第 37 卷 第 2 期
Vol.37 No.2

计 算 机 工 程
Computer Engineering

2011 年 1 月
January 2011

• 开发研究与设计技术 •　　文章编号：1000—3428(2011)02—0278—03　　文献标识码：A　　中图分类号：TP18

基于局部-空间模型的视频摘要研究与设计

王 群，刘 群，向明辉，吴 渝

(重庆邮电大学计算机科学与技术学院网络智能研究所，重庆 400065)

摘 要：介绍视频摘要技术以及缩略视频的 3 类研究方法，提出一种基于局部-空间模型的视频摘要研究方法。通过对视频序列的帧内与帧间信息的分析得到关键帧，对提取出的关键帧进行双线性插值形成视频段，运用 DirectShow 开发软件的 DES 对视频段编辑融合成最终的动态视频摘要。实验结果表明，在不需要人工干预的情况下，自动生成的视频摘要不仅包含视频的主要信息，而且冗余信息少。

关键词：视频摘要；局部-空间模型；关键帧；互信息量

Research and Design on Video Abstraction Based on Local-Space Model

WANG Qun, LIU Qun, XIANG Ming-hui, WU Yu

(Institute of Network Intelligence, College of Computer Science and Technology,
Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

【Abstract】 This paper represents the background and the significance of video abstraction technology, as well as the three types of research method for video abbreviation. A research method is proposed for video abstraction based on local-space model. In the model, the key frame is got by the analysis of the intra-frame and inter-frame video sequence information, and these key frames are extracted to construct the video segment by double-interpolation. With the DES in the DirectShow software, the video segments are edited and fused into a final dynamic video abstraction. Experimental results show that the video abstraction is generated automatically without any article intervention, which contains the main information of the video but little redundant information.

【Key words】 video abstraction; local-space model; key-frame; mutual information

DOI: 10.3969/j.issn.1000-3428.2011.02.098

1 概述

视频摘要，即以自动或半自动的方式对视频的结构和内容进行分析，从原视频中提取出有意义的部分，并将它们以某种方式进行组合，形成简洁的能够充分表现视频语义内容的概要。视频摘要技术涉及通信、视频处理、心理感知、模式识别等领域。视频摘要的主要应用有视频数据的存档及检索、影视行业的应用、家庭娱乐业、军用及公安用途、医学影像用途、航天航空影像分析等。因此，在理论上有很大的研究意义，在实际应用上也有广阔的应用前景。更为明显的是，随着诸如 3G 无线通信等多种新型应用环境的不断涌现，网络视频在手机上传输播放已是现在多媒体手机的发展趋势，因此视频数据会大量增加，而网络传输能力还远远达不到迅速的效果，迫切需要有效的管理以满足用户的信息需求。由于视频数据具有结构复杂、语义丰富、数据类型多样等特点，传统的数据模型如关系模型、对象模型不能担此重任，需要有专门针对视频的数据模型[1]。

视频摘要可分为视频概要和缩略视频，而本文重点研究以缩略视频为代表的摘要形式。从视频摘要的生成算法来看，大致可以分为 3 大类：(1)简单的生成方法。对视频进行时段采样，即每隔一段时间抽取一个代表帧或者一个片段。这种方法比较简单，生成迅速，但完全没有基于视频的内容，效果很不可靠。(2)基于特征信息的生成方法。根据视频中颜色、纹理、形状、运动方向和强度等视觉信息，应用各种视频和图像处理技术，进行镜头探测、关键帧提取、场景聚类、运动特征提取等一系列操作，最终生成具有代表性的关键帧序列或缩略视频。这种算法完全基于视觉特征，而忽略了音频、字幕等信息对表现视频所起的作用。美国 Palo Alto 实验室生成的故事板(漫画书)的研究是其中的典型代表。(3)基于视频语义的生成方法。基于用语义分析视频摘要，首次提出情感单元的概念。情感单元表示一个情节中某一时刻人物或场景的精神状态。情感状态主要有精神状态、期望事件和非期望事件 3 种类型。国内最有代表性的是文献[2]提出的基于 EDU 模型的新闻视频摘要的研究。

文献[3]采用基于互信息量的技术来对镜头进行检测及场景切分。其方法是先计算每帧 RGB 分量的互信息量及总互信息量，然后计算滑动窗口内互信息量的均值，与给定的阈值进行比较判断，从而检测场景边界，其中阈值的选择是该算法的一大难题。而在文献[4]中运用了基于关键帧的视频摘要研究，其研究方法为模糊聚类神经网络提取关键帧的方法及线性插值法形成动态视频摘要，但是对噪音、闪光灯等外界因素的影响却不能避免，造成关键帧提取效率低下。

纵观上述的研究情况来看，目前视频摘要技术已经取得了很大的进展，但在一些关键的技术上还有待突破。为了解

基金项目：国家自然科学基金资助项目(61075019)；重庆市自然科学基金资助项目(CST2007BB2386)；重庆市科委应用基础研究基金资助项目(KJ070504)；重庆邮电大学博士启动基金资助项目

作者简介：王 群(1984－)，男，硕士研究生，主研方向：多媒体信息处理；刘 群，副教授；向明辉，硕士研究生；吴 渝，教授、博士生导师

收稿日期：2010-07-14　**E-mail：**kemmy850128@126.com

决阈值自适应问题、算法效率问题、缩略视频形成问题以及摘要丰富问题及建立对整个摘要过程起指导作用的模型，本文提出了一种基于局部-空间模型的视频摘要。采用局部信息提取分析、空间信息聚类分析以及双线性插值法实现动态视频摘要，最后运用 DES 完成摘要视频整体化。实验结果表明，该方法是可行、有效的。并以此模型为指导，完成以广告视频为例的视频摘要。

2　局部-空间模型

视频摘要追求的目标是：摘要的内容要涵盖视频的主要信息，不仅要冗余信息少，而且要尽量实现自动化或半自动化，以减少人的干预。视频摘要的难点主要是视频数据的复杂性和先验知识的缺乏。基于局部-空间模型的视频摘要方法采用了分层思想，先对视频序列进行建模，并结合视频颜色和纹理特征信息提取视频信息参数，再将这些参数输入空间模块进行聚类分析，通过双线性插值和 DES 非线性视频剪辑处理最后得到该视频的动态视频摘要。此模型的框架如图 1 所示。

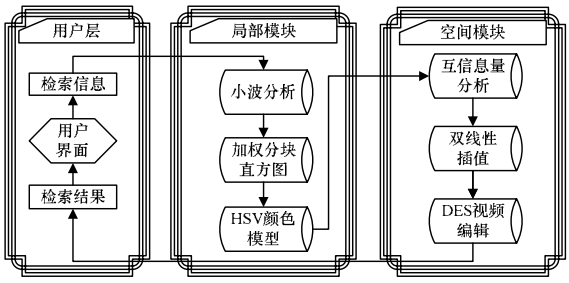

图 1　局部-空间模型框架

2.1　局部模块分析

小波变换提供了一种在不同尺度上研究分析图像特征的工具。小波变换的高频部分表现了图像的纹理特征，而低频部分表现了图像的颜色特征，颜色特征和纹理特征是视频帧的 2 个主要特征。本算法采用了 Db4 小波基，运用离散的小波变换得到图像数据的小波系数，经过一系列变换得到：

$$C_{m,n}^{M}=\sum_{i,j}h(i-2m)h(j-2n)C_{i,j}^{M+1} \tag{1}$$

$$\alpha_{m,n}^{M}=\sum_{i,j}h(i-2m)g(j-2n)C_{i,j}^{M+1} \tag{2}$$

$$\beta_{m,n}^{M}=\sum_{i,j}g(i-2m)h(j-2n)C_{i,j}^{M+1} \tag{3}$$

$$\gamma_{m,n}^{M}=\sum_{i,j}g(i-2m)g(j-2n)C_{i,j}^{M+1} \tag{4}$$

其中，C 为低频信息；α、β、γ 分别为 x 方向和 y 方向的高频信息；$\{h(i)\}$和$\{g(i)\}$分别为低通和高通滤波器。通过上述变换后得到视频帧图像的颜色特征$\{C_{1,1},C_{1,2},\cdots,C_{m,n}\}$和纹理特征$\{D_{1,1,1},D_{1,1,2},\cdots,D_{3,m,n}\}$。

由于传统的直方图特征值不能记录像素点的位置信息，难以反映视频的空间信息，因此国内外已有人提出分块直方图的思想，将整个视频帧均匀分割成 $M\times N$ 块，进行直方图特征值计算。随后出现加权直方图，即在每个子块赋予不同的权值大小，计算后加权平均。本算法采用的是更接近人们对颜色的主观认识 HSV 颜色模型的分块直方图，这样更能体现出图像的位置信息。按照人的颜色感知能力和视觉分辨能力，对 H、S、V 3 个分量进行非等间隔的量化，把色调 H 空间分为 12 份，饱和度 S 分为 4 份，而亮度 V 也分为 4 份。然后，将 3 个颜色分量合成一维特征矢量，即 $L=12H+4S+4V$，这样大大减轻了图像亮度 V 对检索结果的影响，而且也减少了饱和度 S 对检索结果的影响，但是对颜色分布不同的图像能得到较好的检索结果。

颜色直方图描述了图像中的颜色分布，常被应用在镜头边缘检测的特征度量中。为了弥补小波分析中直方图丢失像素空间信息的不足，本文采用文献[5]的分块直方图思想，作了以下改进，如图 2 所示。

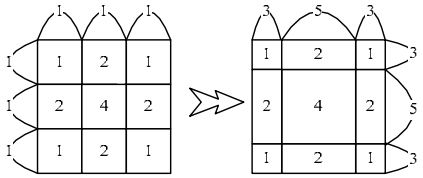

图 2　加权分块直方图的改进

采用 3×3 不均匀分块思想加重图像中间部分权值分布。因为 3×3 图像分块简洁清晰，而不均匀思想既减小了特征的存储空间和图像相似度的运算量，又增大了图像中间部分目标物体信息量的提取，更适合视频摘要技术的研究。

2.2　空间模块分析

经过上述局部模块分析后，增强了视频空间信息量，而且因此排除了闪光灯、抖动等外界因素的影响。对于空间模块，本文采用了互信息量分析[3]获取关键帧。互信息量分析是一种非常有效而且优化的聚类算法。但是文献[3]的算法需要设置阈值，这是其中最为关键的缺陷，因此本算法采用自适应阈值，完全解决阈值难以设定的问题。

2.2.1　互信息量聚类算法

由于颜色特征对于旋转、平移、尺度变化不敏感，因此首先计算颜色特征的互信息量，根据式(2)~式(4)，两帧间的互信息量定义为：

$$I_{t,t+1}^{G}=-\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}C_{t,t+1}[i][j]\times\lg\frac{C_{t,t+1}[i][j]}{C_{t,t+1}[i]C_{t,t+1}[j]} \tag{5}$$

则图像 t,t+1 之间总的互信息量可以表示为：

$$I_{(t,t+1)}=I_{(t,t+1)}^{R}+I_{(t,t+1)}^{G}+I_{(t,t+1)}^{B} \tag{6}$$

纹理信息对光线变化及物体的运动不敏感。纹理熵表示图像中灰度分布的聚集特征所包含的信息量，令 S_i 表示图像中灰度为 i 的像素所占的比例，则图像的灰度熵定义为：

$$E=-\sum_{i=1}^{n}S_i\lg S_i \tag{7}$$

两帧纹理特征的熵差定义为：

$$D_i=|E_{i+1}-E_i| \tag{8}$$

为了消除光线变化及物体运动带来的误检，将 $-\lg(1-D_t)$ 作为权值定义两帧间的相似度为：

$$Q_t=-\lg(1-D_t)\times I_{t,t+1} \tag{9}$$

接着根据式(5)、式(6)和式(9)，计算两帧对应分块之间的相似度：

$$Sim(Q_i,I_i)=\sum_{j=1}^{c}(Q_{ij},I_{ij}) \tag{10}$$

然后确定自适应阈值并划分初始类，其过程如下：

(1)设一个镜头中有 N 帧$\{F_1,F_2,\cdots,F_n\}$连续读入，利用式(10)求相邻两帧的相似度，得到数组 $Dif=\{D_1,D_2,\cdots,D_{N-1}\}$。

(2)以 Dif 中的元素作为一维数据空间的样本进行聚类，分为两类。为提高算法效率，先对 Dif 中的元素由大到小排序，假设排序后有：$D_1\geq D_2\geq\cdots\geq D_{N-1}$，令：

$$T=\arg\min\delta_W^2,\ \delta_W^2=qH\delta_H^2+qL\delta_L^2,\ \mu_H=\frac{1}{q_H}\sum_{i=1}^{T}D_i,$$

$$\mu_L=\frac{1}{q_L}\sum_{i=T+1}^{N-1}D_i,\ \delta_H^2=\frac{1}{q_H}\sum_{i=1}^{T}[D_i-\mu_H]^2,\ \delta_L^2=\frac{1}{q_L}\sum_{i=T+1}^{N-1}[D_i-\mu_L]^2$$

则 D_T 就是所求阈值。

(3)若相邻两帧帧差大于或等于 D_T，则开始新的类；否则，加入此类。

(4)算法停止，得到初始类别数和初始类的划分。

使用基于双向滑动窗口的自适应阈值的方法，取滑动窗口 W 的长度为 15，以当前帧 F_t 为窗口中心，计算 F_t 和 F_{t+1} 之间的互信息量 $I_{t,t+1}$ 以及整个窗口内相邻间的互信息量的均值 $\overline{I}$。若 $\overline{I}/I$ 大于此阈值，就认为镜头发生突变。

最后利用帧间互信息量作为提取关键帧的依据，由此可得镜头 S_i 序列为 $S=\{F_1,F_2,\cdots,F_n\}$，镜头 S 相应的相邻帧间的互信息量为 $Is=(I_{1,2},I_{2,3},\cdots,I_{N-1,N})$。

$$\overline{I_s}=\frac{\sum_{t=1}^{N-1}Q_{t,t+1}}{N} \tag{11}$$

$$\overline{I_{s_i,s_{i+1}}}=\overline{I_{s_i}}-\overline{I_{s_{i+1}}} \tag{12}$$

其中，$\overline{I_s}$ 表示镜头 S 互信息量的平均值；$\overline{I_{s_i,s_{i+1}}}$ 表示镜头 S_i、S_{i+1} 互信息量平均值的差异。

2.2.2 双线性插值法

建立动态视频摘要必须考虑摘要持续时间的长短，以满足不同用户对摘要详细程度的需要。在实现时可以设计一个按钮或滑动条，让用户自己控制视频摘要播放时间的长短。但持续时间不能太短，比如建立一个 2 s 的摘要几乎没有任何意义，因为持续时间太短不能满足人们的感官需要。对于一个连续播放的视频，能够被人脑完全处理的最少播放时间是 3.5 s。因此，在建立视频摘要时，对于摘要的持续时间应考虑以下 2 点：(1)摘要的持续时间不能小于 3.5 s。(2)摘要的持续时间可以根据需要进行调整。

基于以上 2 点，以及考虑到视频编码效率问题，本文采用双线性插值法插入视频帧。因为双线性插值法在考虑场景后，避免摘要重复冗余，大大提高了摘要信息的涵盖度。在插入视频帧时，为了使用户感觉更为自然，考虑了视频场景和关键帧间距的大小。图 3 给出插值法示意图。

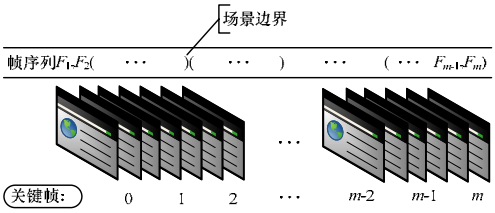

图 3 插值法示意图

在图 3 中，m 为关键帧的个数，用 $d_1,d_2,\cdots,d_m$ 表示两关键帧之间的距离，利用双线性插值法插入视频帧。先通过上述双向滑动窗口算法检测后，向前滑动窗设为 W_g(W_g 为镜头个数)，比较窗内镜头相似度，若镜头相似，则聚类后形成场景，否则，设为场景边界；向后活动窗设为 W_e(W_e 为镜头个数)，比较窗内镜头相似度，若镜头相似，则聚类形成场景，否则作为新场景开始，继续检测。接着判断形成后的场景内关键帧的播放时间是否小于 3.5 s，如果小于，则将场景内所有关键帧按原始顺序插值拷贝，直至其播放时间不再小于 3.5 s，并检测一个场景内关键帧的个数，如果场景内关键帧数 $n\leq 2$，将此场景合并进相邻关键帧相对较少的场景中去，并继续检测相邻场景中关键帧个数。如果 $d_1,d_2,\cdots,d_m$ 序列在一个场景中，则在其中选取最小距离 $d_{\min}=\min(d_i)$, $i=1,2,\cdots,m$。其中，$\Delta d=\lfloor d_{\min}/l \rfloor$(取 $l=2$)；在关键帧之间以 Δd 的间距插入中间帧，插入的帧序列为 $F_1+\theta\times\Delta d$, $\theta\in[1,M/\Delta d]$, $l=l+1$，直至插值结束。

2.2.3 DES 视频编辑

DES(DirectShow Editing Services)是一套基于 DirectShow 核心框架的编程接口。DES 的出现，简化了视频编辑任务，弥补了 DirectShow 对于媒体文件非线性编辑支持的先天性不足。但是，就技术本身而言，DES 并没有超越 DirectShow Filter 架构，而只是 DirectShow Filter 的一种增强应用。本文采用 DES 对摘要视频段进行效果和融合过渡处理，最后生成一个摘要视频段。

3 实验分析

该模型综合了视频底层特征与中层关键帧，优化了算法结构。为了实现摘要 3G 网络传输播放，本文还实现了视频摘要整体化，即把摘要视频段联接起来。实验采用 VC++6.0 作为平台，DirectShow 作为辅助工具。在实验测试中，为了检测关键帧的提取效果，选取 10 多段广告视频片段测试，与文献[3-4]作比较，表 1 为 2 段视频测试部分数据结果。

表 1 关键帧提取结果

视频名	算法	镜头数	总帧数	提取的关键帧数	目测关键帧数
Cool	文献[3]模型	12	269	19	15
	文献[4]模型			21	
	本文模型			17	
JPL	文献[3]模型	25	443	35	31
	文献[4]模型			41	
	本文模型			36	

上述实验结果表明，本文算法在解决阈值设定问题的前提下，提取的关键帧内容丰富，冗余度小，能更好地表现出视频内容。由于本文研究的是动态视频摘要，因此采用线性插值算法，保持了原有关键帧的时间顺序和动态信息，更好地完成插入帧的多少，使视频摘要时间更符合用户需求。

为了检测该模型生成视频摘要结果，仍选取上述 2 段视频与文献[4]进行比较，表 2 为视频数据测试结果。

表 2 视频数据测试结果

视频片段	文献[4]模型		本文模型		视频总时间/s
	摘要时间/s	摘要视频段数	摘要时间/s	摘要视频段数	
Cool	17	4	14.4	3	44
JPL	25	6	17.8	4	44

参考了文献[6]的摘要性能指标评价本文提出的摘要生成方法的内容涵盖度，用文献[7]评价摘要中关键帧集合是不是更好地表达出视频内容。经过对比后，实验结果表明 2 种摘要方法内容涵盖度都很高，且选出的关键帧都具有较好的代表性，但是本文模型的视频摘要更紧凑精要，概括性更好，内容冗余度很低，更重要的是本文采用的是统一、可扩展的模型思想。最后给出本文模型界面及 JPL 视频数据结果，如图 4 所示。

图 4 局部-空间模型框架

(下转第 283 页)