

# 中山大学硕士学位论文

## 基于查询自适应的视频缩略图自动生成方法 与应用

### Approach and Application on Automatic Generation of Video Visual-Text Thumbnail Based on Content Query

学位申请人： 戚 鑫

指导教师： 罗笑南 教授 林淑金 副教授

专业名称： 软件工程

答辩委员会主席(签名): \_\_\_\_\_

答辩委员会委员(签名): \_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

---

二零一七年五月二十三日

# 论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究作出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：

# 学位论文使用授权声明

本人完全了解中山大学有关保留、使用学位论文的规定，即：学校有权保留学位论文并向国家主管部门或其指定机构送交论文的电子版和纸质版，有权将学位论文用于非赢利目的的少量复制并允许论文进入学校图书馆、院系资料室被查阅，有权将学位论文的内容编入有关数据库进行检索，可以采用复印、缩印或其他方法保存学位论文。

学位论文作者签名：

日期： 年 月 日

导师签名：

日期： 年 月 日

论文题目： 基于查询自适应的视频缩略图自动生成方法与应用

专 业： 软件工程

博 士 生： 戚 鑫

指导教师： 罗笑南 教授 林淑金 副教授

## 摘要

摘要概括论文的主要信息，包括研究目的、方法、成果及最终结论。字数控制在满一页但不超过两页,硕士论文摘要一般不超过1200字。博士论文摘要一般不超过2000字。关键词是供检索用的主题词条，应采用能覆盖论文主要内容的通用词。关键词本科3-5个，硕士5-7个，博士7-9个，关键词之间用”,”分割，关键词不能是英文简写。

本科、硕士论文摘要一般采用2段，第一段研究背景（包括理论背景、应用背景）、研究环境、方法手段、影响和前景，第二段研究的内容、成果、价值、意义和不足之处。

博士论文摘要可分段介绍创新点。

**关键词：**学位论文，格式，模板

Title: Approach and Application on Automatic Generation of Video Visual-Text Thumbnail Based on Content Query  
Major: Software Engineering  
Name: Name  
Supervisor: Prof. Xiaonan Luo Shujin Lin

## Abstract

The American students are part of one of the most ambitious undertakings in the history of education: the American effort to educate an entire national population. The goal is-and has been since the early decades of the republic-to achieve universal literacy and to provide individuals with the knowledge and skills necessary to promote both their own individual welfare as well as that of the general public. Though this goal has not yet been fully achieved, it remains an ideal toward which has been made is notable both for its scope and for the educational methods which have been developed in the process of achieving it.

About 85% of American students attend public schools. The other 15% attend private schools, for which their families choose to pay special attendance fees. Four out of five private schools in the United States are run by churches, synagogues or other religious groups. In such schools, religious teachings are a part of the curriculum, which also includes the traditional academic courses of reading, mathematics, history, geography and science.

**Keywords:** Mesh Editing, Mesh Deformation, Sketch-Based Interface, Linear Constrained Deformation, Least-Squared Editing

# 目 录

摘要.....	I
Abstract.....	1
目录.....	2
引言.....	1
第 1 章 综述 .....	1
1.1 研究背景与意义 .....	1
1.2 国内外研究现状 .....	3
1.3 本文研究工作及创新点 .....	10
1.4 本文的组织安排 .....	11
第 2 章 基于双通道信息融合的视频主题边界检测方法 .....	1
2.1 视频双通道信息预处理方法 .....	1
2.2 视觉通道的主题边界检测算法 .....	3
2.3 听觉通道的主题边界检测算法 .....	5
2.4 基于双通道线索融合的主题边界检测 .....	10
2.5 本章小结 .....	12
第 3 章 基于查询自适应的缩略图自动生成方法 .....	14
3.1 视频内容与用户查询相关性度量 .....	14
3.2 缩略图图文内容提取 .....	14
3.3 缩略图排版方法 .....	14
3.4 实验结果及评估 .....	14
3.5 本章小结 .....	14
第 4 章 基于查询自适应的视频缩略图生成方法应用 .....	15
4.1 视频缩略图生成系统应用设计 .....	15
4.2 视频缩略图生成系统应用实现 .....	15
4.3 视频缩略图生成系统应用评估 .....	15
4.4 本章小结 .....	15
第 5 章 总结与展望 .....	16
5.1 本文工作总结 .....	16

5.2 今后工作展望 .....	16
参考文献 .....	17
附录 .....	20
作者简介 .....	21
致谢 .....	23

# 引言

从引言起为论文正文主体部分。页码从1开始编排。引言(前言)部分内容主要包括5个方面:为本研究课题的学术背景与环境、存在的问题、意义(不需要详细解释、只需几种说明解决了哪几个问题)、突破点、比较结果及优缺点。

注意不要与摘要内容雷同。

此部分是第一章(综述)和第五章(总结)的两章内容的总结

如果引言部分省略，可以合并到第一章综述中去。

## 1.正文书写格式说明:

每段落首行缩进2字;或者手动设置成每段落首行缩进2字,宋体,小四,:多倍行距(1.5倍),前段、后段均为0行,取消网格对齐选项。可以采用样式和格式里面的”正文格式”来格式化正文文本。注意:每两级标题之间一定要有过渡性的文字,避免两级标题直接相连。一般而言,硕士论文正文页数在50页以上,而博士论文页数在100页以上。

脚注书写格式说明:

①一般而言,网页地址、国家\地方标准都只能作为脚注,而并非出现在参考文献中。

---

① 网页标题,<http://www.sysu.edu.cn>



# 第1章 综述

近年来,随着网络带宽的不断提高,社会化网络和网络流媒体技术的发展,多媒体信息特别是视频已经成为当今信息时代主要的数据来源形式。视频缩略图在呈现视频内容上扮演着非常主要的作用,它直接决定了用户是否要点击视频以观看视频具体内容。好的视频缩略图能让用户第一时间了解视频内容,极大提高了用户检索效率。

本章首先介绍视频缩略图自动生成方法的研究背景与意义,接着从视频内容的提取,视频语义分割,视频缩略图的自动生成三个角度出发分析国内外研究现状,然后介绍本文的主要工作和创新点,最后简述本文的组织安排。

## 1.1 研究背景与意义

自上世纪九十年代以来,随着互联网技术,多媒体技术的发展和人们对信息需求的不断增长,越来越多的信息通过多媒体的形式展现在用户面前,例如视频,音频,动画,图像等。相比于其他多媒体数据形式,视频以其生动性、直观性、信息的丰富性备受用户的喜爱,特别是网络视频用户规模不断扩大,截至2016年12月,中国网络视频用户规模达5.45亿,较2015年底增加4064万人,增长率为8.1%<sup>①</sup>。网络视频已成为一种人们分享信息,想法,趣事的主要媒介形式。

随着视频数据的爆发式增长,以提供视频分享为主要业务的视频门户网站也在蓬勃的发展,国内有优酷<sup>②</sup>、爱奇艺<sup>③</sup>、腾讯、搜狐等数百家视频门户网站,国外视频门户网站有youTube<sup>④</sup>、Yahoo Video<sup>⑤</sup>、AOL Video等等。数量众多的视频门户网站为用户提供了海量的视频信息,极大的满足了用户对视频数据的需求,但同时也增加了用户检索视频的难度。这些视频分享网站中的视频的来源大多是由众多用户上传,上传的同时也伴随着一些视频的元信息以便于视频搜索引擎检索或者吸引其他用户点击,例如视频标题,描述,标签,缩略图等等,其中相比于标题,描述这些视频文本元信息,视频缩略图更生动,并能达到预览视频内容的效果,用户能通过缩略图直观地了解视频内容,因此视频

<sup>①</sup> 数据来源前瞻产业研究院:<https://bg.qianzhan.com/report/detail/459/170313-223982fb.html>

<sup>②</sup> <http://www.youku.com/>

<sup>③</sup> <http://www.iqiyi.com/>

<sup>④</sup> <https://www.youtube.com/>

<sup>⑤</sup> <http://video.search.yahoo.com/>

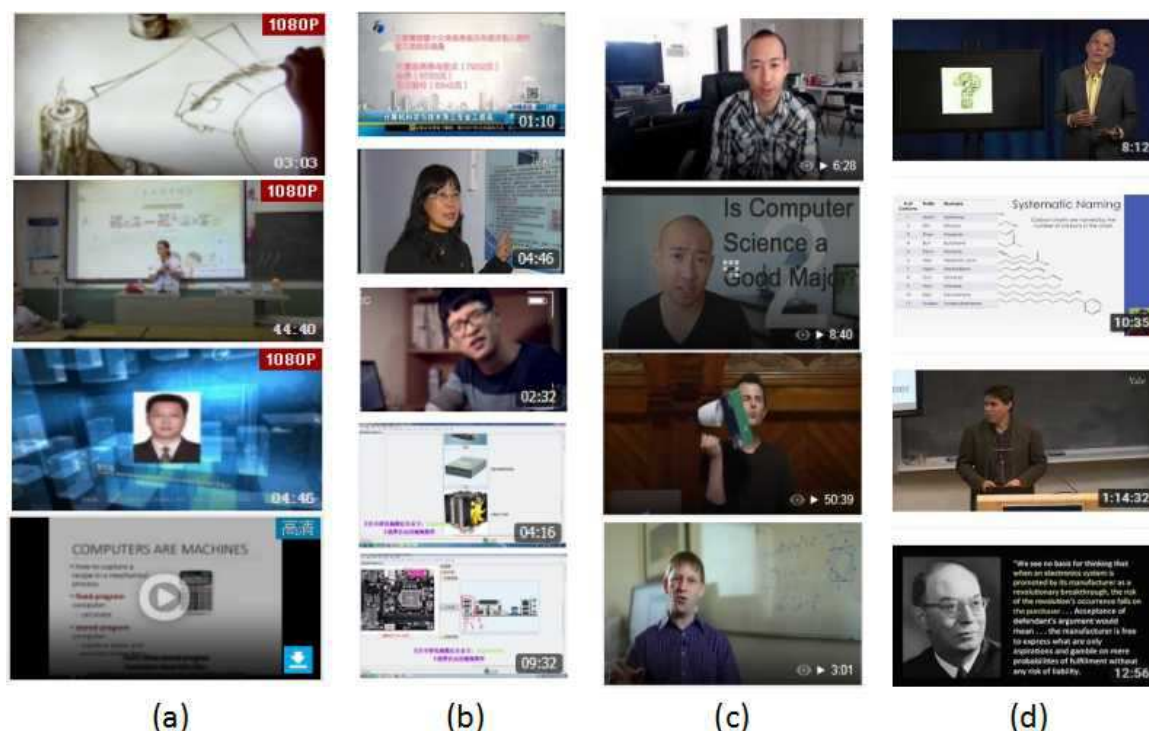


图 1-1 国内外知名视频网站以“计算机科学”/“computer science”为关键词检索的部分结果：(a)优酷视频。(b)爱奇艺视频。(c)YouTube Video。(d)Yahoo Video。

缩略图已成为一种常用的视频和用户交互的技术。Yuli<sup>[1]</sup>等人指出视频缩略图对用户浏览行为上有很强的影响力。Michael<sup>[2]</sup>等人的研究也表明高质量的缩略图大大提高了视频检索效率和用户满意度。

视频上传用户一般选择视频中的一帧作为视频缩略图，以youtube视频网站为例，当用户上传视频后，系统随机生成视频三帧画面供用户选择其中一帧作为视频缩略图，然而这样的缩略图往往无法反映视频内容，质量较差。如图1-1所示，用户以“计算机科学”或者“computer science”为关键词在国内外知名视频网站检索到的部分结果，可以看出仅从现在视频网站提供的视频缩略图用户几乎得不到任何任何有效的信息。如果用人工挑选精心生成缩略图，固然可以提高缩略图的质量，但是这样做太过费时，为每一个视频都精心打造一个缩略图是不现实的<sup>[3]</sup>。而且，视频缩略图和用户查询意图存在鸿沟，无法满足用户的要求。如图1-3所示，图中的3张子图是同一视频的3帧图片，都可以成为视频的缩略图，但它们包含有截然不同的信息：从第一张子图得出这个视频是一个TED演讲类视频；第二张图片表明视频中有讲到蒙娜丽莎的微笑；第三张子图则表明视频有讲到关于健身减肥的内容。不同的用户或者同一用户不同时间在检索视频时显然带有不同的意图，所以当两次检索结果有同一视频时，他

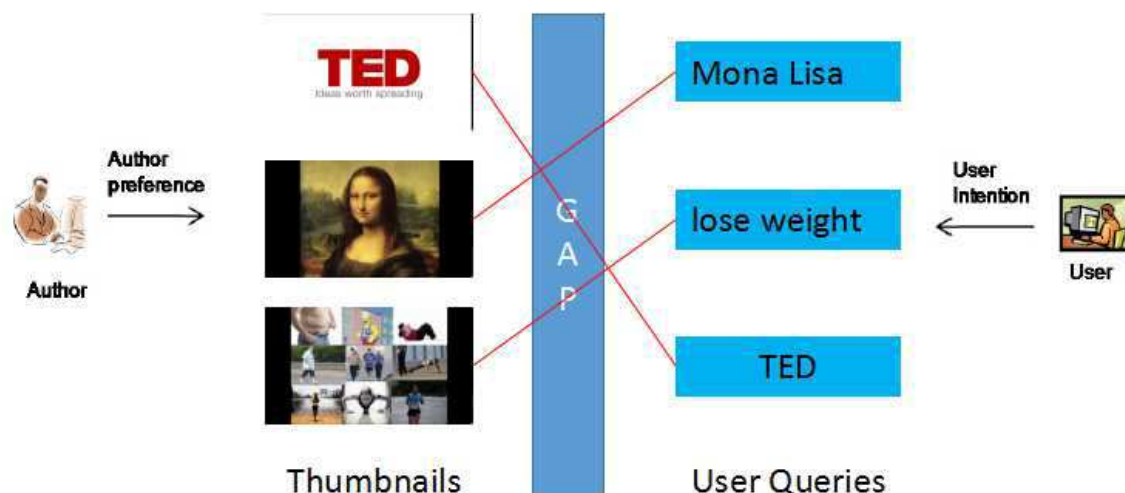


图 1-2 用户查询意图与视频缩略图之间存在的鸿沟

们关注的内容也不尽相同。但是传统的视频缩略图具有静态性，一旦生成就不可改变，且生成缩略图的时候并不知道将来用户会以什么查询词检索视频，我们当然希望以“Mona Lisa”为关键词是检索到的视频缩略图为图1-2中的第二个子图，遗憾的是传统的视频缩略图生成方法显然满足不了需求，很可能会出现用“Mona Lisa”为关键词检索到的视频的缩略图为图1-3中第一个或者第三个子图，这时视频缩略图的内容显然不合适的，与用户的检索意图背道而驰。因此用户查询意图与视频缩略图之间存在巨大鸿沟，用户在缩略图上找不到自己感兴趣的视频内容，导致用户检索效率降低，用户满意度下滑。

本文的目标就是基于视频多通道内容信息，跨越用户查询意图与视频内容的鸿沟，生成动态的并且自适应用户查询的视频缩略图。通过该目标，帮助用户快速了解自己视频中感兴趣的内容，提高检索视频的效率。

## 1.2 国内外研究现状

视频是一种非结构化的流媒体数据，其内容信息复杂而丰富。从视觉上看，视频是由一系列视频帧组成；从听觉上看，视频也含有大量嵌入视频的语音信号。好的视频缩略图首先要能很好的反映视频内容，所以生成视频缩略图之前首当其冲要解决的问题是提取视频内容，并理解视频有效信息，即要清楚地知道视频由几部分组成，每部分蕴含了什么信息，这就涉及到视频主题边界检测，以及视频语义内容概括。近年来，国内外学者对以上几个方面做了大量的研究，本小节下面的内容主要从视频内容提取与分析研究现状，视频主题分割研究现状，视频缩略图生成方法研究现状三个角度出发，对前人的研究加以归纳总

结。

### 1.2.1 视频内容提取与分析研究现状

视频内容提取与分析的目标是将复杂抽象的数据转换成计算机容易处理的格式，并提取、整合、挖掘有效的信息，其中涉及的领域非常广泛，诸如图像处理，语音识别，模式识别等。通过归纳关于视频内容提取与分析的相关文献，将提取分析的关键过程归纳为：镜头分割与关键帧提取方法，基于OCR<sup>①</sup>的视频文字提取方法，视频语音识别方法。

#### (1) 镜头分割与关键帧提取方法

镜头分割的方法主要包含基于颜色直方图的镜头分割方法<sup>[4,5]</sup>，基于像素颜色差异的镜头分割方法<sup>[6]</sup>，基于运动的镜头分割方法<sup>[7,8]</sup>，基于边缘特征的镜头分割方法<sup>[9]</sup>。基于颜色直方图的镜头分割方法统计视频图像的像素灰度分布或者颜色分布，核心原理是不同镜头之前图像的灰度和色彩会发生剧烈变化。该方法简单明了，计算复杂性低，缺点对光照强度和镜头运动速度太过敏感，基于运动的镜头分割方法解决了这一缺点，主要原理是基于点或块的运动矢量的估计，可以很好地检测镜头的变化，但是该方法计算复杂性太高。基于边缘特征的镜头分割方法考虑的是边缘在局部照明变化下大部分是不变的，该方法减少了由于运动和照明导致的不变性问题，但是对于图像内容复杂的视频镜头分割不甚理想。关键帧提取方法包括：特定帧法<sup>[10]</sup>、直方图平均法<sup>[11,12]</sup>、基于运动特征的关键帧提取方法<sup>[13,14]</sup>、基于内容的分析方法<sup>[15]</sup>。

#### (2) 基于OCR的视频文字提取方法

光学字符识别（Optical Character Recognition, OCR）通过光学机制识别字符，可以将数字图像中的文字信息转换为可编辑文本的过程。OCR<sup>[16]</sup>是一个计算机视觉/图像中活跃的研究领域，它通常包含两个步骤，首先是定位图像中的包含文本的区域，然后是识别区域中的文本。Neumann<sup>[17]</sup>等人提出了用于定位图像中文本的算法。OCR技术也已经发展的比较成熟，市面上已经有非常成熟的OCR识别引擎，例如Tesseract, OmniPage, Readiris等等。

#### (3) 视频语音识别方法

视频语音识别是提取语音特征，将视频中复杂的语音信息转换为可编辑易处理的文本数据。语音识别的方法主要包括：基于统计模型的方法，基于学习的方法，基于规则的方法<sup>[18,19,20]</sup>。现在有一些高级的语音识别工具，英文的工具

<sup>①</sup> 维基百科OCR介绍：<https://en.wikipedia.org/wiki/OCR>

有MSR<sup>①</sup>,CMU Sphinx<sup>[21]</sup>,Nuance Dragon<sup>②</sup>等,中文的工具主要有科大讯飞的讯飞开放平台<sup>③</sup>。

### 1.2.2 视频主题分割研究现状

视频主题分割不同于视频镜头分割和场景分割,它的目标是检测视频语义主题边界,把视频按语义分割成一个个语义独立,内容连贯的主题片段。通过视频主题分割,用户能清楚地了解视频结构,掌握视频内容。视频语义分割问题可以转化为文本主题分割问题,两者的本质都是把一段内容丰富,结构复杂的信息分割成内容独立的片段,研究学者对这方面进行了深入的研究。

Hearst<sup>[22, 23]</sup>提出的Textiling算法先把文本数据分成连续的,非重叠的文本块,提取文本块的“bag of word”特征,用余弦相似度衡量文本块的相似度,然后比较文本块间的“bag of word”特征间的差异,Hearst认为差异较大的文本块就是主题的边界。Satanjeev<sup>[24]</sup>等人将Textiling算法应用在会议主题分割,取得了比较好的结果。Textiling算法使用词袋向量来描述文本块的特征,但是词袋模型往往非常稀疏,而且词袋模型不能反映近义词语义的相似,例如“计算机”和“电脑”两个词在语义上很相似,但表示成词袋向量后则截然不同,所以词袋向量不利于文本语义的表示。Gally<sup>[25]</sup>等人提出了一个基于Textiling的算法,不同于Textiling算法,它用词的tf-idf<sup>④</sup>权重代替了单一的词频权重从而取得了更好的效果。Martin<sup>[26]</sup>等人提出了TopicTiling算法,改进了Textiling算法。TopicTiling算法基于LDA<sup>[27]</sup>主题模型,LDA是一个三层贝叶斯模型,可以认为一篇文章的每个词都按一定概率属于某个主题。如图1-3所示,TopicTiling算法首先对语料库进行LDA训练,挖掘出语料库潜在的主题,并得出每一篇文章每一词属于某个主题的概率,然后为每一个主题分配一个TopicID,这些ID被用来计算相邻文本块的余弦相似度,即把“bag of word”特征替换为“bag of TopicID”,然后再进行分割流程。TopicTiling算法利用“bag of Topic”既将冗长的“bag of word”特征降了维,又因为近义词往往属于同一主题,解决了近义词特征不同的问题。

基于Textiling算法是具有线性复杂度,且取得了较好的效果,但是它会出现过度分割或者分割不足的问题,原因是在确定主题边界的阈值难以确定,不能自适应文档的长度。而且把Textiling算法应用在视频主题分割时,不但可以利用由音频转换而来的文本信息,还可以结合视频的图像信息加以优化,本文下

① Microsoft Speech Recognition(MSR) API: <https://msdn.microsoft.com/library/ee125663.aspx>

② Nuance Dragon Speech Recognition Software(Nuance) url: <http://www.nuance.com>

③ 讯飞开放平台网址: <http://ai.xfyun.cn/>

④ tf-idf百度百科<https://baike.baidu.com/item/tf-idf/8816134?fr=aladdin>



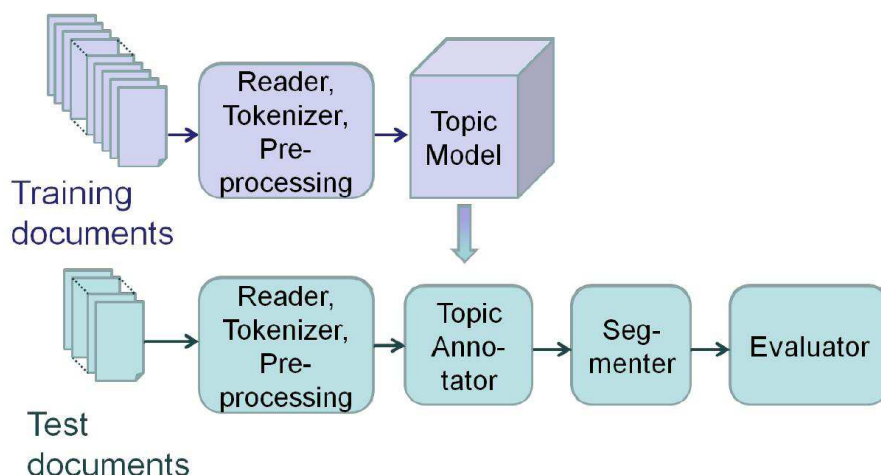


图 1-3 TopicTiling算法主题分割流程

面的章节会详细介绍。

除了无监督的方法，Nguyen<sup>[28]</sup>等人训练了一个无参数的层次化的模型解决了多人谈话的主题分割主题分割问题，认为每个文本块都与一个说话人有关，通过训练出的模型判断文本块的说话人，从而判断主题的转移。Chen<sup>[29]</sup>等人提出了一种基于自我验证的声学分割方法（SACuts）来把口语文档分割成主题片段。相比于其他方法，他们的方法仅用了声学级别的信息确定主题数量，在没有任何的额外的计算负担下解决了过度分割和分割不足的问题。Fragkou<sup>[30]</sup>利用NER（Entity Annotation Recognition）对语料库进行标注，算法首先对语料库每一个词、短语都被归类为提前定义后的实体类型，随后词、短语被替换成替换成实体标识符，然后用已有的分割算法，例如Choi<sup>[31]</sup>提出的C99b算法，Utiyama<sup>[32]</sup>等提出的基于统计学模型的分割算法，Kehagias<sup>[33]</sup>等人最小化全局分割代价优化算法等，进行主题分割。以上算法各有优劣，且取得了较好的效果，但应用到视频主题分割问题上会出现不适用的问题。视频数据复杂丰富，往往包含很多说话人角色，若利用Nguyen、Chen的方法进行声学特征或者说话人的识别进行主题分割，会出现把一个主题过度分割成很多小碎片的问题。且当今视频数据量增长迅猛，很难训练出鲁棒性很好的模型适用所有的视频数据，而且标注，训练语料库往往非常耗时，不能大规模推广。

### 1.2.3 视频缩略图生成方法研究现状

视频缩略图是视频分享网站给用户提供的友好接口，质量高的缩略图能帮助用户快速了解视频内容，吸引用户点击，从而极大提升视频检索系统的检索性能，改善用户检索视频体验。目前现有的视频分享网站往往选取第一帧或随

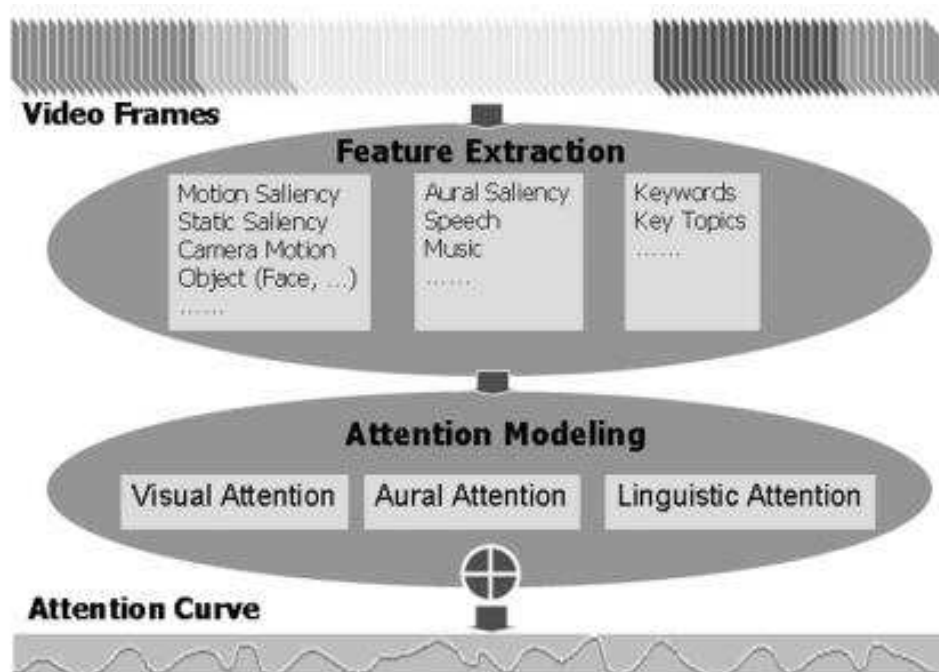


图 1-4 用户关注度模型框架

机选取一帧作为视频缩略图，或者像youtube的做法随机抽取多帧让用户选择一帧作为视频缩略图。这类方法虽然简单易行，但很难达到好的效果。视频缩略图生成研究涉及领域很广，例如计算机视觉、机器学习、数据挖掘、美学设计等等，近年来关于这方面的研究也越来越多，下面详细分析视频缩略图的研究现状。

Wolf<sup>[34]</sup>等人认为运动变化最小的视频帧最具有代表性的帧，算法首先基于光学流动分析衡量每一个视频帧的运动量，然后选取运动量最小的帧。Wolf认为拍摄视频过程中经常会为了一个镜头使用一系列相机镜头的运动来构建复杂的信息，所以往往非常重要的帧算法中描述的运动量最小，但是这只符合一些例如电影类，纪录片类视频，对一些新闻类，演讲类，教育类等镜头变化不显著的视频不是很适用。Ma<sup>[35]</sup>等人提出了一个用户关注度模型，如图1-4所示，首先对视频视觉信息，听觉信息以及视频文本信息进行特征提取，根据这些特征和用户关注度模型计算用户视觉关注度，听觉关注度，语言关注度，然后将多通道关注度融合成用户关注度曲线，取曲线最大值点的关键帧作为视频缩略图。Cong<sup>[36]</sup>等人从稀疏编码和数据重建的角度出发，核心思想是将视频帧编码成字典，如果一个视频帧的字典能够最好的重建原视频，那么就认为这一帧就是最有代表性的视频帧。相似地，Guan<sup>[37]</sup>等人提出了一个稀疏重建框架选取最有代表性的视频帧。Kang<sup>[38]</sup>等人具体度量了视频帧的代表性，认为帧的质量，图像

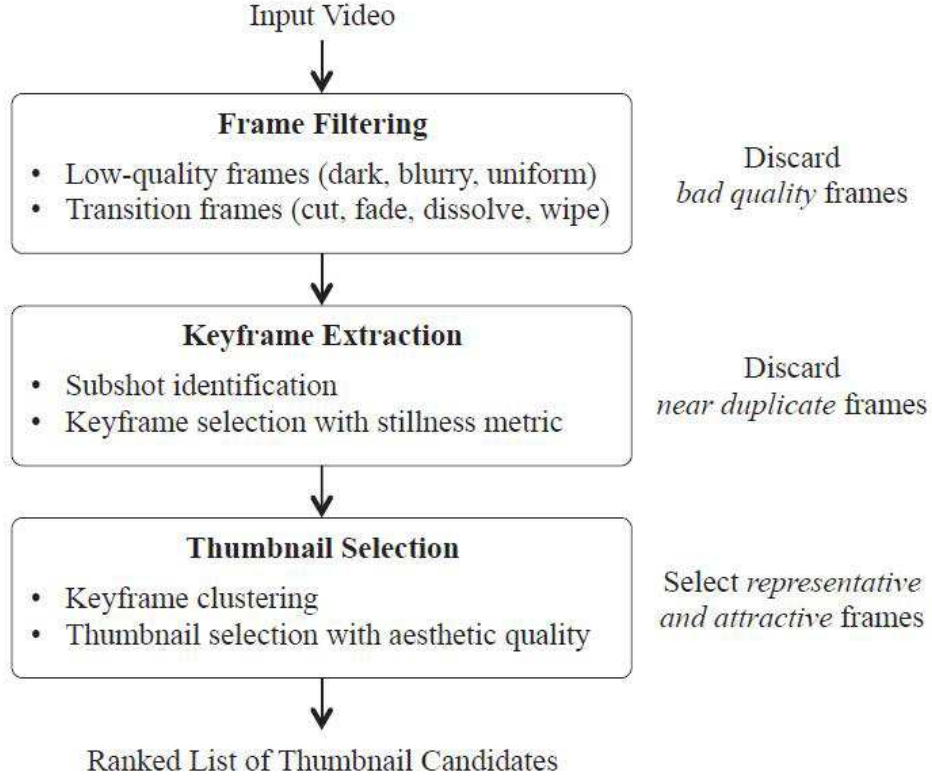


图 1-5 用户关注度模型框架

细节，内容相关性，用户关注度是衡量视频帧是否具有代表性的关键特征，然后基于高斯混合模型对视频帧代表性训练，建模，从而找出最具代表性的视频帧作为缩略图。如图1-5所示，Zhang<sup>[39]</sup>等人从信息丰富度，用户关注度，和美学特征3个方面定义了12个特征来衡量视频缩略图的质量，选取特征得分最高的视频帧作为视频缩略图。Song<sup>[40]</sup>等人的视频缩略图系统既考虑了视频内容相关性，又考虑了视觉美学质量。具体过程图1-6所示，首先过滤掉模糊，灰暗等低质量的帧；然后进行视频关键帧检测去掉重复的帧；最后进行关键帧聚类并结合美学质量评估，选取最有代表性又能吸引用户的视频帧作为视频缩略图。

以上方法都或多或少提升了视频缩略图的质量，但都忽略了用户的查询意图。Liu<sup>[41]</sup>等人的研究改进了这一点，他们的方法首先使用Joshi<sup>[42]</sup>的方法对视频关键帧序列排序，然后计算用户查询词与图片之间的关联性，将其融入到对关键帧的排序过程中，选取排名最高的视频帧作为视频缩略图。在Liu的方法中，词与图片的关联度的计算公式1.1，

$$S(I_u, w) = P(I_u, w) \approx P\left(\frac{I_u}{w}\right) = \frac{1}{n} \sum_{i \in n} s(I_u, v_i) \quad (1.1)$$



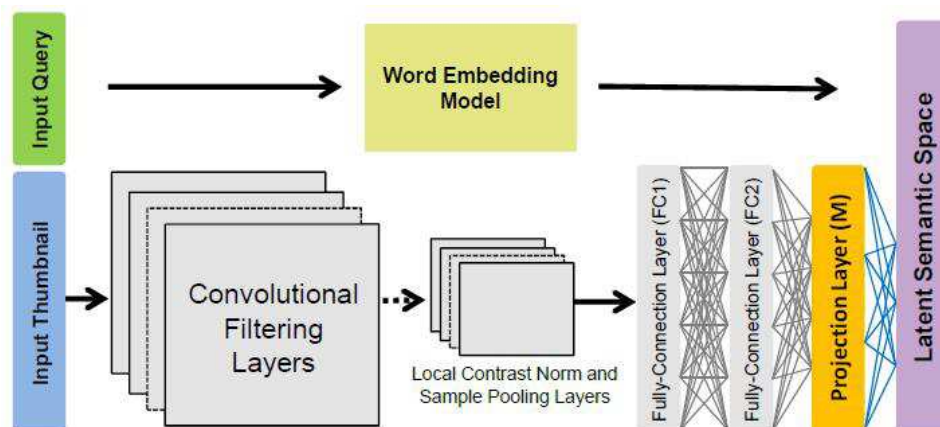


图 1-6 深度视觉-语义嵌入模型架构

其中 $S(I_u, w)$ 表示图片 $I_u$ 和查询词 $w$ 的关联度， $P(I_u, w)$ 表示表示图片 $I_u$ 和查询词 $w$ 关联的可能性， $v_1, v_2, \dots, v_n$ 表示用查询词 $w$ 在图片搜索引擎搜索到的topN个结果图片。由公式1.1可见，图片与查询词之间的关联度计算依赖于第三方图片搜索引擎，这种在线搜索图片的过程充满不确定性，且往往是非常耗时的。另外，关联度计算的有效性完全取决于第三方搜索引擎的有效性，所以可靠性不强。Liu<sup>[43]</sup>等人利用深度卷积神经网络（CNN）实现了一种基于多任务深度视觉-语义嵌入模型，该模型架构如图1-7所示，模型可以把用户查询词和视频缩略图都映射到潜在语义空间，通过计算查询词和缩略图在潜在语义空间中向量表示的相似度得出二者之间的关联度，跨越了用户输入的查询词语义与视频缩略图视觉的鸿沟，生成符合用户查询的视频缩略图。这个方法新颖有效，但是在计算缩略图和查询词关联度时仅仅考虑了缩略图的视觉信息，然而视频有多通道的信息，不但有视频帧等视觉信息，而且还有通过时间轴与视频帧相关联的音频信息，音频信息可通过语音识别转化为文本信息，这些与视频帧关联的文本信息能直观地反映视频帧的语义，在计算查询词和帧语义关联度时具有很大的价值。

通过对上述方法总结可以发现：（1）很多方法的本质是根据图片特征对视频的关键帧打分，然后取分数最高的视频帧作为缩略图；（2）缩略图的生成选取的标准主要考虑了视频内容相关性、图片质量、用户关注度、美学设计等等；（3）少部分方法考虑了用户的查询意图信息。但是几乎所有的方法都是只选择视频中的一帧作为缩略图，即使帧的质量很高，但由于一帧图像包含的图像信息非常有限，往往无法给用户足够有效的信息。而且，只有很少的方法考虑了用户的查询意图，他们的方法比较新颖，但仍存在一些问题，如上文所述，

文章<sup>[41]</sup>的方法具有不确定性而且非常耗时, 文章<sup>[43]</sup>的方法忽略了视频的音频信息, 以上方法都有较大的改进空间。

### 1.3 本文研究工作及创新点

视频缩略图自动生成方法是一项极具挑战的研究课题, 设计数据挖掘, 计算机视觉, 图像处理, 美学设计等多个学科, 是一个综合性很强的研究领域。本文提出了一种新颖的基于内容查询的图文视频缩略图生成算法, 它基于视频多通道信息, 又结合用户查询内容, 动态生成出既能深刻反映视频内容, 又能迎合用户查询意图的视频缩略图, 极大提升了用户检索视频的效率和满意度。

本文主要研究内容:

(1) 视频多通道内容的分析与处理。在视觉通道上, 进行镜头分割, 关键帧提取, 图片显著性区域检测, OCR识别; 在听觉通道上, 进行语音识别, LDA主题挖掘, 关键词提取。

(2) 视频主题分割与整合算法。对视频语音识别的结果进行主题分割, 结合视频镜头分割的结果, 检测视频主题边界, 将视频分割成一个个内容连贯, 主题鲜明的片段。分析每个片段的内容, 提取每个片段最具代表性的图文, 将每个片段的视频内容整合成图文二元组。

(3) 视频缩略图的生成算法。基于用户查询内容和由(2)得到的每个视频片段的图文二元组, 计算二者之间的关联度, 选取和用户查询最相关的视频内容, 再结合美学设计, 生成出美观的, 视频相关的, 自适应用户查询的视频缩略图。

本文主要的创新点在于:

(1) 提出了基于TopicTiling的视频主题分割算法, 它能准确的检测出视频内容的主题边界。算法首先改进了TopicTiling深度分数阈值难以确定的问题, 使其能自适应文本的长度选取合适的阈值进行主题分割; 将TopicTiling这种文本主题分割算法应用到视频主题问题上来, 结合视频多通道信息, 对视频进行主题分割。

(2) 提出了基于内容查询的图文视频缩略图的算法, 其创新点在于: 不在像传统思路那样选取一帧图像作为视频缩略图, 而是生成由若干张图像的显著性区域和相应的文字拼合的缩略图; 提出了将视频主题片段的整合成图文二元组的方法, 并实现了视频主题片段二元组与用户查询词的关联度量算法; 缩略图不仅反映了视频多个主题的内容, 而且自适应用户查询意图。

## 1.4 本文的组织安排

本文主要内容安排如下：

第1章，本文综述。首先介绍视频缩略图的研究背景和意义，接着从视频内容提取与分析，视频主题分割，视频缩略图生成方法三个角度分析了国内外研究现状，最后针对现有方法的局限性，提出了本文的主要研究工作和创新点。

第2章，提出了基于视频多通道内容的主题分割与内容整合算法。首先介绍视频多通道内容提取分析的方法；接着阐述基于TopicTiling的视频主题边界检测算法，根据主题边界，将视频分割成内容连贯的主题片段；最后阐述整合这些片段内容的方法。

第3章，提出基于查询的视频图文缩略图生成算法。首先介绍用户查询与视频内容相似性度量算法，查找与用户查询最相关的若干个视频主题片段图文内容；然后阐述如何利用美学设计知识将上述图文内容排版到一张视频缩略图中；最后是视频缩略图算法的实验结果和评估。

第4章，基于内容查询的视频缩略图生成算法应用。首先设计了一个视频检索系统；然后详尽的阐述该系统包括整体框架，模块布局以及交互应用设计的实现等；最后是该系统的实验评估。

第5章，本文的工作总结和今后工作的展望。

## 第2章 基于双通道信息融合的视频主题边界检测方法

### 方法

视频缩略图图文内容的代表性，缩略图内容与用户查询的相关性是衡量视频缩略图质量的关键指标。因此，自动理解视频内容是生成视频缩略图的基石，理解视频内容一般分为两步：第一步是检测视频主题边界，第二步是概括每一部分的主题内容。本章主要解决第一个问题，即检测视频主题边界，将视频分成内容独立的主题片段。

视频由视觉通道和听觉通道双通道信息组成，本文基于视频双通道线索检测视频的主题边界。首先对于视频视觉通道一系列视频帧，根据视频帧图像特征对视频进行场景分割，根据视频帧场景的切换位置推测视频主题的边界；接着对于视频听觉通道的视频语音文本特征进行基于TopicTiling算法的主题边界检测；接着，本文根据双通道线索的边界位置，提出了基于双通道信息融合的视频主题边界检测方法；最后，本文给出了实验结果评价方法并且进行了实验结果对比，证明了本文的视频主题边界检测方法准确，有效。

### 2.1 视频双通道信息预处理方法

视频数据是一种由视觉通道和听觉通道双通道信息共同组成的综合媒体数据，具有数据量大，信息结构复杂等特点。因此，在分析视频数据之前，需要对其做相应的预处理，将复杂无结构的视频流转化为计算机易于处理的结构化数据，本节从视觉通道和听觉通道的两个角度分别讨论视频数据的预处理方法。

#### 2.1.1 视觉通道信息的预处理方法

视频视觉通道上的信息是一系列的视频帧，本文对这些视频帧的预处理方法主要分为两部分：第一部分是镜头分割，第二部分是OCR识别，下面具体阐述这两部分工作。

视频帧具有极大的重复性，一秒视频通常包含20到40帧画面，时长为一小时的视频包含超过7万张以上的帧画面，如果直接分析这些数量巨大，重复性极高的视频帧内容，不仅耗时，而且很难挖掘到重要信息。镜头分割根据视频帧画面的差异，将内容发生转变的帧画面划分为不同的镜头，相同镜头只保留一

帧作为这个镜头的代表帧，过滤掉绝大多数冗余的视频帧。学界关于镜头分割的算法有很多，本文采用Apostolidis<sup>[44]</sup>等人提出的算法，其优点不但精度较高，而且通过GPU加速检测过程，可以极大提高镜头分割的效率。

视频帧画面时常会包含一些文字信息，这些文字信息能直观反应视频图像的语义信息，具有很重要的价值。光学字符识别技术（OCR）能提取图像的文字特征，识别图像上的文本内容。光学字符识别技术已比较成熟，目前市面上有很多商业性的OCR识别引擎，如OmniPage<sup>①</sup>,Readiris<sup>②</sup>，Tesseract等等。本文采用Matlab 2015b集成的Tesseract OCR引擎<sup>[45, 46]</sup>，其优点是文本识别准确率很高，而且开源，免费。图2-1给出了本文OCR的识别一个结果。

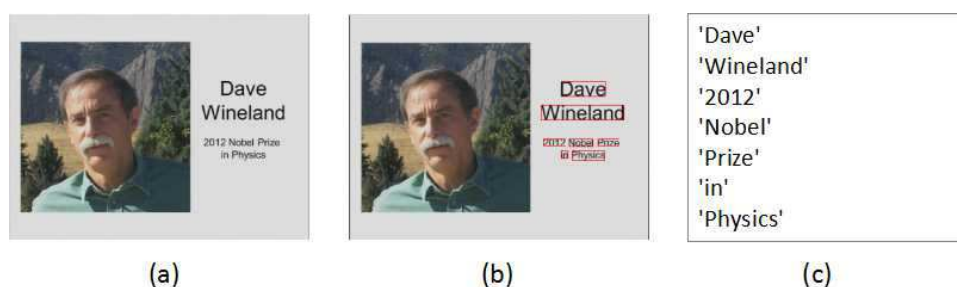


图 2-1 OCR识别结果：(a)为原视频帧，(b)为OCR识别的文本位置，(c)为OCR识别的文本内容

### 2.1.2 听觉通道信息的预处理方法

视频听觉通道上包含大量的语音信息，语音信息直接反映视频内容。自动语音识别（Automatic Speech Recognition, ASR）能将声音特征转化为计算机易于处理的文本。目前，ASR技术已经有了长足的发展，很多公司和机构提供功能强大的自动语音识别API，例如：美国卡内基梅隆大学开发的CMU Sphinx<sup>③</sup>开源工具包，其特点是非特定人，词汇量大，且能连续识别语音；Nuance Dragon<sup>④</sup>工具包提供语音听写功能，使得人们不需要打字就可以创建文档，电子邮件，填写表格等等。本文使用的是微软提供的语音识别工具包<sup>⑤</sup>。微软语音识别工具包已比较成熟，识别准确率极高，越来越多的学者将微软语音识别应用到各个领域，并取得了丰厚的研究成果<sup>[47, 48]</sup>。

上文通过微软语音识别工具包得到视频语音文本，但是往往这些文本中存

① OmniPage url:<http://www.nuance.com/for-business/by-product/pmnipage/index.htm>

② Readiris url:<https://readiris-pro.en.softonic.com/>

③ CMU Sphinx, url:<http://cmusphinx.sourceforge.net/>

④ Nuance Dragon Speech Recognition Software(Nuance), url:<http://www.nuance.com>

⑤ Microsoft Speech API, url:[https://msdn.microsoft.com/en-us/library/ee125663\(v=vs.85\).aspx](https://msdn.microsoft.com/en-us/library/ee125663(v=vs.85).aspx)

在大量的冗余信息，本节对这些文本信息做进一步的处理，挖掘出更加有价值的信息。首先过滤掉停止词（Stop Words）<sup>①</sup>和语音识别输出的冗余的时间戳信息，接着使用Porter算法<sup>[51]</sup>对文本进行提干处理。在英文中，一个单词常常是另一个单词的“变种”，如“happy”和“happiness”，“play”和“playing”，Porter算法能在线性时间对单词进行提干处理，把“happiness”处理成“happy”，“playing”处理成“play”。最后对文本提取关键词，本文使用Rose<sup>[52]</sup>等人的提出的RAKE（Rapid Automatic Keyword Extraction）算法。rake算法提取的关键词并不是一个单一的单词，有可能是一个短语。每个短语的得分由组成短语的单词累加得到，而单词的得分是单词的“度”与单词出现的频率的比值，其中，当与一个词共现的词越多，则该词的度就越大。表2-1给出了使用rake算法对youtube网站视频ID为“40riCqvRoMs”语音文本进行关键词提取的部分结果。

表 2-1 rake算法提取关键词结果

关键词	权值
potentially revolutionary technologies	9.0
explore unseen frontiers	9.0
highly connected neurons	9.0
visual processing apparatus	8.2
typical neural network	8.0
worldwide research community	8.0
neural network	5.0
vision lab	4.375
computer vision	3.91

## 2.2 视觉通道的主题边界检测算法

视频视觉通道由一系列视频帧组成，这些视频帧蕴含着视频语义。如图2-2所示，视频视觉通道信息包含视频帧、镜头、场景和视频四个层次。视频帧是组成视频的最小单位，本文以1秒为步长采样视频帧，即一个小时的视频包含3600张视频帧。镜头是一组连续的视频帧，代表视频拍摄过程中某一个镜头下的图像，同一个镜头下的视频帧图像特征相似，内容变化小。场景有一组语

<sup>①</sup> 本文使用的停止词列表：<http://xpo6.com/list-of-english-stop-words/>



义相关的镜头组成，多个场景就组成了整个视频。不同场景的视频片段一般蕴含了不同的视频语义主题，所以视频场景边界也暗示着视频主题边界。因此，若仅利用视频视觉通道信息，本文将视频主题分割问题转化为视频场景分割的问题。

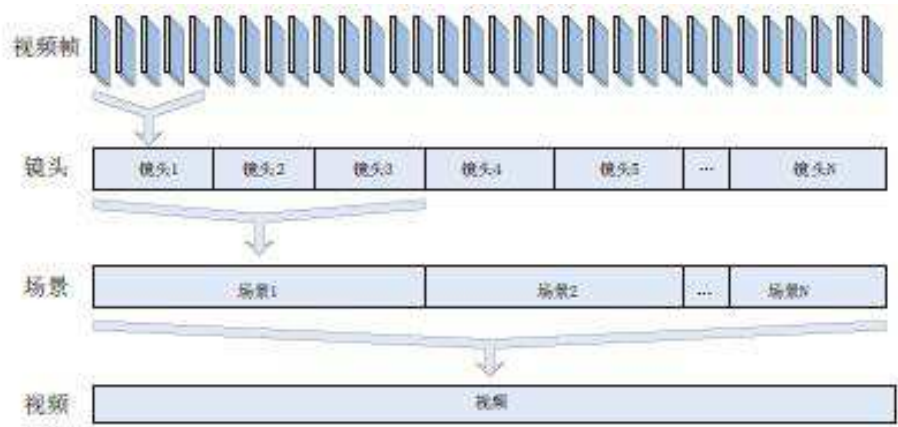


图 2-2 视频视觉通道信息基本结构

本文使用Jeong-Woo<sup>[49]</sup>等人提出的视频场景分割方法，不同于其他常规方法，在对视频帧提取特征时，不仅考虑了图像特征，还考虑了视频音频特征和文字特征，这些特征能更好的反映视频语义，比较符合本文对视频语义主题边界的需要。如图2-3所示，算法首先对视频原始的视频帧序列进行镜头分割以去掉太多重复的帧，然后对过滤后的视频帧进行特征提取，特征包含视觉特征，听觉特征，文字特征3部分。接着根据这些特征进行谱聚类得到初步结果，但是这个结果并不能保证每个类的视频帧在时间轴上是连续的，这不符合场景分割的结果。因此再把每个类的视频帧序列分成在时间轴上连续若干子类，然后对这些子类再进行k-means聚类，选取类中心作为最终场景的片段。

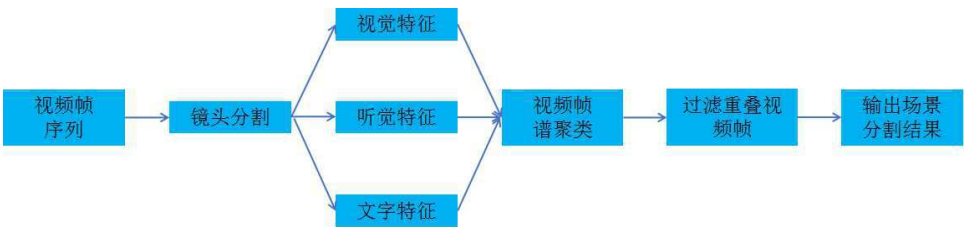


图 2-3 视频场景分割流程图

不同场景一般对应着不同的视频语义主题，但是有些视频场景比较单一，例如有些演讲类视频，它的场景可能就只涉及一个演讲的地点，有些体育类视

频，它的场景可能就一直是某个篮球场或者足球场，即同一场景也可能包含不同的视频语义主题，所以如果仅用场景边界作为视频主题边界，往往会造成视频主题分割不足的问题。

## 2.3 听觉通道的主题边界检测算法

视频听觉通道信息即视频的音频信息，在2.1.2小节本文用微软提供的语音识别工具包将视频音频转换成易于处理的自然语言文本，本小节介绍基于TopicTiling算法的文本主题分割算法。

### 2.3.1 视频语料库主题模型训练

本文的视频数据来源于YouTube<sup>①</sup>网站，利用youtube-dl<sup>②</sup>爬虫工具爬取了188841个视频，利用语音识别技术将视频语音信息转换成自然语言文本构成了本文视频语料库。本节介绍利用LDA<sup>[27]</sup>（Latent Dirichlet Allocation）训练语料库，挖掘视频内容潜在的主题模型。

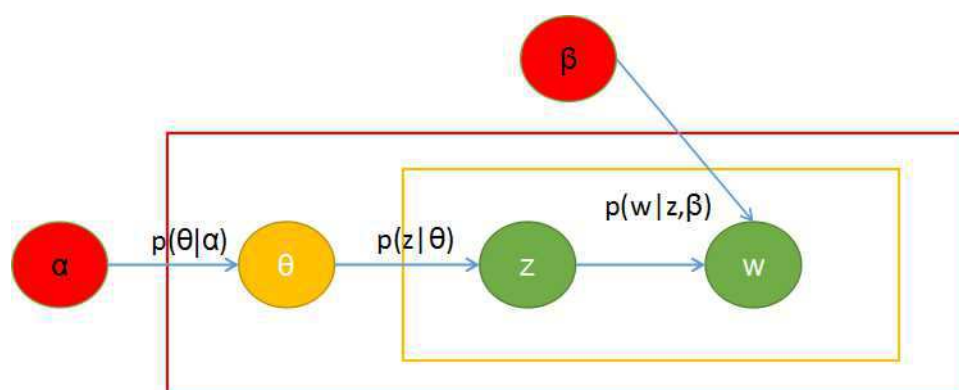


图 2-4 LDA三层模型

LDA是一种无监督机器学习技术，它基于贝叶斯概率模型，包含词，主题，文档三层结构。如图2-4所示，LDA认为每一篇文档 $\theta$ 的每个单词 $w$ 的生成过程都是先以一定概率 $p(z|\theta)$ 选择了某个主题，每个主题下面包含若干个和主题以一定概率的单词，然后在主题的单词表下以一定概率 $p(w|z, \beta)$ 生成了这个单词。图中红色标识的部分是语料库表示层， $\alpha, \beta$ 表示语料库级别的参数，每个文档参数都一样；图中黄色标识的部分是文档表示层， $\theta$ 表示文档别的变量，每个文档对应一个 $\theta$ ；图中绿色标识的是单词表示层，主题 $z$ 有 $\theta$ 生成， $w$ 由 $z$ 和 $\beta$ 共同生成，一个

<sup>①</sup> YouTube url: <https://www.youtube.com/>

<sup>②</sup> youtube-dl url: <http://rg3.github.io/youtube-dl/>



单词 $w$ 对应一个主题 $z$ 。由图2-4知, LDA联合概率公式为:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^N p(z_n|\theta)p(w_n|z_n, \beta) \quad (2.1)$$

LDA模型主要是从给定的语料中学习训练两个控制参数 $\alpha$ 和 $\beta$ ,  $\alpha$ 是Dirichlet分布的参数, 用于生成主题 $\theta$ 向量,  $\beta$ 表示各个主题对应的单词概率分布矩阵 $p(w|z)$ , 确定了 $\alpha$ 和 $\beta$ 两个全局控制参数, 就确定了符合语料库的LDA模型。训练主要思想主要是通过EM算法<sup>①</sup>把 $w$ 当做观察变量,  $\theta$ 和 $z$ 当做隐藏变量不断迭代直到收敛。得出 $\alpha$ 和 $\beta$ 。

本文采用JGibbLDA<sup>②</sup>做LDA主题模型实现, 它是用Gibbs采样做参数估计的LDA算法java版本的实现。通过Griffiths<sup>[50]</sup>提出的交叉验证方法得出当主题个数为40时, 各主题相似度最小, 算法迭代了1500次, 得到LDA训练结果, 由于篇幅原因, 只给出部分结果如表2-2所示:

### 2.3.2 TopicTiling算法主题边界检测

TopicTiling<sup>[26]</sup>算法基于TextTiling<sup>[22]</sup>算法, TextTiling算法使用词袋模型(bag of word vectors)作为文本块的特征, 由于文档的词典维度很大, 通常有几千维甚至上万维, 而局部的一个文本块包含的单词数量往往只包含几十到一百多的单词, 所以文本块的词向量往往很稀疏, 很多维度的值为0, 这样不但耗费机器资源, 而且不利于反映文本的语义。TopicTiling算法主要改进了文本块特征提取的方法, 不再使用bag of word vectors作为文本块的特征, 转而使用bag of TopicId vectors, 具体来说就是把文本块中每一个词映射到和这个词对应概率最大的主题Id, 在上一小节, 我们已经对视频语料库做了LDA训练, 挖掘出40个潜在的主题, 分别对应0-39个TopicId, 对于每一个视频的音频文本的单词都对应一个概率最大的TopicId, 表示这个单词和这个主题最有可能相关。bag of TopicId有两个明显的优点: 第一, 相较于bag of word特征, 它的维度大大减少, 就本文而言一共有40个TopicId, 所以特征向量的维度只有40维; 第二, bag of TopicId特征考虑了文字背后的语义关联, 例如“iphoneX is selling well in China”, “Tim Cook visits Foxconn in TaiWan”两个句子虽然几乎没有单词相同, 但句子背后的语义却非常相关。两个句子的bag of word特征截然不同, 而它们的bag of TopicId特征却非常接近, 显然后者更能挖掘出文本潜在的语义主题。

① 百度百科em算法: <https://baike.baidu.com/item/em算法>

② JGibbLDA url: <http://jgibblida.sourceforge.net/>

表 2-2 视频语料库LDA训练结果

Topic Id	Topic Word
0	roman,greek,temple,emperor,ancient,greeks,greece,forum...
1	protein,gene,population,genetic,melody,evolution,transcription,fitness...
2	india,afghanistan,indian,pakistan,u.s.,american,national,world...
3	water,species,ocean,population,fish,sea,animal,earth...
4	disease,cancer,health,blood,drug,medicine,hospital,doctor...
5	dante,poem,poet,plant,philosophy,tree,leaves,seed...
6	space,time,data,mission,life,years,day,horizons...
7	africa,u.s.,cuban,latin,brazil,international,global,world...
8	universe,energy,mass,density,gravity,vacuum,radiation,galaxy...
...	...
31	language,english,spanish,literature,author,french,fiction,grammar...
32	climate,oil,carbon,fossil,electricity,environmental,greenhouse,pollution...
33	probability,model,linear,matrix,vector,random,conditional,equation...
34	god,jewish,christian,faith,church,gospel,wisdom,temple...
35	carbon,hydrogen,electron,energy,acid,oxygen,atoms,chemical...
36	dog,cat,horse,feet,teeth,animal,good,eye...
37	cell,blood,eye,brain,bone,neurons,stem,muscle,skin...
38	film,theater,camera,dance,director,television,artist,star...
39	internet,technology,google,digital,facebook,social,youtube,computer...

TopicTiling算法以句子为初始块，提取每个文本块的bag of TopicId特征后，以余弦相似度度量相邻文本块的语义相似度。定义 $s(c)$ 表示当前文本块和其上下文的关联度，关联度计算如公式2.2所示：

$$s(c) = \frac{\sum_{t=1}^{40} w_{t,c} w_{t,p}}{\sqrt{\sum_{t=1}^{40} w_{t,c}^2} \sqrt{\sum_{t=1}^{40} w_{t,p}^2}} + \frac{\sum_{t=1}^{40} w_{t,c} w_{t,f}}{\sqrt{\sum_{t=1}^{40} w_{t,c}^2} \sqrt{\sum_{t=1}^{40} w_{t,f}^2}} \quad (2.2)$$

其中， $c$ 表示当前文本块， $p$ 表示和当前 $c$ 文本块相邻的前一个文本块， $f$ 表示和当前 $c$ 文本块相邻的后一个文本块， $w_{t,x}$ 表示 $x$ 文本块第 $t$ 维度bag of TopicId特征的值。公式2.2表明文本块与其上下文关联度 $s(c)$ 是当前文本块与其相邻前后两个文本块bag of TopicId特征余弦相似度之和。深度分数（depthscore）表示文本块两

侧的语义变化的剧烈程度，在其上下文两侧形成“深谷”，其值也就是“深谷”的深度和文本块与其两侧上下文关联度峰值的差值成正比，如图2-5所示，d4的深度分数为 $1/2*((d2-d2)+(d6-d4))$ ，即红色虚线长度之和的一半。公式2.3给出了

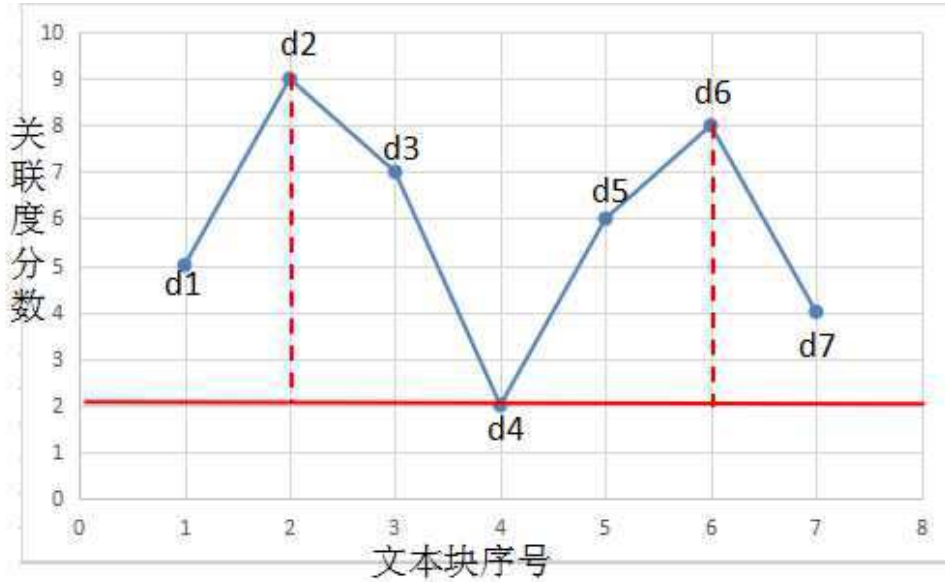


图 2-5 文本块深度分数

深度分数 $d_c$ 计算方法：

$$d_c = \frac{1}{2}(hl(c) - s(c) + hr(c) - s(c)) \quad (2.3)$$

其中 $hl(c)$ 表示从 $c$ 文本块左边找到的第一个关联度分数最高的峰值，右边 $hr(c)$ 同理。深度分数越高表明文本块关联度变化的趋势越剧烈，越有可能是主题边界。用公式2.3算出每一个文本块的深度分数并排序，然后计算深度分数的平均值 $s$ 及标准差 $\sigma$ ，以 $s-\sigma/2$ 为阈值，深度分数大于该阈值的文本块即为主题边界。本文将TopicTiling算法应用在视频主题边界分割的问题上，将视频语音转文本的结果作为TopicTiling算法的输入，得到视频主题边界的位置，算法1给出了基于TopicTiling的视频主题边界检测算法。

TopicTiling算法在确定深度分数阈值时是基于统计学的方法，即使用 $s-\sigma/2$ 为阈值来确定主题边界，其中 $s$ 和 $\sigma$ 分别为深度分数的平均值和标准差。深度分数阈值的而确定直接影响主题边界检测结果的质量，如果阈值过大，那么文本会分割不足，需要进一步分割；而如果阈值过小，那么文本会被过度分割成很多细碎的小片段，需要进行合并。由于视频内容的多样性和结构的复杂性，使用TopicTiling算法中确定阈值的方法很难确定合适的阈值，常常造成分割不足或

**算法 1** 基于TopicTiling的视频主题边界检测算法**输入:** 视频语音转换后得出的自然语言文本text**输出:** 视频主题边界的位置

---

```

1: function Detect_Video_Topic_Boundary(text)
2:   //对文本text进行分句得到blocks[]数组, 每个句子初始化为一个文本块
3:   blocks[]  $\leftarrow$  initialize(text)
4:   初始化features[][40]数组, 用来记录文本块bag of TopicId特征
5:   for block in quadblocks[] do
6:     //对文本块进行提干
7:     stemmer(block);
8:     //对文本块去停词
9:     porter(block);
10:    //对文本块提取bag of TopicId特征
11:    features[][40]  $\leftarrow$  extractFeatures(block)
12:  end for
13:  初始化contextScores[]数组, 记录文本块上下文关联度分数
14:  for feature in features do
15:    //计算文本块上下文关联度分数
16:    contextScores[]  $\leftarrow$  computeContextScore(feature)
17:  end for
18:  初始化depthScore[]数组, 记录文本块深度分数
19:  for score in contextScores do
20:    //计算文本块深度分数
21:    depthScore  $\leftarrow$  computeDepthScore(score, contextScores[])
22:  end for
23:  //对深度分数进行排序
24:  sortdepthScore[]
25:  //根据公式2.4确定深度分数阈值
26:  threshold  $\leftarrow$  determineThreshold()
27:  初始化results[], 记录视频主题边界位置
28:  for ds in detphScore do
29:    if ds > s -  $\sigma/2$  then results[]  $\leftarrow$  position(ds, depthScore[])
30:    end if
31:  end for
32:  return results
33: end function

```

---

者过度分割。

## 2.4 基于双通道线索融合的主题边界检测

视频数据由视觉通道线索信息和听觉通道线索信息组成,本章2.2节基于视频视觉线索对视频进行了视频场景分割,由视频场景边界位置推测视频主题边界位置,该方法的问题在于常常造成分割不足的问题;本章2.3节基于视频听觉线索先将视频语音转换为易于处理的自然语言文本,然后针对处理后的文本信息对视频进行主题分割,该方法忽略了明显的视频视觉边界,而且其深度分数的阈值很难确定,影响主题分割的质量。本节综合以上两种方法,提出了基于视频双通道线索融合的主题边界检测方法。

### 2.4.1 确定视频主题的个数

在检测视频主题边界位置之前,首要的工作是确定视频主题的个数。本文在确定视频主题个数考虑了三个方面:第一,本文在2.2节仅利用视频视觉通道特征视频进行了主题分割,该方法的原理是视频的场景分割,如上文所述,该方法存在分割不足的问题,所以算法检测的视频主题个数往往会比真实主题个数少,即如果视频有 $n$ 个场景,那么实际上视频有 $\alpha * n$ 个主题边界,其中 $\alpha > 1$ ,本文经大量研究总结出经验参数 $\alpha$ 为1.25。第二,视频主题边界个数与视频时长有关,显然越长的视频能讲述越多的内容,其所包含的主题也就越多,一般来说,绝大多数视频一个故事单元的时长在3分钟到5分钟不等,取平均值4分钟,用视频时长除以每个故事单元的平均时长可以估算出视频的主题个数;第三,TopicTiling算法从统计学的角度确定深度分数的阈值从而确定了视频的主题个数,这个方法考虑了不同视频的差异性。综合以上三个因素,本文给出了深度分数阈值的公式2.4:

$$threshold = \alpha * 1.25n + \beta * \frac{t}{st} + \theta * count(depthScore[i]) \text{ if } depthScore[i] > threshold \quad (2.4)$$

其中 $n$ 为视频场景的个数, $t$ 为视频总时长, $st$ 为视频一个故事单元的平均时长240秒, $\alpha, \beta, \theta$ 是3个经验参数,就本文而言,它们的值分别为0.80, 0.05, 0.15。

### 2.4.2 双通道线索融合的主题边界检测

本文提出了基于双通道线索融合的视频主题检测方法,在听觉通道线索上,

我们在本文2.3节的基础上对TopicTiling算法在确定深度分数阈值上做出了改进，不在从统计学的角度以深度分数的平均值-1/2方差为阈值，而改用公式2.4的阈值计算方法，而且为了后面双通道线索边界融合的需要，我们增加了深度分数的阈值，使得边界的个数是2.4.1小节确定视频主题个数的1.5倍，从而获得了根据听觉通道线索得出的视频主题边界，这些边界为视频语音文本的位置。在视觉通道线索上，本文根据视频场景切换的位置推测视频主题边界的位置，获得根据视频场景分割得出的视频主题边界，这些边界为视频一系列视频帧的位置。视频听觉通道和视觉通道通过时间轴关联，本文给出的视频主题边界为视频一系列的时间点，所以在融合双通道线索主体边界之前首先将上文所述听觉通道上若干个文本边界位置和在视觉通道上的若干个视频帧边界位置都分别映射到它们视频时间点位置。

获得双通道线索独立的主题边界后，本文接下来阐述如何将根据双通道线索获得的主题边界融合成最优的边界位置。融合过程主要考虑以下两条原则：第一，通过视觉通道线索获得的视频主题边界比通过听觉通道线索获得的边界更可靠，虽然视频一个场景可能包含多个视频语义主题，但是明显地，不同视频场景的切换一定对应着视频语义主题的切换，所以本文优先考虑根据视频视觉通道线索检测到的视频主题边界位置。第二，本文认为最有可能是视频主题边界的位置是根据双通道线索都能检测到的位置，即如果根据视频双通道线索检测到的视频主题边界时间点重合或者说非常接近，那么此时间点极有可能是真正的视频主题边界。下面详述融合过程，通过视频视觉通道线索获得的 $a$ 个视频主题边界位置设为 $\{x_1, x_2, \dots, x_a\}$ ，通过视频听觉通道线索获得的 $b$ 个视频主题边界位置设为 $\{y_1, y_2, \dots, y_b\}$ ，本章2.4.1节确定了视频主题的个数 $n$ ，由公式2.4可知， $n$ 一定大于 $a$ 。接下来的步骤是一个贪婪选择的过程，对于每一个视觉边界 $x_i$ 在听觉边界列表 $\{y_1, y_2, \dots, y_b\}$ 找到一个“最相关”的位置 $y_j$ ，这里“最相关”的定义为以 $x_i$ 为原点以 $(-60s, 60s)$ 为领域找到 $y_j$ ，使得 $\|x_i - y_j\|$ 最小，如果在 $(-60s, 60s)$ 范围内找不到 $y_j$ ，那么令 $y_j$ 为 $x_i$ 本身。贪婪选择使得每一个视觉主题边界在听觉主题边界列表找到一个映射，如图2-6所示，黑色线条代表时间轴，

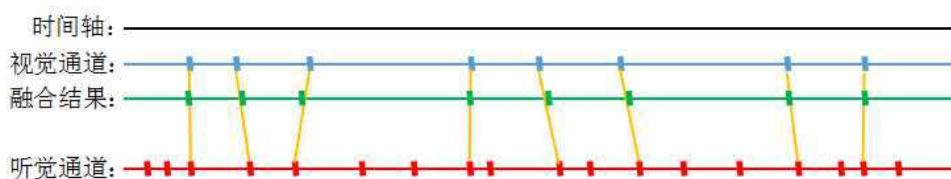


图 2-6 视觉通道边界和听觉通道边界融合



它是视觉通道和听觉通道的联系，蓝色线条代表视觉通道，其上的点代表我们在视觉通道所检测到的视频主题边界，类似地，红色线条代表听觉通道，其上的点代表我们在听觉通道所检测到的视频主题边界。在确定视觉通道边界与听觉通道边界的映射后，公式2.5给出了基于双通道线索融合的最终位置 $p_k$ 的计算方法， $p_k$ 位置介于 $x_i$ 和 $y_j$ 位置之间，其中 $1 \leq k \leq a$ ， $\lambda$ 为权值，它决定了 $p_k$ 的位置与 $x_i$ 更近还是与 $y_j$ 更近，本文认为视觉通道的边界更加可靠，取 $\lambda$ 为 $\frac{1}{3}$ 。

$$p_k = \begin{cases} y_j + \lambda * \|x_i - y_j\|, & \text{if } x_i \geq y_j \\ y_j - \lambda * \|x_i - y_j\|, & \text{if } x_i < y_j \end{cases} \quad (2.5)$$

在2.4.1本文确定了视频的主题个数 $n$ ，上文在通过融合视觉通道线索和听觉通道线索得出了 $a$ （ $a$ 为视觉通道边界的个数）个最终主题边界后，我们还需要检测出剩余的 $n-a$ 个边界，这 $n-a$ 个边界来自于听觉通道边界且没有映射到视觉通道边界的位置，如图2-7所示黑色点所在位置。我们将听觉通道上的主题边界中没有与视觉通道主题边界形成对应关系位置单独放到一个集合中，并且对这些边界的深度分数排序，选取 $n-a$ 个最大深度分数的位置（图中绿色线条上黑色的点）作为视频主题边界，这 $n-a$ 个位置和上文所述的 $a$ 个位置（图中绿色的点）组成了本文最终的 $n$ 个视频主题边界。

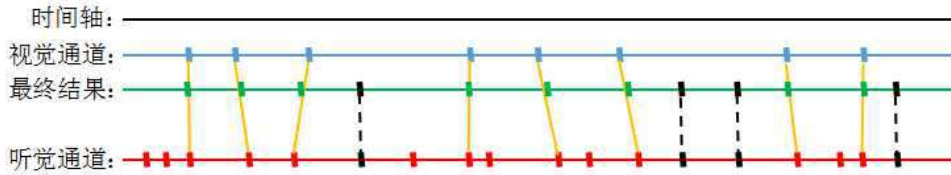


图 2-7 视频主题边界最终结果

### 2.4.3 实验结果及评价

## 2.5 本章小结

本章在2.1小节首先介绍视频视觉，听觉双通道信息的预处理方法，对于视觉通道，主要进行镜头分割，关键帧提取，OCR识别；对于听觉通道，主要进行语音识别，并对语音识别的自然言语文本进行去停词，提干，关键词提取等处理。2.2小节介绍了根据视频视觉通道线索检测视频主题边界的过程。2.3小节介绍了根据视频语音转换而来的文本语料库训练过程，及基于TopicTiling算法的

视频主题边界检测算法。2.4小节介绍了视频双通道主体边界的融合过程，得出了最终的检测结果，并且对实验结果做对比，证明了本文方法的准确性和有效性。



## 第 3 章 基于查询自适应的缩略图自动生成方法

### 3.1 视频内容与用户查询相关性度量

### 3.2 缩略图图文内容提取

### 3.3 缩略图排版方法

### 3.4 实验结果及评估

此处省略N个字……

### 3.5 本章小结

此处省略N个字……

## 第4章 基于查询自适应的视频缩略图生成方法应用

内容概括。

### 4.1 视频缩略图生成系统应用设计

此处省略N个字……

### 4.2 视频缩略图生成系统应用实现

此处省略N个字……

### 4.3 视频缩略图生成系统应用评估

此处省略N个字……

### 4.4 本章小结

此处省略N个字……

## 第 5 章 总结与展望

内容概括。

### 5.1 本文工作总结

此处省略N个字……

### 5.2 今后工作展望

此处省略N个字……

## 参考文献

- [1] Gao Y, Zhang T, Xiao J. Thematic video thumbnail selection[C]//Image Processing (ICIP), 2009 16th IEEE International Conference on. IEEE, 2009: 4333-4336.
- [2] Christel M G. Evaluation and user studies with respect to video summarization and browsing[C]. SPIE, 2006.
- [3] Song Y, Redi M, Vallmitjana J, et al. To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016: 659-668.
- [4] Nagasaka A, Tanaka Y. Automatic Video Indexing and Full-Video Search for Object Appearances.[C]//Visual Database System II, Elsevier. Dallas, United States: ACM, 1991:113 - 127.
- [5] Swain M, Ballard D. Color indexing[J]. International Journal of Computer Vision, 1991,7(1):11 - 32.
- [6] Zhang H, Kankanhalli A, Smoliar S. Automatic partitioning of full-motion video[J]. Multimedia Systems, 1993, 1(1):10 - 28.
- [7] Boutheymy P, Gelgon M, Ganansia F. A unified approach to shot change detection and camera motion characterization[J]. IEEE Transactions on Circuits and Systems for Video Technology, 1999, 9(7):1030 - 1044.
- [8] Courtney J. Automatic video indexing via object motion analysis[J]. Pattern Recognition, 1997, 30(4):607 - 625.
- [9] Zabih R, Miller J, Mai K. A feature-based algorithm for detecting and classifying production effects[J]. Multimedia systems, 1999, 7(2): 119-128.
- [10] Divakaran A, Sun H. Descriptor for spatial distribution of motion activity for compressed video[J]. 2000, 2:392 - 398.
- [11] Yeung M M, Liu B. Efficient matching and clustering of video shots[C]//Image Processing, 1995. Proceedings., International Conference on. IEEE, 1995, 1: 338-341.
- [12] Zhang H J, Wu J, Zhong D, et al. An integrated system for content-based video retrieval and browsing[J]. Pattern Recognition, 1997, 30(4):643 - 658.
- [13] 董晨晨. 镜头边界检测与关键帧提取技术研究[D]. 南京: 东南大学, 2010.
- [14] 朱曦, 林行刚. 视频镜头时域分割方法的研究[J]. 计算机学报, 2004, 27(8):1027 - 1035.
- [15] Jeannin S, Jasinschi R, She A, et al. Motion descriptors for content-based video representation[J]. Signal Processing Image Communication, 2000, 16(1-2):59 - 85.
- [16] Ciresan D C, Meier U, Gambardella L M, et al. Convolutional neural network committees for handwritten character classification[C]//Document Analysis and Recognition (ICDAR), 2011 International Conference on. IEEE, 2011: 1135-1139.
- [17] Neumann L, Matas J. Scene text localization and recognition with oriented stroke detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Sydney, Australia:IEEE, 2013:97 - 104.
- [18] 赵亚琴. 基于内容的视频片段检索技术研究[D]. 南京:南京理工大学, 2007.
- [19] 冯哲. 基于内容的视频检索中的音频处理[D]. 上海:复旦大学, 2004.
- [20] 闫乐林. 基于视听信息的视频语义分析与检索技术研究[D]. 北京:北京邮电大学, 2012.

- [21] Walker W, Lamere P, Kwok P, et al. Sphinx-4: a flexible open source framework for speech recognition[J]. Sun Microsystems, 2004:1 - 18.
- [22] Hearst M A. TextTiling: Segmenting text into multi-paragraph subtopic passages[J]. Computational linguistics, 1997, 23(1): 33-64.
- [23] Hearst M A. Multi-paragraph segmentation of expository text[C]//Proceedings of the 32nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 1994: 9-16.
- [24] Banerjee S, Rudnicky A I. A TextTiling based approach to topic boundary detection in meetings[J]. 2006.
- [25] Galley M, McKeown K, Fosler-Lussier E, et al. Discourse segmentation of multi-party conversation[C]//Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1. Association for Computational Linguistics, 2003: 562-569.
- [26] Riedl M, Biemann C. TopicTiling: a text segmentation algorithm based on LDA[C]//Proceedings of ACL 2012 Student Research Workshop. Association for Computational Linguistics, 2012: 37-42.
- [27] Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
- [28] Nguyen V A, Boyd-Graber J, Resnik P. SITS: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations[C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012: 78-87.
- [29] Chen H, Xie L, Feng W, et al. Topic segmentation on spoken documents using self-validated acoustic cuts[J]. Soft Computing, 2015, 19(1): 47-59.
- [30] Fragkou P. Text Segmentation using Named Entity Recognition and Co-reference Resolution in English and Greek Texts[J]. arXiv preprint arXiv:1610.09226, 2016.
- [31] Choi F Y Y. Advances in domain independent linear text segmentation[C]//Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. Association for Computational Linguistics, 2000: 26-33.
- [32] M. Utiyama and H. Isahara. A statistical model for domain independent text segmentation. In Proceedings of the 9th EACL, pages 491 - 498, 2001.
- [33] R. Kern and M. Granitzer. Efficient linear text segmentation based on information retrieval techniques. In Proceeding of the International Conference on Management of Emergent Digital EcoSystems, 2009.
- [34] W. Wolf, "Key frame selection by motion analysis," in IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 2. IEEE, 1996, pp. 1228 - 1231.
- [35] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Trans. on Multimedia, vol. 7, no. 5, pp. 907 - 919, 2005.
- [36] Y. Cong, J. Yuan, and J. Luo, "Towards scalable summarization of consumer videos via sparse dictionary selection," IEEE Trans. on Multimedia, vol. 14, no. 1, pp. 66 - 75, Feb 2012.
- [37] G. Guan, Z. Wang, S. Lu, J. D. Deng, and D. D. Feng, "Keypointbased keyframe selection," IEEE Trans. on Circuits and Systems for Video Technology, vol. 23, no. 4, pp. 729 - 734, 2013.

- [38] H.-W. Kang and X.-S. Hua, "To learn representativeness of video frames," in ACM International Conference on Multimedia, Singapore, November 2005, pp. 423 – 426.
- [39] Zhang B, Wang Z, Tao D, et al. Automatic Preview Frame Selection for Online Videos[C]//Digital Image Computing: Techniques and Applications (DICTA), 2015 International Conference on. IEEE, 2015: 1-6.
- [40] Song Y, Redi M, Vallmitjana J, et al. To Click or Not To Click: Automatic Selection of Beautiful Thumbnails from Videos[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. ACM, 2016: 659-668.
- [41] C. Liu, Q. Huang, and S. Jiang. Query sensitive dynamic web video thumbnail generation. In ICIP, pages 2449 – 2452, 2011.
- [42] D. Joshi, J. Wang, and J. Li, "The story picturing engine: finding elite images to illustrate a story using mutual reinforcement," ACM SIGMM workshop on MIR, 2004.
- [43] Liu W, Mei T, Zhang Y, et al. Multi-task deep visual-semantic embedding for video thumbnail selection[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3707-3715.
- [44] Apostolidis E, Mezaris V. Fast shot segmentation combining global and local visual descriptors[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 6583-6587.
- [45] Smith R. An overview of the Tesseract OCR engine[C]//Document Analysis and Recognition, 2007. ICDAR 2007. Ninth International Conference on. IEEE, 2007, 2: 629-633.
- [46] Smith R, Antonova D, Lee D S. Adapting the Tesseract open source OCR engine for multilingual OCR[C]//Proceedings of the International Workshop on Multilingual OCR. ACM, 2009: 1.
- [47] Hassan Z, Mohamad A R, Kalil M R, et al. Evaluation of Microsoft speech recognition in controlling robot soccer[C]//Pattern Analysis and Intelligent Robotics (ICPAIR), 2011 International Conference on. IEEE, 2011, 1: 114-118.
- [48] Sultana S, Akhand M A H, Das P K, et al. Bangla speech-to-text conversion using SAPI[C]//Computer and Communication Engineering (ICCCE), 2012 International Conference on. IEEE, 2012: 385-390.
- [49] Son J W, Lee S Y, Park S Y, et al. Video scene segmentation based on multiview shot representation[C]//Information and Communication Technology Convergence (ICTC), 2016 International Conference on. IEEE, 2016: 381-383.
- [50] Griffiths T L, Steyvers M. Finding scientific topics[J]. Proceedings of the National academy of Sciences, 2004, 101(suppl 1): 5228-5235.
- [51] Porter M F. An algorithm for suffix stripping[J]. Program, 1980, 14(3): 130-137.
- [52] Rose S, Engel D, Cramer N, et al. Automatic keyword extraction from individual documents[J]. Text Mining: Applications and Theory, 2010: 1-20.

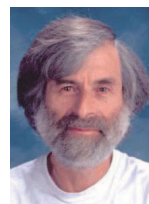
## 附录

## 作者简历

列举发表论文、专著、专利、标准及参加科研项目情况（版权所有必须归中山大学）

可以按如下模板：

1985 年，LaTeX 问世，它构筑在TeX 的基础之上，并且加进了很多新功能，使得用户可以更为方便地利用TeX 的强大功能。右图就是其开发者，美国著名计算机科学家、数学家Lamport 博士，他曾在康柏和惠普工作，目前任职于微软公司。



发表的论文：

- [1] Xin Chen, Hefeng Wu, Xiang Li, Xiaonan Luo, and Taisheng Qiu, Real-time visual object tracking via CamShift-based robust framework, International Journal of Fuzzy Systems, 14(3), 2012, 262 269
- [2] Xin Chen, Hefeng Wu, Xiang Li, Xiaonan Luo, and Taisheng Qiu, Real-time visual object tracking via CamShift-based robust framework, International Journal of Fuzzy Systems, 14(3), 2012, 262 269

专利/标准：

- [1] 陈欣，罗力耕，陈湘萍，一种基于数字电视中间件的视频点播方法及系统，中国发明专利，授权号：ZL201110130102.1，授权日期：2012.09.05
- [2] 陈欣，罗力耕，陈湘萍，一种基于数字电视中间件的视频点播方法及系统，中国发明专利，授权号：ZL201110130102.1，授权日期：2012.09.05

参与课题（注明项目编号）

- [1] 课题名称：广东联合基金项目《数字几何媒体智能技术及应用研究》  
项目编号：U0935004  
主要工作：参与视频目标检测与跟踪技术的理论研究，撰写并发表相关国际会议论文2篇
- [2] 课题名称：广东联合基金项目《数字几何媒体智能技术及应用研究》  
项目编号：U0935004  
主要工作：参与视频目标检测与跟踪技术的理论研究，撰写并发表相关国际会议论文2篇

获奖情况：



- [1] 2012年度中山大学研究生国家奖学金
- [2] 2013年度中山大学研究生国家奖学金

## 致谢

由衷感谢我的导师罗笑南教授，本文是在他的指导下完成的。

某某人

某年某月某日