基于音视频特征的新闻条目自动分割*)

Automatic Segmentation of News Items Based on Video and Audio Features

王伟强[1] 高 文[1,2] 马继涌[1] 林守勋[1] 李锦涛[1]

(中科院计算所数字化技术实验室 北京100080)[1]

(哈尔滨工业大学计算机科学与工程系 哈尔滨150001)[2]

Abstract Automatic segmentation of news items in a MPEG-2 stream is a significant research topic for implementing an automatic cataloging system of news video. This paper presents an approach which employes audio and video feature information to automatically segment news items. Combining the analysis techniques of audio and video can overcome the weakness of the approach which only uses the image analysis techniques. This combination makes our appoach more widely adaptable to variable existence situations of news items. The proposed approach detects silence clips in accompanying audio, and integrates with shot segmentation results, as well as anchor shot detection results, to determine boundaries between two news items. Experimental results show that the integration of audio and video features is an effective approach to solve the problem of automatic news items segmentation.

Keywords MPEG-2 stream, Automatic segmentation of news items, Audio-video information analysis, Anchor shot

1. 引言

目前对视频文档结构化分析的大多数研究都集中在对视频帧中包含的视觉信息的分析基础上。为了有效组织、检索视频数据库中的文档，视频文档被分割成有语义意义的单元，如场景[1,2]，此外作为对场景细节的勾画，组成场景的各镜头被探测出来[3-5]，并从中选择出合适的关键帧[6-9]进行索引。

音频作为视频文档中包含的另外一种类型时间媒体，是一种可为视觉信息提供重要补充的信息源，从中可获得一些通过视觉无法获得的信息。如在CCTV新闻的播音员镜头中，视觉内容基本保持不变，但该视频段的伴音却可能包含多条新闻条目的播报。一部电影中有时会存在视觉上内容具有较大差异的若干相邻镜头序列，但作为它们伴音的连续音乐却可以将其归并于同一个语义片断。

将音频内容分析应用于视频内容刻画方面的研究目前还比较有限，文[10]中提出了利用MPEG流中的子带信息计算音频特征的方法，并将音频分为对话与非对话两类用于对新闻条目的索引。文[11]通过对音频的分析来区分五类不同的视频场景，包括新闻报道、天气报道、篮球赛、足球赛、广告。文[12]通过将音频、图像特征结合以分割视频的镜头，Boreczky与Wilcox通过一个隐含Markov模型来描述视频中镜头变化与音频、视觉特征变化的关系，在模型中包含了镜头(shot)、切变(cut1、cut2)、叠化(dissolve)、镜头平移(pan)等7种状态，观测样本包括相邻帧直方图差、相邻音频片断差、镜头运动的特征，通过训练建立模型的参数，并通过Viterbi算法实现镜头分割。

新闻条目自动分割是实现新闻自动编目系统的一项重要研究内容，它使用户可以在更高语义层次上快速浏览一段新闻节目中包含的信息。文[13]利用领域知识，通过图像分析方法来自动分割新闻条目，其算法核心是定位判别播音员镜头。由于假设每个新闻条目均由前导的播音员镜头与随后的一系列新闻镜头组成，其算法无法处理同一个播音员镜头中包含的多条

*)本文的研究工作得到国家自然科学基金重点项目(69789301)、国家863计划项目(863-306-ZT03-01-2)和中科院百人计划的资助。王伟强 博士生，主要研究领域：多媒体技术、人工智能。高 文 教授，博士生导师，主要研究领域：多媒体数据压缩、图像处理、计算机视觉、多模式接口、人工智能、虚拟现实等。马继涌 博士，主要研究领域：语音处理，模式识别等。林守勋 研究员，博士生导师，主要研究领域：CSCW，数据库，软件构件技术和数字图书馆。李锦涛 研究员，博士生导师，主要研究领域：智能信息网络等。

新闻，以及有的新闻条目不具有前导播音员镜头的情况。这种局限性仅通过对视觉信息的分析是无法克服的。本文提出一种结合音视频信息自动分割新闻条目的方法，克服了文[13]中算法无法对两种情况进行处理的不足，该方法提出了停顿频率概念，并结合静音率确定出在音频流上存在的静音片断组，然后与分析视觉信息获得的镜头分割结果、播音员镜头检测结果进行融合，确定出新闻条目间的分割点。

2. 基于视听信息分割新闻条目

2.1 概述

电视新闻节目是一类具有很强先验时间结构模型的视频节目，码流中包含的视音频信息均有自身的特点。对于视觉信息，通常镜头间的过渡为切变类型，虽然也会存在一些渐变式的镜头过渡类型如叠化，但它们的过渡时间一般较短，一般不超过5帧。这使得应用目前镜头分割技术会得到良好的分割结果，即较高的查全率与正确率。但同时由于新闻具有演播时间短，信息量大的特点，在制作剪辑时会尽量降低重复性的镜头出现，这样文[1,2]中提出的依赖对关键帧内容重复事件检测的场景分割方法，对于新闻条目的抽取效果不很显著。对于新闻节目的音频信息，通常它主要由音乐、语音、静音三种类型的片断来构成，其中最主要的是语音段，其内容是播音员对各项新闻条目的播报。有的新闻节目在新闻中间会插播一些广告，这里我们可以假设它不存在，因为可以通过一个预处理过程将其滤掉。通常在两个相邻新闻条目之间会存在一段相对较长的静音，同时与该静音音频片断相同步的视频帧片断中会出现镜头的过渡，值得注意的一种例外是，在播音员镜头中，新闻条目间的静音段并没有对应着镜头的改变。基于上面的特征，我们可以通过对音频信号的分析并结合对视频信号的分析来自动分割新闻条目。整个系统的结构图如图1所示。

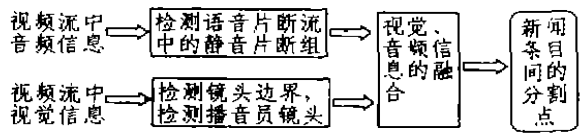

图1 新闻条目分割算法的结构图

2.2 音频流上静音片断组的检测

MPEG标准采用高性能的基于感知的编码方案来压缩编码音频流。标准定义了三种层次的编码方法，层次越高，编解码的计算复杂度也越高，但可获得更高的压缩率[14,15]。目前数字电视节目的音频流普遍采用层次Ⅱ，我们假定后面讨论中涉及的音频流均为层Ⅱ的。MPEG的音频基本流由一系列音频帧构成，每个音频帧包含有一定数目的采样，对于层Ⅱ为1152个采样值，对于每一个音频帧，我们可以利用式(1)计算出其短时平均幅度。

$$M_m=\frac{1}{N}\sum_{n=0}^{N-1}|x(n)| \tag{1}$$

其中 M_m 是第 m 帧的短时平均幅度，$x(n)$ 为帧内的音频采样值，N 为音频帧内的样本数。由于在MPEG音频解码过程中子带滤波数据的合成是一个线性过程，我们可以在频域中利用式(2)直接近似计算一音频帧的短时平均幅度。

$$M_m=\frac{1}{32*K}\sum_{i=0}^{32}\sum_{n=0}^{K-1}|s_i(n)| \tag{2}$$

其中 $s_i(n)$ 为第 i 个子带上的第 n 个采样值，K 为音频帧中各子带中包含的样点数。若 $M_m<\lambda$，λ 为一个门限值，则我们可判断第 m 帧处于一种短时静音状态。对于一个具有48k/s采样率的音频流，每个音频帧的持续时间约为24ms。为了分割出音频流上各种类型的音频段，需考察具有合适大小的小音频片断的特征属性。设小音频片断的预定长度为 T 秒，音频流中音频帧的长度为 τ 秒，则每个小音频片断包含约 $L=INT(T/\tau)$ 个音频帧，其中 $INT(\cdot)$ 为上取整函数。我们选取的小音频片断长约1秒。在播音员按照正常的语速播报新闻时，在字、词、句、不同新闻条目之间均会存在不同程度的停顿，对上述情况进行区分，并探测包含有音乐背景的音频片断，对于实现音频信息的分割具有重要的意义。为此，我们在小音频片断上定义了两个特征量：停顿频率 $PauseRate$，与静音率 $SilenceRatio$。设 $AC_i(i=0,1,\cdots)$ 为一个小音频片断，$af_{i,j}(j=0,1,\cdots,L-1)$ 表示属于 AC_i 的第 j 个音频帧，$M(af_{i,j})$ 表示音频帧的短时平均幅度，则定义函数 $Tag(i,j)$ 为：

$$Tag(i,j)=\begin{cases}1 & \text{若 } M(af_{i,j})\geq\lambda\\ 0 & \text{若 } M(af_{i,j})<\lambda\end{cases} \tag{3}$$

由此我们可以计算 AC_i 的停顿频率 $PauseRate(i)$：

$$C(i,j)=\begin{cases}1 & \text{若 } Tag(i,j)=0\text{且 } Tag(i,j-1)=1,j\geq1\\ 0 & \text{否则}\end{cases} \tag{4}$$

$$PauseRate(i)=\sum_{j=0}^{L-1}C(i,j) \tag{5}$$

以及静音率 $SilenceRatio(i)$：

$$SilenceRatio(i)=1-(\sum_{j=0}^{L-1}tag(i,j))/L \tag{6}$$

我们可以利用停顿频率 $PauseRate(i)$ 来分割出新闻视频中包含音乐或包含音乐背景的片断。对于一个音频片断，若它包含的所有小音频片断 $AC_i(i=s,s+$

$1,\cdots,t-1,t$)均具有如下属性 $PauseRate(i)=0$，$SilenceRatio(i)=0$，则可判断其为音乐类型片断。对于非音乐类型片断，通过算法1来判断可能是属于新闻条目间静音的音频小片断。

算法1 候选新闻条目间静音音频小片断的选取

```
IF(PauseRate(i)!=0)
   IF(SilenceRation(i)/PauseRate(i)>α)
      AC, will be chosen as a candidate silence clip
   ELSE
      AC, is not a candrdate silence clip
ELSE
   IF(SilenceRatio(i)>β)
      AC, wul be chosen as a candidate silence clip
   ELSE
      AC, is not a candidate silence clip
```

算法1中的 α,β 是门限值，$\beta>\alpha$，其取值与 $L*\tau$ 的值有关，$L*\tau$ 的值越大，α,β 的取值相应地变小。我们认为小音频片断时间长度 $L*\tau$ 的取值约为1秒，相应地经验选取 $\alpha=0.27$，$\beta=0.85$。对于一个小音频片断，在图2中我们给出了几种典型的 $Tag(i,j)$ 图像波形，及相应的 $PauseRate(i)$，$SilenceRatio(i)$ 的值。

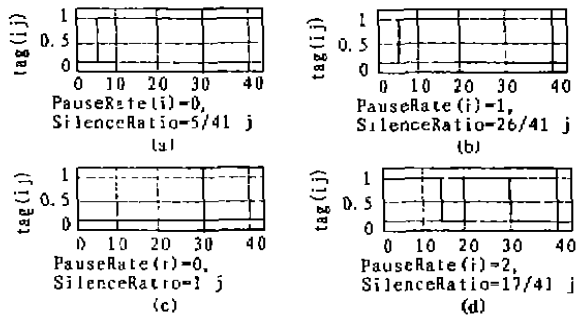

图2 几种典型 AC_i 的 $Tag(i,j)$ 及相应 $PauseRate(i)$，$SilenceRatio(i)$ 值

当我们找出那些候选静音小音频片断后，可以进一步确定出静音片断组 $SG(n)$，它可用一个二元组表示，即 $SG(n)=(s_n,e_n)$，$n,s_n,e_n=1,2,\cdots$ 且 $s_n\leqslant e_n$，其含义为由 $AC_{s_n-1}AC_{s_n}\cdots AC_{e_n}AC_{e_n+1}$ 构成的音频片断中，有且仅有 AC_{s_n-1}，AC_{e_n+1} 为非候选静音小音频片断。我们认为与每个静音片断组 $SG(n)$ 对应的视频片断是最可能发生新闻条目切换的地方。

2.3 镜头分割与播音员镜头的探测

在新闻节目中，通常有两种新闻内容播报方式，一种是在播音员画面下，通过声音流来承载新闻条目的内容；另一种在非播音员画面下，通过同步的声音与画面共同来刻画新闻的内容。对于前者，若同一播音员镜头包含多条新闻，则视觉信息不能为新闻条目的分割提供任何线索。而对于后一种情况，一条新闻条目的结束与下一新闻条目的开始间必然会伴有镜头的切换（但反之则不成立）。在该情况下镜头的分割点可为我们提供用于新闻条目分割的重要信息。我们可以利用目前镜头分割技术[3~5]来获得这些潜在的新闻条目分割点。

在新闻条目分割的计算中，视觉内容分析的一项重要内容是对播音员镜头的检测，因为检测的结果可用于区分在不同的情况下依赖哪些信息源来分割新闻条目。Zhang 等人在文[13]中给出了一种检测播音员镜头的算法，该方法的检测性能受镜头分割结果准确性的影响，在空域上进行分析计算，计算相对复杂。我们提出了一种基于背景色彩及人脸肤色模型的播音员镜头检测方法，该算法具有不涉及镜头分割预处理，计算简单且在压缩域中进行分析的优点。在大数据测试集上的评估实验表明，该算法具有非常理想的探测能力，具有100%的查全率，98.9%的正确率，同时具有约77.55帧/秒的超实时探测速度。由于本文的主旨，这里限于篇幅不对该算法进行详细描述，细节将另撰文讨论。这里可假设正确的播音员镜头检测信息已经获得。

2.4 新闻条目的自动分割

利用2.2、2.3节描述的过程，我们可以获得用于分割新闻条目的重要视、音频信息。对于音频流的每个静音片断组 $SG(n)=(s_n,e_n)$，设与其同步播放的视频流片断为 (Vs_n,Ve_n)，Vs_n、Ve_n 分别为其起始帧号与结束帧号。我们采用两个步骤来实现新闻条目的自动分割。首先，通过对播音员镜头的检测过程将整个视频流分成两类片断，即：①播音员镜头片断，主要由声音流来表现新闻内容；②非播音员镜头新闻片断，由同步的声音、画面共同表现新闻内容。图3给出了一天 CCTV 新闻经上述过程处理后的结果示意图。

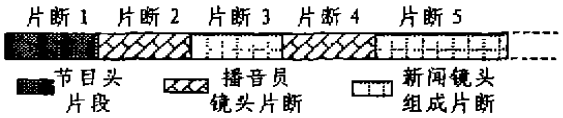

图3 利用播音员镜头的检测形成的粗分割结果示意图

然后我们基于如下准则对前面产生的片断作进一步的细化处理，判定每一个片断中是否存在新闻条目的分割点。

准则1 若与 $SG(n)=(s_n,e_n)$ 对应的视频流片断 (Vs_n,Ve_n) 属于同一个播音员镜头，且该静音片断组 $SG(n)$ 的总静音率 $\eta=\sum_{i=s_n}^{e_n}SilenceRatio(i)$ 满足 $\eta>\zeta$，ζ 为门限值，则将序号为 $INT((Vs_n+Ve_n)/2)$ 的帧作为新闻条目的一个分割点。

准则2 若与 $SG(n)=(s_n,e_n)$ 对应的视频流片断 (Vs_n,Ve_n) 跨越两个不同的镜头 $shot(k)$ 与 $shot(k+1)$，

$k=0,1,\cdots$，且其总静音率 $\eta=\sum_{i=i_1}^{i_2} SilenceRatio(i)$ 满足 $\eta>\sigma$，σ 为门限值，则判定 $shot(k)$ 是某个新闻条目的最后一个镜头，$shot(k+1)$ 是随后新闻条目的起始镜头。

若不存在静音片断组 $SG(n)$ 在某个播音员镜头中间，则该播音员镜头片断构成一个新闻条目。若某个静音片断组 $SG(n)$ 存在于一个非播音员镜头中间，则该静音片断组 $SG(n)$ 仅代表它是某个新闻条目播报中间的一种语音停顿。

事实上，由于新闻节目内容安排一般比较紧凑，声音流中处于静音状态的时间一般比较短暂，一般不会超过3.5秒。同时，画面上为了给观众留下深刻印象和视觉上的舒适感，在3.5秒中不会出现两次镜头的切换，因此可以假设 $SG(n)$ 对应的视频流片断 (Vs_n,Ve_n) 至多跨越两个镜头。

3. 实验结果及分析

基于上述思想，我们实现了基于视音频信息的新闻条目分割算法，并从我们建立的视频数据库中，随机选取了4天的 CCTV 新闻联播节目作为测试数据集，共包含184,100帧。测试前，通过多次观看这些新闻节目，并利用工具对各新闻条目的起始位置进行了手工标注，作为标准参照。在标注的数据中，我们将同一报道内容的播音员镜头与非播音员镜头看作是不同的新闻条目。

表1 算法对新闻条目分割点检测性能的测试实验结果

视频流	实际包含的镜头数目	实际新闻条目分割点(S)	探测出的新闻条目分割点(D)	探测错误的新闻条目分割点(E)	未探测到的新闻条目分割点(U)
NewsA	376	33	40	8	1
NewsB	243	24	34	11	1
NewsC	305	28	33	7	2
NewsD	274	18	24	7	1
总计	1098	103	131	33	5

在测试实验中，我们在参数选取时，倾向于在获得较高的查全率的基础上，获得合适的正确率。为此，我们选取 $\alpha=0.27$，$\beta=0.85$，$\zeta=1.6$，$\sigma=1.13$，相应的测试实验的结果列于表1中。从中我们可以得到有关系统对新闻条目分割点检测的平均正确率 $P=1-E/D=1-33/131=74.8\%$，及查全率 $R=1-U/S=1-5/103=95.1\%$。我们在实验中发现，一般地，ζ，σ 取值越小，查全率越高，正确率则越低。在获得较高的查全率的同时，也会引入了较多的错误检测。

未探测到的分割点，主要发生于两个非播音员镜头的新闻条目衔接时出现短暂的背景声音的干扰，使得静音段检测不到；而检测错误的情况同样多发生于非播音员镜头情况下真实场景声音的存在，例如采访场景，采访者与被采访者之间说话者的变换造成了候选静音片断组的存在，若同时伴有镜头的切换，则会出现误检测。

属于文[13]中算法无法检测新闻条目的两种情况在 CCTV 新闻中普遍存在。本文算法对它们探测性能的统计数据列于表2。从表2不难看出，本算法可以对绝大多数上述两种情况作出正确检测，说明了本算法的有效性，但同时也存在许多虚假的检测(见表1)，表1中的全部虚假检测也是在针对 A、B 两种情况的检测中产生的。

表3中给出了仅通过音频信号分析产生的静音片断组的数目，以及结合视觉上的镜头分割信息产生的有效静音片断组(即跨越镜头边界)的数目。通过两者的对比，表明视觉特征信息可帮助有效地过滤掉绝大多数无效的静音片断组，两种媒体信息综合分析的效果是明显的。

表2 算法对两种复杂情况的检测结果

视频流	实际新闻条目分割点(S)		探测到的新闻条目分割点(U)	
	A	B	A	B
NewsA	3	3	3	2
NewsB	2	5	2	4
NewsC	2	3	1	2
NewsD	0	6	0	5
总计	7	17	6	13

说明，A—同一个播音员镜头中包含多条新闻的情况；
B—新闻条目不具有前导播音员镜头的情况

表3 视音频媒体信息的融合对音频信息的过滤增强作用

视频流	NewsA	NewsB	NewsC	NewsD
音频分析产生的静音片断组的数目	187	160	174	173
结合镜头分割结果所获得的有效静音片断组的数目	33	33	27	22

结束语 本文尝试采用视觉、声音分析相结合的方式来实现新闻条目的自动抽取。提出的方法克服了文[13]中算法对一些情况无法处理的不足，对于新闻条目自动分割具有更广泛的适应性。我们的实验结果表明了算法在该方面的有效性，获得了对新闻条目分割点的95.1%查全率，74.8%的正确率。同时表明采用

视音频综合分析方式是实现对新闻节目进行高层语义结构分析的一种有效思路。虽然本方法是为解析新闻视频而专门设计的，但其中一些音频信号分析方法，以及视音频信息融合策略也可应用到其他视频节目的场景分析中去。

参 考 文 献

1 Yeung M M, Yeo B L. Time-constrained clustering for segmentation of video into story units. In: Int. Conf. Pattern Recognition(ICPR'96), Aug. 1996. 375～380

2 Hanjalic A, Lagendijk R L, Biemond J. Automatically Segmenting Movies into Logical Story Units. In: Proc. of the Third Intl. Conf. VISUAL'99, Amsterdam (NL), June 1999

3 Yeo B L, Liu B. Rapid Scene Analysis on Compressed Videos. IEEE Transactions on Circuits and Systems for Video Technology, 1995, 5(6): 533～544

4 Boreczk J S, Rowe L A. Comparison of Video Shot Boundary Detection Techniques. In: Proc. SPIE Conf. Storage and Retrieval for Video Databases IV. San Jose, CA, Feb. 1995

5 Zabih R, Miller J, Mai K. A Feature-Based Algorithm for Detecting and Classifying Production Effects. Multimedia System, 1999, 7: 119～128

6 Wolf W. Key frame selection by motion analysis. In: Proc. IEEE Int. Conf. Acoust. Speech, and Signal Proc. 1996

7 Gresle P O, Huang T S. Gisting of video documents: A key frames selection algorithm using relative activity measure. In: The 2nd Int. Conf. on Visual Information Systems, 1997

8 Zhuang Y, Rui Y, Huang T S, Mehrotra S. Adaptive Key Frame Extraction Using Unsupervised Clustering. In: Proc. of IEEE Int Conf. on Image Processing. Chicago, IL, 1998. 866～870

9 Girgensohn A, Boreczky J. Time-Constrained Keyframe Selection Technique. In IEEE Multimedia Systems'99, IEEE Computer Society, 1999, 1: 756～761

10 Patel N, Sethi I. Audio Characterization for Video Indexing. In: Proc SPIE on Storage and Retrieval for Still Image and Video Databases, 1996, 2670: 373～384

11 Liu Z, Huang J, Wang Y, Chen T. Audio Feature Extraction & Analysis for Scene Classification. IEEE Signal Processing Society 1997 Workshop on Multimedia Signal Processing, Princeton, New Jersey, USA, 1997. 23～25

12 Boreczk J S, Wilcox L D. A Hidden Markov Model Framework for Video Segmentation Using Audio and Image Features. In: Proc. of ICASSP'98, Seattle, 1998. 3741～3744

13 Zhang H J, Tan S Y, Smoliar S W, Gong Y. Automatic Parsing and Indexing of News Video. Multimedia Systems, 1995, 2: 256～266

14 马小虎，张明敏，严华明. 多媒体数据压缩标准及实现. 清华大学出版社，1996

15 钟玉琢，乔秉新，祁卫译. 运动图像及其伴音通用编码国际标准-MPEG-2(ISO/IEC13818). 清华大学出版社，1997

(上接第123页)

策2，…，决策n)，可用此N项决策的民主化程度表示一个群体的民主化程度。

从以上讨论可以看到，我们将人类历史上从来没有量化的概念，民主概念量化了，这样量化了的概念才能让我们最终在计算机上实现处理。尽管人文和社会科学中的概念具有很大的模糊性，但这正好说明问题的复杂性，模糊概念具有模糊性是很自然的事，我们的任务是如何把它用数学语言表达出来，量化的过程也是一个研究的过程，相信将来会有越来越多的定性概念可以进行定量研究。

参 考 文 献

1 Zadeh L. A. Fuzzy sets. Information and Control, 1965, 8: 338～353

2 Huntington S P. The Third Wave, Democratization in the Late Twentieth Century. University of Oklahoma Press, 1991

3 Freud S. translation by A. A. Brill. The Interpretation of Dreams. Being: China Social Sciences Publishing House, 1999

4 Gupta M M, Sanchez E. Fuzzy Information and Decision Processes. New York, Elsevier Science Pub. Co., 1982