列,然后基于某些特征,如颜色、纹理、形状,将视频库中与用户提交的查询图像相似的图像按顺序排列出来.实际上这就退化成了一个静止图像库的检索问题,由于没有利用视频序列中图像间有很大冗余这个特性,这种方法的效率显然是非常低的.所以人们利用关键帧代表一个镜头来与检索图像进行相似度计算.通常,一个镜头都有多个关键帧,这样,检索问题就变成了一个分类问题,每个镜头的关键帧组成一个类,按照检索图像在特征空间与各个类的距离作为排列查询结果的依据.因此常用的方法是最近邻法(nearest neighbor,简称 NN)和最近中心法(nearest center,简称 NC).

* 收稿日期: 2000-05-11; 修改日期: 2000-11-21
基金项目: 国家重大基础研究 973 发展规划资助项目(G1999032704);清华大学-微软公司多媒体技术实验室资助项目

作者简介: 赵黎(1975－),男,湖北沙市人,博士生,主要研究领域为多媒体,视频分析与传输,多媒体数据编码算法,视频信息基于内容的检索;祁卫(1971－),男,北京人,博士,副研究员,主要研究领域为多媒体技术,基于对象压缩编码方法中的视频分析算法,视频信息基于内容检索;李子青(1958－),男,湖南株洲人,博士,研究员,主要研究领域为图像与视觉的建模、处理与理解,模式识别,统计学习,优化算法,多媒体信息分类和检索;杨士强(1952－),男,山东高唐人,教授,博士生导师,主要研究领域为分布式多媒体,多媒体数据压缩技术,视频数据的存储和检索技术;张宏江(1960－),男,黑龙江哈尔滨人,博士,主要研究领域为视频和图像的分析、处理和检索,媒体压缩和传输,互联网多媒体及计算机视频的研究开发工作在家庭和工业中的应用.

但这里的问题是,几幅关键帧其实很难"覆盖"整个镜头的内容(这和关键帧提取的方法有很大关系),而 NN 法和 NC 法实际上都很依赖于类集合中样本点的数目和分布.因此需要一种新的方法,在已知类别的样本点很少的情况下,仍能较准确地进行分类.

在 Stan.Z.Li 等人所作的工作中[7,8],提出了一种新的聚类算法,称为最近特征线法(nearest feature line,简称 NFL),并用于语音分类和人脸识别.该算法将某一类中的任何两个样本点间连接成一条直线,这根直线就是特征线(feature line).依据查询点(query point)到特征线(feature line)的最小距离就可以进行检索和分类.

在本文中,我们将 NFL 方法用于基于镜头的检索,并将其进行改进以适应视频序列的情况.我们认为,镜头中的每一帧都是特征空间中的一个点,当帧连续发生变化的时候,其对应的点就会在特征空间留下一条连续的轨迹.相隔很近的帧则认为它们在特征空间中的轨迹近似为直线.为了更适于视频的分类与检索以及考虑到特征点的分布范围,我们在原来的 NFL 算法上作了进一步的改进.这种改进的算法将局部序列的特征点集线性化,同时也考虑到了特征点的分布范围.在第 1 节中,我们将集中讨论有关改进的 NFL 算法.第 2 节将给出有关的实验结果.第 3 节为结论.

## 1 用于镜头检索的改进的最近特征线算法

NFL 方法假定每个类中至少可以得到两个样本(即特征点),然后通过对这些已知样本点进行线性内插和外插来近似类中其他未知的元素,即通过已知样本点间的连线(即特征线)来近似整个类的集合.

我们将镜头中的关键帧作为特征空间中已知的样本点.由于本文主要侧重于研究分类方法,因此只简单地选取颜色直方图空间作为讨论的特征空间.

在第 1.1 节中我们首先介绍 NFL 方法如何用于镜头检索,第 1.2 节将着重介绍对原有的 NFL 算法如何加以改进,使之更符合视频检索的特点.

### 1.1 NFL方法用于镜头检索

由于在一个镜头的内部,可以认为相邻两个关键帧之间在颜色直方图特征空间中的距离主要是由于摄像机的运动或操作以及在摄像机前较大物体的大幅移动所造成的,这样可以利用两个关键帧的特征点间的连线来近似它们之间的图像帧序列在特征空间的轨迹.

在特征空间中,两幅关键帧图像 $F_i$ 和 $F_j$ 对应于两个特征点 $f_i$ 和 $f_j$ .定义

$$f_k = \{f_{1k}, f_{2k}, ..., f_{Mk}\}. \quad (1)$$

这里,$M$ 为特征空间的维数.帧 $F_i$ 和 $F_j$ 之间的相似度可以用 $f_i$ 到 $f_j$ 的距离来衡量,表示为 $\Delta f = \|f_i - f_j\|$ .穿过 $f_i$ 和 $f_j$ 的直线,即特征线(feature line,简称 FL),可以表示为 $\overline{f_i f_j}$ (如图 1 所示).

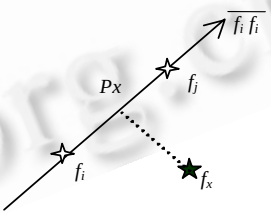



Fig.1 Feature points $f_i$ and $f_j$, feature line $\overline{f_i f_j}$, and query feature point $f_x$

图 1 特征点 $f_i$ 和 $f_j$,特征线 $\overline{f_i f_j}$,查询特征点 $f_x$

令 $F^c = \{f_i^c \mid 0 < i \le N_c\}$ 表示含有 $N_c$ 个原型特征点的集合 $C$.这样一共可以构造 $K_c = \frac{N_c(N_c-1)}{2}$ 条直线来表示集合 $C$.对于集合 $C$ 的特征空间可以由这 $K_c$ 条特征线组成:

$$S^c = \{\overline{f_i^c f_j^c} \mid 0 < i, j \le N_c, i \ne j\}. \quad (2)$$

这是全部特征空间的一个子集.当数据库中有 $M$ 个这样的集合,就有 $M$ 个这样的 FL 空间建立,其总的特征线的数目为 $N_{\text{total}}$ ,其中 $N_{\text{total}} = \sum_{c=1}^{M} K_c$.

查询点 $f_x$ 到特征线 $\overline{f_i f_j}$ 的距离可以定义为 $Dist(f_x, \overline{f_i f_j})$ .令 $p_x$ 表示 $f_x$ 在特征线 $\overline{f_i f_j}$ 向上的投影:

$$p_x = f_i + \mu(f_j - f_i), \quad (3)$$

其中

$$\mu = \frac{(f_x - f_i) \bullet (f_j - f_i)}{(f_j - f_i) \bullet (f_j - f_i)}. \tag{4}$$

所以,

$$Dist(f_x, \overline{f_i f_j}) = \| f_x - p_x \|. \tag{5}$$

此处,$\|\bullet\|$ 表示欧式距离.

设用户查询图像 $F_x$ 在特征空间对应点 $f_x$ (如图 1 所示),则通过计算该点与特征空间中所有镜头的距离来决定其与视频库中各个镜头的相似度.查询点与某个镜头的距离就是查询点与该镜头中所有特征线的距离中的最小值.然后将所得到的查询点与所有镜头的距离按从小到大排序.距离最短的那个镜头就是在此视频库中与查询点最相近的镜头.

### 1.2 改进的NFL算法

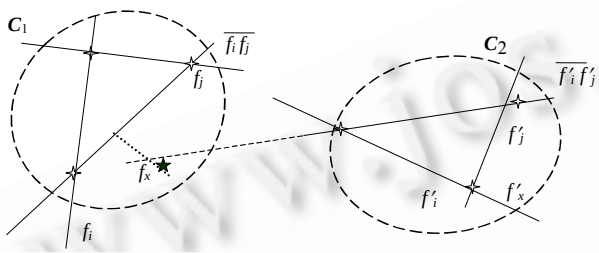



Fig.2 The extension of feature line $\overline{f'_i f'_j}$ passing through the region around feature line $\overline{f_i f_j}$

图 2 特征点 $f_i$ 和 $f_j$ 以及穿过这两个特征点所构成的特征线 $\overline{f_i f_j}$,特征线 $\overline{f_j f_i}$ 同特征线 $\overline{f'_i f'_j}$ 发生了交叉

上节中介绍的一般的 NFL 算法存在一个严重的问题:由于在 NFL 中特征线是无限长的,查询点与某条特征线的延长线距离很近,而实际特征点间却相距甚远(如图 2 所示).如果仅仅从距离上考虑,离特征点 $f_x$ 最近的特征线为 $\overline{f'_i f'_j}$,而实际的情况是 $\overline{f_i f_j}$ 离 $f_x$ 最近.可见,必须考虑限制特征线的范围.为了评估这个特征线的范围,我们引进“镜头活动性”的概念(shot activity)[9],当人们在观看视频或动画的时候,总是有一种节奏感,比如感到情节很舒缓或是节奏很快.活动性的概念正是反映了这种很舒缓或是很紧张的感觉.

这里用一个描述子来表示视频镜头的活动性.具体的定义为

$$Act_c = \underset{i}{\mathrm{MAX}}(\| f_i - f_{\text{center}} \|), \tag{6}$$

这里,$1 \le i \le N_c$, $f_{\text{center}} = \frac{1}{N_c} \sum_{i=1}^{N_c} f_i$ .从而得到特征线的限制范围:

$$SearchRange = C \cdot Act_c, \tag{7}$$

这里,$C$ 是一个常数,取试验值为 0.65.只有当

$$\| f_x - f_{\text{center}} \| \le SearchRange, \tag{8}$$

结果才能被接受,否则就被忽略.

所以,镜头的活动性反映了一个镜头中特征点的分布.假如一个查询点离这个集合的分布非常远,就算离这个集合的某一特征线很近,仍旧认为这个特征点不属于这个集合.

## 2 实验结果

在本文的实验中,选用了颜色直方图作为特征,具体使用的是 1976 年定义的 CIE $u^*v^*$颜色空间:将 $u^*v^*$分量分别量化为 16 级,在 $u^*$方向上将 0.1612~0.2883 的范围拉升为 0~1,在 $v^*$方向上将 0.4361~0.5361 的范围拉升为 0~1,超过这个区域的颜色则合并到离它最近的级别中.

为了验证所提出的改进的 NFL 算法的效果,建立了一个包含 160 个镜头的视频库.这些镜头取自于 40 分钟的体育新闻,包括田径、游泳、足球、赛艇以及插播的广告节目.下面实验的目的就是为了比较改进的 NFL 算法、NFL 算法以及其他聚类算法(如 NC 和 NN 算法)的检索效果.图 3 给出了本文实验环境的用户界面,上面一行是浏览区域,显示整个视频中每个镜头的第 1 个关键帧,用来代表这个镜头,用户可以从中选出想要进行查询

的图像.下面一行是结果区域,按照 NFL 距离从小到大显示查询结果.

首先,用户提交一个查询关键帧,系统将返回视频库中查到的相匹配的镜头列表,并且按照距离从小到大排列.本文采用一个加权的得分来衡量查询效果[8],定义

$$\eta(q,m)=\sum_{k=1}^{m} w_k Match(q,r_k). \tag{9}$$

$$Match(q,r_k)=\begin{cases}1, & \text{如果 } r_k \text{ 同 } q \text{ 相关}\\ 0, & \text{如果 } r_k \text{ 同 } q \text{ 不相关.}\end{cases}$$

这里,$q$ 代表查询图像; $r_1,r_2,...,r_m$ 表示系统返回的认为最相似的前 $m$ 个镜头.目前由人对查询结果是否与查询图像相关进行主观判断,即决定式(9)中 $Match(q,r_k)$ 的取值.

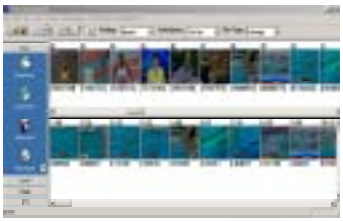


Fig.3 Interface of the retrieval program
图 3 可视化的视频检索程序的界面

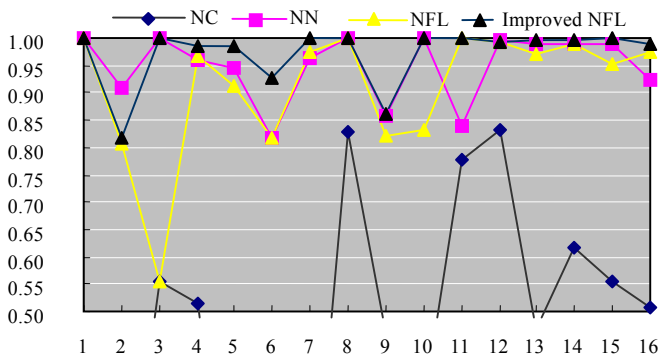
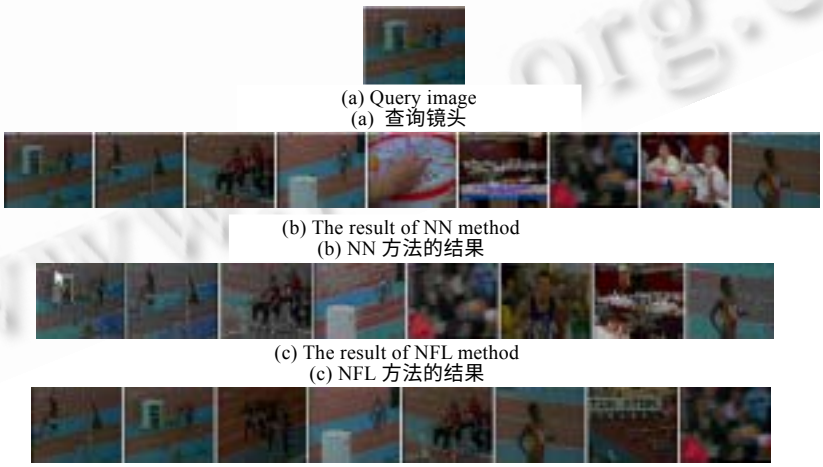



Fig.4 The score of different methods
图 4 各种方法的检索得分

$w_k=W\frac{1}{k}$ 是一个递减的权重序列 $(k=1,2,...)$,其中 $W=1\Big/\sum_{k=1}^{N_q}\frac{1}{k}$, $N_q$ 代表查到的镜头中与查询图像 $q$ 相似的数目.因为权重 $w_k$ 随着位置 $k$ 的增加而减小,所以正确的排名越靠前,则对 $\eta(q,m)$ 的贡献越大.因子 $W$ 是为了将结果归一化:如果前 $N_q$ 个镜头完全正确,则 $\eta(q,m)$ 达到可能的最高值 1.

图 4 给出了本文的实验结果.结果表明,NFL 算法与其他算法相比取得较好的效果,而且改进后的 NFL 算法则更加适合于视频的检索和分类.图 5 给出了一个查询的例子.在这个例子中可以看到使用 NFL 方法,特征线在特征空间产生了交叉(如图 5(c)所示),如果采用改进后的 NFL 方法(如图 5(d)所示),则不存在这种问题,并且表现优于其他的方法.



(a) Query image
(a) 查询镜头

(b) The result of NN method
(b) NN 方法的结果

(c) The result of NFL method
(c) NFL 方法的结果

(d) The result of improved NFL method
(d) 改进后 NFL 方法的结果

Fig.5 An example of retrieval result
图 5 一个检索结果的例子

## 3 结 论

本文中,我们将一种新的聚类算法(NFL)用于视频的检索和分类中,并且根据视频的特点,在原来的算法基础上提出了改进的 NFL 算法.这种算法克服了常规算法中将视频序列中关键帧看做是独立图片的缺点,认为连续的视频帧中存在着内容相似的关系,而特征线则代表了这种关系.这是真正基于内容的检索和分类.

最后的实验结果显示,NFL 方法以及改进的 NFL 方法在视频检索和分类方面表现突出.在本文的实验中仅用到颜色的特征,如果采用更多的特征,如:纹理、形状等,相信可以取得更好的效果.这也是我们下一步的工作之一.

**References:**


[1] Rui, Y., Huang, T.S., Mehrotra, S. Exploring video structure beyond the shots. In: Proceedings of the ICMCS'98 IEEE Conference on Multimedia Computing and Systems. 1998. 237~240.

[2] Zhang, H.J., Zhong, D., Smoliar, S.W. An integrated system for content-based video retrieval and browsing. Pattern Recognition, 1997,30(4):643~658.

[3] Smoliar, S.W., Zhang, H.J. Content-Based video indexing and retrieval. IEEE Multimedia, 1994,1(2):62~72.

[4] Shan, M.K., Lee, S.Y. Content-Based video retrieval based on similarity of frame sequence. In: Proceedings of the ICMCS'98 IEEE Conference on Multimedia Computing and Systems. 1998. 90~97.

[5] Hanjalic, A. Visual-Content analysis for multimedia retrieval system [Ph.D. Thesis]. Delft University of Technology, 1999.

[6] Gresle, P.O., Huang, T.S. Gisting of video documents: a key frames selection algorithm using relative activity measure. In: Proceedings of the 2nd International Conference on Visual Information Systems. 1997. 279~286.

[7] Li, S.Z., Lu, J. Face recognition based on nearest linear combinations. IEEE Transactions on Neural Networks, 1999,10(2): 439~443.

[8] Li, S.Z. Content-Based classification and retrieval of audio using the nearest feature line method. IEEE Transactions on Speech and Audio Processing, 2000,8(5):619~625.

[9] Hsu, P.R., Harashima, H. Detecting scene changes and activities in video databases. In: Proceedings of the ICASSP'94 IEEE International Conference on Acoustics, Speech, and Signal Processing. 1994. 33~36.


# Content-Based Retrieval of Video Shot Using the Improved Neatest Feature Line Method*


ZHAO Li[1], QI Wei[2], LI Zi-qing[2], YANG Shi-qiang[1], ZHANG Hong-jiang[2]

[1](*Department of Computer Science and Technology*, *Tsinghua University*, *Beijing* 100084, *China*);

[2](*Microsoft Research Asia*, *Beijing* 100080, *China*)

E-mail: li.zhao@media.cs.tsinghua.edu.cn

http://imlab.cs.tsinghua.edu.cn/~zhaoli



**Abstract:** The shot based classification and retrieval is very important for video database organization and access. In this paper, a new approach NFL (nearest feature line) used in shot retrieval is presented. Key-Frames in shot are looked as feature points to represent the shot in feature space. Lines connecting the feature points are further used to approximate the variations in the whole shot. The similarity between the query image and the shots in video database are measured by calculating the distance between the query image and the feature lines in feature space. To make it more suitable to video data, the original NFL method by adding constrains on the feature lines is improved. Experimental results show that the improved NFL method is better than the traditional classification methods such as the nearest neighbor (NN) and the nearest center (NC).



**Key words:** content-based retrieval; nearest feature line (NFL); video retrieval; video classification; video shot

* Received May 11, 2000; accepted November 21, 2000

Supported by the National Grand Fundamental Research 973 Program of China under Grant No.G1999032704; Tsinghua University-Microsoft Multimedia Laboratory