镜头内容分析及其在视频检索中的应用*

林通[1], 张宏江[2], 封举富[1], 石青云[1]

[1](北京大学 视觉与听觉信息处理国家重点实验室,北京 100871);
[2](微软中国研究院,北京 100080)
E-mail: jos@iscas.ac.cn; jos@admin.iscas.ac.cn
http:// www.jos.org.cn

摘要: 本文提出了一种新的镜头内容分析方法,及其在视频检索中的两个应用:镜头检索与场景结构提取.为了刻画一个镜头的内容变化,首先引入两个新的内容描述子:主色直方图和空间结构直方图.主色直方图能够捕捉那些持续时间最长的颜色,而这些颜色是这段视频所关注的对象或背景的主要颜色.从颜色块图提取的空间结构直方图是描述图像空间信息的一组特征.一个变化较大的镜头可以划分为几个内容一致的子镜头,两个镜头的相似性可以从对应子镜头的相似性计算得到.镜头相似性度量可以直接用于镜头检索,还可用于场景结构提取.本文提出分裂与合并力量竞争的场景结构提取方法.在大容量视频数据库上所进行的实验证实了本文方法在镜头检索和场景提取的优异表现.

关 键 词: 基于内容的视频检索;镜头内容分析;镜头相似性度量;场景结构提取

1 引言

随着在多媒体数据制造,存储,与传播方面取得的重大技术进步,数字视频已经成为人们的日常生活中不可或缺的一部分.数字视频也是数字图书馆计划中的核心内容.如何管理和检索海量的视频数据已经成为近十年来全球学术界和工业界一个富有挑战性的热门话题之一.因此,基于内容的视频检索(CBVR, Content-based Video Retrieval)方面的研究和被称为多媒体内容描述接口的国际标准 MPEG-7 的制定也就引起了人们广泛的关注.

通常一段视频数据可以划分为几个场景(也叫做故事单元),每个场景又包含一个到多个镜头.一个镜头是指一系列连续纪录的图像帧,用于表示一个时间段或相同地点连续的动作.镜头由摄像机一次摄像的开始和结束所决定.一个视频场景结构指一连串语义相关的镜头,它们一般发生在相同的时间和地点,出现相同的人物或事件.所以视频数据可以按照由粗到细的顺序划分为四个层次结构:视频(video),场景(Scene),镜头(Shot),和图像帧(Frame).

目前大多数研究主要集中于镜头边界检测和关键帧选取,对镜头的内容分析才刚刚开始.镜头是视频的自然结构单元,镜头内容分析将是基于内容的视频检索的核心技术之一.在当前文献中,视频镜头通常用几个关键帧来表示[1],颜色,纹理,和形状等低级特征直接从关键帧提取出来用于索引与检索.一般采用聚类算法进行关键帧选取,也可以根据不同的镜头类型进行关键帧构造.比如,一个变焦(zoom)镜头可以简单表示为变焦之前和之后的两个关键帧[2],一个扫描(pan)镜头可以通过构造一个全景图(panoramic)来表示[3].最近,文献[4]提出一种基于最近特征线(NFL, nearest feature line)的端点检测算法用于选取关键帧.由于计算方面的

* **收稿日期:** 2001-01-08; **修改日期:** 2001-05-09
基金项目:

作者简介: **林通**(1974－),男,四川南充人,博士生,主要研究领域为视频处理;**张宏江**(1960－),男,河南郑州人,博士,研究员,主要研究领域为多媒体技术;**封举富**(1967－), 男,湖南长沙人,博士,副教授,主要研究领域为模式识别;**石青云**(1936－),女,四川合川人,中科院院士,教授,博士生导师,主要研究领域为模式识别,生物度量学.

考虑,视频检索方面的技术通常类似于图像检索.但是,上述基于关键帧的镜头表示方法最大的问题是,不能对存在于视频中的时间信息进行充分利用.

场景结构提取方面的工作包括文献[2][5][6][7][8][10].文献[5]通过匹配关键帧图像块计算镜头相似度,然后分三种情况检测场景边界:(1)当前镜头与后面的某个镜头相似;(2)当前镜头与前面的镜头相似;(3)当前镜头与其它镜头都不相似,但是它前面的镜头与它后面的镜头相似.文献[6]通过比较关键帧的累积颜色直方图而计算镜头相似度.文献[7]比较两个镜头关键帧的颜色直方图和累积的活动量,然后构造一个中间结构叫镜头组(group),再把那些相似的但不相邻的镜头组合并为一个场景.文献[8]对两个镜头中的所有图像帧两两比较,然后计算镜头前后相似度的大小以确定场景边界.

本文提出一种新的镜头内容分析方法.首先,一个镜头内的内容变化被分解为几个内容一致的子单元,称为子镜头(subshot).要描述那些视觉内容有重大变化的镜头,比如一个镜头从室内转移到窗外,子镜头是必不可少的.文献[2]利用相机运动信息来提取子镜头;我们利用的是视频颜色对象,因为描述视频内容的时空变化需要更多的语义考虑.我们定义一个“颜色对象”同时为 HSV 颜色空间中的一个颜色球和图像帧中的一个颜色区域(颜色对象是一个准对象,因为目前技术还不能准确自动的提取出语义对象).为了度量视觉内容变化用于提取和表示子镜头,我们构造了两个内容描述子:

(1) 主色直方图(Dominant Color Histogram)通过把时间信息融入到颜色直方图中,用来描述一组图像帧(GoF, group of frames)中最重要的那些颜色;

(2) 空间结构直方图(Spatial Structure Histogram)用来描述一帧图像中的空间结构信息,它能对颜色直方图提供互补功能,因为颜色直方图缺少颜色的空间分布信息.

子镜头提取出来以后,就可以根据对应子镜头的相似性计算两个镜头的相似性了.镜头相似性度量可以直接用于镜头检索,还可以用于场景结构提取.我们还提出了新的基于力量竞争的场景提取方法.

本文第 2 节讲述镜头内容表示与相似性度量,包括主色直方图与空间结构直方图的构造,以及在此基础上的子镜头提取与镜头相似性度量.第 3 节给出场景结构提取的算法.第 4 节是镜头检索和场景提取的实验结果.

2 镜头内容表示与相似性度量

2.1 主色直方图

颜色直方图由于其简单有效而广泛用于基于内容的图像检索(CBIR, content-based image retrieval)中.很自然的,颜色直方图也在基于内容的视频检索中广泛使用.文献[9]提出了一组颜色直方图叫阿尔法裁减的平均直方图(alpha-trimmed average histogram),包括平均直方图和中值直方图.如果对每帧图像都提取一个颜色直方图,那么一组图像帧就有一系列颜色直方图.为了把这一系列颜色直方图合并成一个颜色直方图,先对相同直方图格(bin)的值先做排序,去除最高的和最低的几个值,再作平均.其思想相当于歌咏比赛的计分方案,比如有十个裁判打分,去除两个最高分和两个最低分,然后再计算平均分.

我们提出通过主颜色提取与跟踪,用主颜色直方图来描述一组图像帧(GoF).描述一幅图像的主颜色直方图不单减少了直方图的尺寸,而且增强了直方图匹配的表现,因为它抓住了最主要的颜色内容而不易受到噪声的影响[11].用于描述一组图像帧的主色直方图不仅考虑了单幅图像的主颜色,而且还考虑了这组图像主颜色的时间变化.这种表示方法抓住了视频作为连续时间媒体的本质.主颜色直方图与其它直方图不同之处在于它利用了时间信息并且有某些语义方面的考虑.“一组图像帧”是一个一般概念,它可以是镜头,子镜头,或者一组镜头.在下文叙述中,我们将把一个镜头作为一组图像帧,并假设这个镜头只包含单一的主题（即内容一致）.否则,我们可以把这个镜头切分为几个内容一致的子镜头作为一个图象帧组.

包含单一主题的镜头一般可以分为两种类型:(1)关注环境背景,比如一条街道,并没有要刻画的主要前景对象;(2)聚焦一个静止或者运动的物体,比如一辆车或者一个人.关注的前景/背景对象应该占据屏幕中心并且持续时间最长.颜色是一个有效的并且计算相对简单的特征.考虑到人类感知,并非在镜头中出现的所有颜

色,而是那些被关注的前/背景对象的颜色,主导了人们关于两个镜头相似性的度量.通过镜头内的时间变化,我们可以抓住并突出那些关注对象的颜色.描述一个镜头内容的主颜色直方图强调了主要前/背景对象的特征,这与以前的直方图完全不一样:通过颜色跟踪,得到每个颜色的持续时间,然后把那些持续时间短的颜色去除因而得到镜头主颜色,最后再按照持续时间长短对主颜色进行加权.下面简要给出主颜色直方图构造的方法,细节见于文献[10].

首先对每一帧图像计算颜色直方图(也可只用 MEPG1/2 的 I 帧 DC 图),然后找出这一帧的主颜色.我们使用的颜色模型是 HSV 颜色锥[11],因为按照欧氏距离,HSV 颜色空间的颜色分布比较均匀.采用三维笛卡尔坐标系进行量化,X 和 Y 轴量化为 20 个值,Z 轴(亮度)为 10 个值,如图 1 所示.每一帧的像素,或者 I 帧的 DC 块(使用 MPEG1/2 数据时),投影到量化的 HSV 颜色空间中.在这个三维颜色空间中归一化后的分布形成归一化的三维颜色直方图.在三维颜色直方图找出所有重要的局部最大值点;我们定义包含每个局部最大值点的直径为 3 个量化单位的小球为一个“颜色对象”.这些颜色对象中像素较多的(在实验中取前 20 名)被认为是一帧中的“主颜色对象”.注意它们并不对应图像帧中的一个空间对象.

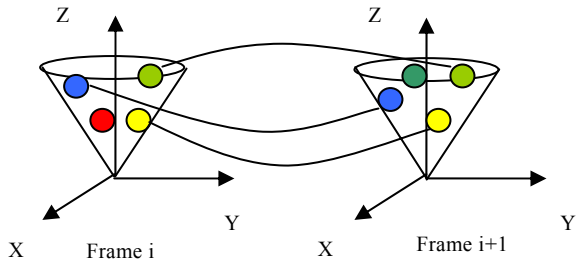

Fig.1 Color objects extraction and tracking

图 1 主颜色提取与跟踪

在连续的图像帧中,按上述定义的主颜色对象在 HSV 颜色空间中进行跟踪以便得到一个镜头的主颜色.如果在连续两帧的 HSV 颜色空间中,两个主颜色对象的位置差异足够小,那么这两个颜色对象被认为是同一个颜色.颜色跟踪过程一直持续到这个镜头结束为止.颜色对象跟踪以后,只有那些持续时间较长的颜色才被保留下来作为这个镜头的主颜色.换言之,对每个镜头可以构造一个主颜色直方图,*histda(x,y,z)* (*a* 表示一个镜头),它所包含的主颜色不但在单个图像帧中很突出,而且在整个镜头中都占主导.为了突出那些持续时间长的颜色,因为它们在感知上更重要,所以可以对它们赋予更多的权重.具体地,对应每个主颜色的直方图格(bin)按照其相对持续时间进行加权:

$$hist_d^A(x,y,z) = hist_d^a(x,y,z) \times d_1 / d_0$$

其中 *d0* 是镜头的持续时间,*d1* 是主颜色格 *(x, y, z)* 的持续时间.然后再对 *histdA(x,y,z)* 进行归一化,就得到了镜头的主颜色直方图.因此,一个镜头的主颜色直方图既表示单个图像帧的结构内容,又表示了整个镜头的时间内容.两个镜头 a 和 b 之间的相似性是对两个镜头的主颜色直方图作直方图交计算得到:

$$DchSim(a,b) = \sum_x \sum_y \sum_z \min[hist_d^A(x,y,z), hist_d^B(x,y,z)]$$

2.2 空间结构直方图

描述一幅图像全局或局部空间格局的图像特征是非常重要的.文献[12] 引入了一个新概念叫做图像结构特征,它是介于纹理和形状之间的一个一般特征.文献[12]提出用 Water-Filling Algorithm 从边缘图中提取基于边缘长度和分叉数的图像结构特征.本文将从颜色块图(color-blob map)中提取一系列新特征叫做空间结构直方图(Spatial Structure Histogram),用来描述一幅图像的空间信息.

首先用颜色量化计算出图像帧的颜色块图(实验结果见图 2).基本方法是在三维 HSV 颜色锥空间中用 K 均值聚类算法提取主要的颜色类.最优类数 *k* 用如下聚类可分性度量计算得到:

$$\rho(k) = \frac{1}{k}\sum_{i=1}^{k} \max_{1 \le j \le k, j \ne i} \{\frac{\eta_i + \eta_j}{\xi_{ij}}\}$$

其中 ηj 是类 *j* 的类内距离, ξij 是类 *i* 与 *j* 的类间距离.注意上述可分性度量不能处理 *k=1* 时的情形.K 均值

聚类算法分别对情形 $k=\{1, 2, 3, ..., 10\}$ 进行测试.如果当 $k=1$ 时类内距离小于某个给定的阈值,则类数就设为 1.否则选取使 $\rho(k)$最小的 k 为最终的类数.我们的实验程序直接在 MPEG 压缩域上实现,只使用了 I 帧的 DC 块图像.十个颜色类足以表示一般的 DC 块图像中的颜色分布.潜在的问题是纹理区域,它可能被分为若干个小的颜色区域.实验证实,上述聚类有效性分析有效地抑制了纹理碎区域,因为它偏好于选取大的颜色类.

从上述计算得到的颜色块图可以得到几个分布特征,包括面积直方图 *Harea*,位置直方图 *Hpos*,在 X 和 Y 方向上的区域方差直方图 *Hvx*, *Hvy*,以及在 X 和 Y 方向上的区域长宽直方图 *Hsx*, *Hsy*.它们分别定义为

(1) $$H_{area}(i)=\sum_{Rj\in\Omega i} Area(R_j), i=0,1,...,7$$

$$\Omega_i=\{R_j \mid Area(R_j)\in[A_i, A_{i+1})\}, i=0,1,...,7 \qquad A_i=1/2^{8-i}, i=1,2,...,8; A_0=0$$

(2) $$H_{pos}(i)=\sum_{Rj\in\Omega i} Area(R_j), i=0,1,2,...,15$$

$$\Omega_i=\{R_j \mid Center(R_j)\in Block(i)\}, i=0,1,2,...,15$$

(3) $$H_{vx}(i)=\sum_{Rj\in\Omega i} Area(R_j), i=0,1,2,...,7$$

$$\Omega_i=\{R_j \mid \sigma_x(R_j)\in[B_i, B_{i+1})\}, i=0,1,2,...,7 \qquad B_i=1/2^{8-i}, i=1,...,7, 8; \quad B_0=0$$

(4) $$H_{sx}(i)=\sum_{Rj\in\Omega i} Area(R_j), i=0,1,2,...,7$$

$$\Omega_i=\{R_j \mid Width(R_j)\in[B_i, B_{i+1})\}, i=0,1,2,...,7 \qquad B_i=1/2^{8-i}, i=1,...,7, 8; \quad B_0=0$$

其中 *Area(Rj)* 是颜色区域 *Rj* 的面积百分比,*Center(Rj)* 是 *Rj* 的中心,*Block(i)* 是第 *i* 个小块(图像被均匀地分为 16 个小块).*σx(Rj)* 是 *Rj* 在 X 方向上的标准方差并用图像宽度作归一化.*Width(Rj)* 是 *Rj* 的最小包含矩形(MBR, minimum bounding rectangle)的宽度除以图像宽度.*Hvy* 和 *Hsy* 的定义分别类似于 *Hvx*, *Hsx*,只不过是用图像高度作归一化.

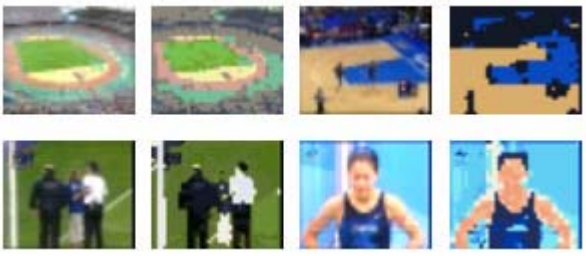

Fig.2 Color-blob images after color quantization
图 2 颜色量化后得到的几幅颜色块图

面积直方图描述了图像的空间复杂度.位置直方图有利于确认相似的空间格局,比如一个对运动员头部放大聚焦的镜头.方差直方图和长宽直方图表示了颜色块图的区域形状分布.实验结果发现长宽直方图的表现不如其它几个直方图那么稳定,因为它不是旋转不变的.因此在我们实验中并没有用长宽直方图.关于两幅图像 a 和 b 的空间相似性定义为

$$SshSim(a,b)=W0\times Sim0+W1\times Sim1+W2\times Sim2+W3\times Sim3$$

其中 *Sim0*, *Sim1*, *Sim2*, 和 *Sim3* 是分别对 *Harea*, *Hpos*, *Hvx*, 和 *Hvy* 用直方图交计算得到的直方图相似度.*W0*, *W1*, *W2*,和 *W3* 是对应的权重,在我们实验中它们被均匀地设为 0.25.

2.3 子镜头提取与镜头相似性度量

对一个有重要内容变化的镜头,比如从室内转移到室外,最好将它切分为几个内容一致的子镜头,因为如果我们把所有的内容变化合成到一个特征矢量上,结果将是无法想象的.考虑到提取出的子镜头应该是视觉内容一致的,我们提出一种简单的基于颜色和空间结构变化的子镜头提取方法.假设新出现的主颜色对象所

占的百分比是 *P1*,在当前 I 帧和上一个 I 帧的空间结构直方图的差异是 *P2*.当如下条件都同时满足时,我们就认为新的子镜头出现:

P1 > T1, P2 > T2, P1 + P2 > T3

其中 *T1*, *T2*,和 *T3* 是预先定义的阈值(在实验中它们被经验性地设为 0.2, 0.2, 0.6).两个镜头 *a* 和 *b* 之间的相似性度量定义为

$$Sim(a,b)=\max_{i,j}(Sim(a_i,b_j))$$

其中 *Sim(ai,bj)* 是镜头 *a* 中的子镜头 *i* 和镜头 *b* 中的子镜头 *j* 的相似度,可以按如下两种方式（加权和方式以及乘积方式）计算：

$$Sim(a_i,b_j)=Wc\times DchSim(a_i,b_j)+Ws\times SshSim(a_i,b_j) \quad (1)$$

$$Sim(a_i,b_j)=DchSim(a_i,b_j)\times SshSim(a_i,b_j) \quad (2)$$

*DchSim(.)*是两个子镜头主颜色直方图之间的相似度,*SshSim(.)*是两个子镜头平均的空间结构直方图之间的相似度.式子(1)中的 *Wc* 和 *Ws* 是对应的权重,在实验中设为相等的 0.5.

3 场景结构提取

用上述定义的镜头相似性度量,可以通过分裂和合并力量竞争的方法,将一系列连续的镜头归纳为一个场景结构.每个镜头受到两种力的作用.一种力是来自前后两个方向的分裂力.如果这两个方向的分裂力的比值较大,那么这个镜头很可能处在一个新的场景的边界上.另一种力是合并力,它阻止当前镜头被单方向的分裂力所吸引.一种简单的阈值算法用于检测场景边界.

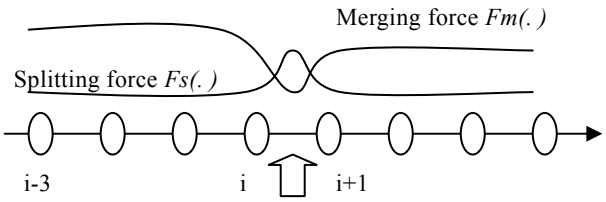

Fig.3 Scene boundary detection by force competition

图 3 用分裂和合并的力量竞争进行场景边界检测

考虑到时间约束, 即相邻的镜头更加可能属于同一个场景, 两个镜头的相似性度量按照时间吸引因子 *w =1/(1+d/C)* 进行加权, 其中 *d* 是两个镜头的最小时间差距(即从前面镜头的结束帧到后面镜头的开始帧的时间), *C* 是一个常数(实验中设为 20 秒钟).

在图 3 中每个圆圈代表一个镜头, 画向上箭头的地方是当前考察的候选场景边界*(i|i+1)*. 定义一个镜头 *i* 受到的分裂力为 *Fs(i) = left(i) / right(i)*, 其中

left(i) = max{sim(i,i-1), sim(i,i-2), sim(i,i-3)}

right(i) = max{ sim(i,i+1), sim(i,i+2), sim(i,i+3)}

当前候选场景边界*(i|i+1)*的分裂力定义为 *Fs(i|i+1) = (Fs(i)+1/Fs(i+1))/2*, 它的物理意义是使镜头 *i* 和镜头 *i+1* 分裂的平均力量. 定义镜头 *i* 的受到的合并力为(其物理意义是跨过镜头 *i* 的平均吸引力)：

$$\mathrm{Fm(i)}=\frac{1}{3}\sum_{j=i-3}^{i-1}\max(sim(j,i+1),sim(j,i+2),sim(j,i+3))$$

当前可能的场景边界*(i|i+1)*的合并力定义为

Fm(i|i+1) = (Fm(i)+Fm(i+1))/2

再分别对 *Fs(i|i+1)* 和 *Fm(i|i+1)* 进行如下基于单位方差的线性伸缩归一化（从 *x* 映射到 *x'*）：

$$x'=\frac{\frac{x-\mu}{3\sigma}+1}{2}$$

得到 *Fs'(i|i+1)* 和 *Fm'(i|i+1)*, 它们的取值范围在[0, 1]之间的概率将大于 99%.

一个理想的场景边界*(i|i+1)*应该使 *Fs'(i|i+1)*达到极大值并且同时使 *Fm'(i|i+1)*达到极小值.当如下条件(1)或条件(2)满足时,我们认为镜头边界 *(i|i+1)* 是一个场景边界:

(1) *Fs'(i|i+1)*达到极大值而同时 *Fm'(i|i+1)*达到极小值,且 *Fs'(i|i+1)>T1*.

(2) *Fs'(i|i+1)*达到极大值,或 *Fm'(i|i+1)*达到极小值,但是要求 *Fs'(i|i+1)>T2* 并且 *Fs'(i|i+1) - Fm'(i|i+1) > T3*.

其中 *T1*,*T2*,*T3* 为预定义阈值（实验中分别设为 0.6,0.7,0.2）,判断极值点时的领域设为[-2,2].

4 实验结果

4.1 镜头检索

镜头检索的实验数据是从电视录制的二十段体育节目,总共有 110 分钟,815 个镜头,163880 帧图像.这个视频数据库非常具有挑战性,因为它包含多种体育项目,如各种球类运动,田径,跳水,拳击等.图 4 是实验程序的用户界面.我们挑选了 8 个语义类作为查询镜头,分别是:篮球比赛远景,网球远景,跳水,游泳,运动员头肩像的特写镜头,观众,水下拍摄的游泳镜头,田径体育场(实例见图 5).对每个类随机挑选出 20 个镜头作为查询样本.我们测试了以下八种不同的镜头表示与相视性度量的方法做实验比较:

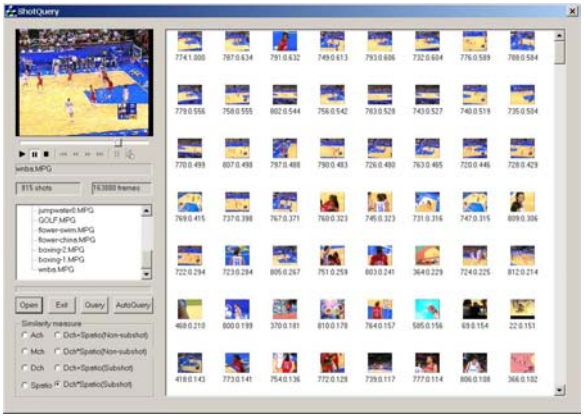

Fig.4 User interface for shot retrieval experiments

图 4 镜头检索实验的用户界面

Fig.5 Sample images for eight semantic video classes

图 5 八个语义类的镜头图像帧例子

I. 平均颜色直方图;

II. 中值颜色直方图;

III. 主颜色直方图;

IV. 空间结构直方图;

V. 主颜色直方图和空间结构直方图相似度的加权和;

VI. 主颜色直方图和空间结构直方图相似度的乘积;

VII. 主颜色直方图和空间结构直方图相似度的加权和,以及子镜头提取;

VIII. 主颜色直方图和空间结构直方图相似度的乘积,以及子镜头提取.

实验采用了两种在 MPEG-7 标准化活动中的评价指标:平均归一化调整后的检索秩 ANMRR(average normalized modified retrieval rank),和平均查全率 AR(average recall)(细节见本文附录).ANMRR 值越小,意味着检索得到的同一类镜头的排名越靠前;AR 值越大,意味着在前 K(K 是检索结果的截断值)个查询结果中相关镜头占所有相关镜头的比例越大.表 1 和表 2 分别是 ANMRR 和 AR 对不同镜头类(从 0 到 7)用不同方法(从 I 到 VIII)的实验结果.整体来看,几种方法中单独使用空间结构直方图的表现最差,因为在一个体育镜头中空间结构变化比颜色变化要大得多,而且没有考虑子镜头提取.在三种颜色直方图方法中,主颜色直方图的表现稍微比平均颜色直方图和中值颜色直方图好一些,它们三者的 ANMRR 值几乎一样,而主颜色直方图的 AR 值比其它两者高出 1%.实验证实,对于那些有对象跟踪或者内容变化较大的镜头,主颜色直方图的检索效果要比其它两种好得多.在方法 V, VI, VII, VIII 中,基于加权和的相似性度量明显不如乘积形式效果好,说明对于主颜色直方图和空间结构直方图而言,加权和不是一种好的融合方式.在八种方法中,方法 VIII 同时在 ANMRR 和 AR 两种度量上取得了最佳

表现,说明本文提出的镜头分析和表示方法在镜头检索中的有效性.

Table 1 AR values for different shot classes (0,1,…,7) with different methods (I,II,…,VIII)

表 1 对不同镜头类(从 0 到 7)用不同方法(从 I 到 VIII)实验得到的 AR 值

	I	II	III	IV	V	VI	VII	VIII
0	0.3951	0.3996	0.3916	0.3637	0.3925	0.4018	0.3850	0.3965
1	0.4417	0.4514	0.4556	0.2694	0.4722	0.4889	0.4708	0.4792
2	0.4967	0.4967	0.5025	0.5475	0.5558	0.5125	0.5667	0.5150
3	0.8485	0.8515	0.8758	0.4727	0.7742	0.8894	0.7682	0.8939
4	0.4737	0.5158	0.5000	0.2553	0.4053	0.5000	0.4211	0.4947
5	0.8227	0.7818	0.7818	0.6409	0.7818	0.7909	0.8318	0.8636
6	0.7387	0.7403	0.8016	0.3177	0.7145	0.7919	0.7129	0.7823
7	1.0000	1.0000	1.0000	0.7000	1.0000	1.0000	1.0000	1.0000
AR	*0.7453*	*0.7482*	*0.7584*	*0.5096*	*0.7281*	*0.7679*	*0.7366*	*0.7750*

Table 2 ANMRR values for different shot classes (0,1,…,7) with different methods (I,II,…,VIII)

表 2 对不同镜头类（从 0 到 7）用不同方法（从 I 到 VIII）实验得到的 ANMRR 值

	I	II	III	IV	V	VI	VII	VIII
0	0.6919	0.7046	0.7093	0.7405	0.7066	0.7031	0.7097	0.7075
1	0.6925	0.6780	0.6930	0.8145	0.6882	0.6810	0.6843	0.6801
2	0.4691	0.4724	0.4690	0.5687	0.4558	0.4711	0.4502	0.4696
3	0.3023	0.2970	0.2932	0.6853	0.3477	0.2629	0.3274	0.2516
4	0.6276	0.5903	0.6163	0.8230	0.6699	0.6214	0.6620	0.6212
5	0.2526	0.2514	0.2706	0.4552	0.2521	0.2539	0.2650	0.2620
6	0.3361	0.3401	0.2945	0.7630	0.3590	0.3000	0.3517	0.3029
7	0.0142	0.0108	0.0128	0.4357	0.0057	0.0037	0.0060	0.0045
ANMRR	*0.4838*	*0.4778*	*0.4800*	*0.7552*	*0.4978*	*0.4710*	*0.4937*	*0.4714*

4.2 场景结构提取

实验数据是两段 MPEG-7 测试视频 lgerca_lisa_1.mpg 和 lgerca_lisa_2.mpg.它们都是 32000 帧左右的家庭录像.在表 3 和表 4 是两段视频主观确认的场景以及实验结果.图 5 和图 6 是两段视频的分裂力与合并力示意图,画竖线的地方是主观确认的场景边界.

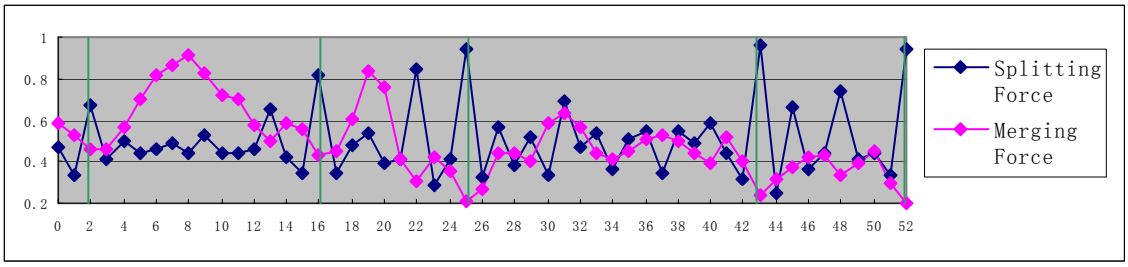

Fig.6 Splitting and merging forces for lgerca_lisa_1.mpg

图 6 lgerca_lisa_1.mpg 分裂力与合并力，画竖线的地方为主观确认的场景边界

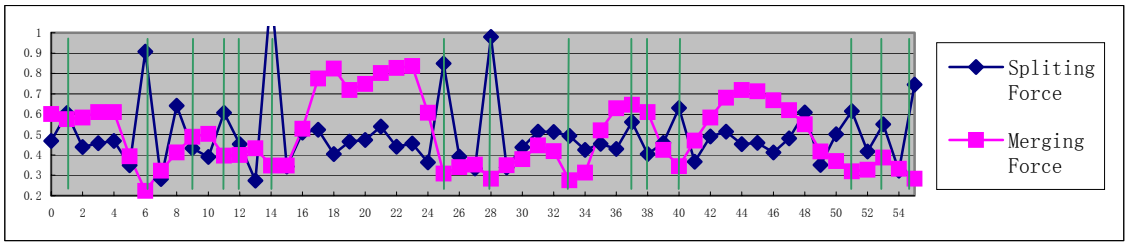

Fig.7 Splitting and merging forces for lgerca_lisa_2.mpg

图 7 lgerca_lisa_2.mpg 分裂力与合并力，画竖线的地方为主观确认的场景边界

可以看出,在大多数的场景边界处,分裂力达到极大值并且合并力达到极小值.lgerca_lisa_1.mpg 中的 3 个假的场景边界是由于拍摄角度或者光线差异造成的.lgerca_lisa_2.mpg 中漏掉的 5 个场景边界是有争议的,因为场景是一个主观概念,它的选取还要同时考虑尺度因素(即多大的时空范围作为一个场景才比较合理).例如,场景 1 至 5 都是在体育馆拍摄的,场景 9 至 11 拍摄的都是舞台演出,场景 13 和 14 都是在游泳池拍摄的.总体说来,对于场景这样抽象的语义概念能获得如此效果,还是令人十分兴奋的.

Table 3 Experimental results for scene structures in lgerca_lisa_1.mpg

表 3 lgerca_lisa_1.mpg 的场景结构提取结果，"\"表示无定义

Scene	Scene Description	Shots	Correct	Missed	False
0	Kids learing roller-skater	0-1	1|2		
1	Kids playing in gym	2-15	15|16		12|13
2	Kid playing with water with parents	16-24	24|25		21|22
3	Hot balloon event	25-42	42|43		
4	Kids playing with parent on lawn	43-51	51|52		47|48
5	Indoor exercise	52	\	\	\

Table 4 Experimental results for scene structures in lgerca_lisa_2.mpg

表 4 lgerca_lisa_2.mpg 的场景结构提取结果，"\"表示无定义

Scene	Scene Description	Shots	Correct	Missed	False
0	Kids at home with cat	0	0|1		
1	Kids in gym	1-5	5|6		
2	Two kids playing high-bar (Over-illuminated)	6-8		8|9	7|8
3	Kid + teacher with high-bar	9-10	10|11		
4	Kids jumping	11		11|12	
5	Kids in Gym	12-13	13|14		
6	Kids playing games in Gym (Over-illuminated)	14-24	24|25		
7	Kids playing at home (Dim lighting)	25-27	27|28		
8	Kids driving outside home	28-32	32|33		
9	Kids dancing on stage (Part I of Play)	33-36		36|37	
10	(Part II of Play)	37		37|38	
11	(Part III of Play)	38-39	39|40		
12	After Play	40-50	50|51		
13	Swimming Pool	51-52		52|53	
14	Crowded Swimming Pool	53-54	54|55		
15	Kid's party	55	\	\	\

5 结论

本文提出了一种新的镜头内容分析方法,及其在镜头检索和场景提取中的应用.我们讨论了子镜头提取的必要性,提出了子镜头提取的具体算法,引入了两个新的内容描述子用于度量镜头内容变化和表示子镜头:主颜色直方图和空间结构直方图.通过对八种不同的镜头表示和相视性度量方法做比较,本文提出的方法(用主颜色直方图和空间结构直方图表示子镜头)按照 ANMRR 和 AR 两种指标取得了最好的镜头检索表现.本文还提出了基于分裂与合并力量竞争的镜头分组方法用于提取场景结构,在两段 MPEG-7 测试视频上进行的场景提取实验取得了较为满意的结果.我们未来的工作是提出新的运动描述子,以便能够把运动信息与颜色以及空间信息融合起来,并研究新的信息融合方式,进一步提高镜头分析算法在镜头检索和场景提取中的表现.

References:

[1] Rui Y. and Huang T. S., A Uniform Framework for Video Browsing and Retrieval, The Image and Video Processing Handbook, edited by Alan Bovik, Academic Press, 2000, 705~715.

[2] Ngo C. W., Pong T. C., Zhang H. J., and Chin R. T., Motion-based video representation for scene change detection, ICPR'00, Barcelona, Spain, 2000.

[3] Irani M. and Anandan P., Video indexing based on mosaic representations, Proceedings of IEEE, 1998, 86: 905~921.

[4] Zhao L., Qi W., Li S. Z., Yang S. Q., and Zhang H. J., Key-frame extraction and shot retrieval using Nearest Feature Line (NFL), International Workshop on Multimedia Information Retrieval, in conjunction with ACM Multimedia Conference 2000, Los Angeles, USA, 2000.

[5] Hanjalic A., Lagendijk R. L., and Biemond J., Automated high-level movie segmentation for advanced video-retrieval systems, IEEE Transactions on Circuits and Systems For Video Technology, 1999, 9(4): 580~588.

[6] Corridoni J. M. and Bimbo A. Del, Structured representation and automatic indexing of movie information content, Pattern Recognition, 1998, 31(12): 2027~2045.

[7] Rui Y., Huang T. S., and Mehrotra S., Exploring video structure beyond the shots, Proc. IEEE Conf. on Multimedia Computing and Systems, 1998. 237~240.

[8] Kender J. R. and Yeo B. L., Video scene segmentation via continuous video coherence, Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 1998. 367~373.

[9] Ferman A. M., Krishnamachari S., Tekalp A. M., Mottaleb M. A., and Mehrotra R., Group-of-frames/pictures color histogram descriptors for multimedia applications, ICIP2000, 2000.

[10] Lin T. and Zhang H. J., Automatic video scene extraction by shot grouping, ICPR'00, Barcelona, Spain, 2000.

[11] Ma W. Y. and Zhang H. J., Content-based image indexing and retrieval, in Handbook of Multimedia Computing, Borko Furht ed. CRC Press, 1998.

[12] Zhou X. S., Rui Y., and Huang T. S., Water-Filling: a novel way for image structural feature extraction, ICIP'99, 1999.

附录:

平均归一化调整后的检索秩 ANMRR(average normalized modified retrieval rank)和平均查全率 AR(average recall)的计算细节如下.首先挑选出一个查询镜头的集合 Q,对每个查询镜头主观地选取一组视觉相似的镜头作为标准正确答案(ground truth).设查询镜头 q 的相似镜头的个数为 $ng(q)$.对于查询镜头 q,检索结果的截断值 K 定义为 $min\{4\times ng(q), 2\times GTM\}$,其中 GTM 是在所有查询镜头中最大的相似镜头个数,即 $GTM = max\{ng(q)\}$.对于查询镜头 q,在前 K 个检索结果中正确的个数记为 $nr(q)$,漏掉的个数记为 $M(q) = ng(q) - nr(q)$.查全率记为 $R(q) = nr(q)/ng(q)$.每个正确答案在检索结果中都有一个秩(rank) $r(i)$, $I = 1,...,ng(q)$.在前 K 个检索结果中正确的镜头的秩 $r(i)$就是它的序号,其余被漏掉的镜头的秩 $r(i)$都设定为 $K+1$.对于某个查询镜头 q,它的平均检索秩和调整后的检索秩分别定义为

$$ARR(q) = \sum_{i=1}^{ng(q)} \frac{r(i)}{ng(q)}, \quad MRR(q) = ARR(q) - \frac{ng(q)}{2} - 0.5$$

将 $MRR(q)$归一化至 [0,1] 范围内,得到归一化调整后的检索秩 $NMRR(q)$:

$$NMRR(q) = \frac{MRR(q)}{K - \frac{ng(q)}{2} + 0.5}$$

对 Q 中所有的查询镜头 q 的 $NMRR(q)$和 $R(q)$作平均,得到 $ANMRR$ 和 AR:

$$ANMRR = \frac{1}{Q}\sum_{q=1}^{Q} NMRR(q), \quad AR = \frac{1}{Q}\sum_{q=1}^{Q} R(q)$$

Shot Content Analysis for Video Retrieval Applications

LIN Tong[1], ZHANG Hong-jiang[2], FENG Ju-fu[1], SHI Qing-yun[1]

[1](*National Laboratory on Machine Perception, Center for Information Science, Beijing University, Beijing* 100871, *China*);

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

Abstract: In this paper we present a novel scheme on shot content analysis for two video retrieval applications: shot retrieval and scene structure extraction. To characterize the temporal content variations in one

shot, we developed two descriptors: Dominant Color Histograms and Spatial Structure Histograms. By fusing temporal information into color content, Dominant Color Histograms for a group of frames are trying to capture the dominant colors with longer durations, which would be the colors of the focused objects or background. Spatial Structure Histograms is a set of features extracted from color-blob maps to describe spatial information for one individual frame. A shot with significant content changes can be segmented into several subshots that are of coherent content, and shot similarity measure can be computed from the similarity between corresponding subshots. Scene structure is extracted by analyzing the competition of splitting and merging forces. Experimental results on real-world sports video prove that our proposed approaches achieve the best performance on shot retrievals and promising results on scene structure extraction.

Key words: Content-based Video Retrieval; shot content analysis; shot similarity measure; scene structure extraction

* Received January 8, 2001; accepted May 9, 2001
Supported by the