一种基于多模态特征的新闻视频语义提取框架*

闫建鹏[1]，封化民[1,2]，刘嘉琦[1]
（1. 西安电子科技大学 通信工程学院，西安 710071；2. 北京电子科技学院，北京 100070）

摘　要：为提高视频语义信息提取准确率，提出了一种基于多模态特征的新闻视频语义提取框架。在视频中提取主题字幕信息，对音频进行分类和语音识别，根据主题字幕信息借助搜索引擎得到与新闻视频相关的网页；最后利用网页文本对语音识别的结果进行纠错，从而通过视频字幕信息和语音脚本的跨模态融合提高视频语义提取的准确率。在中等规模的新闻视频（含新闻网页）库测试表明了该方法的有效性，经纠错后的语音识别准确率达到了 65% 左右。

关键词：多模态特征；语义分析；视频检索
中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1001-3695(2012)07-2725-05
doi:10.3969/j.issn.1001-3695.2012.07.089

News video semantic extraction framework based on multimodal information

YAN Jian-peng[1], FENG Hua-min[1,2], LIU Jia-qi[1]
(1. *School of Telecommunication Engineering, Xidian University, Xi' an* 710071, *China*; 2. *Beijing Electronic Science & Technology Institution, Beijing* 100070, *China*)

Abstract: This paper proposed a framework to analyse video semantic based on multimodal information. Firstly, this method detected and extracted captions. Secondly, it classified audios and obtained scripts via speech recognition. Then according to the caption information it obtained Web pages related to the video with the aid of search engine. Finally, the result of speech recognition were corrected error by Web information. Thereby, caption and audio information were cross-modally integrated. The experimental result on a middle scale testing set shows that the framework is feasible, achieves the accuracy of about 65% by error correction.

Key words: multimodel feature; semantic analysis; video retrieval

0　引言

随着计算机技术、多媒体技术的飞速发展，基于内容的视频检索（content-based video retrieval，CBVR）成为目前国内外最为活跃的研究课题之一。近年来，在视频感知特征的提取与表达、视频结构化分析、视频摘要、视频索引建立等多个方面都取得了长足的进步，支持根据不同的底层特征、草图、示例图片或视频片段、关键词来进行视频查询。然而，语义鸿沟（semantic gap）阻碍了 CBVR 的进一步发展。

基于语义的多媒体信息检索（semantics based information retrieval，SBIR）相对于基于底层特征的多媒体信息检索更符合人们的思维和检索习惯，具有更广泛的应用前景。能否准确提取出视频高层语义成为该检索技术能否走向应用的关键。考虑到视频中的多种模态（如图像、音频和文本）在本质上呈现的时序关联共生特性，在视频语义信息理解和挖掘中，充分利用多模态媒质之间的交互关联是非常重要的研究方向。

近年来，国内外很多研究人员在该领域做了大量的工作，并取得了一定的成果。Yang 等人[1]通过从视频的语音识别（ASR）以及视频图像光学识别（OCR）中提取文本信息，将视频检索转变为文本信息检索。冀中[2]提出了一种融合音频、字幕以及视觉等多模态信息的新闻故事单元分割算法。刘亚楠等人[3]提出了一种基于多模态子空间相关性传递的语义概念检测方法来挖掘视频的语义信息，该方法对所提取视频镜头的多模态底层特征，根据共生数据嵌入和相似度融合进行多模态子空间相关性传递而得到镜头之间的相似度关系。赵志城[4]在视频语义的提取方面，提出了一个基于社会网络分析（SNA）和电影本体（ontology）的影片内容理解框架和一套语义提取算法。郑李磊等人[5]设计、开发了一个中文新闻视频字幕自动生成系统，将音频分为语音、静音、音乐三种类型，利用语音识别工具包 HTK，通过大词汇量连续语音识别技术对广播音频进行自动抄本。

传统的新闻视频处理系统存在着以下一些缺陷：a）它受基于内容检索技术发展的限制，更多地关注于对新闻视频进行镜头分割、故事单元分割等结构化的分析，而很少从支持用户语义检索的角度去考虑处理问题；b）在多模态特征的融合问题上，往往是对多个模态特征的叠加累积，在造成高维灾难的同时，没有充分利用各个模态的语义在持续时间内的彼此关联耦合性。

针对传统新闻视频处理系统的缺陷和新闻视频自身的特点，本文提出了一种基于多模态特征的新闻视频语义分析框架，将新闻视频的语义分析分为故事单元分割、语义信息提取以及

收稿日期：2011-11-21；**修回日期**：2011-12-23　　**基金项目**：国家自然科学基金资助项目（60972139）；北京市自然科学基金资助项目（4092041）

作者简介：闫建鹏（1986-），男，山西晋中人，硕士，主要研究方向为视频检索（jianpengyan@sohu.com）；封化民（1963-），男，陕西渭南人，教授，硕导，博士，主要研究方向为多媒体智能信息处理、网络安全；刘嘉琦（1986-），男，山东威海人，硕士，主要研究方向为视频检索.

视频分类、检索三块,本文的主要工作为视频语义信息提取。

1 视频语义分析总体框架

1.1 研究基础和前期成果

基于内容的视频分析(content-based video analysis,CBVA),旨在通过对视频结构和语义内容的分析,将非结构化的视频数据结构化,并提取其中的语义内容单元,在此基础上建立视频的索引、浏览和检索等应用系统。

北京电子科技学院信息安全重点实验室多媒体智能信息处理研究室在相关方面有步骤地进行了大量的研究工作。Feng 等人[6]提出了基于支持向量机(SVM)的时域多尺度视频分割算法,取得了较为满意的镜头分割效果。范竞往等人[7]采用镜头层和故事层的双层混合模型,再利用中文新闻视频中的主题标题特征,对视频进行故事分割,大大提高了中文新闻视频的故事分割准确率。Feng 等人[8]还提出了一种基于规则和 SVM 的音频分类方法,在此基础上使用 Sphinx 工具包对语音成分进行自动语音识别。这些相关的研究为本文的工作奠定了基础。

1.2 系统总体框架和本文的主要内容

本文的主要工作是视频语义信息的提取,即通过提取视频字幕信息、音频信息,再提取字幕信息中的关键词,并借助搜索引擎得到与新闻视频相关的网页,且借助网页信息对语音识别的结果纠错。通过字幕信息和语音信息在高级语义层次上的跨模态融合,实现对新闻视频高层语义的准确提取,为基于语义信息的视频分类和检索奠定基础。

基于多模态特征的新闻视频语义分析框架如图 1 所示。

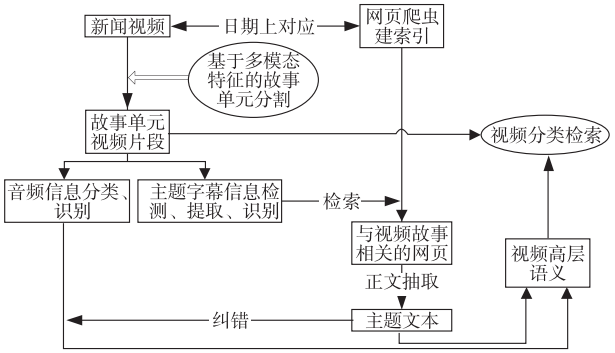

图1 基于多模态特征的新闻视频语义分析框架

2 新闻视频语义信息提取

2.1 新闻视频主题字幕的定位提取和识别

在新闻视频制作后期,往往会人工加入一些字幕,作为一种高层语义特征,该字幕中的文字信息对视频内容的理解、索引和检索具有重要意义。因此,如果能够自动检测、提取、识别出这些文字信息,就可以为视频语义分析提供一条有效的途径和解决方法。本文综合该领域常用的一些方法,并不断地实验,得出如图 2 所示的视频主题字幕信息提取框架。本框架可针对一段视频流准确地检测主题字幕的位置和字幕存在时间,将字幕区域以文字框形式抽取出来,在文字框分类的基础上进行二值化,并借助 OCR 识别技术转换为文本。

视频字幕提取系统可分为视频预处理、文字区域检测、文字区域过滤、文字框合并和时间定位、文字识别五部分。

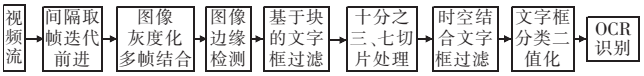

图 2 主题字幕信息提取流程

2.1.1 视频预处理

新闻视频一般为 20 ~ 30 fps,视频中主题字幕持续至少 2 ~ 3 s。综合考虑处理效率以及下一步的多帧综合效果,该框架采取间隔取帧处理的方式,经由实验确定采用每隔十帧取一帧的方法。

视频文件多为彩色视频,本框架采用的基于边缘的文字检测算法需要对图像进行灰度化处理。本系统采用更常用和更符合人眼视觉特点的不等权策略,对图像中坐标为(x,y)的点,计算式为

$$f_g(x,y)=0.3R(x,y)+0.59G(x,y)+0.11B(x,y) \quad (1)$$

其中:$R(x,y)$、$G(x,y)$、$B(x,y)$分别为输入的彩色图像的对应坐标点 R、G、B 三色分量;$f_g(x,y)$为输出灰度图像灰度值。

在视频中出现的主题字幕,一般至少持续 2 s 左右的时间,文字标题以外的像素点则不能保持长时间连续不变。多帧综合(multiple frames integration,MFI)算法就是基于以上原则,通过求相邻若干帧像素灰度值平均值或灰度值最小值来降低文字背景复杂度。本文采取了李雪龙等人[9]提出的一种改进的 MFI 算法处理流程,即分别求多帧图像的灰度最小值和平均值,单独进行边缘检测,再对检测结果综合。

2.1.2 图像文字区域检测

图像中的文字区域通常有较鲜明的颜色对比度和亮度对比度,同时文字区域有较丰富的边缘点。利用文字的边缘特征、颜色特征、连通性特征进行文字检测,都是较常用的方法。本文分析比较基于以上几种特征的文字区域检测算法,综合考虑系统的计算效率和鲁棒性能,选取利用文字的边缘特性(灰度不连续性)进行文字区域检测。通过实验比较,在 Robert、Sobel、Canny、Prewitt 等边缘检测算子中选取 Sobel 垂直边缘检测算子,其算子模板如图 3 所示。

-1	0	1
-2	0	2
-1	0	1

图 3 Sobel 垂直边缘检测算子模板

用$g(x,y)$表示输出图像,$f(x,y)$表示输入图像对应坐标灰度值,则计算式为

$$g(x,y)=-f(x-1,y+1)+f(x+1,y+1)-2f(x-1,y)+2f(x+1,y)-f(x-1,y-1)+f(x+1,y-1) \quad (2)$$

图 4 为 Sobel 边缘检测结果示例图。Sobel 边缘检测后的图像为灰度图像,为了便于下一步处理,在此对图像进行二值化,这里采用全局阈值算法中的 Otsu 算法。

(a) (b) (c) (d)

图4 Sobel 边缘检测结果

2.1.3 文字区域过滤

对边缘检测后的图像,采用对图像进行边缘点强度和密度过滤确定候选文字框;再采用对候选文字框进行垂直和水平方向直线迭代分解的方法得到更加紧凑、精确的文字外接矩形框;最后结合新闻视频字幕特点,对图像进行十分之三、七分

割,分别处理,删除误判结果,得到更为精确的文字框。下面对文字区域过滤规则作详细介绍。

对图像进行边缘点密度过滤,采用一个 $M \times N$ 大小的扫描窗口,从上到下、从左到右扫描图像,若一个窗口内边缘点密度大于某个预设最小密度阈值而小于一个最大密度阈值,则它被认为是候选文字块,保留其所有像素点的灰度值;若边缘点密度相对上述阈值过低或过高,则认为是非文字块,把非文字块像素点置为某个定值(这里取为 100)。过滤过程取 $M = 20$, $N = 10$,密度阈值 min = 20,max = 150。图 4(a)(b)经密度过滤结果如图 5(a)(b)所示。由图可见,边缘点密度、强度过滤得到的候选区域过于粗糙,需要进一步处理。这里采取 Hua 等人[10]提出的水平直线扫描的方式,即对每个候选文字区域,首先进行水平直线扫描,得到一个或多个线段 $l_i(i = 1,2,\cdots,n)$。l_i 上面 r 行和下面 r 行内的 Sobel 边缘点数为 $ETopv(l_i,r)$ 和 $EBtmv(l_i,r)$,l_i 长度为 $|l_i|$,扫描区域内的边缘点个数 $ENv(l_i,r)$ 和密度 $EDv(l_1,r)$ 分别为

$$ENv(l_i,r) = \max(ETopv(l_i,r),EBtmv(l_i,r)) \tag{3}$$

$$EDv(l_i,r) = ENv(l_i,r)/(|l_i| \times r) \tag{4}$$

保留 $ENv(l_i,r)$ 和 $EDv(l_1,r)$ 大于一定阈值的线段,否则将其置为指定灰度(100)。之后再进行一次相同步骤的垂直方向上的分解。分解后结果如图 5(c)(d) 所示。

最后结合新闻视频字幕特点(图 4(a)(b)分别代表两种特点的字幕),对图像进行十分之三、七分割,分别处理,即各部分的非候选区域小于一个阈值,则判断为该区域不存在字幕框,从而删除误判结果,得到更为精确的文字框,如图 5(e)(f)所示。

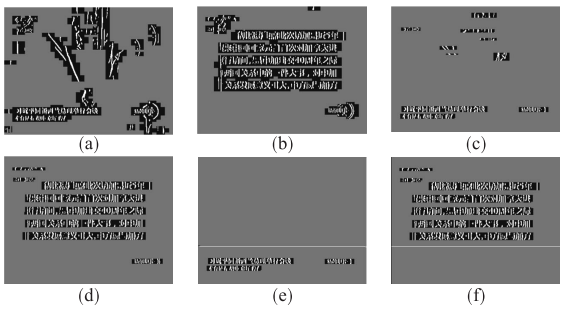

图5　文字框过滤示例图像

2.1.4　文字框合并和时间定位

经过文字框过滤,候选文字框已基本正确,但依旧有部分误判情况,可以通过一些启发式规则过滤。

规则 1　文字框的长度和宽度大于一定阈值。

规则 2　文字框长宽比在一定范围内。

规则 3　文字框持续时间须有一定时间,且不能持续过长时间(台标、演播室背景等信息)。

规则 4　文字框距离图像边缘需大于一定阈值。

之后合并相邻时间上的相同文字框,得到视频主题字幕文字框。

2.1.5　文字识别

确定视频文字框的位置和时间,再对其中的文字进行提取和识别,这里借用文档图像 OCR 技术。文档图像的 OCR 技术通常只对二值化图像具有较好的识别效果,因此主题字幕文本框还必须经过二值化处理。视频中得到的文字框既有背景颜色单一并与前景颜色色差大的,也有背景复杂且与前景颜色色差小的,这为图像二值化带来了较大的难度。对于前者(如图 6(a))采用全局阈值二值化算法(这里采用 Otsu 算法)就能得到比较好的效果,对于后者(如图 6(b))采用局部阈值二值化效果中的 Niblack 算法,效果通常比全局阈值化算法好。将二值化后的结果图像送入 OCR 识别器识别,本文使用商用 OCR 识别软件——清华 TH-OCR 9.0。

(a)　(b)

图6　图像二值化示例

2.2　音频信息提取

音频作为伴随视频的一种重要媒质,包含了丰富的语义信息。对音频信息的研究主要包括音频分类和语音识别。早期的音频分类算法通常是选择几种特征(如过零率、短时能量、带宽等),然后人工或自适应地选择一个阈值,按照不同的音频类型在这些特征方面的不同特点对音频加以分类。近年来,随着人工智能、模式识别等领域的快速发展,越来越多的研究者将 SVM、HMM 等统计学习算法应用到音频分类研究中,并取得了一定的成效。文献[9]提出了如图 7 所示的基于规则和 SVM 结合的音频分类方法,将视频中的音频分为静音、语音和音乐。根据音频类别信息设计切分算法,对相同类别的音频进行聚类,得到音频切分信息。在此基础上,使用 Sphinx 工具包对语音成分进行自动语音识别。对中等规模的视频库实验,音频切分的准确率为 91.36%,语音识别的准确率为 61.86%。语音识别中的错误结果会影响对视频语义的正确理解和检索,所以需要对识别结果进行纠错。

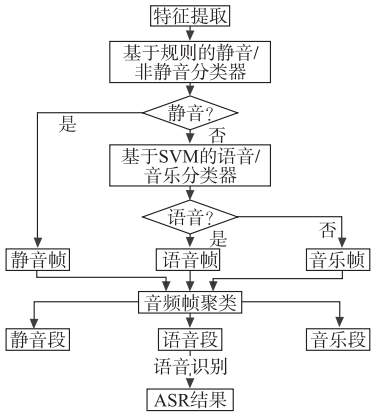

图7　视频中音频信息提取流程

2.3　与视频相关的网页检索

考虑到互联网在新闻传播方面的快捷性和广泛性,视频中的新闻故事通常会在网上找到相同或相似的新闻网页。而新闻网页的文本能够更准确、详细地表述相对应的视频内容,甚至涵盖更多的相关信息。在互联网上要寻找与新闻视频故事单元相对应的网页,需要借助搜索引擎,而搜索关键词是这两部分的中间枢纽。关键词的准确度决定能否找到与视频内容相关的网页。结合新闻事件的特点,搜索关键词主要是新闻发生的时间、地点、人物和事件。本系统的搜索关键词来自视频主题字幕。视频主题字幕大多是简短的概括性短语,又因搜索引擎支持双引号搜索功能,即针对全文搜索,如果文中有与引号内容完全匹配的句子,则呈现给用户搜索结果。本文将整个字幕标题以双引号内容输入搜索引擎,发现有些新闻可以直接找到与视频高度相关的网页,而有些视频利用主题字幕信息的全文匹配无法找到相关网页,需要提取字幕信息中的关键词进行搜索。对主题字幕的语句进行分词以后,进行命名实体词的

抽取,作为视频字幕信息关键词。在这一部分本文借助了哈尔滨工业大学信息检索研究室研究开发的 LTP 系统中的中文分词及词性标注模块和命名实体识别模块[11]。

为适应本课题的研究,本文部署搭建了 Apache 下的开源的搜索引擎 Nutch-1.0[12],并进行了中文分词模块添加,包括网页定向爬虫模块和搜索模块,采集数据阶段每日定时对国内权威新闻网站新华网进行爬取(爬取深度为 3),每日爬取网页约 1 800 个,爬虫索引以日期存储。

对每一个新闻故事段,根据视频主题字幕信息,首先进行全字幕信息双引号搜索,若有返回结果则作为视频对应的网页;否则对主题字幕信息命名实体抽取得到关键词进行检索,得到相关的网页。对搜索结果取排名最靠前的两个网页的正文文本作为视频相关的网页文本。

3 利用相关网页文本对 ASR 结果纠错

分析本系统的 ASR 识别错误,包括音错字错和音对字错两种。音错包括吞音、添音、错音三种错误。针对这些识别错误,本文设计了如图 8 所示的语境知识约束下基于拼音相似度规则的纠错系统,借助与视频相关的网页文本,可以纠正语音识别结果中的部分错误。

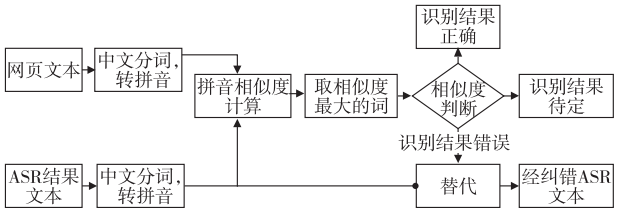

图8 借助网页文本对语音识别结果纠错框架

对同一视频的相关网页文本和语音识别文本,首先使用同一分词工具进行中文分词。以词语为单位,进行检错纠错。对字数相同的两个词 A 和 B,进行相似度距离度量,单独地分析组成其每一个字的相似度,词中每一字的相似度的总和为两个词的相似度。字 A_1 和 B_1 的相似度可以细化为以下几个值:汉字相似度 m_{whz}、拼音相似度 m_{wpy}、声母相似度 m_{ws}、韵母相似度 m_{wy} 和声调相似度 m_{wd}。对每个相似度的值定义如下:

a)若 A_1 和 B_1 的汉字相同,赋值 $m_{whz}=1$,同时对 m_{wpy}、m_{ws}、m_{wy} 和 m_{wd} 全部赋值为 0;否则赋值 $m_{whz}=0$,转到 b)去判断拼音相似度。

b)若 A_1 和 B_1 的拼音完全一致(包括声母、韵母和声调),则赋值 $m_{wpy}=1$,同时对 m_{ws}、m_{wy} 和 m_{wd} 全部赋值为 0;否则赋值 $w_{py}=0$,转 c)判断声母、韵母、音调值是否相同。

c)若 A_1 和 B_1 声母一致,则 $m_{ws}=1$,否则 $m_{ws}=0$;若 A_1 和 B_1 韵母一致,则 $m_{wy}=1$,否则 $m_{wy}=0$;若 A_1 和 B_1 音调相同,则 $m_{wd}=1$,否则 $m_{wd}=0$。

两个字的汉字、拼音、声母、韵母和音调分别给予不同的权值,经过不断实验和分析语音识别中的一些错误,对于权值定义如下:若汉字相同,给予最大权值 1;若拼音相同,给予权值 0.7;若声母相同,给予权值 0.3;若韵母相同,给予权值 0.3;若音调相同,给予权值 0.1。那么 A_1 和 B_1 相似度 $\mathrm{sim}(A_1,B_1)$ 计算式为

$$\mathrm{sim}(A_1,B_1)=1\times m_{whz}+0.7\times m_{wpy}+0.3\times m_{ws}+0.3\times m_{wy}+0.1\times m_{wd} \tag{5}$$

两个词 A 和 B 的相似度 $\mathrm{sim}(A,B)$ 计算式为

$$\mathrm{sim}(A,B)=\frac{1}{N}\sum_{K=1}^{N}\mathrm{sim}(A_K,B_K) \tag{6}$$

其中:N 为词 A 中含有的汉字个数;$\mathrm{sim}(A_K,B_K)$ 为词 A 和 B 中第 K 个字的相似度。

为方便表述,这里规定语音识别结果对应的文本为文档 1,与视频相关的网页文本为文档 2。对于文档 1 中的每一个词 $m_1,m_2,\cdots,m_k$,分别计算其与文档 2 中的每一个词 $n_1,n_2,\cdots,n_l$ 的相似度(只取字数相同的词),然后取相似度最大的词,判定其相似度所属范畴,进行下一步处理。

以 m_1 为例,判断其是否为正确的语音识别结果,并对错误纠错,采取以下方式:寻找文档 2 中与 m_1 相似度最大的词(假设为 n_t),即

$$\mathrm{sim}(m_1,n_t)=\max(\mathrm{sim}(m_1,n_1),\mathrm{sim}(m_1,n_2),\cdots,\mathrm{sim}(m_1,n_l)) \tag{7}$$

若 $\mathrm{sim}(m_1,n_t)=1$,即在网页文本中存在与视频语音识别完全相同的词,基本可以判定词 m_1 为正确的语音识别结果。

若 $\mathrm{sim}(m_1,n_t)\neq 1$,则将 $\mathrm{sim}(m_1,n_t)$ 与指定阈值 T 进行比较,若 $\mathrm{sim}(m_1,n_t)<T$,则说明在网页文本中不存在与 m_1 发音相似的词,m_1 正确与否有待下一步确定。

若 $T<\mathrm{sim}(m_1,n_t)<1$,则说明在网页文本中存在与 m_1 发音接近的词,但并不是同一个词。由于网页文本的准确度和语音识别结果的高错误率,可以认为语音识别的结果是错误的,则以 n_t 代替 m_1。

其中阈值 T 的大小为影响该检错纠错系统的关键,若 T 值设置过小,则容易发生虚惊,将正确的语音识别结果错判;若 T 值设置过大,则会导致检错过程发生过多的漏检,使本方法无法发挥作用。经过对错误词汇的数据统计和不断实验,本文选定阈值 $T=0.6$。

4 实验数据和结果分析

实验数据是从一周内的央视新闻类节目(包括 CCTV-1、CCTV-13 频道的《新闻联播》、《新闻 30 分》、《朝闻天下》等节目)中随机选取 30 段,共计 48 min,视频分辨率为 640×480,作为实验视频数据集;对应时期内,采用本文 2.3 节提到的 Nutch 搜索引擎对新华网进行每日定时爬虫,每日爬取网页为 1 800 个左右作为实验的网页库。根据本文内容对以下三个部分进行实验测试,并对实验结果作简要分析。

a)主题字幕信息提取。表 1 为主题字幕信息提取结果。由表 1 可知,前文所述的主题字幕检测提取框架,在主题字幕检测和背景颜色单一且与前景颜色色差大的文本框的识别准确率方面都取得了较好的效果,但对背景复杂或与前景颜色色差小的文字框的 OCR 识别结果不够理想,对这一部分需要进一步采取去噪等图像增强方式后再进行二值化。存在的问题主要有:(a)对于台标字幕,尽管在本文的文字框过滤系统中给予了排除规则,但仍有部分台标误判为文本框;(b)文本框的背景复杂与简单在这里是人工判别,需要研究适合的方法实现自动分类。此外,本文的重点工作是多模态特征的融合,在主题字幕信息提取方面只是选取了比较合理的方法,所采取的方法和取得的结果并没有去与该领域内最高水平及最新进展作比较。

表 1 主题字幕信息提取结果

文字总数	正确检测数	背景简单/复杂	OCR 识别正确字数
442	418	背景简单,字数:274	255
		背景复杂,字数:144	78

b)网页文本与视频故事的相关度。网页文本作为一种扩充的语义,如果准确度不高则会产生语义偏离,且对 ASR 结果的纠正不会起到太好的效果。这里采用多人打分、加权平均的人工方式,用优、中、差三个档次评价网页文本与新闻视频的相关度,结果如表 2 所示。对实验视频分析可得,对于主题字幕命名实体较强的国内外政治新闻,特别是主要领导人活动,网页与视频相关性较好,如“×××会见世界银行行长”;而文化、体育类新闻,特别是在一段时间内多次发生但视频字幕信息不是很完善的视频,如“北方多数地区持续大雨天气”,得到的网页与新闻视频故事相关性较差。

表 2　搜索到的网页与新闻视频的相关性

视频段	相关度:优	相关度:中	相关度:差
30	19	6	5

c)音频信息提取。音频分类是为语音识别做准备,这里不作单独的数据统计。在上面所得的网页文本与视频相关性优、中、差三部分视频中各取一段,作 ASR 识别和纠错结果(表 3)统计。由表 3 可知,由 3.3 节所提方法可提高语音识别结果的准确率,且网页与视频相关度越好纠错性能越好。对于表 3 中的视频 E、F,纠错几乎没有发挥作用,甚至有将正确结果变成错误的词语的可能。此外,语音识别错误导致分词不准确也是影响纠错效果的一个主要原因。借助与视频相关的网页文本对语音识别结果纠错是可行的,但本文采取的语音相似度规则作用有限,应该在本文工作的基础上继续从句法分析、句子成分依存关系的角度去分析问题。另外,在纠错算法中,应采用更智能的方式弥补启发式规则的不足。

表 3　ASR 识别及纠错结果

视频与网页相关性	语音汉字总数	ASR 识别正确字数	纠错后正确字数
视频 A　优	580	348	406
视频 B　优	632	391	454
视频 C　中	648	382	412
视频 D　中	550	335	367
视频 E　差	480	280	276
视频 F　差	530	320	324
总数及正确率	3 420	2 056　60.1%	2 239　65.4%

由实验结果可知,本文提出的基于多模态特征的新闻视频语义分析框架是可行的,在主题字幕信息提取、音频信息提取、利用主题字幕信息搜索到相关网页、借助网页信息对语音识别结果进行纠错等方面都取得了较好的效果,在多模态特征融合基础上实现了新闻视频高层语义较为准确的提取。

5　结束语

本文提出一种新闻视频语义分析框架,对视频中的主题字幕进行检测提取,借助 OCR 工具进行识别;在音频分类的基础上进行语音识别。在此基础上根据主题字幕得到的文字信息借助搜索引擎得到与新闻视频故事相关的新闻网页,基于网页文本对语音识别的结果进行纠错,从而得到更准确的视频语义信息。通过搜集到的中等规模的新闻视频的实验验证了框架的可行性。未来的工作是根据提取出的视频语义对视频进行分类、索引、内容安全性分析等工作。

参考文献:

[1] YANG Jun, HAUPTMANN A G. Naming every individual in news video monologues[C]//Proc of the 12th ACM International Conference on Multimedia. 2004:580-587.

[2] 冀中. 基于多模态信息的新闻视频内容分析技术研究[D]. 天津:天津大学,2007.

[3] 刘亚楠,吴飞,庄越挺. 基于多模态子空间相关性传递的视频语义挖掘[J]. 计算机研究与发展,2009,46(1):1-8.

[4] 赵志城. 故事视频的语义分析与提取[D]. 北京:北京邮电大学,2008.

[5] 郑李磊,谢磊,王晓璇,等. 中文新闻字幕自动生成系统的设计与实现[C]// 第十八届中国多媒体学术会议. 2009:232-239.

[6] FENG Hua-min, FANG Wei, LIU Sen, *et al*. A new general framework for shot boundary detection and key-frame extraction[C]//Proc of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval. New York:ACM Press,2005:121-126.

[7] 范竞往,翟晓飞,封化民,等. 一种双层新闻逻辑单元分割框架[C]//第十四届中国多媒体学术会议. 2005:75-80.

[8] FENG Hua-min, JIANG Chao, YANG Xin-hua. An audio classification and speech recognition system for video content analysis[C]// Proc of the 2nd International Conference on Multimedia Technology. 2011:5271-5276.

[9] 李雪龙,封化民,刘飚,等. 一种改进的视频标题检测与提取方法[J]. 江西师范大学学报,2008,32(2):157-164.

[10] HUA Xian-sheng, LIU Wen-yin, ZHANG Hong-jiang. An automatic performance evaluation protocol for video text detection algorithms[J]. Circuits and Systems for Video Technology, 2004, 14(4): 498-507.

[11] LTP 使用文档 v1.4[EB/OL]. http://ir.hit.edu.cn/ltp/#ner.

[12] Welcome to apache nutch[EB/OL]. http://nutch.apache.org/.