新闻视频故事单元分割技术综述

冀 中 张春田 苏育挺
(天津大学电子信息工程学院，天津 300072)

摘 要 新闻视频的故事单元分割一般采用统计学或者信息论的方法，将新闻节目分割成一系列有各自主题内容的故事单元。这些单元反映的是视频流的高层语义，是建立视频索引的最佳层次。该文对这一技术进行了综述，将现有方法根据利用信息的角度分为 3 类：单模态的分割方法、多模态融合的分割方法和基于上下文信息的分割方法，并且详细讨论了每一类方法的特点。此外，还分析了一些分割错误的原因和今后的发展趋势。

关键词 故事单元分割 基于内容的视频检索 新闻视频 上下文信息
中图法分类号：TP391 **文献标识码**：A **文章编号**：1006-8961(2007)11-1952-09

News Video Story Segmentation Technique: An Overview

JI Zhong, ZHANG Chun-tian, SU Yu-ting
(*School of Electronic and Information Engineering, Tianjin University, Tianjin* 300072)

Abstract As a challenging technique, news video story segmentation divides news programs into a series of stories by statistical techniques or information theory methods. It provides the best semantic units for video indexing. According to the information utilized, this paper gives an overview of the existing methods and classifies them into three categories: the single-modality approach, the multi-modality fusion approach and the context-based approach. Their relative advantages and limitations are discussed in detail and segmentation errors are analyzed. In the end, some trends in this field are introduced.

Keywords story segmentation, content-based video retrieval, news video, contextual

1 引 言

近年来数据压缩、通讯、存储技术的飞速发展，以及计算机性能的不断提高和网络的广泛普及，使得多媒体视频的应用得到了极大的发展。数字图书馆、远程教育、家庭娱乐以及视频广播等使用了大量的视频数据，如何快速有效地浏览这些视频以及如何自动地从这些海量数据中检索到某一感兴趣的视频片段，在时间和效率就是生命和效益的现代社会，已成为亟待解决的问题。视频摘要和视频检索是解决这些问题的主要方法，而其中一个重要的步骤就是视频结构分割。实际上，因为视频是一种非结构化的媒体，因而视频结构分割是执行任何数字视频内容管理的前提环节，目的是将视频分解为一系列有意义可管理的片段，作为进一步分析和处理的基本元素。通常按照视频内容粒度可以把视频分解为两层基本单元——镜头(shot)层和故事单元(story)层。镜头是指由一个摄像机镜头连续拍摄的一组内在相关的连续帧，它用来表现在时空上连续的一组运动。镜头分割是视频处理的第一步，具有十分重要的意义，有关镜头分割的研究很多，也较为深入[1,2]。然而一个镜头往往不能表达完整的语义信息，这就需要在镜头之上再划分一个粒度较大的单元——故事单元，使每个故事单元能够表达同一主题，这样更方便人们的理解。故事单元反映的是视频流的高层语义，更符合人类的思维模式，是建立视频索引的最佳层次。因此，基于内容的故事单元层

收稿日期：2006-05-26；改回日期：2006-06-26

第一作者简介：冀中(1979～)，男。天津大学信号与信息处理专业博士研究生。主要研究方向为视频内容分析、图像处理、视频压缩。E-mail: ji.zhong@hotmail.com

的正确分割,对建立视频数据库系统,实现基于内容检索具有重要的意义。

根据视频内容所关联的领域的不同,故事单元分割的方法也不尽相同,可以分为新闻节目故事单元分割、电影故事单元分割以及家庭娱乐视频场景分割等。新闻视频节目因其内容的丰富、实用以及结构的有序、固定等诸多有益特点,已经成为基于内容的故事单元分割的研究热点和重点。典型的新闻节目的结构如图 1 所示。

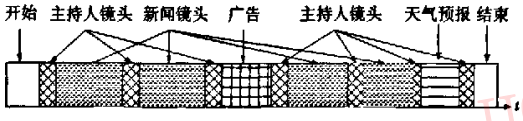


图 1　新闻视频节目的时序结构

Fig. 1　The temporal structure of a typical news program

新闻视频的故事单元分割是一种新兴的技术,吸引了来自多媒体、视频处理、信息检索等不同领域学者的极大兴趣,是一项极具挑战性的工作。作为早期工作之一,Zhang 等人于 1994 年提出的新闻视频分析系统将新加坡 SBC 新闻按类分割为主持人镜头、新闻镜头、广告镜头和起始结束镜头等[3]。Merlino 等人利用故事单元分割技术研究了新闻节目的编辑和导航系统[4]。而自国际影视检索评测 TRECVID(TREC Video Retrieval Evaluation)于 2003 年提供了标准的数据库和评测标准之后一直到现在,才掀起了研究新闻视频故事单元分割的高潮。许多研究人员在该方向都投入了大量的研究,并且取得了相当大的进步。

2　TRECVID 中的故事单元分割任务简介

TRECVID 是影视检索领域的国际性权威评测,目的是通过公开公正的评测机制,评测每个参评单位所提交任务系统的性能情况,促进基于内容的视频检索技术的发展。2003 年,TRECVID 脱离于 TREC(text retrieval conference)成为一项独立的评测活动,并于 2003 年和 2004 年连续两年将新闻视频的故事单元分割作为一项独立的任务[5,6]。新闻故事单元(news story)的定义为一个在新闻内容上相关的,至少带有两个独立声明性语句的片段。非新闻片段标记为"其他"(miscellaneous),当它们相邻时予以合并,标注为一个故事单元[4,5]。该定义源自语言数据联盟(linguistic data consortium, LDC)对故事单元的定义[7],也就是说,该定义针对的是新闻文本,但是 TRECVID 把它借用到了视频中。可以这样理解,新闻视频中的新闻故事单元是一个在新闻内容上相关、描述一个完整事件的视频片段,包括一些政治事件、财经报道、天气预报、体育报道等等。而非新闻故事单元就是除了新闻单元之外的部分,如广告、新闻片头等。一个故事单元可能有多个镜头组成,比如主持人先对某段新闻做大概介绍,然后镜头转移到现场做详细报道,最后镜头又转回直播间由主持人做某些评论等。另一方面,一个镜头也可能包含多个故事单元,如在一个主持人镜头中连续播报几条新闻。实验和评估采用的数据是"CNN Headline News"和"ABC World News Tonight",格式为 MPEG-1。2003 年的测试集中包含 104 段新闻(52h)共有 2 929个故事边界,而 2004 年的测试集不同于上一年,包含 118 段(59h)共3 105个故事边界。通常一段新闻的故事单元个数在 14 ~ 42 个之间[8]。这些数据中一些常见的故事单元的类型如图 2 所示[9]。

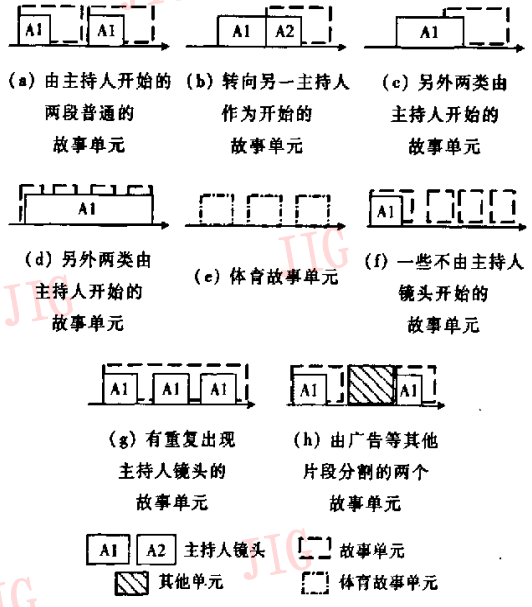


图 2　TRECVID 测试集中常见的故事类型

Fig. 2　Common story types seen in TRECVID test data

在 TRECVID 的故事单元分割任务中,根据利用的信息资源,每个参评单位必须提交 3 类测试结果:(1)利用音频和视频(AV)信息;(2)利用音频、视频和自动语音识别来得到 ASR(automatic speech recognition) transcripts 文本信息(AV + ASR);(3)

只利用 ASR transcripts 信息。每类至少提交一项结果,根据方法或者参数的不同,也可以提交多项结果。数据集中的所有故事单元均由人工分割,得到的故事边界点作为参考边界点。在每个参考边界点前后两个方向上分别延拓 5s,这样就形成一个 10s 的模糊窗口。当计算所得的边界点落在这个窗口内的时候,就认为检测到该参考边界点,否则认为是一个误检测。

每种方法用查准率(precision,也称精确率)、查全率(Recall,也称召回率)和 F-Measure(F)评估性能。查准率 P 定义为正确检测到的边界数量 r 与全部检测到的边界数量 N(包括正确检测的 r 和误检的 n)之比。查全率 R 定义为正确检测到的边界数量 r 与参考边界数量 M(包括正确检测的 r 和误检的 m)之比。F 是信息检索等领域的一种系统性能测试指标,是综合查准率和查全率的一种系统评价指标。从式(3)中可以发现,需要查准率和查全率都很高时,F 才能得到较高的值,任何一者都不能偏废。

$$P = \frac{r}{r+n} \tag{1}$$

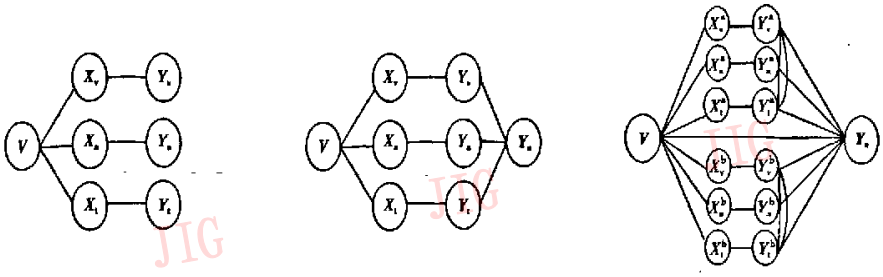
$$R = \frac{r}{r+m} \tag{2}$$

$$F = \frac{2 \times P \times R}{P+R} \tag{3}$$

实际上,大部分 2003 年以后的工作都是采用 TRECVID 提供的数据和评测标准进行实验的,而在此之前,虽然评测标准大多采用查全率和查准率,但实验使用的数据一般比较少,并不能客观的反映实验方法的有效性。

3 新闻视频故事单元分割技术

对新闻视频故事单元分割技术而言,可以利用 3 类信息,即音频、视频和文本信息。其中,文本信息又包括 ASR 文本信息和字幕(caption)信息。近年来对新闻视频故事单元分割技术的研究产生了大量的方法,根据利用信息的角度我们将之分为 3 大类:单模态的分割方法、多模态融合的分割方法以及基于上下文(context)信息的分割方法如图 3 所示[10]。图中符号 X_v、X_a、X_t 分别表示视觉信息、听觉信息、文本信息,Y_v、Y_a、Y_t 为相对应的求得的故事单元边界,V 表示输入的新闻视频,Y_a 表示最终分割的边界集合,上标 b 和 a 分别表示候选边界点前边和后边的视频片段。



(a) 单模态的方法 (b) 多模态融合的方法 (c) 基于上下文信息的方法

图 3 3 种新闻视频故事单元分割模型

Fig. 3 Various story segmentation models

3.1 单模态的分割方法

这种方法单独使用音频、视频和文本信息中的一种,从中提取特征用来分割故事单元。因此,可以分为 3 小类:基于文本信息的方法,基于音频信息的方法和基于视觉信息的方法。

3.1.1 基于文本信息的分割方法

基于文本信息的方法是最早应用到新闻故事单元分割中的方法,开始研究对象并不是新闻视频,而是新闻专线(newswire)或广播新闻文本。自从应用了 LIMSI 提供的 ASR 系统[11],新闻视频中的语音得以较顺利的转化为文本,才开始了大规模的对新闻视频中 ASR 文本的故事单元分割的研究。但无论处理对象是新闻专线还是 ASR 文本,它们应用的分割方法基本上是相同的。不同的是,对 ASR 文本分割后得到的边界点还要进一步对应到新闻视频中去。

典型的基于文本信息的分割方法是检测文本块中词汇内容间的相关性,以之判断新闻故事边界,进

行主题划分[12,13]。还有一类使用机器学习的分割方法,例如隐马尔科夫模型(HMM)[14],最大熵模型(ME)[15],以及最大熵与决策树(DT)的结合[16]等。

基于文本信息的方法 F 一般在 30% ~50% 左右[5,6,8],这对新闻故事分割而言是远远不够的。而且新闻视频中除语音外还有非语音(音乐,背景噪声等)的存在,有时还会同时出现多人的声音,所以对新闻视频的自动语音识别要比传统的对语音的识别难度大的多,因此单纯根据 ASR 文本来确定视频中新闻单元的边界效果肯定要差一些。实验表明与利用音视频复合特征的方法相比较,基于文本信息的方法性能明显不好。

3.1.2　基于音频信息的分割方法

该方法是对视频中的音频进行分割得到音频中故事单元的边界点,以此达到分割或者辅助分割视频故事单元的目的。基于音频的新闻故事单元分割的一般的方法如图 4 所示。首先从输入的新闻视频中提取有效的音频特征;其次根据这些特征对音频流分段并按照听觉特征判断每段类别,如语音、音乐、噪声、静默等等;最后根据这些片段的类别信息结合新闻视频在内容上和时间上的特性对故事单元分割并予以分类,如含有语音、歌声,并且有背景音乐的一段连续的片段一般是广告单元。

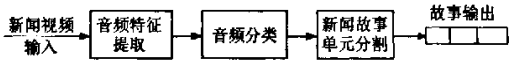


图 4　通过音频分类进行新闻视频故事单元分割的一般方法

Fig. 4　A unified approach to content-based audio analysis for news video segmentation

常见的音频特征主要有:过零率(zero crossing rate, ZCR),短时能量(short time energy, STE),响度(loudness),音调(pitch),频谱通量(spectrum flux, SF),线性预测系数(linear prediction coefficient, LPC)以及梅尔频率倒频系数(mel frequency cepstrum coefficient, MFCC)等。其中,短时能量是检测静默区的一个重要的特征,而基于频谱的那些特征有较强的区分语音和音乐信号的能力。

由图 4 可知,音频分类是其中的一个关键的步骤。作为音频处理的一个重要方向,音频分类主要应用于以音频形式存储的媒体,如采访录音,广播新闻等,但同时也为新闻视频中的故事单元分割提供了重要的手段。大多数研究人员采用基于监督学习的方法,如隐马尔科夫模型、支持向量机(SVM)等来进行音频分类。Liu 等人在 8 个基本音频特征中计算了 14 个高层次的音频特征,用含有 5 个状态的隐马尔科夫模型把电视节目划分为新闻、广告、足球、篮球和天气预报几个部分,不同条件下的最高准确率可达 84.7%[17]。支持向量机是一个有很强分类能力的技术,Lu 等人使用它进行新闻节目中音频的分割和分类并且得到了比较满意的结果[18]。此外,他们还对 3 类音频类型:语音/非语音、纯语音/非纯语音以及音乐/环境声音,分别用 K 近邻法(KNN)和混合高斯模型法(GMM)与支持向量机法进行了比较,实验结果表明支持向量机的方法明显好于其他两种方法。

虽然支持向量机、隐马尔科夫模型等这些监督学习的方法在音频分类方面比较有效,但是还是存在着较强的局限性。首先,对训练样本依赖性很强。如果这些训练样本不充分或者分布不全面,系统的性能就会显著下降。其次,在某些实际应用中,很难把所有的音频类别都列举出来。比如在监控领域和普适计算领域,事先不能确定音频中会有一些什么内容,所以无法对类别进行事先划分也不能获得数据来训练模型。

基于音频信息的方法从音频的角度对新闻视频分割,对于广告,天气预报以及新闻片头等检测比较有效。但它只能作为新闻视频故事单元分割的一个有效的辅助手段,并不能达到满意的分割效果。

3.1.3　基于视觉信息的分割方法

从图 1 中可以看到,一般的新闻故事单元都由"主持人镜头 + 新闻镜头"这样的结构构成。通常新闻镜头不具有固定结构,而主持人镜头却具有明显的空间特征,一般在整个镜头中均由主持人,台标以及较为固定的背景等构成,因而基于视觉信息的方法一般集中在主持人镜头检测方面,认为每个主持人镜头表示着一个新闻故事单元的开始。

一般的主持人镜头检测的过程如图 5 所示。首先,通过某种镜头检测的方法将视频分割成镜头;其次,通过一些规则,比如主持人镜头一般持续时间长于 2s 等滤除一部分镜头以便减小计算量;然后,从那些剩余的镜头中提取关键帧代表镜头作为分析处

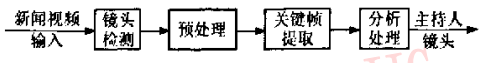


图 5　主持人镜头检测的一般方法

Fig. 5　Diagram of anchorperson shot detection scheme

理的对象;最后,从那些关键帧中提取一些视觉特征利用模板匹配[3,19]、聚类[20]等方法检测主持人关键帧,就可以把新闻视频分割成以主持人镜头开始的故事单元。

显然,基于主持人镜头检测的新闻视频故事单元分割的方法只能够检测到图 2 所示的部分类别的故事单元,对其他几种类型存在着严重的误检(如图 2(d)、图 2(g)所示)和漏检(如图 2(e)、图 2(f)所示)。但是,在大多数新闻视频中,由主持人镜头开始的故事单元最少可以占到全部类型的三分之一以上[21],是所有类型中的主要部分,所以基于视觉信息的主持人镜头检测仍是故事单元分割中的重要方法之一。

总之,单一特征用于新闻故事单元分割的方法有它各自适用的范围,同时也为对新闻视频的分割从不同角度提供了分割的方法。但是,因为使用的特征单一、信息不充分,所以分割效果不理想。为了提高检测的准确性,必须利用多模态融合的方法。

3.2 多模态融合的分割方法

该类方法利用视觉、听觉和文本 3 种特征中的两种或者全部进行基于内容的新闻视频的故事单元的分割。主要思想就是首先提取尽可能多的有效特征;然后建立一个框架,利用新闻视频的一些先验规则,有效地进行多模态特征的数据融合;最后通过一些决策机制判断故事单元的边界点位置。可将这类方法细分为基于规则的和基于统计的分割方法。

3.2.1 基于规则的多模态融合的分割方法

一些重要的和新闻故事边界相关的领域知识有:主持人镜头,静默区,语音的韵律(prosody)以及文本中的线索性短语等。有关主持人镜头已经在 3.1.3 小节中介绍过,静默区指那些在新闻视频中有较长听觉停顿的部分,通常指示可能的镜头或故事单元的转换。文本中的线索性短语指诸如"欢迎观看××新闻"、"接下来请看天气预报"等经常出现在某类故事单元开始部分的短语。而语音的韵律指新闻主持人的语速和语调,这也是检测故事单元变换的一个重要的线索。例如,在图 2(d)所示的类型中,同一主持人在一个镜头中会播报几段新闻,构成几段新闻故事单元。在一段新闻播报将近结束时,主持人一般会降低语调、放慢语速,经过短暂的停顿后,主持人一般会提高语调、加快语速来播报下一段新闻。实验还证明了该特征不受语言和主持人性别的约束[22],因此适用性很广。而基于规则的方法就是利用这些先验知识中的一种或几种作为分割的主要依据[23-26]。

Eichmann 等人在镜头检测的基础上,认为静默区和线索性短语同时出现处就是故事单元的边界点[23],当然这种方法过于简单,F 只能达到 38%。Wang 等人假设每个故事单元存在一个主题字幕帧,并且在此字幕帧的前边或者后边必定存在一个或者几个静默区。在两个主题字幕帧之间,如果镜头边界在静默区内,就认为该点是故事边界点[24]。虽然实验得到了较好的效果,但是测试数据只有 1.5h 的 CCTV 新闻,数据远远不够。

主持人镜头检测同样是基于规则方法中的一个重要部分,同 3.1.3 小节不同的是,除了视觉特征外,还要利用音频或文本的一些特征,在加大计算复杂度的前提下提高了检测性能[25,26]。不过有实验表明,如果主持人镜头检测能够达到 100% 的准确率,F 也就只能达到 51%[9] 或 62%[27]。

一般来说,基于规则的方法相对简单,实现也比较容易,但是性能差一些。而且由于新闻节目非常多,不同新闻的特征或者规则不尽相同,所以基于规则的方法适用性弱,推广性差。例如,国内新闻很少含有广告,而国外新闻中常有广告镜头;CCTV-1 中的大多数新闻类型如图 2(a)和图 2(f)所示,而"ABC World News Tonight"中类型在图 2(a)和图 2(g)中占大部分;此外,CCTV-1 和 ABC 的新闻中故事单元一般有多个镜头组成,而"CNN Headline News"中经常会出现一个镜头中存在几个故事单元的情况。

3.2.2 基于统计的多模态融合的分割方法

为了达到比较满意的分割效果,除了利用规则之外,还必须采用一些更有效的统计学习的手段,有效地融合各个模态的多个特征。实际上,基于统计的方法是目前新闻故事单元分割中最常用的一种方法。

基于支持向量机的方法首先分割镜头;然后从每个镜头中提取特征形成"镜头矢量"作为支持向量机的输入,训练时认为每个含有故事边界的镜头为正,不含有的为负;测试时,将待测试镜头的"镜头矢量"输入到支持向量机中,进而判断该镜头是否包含故事边界[28,29]。Browne 等人通过主持人检测、人脸检测和运动分析这 3 个模块分别给每个镜头计算一个值(在 0 和 1 之间),还通过文本分析对每个镜头产生一个二进制的值(非 0 即 1)再加上镜

头长度(归一化)等特征一同输入支持向量机,进而判断每个镜头是否含有故事边界。实验的查全率为31%,查准率为45%,F 为38%[28]。Hoashi 等人采用了类似的方法检测一个镜头中是否含有故事边界,不过只使用了视觉和听觉的特征[29]。此外,为了提高分割的性能,还采用了3个后处理过程来处理某些特殊的故事单元。实验取得了较好的效果,F 可达69%。

Chaisorn 等人提出了一个两级的多模态融合的故事单元分割框架[13,30](如图6所示)。由图6可以看出,最基本的处理单元是镜头,关键的步骤是镜头分类。在镜头层,首先使用一系列特征的组合来表征镜头,如颜色直方图、背景变化、说话人变化、运动、音频类型和镜头持续时间长度,以及人脸,线索性短语,镜头类型等高层次的特征。然后通过广告单元检测器等一些专门的检测过程确定一些镜头种类,进而采用决策树对所有镜头进行分类并予以标注。镜头分类是一个比较热门的研究方向,较早的研究工作只把镜头分成主持人、体育、天气和广告等几类。为了比较全面地对新闻镜头分类,作者把所有镜头划分为17类:新闻片头、主持人、双主持人、人物、讲演/采访、现场报道、体育、文本、经济、天气、广告等等。在故事单元分割层,根据镜头类型、线索性短语以及镜头间场景变化等特征应用隐马尔科夫模型进行故事单元分割。实验结果表明只使用视觉和听觉特征时 F 为75%,而再利用文本特征后,F 可达77%的良好效果。该方法将问题分解为镜头分类和故事分割两步来做,有效地消除了机器学习过程中的数据稀疏(data sparseness)问题,而且算法比较直观,效果也比较好。但是这种方法以镜头作为最小的故事单元单位,致使同一镜头中包含多个故事的情况无法分割。再有,因为需要人工定义镜头种类并且需要特殊的检测器检测,导致算法的推广能力较弱。

以上两种方法都是以镜头为研究对象,或者判断镜头中是否含有故事边界,或者在镜头分类的基础上分割故事,而 Hsu 等人提出了一种不同的方法——基于边界的方法[9,16,31]。这种方法首先选择

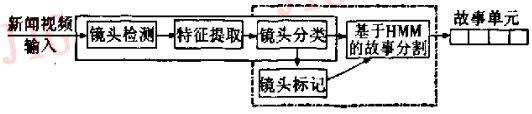


图6　两级多模态融合的故事单元分割框架

Fig. 6　A two-level, multi-modal news video segmentation framework

一些边界候选点,然后根据这些点周围的特征为这些点建立最大熵模型,最后判断哪些点是故事边界点。候选边界点必须有很高的查全率,也就是说必须包括所有的参考故事边界点,而且还应该是那些比较明显的有效特征出现的地方。镜头边界比较符合这些要求,但不能包含全部的边界点,所以作者采用镜头边界和静默区的组合作为候选边界点,并且应用一个5s大小的窗口将窗口内有重合的多余点去掉。然后,在候选点的周围选取一些诸如人脸、运动、语调、文本等的原始特征送入一个特征处理器中产生一个二进制特征函数 $f_i(x,b)$。最后,使用最大熵模型将这些二进制特征进行融合,把判断该候选边界点是否是故事边界的问题转换成求后验概率 $p_\lambda(b|x)$,其中,x 表示该候选边界点周围的原始特征,$b\in\{0,1\}$ 是当 x 出现时与存在或者不存在故事边界相对应的随机变量。故事单元边界存在的后验概率可以表示为

$$p_\lambda(b|x)=\frac{1}{Z_\lambda(x)}\exp\left\{\sum_i \lambda_i f_i(x,b)\right\} \tag{4}$$

其中,$Z_\lambda(x)$ 是一个标准化因子,用来确保方程是有效的条件概率分布。λ_i 是第 i 个二进制特征的权重因子。实验结果证明了方法的有效性:只使用视觉和听觉特征时,F 为69%,而使用视觉、听觉和文本特征时,F 可达74%。通过最大熵这种灵活有效的统计学框架,该方法强调了一种自动的从原始特征集中选取有效特征并且进行多特征融合的系统的解决方案。这种基于信息论的模型避免了采用专门领域知识进行手动分类的过程,有较好的推广能力。不过,因为分割的目的在于索引和检索,所以这种方法还要在分割之后进行故事单元的语义标注。

Hsu 等人比较了支持向量机、最大熵、Boosting 以及最大熵与 Boosting 相结合(BoostME)4种方法的数据融合和故事单元分割能力[21]。结果表明,基于支持向量机的方法分类能力最强,而基于最大熵的方法在选取特征方面最有效。此外,在故事分割方面,基于支持向量机的方法稍好于基于最大熵的方法,明显好于单纯的基于 Boosting 的方法。

3.3　基于上下文信息的分割方法

人类具有理解上下文语境含义的能力,这种能力使我们能够理解一个词在某个句子中的确切含义,能够想象出一个模糊的图片可能表示的意义。同样故事单元分割过程中的上下文知识也能够提高现有方法的性能。上下文可以定义为一系列连接语

义概念的关系，为故事单元分割提供了除文本、听觉、视觉特征外的其他有用信息。一些语义概念经常会同时出现，例如，根据视觉信息标注为“主持人”的片段，其音频一般会标注为“语音”。另外，还有一些语义概念几乎不会同时出现，例如，根据音频信息标注为“噪声”的片段，其文本不可能标注为“天气”。通过明确的定义新闻片段间的联系，所有这些上下文信息都可以应用到新闻视频的故事单元分割中提高现有方法的性能。

实际上，基于上下文的方法早已应用在了文本的故事单元分割中[32]。而对于新闻视频的故事分割却刚刚开始。基于统计的方法多少也用到了一些上下文的信息，但是更多的还是强调的特征间的融合，没有明确的应用上下文信息[9,13]。

Janvier 等人在 2006 年首次将上下文的概念用于新闻视频故事单元分割[10]，他们建立了一个分割新闻故事的上下文模型（如图 3（c）所示），并用 BRF（boosted random fields）技术估计其参数。首先，采用了类似文献[16]的候选边界点的选取方法，并在候选边界点周围提取视觉、听觉和文本特征。然后，将视觉内容分为以下 9 类：新闻主题独白、演播室、户外、人造场景、卡通、天气、体育、文本以及图表和广告，将听觉内容分为以下 6 类：语音、音乐、噪声、语音＋音乐、语音＋噪声以及其他，可将这些类别作为上下文的标注。对于文本内容，先使用 Information Bottleneck 算法[33]自动地从每个片段中提取 20 个主题词作为上下文标注；然后利用这些标注信息，根据上下文故事分割模型，对那些候选边界点进行是否故事边界的判断，并且对候选边界点前后的每个片段进行语义标注，同时确定每个故事是“新闻单元”还是“其他单元”。从图 3（c）可以看到，所有候选边界点前后的标注都各自相互作用，同时前后片段标注间还有一定连续性，最终决定故事边界的划分。例如对某一片段，当听觉和文本信息都标注为“体育”时，视觉信息标注为“天气”的概率就应该减小。通过与仅使用 Boosting 而没有利用上下文的方法相比较，基于上下文的方法在性能方面有较为明显的提高（如表 1 所示）[11]。

基于上下文的方法不仅使用了统计的方法有效地对多模态特征进行了融合，而且还强调了各模态之间的以及候选边界点前后的上下文关系，对有限的特征信息利用的更充分。这是新闻视频故事单元分割的一个新的方法，有较好的发展前景。

表 1 基于上下文的方法与基于 Boosting 方法性能的比较

Tab.1 Performance comparison for the new story segmentation with a contextual and Boosting model

	查准率（%）	查全率（%）	F（%）
基于 Boosting 的方法	57	62	59
基于上下文的方法	64	66	65

4 性能及错误分析

Arlandis 等人从 TRECVID2003 和 2004 两届提交的实验结果中，选择出了一些有代表性的取平均，分别对使用 AV、AV＋ASR 和 ASR 3 种条件下的分割性能做了比较（如表 2 所示）[8]。从中可以看到：

（1）单独使用 ASR 信息的方法性能最差；

（2）使用 AV＋ASR 信息的方法较之使用 AV 信息的方法保守，反映在查全率比较低而查准率较高。

（3）使用 AV 信息方法的性能不一定比使用 AV 加上 ASR 的方法（AV＋ASR）差，也就是说，在 AV 信息上多使用 ASR 信息不一定会使分割性能提高。

表 2 TRECVID 2003 年和 2004 年不同条件下的系统性能比较

Tab.2 Results by condition each year. Recall and precision are averages by condition

条件	TRECVID2003				TRECVID2004			
	选取系统个数	查全率（%）	查准率（%）	F（%）	选取系统个数	查全率（%）	查准率（%）	F（%）
AV	5	58.7	53.8	56.1	6	56.6	40.3	47.1
AV＋ASR	5	47.4	65.4	55.0	6	48.9	55.0	51.8
ASR	5	44.6	47.8	46.1	6	46.0	38.2	41.7
全部	15	50.2	55.7	52.8	18	50.5	44.5	47.3

从表 2 中还可以看出分割的性能并不十分理想，主要有以下两类错误对其有影响[35]：其一是在特征提取阶段，底层特征选取的不充分或准确性不高直接影响到中层特征的确定，使得对片段的语义标注和分类有偏差，导致故事分割的漏检或误检发生。分析表明，人脸、运动和音频等是一些最为重要的特征；其二在故事分割阶段，由于新闻中故事类型多种多样没有固定的模式，而且有的类型出现的次数比较少，导致一些监督学习的方法在训练时对此

类故事学习的不充分,因而在测试时不能检测到此类故事,致使漏检发生。另外一种情况把一个故事单元分割成两个甚至多个,比如将同一段在不同场景的报道分割为几个故事单元,致使误检发生。

Hsu 等人分析比较了 TRECVID2003 数据集中的 CNN 新闻的每种类型所占的百分比以及相应的查全率[21]。通过用 SVM 的方法发现图 2 所示的(e)、(f)、(b)3 种类型的查全率分别为 45.6%、48.8% 和 54.2%,占全部类型的百分比分别为 15%、21.3% 和 6.3%。这几种类型查全率比较低的原因之一是这些故事边界处的静默区持续时间比较短。这 3 种类型一共占到全部类型的 42.6%,是一个相当大的比重,如何提高查全率将是今后的研究方向之一。

5　结论及发展趋势

本文对新闻视频故事单元分割技术进行了综述并将其分为 3 大类,分别予以了详细地介绍与分析。可以看出,无论是单模态的方法还是多模态信息融合的方法,应用统计学的方法诸如 HMM、SVM、ME、DT 等结合一些规则,是处理新闻视频故事单元分割问题的有效方法。在这些方法基础之上,最大限度的利用上下文信息是一类新兴的有前途的方法。

由前文可知,目前故事分割的性能在 $F=70\%$ 左右,并非完善,仍有很大的改进空间。一些可能的发展趋势如下:

(1) 如 3.3 小节所述,使用基于上下文的多模态融合方法能够更大限度的利用有限信息,提高分割效果。基于上下文的文本分割方法要比基于上下文的视频分割方法成熟的多,所以可以借鉴基于上下文的文本分割方法。

(2) 分割的一个重要步骤是提取视频中的各模态的低级特征,只有得到了丰富的特征数据,才能有效地利用各种方法进行分割。所以,如何恰当地选择并有效地提取尽可能多的特征,还有待深入研究。此外,如何利用一些高层特征,如 Video OCR 等用于故事分割,同样是研究方向之一。

(3) 将一些新的有较强学习和分类能力的理论,尤其是统计学和信息论理论,如 Hierarchical HMM、Information Bottleneck 等应用到故事单元分割,对各类特征进行有效地融合。

(4) 如何更有效的综合多种方法,如基于规则的方法与基于统计的方法、不同基于统计方法的混合应用等,是不同方法发挥各自最大的优势,是一个值得尝试的方向。

(5) 由于新闻节目不计其数,每个节目中的故事单元构成各有特色,所以概念驱动的方法有相当的局限性,而数据驱动的方法有相对较强的推广能力。实际上,许多概念驱动的方法都可以由数据驱动的方法实现,所以如何有效的利用数据驱动的方法进行故事单元分割是一个重要的发展方向,同时也是一项极具挑战性的工作。

参考文献(References)

1 Gargi U, Kasturi R, Strayer S H. Performance characterization of video-shot-change detection methods [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2000, 10(1): 1 ~ 13.

2 Hanjalic Alan. Shot-boundary detection: unraveled and resolved[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, 12(2): 90 ~ 105.

3 Zhang H J, Gong Y, Smoliar S W, *et al*. Automatic parsing of news video [A]. In: Proceedings of the International Conference on Multimedia Computing and Systems[C], Boston, NJ, USA, 1994: 45 ~ 54.

4 Merlino A, Morey D, Maybury M. Broadcast news navigation using story segmentation [A]. In: Proceedings of ACM Multimedia' 97 [C], Bedford, MA, USA, 1997: 381 ~ 391.

5 Smeaton A F, Kraaij W, Over P. TRECVID 2003-An Overview[EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tvpapers03/tv3intro. paper. ps, 2003-12-04.

6 Kraaij W, Smeaton A F, Over P, *et al*. TRECVID 2004-An Overview[EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tvpapers04/tv4overview. pdf, 2005-02-23.

7 TDT-4 Corpus Annotation Specification[EB/OL]. http://projects. ldc. upenn. edu/TDT4/Annotation/annot_task_def_V1. 4. pdf, 2002, 11.

8 Arlandis J, Over P, Kraaij W. Boundary error analysis and categorization in the TRECVID news story segmentation task[A]. In: Proceedings of International Conference on Image and Video Retrieval [C], Singapore, 2005: 103 ~ 112.

9 Hsu W, Kennedy L, Huang C W, *et al*. News video story segmentation using fusion of multi-level multi-modal features in TRECVID 2003 [A]. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing [C], Montreal, Quebec, Canada, 2004: Ⅲ645 ~ Ⅲ648.

10 Janvier B, Bruno E, Marchand-Maillet S, *et al*. Performance evaluation of a contextual news story segmentation algorithm[A]. In: Proceedings of International Conference on Multimedia Content Analysis, Management, and Retrieval 2006[C], San Jose, CA, US, 2006: 60730X-1 ~ 60730X-10.

11 Gauvain J, Lamel L, Adda G. The LIMSI broadcast news

transcription system[J]. Speech Communication, 2002, 37(1-2): 89 - 108.

12 Sugano M, Hoash K, Mutsumato K, *et al*. Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2003[EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tvpapers03/kddi. final2. paper. pdf.

13 Chaisorn L, Chua T S, Koh C K, *et al*. A Two-Level Multi-Modal Approach for Story Segmentation of Large News Video Corpus[EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tvpapers03/nus. final. paper. pdf.

14 Allan J, Carbonell J, Doddington G, *et al*. Topic detection and tracking pilot study final report[A]. In: Proceedings of DARPA Broadcast News Transcription and Understanding Workshop[C], Lansdowne, Virginia, USA, 1998: 194 - 218.

15 Beeferman D, Berger A, Lafferty J. Statistical models for text segmentation[J]. Machine Learning, 1999, 34(1): 177 - 210.

16 Hsu W, Chang S F, Huang C W, *et al*. Discovery and fusion of salient multi-modal features towards news story segmentation[A]. In: Proceedings of International Conference on Storage and Retrieval Methods and Applications for Multimedia 2004[C], San Jose, CA, USA, 2004: 244 - 258.

17 Liu Z, Huang J C, Wang Y. Classification of TV programs based on audio information using hidden Markov model[A]. In: Proceedings of IEEE Workshop on Multimedia Signal Processing[C], Redondo Beach, CA, USA, 1998: 27 - 32.

18 Lu L, Zhang H J, Li S Z. Content-based audio classification and segmentation by using support vector machines[J]. Multimedia Systems, 2003, 8(6): 482 - 492.

19 Hanjalic A, Lagensijk R L, Biemond J. Template-based detection of anchorperson shots in news programs[A]. In: Proceedings of International Conference on Image Processing[C], Chicago, IL, US, 1998: 148 - 152.

20 Gao X B, Li J, Yang B. A graph-theoretical clustering based anchorperson shot detection for news video indexing[A]. In: International Conference on Computational Intelligence and Multimedia Applications[C], Xi'an, China, 2003: 108 - 113.

21 Hsu W, Chang S F. Generative, discriminative, and ensemble learning on multi-modal perceptual fusion toward news video story segmentation[A]. In: Proceedings of IEEE International Conference on Multimedia and Expo[C]. Taipei, China, 2006: 1091 - 1094.

22 Shriberg E, Stolcke A, Hakkani-Tur D, *et al*. Prosody-based automatic segmentation of speech into sentences and topics[J]. Speech Communication, 2000, 32(1): 127 - 154.

23 Eichmann D, Park D J. Experiments in Boundaries Recognition at the University of Iowa[EB/OL]. http://www. itl. nist. gov/iaui/894. 02/projects/tvpubs/tvpapers03/uiowa. paper. pdf.

24 Wang C, Wang Y, Liu H Y, *et al*. Automatic story segmentation of news video based on audio-visual features and text information[A]. In: Proceedings of International Conference on Machine Learning and Cybernetics[C], Xi'an, China, 2003: 3008 - 3011.

25 Qi W, Gu L, Jiang H, *et al*. Integrating visual, audio and text analysis for news video[A]. In: Proceedings of International Conference on Mage Processing[C], Vancouver, BC, Canada, 2000: 520 - 523.

26 Lan D J, Ma Y F, Zhang H J. Multi-level anchorperson detection using multimodal association[A]. In: Proceedings of International Conference on Pattern Recognition[C], Cambridge, UK, 2004: 890 - 893.

27 Chua T S, Chang S F, Chaisorn L, *et al*. Story Boundary Detection in Large Broadcast News Video Archives-Techniques, Experience and Trends[A]. In: Proceedings of ACM Multimedia'2004[C]. New York, US, 2004: 656 - 659.

28 Browne P, Czirjek C, Gaughan G, *et al*. Dublin City University Video Track Experiments for TREC 2003[EB/OL]. http://www-nlpir. nist. gov/projects/tvpubs/tvpapers03/dublin. lee. paper. pdf.

29 Hoashi K, Sugano M, Naito M, *et al*. Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID'2004[EB/OL]. http://www-24. nist. gov/projects/tvpubs/tvpapers04/kddi. pdf,2004,11.

30 Chaisorn L, Chua T S. The segmentation and classification of story boundaries in news video[A]. In: Proceedings of International Conference on Visual and Multimedia Information Management[C], Brisbane, Australia, 2002: 95 - 109.

31 Hsu W, Chang S F. A statistical framework for fusing mid-level perceptual features in news story segmentation[A]. In: Proceedings of IEEE International Conference on Multimedia and Expo[C]. Baltimore, MD, USA, 2003: Ⅱ-413 - 416.

32 Yamron J P, Gillick L, Knecht S, *et al*. Statistical models for tracking and detection[A]. In: Proceedings of the DARPA Topic Detection and Tracking Workshop[C]. Gaithersburg, Maryland, US, 2000:139 - 144.

33 Slonim N. The Information Bottleneck: Theory and Applications[D]. Jerusalem, Israel, Hebrew University, 2002.