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

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

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：　　导师签名：

日期：　年　月　日　　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
专 业： 软件工程
博 士 生： 戚鑫
指导教师： 罗笑南 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: Xin Qi

Supervisor: Prof. Xiaonan Luo

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 primate 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

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%[①]。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[②]、爱奇艺[③]、腾讯、搜狐等数百家视频门户网站，国外视频门户网站有youTube[④]、Yahoo Video[⑤]、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

① 数据来源前瞻产业研究院:https://bg.qianzhan.com/report/detail/459/170313-223982fb.html
② http://www.youku.com/
③ http://www.iqiyi.com/
④ https://www.youtube.com/
⑤ http://video.search.yahoo.com/

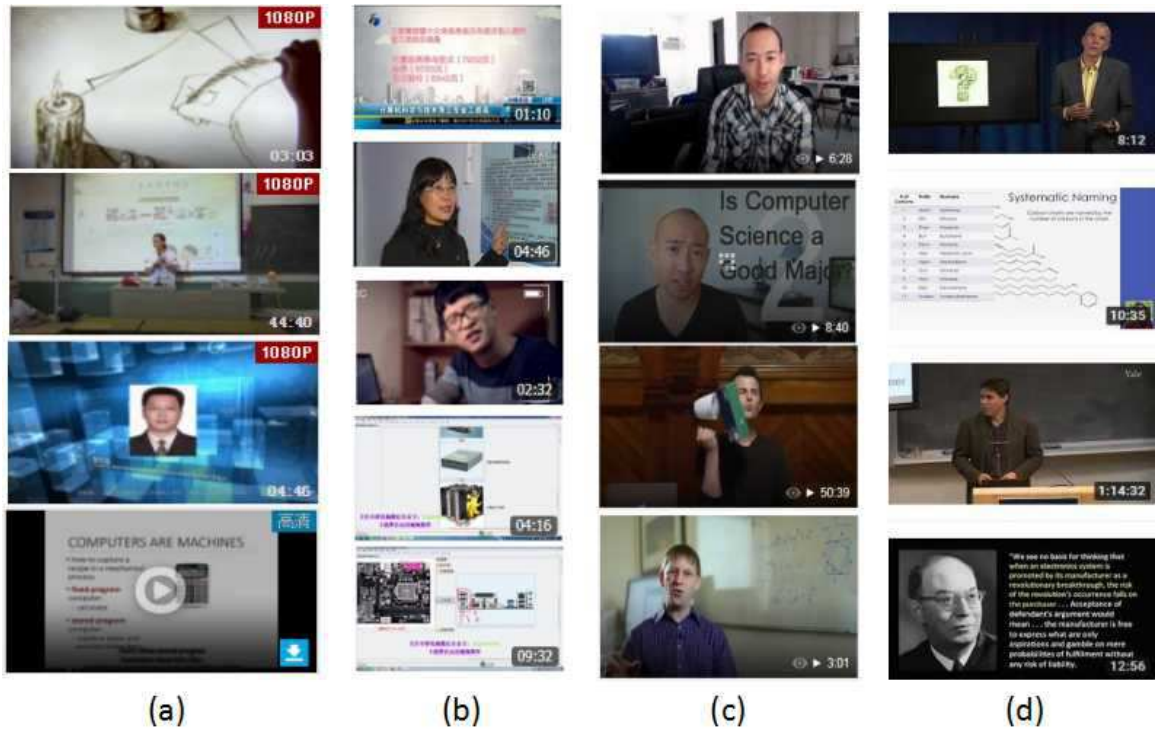


图 1-1 国内外知名视频网站以“计算机科学”/”computer science”为关键词检索的部分结果：(a)优酷视频.(b)爱奇艺视频.(c)YouTube Video.(d)Yahoo Video.

缩略图已成为一种常用的视频和用户交互的技术。Yuli[1]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[2]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[3]。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

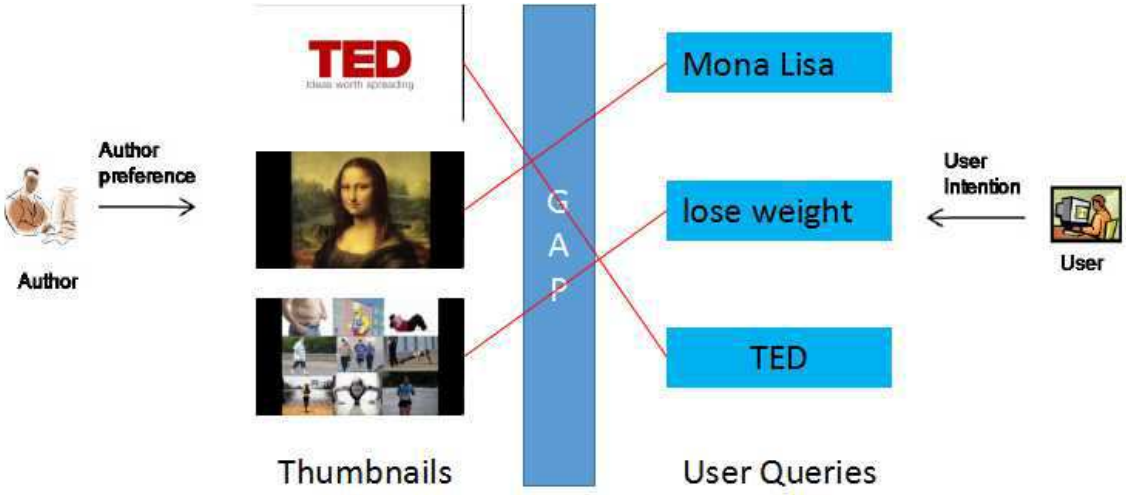


图 1-2 用户查询意图与视频缩略图之间存在的鸿沟

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OCR①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[4, 5]，基于像素颜色差异的镜头分割方法[6]，基于运动的镜头分割方法[7, 8]，基于边缘特征的镜头分割方法[9]。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[10]、直方图平均法[11, 12]、基于运动特征的关键帧提取方法[13, 14]、基于内容的分析方法[15]。

（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[16]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[17]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[18, 19, 20]。现在有一些高级的语音识别工具，英文的工具

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

有MSR①,CMU Sphinx[21],Nuance Dragon②等，中文的工具主要有科大讯飞的讯飞开放平台③。

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[22, 23]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[24]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[25]等人提出了一个基于TextTiling的算法，不同于TextTiling算法，它用词的tf-idf④权重代替了单一的词频权重从而取得了更好的效果。Martin[26]等人提出了TopicTiling算法，改进了Textiling算法。TopicTiling算法基于LDA[27]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TextTiling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

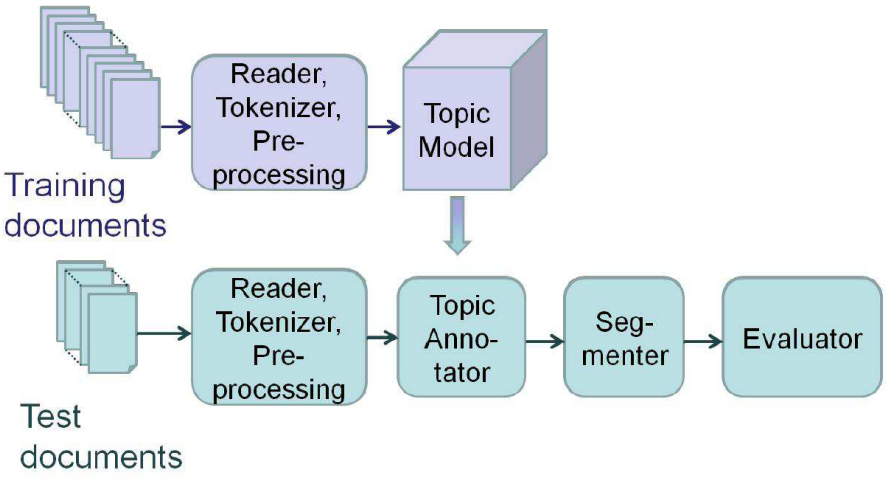


图 1-3 TopicTiling算法主题分割流程

面的章节会详细介绍。

除了无监督的方法，Nguyen[28]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[29]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[30]利用NER（Ebtity 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[31]提出的C99b算法，Utiyama[32]等提出的基于统计学模型的分割算法，Kehagias[33]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

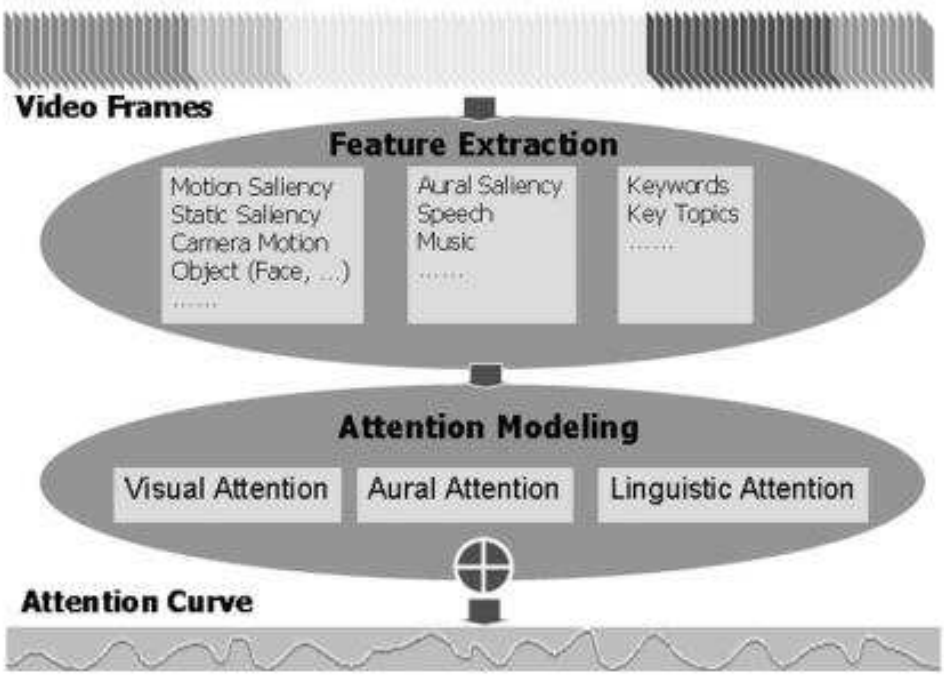


图 1-4 用户关注度模型框架

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[34]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[35]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[36]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[37]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[38]等人具体度量了视频帧的代表性，认为帧的质量，图像

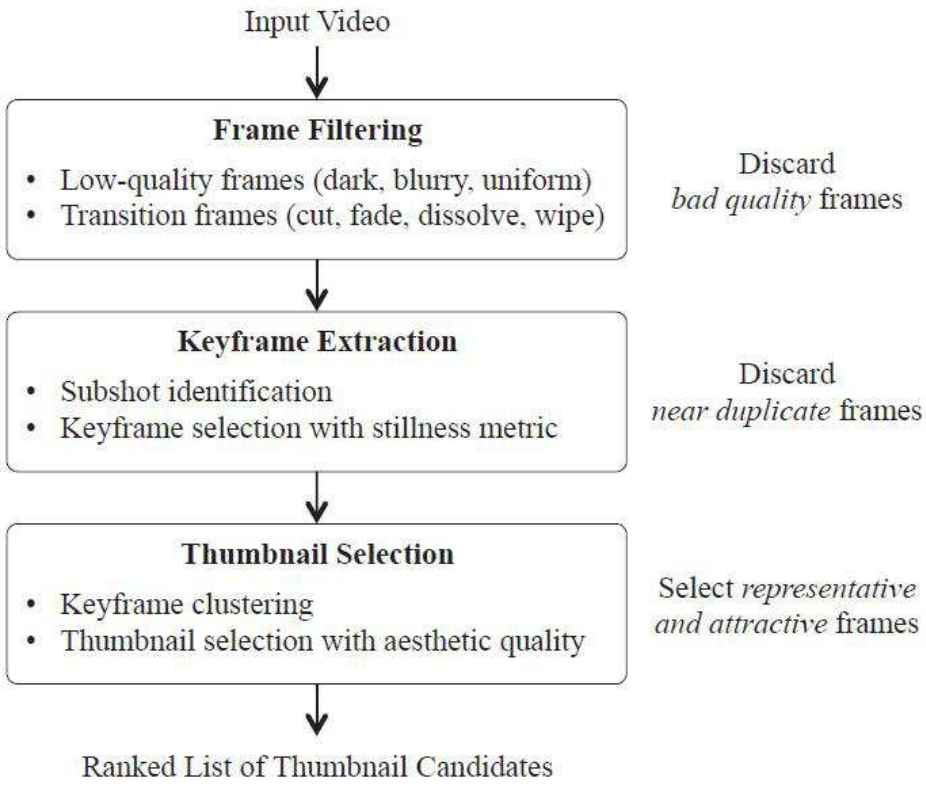


图 1-5 用户关注度模型框架

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[39]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[40]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[41]等人的研究改进了这一点，他们的方法首先使用Joshi[42]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,

$$S(I_u, w) = P(I_u, w) \approx P(\frac{I_u}{w}) = \frac{1}{n}\sum_{i\epsilon n} s(I_u, v_i) \tag{1.1}$$

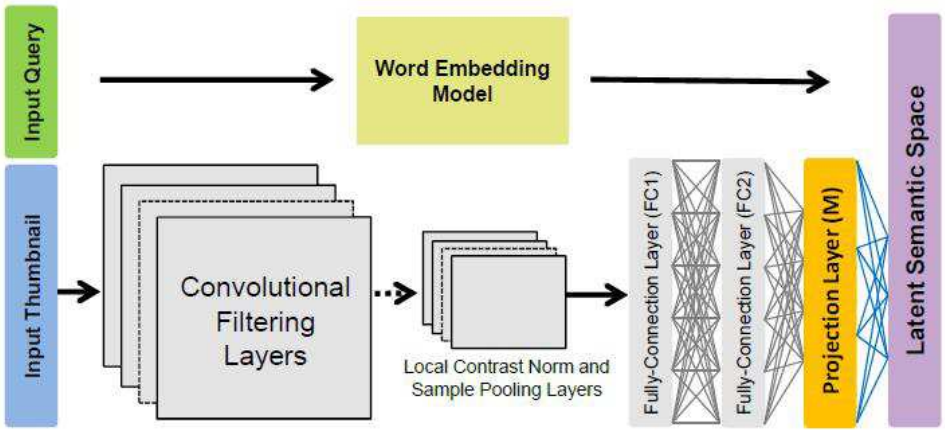


图 1-6 深度视觉-语义嵌入模型架构

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，$v_1,v_2,...,v_n$表示用查询词w早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[43]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[41]的方法具有不确定性而且非常耗时，文章[43]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，涉及数据挖掘，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

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[44]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OmniPage①,Readiris②，Tesseract等等。本文采用Matlab 2015b集成的Tesseract OCR引擎[45, 46],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 Sphinx③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 Dragon④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⑤。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[47, 48]。

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

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）①和语音识别输出的冗余的时间戳信息，接着使用Porter算法[51]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”。

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。

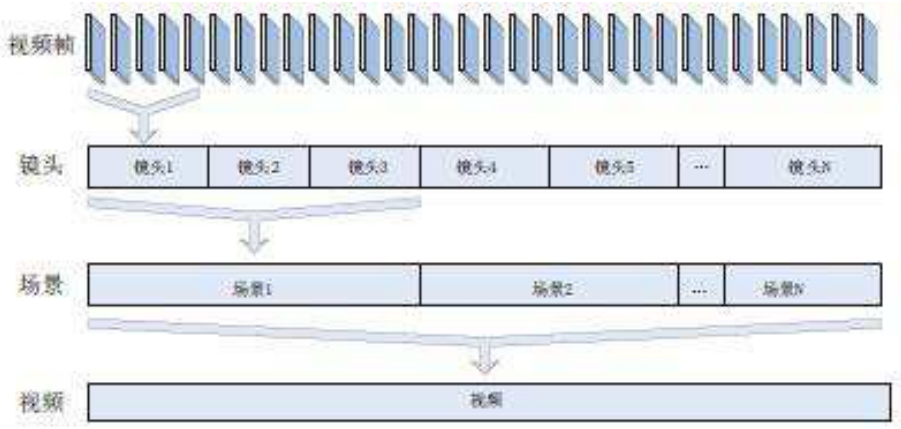


图 2-2　视频视觉通道信息基本结构

本文使用Jeong-Woo[49]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

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

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。

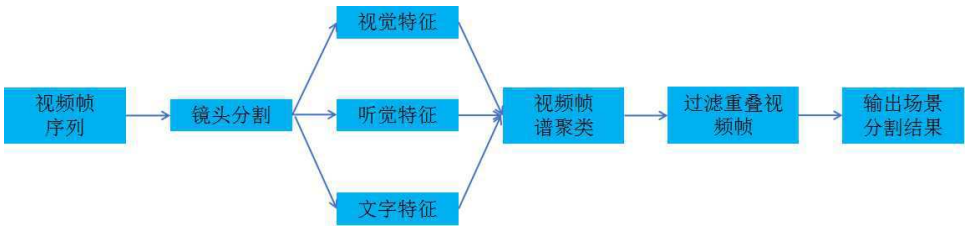


图 2-3　视频场景分割流程图

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YouTube①网站，利用youtube-dl②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[27]（Latent Dirichlet Allocation）训练语料库，挖掘视频内容潜在的主题模型。

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θ的每个单词w的生成过程都

① YouTube url：https://www.youtube.com/
② youtube-dl url:http://rg3.github.io/youtube-dl/

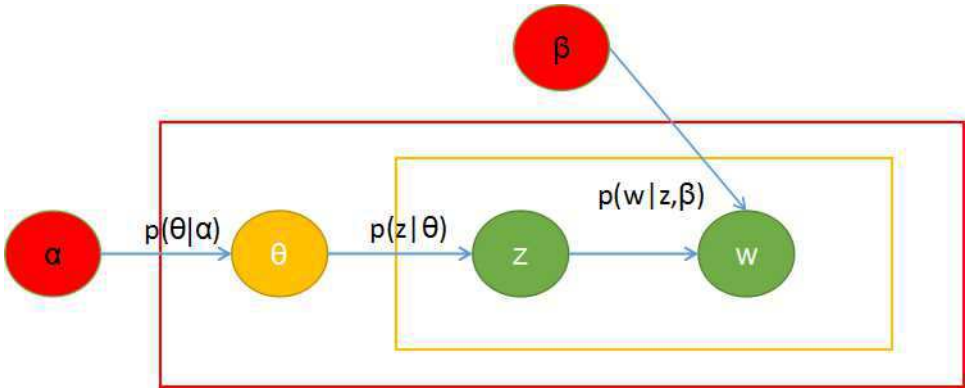


图 2-4　LDA三层模型

是先以一定概率p(z|θ)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,β)生成了这个单词。图中红色标识的部分是语料库表示层，α,β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，θ表示文档别的变量，每个文档对应一个θ;图中绿色标识的是单词表示层，主题z有θ生成，w由z和β共同生成，一个单词w对应一个主题z。由图2-4知，LDA联合概率公式为：

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

LDA模型主要是从给定的语料中学习训练两个控制参数α和β,α是Dirichlet分布的参数，用于生成主题θ向量，β表示各个主题对应的单词概率分布矩阵p(w|z),确定了α和β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①把w当做观察变量，θ和z当做隐藏变量不断迭代直到收敛。得出α和β。

本文采用JGibbLDA②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[50]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[26]算法基于TextTiling[22]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

① 百度百科em算法：https://baike.baidu.com/item/em算法
② JGibbLDA url:http://jgibblda.sourceforge.net/

表 2-1　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...

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 verctors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。

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：

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

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

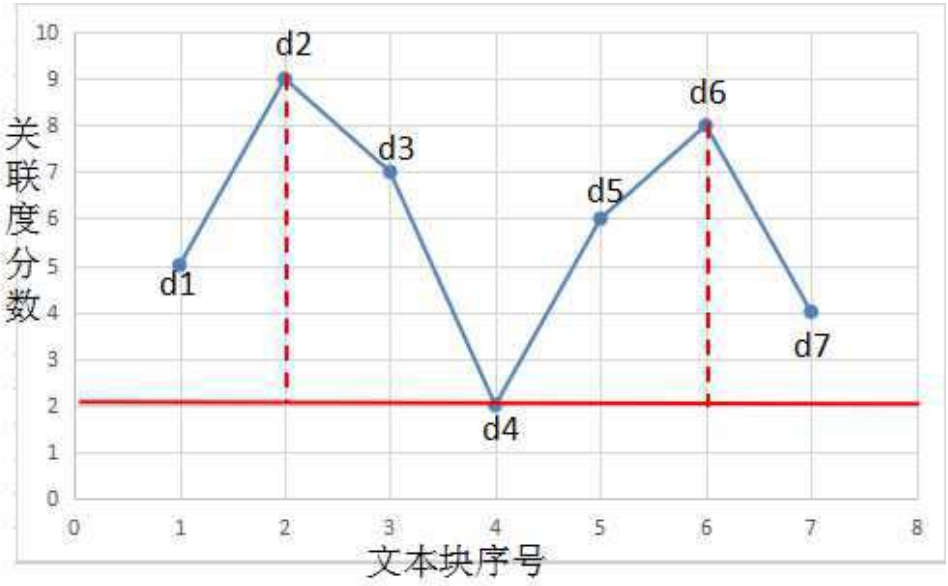


图 2-5　文本块深度分数

深度分数d_c计算方法：

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

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σ,以s-σ/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-σ/2为阈值来确定主题边界，其中s和σ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

算法 1 基于TopicTiiling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

function $Detect\_Video\_Topic\_Boundary$(text)
　//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
　$blocks[\,] \leftarrow initialize(text)$
　初始化$features[\,][40]$数组，用来记录文本块bag of TopicId特征
　for $block\ in\ quadblocks[\,]$ **do**
　　//对文本块进行提干
　　$stemmer(block)$;
　　//对文本块去停词
　　$porter(block)$;
　　//对文本块提取bag of TopicId特征
　　$features[\,][40] \leftarrow extractFeatures(block)$
　end for
　初始化$contextScores[\,]$数组，记录文本块上下文关联度分数
　for $feature\ in\ features$ **do**
　　//计算文本块上下文关联度分数
　　$contextScores[\,] \leftarrow computeContextScore(feature)$
　end for
　初始化$depthScore[\,]$数组，记录文本块深度分数
　for $score\ in\ contextScores$ **do**
　　//计算文本块深度分数
　　$depthScore \leftarrow computeDepthScore(score, contextScores[\,])$
　end for
　//对深度分数进行排序
　$sortdetphScore[\,]$
　//根据公式2.4确定深度分数阈值
　$threshold \leftarrow determineThreshold()$
　初始化$results[\,]$,记录视频主题边界位置
　for $ds\ in\ detphScore$ **do**
　　if $ds > s - \sigma/2$ **then** $results[\,] \leftarrow position(ds, depthScore[\,])$
　　end if
　end for
　return $results$
end function

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α>1,本文经大量研究总结出经验参数α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：

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

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，α,β,θ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$\{x_1, x_2, ..., x_a\}$,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$\{y_1, y_2, ..., y_b\}$，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$\{y_1, y_2, ..., y_b\}$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，

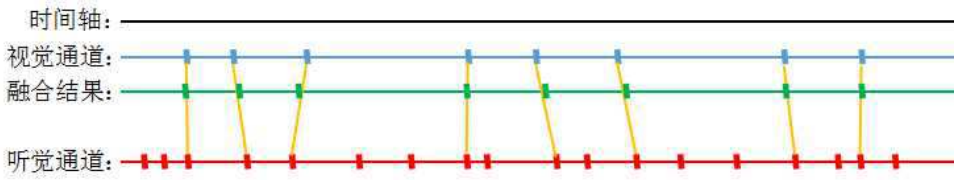


图 2-6　视觉通道边界和听觉通道边界融合

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$1 \le k \le a$,λ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λ为$\frac{1}{3}$。

$$p_k = \begin{cases} y_j + \lambda * \|x_i - y_j\|, & if\ x_i \ge y_j \\ y_j - \lambda * \|x_i - y_j\|, & if\ x_i < y_j \end{cases} \tag{2.5}$$

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。

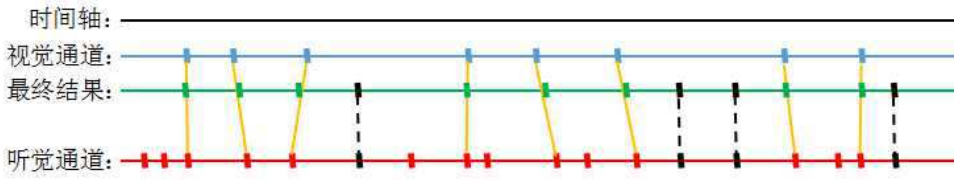


图 2-7　视频主题边界最终结果

2.5　实验结果与分析

本节评估基于双通道信息融合的视频主题边界检测算法的性能，首先介绍实验流程及评价实验结果的性能指标，接着展示相关实验结果，最后是对实验结果的分析。

2.5.1　实验流程及性能指标

本文选择了youtube网站中10个视频作为实验数据，视频时长在12分钟到55分钟之间。由于评价视频边界具有一定的主观性，本文选择的实验视频主要是一些新闻类，教育类视频，此类视频具有较为明显的视频边界，易于人工判断。首先，邀请了5个参与实验的同学，让他们认真观看视频，理解视频内容，然后标注出视频主题切换的时间节点，这些时间节点作为评价实验结果的参照。本文评价算法的有效性的方法是将算法检测的主题边界结果与人工标注的主题边界结果作对比，如果两个边界的时间戳差值在5秒以内，就认为该边界准确有效。

本文采用信息检索领域常用的查准率（Precision），查全率（Recall），和F-measure三个常用的性能指标来评价算法的好坏。公式2.6-2.8给出了它们在本文中的计算方法。

$$Precision = \frac{\text{检测出的准确边界数}}{\text{检测出的边界总数}} \tag{2.6}$$

$$Recall = \frac{\text{检测出的准确边界数}}{\text{人工标注的边界总数}} \tag{2.7}$$

$$Precision = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.8}$$

其中，查准率反映了实验结果的准确性，查全率反映了检测的全面性，F-measure是查准率与查全率的调和平均值，综合放映了实验结果的好坏。这些性能指标的数值在0到1之间，越大说明算法效果越好。

2.5.2　实验结果及分析

表2-2给出了本文主题边界检测实验结果，T0是仅用视频视觉通道特征得出的结果，T1是仅用视频听觉通道特征得出的结果，T2是基于双通道线索融合得出的结果。可以看出，相比于仅利用视觉通道或者仅利用听觉通道特征，基于双通道线索融合的方法有较大的优势。原因在于，有些视频场景比较单一，如教育类视频，场景大多是在教室，这样仅用视觉通道特征往往效果较差；还有些视频语音信息语义相似度很大，如一些体育视频，不利于语义的主题分割。本文基于双通道线索融合的主题边界检测方法很好的融合视频双通道信息特征，算法在10个视频上获得了平均查准率0.83、平均查全率0.68和平均F-measure0.74的效果，基本能够获取准确有效的视频主题边界。

表 2-2　主题边界检测实验结果

VideoID	T0			T1			T2		
	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure	Precision	Recall	F1-Measure
1	0.63	0.56	0.59	0.50	0.40	0.44	0.75	0.67	0.71
2	0.88	0.54	0.67	0.75	0.55	0.63	0.88	0.78	0.82
3	1.00	0.56	0.71	0.80	0.57	0.67	1.00	0.71	0.83
4	0.78	0.64	0.70	0.67	0.67	0.67	0.78	0.78	0.78
5	0.63	0.42	0.50	0.88	0.54	0.67	0.88	0.64	0.74
6	0.71	0.36	0.48	0.86	0.55	0.67	0.86	0.50	0.63
7	0.83	0.38	0.53	0.67	0.50	0.57	0.83	0.63	0.71
8	0.60	0.55	0.57	0.70	0.58	0.64	0.80	0.80	0.80
9	0.56	0.83	0.67	0.67	0.46	0.55	0.78	0.64	0.70
10	0.64	0.44	0.52	0.55	0.46	0.50	0.73	0.67	0.70
average	0.72	0.53	0.59	0.70	0.53	0.60	0.83	0.68	0.74

2.6　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

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。目前各视频分享网站为视频添加缩略图主要有两种方法：其一是人工定制，这种方法生成的缩略图质量图很高，但这个过程需要人工浏览并理解视频内容，从复杂的视频信息中提炼重要信息以生成合适的缩略图，所以非常耗时，为每一个视频人工定制一张缩略图显然是不现实的；其二是随机选择一张视频帧作为缩略图，随机的方法多种多样，可以是视频第一帧或者是视频50%处的视频帧，例如，youtube网站随机选择3张视频帧供视频上传用户选择，这种方法固然效率很高，但是很容易导致视频缩略图质量较低。

本章提出了一种基于查询自适应的视频缩略图生成方法，该方法能感知用户的查询意图，并结合视频的图文内容生成出信息丰富，布局美观，自适应用户查询意图的缩略图，从而让用户能了解视频中其最感兴趣的部分，帮助用户做出正确的相关性判断，提高用户体验。本章内容安排如下：3.1节介绍视频内容结构化方法；3.2节提出了视频图文内容的质量评价模型，从而为视频缩略图选取合适的图文内容；3.3节根据已选取的图文内容生成布局美观的视频缩略图；3.4节展示了视频缩略图的生成结果，并对结果进行评估，验证了本文方法的有效性；最后是本章的总结。

3.1　视频内容结构化方法

视频数据一种信息丰富的多媒体数据，由于视频数据的复杂性和无结构性，处理视频数据之前需要对视频数据进行有效的组织，将无序的视频流媒体的数据整理成有序的结构化数据，以便高效快速的分析视频数据。视频结构化是指通过视频分割、特征提取、目标识别等方式提取视频内容信息，根据语义关系对视频内容进行分类组织，形成方便计算机及用户检索和理解的文本信息[53]。因此，视频内容结构化对视频数据的处理具有重要的意义。视频内容结构化的方式多种多样，没有一个统一的标准，本文提出了一种基于主题单元分割的视频结构化方法，即先根据视频语义检测是视频主题边界，将视频分割成语义连贯的视频主题单元，这部分工作本文第二章已给出了解决方法，然后提取、挖

掘每个主题单元的显著、重要的图文信息，并加以有效的浓缩、组织、归纳。通过本文视频结构化方法，可以是计算机及用户快速，清晰地掌握视频主题脉络，并理解各个主题单元的内容。

3.1.1　听觉通道内容结构化

对于视频听觉通道内容，先要把难以处理的音频特征转化为易于处理的文本信息，对于有字幕文件的视频，直接使用字幕文本，对于没有字幕文件的视频，本文先用2.1.2小节方法介绍的微软语音转化工具包将视频语音转化为文本信息。对于语音转换来的文本或者字幕文件，首先浓缩文本信息，提取关键词。表3-1给出了本文对youtube网站视频ID为”40riCqvRoMs”语音文本进行关键词提取的得分最高的部分结果。

表 3-1　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

本文使用Rose[52]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，并且倾向于较长的短语。RAKE算法首先使用标点符号（如半角的句号、问号、感叹号、逗号等）将一篇文档分成若干分句，然后对于每一个分句，使用停用词作为分隔符将分句分为若干短语，这些短语作为最终提取出的关键词的候选短语。对于每个候选短语，公式3.1给出了候选短语的重要性得分phraseScore计算方法：

$$phraseScore = \sum_{i}^{n} \frac{wordDegree(w_i)}{wordFrequency(w_i)} \tag{3.1}$$

其中n为候选短语包含单词的个数，wordDegree为单词的度，当与一个单词共现

的单词越多，则该单词的度就越大，wordFrequency为单词出现的频率，每个候选短语的得分由组成短语的单词得分累加得到，而单词的得分是单词的“度”与单词出现的频率的比值。上面的步骤对文本信息做了提炼、浓缩，接下来结合本文第二章主题边界检测算法得出的视频主题边界及文本信息的时间戳得出各个主题的关键文本。图3-1展示了视频听觉通道结构化的过程。

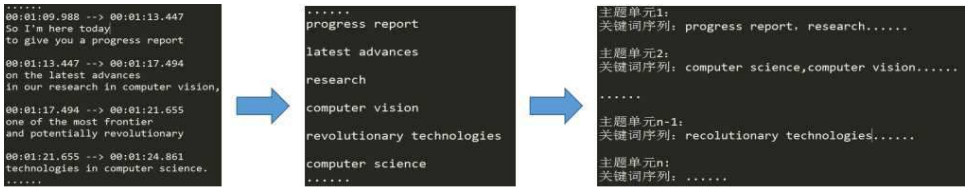


图 3-1　视频听觉通道内容结构化过程：(a)字幕文件或者语音转换的文本信息;(b)关键词提取后的文本;(c)最后的结构化文本

3.1.2　视觉通道内容结构化

由于网络视频大多由非专业人员拍摄，其拍摄设备和拍摄技术参差不齐，导致视频模糊画面较多，视频质量往往较低，而且视频中如果含有较多运动场景，这些摄像机变焦产生的画面运动常常引起画面的模糊，所以首先要过滤掉这些模糊、昏暗、低质量的图像，只有清晰高质量的帧才能作为组成缩略图的候选帧。本文遇到的图像质量评价问题是一个典型的无参照问题，不适合采用传统图像还原问题与高质量原图对比的方式，本文自动聚焦函数（Autofocusing function）来判定视频帧质量的好坏。自动聚焦函数是显微成像领域计算图像模糊常用的方法，其输出值越小，则图像越模糊，图像质量越差，相反，其输出值越大，则图像越清晰，图像质量越高。自动聚焦函数多种多样，例如归一化方差函数[54, 55]，拉普拉斯能量函数，Brenner梯度函数[56]等等,其中归一化方差函数的性能是最好的[54],所以本文使用归一化方差函数衡量个视频帧的模糊程度，公式3.2给出其计算方法：

$$quality(i) = \frac{1}{H \times W \times \mu_i} \sum_{W}^{x=1} \sum_{H}^{y=1} (S_i(x,y) - \mu_i)^2 \tag{3.2}$$

$$\mu_i = \frac{1}{H \times W} \sum_{W}^{x=1} \sum_{H}^{y=1} \tag{3.3}$$

其中H和W分别为视频帧的高度和宽度，$S_i(x,y)$为视频帧i像素点坐标为（x,y）

的亮度值，即HSV色彩空间S通道的值，μ_i为视频帧i的亮度均值，计算方法如公式3.3所示。本文优先选取自动聚焦函数输出值quality(i)大的视频帧作为视频缩略图视觉内容的候选帧。

过滤掉模糊的低质量的视频帧后，后续的处理过程如图3-2所示，首先对视频帧进行镜头分割，去除掉重复冗余的视频帧，然后利用本文第二章主题边界检测算法得出的视频主题边界及视频帧对应的时间戳将视频帧序列分为若干个主题单元，为了找出每个主体单元最显著的视频帧，本文对每个主题单元内的视频镜头根据其视觉特征相似性进行聚类，选取每个类簇的类中心作为这个主题单元的代表帧。在聚类之前，计算每个镜头的颜色直方图用来表示镜头的视觉内容，颜色直方图计算简单，对于视频拍摄时摄像机视角的小变化不敏感，同时对物体的自身运动也有较好的鲁棒性。本节方法中颜色直方图计算方法为16*4*4(H:16,S:4,V:4)的256维的归一化HSV颜色直方图，之所以将H通道的比重划分的更为细致，是因为HSV颜色空间对应于色度、饱和度和亮度人眼色彩视觉特征3要素，而相对于饱和度、亮度两个分量，色度分量更加影响人类的视觉判断。提取所有镜头特征向量后，使用经典的K-Means聚类算法对这些特征向量进行聚类即可。

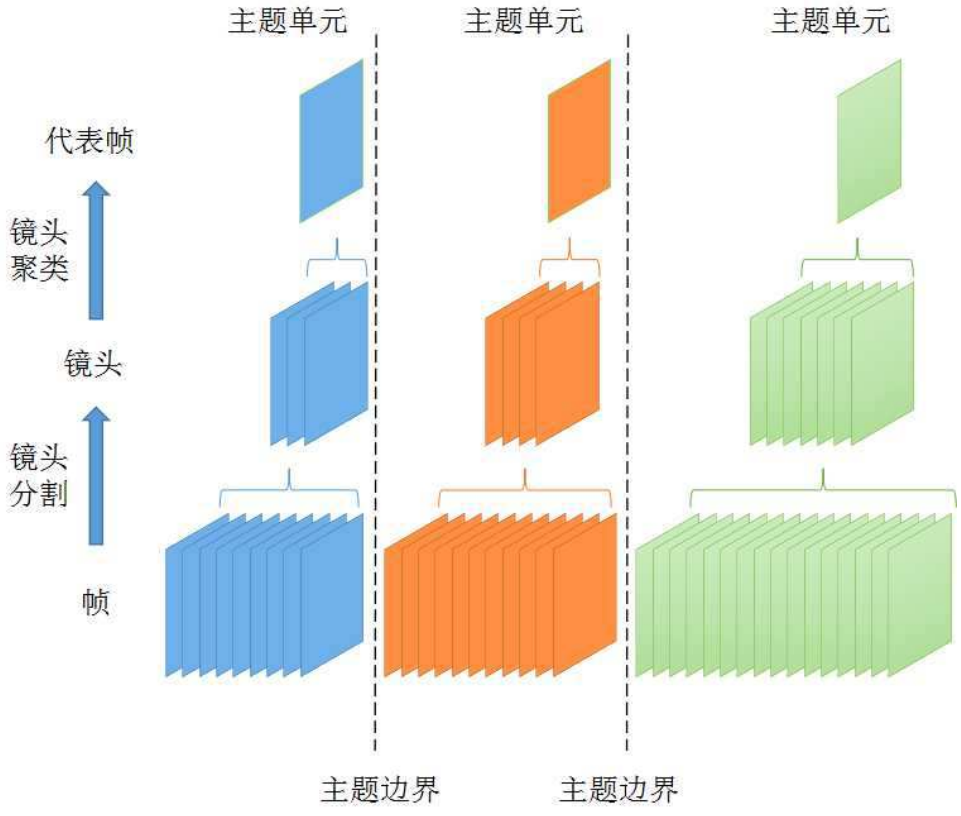
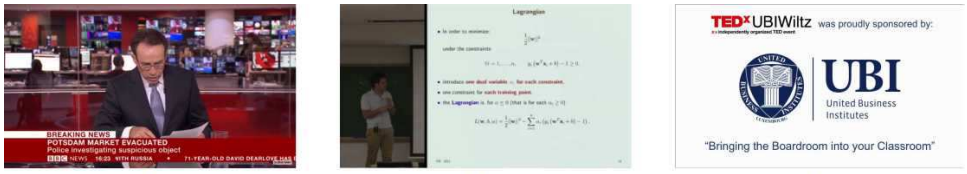


图 3-2　视频视觉通道内容结构化过程

视频镜头图像中可能还包含一些非常有价值的文字信息，特别是新闻类、教育类、演讲类视频，如图3-3所示。这些文字信息精炼，具有高度的概括性，直接反映了视频的内容，本文使用OCR技术提取视频帧的文字内容，以便后续处理。



(a)　(b)　(c)

图 3-3　有些视频帧含有丰富的文字信息：(a)bbc新闻视频帧；(b)斯坦福公开课视频帧；(c)TED演讲视频帧

3.1.3　主题单元内容结构化

本文提出了一种基于主题单元分割的视频内容结构化方法，首先用本文第二章视频主题分割方法将复杂无结构的视频流数据先分割成一个个语义内容外部独立、内部连贯的主题单元片段，对与每个主题单元片段双通道内容进行提炼、浓缩，提炼过程如图3-4所示，最终主题单元被结构化为一个精炼、具有高度概括性和代表性的图文二元组（I,T）,其中’I’代表图像内容，它是从视频原始帧序列经过质量评估、镜头分割、镜头聚类等步骤选取一帧最有代表性的帧得出的；’T’代表文字内容，文字内容来源于视频语音字幕和图像OCR识别得到的文本信息，然后提取文本关键词得到关键词序列。I,T在时间维度上重合，分别以图像和文字的方式概括着视频主题单元的内容。

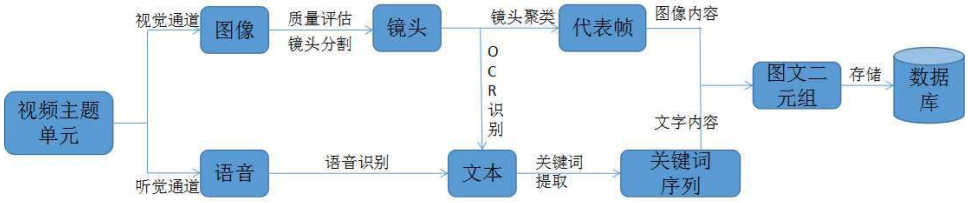


图 3-4　视频主题单元结构化过程

最终，本文将视频内容结构化为如图3-5所示，视频有若干个主题单元组成，每一个主题单元都由其代表视频帧和关键词序列图文二元组（I,T）表示，这些图文二元组（I,T）以代表帧和关键词序列的形式概括了主题单元最显著的图文信息，最终被存储在数据库中，以便后续计算的需要。

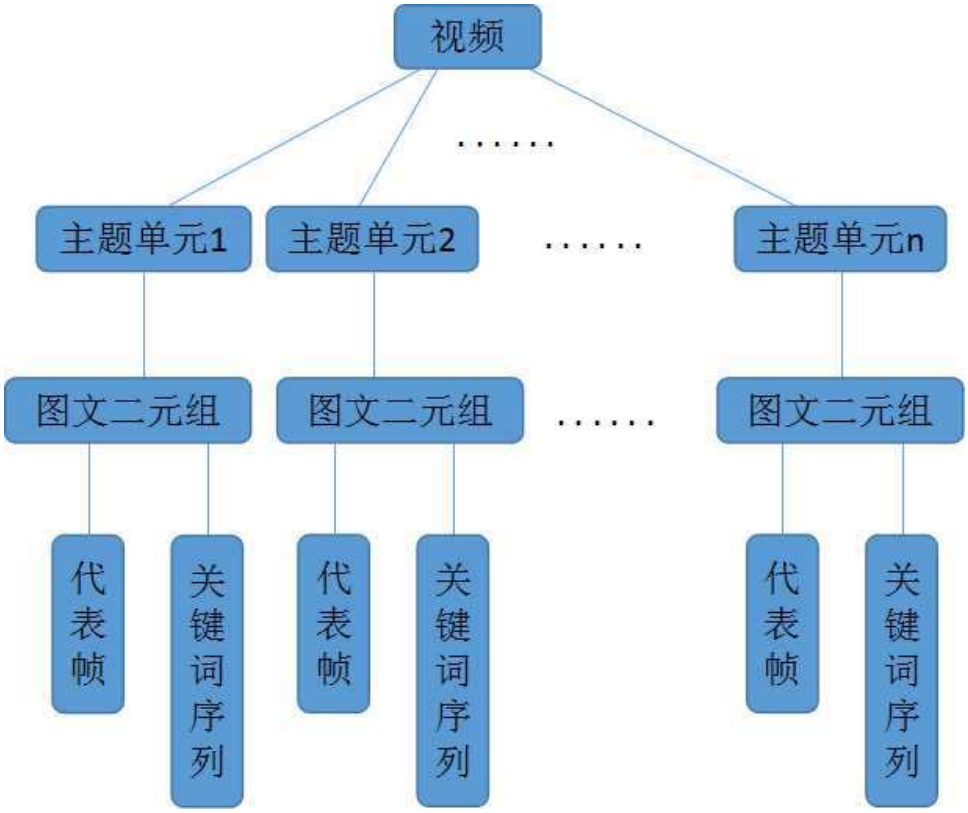


图 3-5　视频结构表示

3.2　视频缩略图图文内容提取

本节介绍视频缩略图图文内容提取方法，图文内容的质量直接决定了缩略图的质量，本文提取缩略图图文内容时主要考虑其代表性和与用户查询意图的相关性。关于代表性，上一节本文将无结构的视频数据结构化为一个个精炼、具有高度概括性和代表性的图文二元组，本节将这些图文二元组作为视频缩略图候选图文内容。本节后续内容将介绍如何在这些候选图文二元组中选取符合用户查询意图的图文内容。

用户在视频分享网站里搜索自己感兴趣的视频时，一个典型的搜索过程如下：用户输入代表自己搜索意图的查询词，网站搜索引擎根据查询词返回相关视频给用户。本文计算用户查询词与视频主题单元图文二元组的相关性，以倒序相关性排序图文二元组，选取排名靠前的图文二元组作为视频缩略图内容。

图像内在的语义比较抽象，虽然Karpathy[57]等人基于深度网络建立了图片与语义空间的映射从而描述了图片的语义，而本文希望能在用户浏览的过程中在用户可接受的时间内生成出动态的缩略图，但是这种方法是相对耗时的，显然不能满足本文需求。图文二元组的图文在时间维度上是重合的，因此它们的

语义是紧密相连的，所以本文计算图文二元组中关键词序列与查询词的相关性来衡量它们之间的相关度。要解决这个问题，首先要把这些自然语言符号数学化，在自然语言处理领域中通常用词向量表示自然言语中的单词。一种最简单的词向量方式是”one-hot representation”，向量的维度为预料库词典的长度，向量的分量只有一个1，其他全为0。这种方法虽然简单，但缺点也很明显，第一，向量维度冗长，容易造成维度灾难；第二，这种向量有”词汇鸿沟“的问题，不能很好的刻画词与词的相似性。本文采用Hinton[58]提出的词向量”Distributed Representation”的表示方法，其基本思想在于将语料库训练成一个向量空间，使得语料库中每一个词都对应向量空间中一个固定长度的向量，当计算两个词之间的相似度时，则根据这两个词对应的向量的距离来判断它们之间的相似性。

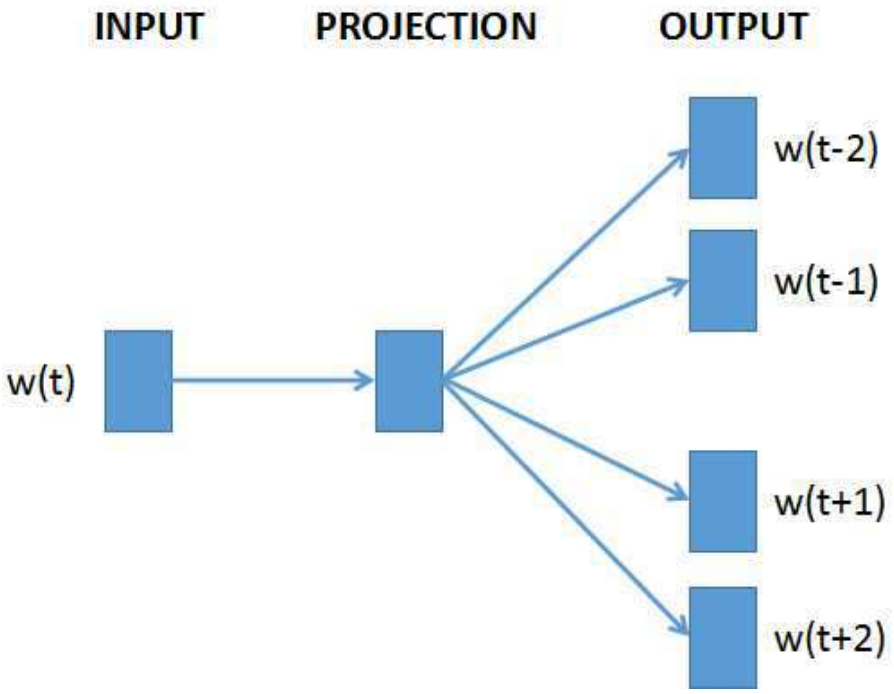


图 3-6　Skip-gram Model

本文采用Mikolov[59, 61]等人提出的word2vec词向量空间计算方法，word2vec能在百万乃至上亿数量及的语料库上进行高效的训练[61]。他们使用名为一个浅层神经网络语言模型“Skip-gram Model”对语料库中的每一个词做向量表示，如图3-6所示，模型分为输入层，投影层，输出层。Skip-gram模型在输入已知的w(t)前提下，预测其上下文w(t-2),w(t-1),w(t+1),w(t+2)。在训练过程中，使得每个词向量与其上下文对数概率P最大化，公式3.4给出了P的计算方法，给定单词序列$w_1, w_2, ..., w_T$，nb(t)是单词w_t上下文单词集合，$p(w_j|w_t)$是w_j和w_t两个单词

词向量的层次softmax。

$$P = \frac{1}{T}\sum_{t=1}^{T}\sum_{j\epsilon nb(t)} logp(w_j|w_t) \tag{3.4}$$

本文以大约18万个视频文本内容作为训练数据，得到语料库中每一个词汇的特征向量，每个向量50维，两个词语义相关度以两词向量余弦距离衡量，表3-2展示了部分单词以上述方法得出的最相关的词汇。

表 3-2　基于Wordvex最相关词汇挖掘

word	related word
computer	computers,software,technology,electronic,internet,computing,devices,digital
human	animal,rights,aids,nature,particular,body,that,causes
apple	blackberry,chips,iphone,microsoft,ipad,ipod,intel,software
movie	movies,film,films,comedy,hollywood,drama,sequel,animated
network	network,networks,cable,channel,channels,internet,television,media,satellite
smile	grin,smiles,eyes,smiling,laugh,touch,gentle,smirk,tears,whisper
music	musical,dance,songs,recording,folk,studio,contemporary,artists
history	historical,tradition,great,career,ever,life,greatest,literature,decades,origins
language	languages,word,spoken,vocabulary,translation,arabic,refers,dialect
frog	snake,toad,monkey,spider,lizard,spiny,orchid,rattlesnake,snakes,salamanders

用户查询词与图文二元组关键词序列相关度计算方法主要思想是将两部分词序列基于word2vex语义相关度做一个映射，如图3-7所示，首先对于用户查询词序列中的每一个词，分别计算其与图文二元组关键词序列中每一个词中的语义相关度，那么语义相关度最大对应的词就是其在图文二元组词序列中的映射词，如图3-7中红色线条所示；同理对视频图文二元组中词序列做同样的操作，得到它们在用户查询词序列中的映射，如图3-7绿色线条所示。将两部分映射词语义相关度的值累加除以总词数就是两个词序列之前的语义相关度。算法2给出了计算用户查询词与图文二元组关键词序列的具体算法流程。

3.3　视频缩略图图文内容布局方法

上文基于图文内容质量，代表性，与用户查询词的语义相关性，从视频图

算法 2 文本相关度衡量算法

输入: 用户查询词序列，视频关键词序列序列
输出: 文本相关度

```
function Compute_Text_Relation(String[] text1,String[] text2)
    double sum=0;
    for s1 in text1[] do
        double max=0;
        for s2 in text2[] do
            double score=Word2VexScore(s1,s2);
            if score > max then
                max ← score
            end if
        end for
        sum ← sum + max
    end for
    for s1 in text2[] do
        double max=0;
        for s2 in text1[] do
            double score=Word2VexScore(s1,s2);
            if score > max then
                max ← score
            end if
        end for
        sum ← sum + max
    end for
    result ← sum / (text1.length+text2.lengtg)
    return result
end function
```

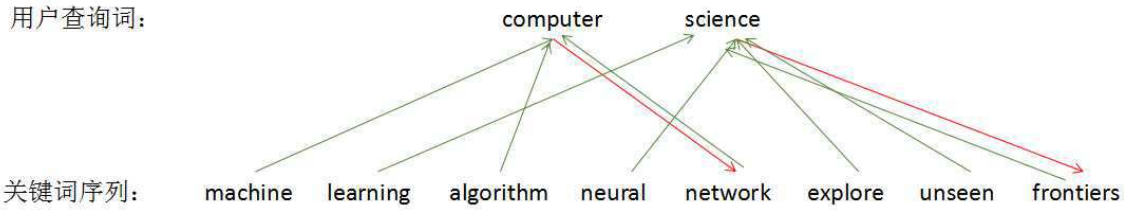


图 3-7　用户查询词与视频关键词序列相关度计算

文内容中选出了最合适的图文二元组作为缩略图的候选图文内容，本节介绍缩略图图文二元组布局方法，将这些候选图文内容合理的组织，从而生成美观的缩略图。

在生成缩略图之前，本文首先采用Song[3]等人的方法检测候选图像显著性区域，该方法将图像的空间布局结构特征化为图像边界，并且结合多个底层线索优化图像边界，从而得到圆润的显著性灰度图，接着确定一个自适应阈值将图像二值化，最后通过检测轮廓得到图像显著性区域，本文选择面积最大的轮廓作为图像显著性区域。

3.4　实验结果及评估

3.5　本章小结

此处省略N个字……

第 4 章　基于内容的视频可视化浏览系统

随着网络视频数量越来越多，用户在视频海中找出自己感兴趣视频的压力也越来越大；当用户找到自己感兴趣的视频后，由于视频数据的复杂性，用户理解视频内容、定位感兴趣视频片段也是非常耗时的。视频浏览系统为了更好的服务用户主要解决以下两个问题：第一，帮助用户快速检索出自己感兴趣的视频；第二，在浏览视频内容时，帮助用户快速定位到感兴趣片段、理解掌握视频内容。

本文第二章基于双通道线索对视频进行了主题分割，将视频分割为语义独立、内容连贯的主题片段；第三章对视频各个主题片段进行内容精炼，结构化，并为视频生成了基于用户查询自适应的视频缩略图；本章基于上述两章的研究结果设计并实现了一种基于内容的视频可视化浏览系统VideoVis，该系统较好地满足用户在线浏览和定位查找视频内容的效率需求，其特点在于：第一，对视频双通道内容进行索引，提高了视频检索精度；第二，对视频检索结果进行内容聚类并可视化，提高视频检索效率；第三，对视频双通道内容进行精炼、结构化，可视化，帮助用户高效定位感兴趣片段、并理解其内容。

4.1　视频可视化浏览系统设计

本节介绍视频可视化浏览系统VideoVis设计方案，系统主要分为视频检索和视频内容可视化两个模块，第一个模块旨在帮助用户在视频库中快速检索出感兴趣的视频，提高用户检索效率，第二个模块旨在帮助用户快速定位到感兴趣片段、理解掌握视频内容。下面详细介绍这两个模块设计方案。

4.1.1　视频检索可视化模块设计

视频检索模块的目标是帮助用户快速找到感兴趣的视频，缩短用户搜索时间，提高检索效率。用户传统的搜索过程如下：用户输入代表自己搜索意图的查询词，网站搜索引擎根据查询词返回相关视频列表，若视频列表过长，则采用分页的方式呈现视频列表给用户，有时页数达到几十页甚至上百页，用户需要通过翻页的方式逐条筛选出感兴趣的视频并点击观看，这种方式极大的影响了检索效率，降低用户体验。本文用一种层次化可视模型将返回的视频搜索结果以可视化的方式呈现给用户，其与用户交互的web界面如图4-1所示。

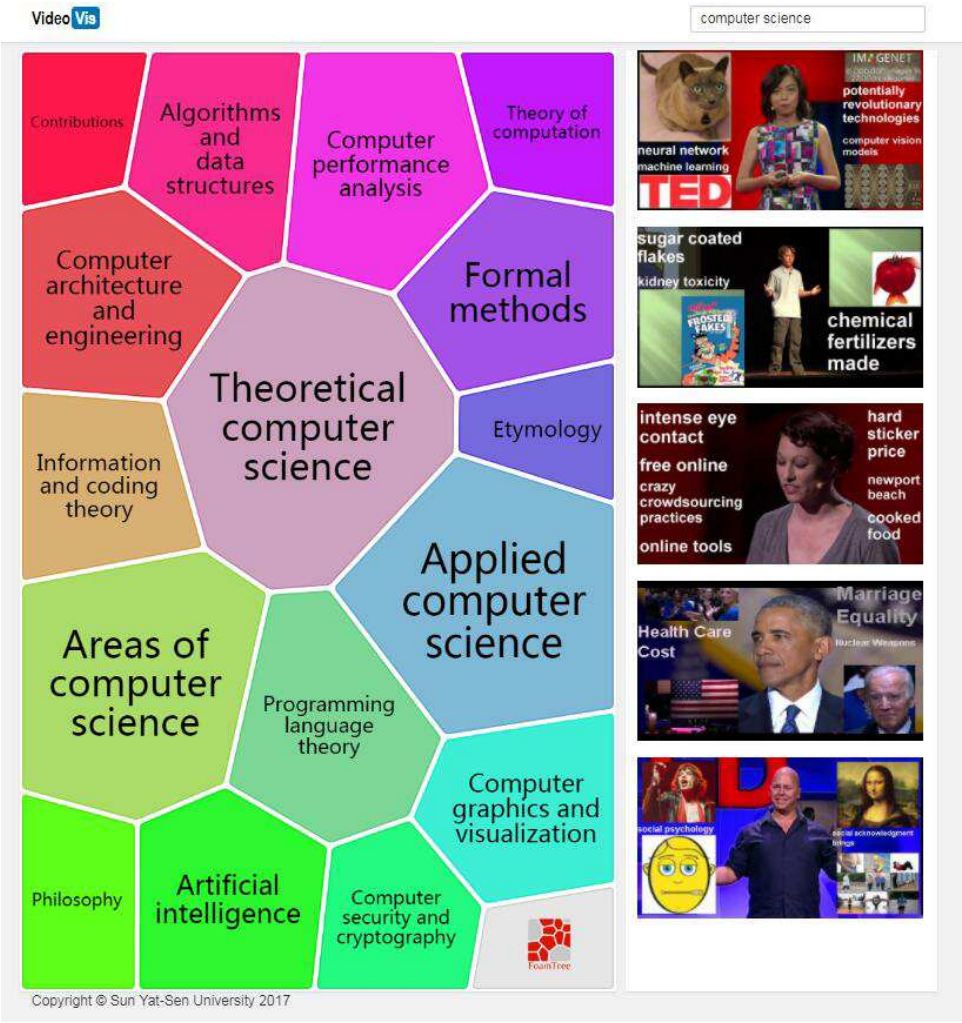


图 4-1　视频检索结果可视化模型

用户输入的查询词概念通常比较宽泛，有很多方面[61],有的查询词包含的非常复杂的结构。例如以“computer science”搜索视频时，搜索引擎通常返回关于计算机科学的方方面面的视频，如算法和数据结构、计算机系统架构、编程语言理论、人工智能等相关视频；又如以“Steven Spielberg”搜索视频时，搜索引擎通常会返回关于Steven Spielberg的电影，采访，成就，个人生活等等相关视频。这种情况下，传统的以一个视频列表包含着不同方面主题视频作为搜索结果返回给用户的方式是低效的，用户会非常难以定位到自己感兴趣的视频。因此，VideoVis系统在得到搜索结果后，先基于视频内容对视频搜索结果进行聚类分析，挖掘用户查询词语义相关话题各个方面的子话题，每个子话题对应于图4-1的各个标签，然后将视频检索结果根据视频内容映射到搜索词的各个方面，最后将层次化视频检索结果可视化呈现给用户。呈现方式采用FoamTree①的方

① https://carrotsearch.com/foamtree/

式，如图4-1所示，FoamTree生动、形象，是一个基于JavaScript的用于展示层次化数据的可视化模型。这种可视化方式细化了用户查询意图，起到了类似目录的作用，大大提高了用户检索视频的效率。当用户点击每个标签时，右边的视频缩略图列表会相应的刷新。图4-1右边的部分用来展示每个标签下对应的具体视频，展示的方式是用本文第三章为视频生成的与用户查询自适应的视频缩略图,它很好的展示了各个视频中用户感兴趣的内容，指导用户是否继续点击以观看视频。

4.1.2　视频内容可视化模块设计

当用户点击图4-1中右边部分的视频缩略图时，VideoVis系统便会进入视频内容可视化模块，此模块用来展示视频双通道信息，以便用户快速定位到自己感兴趣的片段并快速读懂理解相关视频内容。

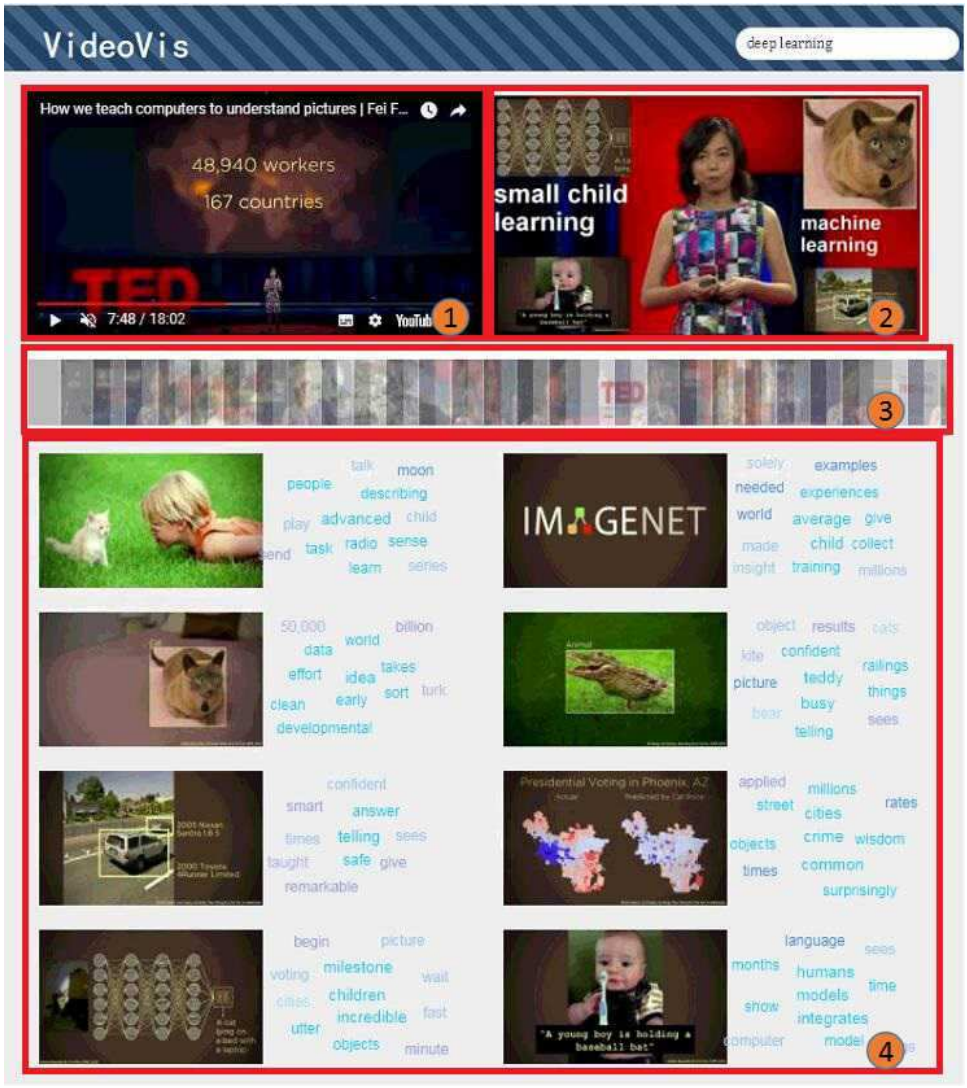


图 4-2　视频内容可视化模块设计

该模块与用户交互的web界面如图4-1所示，模块分为4个部分，第一部分是视频播放器，它位于整个页面的左上方，用于控制视频的播放，同时在与其它模块交互的过程中，跳转到视频相应的时间节点；第二部分是视频缩略图部分，它位于页面的右上方，用于概括视频中与用户查询相关的内容；第三部分用来展示视频的各个镜头图像内容，镜头是视频视觉通道上最直观最重要的信息，当用户点击镜头图像时，视频播放器会跳转到该镜头对应的时间节点；第四部分展示视频各个主题单元的内容，本文第二章检测了视频主题边界，3.1.3小节将视频内容结构化成图文二元组，该部分便从左到右展示了各个主题单元的图文信息。每个主题单元分为图文两部分，图片即为该主题单元的代表帧图片，文字为该主题单元关键词序列的词云。当用户点击各个主题代表帧图片时，视频播放器会跳转到主题单元起始的时间点。

4.2　视频可视化浏览系统实现

本节介绍VideoVis系统实现细节，系统实现流程图如图4-3所示。实现流程主要分为3个大步骤：第一，视频内容预处理，通过镜头分割、OCR识别、语音识别、镜头聚类、关键词提取等步骤把复杂的视频流数据转化为易于计算机处理的数据；第二，视频内容分析，检测视频主题边界，对视频内容进行结构化处理，基于视频内容和用户查询词生成视频缩略图，对视频内容进行聚类分析；第三，将第二步挖掘的内容以可视化的方式友好的呈现给用户。其中，有些过程的实现在上两个章节已经详述，本节下面的内容将介绍视频搜索引擎、视频内容聚类的实现细节。

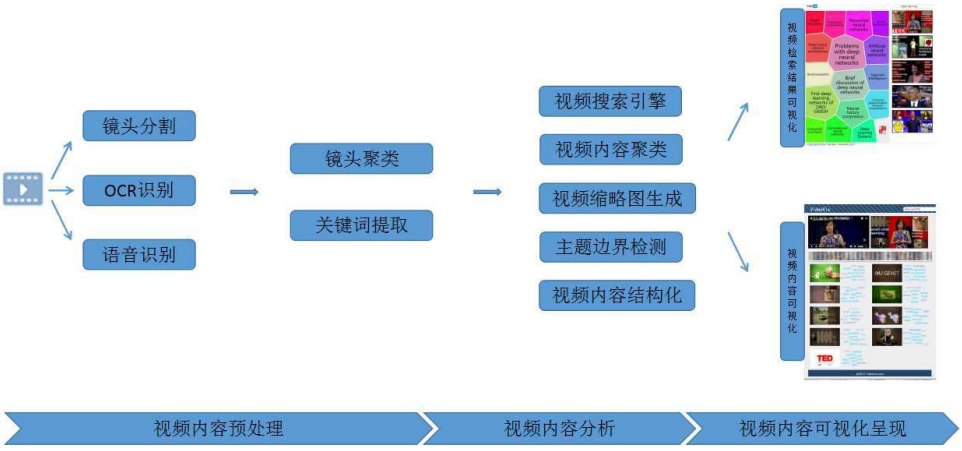


图 4-3　VideoVis系统实现流程图

VideoVis系统搜索引擎架构如图4-4所示，引擎分为3个模块，第一个模块为

离线处理模块，对视频进行OCR识别和语音识别得到文本信息，接着精炼这些文本信息，提取关键词序列。第二个模块是系统的后台，首先对各个视频提取关键词进行全文索引，接着将索引放在服务器目录下，等待用户访问。本文采用高性能文本搜索引擎Lucene①,lucene是Apache基金会旗下一款开源的文本搜索引擎，具有性能高，接口简单等优点。当用户通过浏览器搜索视频时，服务器即通过Lucene索引匹配用户查询词找出相关视频返回给用户。第3个模块为系统客户端浏览器，图4-1展示了系统和用户交互的web页面。

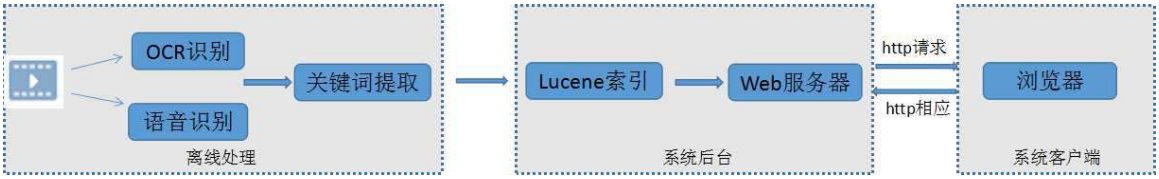


图 4-4　VideoVis系统搜索引擎架构图

VideoVis系统在得到搜索引擎结果后不直接向用户展示视频结果列表，而是将结果视频进行聚类分析，将视频聚成一簇簇内容相近的类簇，聚类流程图如图4-5所示。本文聚类方法主要参考了Jiang[61]等人的方法，首先挖掘用户输入的搜索词的语义层次结构，希望这个语义结构能概括用户搜索词的方方面面，维基百科②很好的满足这一需求。维基百科是一个庞大的众包语料库，本文首先下载维基离线语料库③，对其中每个词条及其语义层次结构结点内容建立Lucene索引，便于快速查阅。当用户输入查询词搜索时，分别检索维基百科索引和视频内容索引得到查询词语义层次结构和视频搜索结果，紧接着，本文参照Jiang[63]等人的方法并结合系统实际的需求将视频映射到维基语义结点上，映射过程中主要考虑视频与结点的相关性、视频从属结点的唯一性、结点中视频的多样性，它们的计算方式如公式4.1,4.2,4.3所示：

$$Rel(v,n) = Sim(v,n) \tag{4.1}$$

$$Uniq(v,n_i) = \frac{Sim(v,n_i)}{\sum_{j=1}^{N} Sim(v,n_j)} \tag{4.2}$$

$$Div(v,n_i) = max_{v_j \epsilon V}(Sim(v,v_j)) \tag{4.3}$$

① https://lucene.apache.org/
② https://en.wikipedia.org/wiki/Wiki
③ 下载地址：https://dumps.wikimedia.org/

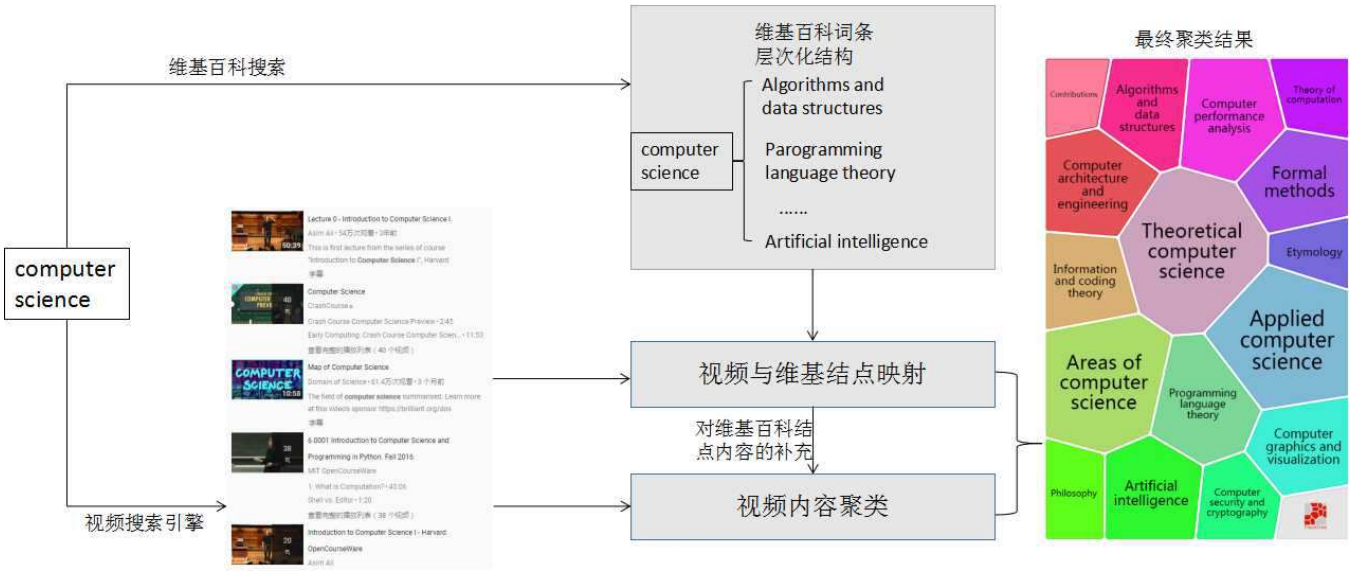


图 4-5　视频内容聚类流程图

公式4.1给出了视频v与结点n相关度的计算方法，$Sim(v,n)$表示视频v关键词序列与结点n文字内容基于word2vex的语义相关度，具体算法与算法3-2一致；公式4.2给出了视频v结点n_i的唯一性计算方法，显然相较于一个视频只属于多个结点，一个视频只属于一个唯一的结点的效果更好，其中N为结点总个数,视频v与结点n_i的唯一性定义为v与n_i语义相关度除以视频v与所有结点的语义相关度之和；公式4.3给出了结点n_i多样性的计算方法，其中V是结点结点n_j中已有的视频集合，$Sim(v,v_j)$表示视频v与视频v_j的语义相关度，用当前视频v与结点中视频语义相关度最大值作为视频v与结点n_i的多样性，考虑多样性主要是为了去除结点内内容特别相近的视频。基于相关性，唯一性，多样性定义目标函数F计算方法如公式4.4所示：

$$F = max_{v_i \epsilon V, n_j \epsilon N} \sum_i \sum_j \beta * Rel(v_i, n_j) * Uniq(v_i, n_j) - (1-\beta) Div(v_i, n_j) \quad (4.4)$$

目标函数F是一个全局优化优化函数，使得视频与结点的映射结果在相关性、唯一性、多样性上做到全局上做到最优，尽管可以用穷举的方式得到全局最优解，但考虑到计算效率，本文采用一种贪婪的方式，即添加一个视频到结点中就使得F局部最优，进而得到最终的相对较优的可行解。其中β是调节相关性、唯一性、多样性的参数，本文使用交叉实验来确定β的具体值，如图4-6所示，本文选取20个用户搜索最为频繁的检索词，β从0到1以步长0.1取值，分别计算每个查询词结果视频集每个类别下视频与结点的相关度，进行交叉实验。基于图4-6所

示交叉实验的结果，本文β取0.7。

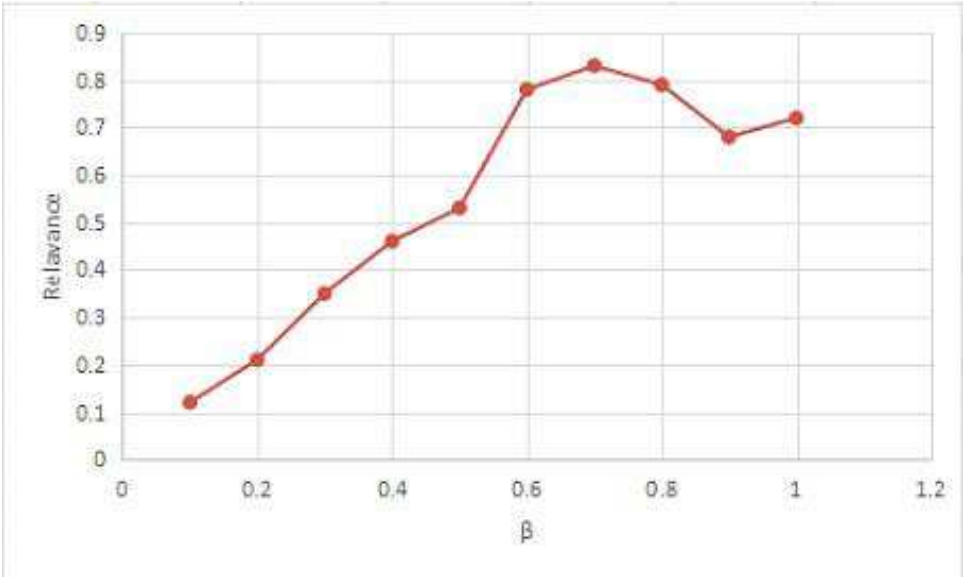


图 4-6　β参数交叉实验折线图

然而并不能保证维基词条包含用户每个查询词，而且即使包含用户查询词也不能保证其语义层次结构能覆盖查询词的方方面面，因此，作为对维基词条语义结构的补充，本文在维基词条语义结构的基础上挖掘用户查询词更全的语义结构。当维基词条没有收录用户查询词或者当视频内容与维基词条语义结构相关度不高时，本文使用carrot2①开源框架对视频搜索结果进行聚类分析，carrot2框架能实时对文档集基于语义聚类，并且为每个类别生成一个具有概括性且用户可读的文本标签，其原理基于Osinski[63]等人提出的lingo算法。lingo算法首先介于SVD分解挖掘文档背后的抽象概念，接着这些抽象概念构造用户可读的类标签，最后将文档分配到这些类标签中。图4-7展示了VideoVis系统一些基于不同搜索词得到的结果视频的聚类分析结果。

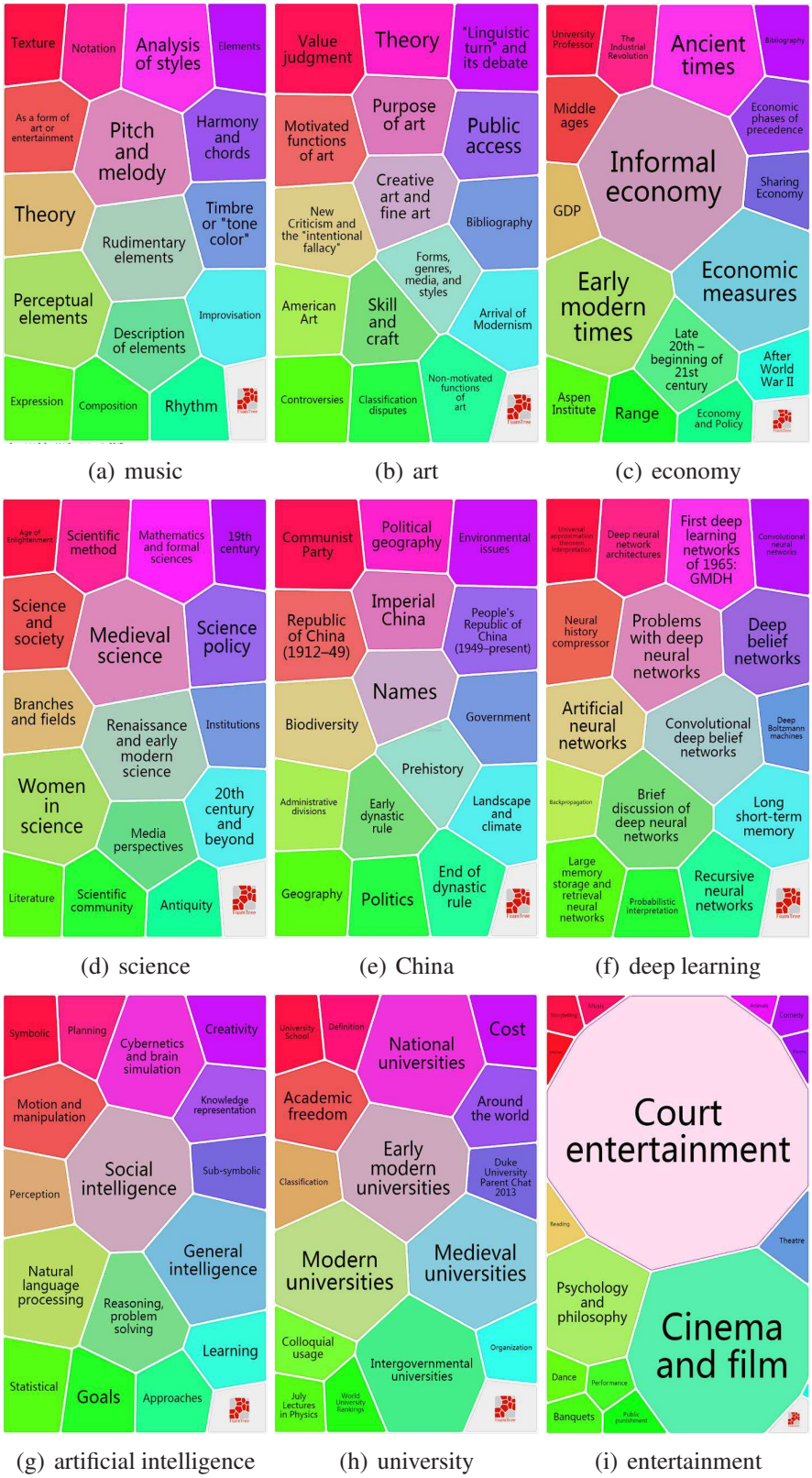
4.3　视频可视化浏览系统实验评估

4.3.1　视频检索可视化模块实验评估

4.3.2　视频内容可视化模块实验评估

4.4　本章小结

① http://project.carrot2.org/



(a) music　(b) art　(c) economy

(d) science　(e) China　(f) deep learning

(g) artificial intelligence　(h) university　(i) entertainment

图 4-7　视频聚类分析结果

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] Bouthemy 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 compresse dvideo[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.

[53] 袁祉赟. 基于内容的视频结构化方法研究[D].电子科技大学,2016.

[54] Sun Y, Duthaler S, Nelson B J. Autofocusing in computer microscopy: selecting the optimal focus algorithm[J]. Microscopy research and technique, 2004, 65(3): 139-149.

[55] Santos A, Ortiz de Soló rzano C, Vaquero J J, et al. Evaluation of autofocus functions in molecular cytogenetic analysis[J]. Journal of microscopy, 1997, 188(3): 264-272.

[56] Brenner J F, Dew B S, Horton J B, et al. An automated microscope for cytologic research a preliminary evaluation[J]. Journal of Histochemistry and Cytochemistry, 1976, 24(1): 100-111.

[57] Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.

[58] Hinton G E. Learning distributed representations of concepts[C]//Proceedings of the eighth annual conference of the cognitive science society. 1986, 1: 12.

[59] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[60] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in neural information processing systems. 2013: 3111-3119.

[61] Jiang Y G, Wang J, Wang Q, et al. Hierarchical visualization of video search results for topic-based browsing[J]. IEEE Transactions on Multimedia, 2016, 18(11): 2161-2170.

[62] Tan S, Jiang Y G, Ngo C W. Placing videos on a semantic hierarchy for search result navigation[J]. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 2014, 10(4): 37.

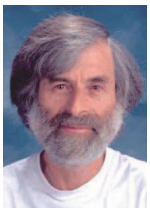
[63] Osinski S, Weiss D. A concept-driven algorithm for clustering search results[J]. IEEE Intelligent Systems, 2005, 20(3): 48-54.

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