一种层次视频摘要生成方法

程文刚 须 德
(北方交通大学计算机与信息技术学院计算机研究所,北京 100044)

摘 要 视频摘要是视频内容的一种压缩表示方式。为了能够更好地浏览视频,提出了一种根据浏览或检索的粒度不同来建立两种层次视频摘要(镜头级和场景级)的思想,并给出了一种视频摘要生成方法:首先用一种根据内容变化自动提取镜头内关键帧的方法来实现关键帧的提取;继而用一种改进的时间自适应算法通过镜头的组合来得到场景;最后在场景级用最小生成树方法提取代表帧。由于关键帧和代表帧分别代表了它们所在镜头和场景的主要内容,因此它们的序列就构成了视频总结。一些电影视频片段检验的实验结果表明,这种生成方法能够较好地提供粗细两种粒度的视频内容总结。

关键词 视频摘要 视频总结 视频预览 关键帧 代表帧
中图法分类号: TN941.1 **文献标识码**: A **文章编号**: 1006-8961(2004)01-0118-06

A Novel Approach of Generating Hierarchical Video Abstraction

CHENG Wen-gang, XU De
(*Institute of Computer, School of Computer & Information Technology, Northern Jiaotong University, Beijing* 100044)

Abstract Video abstraction is a short summary of the content of a longer video document, which helps to enable a quick browsing of a large collection of video data and to achieve efficient content access and representation. In this paper, we propose a novel approach of generating video abstraction at two levels, namely, the shot level and the scene level. The hierarchical video abstraction can facilitate video browsing and retrieval at different granularities. Firstly, the video stream is segmented into shots. A new key frame extraction algorithm, which does not rely on threshold, is put up to extract key frames from shots based on the content variation. An updated time-adaptive algorithm is used to group the shots into scene. Based on the defined shot similarity formula, the video scene structure is constructed after shot clustering and adjustment. Representative frames are extracted from the key frames of each scene using the method of generating Minimum Spanning Tree. Key frames and representative frames can represent the content of shot and scene, respectively. The sequence of key frames and representative frames is the two-level video abstraction. Experiments based on real-world movies show that the method above can provide users with better video summary at different levels.

Keywords Video abstraction, Video summary, Video skimming, Key frame, Representative frame

1 引 言

随着多媒体技术、数字电视和网络的发展,产生了大量的视频文档,如何对其进行管理和利用就成为迫切需要解决的问题,而基于内容的视频存取则成了对其进行有效处理的一个基础。对于一个视频文档,如电影故事片,用户可能需要在短时间内,无须浏览整部影片,而仅通过了解其主要内容,即能决定是否值得进行详细的观赏,因此如何快速地浏览视频,并且了解其主要内容已经成为一个重要的研究课题。视频摘要(Video Abstraction),是解决这种问题的一个途径。由于视频摘要同时也能辅助建立视频检索和索引系统,因此对于视频摘要的研究有着重要的意义。

所谓视频摘要就是对长时间视频文档的简短的

收稿日期:2002-11-25;改回日期:2003-07-15

内容总结,在表现形式上,它是一个静止或者运动的图片序列。视频摘要可以通过人为地选取视频中具有代表性的图片序列来产生,但是由于视频数量巨大和手工建立的低效,尽量减少人为的干预,以实现自动建立就成为一种必然的趋势。通常有两种基本形式的视频摘要:视频总结(Video Summary)和视频预览(Video Skimming)[1],其中,视频总结是从视频中提取的一个静止图片(帧)的集合;而视频预览则是从原始视频序列中提取的由一些简短的视频片段组成的集合,它是动态的,也可能包含相应的音频和文本信息;由于视频总结需要通过提取一些最能反映视频内容的帧(关键帧)来组成,因此提取和生成这些帧就成为系统建立的核心;视频预览建立的目的或者是为了给用户以整个原始视频的内容的简单印象(Summary Sequence),或者是为了给出其中的一些精彩片段(Highlight)[2]。

因为这两种形式的视频摘要具有不同的特点,所以建立它们所使用的算法也相应地有所不同。鉴于视频总结建立过程快速简单,并且一旦建立就可以很方便地显示和组织,因此目前在这方面的研究比较多。由于关键帧能反映视频的内容,相应的,关键帧的提取算法研究就成了中心内容。如今已经有很多自动提取关键帧的方法。其中,最简单的方法是在每隔一定时间抽取一帧作为关键帧,其结果是一些短小,但比较重要的片段可能没有相应的关键帧,相反一些比较长的片段则可能有很多相似内容的帧。由于镜头是一个连续的拍摄过程,故选取镜头中的第 1 帧作为关键帧就成为一种比较直接的方法,但是由于这种方法有时不能反映镜头内容的变化,于是产生了选取一个镜头内的多帧作为关键帧的算法,其基本思想是先选取第 1 帧作为关键帧,而将其后的帧按顺序与这一帧进行比较(用一些特征之间距离来度量),当第 k 帧与前一关键帧的差超过一个阈值 T 后,则第 k 帧成为新的关键帧,后面的帧依次与新的关键帧进行比较,直至本镜头的最后一帧为止。这种算法还产生了一些变种,其主要区别在距离的度量上。另外一些算法则采用镜头内帧的聚类来选取关键帧,但这些算法一个共同的特点是需要设置阈值。由于阈值的设定一般来说比较困难,而不同的阈值产生的关键帧数目又不同,且阈值随不同类型的视频而变化,同时一个固定的阈值也很难适应于同一视频中的不同片段,因此阈值直接关系着提取的效果。

文献[3]和文献[4]中都提出了不依靠阈值的提取算法。在文献[3]算法中,就不需要进行镜头探测,其视频序列由单元组成(每 L 帧组成一个单元),而单元变化则用单元内的尾帧和首帧的特征距离来衡量,并根据一个比率 r 将单元变化分成如下两类:变化大的一类和变化小的一类。提取关键帧时,抛弃后者,而将前一类中的所有帧作为关键帧,如果得到的关键帧数目多于设定的数目 N,则将所有的关键帧再重新组织进行新一轮的分类,直至满足要求为止。这种方法的一个缺陷是单元变化的度量问题,如果单元内的首尾帧恰好具有相似的特征(如颜色分布),那么即使单元内部变化很大也检测不出来,另外,分成两类过于粗糙。在文献[4]算法中,则首先设定所有关键帧的数目,并根据镜头内容变化与所有镜头内容变化累计值的比值来得到本镜头应该获取的关键帧数目;然后提取关键帧。该算法的一个缺点是描述镜头内容的变化使用了累计的帧间差,这样即使一个帧间变化比较小,但比较长的镜头也有可能分配给较多的关键帧数目,这显然是不合理的。

另外,还有一些利用镶嵌图案(mosaic)和基于视频对象(Video Object)等关键帧生成算法来产生视频总结的方法。

在生成视频预览方面,由于判断视频中的精彩片段是一个非常主观的问题,并且把人的认知转化为计算机自动提取的过程是很困难的,因此目前的视频预览研究集中在产生一些视频内容的简单总结上。如今有的视频预览原型系统仅建立在快速回放的基础上,而有的则综合利用声音、文字、视觉信息来建立,如文献[5]中提出了一种综合利用语言和图片理解来提取重要声音和视频信息的方法。

为了能够提供不同粒度的视频信息,更好地满足用户浏览和检索的需求,本文提出了一种建立层次视频摘要的思想,并给出了一种构造层次视频总结的方法。

2　层次视频摘要

2.1　层次视频摘要思想

目前的视频摘要基本上都是基于镜头的,这种视频摘要虽然提供了视频内容的详细描述,然而由于有时某个镜头并不是用户所关心的单元,而用户可能更关心高层的语义单元,如场景;同时,由于一个视频序列中可能包含太多的镜头,如一般一部电

影中都有几千个镜头，这样将产生太多的关键帧，这对于用户来说，信息的粒度可能太细，因此本文根据视频内容的层次性，提出了一种建立层次视频摘要的思想，即在镜头级和场景级，分别用关键帧和代表帧表示它们的主要内容，并首先通过镜头检测，将视频序列划分为一个个镜头，在镜头级进行关键帧提取；然后根据提取的关键帧将时间相近镜头进行组合来得到场景，再在场景中提取代表帧；这种两级视频摘要分别由经过处理的关键帧和代表帧的序列构成，这就是层次性的视频摘要。由于层次性的视频摘要能够提供给用户不同粒度的信息，且可以更好地描述视频的内容，从而方便了用户的浏览和检索。图 1 表示了这种结构。

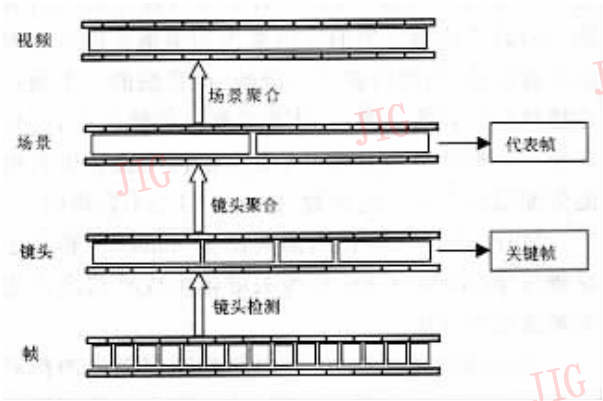

图 1 层次视频摘要

2.2 层次视频摘要生成方法

2.2.1 镜头边界检测

目前，自动镜头边界检测技术可以分为如下 5 类：基于像素、基于统计、基于变换、基于特征和基于直方图的方法。研究表明，其中基于直方图的方法能在精确度和速度的综合上达到更好的性能[6]。

通常描述视频的视觉信息可以用颜色、纹理、形状以及它们的组合来表示，本文选用颜色直方图表示。定义第 k 帧与第 $k-1$ 帧的帧间差为

$$d_k = D(H(k), H(k-1)) \tag{1}$$

式中，D 是一种距离度量，本文实验中使用简单的街区距离(city-block distance)；$H(k)$ 表示镜头内第 k 帧的归一化颜色直方图。

本文采用基于颜色直方图的方法来检测镜头边界，并在镜头边界检测的同时，记录各帧与其前一帧的帧间差。对于渐变镜头，则去掉渐变过渡帧，并将其作为镜头实际包含的帧。

2.2.2 关键帧提取算法

针对现有算法的一些缺点，提出了一种无需设置阈值，只需预先设定一个总的关键帧数目，然后根据内容变化来抽取关键帧的算法。该算法对每个镜头选取的关键帧的数目视其内容变化的程度而定。

对于镜头 K_i 内容的变化，可使用在一镜头 K_i 内的平均帧间差 $\overline{d}_{K_i}$ 来衡量

$$\overline{d}_{K_i} = \left(\sum_{k=2}^{N_{K_i}} d_k\right) / N_{K_i} \tag{2}$$

式中，N_{K_i} 表示第 i 个镜头 K_i 内的帧的数目。

由于平均帧间差 $\overline{d}_{K_i}$ 反映了镜头内容变化的程度，因此可以根据 $\overline{d}_{K_i}$ 的大小在所有镜头的累计平均帧间差中所占的比重来确定其所需要选择的关键帧数目。这样就避免了某个特别长，但内容变化不大(如特写镜头)的镜头，其选取的关键帧数目过多的问题。

设整个视频序列共有 s 个镜头，对它们逐个计算平均帧间差，计算结果就形成了一个长度为 s 个镜头的一个队列。图 2 为一个由 6 个镜头组成的视频片段的平均帧间差变化曲线示例，显然镜头 K_4 内容变化最大，K_1 内容变化最小，它们相应的关键帧数目也应该不同。

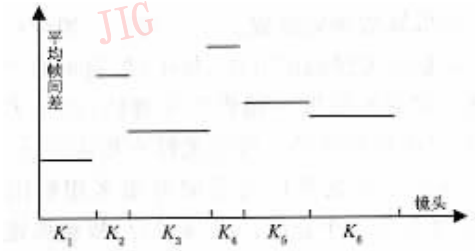

图 2 视频片段镜头的平均帧间差变化示意图

给定总的关键帧数目 N，这样任意一个镜头 K_i 应该分配到的关键帧数目 $N_{\text{keyframe}}(K_i)$ 为

$$N_{\text{keyframe}}(K_i) = \max(1, N \times \overline{d}_{K_i} / \sum_{j=1}^{s} \overline{d}_{K_j}) \tag{3}$$

式中，max 函数是为了保证每一个镜头至少有一个关键帧。

这样，就可以根据指定的镜头关键帧数目 $N_{\text{keyframe}}(K_i)$ 来提取任意一个镜头 K_i 的关键帧了。由于当内容变化不大时，帧间差比较小，相应的，帧间差较大，则说明内容变化较大，即相邻两帧表示的内容有较大的不同，因此，关键帧选取的原则是：选取帧间差最大的 $N_{\text{keyframe}}(K_i)$ 个 d_k，其所对应的帧即为关键帧。图 3 为一个由 10 帧组成的镜头的关键帧选取(选 2 个)示例，由图 3 可见，根据帧间差的大小，将第 2 帧和第 7 帧选为关键帧。

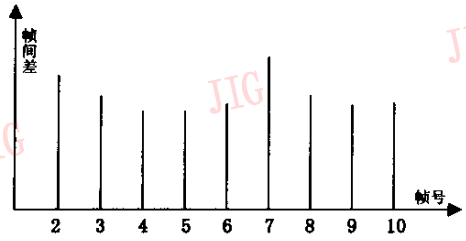

图 3　镜头内帧间差变化示意图

2.2.3　场景生成算法

按照电影编辑的原理,语义相关、视觉相似以及时间相近的镜头组合成场景。文献[7]中提出了一种简单有效的基于时间自适应的方法(time-adaptive grouping approach)来构造视频场景,它取每个镜头简单的首尾两帧作为关键帧,聚类过程中,镜头先组合成组(Group),然后再形成场景。本文对文献[7]的算法进行了改进,本文算法核心是定义镜头之间的相似度。

任意两帧 F_x 和 F_y 之间的相似度定义为

$$S(F_x,F_y)=1-D(H(x),H(y)) \tag{4}$$

根据 2.2.2 节中的算法,一个镜头内,由于已经取出其中的 $N_{\text{keyframe}}(K_i)$ 帧作为关键帧,因此可以这样表示任意一个镜头 K_i

$$K_i=K(o_1,o_2,\cdots,o_{N_{\text{keyframe}}(K_i)},\overline{d}_{K_i},H(o_1),H(o_2),\cdots,H(o_{N_{\text{keyframe}}(K_i)})) \tag{5}$$

式中,$\overline{d}_{K_i}$ 为镜头 K_i 的平均帧间差;o_x 为帧号,$H(o_x)$ 表示 o_x 帧的颜色直方图($x=1,2,\cdots,N_{\text{keyframe}}(K_i)$)。

对于两个任意的镜头,其关键帧之间的相似度形成一个 $N_m\times N_n$ 的距离矩阵。选取其中的一个最大值 $S(F_x,F_y)_{\max}$,设对应的帧号分别为 o_x 和 o_y,则这两个镜头之间的内容相似度可以定义为

$$S(C_{K_i},C_{K_j})=S(T_x,T_y)\times S(F_x,F_y)_{\max} \tag{6}$$

为了描述两个镜头的内容变化情况大小,定义它们的运动相似度为

$$S(M_{K_i},M_{K_j})=S(T_x,T_y)\times\left|\overline{d}_{K_i}-\overline{d}_{K_j}\right| \tag{7}$$

在公式(6)和(7)中,$S(T_x,T_y)$是两个镜头之间的时间相似度,其随相距距离增大而递减。定义

$$S(T_x,T_y)=\max\left(0,1-\frac{|o_x-o_y|}{c\times\overline{L}}\right) \tag{8}$$

式中,c 是一个常数,$\overline{L}$ 是整个视频中镜头的平均长度,以帧为单位。

因此镜头 K_i 与镜头 K_j 的相似度为

$$S(K_i,K_j)=W_1\times\hat{S}(C_{K_i},C_{K_j})+W_2\times\hat{S}(M_{K_i},M_{K_j}) \tag{9}$$

式中,W_1 和 W_2 分别是内容相似度和运动相似度的权重;$\hat{S}(C_{K_i},C_{K_j})$和 $\hat{S}(M_{K_i},M_{K_j})$分别是 $S(C_{K_i},C_{K_j})$和 $S(M_{K_i},M_{K_j})$的归一化值,采用下式归一化(高斯归一化)

$$\hat{V}_i=\frac{\left(\frac{V_i-\mu}{3\sigma}+1\right)}{2}\quad(i=1,\cdots,n) \tag{10}$$

式中,V_i 表示数值序列中第 i 个值;μ 是数值序列的均值;σ 是其标准差。

经过归一化后,镜头内容相似度和运动相似度的值都落在同一区间[0,1]中。W_1 和 W_2 是两者相应的权重,表明了它们各自的重要性。如果一个数值序列的标准差很小,则表明这个数值序列的值都相差不大,因此这个数值序列的分辨能力也较差;反之则具有很强的分辨能力,即应该具有较大的权重。为了减少参数的设定,可从镜头内容相似度和运动相似度的值的统计特征来分析,并采用下式来确定式(9)中的两个参数

$$W_i=\frac{\sigma_i}{\sigma_1+\sigma_2}\quad(i=1,2) \tag{11}$$

式中,σ_1 和 σ_2 分别表示镜头内容相似度和运动相似度的标准差。

运用式(9)中的相似度来计算样本间距离,其中类间距离使用最大距离,并对镜头进行层次聚类[8]。这里设定聚类结束的阈值为 T,输入为镜头序列 $\boldsymbol{K}=\{K_1,K_2,\cdots,K_M\}$,输出(聚类结果)为候选场景序列 $\boldsymbol{Q}=\{Q_1,Q_2,\cdots,Q_N\}$。

视频中可能会有这样的情况,即使两件事情同时发生,但是它们必须顺序地播放。如 Tom 与 Jerry 的对话片段,由于运用上面的算法可能得到 Tom 和 Jerry 分属两个不同的候选场景,因此需要对上面的结果进行进一步处理,其具体处理办法是对于时间上有交叉的(比较关键帧帧号)两个候选场景进行合并。经过处理的结果即为最终要求的场景。

2.2.4　代表帧的选取

对于任意的视频场景可以这样表示

$$\boldsymbol{Q}=(\text{QID},\boldsymbol{S},\boldsymbol{D},\boldsymbol{F}) \tag{12}$$

式中,QID 是场景的唯一标识;$\boldsymbol{S}$ 是场景中包含的镜头集合;$\boldsymbol{D}$ 是场景中镜头的平均帧间差的集合;$\boldsymbol{F}$ 是从场景包含的镜头中抽取出来的所有关键帧的集合。

代表帧选取的数目可根据预选设定的比率 $r(0<r<1)$来定,其选取的原则是内容变化最大的 n

帧，也就是选取相似距离矩阵中 n 个最小的值。如果 F 中关键帧的总数为 N_{keyframe}，则 $n=N_{\text{keyframe}}\times r$。如果把每一个关键帧当作一个顶点，则两个关键帧之间的相似度构成一条边，这样问题就转化为求无向图的最小生成树问题，若取相似矩阵中最小的一个值作为起始的一条边，则最小生成树的前 n 个节点即为所求的结果。

2.2.5 视频总结的建立

由于关键帧和代表帧分别代表了镜头和场景的主要内容，所以通过对得到的关键帧和代表帧的集合进行组织排列（如按时序排列），就可以得到两个层次的视频总结。

要建立视频浏览系统，可以对 2.2.2 节中得到的关键帧或 2.2.4 节中得到的代表帧，按内容相似度进行聚类，若每个类别选取一帧表示该类别的内容，则可通过此帧所在的镜头连接来形成视频浏览。

3 实验结果与分析

为了检验本文算法的效果，选取了几个视频片段对视频摘要的生成算法进行了检验。图 4 是《白雪公主》中白雪公主与 7 个小矮人唱歌跳舞的一个场景片段，其中包含 991 帧 8 个镜头的平均帧间差变化图。根据其内容变化设定 2.2.2 节关键帧提取算法中的 N 值为 20，其关键帧的提取如图 5 所示。如果选取的 N 值太大，则可能出现很多内容相似的关键帧，而太少，则容易错过一些主要的内容。一般可以根据视频的类型来确定 N 的值，在本文的实验中，N 的值取为视频片段中所有帧数的 2% 左右，结果比较满意。

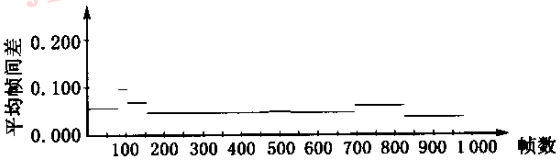

图 4 《白雪公主》片段镜头平均帧间差变化

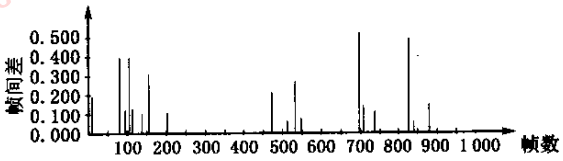

图 5 《白雪公主》片段关键提取

根据 2.2.3 节的场景生成算法进行场景构造。时间相似度的计算公式（式(8)）中，c 是经验值，由对 c 取不同值得到的实验结果进行的分析可见，当 c 取 10 时结果比较满意；层次聚类中的阈值设定为 $T=0.2$，即可得到比较好的结果。

运用场景构造算法来构造场景，上面《白雪公主》片段中的 8 镜头最后组合为一个场景，然后就可以根据 2.2.4 节中的算法进行代表帧的提取，设定比率 $r=0.25$，结果提取出 5 个代表帧被用来表示此场景的内容。这一个场景的关键帧提取和代表帧提取的结果分别如图版 I 图 1 和图 2 所示，它们就构成了此场景的两级视频摘要。

表 1 是《魂断蓝桥》等 4 个视频片段运用上述方法进行分析的实验结果。

表 1 视频片段的关键帧和代表帧提取结果

视频片段	帧数	检测镜头数	实际镜头数	候选场景数	构造场景数	实际场景数	关键帧数	代表帧数
《魂断蓝桥》片段	10 290	79	75	17	8	7	203	50
《简爱》片段	5 292	66	60	16	8	6	118	30
《一夜风流》片段	6 510	88	82	23	13	12	139	35
《走遍美国》第 1 集 Act1	4 558	55	49	11	4	4	96	24

由表 1 的数据分析可见，通过镜头检测得到的镜头数目与实际镜头数目相差不大，而差别的出现主要是由于受渐变（Gradual Transitions）的影响所致。由于关键帧能很好描述镜头内容，因此该提取算法具有良好的效果。因为镜头层次聚类后得到的聚类数目远远多于实际场景数目，所以对候选场景的合并是十分必要的，如在《走遍美国》第 1 集 Act1 实验中，Richard 与 Marilyn 的对话场景，层次聚类后得到 3 个候选场景（3 个聚类类别），但是由于 3 个候选场景在时间上交替，所以合并后则得到一个场景，这符合视频中实际要表达的语义内容。实验数据表明，场景构造结果是有效的，但是对个别视频片段，仍然存在差别。如果对取出的关键帧和代表帧分别按帧号（时间）进行排列，则通过浏览可以发现，关键帧形成的视频摘要能很详细地表达视频内容，而由代表帧形成的视频摘要则反映了视频要表达的大致内容。

4 结　论

本文提出了建立两个层次视频摘要的思想,并给出了一种建立视频总结的方法。该方法首先根据镜头内容变化自动提取关键帧,在此基础上又给出了镜头之间相似度的计算公式,这样通过聚类和调整,即可实现场景生成,然后用最小生成树方法提取代表帧。实验结果表明,这种关键帧和代表帧的提取算法是有效的,其所建立的视频总结能较好地反映原始视频的内容,而对于长时间的视频处理,算法需要较大的存储量,尚待改进;如何更好地实现场景生成以及实现视频预览,也是今后要考虑的问题。

参 考 文 献

1 Li Y, Zhang T and Tretter D. An overview of video abstraction techniques [R]. HP Laboratory Technical Report, HPL-2001-191, 2001.

2 Hanjalic A, Zhang H J. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis [J]. IEEE Transactions on Circuits and System for Video Technology, 1999, **9**(8): 1280～1289.

3 Sun X D, Kankanhalli M S. Video summarization using R-sequence[J]. Real-time Imaging, 2000, **6**(6): 449～459.

4 Hanjalic A, Lagendijk R L, Biemond J. A new method for key frame based video content representation [A]. In: Proceedings of 1st International Workshop on Image Database and Multimedia Search [C], Amsterdam, Holand, 1997:67～74.

5 Smith M A, Takeo Kanade. Video skimming and characterization through the combination of image and language understanding [A]. In: Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Databases[C], Bombay, India, 1998:61～70.

6 Zhang H J, Kankamhalli A, Smoliar S. Automatic partitioning of full-motion video [J]. Association for Computing Machinery Multimedia Systems, 1993, **1**(1): 10～28.

7 Rui Y, Thomas S. Huang and Sharad Mehrotra. Exploring video structure beyond the shots [A]. In: Proceedings of IEEE International Conference on Multimedia Computing and Systems [C], Austin, Texas, USA, 1998:237～240.

8 沈清,汤霖. 模式识别导论[M]. 长沙:国防科技大学出版社, 1991:107～108.

程文刚　1977 年生,1998 年获山东科技大学学士学位,2001 年获中国矿业大学(北京校区)硕士学位,目前在北方交通大学计算机应用专业攻读博士学位。主要研究方向为多媒体信息处理、基于内容的视频检索和机器学习。

须　德　1944 年生,1967 年毕业于中国科技大学应用数学系,1982 获北方交通大学计算机应用专业硕士学位,教授,博士生导师。主要研究方向为基于内容的视频检索、多媒体信息处理和数据库系统。