文章编号 :1001 – 2486(2004)05 – 0099 – 05

层次化新闻视频处理框架的设计与实现*

谢毓湘,栾悉道,吴玲达,老松杨,王卫威
(国防科技大学人文与管理学院,湖南 长沙　410073)

摘　要 :提出了一个通用的层次化新闻视频处理框架,将新闻视频处理分为句法分段、语义标注以及视频摘要三个层次,并给出了三个层次中涉及的故事单元探测、字幕探测、视频摘要等关键技术的解决方案。框架突破了传统的新闻视频处理框架仅局限于句法分段以及单媒体特征进行处理的缺陷,通过对视音频特征进行多模态的综合分析来获取新闻视频高层的语义内容。实验通过一个新闻视频处理原型系统 NVPS 验证了框架的可行性,重点对故事单元探测、标题探测以及口播帧探测三个算法进行了实验,实验结果分别达到 88%,86%和 86%的探测准确率,从而进一步证实了层次框架在新闻视频处理方面的有效性。

关键词 :新闻视频 ;句法分段 ;语义标注 ;视频摘要
中图分类号 :TN941.2　　文献标识码 :A

The Design and Realization of Hierarchical Framework of News Video Process

XIE Yu-xiang, LUAN Xi-dao, WU Ling-da, LAO Song-yang, WANG Wei-wei
(College of Humanities and Management, National Univ. of Defense Technology, Changsha 410073, China)

Abstract :A general hierarchical framework of news video process is presented. It divides the news video process into three levels: syntax segmentation level, semantic labeling level and abstraction level. Some key techniques related to these levels are described and solutions of them are introduced. The proposed framework overcomes the shortcomings of traditional news video process methods, which are limited to the content-based segmentation and process based on the single media feature. It acquires the semantic content by the analysis of audio-visual features synthetically. Experiments are carried out on a news video process prototype called NVPS, which validates the feasibility of the framework. Three methods, namely story detection, caption detection and anchor detection methods are tested on NVPS. The results reach to the detection precision of 88%, 86% and 86% respectively, which prove the efficiency of the layered framework in the semantic content analysis of news videos.

Key words :news video ; syntax segmentation ; semantic labeling ; video abstraction

新闻节目具有实时、准确、信息量大等特点,在许多方面都发挥着重大的作用。然而,由于缺乏有效的新闻视频处理和管理方法,使得用户无法实现有效的新闻视频检索。为解决该问题,设计了一个通用的新闻视频处理框架,并在该框架的指导下进行新闻视频的自动分析与处理,其目的是通过镜头探测、故事单元探测、字幕探测、人脸探测、语音识别等多种技术对新闻视频进行多特征融合的处理,完成新闻视频的自动分析与编目功能,帮助用户解决新闻视频数据无法处理以及难以管理的问题,从而实现用户对新闻视频语义层次上的查询与检索。

1　新闻视频处理系统的层次框架

传统的新闻视频处理系统存在以下一些缺陷[1] :首先,它只是针对视频流中的视觉信息来进行分析,而忽略了新闻视频流中其它丰富的模态信息,如音轨、文字等等的处理,从而造成了处理的结果与用户的实际感知之间存在较大的差距 ;其次,它受基于内容检索技术发展的限制,更多关注于对新闻视频进行结构化的分析,即基于内容的视频分割上,而很少从支持用户语义检索的角度去考虑处理问题。这

* 收稿日期 :2004 – 04 – 02
基金项目 :国家 863 高技术资助项目(2001AA115123)
作者简介 :谢毓湘(1976—),女,博士生。

两个缺陷使得新闻视频处理系统在实用上还存在较大的差距,大部分的工作只是停留在实验室阶段。

针对传统新闻视频处理系统的缺陷以及新闻视频自身的特点,设计了如图1所示的通用的层次化新闻视频处理框架。将新闻视频的处理从低到高分为三个层次,即句法分段层、语义标注层与摘要层。其中,句法分段层主要负责完成视音频流的分段任务,具体包括镜头探测、故事单元探测、关键帧提取、静音探测等。该层侧重于新闻视频的结构划分,它对应着文本处理中的划分段落。语义标注层在句法分段的基础上完成新闻视频语义标注的任务,如字幕探测、人脸探测、人物识别、语音识别等等。该层的处理能够部分地抽取新闻视频中的语义,对应着文本处理中的概括各个段落的中心思想。摘要层负责将句法分段层以及语义标注层的处理结果进行汇总,生成摘要。摘要层的任务是从结构与语义两方面来对新闻视频进行一个汇总,它应该能够浓缩新闻视频中的精华,并且便于用户从整体角度对新闻视频有一个大概的了解。

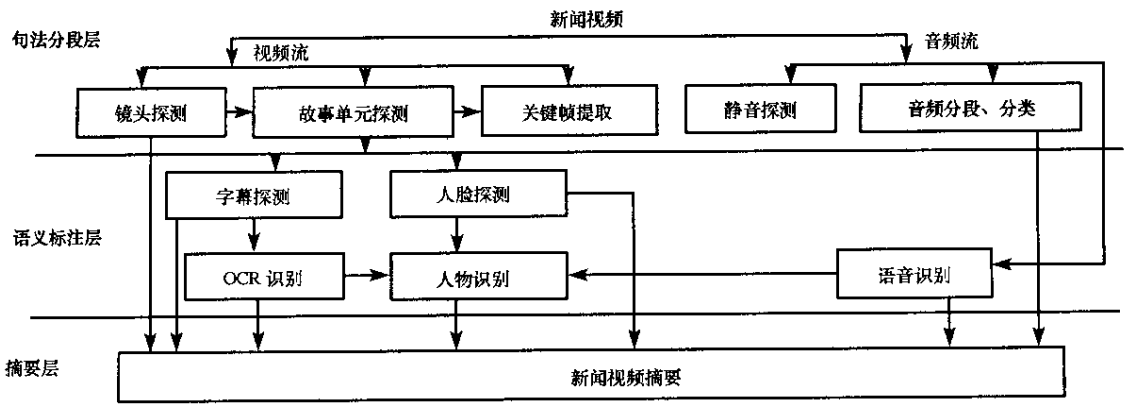


图1 层次化新闻视频处理框架图

Fig.1 Hierarchical framework of news video process

2 新闻视频句法分段层

新闻视频的句法分段实质上就是一个视频结构化的问题。视频的结构从低到高可以分为帧、镜头、场景、故事单元以及视频流等多个层次。新闻视频句法分段的任务就是从新闻视频流中自动地分析出新闻的镜头、场景、故事单元,特别是含有高层语义信息的故事单元。对于新闻视频而言,一个故事单元通常就是一则新闻。因此,新闻故事单元的提取问题也即如何从一段新闻视频中准确地提取出每则新闻。从获取高层语义信息的角度考虑,这里着重讨论新闻故事单元的探测方法,提出了两种解决方案。

2.1 场景探测与静音探测相结合的方法

大多数的场景探测是在镜头探测与镜头聚类的基础上完成的,它将视觉上相似而时间上未必相邻的镜头聚为一类,代表某个场景。这种方法由于忽略了时间关系,使得场景的表示并不能真正地让用户理解视频的内容,因此完全从视觉的角度出发进行故事单元的探测其结果总是不尽人意。音频作为与视频相伴的一种媒体,含有丰富的内容信息,因此结合音频的内容分析进行新闻视频故事单元的切分,将更为合理。基本方法是:首先对视频流进行场景探测,得到若干场景边界点;接下来对音频流进行静音探测,选取静音时间段较长的点(即停顿时间较长的点)作为静音边界。在静音边界附近寻找是否出现场景边界,若是则将其确定为故事单元边界。这种方法综合考虑了视觉以及声音的变化信息,该方法基于这样的观察,即每个新闻故事单元之间会有相对较长的语音停顿以及较大的视觉差异。将视觉与听觉信息融合起来进行分析,有效地改善了仅依靠视觉信息来判断故事单元边界的不足。

2.2 口播帧探测

所谓口播帧,是指新闻故事开始阶段的播音员镜头,口播帧探测[7]是新闻视频结构分析较常采用的方法,该方法基于这样的观察,即每则新闻之间通常都会以播音员的镜头出现作为分隔的标志。如果能

够将播音员镜头(也即口播帧)探测出来,那么新闻故事单元大体也就划分出来了。但是对于某些不以口播帧的出现作为新闻出现标识的视频来说,这种方法可能会漏掉部分新闻。

上述两种方法各有优劣,从准确性的角度而言,第一种方法更为合适。本文采用了在镜头探测[6]的基础上进行口播帧聚类,从而获得新闻视频故事边界的方法。这种方法对中央台的新闻故事能够进行较为准确的划分。

3 新闻视频语义标注层

新闻视频语义标注就是在新闻视频句法分段的基础上对某些重要的语义事件进行标注。在新闻视频中,字幕的出现以及特写人脸的出现都可以看作是重要的语义事件。从用户感知的角度出发,人们对一段新闻故事的了解,很大程度上是从新闻视频的标题字幕以及语音中得到的,特别是标题字幕能够高度提炼一则新闻的内容。因此字幕探测技术成为了新闻视频语义标注的一个重要内容。另外,新闻视频中出现的特写人脸,如某领导人的特写镜头,往往都蕴含着重要的语义信息。因此建立在人脸探测基础之上的人物识别也是新闻视频语义标注的一个研究方向。下面将分别对它们进行说明。

3.1 字幕探测

字幕探测[8]对新闻视频的语义标注格外重要。这里将新闻视频中的字幕分为两类,一类是编辑字幕,如新闻开始阶段的字幕帧,称之为标题帧。这类字幕对理解新闻故事的语义有重要的作用;另一类是场景字幕,即在新闻摄制时就已存在的字幕。由于标题字幕的出现通常可以辅助新闻的故事边界探测,亦可以通过后续的 OCR 识别抽取出该条新闻的内容信息,因此将重点探测这类标题帧。

目前采用的方案是在镜头关键帧的基础上进行字幕的探测,字幕探测的过程包括灰度变换、边缘检测、字幕区域探测、字幕区域合并与过滤以及二值化五个阶段。其中边缘检测采用 Sobel 算子,这样将保证其对垂直和水平边缘都具有较大的响应;字幕区域探测采用水平与垂直扫描相结合的方法;字幕区域合并的条件为两个字幕区域的水平与垂直位移均小于预定阈值;字幕区域过滤的条件则包括宽高比、距边界的距离、高度、面积小于预定阈值等等;最后采用双峰阈值法进行二值化,即在字幕区域的最大最小灰度中间选取方差最大的灰度作为二值化的灰度阈值,从而保证为后续的 OCR 识别提供较高质量的图像源。图 2 是对中央台的新闻字幕帧进行字幕探测的部分实验过程与结果。

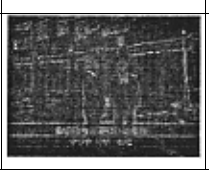
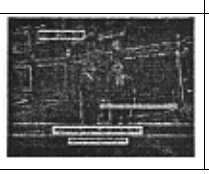
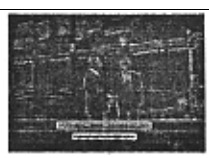
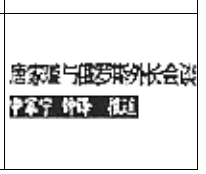
原始图像	灰度图像	边缘检测	字幕区域探测	字幕区域合并与过滤	二值结果
					

图 2 新闻视频帧中字幕探测的过程及结果

Fig.2 The process and results of caption detection in news video frames

3.2 人物识别

新闻视频中大部分的新闻内容都与人物相关,人们在浏览或检索新闻时,也经常希望找到与某一特定人物相关的新闻内容。因此,通过找到用户关心的人脸来检索新闻是一条不错的途径。人物识别[2]作为视频语义特征提取中的关键技术,不仅有助于人们对视频内容的理解,而且可以用来辅助视频内容的检索以及生成多模态融合的新闻摘要。

人物识别的过程如图 3 所示。由图可见,为了完成上述任务,需要着重解决以下四个关键问题:人脸探测、人脸跟踪、人脸鉴别以及人脸与人名的关联。所谓人脸探测,即判断视频流中的图像帧是否存在人脸。如果存在人脸,那么确定所有人脸的位置以及大小等参数。在实际的处理过程中,只对镜头的代表帧进行人脸探测,采用 Haar 对象探测的方法[3]来进行。人脸跟踪即在一组图像序列中实时连续地评估人脸的位置和角度。目的之一在于探测出某一人脸后,通过跟踪将视频中出现该人物的所有帧都找出来,从而为后续的视频检索与摘要打下基础;另一目的在于找到某人的所有人脸后,从中挑选出比较端正的人脸进行人脸鉴别。人脸鉴别就是验证探测到的两个人脸是否是同一个人,即人脸相似度的

验证问题。它可以应用于人脸识别、人脸与人名之间一对多或无法对应的问题以及人脸的验证。

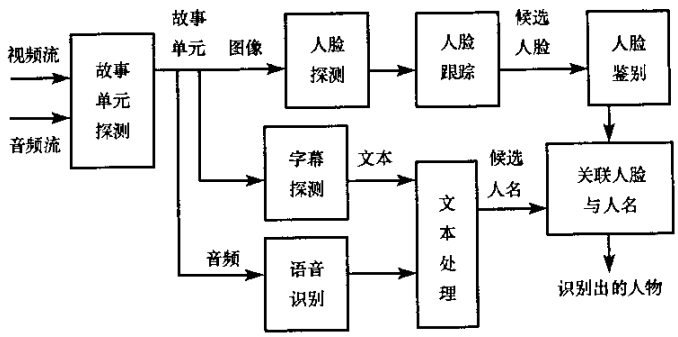


图3 人物识别的过程

Fig.3 The process of face recognition

如图3所示,经过上述人脸探测、跟踪与鉴别得到的人脸必须与语音识别以及字幕探测得到的文本中的人名很好地关联起来,才可以完成人物识别的任务。建立二者之间的关联关键在于制定一系列的规则:(1)首先区分已探测人脸的重要程度,这可以通过人脸出现的位置、大小和频度来判定;然后区分已提取人名的性质与重要程度,比如说,是总统还是消防员,是演讲者还是听众,是会议的主持者还是会议的参加者,这可以通过文本处理的知识来判定。根据两者的重要程度匹配人脸与名字,从而实现人物识别。(2)根据人脸和名字出现的时间相关性来进行人物判定。由于人脸与名字都从视频中得到,并且它们都具有额外的时间信息,某些视频中人脸的出现与音频或字幕中名字的出现基本上同步,因此可以利用这些特点将名字与人脸关联起来,实现人物识别。

4 新闻视频摘要层

视频摘要即以自动或半自动的方式对视频的内容进行分析,从原始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。它通常建立在视频句法分段以及语义标注的基础上(见图1),是一种较高层次的新闻视频处理技术。

4.1 摘要的重要度评判模型

新闻视频摘要的目的在于用尽可能少的时间和空间表现尽可能重要的视频内容。该问题就是视频内容重要度评判模型的设计问题,也即如何选取重要的视频内容作为摘要的素材。通常认为,对于新闻视频而言,重要的视频片段应至少具备以下条件之一:(1)字幕帧。主要是指新闻的标题帧,以及新闻中间某些非常醒目的字幕帧。(2)人脸帧。人脸(尤其是特写人脸)的出现往往意味着比较重要的内容发生。尤其是对新闻视频,领导人物的出现能够从很大程度上说明该则新闻的内容。因此,非常有必要对人脸帧进行探测,并将其作为视频摘要的一个重要素材。(3)镜头关键帧。我们认为在镜头探测的基础上通过关键帧抽取得到的帧,可能是比较重要的视觉镜头,因此该问题的关键也即镜头关键帧的抽取。其它需要满足的条件包括持续时间较长、出现频率较高等。

简单的视觉重要度模型[4]可以用下式表示:$\overline{I_v} = \alpha \overline{I_t} + \beta \overline{I_f} + \gamma \overline{I_k}$。其中$\overline{I_v}$表示视频摘要的视觉重要度,$\overline{I_t}$表示字幕帧的视觉重要度,$\overline{I_f}$表示人脸的视觉重要度,$\overline{I_k}$表示镜头关键帧的视觉重要度,$\alpha$,$\beta$,$\gamma$分别为权值。从式中可以看出,视频摘要的视觉重要度模型可以简单地用字幕帧、人脸帧以及镜头关键帧的视觉重要度的线性组合来表示。

4.2 摘要的可视化

视频摘要的可视化就是如何将处理得到的视频内容用一种可视的、用户可理解、可交互的方式表现出来,从而便于从整体的角度去理解新闻视频的内容以及它们之间的关联。从表现形式上看,主要包括两种不同的摘要形式:静态的视频摘要以及动态的视频摘要。静态的视频摘要是以静态的方式来表现视频的内容,如标题、关键帧、故事板、文本描述等等,它是从视频中抽取或生成的有代表性的图像或文字。动态的视频摘要,如缩略视频(video skim),是图像序列及其伴音的集合,它本身也是一段视频,但

比原视频要短得多。前者通常只考虑视觉和文本信息,不考虑音频以及时间与同步问题,因此它的构建与表现都相对简单。缩略视频则不同,它含有丰富的时间以及音频信息,因而更加符合用户的感知。

5 实验及原型

实验数据主要来自于 2003 年 4 月采集的中央台晚间新闻节目,包括国际新闻与国内新闻的内容,历时 390 分钟。另外,还有部分实验数据来自于凤凰台的新闻节目。整个视频新闻的处理过程包括镜头探测、关键帧提取、字幕探测、人脸探测、故事单元探测、场景探测,以及最后的数据库操作等多个步骤。语音识别过程由于需要处理的数据量非常大,暂时没有对其进行实时的处理。为减少数据量,今后将重点识别播音员镜头部分的语音。表 1 分别给出了故事单元探测、标题探测以及口播帧探测的实验结果。

表 1 实验结果

Tab.1 Experiment result

	长度	E_s/S	H_s	E_c/C	H_c	E_A/A	H_A
News1	657513	2/10	0.8	2/10	0.8	2/10	0.8
News2	904357	0/8	1	1/8	0.88	0/8	1
News3	799113	1/12	0.92	2/12	0.83	2/12	0.83
News4	653313	1/10	0.9	1/10	0.9	1/10	0.9
News5	595713	2/10	0.8	1/10	0.9	2/10	0.8
总计或均值	3610009	6/50	0.88	7/48	0.86	7/50	0.86

注:长度以 ms 为单位,E_s/S 表示误漏判故事单元/实际故事单元,E_c/C 表示误漏判标题数/实际标题数,E_A/A 表示误漏判口播帧数/实际口播帧数,H_s,H_c,H_A 分别表示故事单元探准率,标题探准率以及口播帧探准率。

表 1 的实验结果表明我们的算法可以获得平均 88% 的故事单元探准率,86% 的标题探准率以及 86% 的口播帧探准率,从而进一步验证了提出的框架基本上可以满足新闻视频语义处理的要求。

6 结论与展望

早期开发的新闻视频处理系统侧重于从视频分割的角度来处理新闻视频,而尚未上升到语义这样一个层次,因此无法真正地支持用户的语义检索。本文提出了这种层次化的新闻视频处理框架,并在该框架的指导下开发了 NVPS 这样一套基于视音频多特征融合的新闻视频处理系统,向语义化的视频内容检索迈出了关键性的一步。

目前,字幕探测、人脸探测等新闻视频的语义标注技术已经加入到 NVPS 系统中来了,只是在处理速度和执行效率上还有待提高。特别是字幕探测之后的字幕识别,以及语音识别的准确率将是我们今后需要着重解决的问题。同时,NVPS 新闻视频处理系统的不断完善也将为我们今后即将进行的新闻视频辅助决策分析奠定良好的基础。

参 考 文 献:

[1] Michael R L, Edward Yan, Sam Sze. A Multilingual Multimodal Digital Video Library System[A]. JCDL '02, July 13 – 17,2002, Portland, Oregon, USA. 145 – 153.

[2] Shin'ichi Satoh, Name-It: Naming and Detecting Faces in News Videos[J]. IEEE Multimedia, 1999: 22 – 35.

[3] Lienhart R, Pfeiffer S, Effelsberg W. Video Abstracting[J]. Communications of the ACM, 55 – 62, Dec. 1997.

[4] Ma Y F, Lu L, Zhang H J, et al. A User Attention Model for Video Summarization[A]. Proceeding of ACM Multimedia '02, Juan-les-Pins, France, December, 2002.

[5] Christel M G, Hauptmann A G, Wactlar H D, et al. Collages as Dynamic Summaries for News Video[A]. Proceeding of ACM Multimedia '02, Juan-les-Pins, France, December, 2002.

[6] 谢毓湘,栾悉道,吴玲达,等. 一种基于解压的镜头探测方法[J]. 系统工程与电子技术,2003,25(8):1028 – 1031.

[7] 马宇飞,白雪生,徐光佑,等. 新闻视频中口播帧检测方法的研究[J]. 软件学报,2001, 12(3):377 – 382.

[8] Hua X S, Chen X R, Liu W Y, et al. Automatic Location of Text in Video Frames[A]. Proceeding of ACM Multimedia 2001 Workshops: MIR2001, Ottawa, Canada, October 5, 2001: 24 – 27.