文章编号: 1001- 2486(2003) 03- 0088- 08

# 针对无线移动环境的音频同步视频连环画的自动生成*

王 炜, 张 军, 武德峰
(国防科技大学人文与管理学院, 湖南 长沙 410073)

摘 要: 针对无线移动通信网络的客观制约, 给出了一种音频同步的基于视频连环画模式摘要自动生成的实用视频服务方案, 使得系统在向用户提交连续音频流的同时, 能够随着网络带宽条件的变化, 按照动态采样频率, 基于视频内容选择重要帧递送, 并按照与音轨同步的模式播放, 从而在降低数据量的同时, 满足用户对视频内容综合理解的需求。

关键词: 无线移动视频服务; 配音视频连环画; 视频分割; 代表帧选择
中图分类号: TP26 文献标识码: A

# Automatic Generation of Dubbing Video Slides for Wireless Mobile Environment

WANG Wei, ZHANG Jun, WU De-feng
(College of Humanities and Management, National Univ. of Defense Technology, Changsha 410073, China)

**Abstract**: Aiming at the objective restriction of the current wireless mobile network, a practical video delivery solution based on the automatic generation of dubbing video slides is presented. We guarantee the continuity of the audio stream first, and then select important frames in the dynamic sampling frequency based on the visual content analysis according to the current network bandwidth. Those frames are transmitted and played in a synchronous mode with the corresponding audio track. Our solution can satisfy the user's requirements for video content comprehension while remarkably reducing the quantity of data flow.

**Key words**: wireless mobile video service; dubbing video slides; video segmentation; representative frame selection

随着无线移动网络的迅猛发展和个人数字助理、移动电话等无线终端的普及, 人们对无线移动通信的各种应用需求也与日俱增。在丰富的无线移动数字多媒体服务中, 视频服务是最重要的组成部分之一。

通过无线网络提供移动视频服务的困难和挑战是多样化的。与固定网络相比, 除了接入带宽有限, 还面临网络条件动态变化、稳定可靠性差、传输中继多样化、终端设备解码运算速度、显示屏幕分辨率、色彩深度多样化、传输高功耗和电池容量有限等多方面难题[1]。针对这些挑战的研究正在积极展开, 包括: 新一代无线通信的物理、链路、网络等各层次的高效率传输协议、标准[2], 无线网络上下文中对媒体数据的高效压缩编码[3]、低分辨率图像内插恢复[4], 传输控制、差错保护、自适应服务质量控制[5]、低功耗无线视频传输功率控制[6], 甚至终端设备公共操作系统等。

即使如此, 在目前有限的无线带宽制约下, 针对移动终端的视频服务仍很难传送流畅的多媒体视频。排除文件下载回放的在有限带宽的根本制约下, 目前各类播放软件的实际情况是即便使用缓冲区, 播放中由于带宽的动态变化, 缓冲数据迅速被播放器消耗, 却无法及时从网络获得后续视频数据。遇到类似情况时, 只能简单地随机丢弃跟不上播放速率的视频帧, 向服务器申请后面的帧, 尽可能维持相应音频流的实时连续, 不可避免地造成当前画面与音频严重失去同步, 妨碍用户综合理解。严重情形下将变为类似广播剧的纯粹音频实时接收回放, 不仅浪费了有限带宽, 还经常造成用户放弃。即将到来的 3G 网络保守估计的用户实际带宽也仅在 64Kbps 左右。具有更高带宽和速度, 与 Internet 更好兼容的 4G 技术的研究尚处于萌芽。除了强调更加经济高效地提供传统服务之外, 提供恰当的综合移动互联网数

* 收稿日期: 2002- 10- 25
作者简介: 王炜(1973—), 男, 副教授, 博士。

字多媒体服务,最大程度地满足用户不断增长的信息需求,或者说,解决应用业务内容不足并引导用户消费是目前 3G 普及的前提。与其播放视频时因不能满足帧率需求而被动地随机丢弃数据导致画面与音轨严重失去同步,不如一开始就通过过滤那些可以从信息冗余角度丢弃的视频帧,主动降低帧率来匹配有限信道带宽,因此视频摘要自然地成为目前无线移动视频服务类型的首要选择[7]。

自动视频摘要[8]最初是针对如何快速浏览大型视频数据库来实现有效的内容存取和表现而提出的,分为系列运动图像摘要(视频撇取)[9]或静态图像概要(故事板)[10]两类模式。故事板是从视频源中产生的跳跃的显著图像帧的子集。视频撇取是由较短长度的连续图像的子序列集合构成,也附带相应从原始序列中抽取的音频摘要,是视频片段的集合。目前这两类摘要主要是针对 Internet 应用环境,在较好保留基本信息的基础上,以洗练的数据辅助用户迅速定位需要的原始视频内容。本文需要的是一个音轨画面相结合并反映视频全部概貌的摘要机制。静态故事板概要只是从视觉角度提供关于整个视频内容的全面印象,不包含音频内容;运动故事板中则仅包含原始视频中最吸引人的精彩场面,不涉及全面视频。其次,二者对特定视频产生的摘要结果都是固定的,而无线移动视频需要针对带宽的随时动态变化给出不同颗粒度的视频概要。因此这些已有模式并不适用于本文问题。

## 1　配音视频连环画思想简介

受电影改编连环画、PowerPoint 幻灯及配音 Flash 等类似媒体形式的启发,本文提出一种在带宽和视频质量间折衷的新思想。其核心是在尽可能减少用户对视频理解的干扰的前提下,针对不同的和动态变化的带宽,基于视频内容自动从连续视频流中依照时间顺序抽取动态数目的最可能代表视频内容的基本帧序列,调整帧率和流量大小,按照不同细节程度为动态带宽提供、播放与原始连续音频同步的离散连环视频帧序列。该方法极大地减少了传输数据量,对各种窄带有着很好的自适应能力,从而可以有效地应用于无线移动视频服务环境。类似 Flash,该方法也可能为用户带来新的视频消费模式体验。

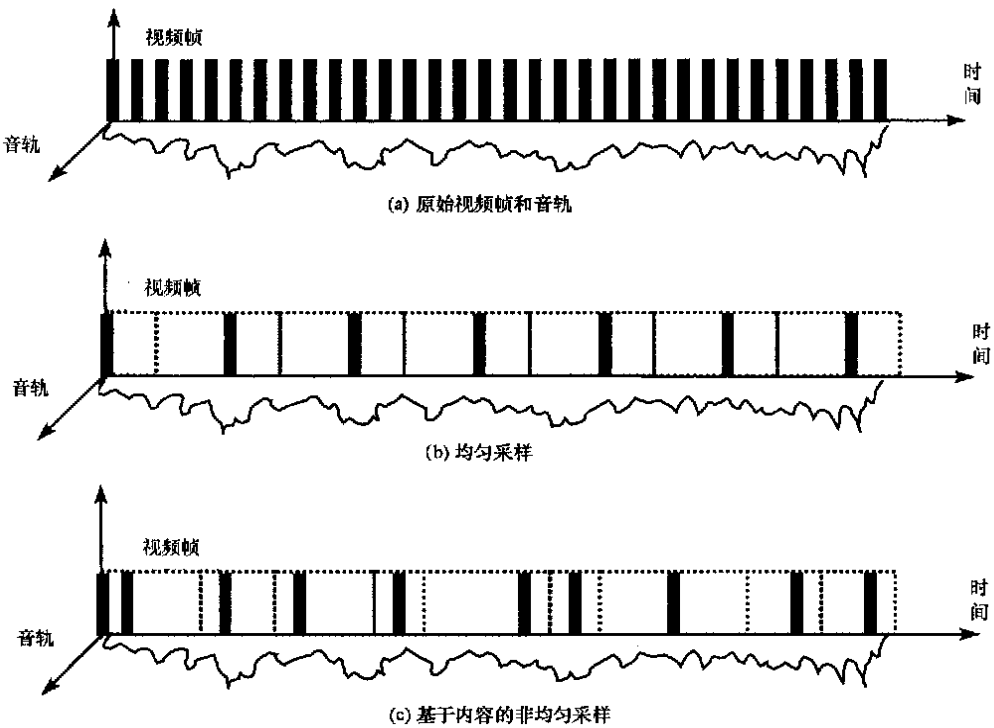


图 1　原始视频和配音视频连环画

Fig. 1　Original video and dubbing video slides

假定视频 $F$ 实时连续播放条件下持续时间为 $T^F$。$F$ 的帧总数为 $N^F$,顺序排列为:

$$f = \{f_1, f_2, f_3, \cdots, f_i, \cdots, f_{N^F}\}, \quad 1 \leqslant i \leqslant N^F$$

前提是在实时连续递送并播放音轨的同时,剩余带宽无法达到画面流畅的要求,但至少能满足某个程度用户可接受的活动图像传输率。此时考虑基于内容过滤画面帧来降低帧率。假定当前网络状态下应保留的帧数为 $N_a$,对应过滤后保留的活动图像序列为:

$$f' = \{f^{1}_{1}, f^{6}_{2}, \cdots, f^{i}_{x}, f^{j}_{x+1} f^{k}_{x+1} \cdots f^{N^F}_{N_a}\}, \quad 1 \leqslant i < j < k \leqslant N^F, 1 \leqslant x \leqslant N_a - 2$$

上标表示特定帧 $f$ 在原来序列中的序号,下标表示在新序列中的序号。使那些被保留的离散帧代表其邻近被丢弃的若干帧构成的区间,如 $f^{i}_{x}$ 代表原来 $\{f_i, \cdots, f_j\}$ 子序列,$f^{j}_{x+1}$ 代表 $\{f_j, \cdots, f_k\}$,并在播放时持续相应时间 $T^F \cdot (j - i + 1)/N^F$ 和 $T^F \cdot (k - j + 1)/N^F$,以便和连续音轨保持同步。

基于这种思想的活动图像序列 $f'$ 和音轨同步构成的视频递送和播放模式就是所谓的配音视频连环画模式。其合理性在于:视频帧的画面变换是一个连续过程,只要视频帧的丢弃限制在局部时间范围内,通过多个诸如 $f'$ 序列中的采样帧强制设定音频同步点,用户仍可依据连续视频片段中的局部上下文信息和先验知识,结合画面和音频流来综合理解。虽然部分局部细节丢失,但在整体上不会影响用户捕捉视频轮廓。它有效地在有限带宽条件下,极大程度地表达了尽可能多的视觉、听觉综合信息。本文重点讨论在已知 $N_a$ 的前提下,选择保留和递送哪些帧,使得这些帧能最大程度地支撑或代表整个画面帧集合并结合音轨综合反映内容轮廓。

自然的想法如图 1(b),均匀采样虽然也能显著降低帧率、满足带宽条件,但视频内容在不同的场景、不同的片段中变换强度不同,有的片段相对静止,有的片段运动剧烈。等间隔采样无法反映内容变化的这种不均匀性,势必在平缓的地方采集过多的样本,却在变换陡急的地方采样不够,不能很好表现视频内容。合理的配音视频连环画自动生成应该基于视频内容分析,实现非均匀活动图像采样,帧采样频度反映原始视频帧序列中的内容变换强度。

我们将按如图 2 所示过程讨论该方案。第 2 节讨论视频段分割,将整个视频分割为基本片段,以片段为单位从中抽取代表帧子集;第 4 节讨论如何得到最大程度代表单个片段的动态数目的代表帧集合。对任一片段的视频帧序列 $f$,抽取表征其图像内容的特征向量,将帧序列转换为一个特征序列。分析过滤该特征序列,得到代表序列 $f$ 的活动图像序列 $f'$。最后在模拟效果实验中运用同步手段,将所有 $f'$ 和原始音轨合成为播放时刻视频连环画,并给出小结。

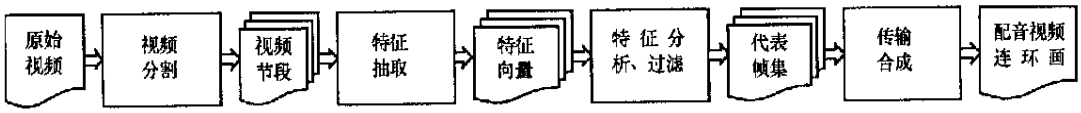


图 2 配音连环画自动生成过程

Fig. 2 Automatic generation procedure of dubbing video slides

## 2 视频分割

将视频分割为若干情节片段,在局部情节范围内依据内容抽取视频片段的若干代表帧。在局部范围内实现帧选择,再按情节片段顺序拼接,可以保证每个局部情节都至少存在一个相应代表帧,从而不漏掉任何语义片段,能有效提高配音连环画的视听效果或可读性。

视频分割是按照某些规则选择视频间隔片段的处理过程,由系列连续帧构成的镜头是视频制作、剪辑过程中形成的最基本单位,其内部组成帧不再包含任何编辑效果,记载着单个摄像机在一个地点对一次连续运动捕捉所产生的图像帧序列,也反映了视频序列中场景的不同变换。作为主要分割规则,镜头分割采用的线索不依赖于具体应用领域的语义,又称为基于语法的分割。

根据上下文需求,使用基于镜头和基于文本字幕的分割相结合的综合分割。字幕,尤其是电影对白,是在后期制作过程中人工生成的,与音轨有很强的语义关联同步,在此基础上得到的连环画代表帧包含的文本也将与音轨保持语义同步,因此是很好的语义分割依据。

### 2.1 镜头分割

镜头分割的基本思想是通过构造量化指标衡量、比较帧间差异, 并设定阈值捕获峰值来判断边界, 但由于后期制作采用的编辑效果不同, 镜头衔接上会有突变或渐变等不同风格。渐变会减弱边界处邻近帧差异, 造成镜头边界漏判。单纯靠降低阈值来捕获渐变边界, 会导致反映相同内容但内部运动相对剧烈的帧序列构成的单一镜头被误判为多个镜头。为了较好解决渐变边界探测问题, 使用如下双阈值增强判定方法。它适用于任何帧间差异指标, 在有效降低渐变镜头漏判的同时, 也能抑制错判导致的过度分割。

假定邻近帧差异为 $D(n, n-1)$, 帧序号 $n$ 满足 $2 \leqslant n \leqslant N^F$。定义两个阈值 $TH > TL$。逐个计算 $D(n, n-1)$。当 $D(n, n-1) \geqslant TH$ 时, 判定 $n$ 对应帧为突变镜头边界; 当 $D(n, n-1) \leqslant TL$ 时, $n$ 和 $n-1$对应的帧显然隶属同一个镜头内部; 当 $TH > D(n, n-1) \geqslant TL$ 时, 激活 $m_0$, 记 $m_0 = n$ 为可能渐变镜头的开始帧序号, 进一步观察后续帧来确认是否存在镜头渐变以及渐变结束位置。当 $m_0$ 有效时, 在计算后续帧 $m$ ($m > n$) 的 $D(m, m-1)$ 的同时, 也同时激活对第 $m$ 帧与 $m_0$ 帧的非邻近帧差异 $D(m, m_0)$ 的计算。一个渐变镜头进入渐变后, 其邻近帧间变化由于渐变编辑效果而削弱, 但即使被弱化, 一般仍大于属于相同镜头内部的邻近帧间变化, 因此邻近帧差 $TH > D(n, n-1) \geqslant TL$ 的假定合理。分析后不难得知, 在渐变过程中, $D(m, m-1)$ 应该继续保持 $TH > D(m, m-1) \geqslant TL$, 但 $D(m, m_0)$ 随着帧号 $m$ 的增加而呈现积累效果, 直至 $D(m, m_0) \geqslant TH$ 并持续保持。当渐变结束进入下一个镜头时, $D(m, m_0)$ 的计算结果显然也继续保持 $D(m, m_0) \geqslant TH$, 但 $D(m, m-1)$ 由于进入相同镜头内部, 将恢复为 $D(m, m-1) < TL$。根据这些规律, 在 $D(m, m_0)$ 计算激活后, 随着 $m$ 的递增, 如果出现 $D(m, m_0) \geqslant TH$ 且 $D(m, m-1) < TL$, 则判定当前 $m$ 帧为渐变镜头结束处, 记为 $m_1$。在渐变探测激活后, 如果在渐变结束被断定之前出现两次以上 $D(m, m_0) < TL$, 或已从 $TH > D(m, m-1) \geqslant TL$ 进入 $D(m, m-1) < TL$, 但 $D(m, m_0)$ 仍未满足 $D(m, m_0) \geqslant TH$ 判定结束条件, 说明这次激活可能是镜头内部运动前景对象误激造成, 取消 $m_0$ 作为这次渐变起点的有效性并继续执行即可。$TH$ 和 $TL$ 的合理取值通过实验数据训练得到。

### 2.2 基于字幕分割

基于边缘提取、平滑消噪、增强、投影、区域分割等方法探测和定位图像中文本区域[11], 判断镜头片段内部每帧画面是否包含文本区域。如果存在文本区域, 文本串图像的宽、高一般符合一定比例, 且一般出现在画面下方, 根据这些类似先验语义, 可进一步确定初步筛选区域的有效性, 将镜头内部视频帧分割为包含字幕和不包含字幕的交替视频片段。

对包含字幕的视频片段, 根据上面得到的文本区域位置, 逐帧提取文本区域图像, 通过灰度化、平滑去噪、二值化、字符分割、规格化等过程, 得到每个字符大小相同且充满空间的顶格标准化黑白图像。送入OCR 识别引擎, 得到对应字符编码。根据字符顺序和字符间的间距, 得到对应字幕的文本编码。实验表明, 结合先验知识的文本区域定位比较精确[11], 但由于背景复杂及图像分辨率较低等主要原因, 字幕文本字符串的识别效果不佳, 正确识别率约在 50% 左右。但足以区分连续视频中不同字幕的节段。

在分割基础上, 得到足以反映最小情节单元且长度不等的视频片段集合。假定分割得到 ρ 个视频片段, 最后的视频片段集合为:

$$\{D_{i,j}^{\Psi} \mid 1 \leqslant i < j \leqslant N^F, 1 \leqslant \Psi \leqslant \rho\}$$

上标 $\Psi$ 表明是第几个视频片段; 下标 $i, j$ 分别表示该片段在原始视频中的起止帧号。其中每个片段或不包含文本, 或包含相同文本, 且片段之间保持原有时间顺序。$N_a$ 的取值跟随网络变化在 ρ 和 $N^F$ 之间变化, 即 $\rho \leqslant N_a \leqslant N^F$。

## 3 活动图像过滤选择

针对特定视频片段的画面帧, 考虑抽取能代表和区分其画面内容的结构化特征向量。在特征向量

空间中比较、分析向量间的异同, 从而实现动态视频摘要。

### 3.1 多维模糊颜色特征向量

问题首先是在保证合理计算复杂度的同时, 如何构造代表视频帧内容的结构化特征向量。颜色的特征与应用无关, 适用于各种场合, 其中的颜色直方图技术公认是典型、简单、抗噪的颜色特征分析技术[12]。但基本颜色直方图严重依赖于颜色区间的刚性划分边界, 在实际应用中存在不足。本文给出基于模糊分类和不同区域比重的模糊颜色直方图法计算特征[13], 其依据是人类对颜色差别的心理认知基于连续变化, 对近似色不敏感; 其次, 人类对图像心理认知取决于各种显著颜色比率及其在平面中的分布位置。一瞥之下的心理认知感受仅对不超过几十种且至少占图像面积 2% ~ 3% 以上的显著颜色敏感。除非对局部聚焦观察, 否则其它细节将被忽略。模糊直方图大致描述如下, 显著颜色提取和更详细的步骤参见文献[13]。

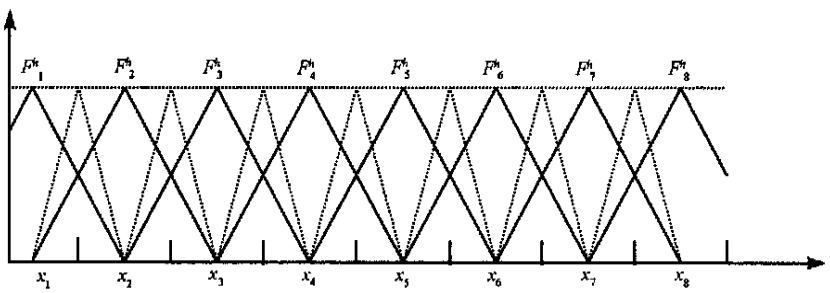


图 3 隶属度函数

Fig. 3 Membership function

假设视频帧 $G$ 是宽 $W$ 高 $H$ 的真彩图像, 向量 $\boldsymbol{V}$ 表征其颜色特征。分别将 3 个不同颜色通道量化为 $n^h$, $n^s$, $n^v$ 个区间段, 得到如下独立通道上的区间分割:

$$\boldsymbol{C}^h = \{C^h_1, C^h_2, \cdots, C^h_i, \cdots, C^h_{n^h}\}, \quad i = 1, 2, \cdots, n^h$$

$$\boldsymbol{C}^s = \{C^s_1, C^s_2, \cdots, C^s_j, \cdots, C^s_{n^s}\}, \quad j = 1, 2, \cdots, n^s$$

$$\boldsymbol{C}^v = \{C^v_1, C^v_2, \cdots, C^v_k, \cdots, C^v_{n^v}\}, \quad k = 1, 2, \cdots, n^v$$

假定 $G$ 中任意像素 $P$ 的值为 $\boldsymbol{x} = (x^h, x^s, x^v)$, $x^h$、$x^s$、$x^v$ 表示像素在三个颜色通道中取值。以 $x^h$ 为例, 讨论如何将其模糊划分到 $\boldsymbol{C}^h$ 的 $n^h$ 个类中去。为类别 $C^h_i$ $(1 \leqslant i \leqslant n^h)$ 分别指定隶属置信度函数 $F^h_i(x)$ $(1 \leqslant i \leqslant n^h)$。任给 $x^h$ 值, 分别计算 $n^h$ 个 $F^h_i(x^h)$ 函数, 得到 $x^h$ 隶属于所有 $n^h$ 个类别的 $n^h$ 维隶属置信度向量。

$$\boldsymbol{C}^{x^h} = \{c^{x^h}_1, c^{x^h}_2, \cdots, c^{x^h}_i, \cdots, c^{x^h}_{n^h}\}$$

隶属置信度函数 $F^h_i(x)$ 有多种构造方式, 模糊数学常用的有三角形、梯形或高斯函数。三角形函数 $F^h_i(x)$ 如公式(1), 值域范围在[0, 1], $x_i$ 是第 $i$ 个区间的中间点的取值。

$$F^h_i(x) = \begin{cases} 1 - \dfrac{|x - x_i|}{\varphi}, & |x - x_i| < \varphi \\ 0, & |x - x_i| \geqslant \varphi \end{cases} \quad (i = 1, 2, \cdots, n^h) \tag{1}$$

或

$$F^h_i(x) = \begin{cases} 1, & |x - x_i| < \dfrac{\varphi}{2} \\ 2 - \dfrac{2|x - x_i|}{\varphi}, & \dfrac{\varphi}{2} \leqslant |x - x_i| < \varphi \\ 0, & |x - x_i| \geqslant \varphi \end{cases} \tag{2}$$

使用简单的三角隶属函数, 如图 3 中实线所示, 虚线为梯形函数。$\varphi$ 表示三角形底边宽度的1/2。针对每个通道, 使用类似定义为类别 $\boldsymbol{C}^h$, $\boldsymbol{C}^s$, $\boldsymbol{C}^v$ 指定三组隶属函数:

$$F^h_i(x^h), \quad 1 \leqslant i \leqslant n^h; \quad F^s_j(x^s), \quad 1 \leqslant j \leqslant n^s; \quad F^v_k(x^v), \quad 1 \leqslant k \leqslant n^v$$

分别得到像素点 $P$ 的值 $\boldsymbol{x}=(x^h, x^s, x^v)$ 的三个独立颜色通道分量 $x^h, x^s, x^v$, 分别隶属于 $\boldsymbol{C}^h$, $\boldsymbol{C}^s$, $\boldsymbol{C}^v$ 的置信度向量, 然后按照如下公式计算不同通道上的模糊直方图统计:

$$\boldsymbol{H}^h=\left(H_1^h, H_2^h, \cdots, H_i^h, \cdots, H_{n^h}^h\right)$$

$$H_i^h=\frac{1}{W*H}\sum_{\forall x\in G}F_i^h(x^h), \quad i=1,2,\cdots,n^h \tag{3}$$

$$\boldsymbol{H}^s=\left(H_1^s, H_2^s, \cdots, H_j^s, \cdots, H_{n^s}^s\right)$$

$$H_j^s=\frac{1}{W*H}\sum_{\forall x\in G}F_j^s(x^s), \quad j=1,2,\cdots,n^s \tag{4}$$

$$\boldsymbol{H}^v=\left(H_1^v, H_2^v, \cdots, H_k^v, \cdots, H_{n^v}^v\right)$$

$$H_k^v=\frac{1}{W*H}\sum_{\forall x\in G}F_k^v(x^v), \quad k=1,2,\cdots,n^v \tag{5}$$

整个颜色空间被 $\boldsymbol{C}^h$, $\boldsymbol{C}^s$, $\boldsymbol{C}^v$ 分割为 $n=n^h*n^s*n^v$ 个量化子空间, 记为 $\boldsymbol{C}=\{C_1, C_2, \cdots, C_l, \cdots, C_n\}(l=1,2,\cdots,n)$。$l$ 与 $i, j, k$ 间定义如下对应关系:

$$l=n^i*n^j*(k-1)+n^i*(j-1)+i \tag{6}$$

根据模糊数学乘法, 在遍历像素 $x\in G$ 后, 得到针对 $G$ 定义的模糊直方图特征向量 $\boldsymbol{V}$ 如下, 其中 $i$, $j$, $k$, $l$ 满足(6)式给出的关系:

$$\boldsymbol{V}=(H_1, H_2, \cdots, H_l, \cdots, H_n), \quad n=n^h*n^s*n^v$$

$$H_l=H_i^h*H_j^s*H_k^v, \quad 1\leqslant l\leqslant n, 1\leqslant i\leqslant n^h, 1\leqslant j\leqslant n^s, 1\leqslant k\leqslant n^v \tag{7}$$

### 3.2　视频片段内部代表帧数目的确定

在 $\rho$ 个长短不等的情节片段 $\{D_{i,j}^{\Psi}\}$ 及对应帧多维特征向量 $\boldsymbol{V}(i)$ 到 $\boldsymbol{V}(j)$ 的基础上, 在分析特征向量并抽取活动图像集合之前, 需要基于待抽取活动图像的总数目 $N_a$ 计算特定片段中待抽取的帧数目。

该值和片段长度及内容变换强度有关。对任意视频片段 $D_{i,j}^{\Psi}$ 内容变化强度的计算依赖从第 $i$ 帧到第 $j$ 帧对应的特征向量 $\boldsymbol{V}(i)$ 到 $\boldsymbol{V}(j)$, 估算为:

$$I^{\Psi}=\frac{1}{j-i+1}\sum_{x=i+1}^{j}\|\boldsymbol{V}(x)-\boldsymbol{V}(x-1)\| \tag{8}$$

基于长度和内容变化强度双重影响因素, 给出对应每个 $D_{i,j}^{\Psi}$ 应抽取帧比率经验公式:

$$\beta^{\Psi}=(j-i+1)^{\frac{1}{2}}*I^{\Psi} \tag{9}$$

将 $\{D_{i,j}^{\Psi}\mid 1\leqslant i<j\leqslant N^F, 1\leqslant\Psi\leqslant\rho\}$ 中所有片段对应的 $\beta^{\Psi}$ 做如下归一化处理:

$$\beta^{\psi}=\frac{\beta^{\Psi}}{\sum_{k=1}^{\rho}\beta^k} \tag{10}$$

得到针对视频 $F$ 的每个片段的相对代表帧数比率 $\beta^{\psi}(1\leqslant\Psi\leqslant\rho)$, 每个 $D_{i,j}^{\Psi}$ 片段中应抽取的活动图像代表帧数目为:

$$N^{\Psi}=\lfloor N_a*\beta^{\psi}+1\rfloor, \quad 1\leqslant\Psi\leqslant\rho \tag{11}$$

$\lfloor\ \rfloor$ 表示向下取整, 因为 $N_a$ 受当前网络带宽状态影响在特定范围内变化, 因此每个片段应抽取的帧数目也跟随 $N_a$ 的变化而变化, 从而生成多个不同精细程度的画面内容序列。

### 3.3　代表帧选择

给定片段 $D_{i,j}^{\Psi}$, 从第 $i$ 帧到第 $j$ 帧的特征向量 $\boldsymbol{V}(i)$ 到 $\boldsymbol{V}(j)$ 以及待抽取帧数目 $N^{\Psi}$, 如何从 $j-i+1$ 个候选帧中确定 $N^{\Psi}$ 个最能代表视觉内容的代表帧? 合理做法是剔除那些在视觉上和前面的帧相比变化不大、从局部上下文中可以推测出内容的次要帧, 同时保留那些相对难以推测或预见的有重要视觉线索的代表帧。

对 $D_{i,j}^{\Psi}$ 中任意帧 $k(i\leqslant k\leqslant j)$, 对应特征向量 $\boldsymbol{V}(k)$ 可认为是 $n^h*n^s*n^v$ 高维空间中的抽象点, 具体位置由各分量确定, 从第 $i$ 帧到第 $j$ 帧的帧序列就被 $j-i+1$ 个特征向量的对应点映射成一条反映内容

变化走势的轨迹曲线。这些特征点构成的连线实际上是非平滑的折线型连贯轨迹, 形状起伏直观上反映了内容的变化。

可直觉断定, 在轨迹线曲率大的地方, 对应帧的内容更难从局部上下文中推断, 因而是主要的; 曲率小的地方, 即较平滑的地方, 对应帧的内容则更容易从局部上下文中推断, 因而是次要的。依次剔除该曲线轨迹中 $j- i+ 1- N^{\Psi}$ 个可以从局部上下文中容易预见的次要点, 将得到仍然能较好表征该曲线轨迹大致轮廓的 $N^{\Psi}$ 个点。这 $N^{\Psi}$ 个点作为选择结果, 其对应帧能够较好表征视频片段的大致内容。

如前所述, 每个片段应抽取的帧数 $N^{\Psi}$ 受网络状态影响在特定范围内变化。假定 $N^{\Psi}$ 最小值为 $N^{\Psi}_{\min}$, 最大值为 $N^{\Psi}_{\max}$。下面讨论如何选择 $N^{\Psi}_{\min}$ 个连环代表帧:

首先使用给定片段 $D^{\Psi}_{i,j}$ 的 $j- i+ 1$ 个视频帧的局部序号顺序, 得到从 1 到 $j- i+ 1$ 的空间点对应的序列。对任意 $\boldsymbol{V}(k)$, $1< k< j- i+ 1$ 对应的点, 定义一个局部上下文相关性测度 $LR(\boldsymbol{V}(k))$, 用来度量该点在空间序列中的可推测性。曲线 $\boldsymbol{V}(k)$ 在对应点处的曲率或曲率半径是一个逻辑上直观的度量标准, 但由于该空间曲线由非光滑的折线段构成, 因此构造如下局部上下文相关性测度:

$$LR(\boldsymbol{V}(k)) = \| \|\boldsymbol{V}(k+1)- \boldsymbol{V}(k-1)\| - (\|\boldsymbol{V}(k)- \boldsymbol{V}(k-1)\| + \|\boldsymbol{V}(k+1)- \boldsymbol{V}(k)\|)\| \quad (12)$$

含义很直观, 3 点构成超平面中的一个三角形关系。如果 $LR(\boldsymbol{V}(k)) = 0$, 则 $\boldsymbol{V}(k)$ 在 $\boldsymbol{V}(k- 1)$ 和 $\boldsymbol{V}(k+ 1)$ 二点的连线上, 说明 $\boldsymbol{V}(k)$ 处的变化不大, 可通过局部上下文推测。否则说明 $\boldsymbol{V}(k)$ 偏离 $\boldsymbol{V}(k- 1)$ 和 $\boldsymbol{V}(k+ 1)$ 二点的连线, $LR(\boldsymbol{V}(k))$ 越大, 说明偏离越大, 该点局部变化也就越大。得到 $j- i- 1$ 个点对应的 $LR(\boldsymbol{V}(k))$ 值后, 对 $LR(\boldsymbol{V}(k))$ 排序, 删除 $LR(\boldsymbol{V}(k))$ 最小值代表的那个空间点, 或者说视频帧。

对剩下的 $j- i$ 个点重复上述操作, 直到剩下的点个数等于 $N^{\Psi}_{\min}$ 时停止。不能在第一次 $LR(\boldsymbol{V}(k))$ 排序后就一次性删除 $j- i+ 1- N^{\Psi}_{\min}$ 个 $LR(\boldsymbol{V}(k))$ 值较小的点来直接得到结果。因为每次删除一个点后, 对任意 $\boldsymbol{V}(k)$, 其邻近点代表的局部上下文可能产生变化, 因此必须逐个剔除最小变化的点来迭代产生最后结果。

在计算需保留的 $N^{\Psi}_{\min}$ 个代表帧的完整循环扫描过程中, 实际上也得到了 $N^{\Psi}$ 从 $N^{\Psi}_{\min}$ 一直到 $N^{\Psi}_{\max}$ 所对应的所有结果序列。为曲线中每个点对应的帧定义一个变量 $\lambda(\boldsymbol{V})$, 定义为在整个迭代过程中的某一步, 当该点应该从序列中剔除时, 目前已经删除的点的总个数, 该变量初始化取值为 $j- i+ 1- N^{\Psi}_{\min}$。最后得到的过滤结果就是按照顺序记载的, 按上述算法产生的 $j- i+ 1$ 个对应点的 $\lambda(\boldsymbol{V})$ 数值。所有结果都隐含在这个 $\lambda(\boldsymbol{V})$ 序列中, 并不需要为每个 $N^{\Psi}$ 取值都专门计算并保留一个帧序号数组。可认为 $N^{\Psi}$ 是一个反映网络动态带宽的实时变化阈值, 传输时用每个帧预先计算好的并反映内容重要性的 $\lambda(\boldsymbol{V})$ 属性和阈值做比较, 符合条件的递送, 否则丢弃。

## 4 模拟实验和结论

动态活动图像反映了不同颗粒度的视频内容梗概。在传输、播放完整音轨的同时, 按顺序传输、缓冲这些图像帧。播放时每个帧在原始视频参考时间轴上按序号对齐, 用每个离散图像置换其邻近被丢弃未传输的若干连续帧构成的区间, 作为区间代表的连环画图像持续时间等同于区间段的持续时间, 并与区间原来的音频同步关系保持一致, 就形成了配音视频连环画的播放效果。我们运用上述方案分别对新闻、动画片和电影等不同类型视频素材进行了实验处理。利用 Adobe Premier 视频编辑工具, 针对不同网络带宽计算视频内容梗概结果, 手工合成具有等价播放效果的配音视频连环画并观察视听效果。

对原始视频的类似梗概计算和播放效果是一个难以按照量化方式客观评估的主观性过程。从模拟播放效果来看, 虽然画面不连贯, 视觉细节丢失, 连环动画方式的画面与音轨的准同步结合仍从整体表达了可理解的内容梗概。有趣的是, 虽然随着 $N_a$ 的减少, 配音视频连环画画面跳跃增大, 细节丢失逐步加重, 对内容的理解难度加大, 但参与主观评估的人员均认为 $N_a$ 值较小时的综合效果优于 $N_a$ 值较大时帧数虽多但仍不流畅的情形, 后者虽然提供更多的细节内容, 但不流畅时画面闪动明显且频繁, 反而容易引起视觉的不适。

模拟实验表明, 在有限带宽的前提下, 当表征清晰度和流畅性的解析度、帧速率指标无法同时满足时, 这种配音视频连环画思想给出了一个提供恰当多媒体视频服务质量的简单、实用的新模式, 使得用户至少可以把握主体内容, 尤其适合于动画片、新闻点播等视频节目。在带宽允许范围内, 上述方案具有很强的动态可伸缩性, 事实上有能力产生代表帧数目在片段总数和完整视频总帧数间的各种概括程度的活动图像序列, 因而具有广泛的适用性, 可运用于各种接入带宽有限条件下缓冲不足时选择性视频帧丢弃等类似应用。有限带宽下缺乏恰当的数字多媒体服务是 3G 进展缓慢的重要原因。多媒体配音视频连环画点播则为类似 3G 无线移动环境提供了可行的折衷业务, 或许有可能带来用户认可的新型视频消费模式。

本文并未完整讨论服务器端内容创建、数据传输, 客户端接收以及同步回放等软件模块在内的整套系统。完整的实现还包括系统在发送连环画活动图像前, 首先端—端协商, 根据网络现状和移动终端显示、计算能力的不同确定本次服务质量, 并确定初始的 $N^{\Psi}$ 值。连环帧过滤需在 $\lambda(\boldsymbol{V})$ 计算结果中按顺序取出所有符合 $\lambda(\boldsymbol{V}) \leqslant (j - i + 1 - N^{\Psi})$ 的帧, 并按原始视频中的子序列顺序发送。系统在递送过程中要跟踪当前带宽的动态变化, 自动调整当前传输对应的 $N^{\Psi}$ 值来动态适应。这些内容的实现必然还涉及压缩编码、格式标准等一系列问题。

服务器端视频连环画媒体的内容创建, 即如何自动生成配音视频连环画, 以便在有限带宽下最大程度地保留更多的视听综合信息是本文核心。做法是将视频分割为情节片段, 然后从中基于内容分析抽取少量离散活动图像, 最后以持续连环画方式与音轨同步播放, 在降低帧率和牺牲连续性的同时有效降低对带宽的需求。这种计算建立在已知全部视频内容及特性的基础上, 在没有得到全局视频内容之前, 无法就局部信息实时给出合理结果。鉴于计算复杂, 在播放之前也必须完成计算, 因此无法适用于类似现场直播的场合。

## 参 考 文 献:


[1] Seckin G. Challenges of Wireless Media Streaming[C]. SSGRR L' Aquila, Aug 2001.

[2] Vass J, Zhuang S. Scalable, ErrorResilient, and High-Performance Video Communications in Mobile Wireless Environments[J]. IEEE Trans. Circuits and Systems for Video Technology, 2001: 833- 847.

[3] Beek P, Tekalp A M. 2D Mesh Geometry & Motion Compression for Efficient Object-based Video Compression[C]. IEEE Intl. Conf. on Image Processing, 1997.

[4] Biork C. Transcoder Architectures for Video Coding[J]. IEEE Trans. Consumer Electronics, 1998, 44(9): 88- 98.

[5] Wang G J, Zhang Y Q. Channel-adaptive Error Control for Scalable Video over Wireless Channel[C]. IEEE MoMuc. 2000, 2000.

[6] Kravets R, Krishnan P. Application-driven Power Management for Mobile Communications[J]. Wireless Networks, 6(4): 263- 277.

[7] Tseng B L, Lin C Y, Smith J R. Video Summarization and Personalization for Pervasive Mobile Devices[C]. Storage and Retrieval for Media Databases, 2002, Proceedings of SPIE Vol. # 4676.

[8] Li Y, Zhang T, Tretter D. An Overview of Video Abstraction Techniques[R]. HP Laboratory Technical Report, HPL- 2001- 191, 2001.

[9] Smith M, Kanade T. Video Skimming for Quick Browsing Based on Audio and Image Characterization[R]. Technical Report CMU- CS- 95- 186, School of Computer Science, Carnegie Mellon University, July 1995.

[10] Wolf W. Key Frame Selection by Motion Analysis[C]. ICASSP' 96, 1996, 2: 1228- 1231.

[11] Cai M, Song J Q, Lyu M R. A New Approach for Video Text Detection[C]. Proc. Intl. Conf. On Image Processing, Rochester, New York, USA, 2002.

[12] Flickner M. Query by Image and Video Content: the QBIC System[J]. IEEE Computer, 1995, 28(7): 23- 32.

[13] 王炜, 武德峰. 基于显著颜色和模糊分类的颜色直方图构造研究[J]. 模糊数学与应用(已录用, 2004 年 2 月发表).