摘 要

近年来，基于内容的多媒体信息检索已经成为了一个热门的研究领域，新闻视频检索作为其中一部分也得到了广泛的研究。新闻视频是人们获取信息的主要媒体，但如何快速、准确地从海量的新闻视频数据中，找到所需的内容成为一个迫切需要解决的问题。本文以新闻视频为研究对象，完成了以下几个部分：镜头分段、关键帧提取、主题字幕提取、新闻故事分段和新闻视频检索，具体工作如下：

(1) 镜头分段是后续视频处理分析的前提，本文改进了基于双重比较的镜头分段算法，将自适应阈值运用于双重比较法，完成了对新闻视频的突变和渐变镜头检测，具有更高的准确性。由于主题字幕最大限度地反映了新闻的主要内容，本文提出了基于主题字幕帧的关键帧提取算法，提取的关键帧基本上可以代表镜头内容。

(2) 主题字幕提取包含三部分内容。首先根据新闻主题字幕的特点，设计了基于 3/10 时空切片的字幕帧检测算法，取得了较好的检测效果；然后采用基于小波变换和支持向量机的字幕区提取算法，实现了主题字幕的准确定位；最后对主题字幕进行插值放大、二值化、字符分割及 OCR 文字识别，识别效果较好。

(3) 针对新闻结构的特点，设计了基于主题字幕提取的新闻故事分段算法，在主题字幕提取的基础上，通过静音检测、主持人镜头检测，把连续的新闻视频分割成一个个的新闻故事，能够达到令人满意的效果；根据新闻分段结果，以及主题字幕的标注，实现了基于关键字的新闻视频检索。

关键词：镜头分段，关键帧，主题字幕提取，新闻故事分段，新闻视频检索

Abstract

Content-based multimedia retrieval has been a hot research field recent years. As a part, news video retrieval has been studied widely. News video is one of the most important media for users to get the information. It is an urgent problem to retrieve the interested news from a huge amount of news video efficiently and correctly. This paper takes news video as a research object, and discusses several problems in the process of content-based video retrieval, including shot segmentation based on cut detection, key frame selection, topic caption retrieval, news story segmentation, news video retrieval based on keywords.

(1) Video segmentation is the first step, this paper offers a video shot detection method which is based on self-adapting dual-threshold compare, to complete the detection of the cut and the wipe, which achieves good experiment results. Because topic caption contains important semantic information, this paper offers a key frame selection method based on the frame of topic caption, which basically can show the content of the shot.

(2) The retrieval of topic caption contains three parts. First based on analyzing of recent researches, as to the caption detection, the caption detection method is designed, which is based on texture features of 3/10 spatio-temporal slice. Secondly, the caption location method is adopted, which is based on wavelet transformation and SVM and achieves good experiment results. Thirdly, it employs the interpolation method to enhance the caption, then binary and partition the caption, and it uses the given Han Wang OCR software to recognize the characters.

(3) Based on the research of news story segmentation, the method based on caption is discussed to the segmentation of news. Based on the retrieval of the caption, and the detection of slice and anchorperson shot, news video is segmented into a series of news story, and the experimental result indicate that this method is efficient. Based on the result of news story segmentation and the label of topic caption, the news retrieval based on key word is realized.

Keywords: Shot Subsection, Key Frame, Topic Caption Retrieval, News Story Segmentation, News Video Retrieval

声　明

本学位论文是我在导师的指导下取得的研究成果，尽我所知，在本学位论文中，除了加以标注和致谢的部分外，不包含其他人已经发表或公布过的研究成果，也不包含我为获得任何教育机构的学位或学历而使用过的材料。与我一同工作的同事对本学位论文做出的贡献均已在论文中作了明确的说明。

研究生签名：王艳　　2008年7月1日

学位论文使用授权声明

南京理工大学有权保存本学位论文的电子和纸质文档，可以借阅或上网公布本学位论文的部分或全部内容，可以向有关部门或机构送交并授权其保存、借阅或上网公布本学位论文的部分或全部内容。对于保密论文，按保密的有关规定和程序处理。

研究生签名：王艳　　2008年7月1日

1 绪论

1.1 课题的研究背景及意义

随着宽带网络、通信器材、存储设备以及数字电视等多媒体载体及处理设备的快速发展，人们已经可以通过个人电脑、数字电视和 3G 手机，随时随地访问视频。由于多媒体信息具有很大的数据量，而且人们习惯运用高层语义概念来查询和浏览多媒体数据库，因此有必要发展多媒体内容的自动语义分析技术，实现多媒体数据库的建立、管理和检索等。

在多媒体信息检索领域，基于内容的信息检索近年来得到了广泛的研究。基于内容的信息检索是指通过对视频数据从低层到高层进行处理和分析以获取其内容，并根据内容进行检索。涉及到直接根据图像和视频内容的含义，对图像和视频进行有效查询、索引、浏览、搜索与提取。

新闻视频作为一种典型的视频事件，是目前人们获取信息的主要途径之一，它可以提供大量内容丰富、直观、形象生动的新闻节目。但随着人们生活节奏的加快和电视新闻数据量的急剧增加，目前的电视技术已无法满足人们的需要。在目前的条件下，观众往往处于被动地位，必须按照电视台制定的时间表来观看，另外目前的新闻视频只是一次性资源，很难再次利用，如果想检索自己需要的新闻视频，就必须在海量的新闻视频里，利用传统的快进快退的方法来寻找，费时费力，且效率低。为了解决这些矛盾，近年来国内外的一些学者开始探讨建立新闻视频库，实现基于内容的新闻视频检索，以方便不同观众不受时间限制、基于内容的选择，观看自己喜爱的电视新闻节目。

新闻视频有自己的特点，具体而言，新闻节目的播放格式较为固定，如各个新闻故事之间有着比较明确的边界，新闻故事与表述新闻故事的主题字幕一一对应。新闻中的主题字幕，不仅仅是新闻视频内容的概述，也是新闻视频结构中的一个重要标识，可以为基于内容的视频索引的建立，提供丰富的信息。然而人们在观看视频的同时可以轻易地识别并理解这些字幕，但对于计算机而言，它们只是视频流中的一些像素而已，如果采用人工方式来实现这些信息的提取，不仅费时，也无法准确定位字幕所对应的新闻视频的开始和结束点。以上分析可知，新闻视频中主题字幕的自动提取，对于基于内容的新闻视频检索，有着极为重要的意义。

1.2 基于内容的视频检索国内外研究现状

目前，基于内容的视频检索，正逐渐成为国际多媒体界的一个研究热点。关于基于内容的视频分析与检索，前人已经取得了很多研究成果。本节对国内外在这一领域的一

些情况和成果，作一个概要的介绍。

目前为止，国外已研发出多个基于内容的视频检索系统，主要有:

QBIC 系统是由 IBM 公司开发的基于内容的检索系统[1]，此系统提供了对静止图像及视频信息基于内容的检索方法。QBIC 提供了对视频数据检索的能力，其中包括镜头探测、运动估计、层描述和代表帧生成等多种视频处理方法。

JACOB 是意大利 Palermo 大学开发的是一个基于内容的视频查询系统[2]，可进行视频自动分段并从中抽取代表帧，并可按彩色及纹理特征以代表帧描述基于内容的检索。

OVID 是一个采用面向对象技术的视频对象数据库系统，此系统采用了面向对象的视频对象模型，系统中建立了面向对象技术的浏览及查询机制，并设计了一种查询语言 Video SQL;

VideoQ 系统[12]，由美国哥伦比业大学研究实现。它扩充了传统的关键字和主题导航的查询方法，允许用户使用视觉特征和时空关系来检索视频。其主要特征有：文本和视觉搜索综合；自动视频对象分割和跟踪；丰富的视觉特征库，包括颜色、纹理、形状和运动；通过 WWW 互联网交互查询和浏览。

VisualSeek 系统，是美国哥伦比业大学图像和高级电视实验室开发的，它实现了互联网上的基于内容的图像视频检索系统，提供了一套工具供人们在 WEB 上搜索和检索图像和视频。

国内的主要研究单位如清华大学，复旦大学，微软业洲研究院，国防科技大学多媒体研究中心等大学等单位，也开展了基于内容的视频检索技术研究，获得了一定的成果。

Ifind 信息检索系统，是微软业洲研究院张宏江博士所带领的小组研制出的系统，该系统取得的成果最为突出。

TV-FI 系统，Tsinghua Video Find It 是清华大学开发的视频节目管理系统。该系统可以提供视频数据入库、基于内容的浏览、检索等功能，并提供多种数据访问模式，包括基于关键字查询、示例查询、按视频结构浏览及按用户自定义类别进行浏览。

NewVideoCA，是国防科技大学多媒体研究开发中心研制开发的新闻节目浏览检索系统。

MIRC，是国防科技大学系统工程系研制开发的多媒体信息查询和检索系统。

1.3 基于内容的视频检索所面临的问题

基于内容的视频检索方面存在的困难，很大程度上是由视频数据本身所具有的特殊性和复杂性决定的。数字视频作为一种多媒体信息，属于一种非字符值(Non alpha numeric)数据，它与传统数据库系统中的字符数值(Alpha numeric)数据不同，有其特有的特征：

(1) 视频数据有复杂的结构和关系

视频数据与文本及文本图像等类的数据的结构有很大的不同，这主要体现在以下方面：

- 视频数据既有空间属性又有时间属性。文本数据是一种纯字符数值型数据，它不牵涉空间(Spatial)及时间(Temporal)双重性，有时称其为一维数据；图像数据是一种具有空间属性而无时间属性的数据，有时称其为二维数据。然而，视频数据还有一个附加维—时间，因此有时称其为三维数据。
- 视频数据单元之间关系不明确。在文本数据库中，各数据单元之间的关系运算是十分明确的，如可以比较两个数据项之间的相等或不等。然而，视频数据段之间的关系是十分复杂的，而且难以确切地下定义，这给视频数据库的建立及操作，带来许多新的问题。例如，两个视频序列之间的相似运算，就难以有一个普遍可接收的定义。

(2) 视频数据含有丰富的信息内容

动态视频包含了极其丰富的内容，一部电影或电视片可表达一个生动的故事；电视新闻可传达国内外大事；体育比赛视频所表达的可能是一场精彩的球赛等等，这些都是用其它媒体所无法表达的。但是，由于以下几点特征，对视频内容的提取十分困难[6]。

- 视频数据有较高的信息分辨率。所谓信息分辨率，是指某种媒体提供的细节的多少。视频数据随着观察的深入，可逐渐获得一些新的细节。如对于一段描述犯罪现场的视频数据，可从中分辨出犯罪地点、背景、犯罪人、犯罪工具，乃至作案手段等细节。这些细节中蕴涵着丰富的信息内容。
- 视频数据内容的多样性。视频数据作为一种表达信息的媒体，其中所含内容可分为两类：一类视频内容称为信息内容(Information Content)，它是指视频中所含有的语义内容，例如，上述描述犯罪现场的视频数据，信息内容指是何种罪行、犯罪地点及手段等等；另一类视频内容称为声视内容(Audio Visual Content)，它是指视频中所含有的可视及声音的外部表示，如视频中所含的颜色、纹理、物体运动、物体之间的关系等等。信息内容可以通过多种声视内容表达出来。
- 视频内容解释的多样性及模糊性。视频数据是连续播放的图像信息，在图像帧中所含有的信息十分丰富。不同的人对一幅图像或一段视频可能有不同的解释，这就不像字符数值型数据有完全确切的客观的解释，视频数据常常有个人主观的因素，如感情、心理及神经生理等。由于视频数据的模糊性，当对其进行查询时，就无法像字符数值型数据，用一个指定的字段作为确切查询一个特定的记录。在视频数据库中，常常只能用相似性进行查询，即只能用近似匹配对视频数据进行查询。

(3) 视频数据有巨大的数量

视频数据与字符数值类数据不同，它有巨大的数据量。视频数据的巨大数据量给视频的存储、分析、处理和传输带来很大的困难。

1.4 视频中的字幕分析

视频流本质上是由字幕、图像、图形、音频和视频等多态媒质交互融合形成的，每一种模态都表示了丰富的语义信息。作为视频高层语义的一种，包含在视频和视频语音中的字幕信息，是目前唯一能够利用现有技术直接提取而不必通过语义推理的视频高层语义内容。字幕是视频中经常包含的一种重要语义信息来源，其对视频内容有很强的描述作用，对视频的高层语义分析(如视频分类、相关主题搜索)有很重要的意义。

过去视频的字幕信息提取主要靠人工注记来完成。但随着视频信息数据量呈几何级数增长，对海量的视频数据进行人工标注所带来的繁琐性、主观性以及人为错误是不可避免的。日益成熟的光学字符识别技术(Optical Character Recognition，简称 OCR)[50]和语音识别技术(Voice Recognition，简称 VR)[51]，为视频文本信息的自动提取提供了可能。然而对视频流中的字幕进行提取、识别，与普通 OCR 或 VR 相比，存在许多不同之处，要复杂、困难得多。

对视频中的字幕的提取、识别而言，主要表现为：① 视频流数据量巨大，字幕定位比较困难；② 视频流中的字幕往往有比较复杂的背景，且可能出现在任何位置、在任何情况下，虽然也有可能出现在较为纯净的背景上，如新闻中的主题字幕，但多数情况下背景比较复杂，且是动态的；③ 视频流中的字幕分辨率比较低，如果不采取任何措施，即使文字的分割度比较好，将分割后的图像送给一般 OCR 软件不会得到较好的输出。

1.5 本文的主要工作和论文组织

1.5.1 研究的主要内容

本文以新闻视频为研究对象，针对新闻视频的特点和节目制作手段，主要讨论了如下几个问题：镜头分段、关键帧提取、主题字幕提取、新闻故事分段，和基于关键字的新闻视频检索。具体工作如下：

(1) 针对新闻视频固有特征，改进了双重比较法，将自适应双阈值运用于双重比较法，来完成镜头突变和渐变的检测。根据主题字幕出现的帧往往是新闻视频镜头的关键帧，提出了基于主题字幕的关键帧算法。

(2) 针对新闻主题字幕的特点，设计了基于 3/10 时空切片的主题字幕帧提取算法，完成了主题字幕帧的检测；采用基于小波变换和支持向量机(SVM)的字幕区定位方法，实现了对新闻视频中主题字幕的准确定位；对主题字幕区进行插值放大、二值化处理，然后采用投影法对字符进行分割，最后进行 OCR 文字识别。

(3) 针对新闻结构的特点，设计了新闻故事分段算法，在主题字幕提取的基础上，根据静音检测、主持人镜头检测，把连续的新闻视频分割成一个个的新闻故事；根据新

闻分段结果，以及主题字幕的标注，实现了基于关键字的新闻视频检索。

(4) 以 Visual C++6.0 为开发平台，实现了对新闻视频的镜头分段、关键帧的提取、主题字幕的提取、新闻故事的分段与新闻视频的检索。

1.5.2 论文组织

本文内容总共六章，具体安排如下：

第 1 章 绪论。本章介绍了课题的研究背景和意义和基于内容的视频检索国内外研究现状以及所面临的问题，接着分析了视频中的字幕，最后介绍了本文的研究内容和论文组织。

第 2 章 字幕提取技术。本章介绍了字幕提取系统，并对视频中字幕提取的几个关键技术进行了综述。

第 3 章 新闻视频的结构化分析。本章分为三部分内容。第一部分介绍了视频结构化分析的概念。第二部分介绍了常用的镜头分段方法，接着改进了双重比较法，将自适应阈值运用于双重比较法，实验表明，本算法对镜头突变和镜头渐变有较好的检测效果。第三部分介绍了常用的关键帧提取方法，提出了基于主题字幕的关键帧算法，计算简单，提取的关键帧基本上可以代表镜头的内容。

第 4 章 新闻视频中的主题字幕提取。本章分为四部分内容。第一部分介绍了新闻视频中的字幕分类。第二部分是字幕帧检测，设计了基于 3/10 时空切片的主题字幕帧提取法，通过对切片的纹理预处理、能量模型分析，确定字幕帧边界。第三部分是主题字幕区定位，采用了基于小波变换和支持向量机(SVM)的字幕区定位方法，实现字幕区提取。第四部分是主题字幕的二值化及识别，采用双三次插值法增强字幕、Otsu 法二值化字幕、投影法分割字符，并对分割时产生的字符错分作相应的处理，最后采用汉王 OCR 软件进行识别，具有较好的识别效果。

第 5 章 基于主题字幕提取的新闻故事分段和视频检索。本章分为两部分。第一部分设计了新闻故事分段算法，在主题字幕提取的基础上，根据静音检测、主持人镜头检测，把连续的新闻视频分割成一个个的新闻故事。第二部分介绍了新闻视频检索的概念、本新闻视频检索系统的数据库，完成了基于关键字的新闻视频检索，用户可以通过输入关键字，实现对相关日期、相关新闻故事的检索，达到了预期的检索效果。

第 6 章 总结。本章对研究内容进行了总结，并提出了进一步的工作方向。

2 字幕提取技术

本章介绍了字幕提取系统，和字幕提取中的几个关键技术，包括字幕帧检测技术、字幕区提取技术、字幕增强技术与字幕分割技术。

2.1 字幕提取系统

字幕提取系统一般由字幕帧检测、字幕区提取、字幕分割、字幕增强和字符识别(OCR)五个模块组成[41]，如图 2.1.1 所示。

字幕帧检测，判断某帧图像是否有字幕存在，字幕帧检测通常用于图像序列中。字幕区提取，就是把字幕中的字符从背景中分割出来，同时产生一个包络字幕的框。尽管字幕的包络框可以给出字幕的精确位置，但是为了便于字幕的识别，还需把字幕从背景中分割出来。由于字幕区域的分辨率可能较低和噪声等原因，提取的字幕图像在输入到 OCR 前必须对它进行增强。

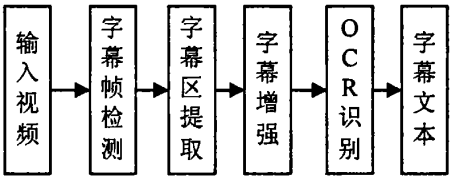


图 2.1.1 视频字幕提取系统结构示意图

由图 2.1.1，视频中字幕的提取主要包括：字幕帧检测、字幕区提取、字幕增强及 OCR 识别等关键模块。由于目前有比较成熟的 OCR 软件(如汉王、尚书 OCR 识别)可直接应用于字符的识别，无需自己再另外开发。因此本节只对前四种关键技术[41]进行回顾和总结。

2.2 字幕帧检测技术

在视频语义提取和索引建立过程中，镜头分段和字幕事件检测往往是同时进行而且互相关联的，字幕事件可能在一个镜头内完成，也可能经历若干个镜头，字幕的区域边界帧可能出现在视频帧中任意位置。

文献[30]利用FCNN的模糊聚类特性，依据帧间差提取出的特性进行视频镜头分段，作为视频分析的基础；然后在镜头内及镜头边界处用量化空域差分密度(Quantized Spatial Difference Density，QSDD)作为特征来判断字幕出现和消失帧。

文献[15]用场景转换检测方法从视频中选择一帧作为包含字幕的候选帧，在场景图像中每隔2秒选取一帧作为含有字幕的候选图像，这种方法适用于视频索引，因为视频

索引只需要从视频中提取关键词，而不需要得到所有完整的字幕。

文献[16]提出了利用MPEG视频P-帧和B-帧图像中内部编码块的数目，来检测字幕出现和消失时的图像帧，算法思想是基于当字幕出现或消失时，其所对应的块常常是内部编码块的假设，此算法没有利用视频中I-帧图像的信息，而且受场景变化和运动的影响。

文献[17]采用模糊聚类神经网络分类器对字幕出现和消失的视频帧进行检测，根据相邻二帧之间局部图像块欧氏直方图的差值检测字幕事件[19]。

文献[18]提出了一种字幕检测方法，他们首先对连续30帧图像的时间特征矢量(TFV，Temporal Feature Vector)进行监督分类，得到一系列二值图像，在每个二值图像中，值为1的像素表示其属于字幕区域或者它是有着与字幕像素相似特征的噪声点；在一系列二值图像中连续二帧图像相减后值为1的点数形成二条曲线：前帧减后帧产生一条，后帧减前帧产生一条，根据这二条曲线的峰点确定字幕出现和消失的帧。用含有中文字幕的美国电影中的连续10000帧图像进行测试，字幕出现帧的检测率为98.46%，字幕消失帧的检测率为89.23%。

从上述文献分析中可以看出，字幕检测的技术主要有两种，一种是根据字幕像素在时间域变化来检测字幕出现或消失的帧，这种方法一般只适用于检测在时间域上位置、形状和大小不变的字幕(例如视频中的字幕和标题)，它的缺点是比较容易受场景转换的影响，它的优点是可以使用字幕出现或消失时的差影图像进行后续的字幕定位，一般地，字幕出现或消失时差影图像的字幕区域比原始图像更加显著，在差影图像中进行字幕定位会更加容易；另一种是用一种快速的字幕定位算法，例如压缩域字幕定位算法[20]来进行字幕检测，这种方法实际上是在字幕提取系统中省去了字幕检测步骤。

2.3 字幕区提取技术

字幕区提取大致可以分为基于区域的方法、基于边缘或梯度的方法、基于纹理的方法、基于学习的方法、基于时空分布特征的方法和压缩域方法等六类。

(1) 基于区域的方法

这种方法中，根据字幕中的字符具有相同或相似的颜色或灰度级，采用自下而上的方式，先提取字符或字符的笔画，然后，根据字幕中字符的空间排列，合并这些字符以形成字幕区域。文献[21]用形态学运算确定背景区域的种子像素，再对种子点执行区域生长算法，以得到前景区域，这些前景区域就是字符的候选区域，用多尺度 K-Mean 算法对每个字符的候选区域进行色彩聚类，然后进行连通元形态分析得到真正的字符。基于区域的方法对于字幕字符较大、字幕与背景的对比度较大、分辨率较高、同一行字幕中字符的颜色变化较小的图像，字幕定位的效果比较好，而对字幕字符较小、字幕与背景的对比度较小、分辨率较低，同一行字幕中字符的颜色变化较大的图像，效果不够理

想。

(2) 基于边缘和梯度的方法

这种方法是根据字幕区域边缘比较密集设计的：字幕笔画与背景的颜色或亮度之间具有的较高对比度，它们之间产生的边缘较为明显，而字幕一般由多个字符组成，而且这些字符的间距相对较小。文献[22]对边缘图像进行水平和垂直投影，然后，根据投影直方图曲线及其差分曲线确定字幕的位置和大小，与其它算法不同的是他们对投影直方图曲线及其差分曲线的分析处理。对 175 幅原图像进行测试，算法结果检测的正确率为 85.62%，recall 率为 84.94%。当图像中存在较多具有强边缘的对象时，基于边缘和梯度的方法产生的虚警率较高，字幕定位的精度也不高。

(3) 基于纹理的方法

这种方法中利用图像中的字幕有着与背景不同的纹理特性，来决定一个像素点或图像块是否属于字幕区域，Gabor 滤波、小波、FFT 和空间变化等技术都可用于检测图像中字幕区域的纹理特性。文献[23]采用小波变换方法定位字幕，Harr 小波分解系数用于计算图像的局部能量变化，阈值化局部能量变化得到二值图像，然后用尺寸和高宽比等几何属性对二值图像进行连通区域滤波，最后把各个尺度检测到的字幕区域合并产生最终结果。基于纹理的方法可以提取不同分辨率图像中不同尺寸、不同语言和不同字体的字幕，具有一定的通用性，但存在着计算量大和定位精度不高的缺点。

(4) 基于学习的方法

在使用基于纹理的方法进行字幕定位时，人为地构造一个适于各种情况的纹理分类器比较困难，为此，人们提出了几种基于学习自动产生纹理分类器的方法。文献[24]等研究人员使用基于学习的纹理鉴别方法分离文档中的字幕、行间空白和图像。文献[25]使用支持向量机(SVM)来分析图像中字幕的纹理属性，其输入的结构与文献[24]方法的输入结构相同，SVM 能够在其构架内部结合一个特征提取器，而且在高维空间内也能产生好的结果，在用 SVM 进行纹理分类后，再用投影分析来提取字幕行。文献[26]的方法与其它基于小波、FFT 和 Gabor 特征提取的方法不同，没有明显的特征提取步骤，神经网络根据输入的颜色判断字幕是否存在，然后用投影分析方法从神经网络的输出中提取字幕的包络框，算法处理大小为320×240的图像需 11.3 秒，定位率为 92.2%。基于学习的方法可以提取不同分辨率图像中不同尺寸、不同语言和不同字体的字幕，具有通用性，效果也比较好。但是，它有一个缺点，就是用学习机对图像进行分割的结果受训练样本集与测试样本集的相似程度的影响。

(5) 基于时空分布特征的方法

这种方法中是利用字幕的时域不变性和字幕与背景具有较强的对比度等空域特性对字幕和标题进行定位。字幕的时域不变性是指视频中的同一字幕或标题通常会在连续的多帧中出现，且它们的位置、形状和尺寸在时间域上几乎不变。文献[19]根据相邻二

帧之间局部图像块欧氏直方图的差值检测文字事件，然后，用 Sobel 算子提取边缘，根据边缘的尺寸对边缘图像进行滤波，接着，用像素的密度对边缘图像进一步滤波，最后，在边缘图像上用投影方法提取字幕区域，用新闻视频进行测试，算法的字幕提取正确率超过 85%。但是，此算法没有利用字幕在连续多帧中存在的信息进一步提高字幕提取的性能。时空方法仅适用于位置、形状和尺寸在时间域上几乎不变的字幕，而且它还容易受场景转换检测结果的影响。

(6) 基于压缩域的方法

压缩域方法处理经过压缩的图像或视频，它 像或视频，而直接在 DCT 等压缩域中对图像或视频字幕进行定位，算法具有很快的速度。文献[27]根据 I-帧的纹理能量和 B-帧或 P-帧的运动能量来初步定位字幕，接着在初步定位的字幕区域内进行彩色空间分割得到一系列二值子图像，在每个二值子图像中进行连通成分分析和排列分析得到字幕行，最后，在时域上用多帧验证方法除去虚警。用美国和台湾的新闻视频进行测试，算法字幕检测的平均精度为 71.6%，Recall 率为 96.9%。压缩域方法最大的优点是速度非常快，可以进行实时处理。算法的缺点是容易漏检具有以下五种情况的字幕：(1)字符的尺寸较大；(2)字符间隙较大；(3)字幕中的字符较少；(4)字幕与背景的对比度较小；(5)在视频中存在时间非常短的字幕。当一些非字幕区域具有与字幕相似的纹理时，算法容易产生虚警，而且得到的字幕包围框的位置精度较低。

综合上述各种字幕定位方法，可以看出视频中字幕定位一般由特征提取、特征分类、特征聚集、候选字幕区域提取和字幕区域验证等五个步骤组成，首先，选择某个或某些能够把字幕与背景区别开来的字幕特征，其次，采用某种算法提取字幕特征，接着，聚集空间相邻的特征点形成区域，然后，用字幕的另一些特征除去一些不可能是字幕的区域得到候选字幕区域，最后，再用字幕的一些特征对候选字幕区域进行验证以得到真正的字幕区域。

2.4 字幕增强技术

在视频图像中，字幕的分辨率较低，字幕与背景的对比度较低，字幕背景比较复杂，字幕增强的目的是提高字幕的分辨率，提高字幕与背景的对比度，降低背景的复杂度，字幕增强方法包括单帧字幕区域增强和多帧字幕区域增强。

文献[28]用双线性插值的方法来提高字幕区域的分辨率，用多帧平均的方法来提高字幕与背景的对比度，同时降低字幕背景的复杂度，他们的多帧平均方法首先选取参考字幕区域，然后，用基于方差和(Sum of Square Digerence)的图像匹配方法跟踪登记字幕区域，最后，通过平均方法融合登记的字幕区域。

文献[29]也使用线性插值的方法来提高字幕区域的分辨率，对含有相同字幕的多帧图像采用“与”操作来最小化背景的变化，用直方图拉伸的方法增强字幕区域的对比

度。

文献[55]考虑到同一字幕出现在连续的视频帧中，在检测出了字幕的出现和消失帧后，采用多帧平均来减小复杂背景的影响。

$$\bar{\gamma}_i = \frac{1}{|C_i|}\sum_{f_i \in C_i} \gamma_i(f_i) \tag{2.1}$$

其中$|C_i|$为字幕事件C_i中的视频总帧数。因为背景比字幕变化快，所以平均的方法能平滑复杂背景的变化。

文献[56]用高对比度帧平均(High Contrast Frame averaging)和高对比度块平均(High Contrast Block Averaging)的方法来增强字幕，他们用字幕区域周围黑像素所占的比例来判断字幕区域的对比度是否高，多帧平均方法只选取那些高对比度的字幕区域进行平均，高对比度块平均方法先把字幕区域分成较小的块，然后，只选取那些清晰的块进行多帧平均。

字幕增强技术包括单帧字幕区域增强和多帧字幕区域增强二类，对于单帧字幕区域，可以用双线性插值的方法来提高字幕区域的分辨率，也可以用直方图拉伸的方法提高字幕区域的对比度，对于多帧字幕区域，可以用求多帧平均、或求各帧对应像素灰度值的最大或最小值来提高字幕与背景的对比度。

2.5 字幕分割技术

字幕分割的目标是从字幕区域提取一个个字符，它的输出是可以输入到 OCR 系统中的二值字符图像。字幕分割主要包含阈值法、聚类法、基于区域的方法、投影法等。

文献[34]在前人的基础上，将 FCM 算法应用到字幕的分割中去，设定迭代停止阈值ε，初始化原型模式$p^{(0)}$，设置迭代计数器$b=0$，由式(2.2)计算更新分类矩阵$U^{(b)}$对于$\forall i,k$，如果$\exists d_{ik}^{(b)} > 0$，则有

$$\mu_{ik}^{(b)} = \left\{ \sum_{j=1}^{c} \left[\frac{d_{ik}^{(b)}}{d_{jk}^{(b)}} \right]^{\frac{2}{m-1}} \right\}^{-1} \tag{2.2}$$

用式(2.3)更新聚类原型模式矩阵$p^{(b+1)}$

$$p_i^{(b+1)} = \frac{\sum_{k=1}^{n} (\mu_{ik}^{(b)})^m \cdot x_k}{\sum_{k=1}^{n} (\mu_{ik}^{(b)})^m} \tag{2.3}$$

如果$\left\| p^{(b)} - p^{(b+1)} \right\| < \varepsilon$，则算法停止并输出划分矩阵和聚类原型，否则令 b=b+1，转向公式(2.2)。由实现算法不难看出，整个计算过程就是反复修改聚类中心和分类矩阵的过程，因此常称这种方法为动态聚类或者逐步聚类法。

文献[59]提出了一种基于区域的字幕提取方法，此算法先用对称邻域滤波器SNF(Symmetric Neighborhood Filter)来增强图像，然后用分级连通区域分析HCCA(Hierarchical Connected Component Analysis)来把字符从背景中分离出来。HCCA把经SNF增强的图像分割为一系列连通区域，在第一级分析中，考虑像素级的连通性，在其它的级中，考虑区域级的连通性。

文献[60]提出了一种基于投影的字幕分割方法，该方法首先采用多帧平均和最小像素搜索法增强字幕区域，其次，使用简单的阈值二值化字幕区域，接着，用数学形态学操作来检测字符的轮廓图像，然后，采用投影方法把字幕区域分离为单个字符图像，最后，使用普通的阈值方法二值化每个字符图像。

2.6 本章小结

本章首先了介绍了字幕提取系统，接着对视频中字幕提取的几个关键技术进行综述，包括：字幕帧检测技术、字幕区提取技术、字幕增强技术与字幕分割技术。

3 新闻视频的结构化分析

人们普遍认为，视频结构的模型化或形式化，是解决基于内容视频检索问题的关键。为此，现有的基于内容视频检索系统，一般先对视频进行结构化分析，包括镜头分段和关键帧提取等，从而为视频检索做好预处理工作。

3.1 视频结构化分析的概述

视频是一种非结构化的数据流，由连续的图像组成，它比文本、图片包含更丰富的信息，但是却无法像文本那样，直接地给出它的内容或者直接地进行内容的比较，也不便于直接管理和检索。可以将视频数据按照由细到粗的顺序，划分为 4 个层次结构：图像帧、镜头、场景和视频，视频结构化流程示意图如图 3.1.1。

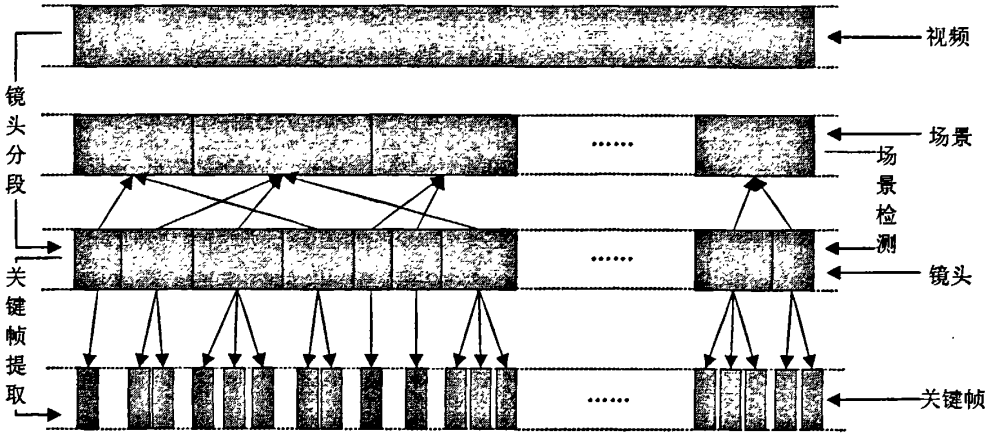


图 3.1.1 视频流结构化流程示意图

其中：

帧是视频流的基本组成单元，每一帧均可看成一个独立的图像。视频流数据就是由连续的图像帧构成，它是构成视频序列的最小单位。

镜头是指一系列连续纪录的图像帧，用于表示一个时间段或相同地点连续的动作。它是视频流进一步结构化的基础结构层。镜头一般是由摄像机一次摄像的开始和结束的所有帧构成，表示一个物理概念。

关键帧有时也称为代表帧，用以描述一个镜头的关键图像帧，它可以用来代表一个镜头的主要内容。关键帧的使用，大大减少了视频索引的数据量，同时也为视频摘要和检索提供了一个组织框架。

场景是指一连串语义上相关和时间上相邻的镜头，它们一般发生在相同的时间和地点，出现相同的人物或事件。场景是视频所蕴涵的高层抽象概念和语义的表达，表

示的是一个语义概念。

如图 3.1.1，连续的视频图像帧，通过视频镜头边缘检测，被分割成长短不一的镜头单元；然后对每个镜头单元提取关键帧，得到可以表征每个镜头单元的关键帧；最后将语义上相关且时间上相近的若干镜头组织成场景。在新闻视频中，场景也称为新闻故事。由于这种表示方法更符合人对事物的认知方式，故越来越得到重视，而有关这方面的研究也将越来越深入。

3.2 镜头分段

视频是由一系列镜头构成的，镜头是视频中最小的物理单元。将视频拆分出镜头是视频结构化的基础，也是视频分析和检索工作中的首要任务，因此这是整个视频检索工作的第一步。本节实现对视频镜头的分段。

3.2.1 镜头分段的概念

在各种自动视频分析方法中，镜头分段 (也称为视频分段)是基础之一，它是指将一系列经过镜头编辑的连续视频。拆分还原为单独的物理镜头(shot)。镜头是视频的基本单元，它对应摄像机的一次拍摄动作。镜头分段的结果正确与否，将直接影响到后续分析步骤的准确性。视频分段是通过检测和识别镜头间的转换模式来实现的。

镜头间的转换模式可以分为两类：突变(abrupt transition)和渐变(gradual transition)。突变也叫切变，是指一个镜头与另一个镜头之间没有任何过渡，两个镜头的切换发生在连续的两帧 k 和 $k+1$ 上；渐变则是指加入了一些空间或时间上的编辑效果，前一个镜头逐步转换为下一个镜头，通常可能延续几帧到几十帧。渐变的类型较多，常见的有淡入(fade in)、淡出(fade out)、溶解(dissolve)和滑变(wipe)等。淡入是指画面不断加强；淡出是指画面逐渐减弱直到消失；溶解是指在上一个镜头画面逐渐减弱的同时，下一个镜头的画面逐渐加强；滑变是指从画面的某一部分开始，逐渐被下一个镜头的画面所代替。

(a) 突变效果

(b) 滑变效果

图 3.2.1 两种镜头转换编辑模式

图 3.2.1 以新闻视频为素材，给出了两种镜头转换模式的例子，其中图(a)为镜头突变，从图中可以看出，下一个镜头的首帧直接取代了上一个镜头的末帧；图(b)为镜头渐变中的滑变效果，它表现为下一个镜头的首帧，从图像的一角或者一边平稳地逐渐取代当前镜头的最后一帧。

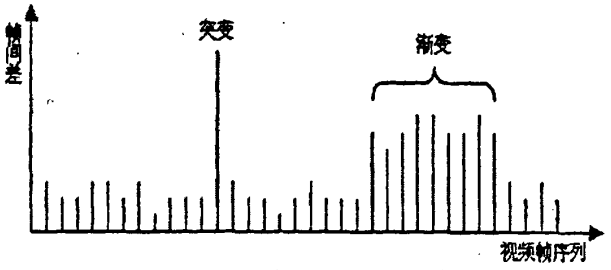


图 3.2.2 两种镜头转换编辑模式

镜头分段的基本方法是比较视频帧间的差异，根据帧间差得到一系列判断数据，再根据这些数据是否超过某个阈值的方法，将镜头转换检测出来。当发生镜头突变时，帧间的差异往往表现为一个单独突起的峰值；而发生镜头渐变时，帧间差往往表现为一系列比较突起的峰值，但是远远没有镜头突变那样明显，而且往往和视频对象作剧烈运动时，产生的帧间差有比较类似的表征。图 3.2.2 为发生镜头突变和镜头渐变时的帧间差示意图。

3.2.2 常用的镜头分段方法

现有的镜头分段算法大致可分为突变检测算法和渐变检测算法两大类，本节将分别介绍突变和渐变的几种检测算法。

3.2.2.1 镜头突变的检测

镜头突变的特点，是人的视觉上可以感觉到一个突然的变化，同样在帧间差别比较上会出现一个尖峰(peak)。因此，突变镜头检测的基本思想，是寻找较好的帧间差别比较方法来检出尖峰，同时减少与其他镜头变换方式间的误检测。目前，常用的方法主要有以下几种：

(1) 模板匹配的方法

该方法对两帧图像的对应像素灰度值或颜色值进行比较，并把差的绝对值之和作为帧间差。当图像大小为 $M \times N$ 时，其计算公式如下：

$$d(I_i, I_j) = \sum_{x=0,y=0}^{x<M,y<N} \left| I_i(x,y) - I_j(x,y) \right| \tag{3.1}$$

若前后两帧的差值超过一定阈值 T，则认为有突变发生。这种模板匹配的方法计算简单，但由于其与像素的位置密切相关，因此对噪声和物体运动很敏感，单独使用易造成误检测。

(2) 统计量的方法

为克服上述基于模板匹配的方法的不足，人们提出了基于统计量的方法，其基本思想是，首先计算视频帧窗口内灰度均值和方差等统计量，然后用这些统计量来比较前后两帧的变化，用以检测镜头突变。如文献[9]等人提出：把各帧划分为8×8像素的小块，并对每个块取平均值，再用这个平均值对前后帧的对应小块进行比较，这种方法可以去掉部分噪声，并对小的物体运动和摄像机运动进行补偿。

(3) 灰度和彩色直方图的方法

灰度和彩色直方图使用的是像素亮度、色彩统计值，抗噪能力比模板匹配强，是使用最多的帧间差计算方法。若两视频帧I_i、I_j的直方图为H_i和H_j，则彩色直方图匹配的计算方法如下式所示：

$$d(I_i,I_j)=\sum_{k=1}^{n}\left|H_i(k)-H_j(k)\right| \tag{3.2}$$

由于直方图丢失了彩色的位置信息，有时会出现两幅图像内容完全不同，但直方图相似的情况。为此研究人员提出了许多改进的方法。如文献[10]提出了一种基于子块划分与匹配的方法：首先将视频帧分割成4×4子块，然后对两帧相应子块进行比较，废弃差别最多的一对，其余的比较结果参与最后的识别。由于子块直方图在一定程度上反映了颜色的位置特征，因而对物体运动、摄像机运动、镜头缩放等有更好的适应性。

(4) 压缩域的方法

压缩域方法的特点是速度快，并且可以直接利用某些压缩过程中已提取的信息，如运动矢量等。其基本思想是：视频经过离散余弦变换(DCT)后，DCT系数的直流分量(DC)系数集中了视频帧的主要基本信息，用DC系数为主反变换得到的压缩图，也就是DC图像，可以比较正确的反映视频全景的变化过程。通过对DC图像序列的测度计算、判决分析，可达到镜头转换检测的目的。如文献[11]利用MPEG中的DC系数来计算帧间差；文献[13]则利用MPEG域中帧的DC系数计算亮度直方图；文献[9]等人则利用MPEG压缩视频中的DCT块和运动矢量来进行镜头边界检测。

3.2.2.2 镜头渐变的检测

由于参与镜头渐变的两个镜头之间的转换是缓慢进行的，帧间差虽然存在，但不会出现明显的峰值。所以镜头渐变比较难于检测，并且渐变过程的起止点检测也是一个难点。为此，人们通过研究镜头渐变的各种特点，提出了许多有针对性的算法。

(1) 基于图像边缘的方法

该方法的基本原理是“在发生镜头转换时，新出现的边缘应远离旧边缘的位置，同样旧边缘消失的位置应远离新边缘的位置”[11]。

首先对前后连续两帧图像I_i、I_j进行边缘检测得到二值图像E_i和E_j，并对E_i和E_j进行加宽得到$E_i^{'}$和$E_j^{'}$，然后进行边缘变换计算出d_{in}和d_{out}：

$$d_{in}=1-\frac{\sum_{x,y}E_i(x+\delta x,y+\delta y)E'_j(x,y)}{\sum_{x,y}E_i(x,y)} \tag{3.3}$$

$$d_{out}=1-\frac{\sum_{x,y}E'_i(x+\delta x,y+\delta y)E_j(x,y)}{\sum_{x,y}E_i(x+\delta x,y+\delta y)} \tag{3.4}$$

其中 d_{in} 为进入像素(新出现并远离旧边缘的像素点)所占比例，d_{out} 为退出像素(新消失并远离新边缘的像素点)所占比例。两帧图像之间的差异由下式计算：

$$diff=\max(d_{in},d_{out}) \tag{3.5}$$

如果 *diff* 大于某一个设定的阈值，则认为出现了镜头转换。

(2) 基于数学模型的方法

前面所介绍的镜头转换识别方法，都是完全基于图像处理技术的，主要是利用帧间差自下而上来进行镜头转换检测。由于其忽略了渐变转换过程中，帧之间结构上的相关性，因此对渐变检测效果并不是很好。于是，人们通过研究镜头渐变过程中，帧之间结构上的相关性，提出了基于模型的自上而下的检测方法。例如对于纯"颜色变换型"渐变中，视频剪辑是通过对某镜头的最后一帧(淡入、淡出)或同时对两个镜头的结束帧和开始帧进行颜色变换操作来实现的。如果假设淡入具有正的颜色变化率，淡出具有负的颜色变化率，而隐现则同时具有正、负两个颜色变化率。据此可建渐变数学模型[9]，如式3.6：

$$f(x,y,t)=\alpha(t)g_1(x,y,t)+\beta(t)g_2(x,y,t) \tag{3.6}$$

其中，$g_1(x,y,t)$ 是即将逐渐消失的镜头，$g_2(x,y,t)$ 是即将逐渐出现的镜头。如果镜头内没有运动或运动很小，则可以分别记为：$g_1(x,y,t)\cong g_1(x,y)$，$g_2(x,y,t)\cong g_2(x,y)$，$\alpha(t)$ 和 $\beta(t)$ 都是时间的线性函数，假设渐变转换的持续时间为 0 到 T。对于慢转换，它们可以表示为：

$$\alpha(t)=\begin{cases}1 & ,\ t<0\\ 1-t/T & ,0\le t\le T\\ 0 & ,\ t>T\end{cases} \tag{3.7}$$

$$\beta(t)=1-\alpha(t) \tag{3.8}$$

对于淡出，则 $g_2=0$；对于淡入，则 $g_1=0$。在变化的过程中，每幅图像上所有的像素都是以线性规律变化。因此图像 $f(x,y,t)$ 对时间的微分为一幅常量图(constant image)。根据这一性质，可以通过检测常量图来检测渐变过程。

(3) 基于模式识别的方法[48]

该方法的基本思路是：首先提出一种模式模板的概念用于对各种渐变转换建立模式模板库；然后在模式模板库的基础上设计了一个通用的匹配算法；最后，通过对匹配结果进行 Hough 变换，检测转换区位置和识别转换模式。该方法对运动不敏感，具有很好

的鲁棒性，对目前存在的各种渐变镜头中的 wipe 转换模式的检测和识别，都有很好的效果。

3.2.3 一种改进的双重比较镜头分段算法

3.2.3.1 改进的双重比较法

文献[31]提出的双重比较法采用两个阈值来检测镜头转换的发生，是一种较实用的镜头转换检测方法。双重比较法的示意图如图3.2.3，横坐标是视频帧，纵坐标是帧间差，帧间差高于高阈值时是镜头突变；帧间差大于低阈值且小于高阈值时，认为是渐变的开始，累计帧间直方图差，当累计的差值大于高阈值时，则认为是潜在的镜头渐变的结束。

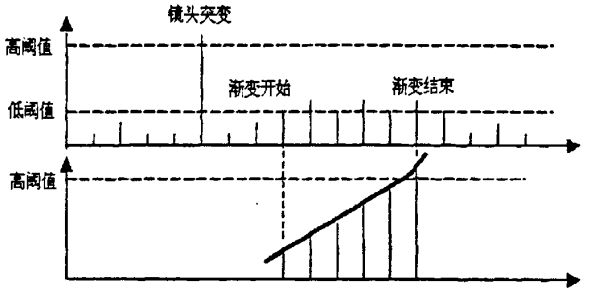


图 3.2.3 双重比较法思想示意图

双重比较法采用固定阈值法。在不同的视频中，镜头转换时产生的帧差大小不会相同，甚至相差很大；即使在同一段视频里，镜头转换时产生的帧差变化范围也很大。因此固定阈值法不能很好的检测出镜头分段点。例如某些时间段里视频对象或摄像机运动比较频繁，从而直方图差值都普遍比较大，该方法可能会造成较多的误检测。为此，本文对双重比较法进行了改进，将自适应阈值应用于双重比较法，实验结果表明，该方法具有较好的分段效果。

同一个镜头内的视频帧内容差不多，帧间差的大小也差不多，在该镜头所有帧间差平均值处上下波动，而镜头边界处的帧间差，要明显大于此镜头的帧间差平均值。因此帧间差相对平均值较大的视频帧就是镜头边界帧。

据此，本文提出的自适应阈值算法如下：首先计算当前六帧的帧间直方图差值，得到五个帧间差值，然后取这五个帧间差的平均值 $Mean$，用 $Mean$ 作为自适应阈值的参考基准，分别求取高低阈值 λ_{high} 和 λ_{low}。

$$\begin{aligned} \lambda_{high} &= k_1 \times Mean \\ \lambda_{low} &= k_2 \times Mean \end{aligned} \tag{3.9}$$

其中 k_1 和 k_2 为比例因子，根据实验，将 k_1 取为6，k_2 取为1.5较为合适。

当镜头内运动强度小时，帧间差比较小，从而计算出来的高阈值 λ_{high} 比较小，此时若镜头内包含字幕边界帧(字幕起始帧、消失帧)，例如字幕起始帧与其前一帧的直方图帧差特别大，容易将字幕起始帧误检为镜头边界。因此当 λ_{high} 低于某固定阈值时，根据

经验，设定高低阈值 λ_{high} 和 λ_{low}。

3.2.3.2 算法步骤

本文将自适应阈值运用于双重比较法，即基于自适应阈值的双重比较法，具体算法步骤如下：

Step1. 初始化，总帧数是 *TotalNum*，帧数 $k=1$。

Step2. 若k=*TotalNum*，转到Step5；否则 $k=k+1$，计算帧 k 到帧 $(k+5)$ 这六帧的帧间直方图差的平均值 *Mean*，按照3.2.3.1提出的方法计算高低阈值 λ_{high} 和 λ_{low}，$k=k+6$。

Step3. 计算当前帧 k 它与前一帧的帧差 $d(I_{k-1},I_k)$。

Step4. 如果 $\lambda_{low}<d(I_{k-1},I_k)<\lambda_{high}$，则累计帧间直方图差 $A_c(j)=\sum_{j=k}^{n}d(I_j,I_{j+1})$。 若不满足 $\lambda_{low}<d(I_{k-1},I_k)<\lambda_{high}$ ，分为三种情况：若满足 $A_c(j)>\lambda_{high}$，则认为发生渐变，记录渐变位置，转到Step2；若不满足 $A_c(j)>\lambda_{high}$，满足 $d(I_{k-1},I_k)>\lambda_{high}$，则认为发生突变，记录突变位置，转到Step2；若不满足 $A_c(j)>\lambda_{high}$，也不满足 $d(I_{k-1},I_k)>\lambda_{high}$，转到Step2。

Step5. 程序结束，镜头分段完毕。

程序流程图如下图3.2.4：

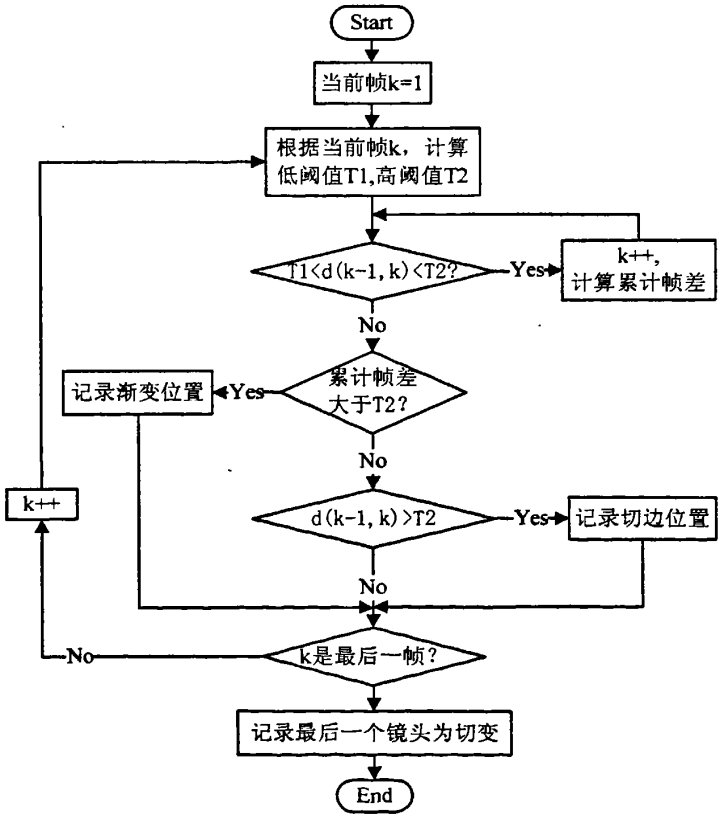
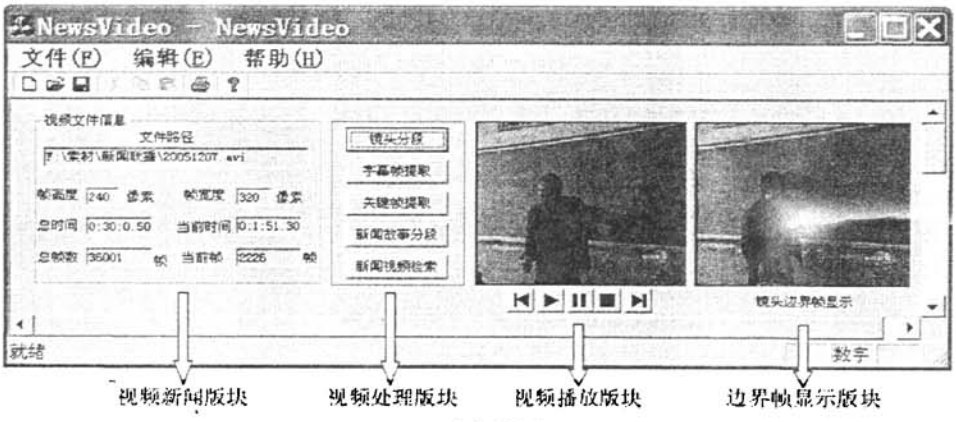


图3.2.4 镜头分段流程图

3.2.4 实验结果

镜头分段界面如图 3.2.5。在介绍镜头分段界面之前，先简要说明本文完成的新闻视频原型系统 NewsVideo 的界面。如图 3.2.5(a)，分为四个版块：视频信息版块，显示了视频文件的路径、帧高度、帧宽度、总时间、总帧数、当前时间和当前帧；视频处理版块，包括镜头分段、字幕帧提取、关键帧提取、新闻故事分段和新闻视频检索这五个功能；视频播放版块，实现对载入的视频播放、暂停等操作；边界帧显示版块，实现了对镜头分段时镜头边界帧的实时显示。本文将分别介绍视频处理版块实现的各个功能。

图 3.2.5 中，图(a)是突变实例，图(b)是渐变实例。选取的素材是长达三十分钟的新闻联播节目“20051207”，当按下“镜头分段”按钮，边界帧显示版块，将实时显示镜头分段时的镜头边界帧，所谓镜头边界帧，是指当前镜头的第一帧。



(a)突变实例

(b)渐变实例

图 3.2.5 镜头分段界面

将中央一台的新闻联播、新闻三十分以及中央五台的体育新闻作为新闻素材，选取其中五段具有代表性的视频进行实验，对突变和渐变进行检测。本文采用查全率和查准

率，衡量镜头分段的效果。查全率和查准率的定义如下：

$$(1)\text{查全率} = \frac{\text{正确检测数}}{\text{正确检测数} + \text{漏检数}}$$

$$(2)\text{查准率} = \frac{\text{正确检测数}}{\text{正确检测数} + \text{误检数}}$$

实验证明具有很好的效果，实验结果如表 3.2.1。由表可知，误判的情况很少，查准率比较高，达到 99%，查全率也达到 90%以上。漏检的主要原因是镜头分段点前后两帧直方图结构比较相似，而误检的原因是闪光灯造成的。另外，由于要计算自适应阈值，系统的实时性比较差，有望进一步改进。

表 3.2.1 镜头检测实验结果

	帧数	突变镜头 / 滑变镜头	检测到	漏检测	误检测	查全率	查准率
视频片断 1	2607	18 / 2	18 / 2	0 / 0	0 / 0	100%	100%
视频片断 2	1641	30 / 3	27 / 2	3 / 1	0 / 0	87.88%	100%
视频片断 3	1726	26 / 0	24 / 0	2 / 0	1 / 0	92.31%	96%
视频片断 4	2693	21 / 4	21 / 4	0 / 0	0 / 0	100%	100%
视频片断 5	2570	18 / 6	17 / 5	1 / 1	0 / 0	91.67%	100%
总计	8444	113/15	107/13	6/2	1 / 0	93.75%	99.17%

3.3 关键帧提取

关键帧(key frame)，有时也称为代表帧，是用于描述一个镜头的关键图像帧，它通常会反映一个镜头的主要内容。在镜头分段的基础上可对每个镜头提取关键帧[46]，并用关键帧简洁地表达镜头。

3.3.1 常用的关键帧提取方法

视频序列分割成镜头后，需要对镜头提取关键帧，因为每个镜头都是在同一场景下拍摄的，同一个镜头中的各帧图像有相当重复的信息。一个镜头的关键帧，就是反映该镜头中主要信息内容的一帧图像或若干帧图像。形象地说，所有的关键帧组成了一部生动的“连环画”。视频数据量巨大，在存储容量有限的情况下，通常仅存储镜头的关键帧，也可以收到数据压缩的效果。另外，关键帧代表镜头，使得对视频镜头可用图像的技术进行检索。

针对关键帧的特点，选取时有两个基本要求：第一，所选帧必须能够反映镜头中的主要事件，描述应尽可能准确完全；第二，为了便于管理，数据处理量应尽量小，计算不宜过于复杂。目前，关键帧提取的方法有很多种：

(1) 基于镜头边界法：将每个镜头的首帧作为关键帧[52]。这是由于同一镜头中后面各帧，可以看作是第一帧在逻辑上和时间上的扩展。但有时镜头内容变化较大，第一帧并不能很好地代表镜头的内容，所以在此基础上，一种改进的方法就是，把每个镜头的第一帧和最后一帧或中间某帧直接作为关键帧选取出来，这样实现起来较为简单，运算量小，适合内容活动性小或保持不变的镜头，但对于摄像机不断运动的镜头，该方法不稳定，无法有效表达镜头的主要内容。

(2) 基于颜色特征的方法[47]：镜头当前帧与最后一个关键帧比较，如有较多内容被改变，则当前帧为新的一个关键帧。但基于颜色特征的方法对摄像机的运动(如摄像机镜头拉伸，造成焦距的变化及摄像机镜头平移的转变)很不敏感，无法量化地表示运动信息的变化。

(3) 帧平均法[45]：从镜头中取所有帧在某个位置上像素值的平均值，然后将镜头中该点位置的像素值最接近平均值的帧作为关键帧。

(4) 直方图平均法[52]：是将镜头中所有帧的统计直方图取平均，然后选择与该平均直方图最接近的帧作为关键帧。

(5) 基于运动的方法[44]：通过计算镜头中某帧的每个像素光流分量的模之和，作为这一帧的运动量，选取运动量的局部最小处作为关键帧，它反映了视频数据中的静止信息，往往表示一种强调的实际情况。这种方法的缺点是计算量很大，算法复杂。

(6) 聚类法[54]：Zhuang 和 Hanjalic 提出了用无监督聚类方法提取关键帧，这种方法首先初始化一个聚类中心，然后根据当前帧与中心的距离，来判断是归为该类还是作为新的聚类中心，最后将各类中离聚类中心最近的帧作为关键帧。

3.3.2 基于主题字幕的关键帧提取

通过反复观察不同电视台的大量电视新闻视频之后，发现新闻视频与一般的视频相比，内容比较紧凑，并且大部分镜头的持续时间较短，镜头内容变化不大，运动强度低，每个镜头可以只选取一个关键帧。

在新闻视频中，有字幕出现的帧往往是一个新闻故事、一个新闻视频镜头的关键帧，这些视频帧最大限度地反映了新闻的主要内容，具有极强的代表性[43]。在第四章的第二节中，将提到新闻视频中的字幕分为主题字幕和非主题字幕，主题字幕与非主题字幕相比，所包含的语义信息更为丰富。一般情况下，同一个镜头中不会出现两个主题字幕。在此基础上提出了基于主题字幕的关键帧提取算法，如果镜头中包含主题字幕，则选取包含主题字幕；如果不包含主题字幕，则结合新闻视频中镜头的特点，采用直方图平均法[52]提取关键帧，该方法所选取的帧具有平均代表意义。

关键帧提取算法步骤如下：

Step1. 初始化，镜头总数为 n，当前镜头为 i，提取主题字幕的出现帧和消失帧 (详

细算法见 4.2 节)；

Step2. 如果镜头 i 中不包含字幕起始帧、消失帧，转到 Step3，否则选取包含字幕的中间帧作为关键帧，转到 Step4；

Step3. 将镜头 i 中所有帧的统计直方图取平均，然后选取与该平均直方图最接近的帧作为关键帧；

Step4. i=i+1，若镜头 i 不是最后一个镜头，转到 Step2，否则转到 Step5；

Step5. 程序结束，关键帧提取完毕。

关键帧提取的流程如图 3.3.1。

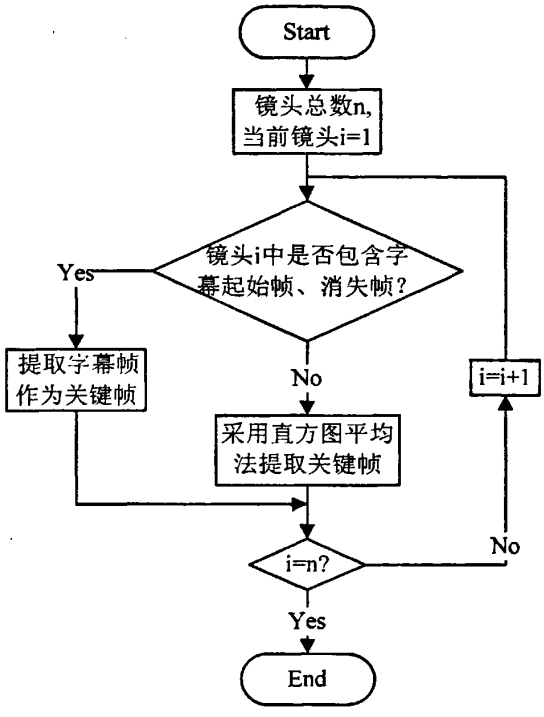
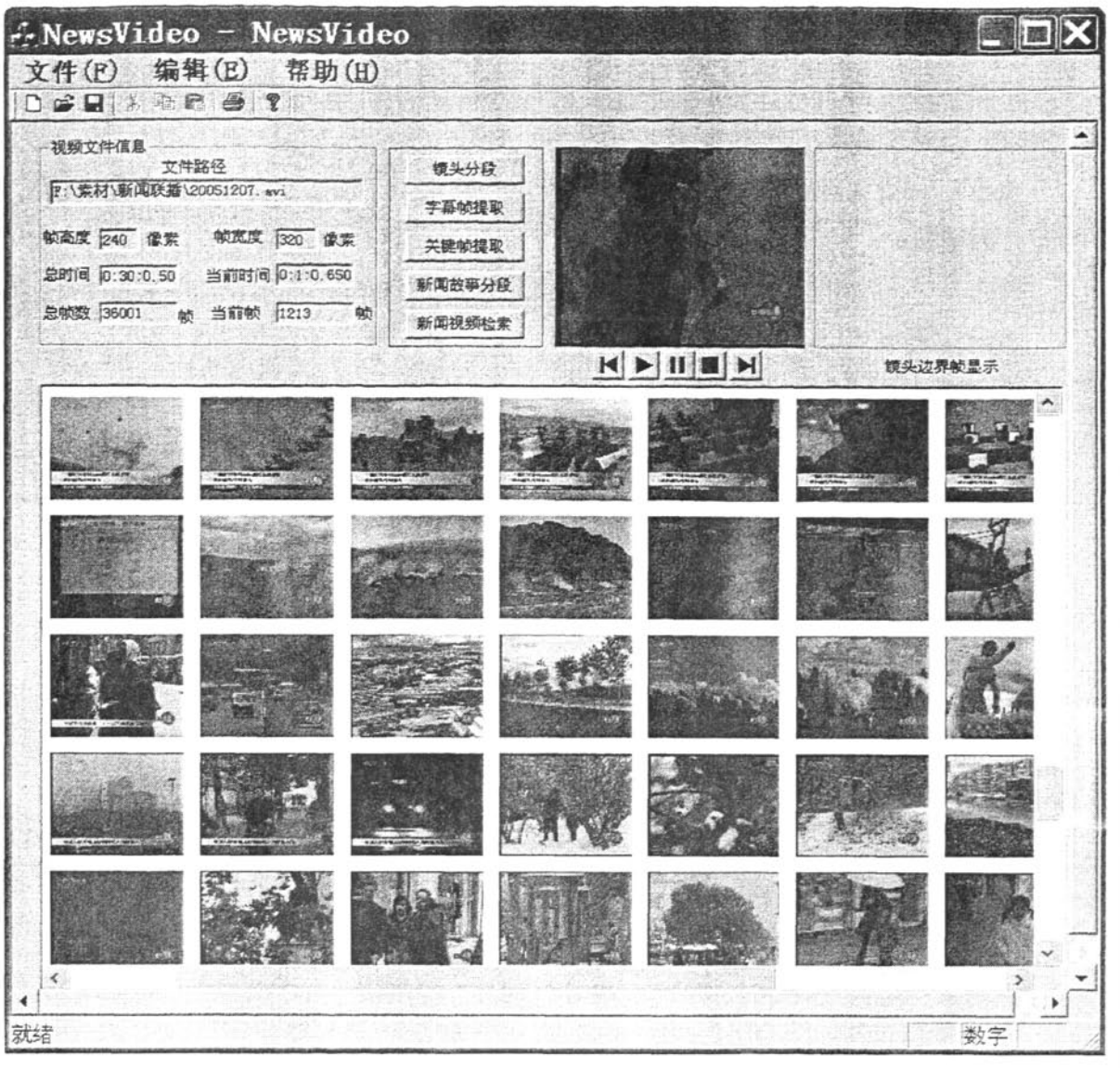


图3.3.1 关键帧提取流程图

3.3.3 实验结果

由于新闻视频内容比较紧凑，并且大部分镜头的持续时间较短，镜头内容变化不大，运动强度低，因此一个镜头可以只提取一个关键帧。本节的关键帧算法，一个镜头中只提取一个关键帧，所提取的帧具有平均代表意义。

关键帧提取的界面如图 3.3.2，对视频进行镜头分段、字幕帧提取后，可以进行关键帧提取。选取的素材是长达三十分钟的新闻联播节目“20051207”，当按下“关键帧提取”按钮后，界面下半部分显示的是新闻视频的关键帧，通过浏览关键帧，可以知道该新闻联播的主要内容。



3.3.2 关键帧提取界面

3.4 本章小结

本章分为三部分，主要完成了新闻视频的结构化。对视频的结构化分析既是视频分析的主要内容，也是解决基于内容视频检索问题的关键。

第一部分介绍了视频结构化分析的概念。

第二部分首先总结了常见的镜头分段方法，接着分析了双重比较法的缺点，并改进了双重比较法，将自适应阈值运用于双重比较法，完成了新闻视频的镜头分段。大量实验证明，该算法应用在新闻这类视频节目中，可以取得理想的分段效果，查准率达到93%以上，查全率达到99%以上。

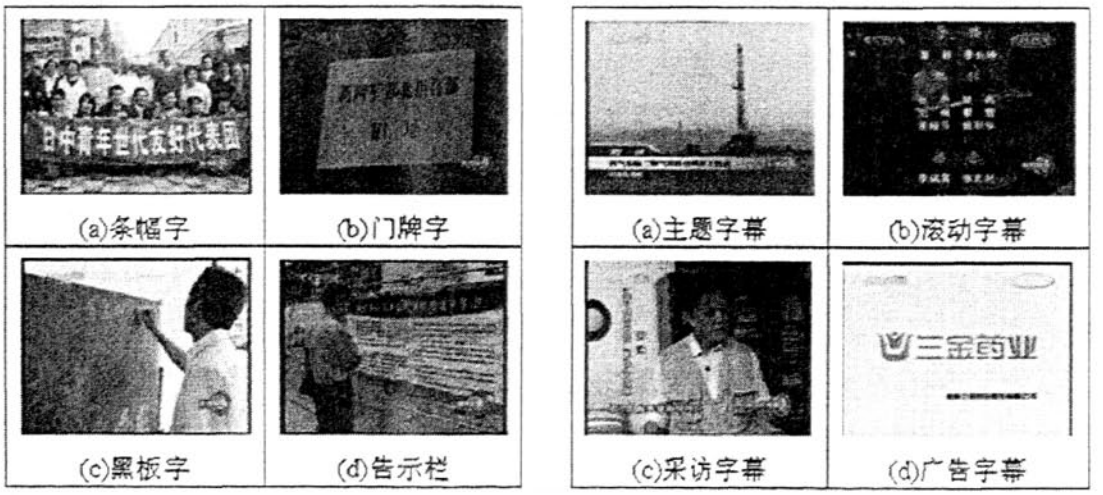
第三部分总结了常见关键帧提取方法，接着提出了基于主题字幕的关键帧提取算法，计算比较简单，基本上可以代表该镜头内容，通过浏览新闻视频的关键帧，可以知道新闻视频的主要内容。

4 新闻视频中的主题字幕提取

本章的主要内容是完成新闻视频中主题字幕提取，分为四部分：新闻视频中的字幕分类、主题字幕帧的检测、主题字幕区的提取和主题字幕的二值化及识别。

4.1 新闻视频中的字幕分类

新闻视频中的字幕可分为两类：场景字幕和标注字幕。场景字幕是原有场景的一部分，多数的场景字幕没有特定意义，而且出现的时间、位置、大小等都是随机不确定的，这种情况下的字幕比较复杂，很难进行检测和识别，如图 4.1.1(a)所示。而标注字幕是通过后期制作合成到视频流中去的，包含了对当前新闻视频内容的高级语义信息描述。节目编辑用这些字幕信息提供相关新闻故事信息，这类字幕有主题字幕、采访字幕、滚动字幕和广告字幕等等，如图 4.1.1(b)所示。



(a) 场景字幕　　(b) 标注字幕

图 4.1.1　字幕实例

新闻中的标注性字幕，按照字幕包含语义信息的重要程度及观众关注的程度，又可以将其分为两大类：非主题字幕和主题字幕。

在新闻采访中，记者与被采访人物的谈话字幕、以及采访者的身份介绍，是帮助观众听懂说话人的发言，或提供一些场景中的任务、地点等信息，称这种含语义信息较少的字幕为非主题字幕。

播音员在报道每个新闻故事单元时，必然会有一行或几行文字出现在新闻故事的开始或中间位置，概括表达正在报道的新闻故事的主题或中心意思，这些字幕即为新闻故事单元的主题字幕。主题字幕的反复出现，是新闻视频区别于其它视频的显著特点，这些主题字幕常常是新闻故事单元的切换标志。

非主题字幕和主题字幕的实例如图 4.1.2，前两张图为非主题字幕，后四张图为主题字幕。

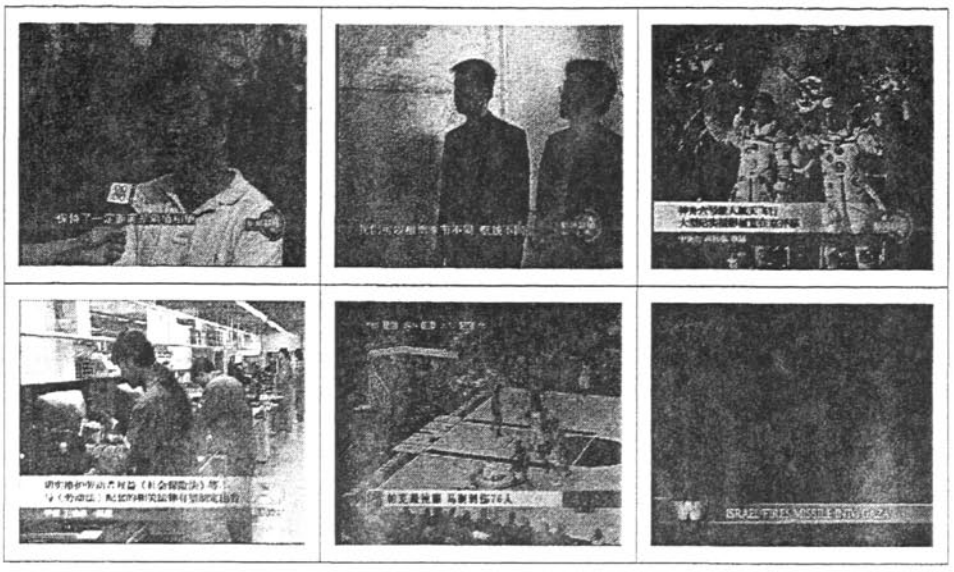


图 4.1.2 新闻字幕实例图

经过观察，发现主题字幕和非主题字幕，主要存在以下几个方面的不同[50]。(1)主题字幕和非主题字幕在视频流上停留的时间长短不同。主题字幕是故事单元的主题，在视频流上停留的时间较长，通常至少在 5 秒以上，以给观众足够的时间注意到并留下印象；而非主题字幕在视频流上停留的时间较短，通常不超过 3 秒。(2)主题字幕以一种特定的颜色，为了更醒目，字幕的背景是一个与字幕颜色相差较大的颜色固定的长矩形框，如新闻联播的主题字幕是蓝色，字幕背景色是白色；而非主题字幕常常是直接嵌入在视频的背景中，没有一定形状的文本背景。由此可知，主题字幕与非主题字幕相比，语义信息更丰富，也更容易提取。

但是，新闻主题字幕的检测，与文件图像的 OCR 相比，仍然面临以下挑战：一是无法判别哪帧是主题字幕帧；二是视频帧的字幕通常嵌入在复杂的背景中，增加了字幕提取和分割的难度；三是为了避免遮挡图像中的感兴趣对象，新闻主题字幕是相对较小的，因此，字符的分辨率较低；四是视频压缩使得低分辨率的字符的清晰度下降。综合上述问题，新闻视频不适合直接用 OCR 系统对主题字幕进行识别。

4.2 新闻主题字幕帧的检测

主题字幕帧的检测是后续字幕定位与识别的基础，通过研究新闻视频单镜头时空切片的纹理特征，设计了基于 3/10 时空切片的主题字幕帧边界检测算法，对主题字幕帧起始帧、消失帧进行检测。

4.2.1 时空切片的字幕纹理分析

4.2.1.1 时空切片的概念

时空切片[58]的概念是E.H.Adelson和J.Bergen于1985年在“Spatiotemporal Energy Models for the Perception of Motion”一文中首次提出的。所谓切片，就是对连续的视频图像序列的同一位置提取出的一行(列)像素组合而成的一副二维图像。如果将视频看作是一个(x,y,t)三维的图像序列，其中(x,y)为图像维、t为时间维，则视频的时空切片，可以看作是由时间维与图像维构成的一副二维图像。时空切片通常有3种取法[49]：垂直切片、水平切片、和对角切片。具体取法如图4.2.1所示。

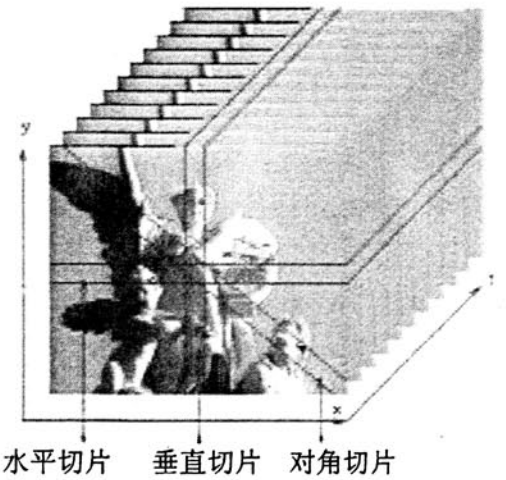


图 4.2.1　时空切片示意图

时空切片的计算方法是：假设一个视频序列总共有T帧图像，图像帧表示为$I(i,p)$，其大小为$M\times N$；h，v，d分别表示水平，竖直和对角线方向上的扫描线。

将二维的图像$I(i,p)$以垂直、水平和对角线方向影射成一条一维的扫描线，按扫描线的类型拼合成不同类型的时空切片。像素i所在的扫描线的表达式如下：

$$h_i=\sum_{p=k_2-j}^{k_2+j}\alpha_p I(i,p) \qquad k_2=\frac{M}{2} \tag{4.1}$$

$$v_i=\sum_{p=k_1-j}^{k_1+j}\alpha_p I(p,i) \qquad k_1=\frac{N}{2} \tag{4.2}$$

$$d_i=\sum_{p=i-j}^{i+j}\alpha_p I(p,i,t) \tag{4.3}$$

其中，α_p是高斯滤波的权重系数，有$\sum\alpha_p=1$。j是一个滤波窗口，它支持α_p，将几条相邻的扫描线合并为一条的目的是平滑去噪，将窗口大小设置为 3，相应的$[\alpha_p]=[0.2236,0.5477,0.2336]$。

取得扫描线之后，将它们按时间顺序依次拼合，得到时空切片：由水平扫描线得到H(尺寸为M×T)，竖直扫描线得到V(尺寸为N×T)，斜扫描线得到D(尺寸为N×T)。图4.2.2

是本文从一段不包含字幕的新闻视频中提取出来的时空切片，仔细分析镜头切片的内在纹理，可以发现切片的纹理中包含有对应视频的大量信息，该视频由五个突变镜头组成。

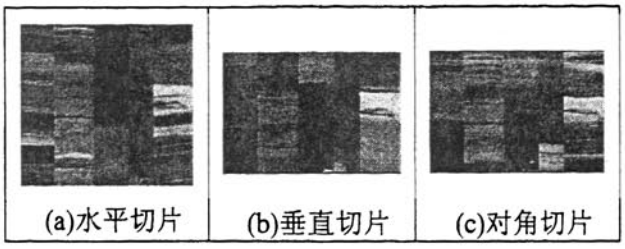


图 4.2.2 三种切片图

4.2.1.2 主题字幕帧检测算法

对包含主题字幕的原视频建立时空切片，如图 4.2.3，由于主题字幕位于视频下部横向排列，从而水平切片一般不包括主题字幕信息，只需建立垂直切片和对角切片。观察大量新闻视频可知，主题字幕一般位于视频下部 3/10 处，如图 4.1.1。因此只需要对视频的下部 3/10 处做垂直切片和对角切片，同时避免了视频上部切片纹理的干扰。

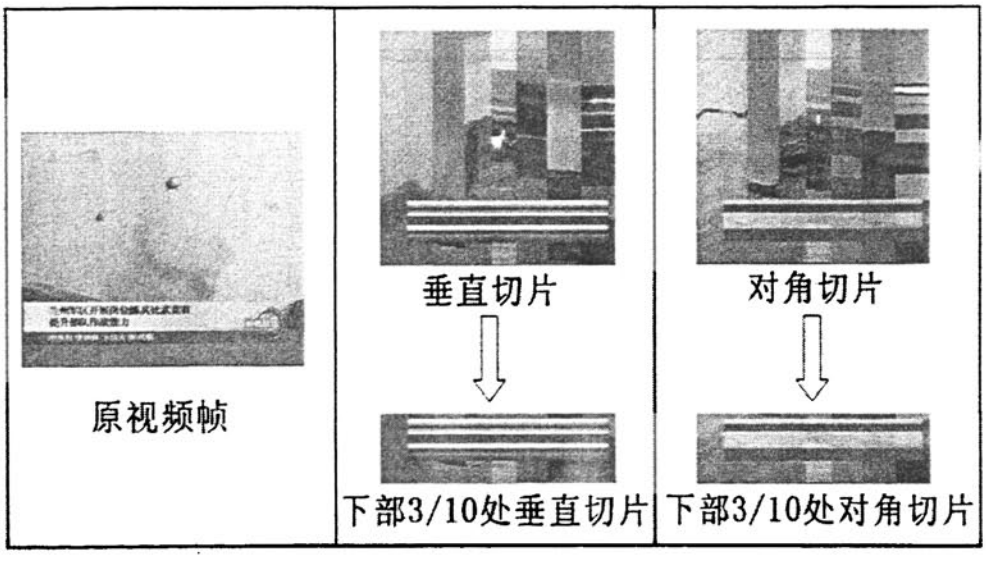


图 4.2.3 主题字幕时空切片图

通过检测单镜头切片中的字幕背景纹理的边缘，如图 4.2.4 就可以大致确定主题字幕的起始帧和消失帧。由图可知，垂直切片和对角切片的纹理，可以清晰地反映字幕帧的区域边界。主题字幕纹理表现为具有一定宽度的矩形，根据字幕的稳定性和时间连续性，一段主题字幕的出现至少要经过 5 秒且在其间字幕无变化，因此只要能确定切片中字幕纹理矩形的两端，也就确定了字幕起始帧和消失帧，而没有必要检测到所有的包含相同字幕的帧，降低了检测的复杂度。

观察图 4.1.2 可知，非主题字幕一般也位于视频下部 3/10 处。下面对一段包含非主题字幕的单镜头建立时空切片，如图 4.2.4。由于非主题字幕没有字幕背景，直接嵌入在视频的背景，字幕区域所占在像素少，从而在切片上只能表现出一个横线似的纹理，导

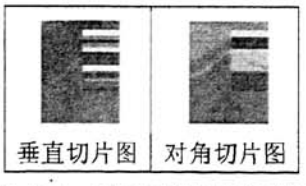
致整个时空切片图上表现出大致相同的颜色特征和纹理，与图 4.2.2 所示不包含字幕的切片纹理相似。以上分析可知，时空切片法不能检测出非主题字幕，从而非主题字幕不会被误检为主题字幕，因此通过时空切片法可以用来完成主题字幕的检测。(注：本文下面提及的时空切片均是视频下部 3/10 处的时空切片)。

(a)原视频帧 (b)垂直切片 (c)对角切片

图 4.2.4 非主题字幕时空切片图

4.2.2 纹理预处理

新闻视频进行镜头分段后，建立镜头的时空切片图，得到单镜头的垂直切片和对角切片图，如图 4.2.5 所示。在无主题字幕出现时候，由于单镜头内纹理基本上是一致的；主题字幕在出现或者消失的时候，单色的字幕背景在切片中，将镜头切片分为字幕区域和非字幕区域，其字幕边界纹理在图 4.2.5 可以清晰地看到。



垂直切片图 对角切片图

图 4.2.5 时空切片图的字幕纹理

4.2.2.1 基于坎尼(Canny)算子的边缘特征提取

因为 Canny 算子可以有效减弱噪声在最终边缘图像中的影响，因此采用该算子对切片图进行边缘检测，从而有效地检测到单镜头时空切片图中的边缘直线。

在众多的边缘检测算子中，坎尼(Canny)算子有其独特的检测效果。Canny 算子的基本思想是：先对处理的图像选择一定的高斯滤波器进行平滑滤波，然后采用一种称之为“非极值抑制”(Nonmaxima Suppression)的技术，对平滑后的图像处理后，得到最后所需的边缘图像。下面列举 Canny 算子的步骤[53]。

Step1. 先对原始切片图像，进行灰度变换，在这里对 R，G 和 B 三个通道的颜色信息都加以考虑，按照式(4.4)进行灰度变换：

$$L(x, y)=0.3*R(x, y)+0.59*G(x, y)+0.11*B(x, y) \tag{4.4}$$

Step2. 用高斯滤波器来对图像滤波，以去除图像中的噪声；

Step3. 用高斯算子的一阶微分对图像进行滤波。对滤波后图像中的每个像素，计算

其梯度的大小 M 和方向 θ，可采用2×2大小的模板作为对 x 方向和 y 方向偏微分的一阶近似：

$$P=\frac{1}{2}\times\begin{bmatrix}-1 & 1\\ -1 & 1\end{bmatrix} \qquad Q=\frac{1}{2}\times\begin{bmatrix}1 & 1\\ -1 & -1\end{bmatrix} \tag{4.5}$$

由此得到梯度的大小 M 和方向 θ：

$$\theta(i,j)=\arctan[Q(i,j)/P(i,j)] \tag{4.6}$$

Step4. 遍历一次灰度矩阵 M，把梯度方向 θ 分为四种：

0 度 (水平方向) 范围：0-22.5 以及 157.5-180 度

45 度 (45 度方向) 范围：22.5-67.5 度

90 度 (垂直方向) 范围：67.5-112.5 度

135 度 (135 度方向) 范围：112.5-157.5 度

求 M 的累积直方图得到 T_h，$T_l=0.4T_h$；

Step5. 非最大抑制：遍历 M，若某个像素的灰度值与其梯度方向上前后两个像素的灰度值相比不是最大的，那么这个像素值置为 0；

Step6. 双阈值操作：遍历 M，大于 T_h 的设为边缘；小于 T_l 的设为非边缘；小于 T_h 但是大于 T_l 的时候，看它周围像素有没有大于 T_h 的边缘，如果有则其是边缘，否则，不是边缘；

按照上述 Canny 算子，就可以完成对切片图中边缘轮廓的提取工作，算法效果如图 4.2.6 所示。

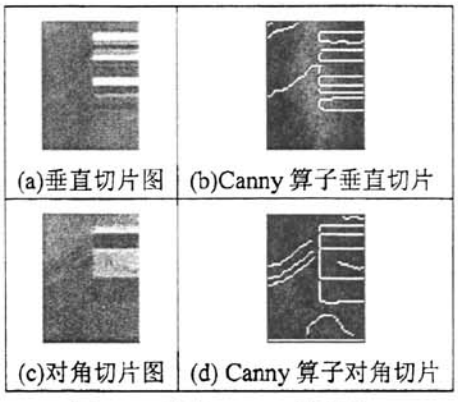


图 4.2.6 坎尼(Canny)算子图

4.2.2.2 基于形态学算子的水平线去除

观察图 4.2.2 所示切片可知，即便是一个相对静止的图像序列，反映在时空切片上也是包含有许多水平的条纹，这是由于视频流图像中不同区域中不同颜色块的边界线引起的。因此通过 Canny 算子将时空切片图像边缘化之后，其中仍然包含大量的水平直线，如上图 4.2.6。

由上文分析可知，主题字幕边界处的纹理，一般只包含垂直线和斜线两种情况，而不包含有长的水平直线。这些水平直线的存在，只会给主题字幕边界纹理特征的提取带来干扰，使主题字幕边界区域变得模糊，因此需要将这些水平线去除。

扫除水平直线的简单算法步骤如下：

Step1. 在 Canny 边缘化的时空切片图像中找到下一个要处理的边界点；

Step2. 观察边界点的 3×3 邻域，如果除了水平方向两个邻域有边界点之外，其余六个方向都没有边界点，则执行 Step3，否则执行 Step4；

Step3. 按照自左至右的顺序，在水平方向数 N 步(这里 N 取 2)，若都是边界上的点，就认为找到了一条水平线，将水平线开头的目标点置为非边界点；否则，执行 Step4；

Step4. 重复上述三步，直到所有的 Canny 边界点都被扫描到为止。

经过实验证明，上述算子对扫除非镜头边界线处的水平直线很有效，而且可以保留大部分的镜头边界上的点。如图 4.2.7 所示，它是图 4.2.6 的后续处理图。

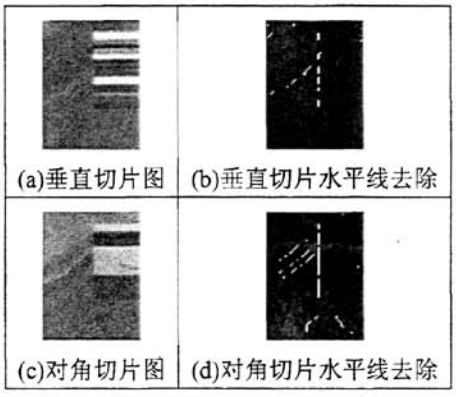


图 4.2.7 水平线去除图

4.2.2.3 基于 Hough 变换的非边缘点扫除

如图 4.2.7 所示，经过上面两小节处理之后，主题字幕边界线已经比较清晰了。但是可以注意到，仍然有不少短的斜线甚至是孤立的小点存在，这些都可能对后面的能量检测形成干扰。这些干扰线来源于两个方面：其一，来源于时空切片处的水平条纹。由于某种原因，使其在经过 Canny 算子之后没有成为标准的水平线，而有一定的斜率；其二，来源于摄像机和视频目标的轻微运动。因此，很有必要将这些干扰排除。

这里采用 Hough 变换检测直线的方法，排除这些小短线的干扰，实验证明，该算法的效果较好。下面首先介绍一下 Hough 变换的概念。

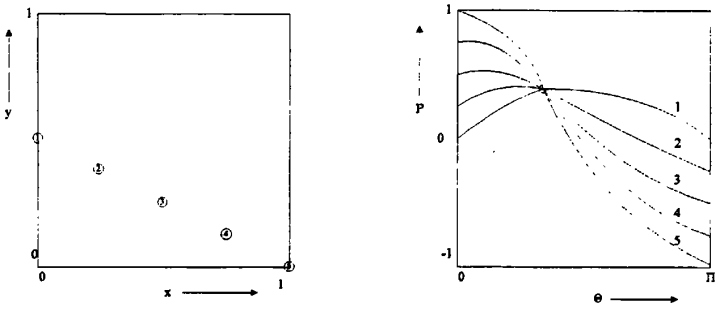
Hough 变换于 1962 年由 Paul Hough 提出，并在美国作为专利被发表。它所实现的是一种从图像空间到参数空间的映射关系。Hough 变换的实质是将图像空间具有一定关系的像素进行聚类，寻找能够把这些像素用某一解析形式联系起来的参数空间累积对应点。

直线 Hough 变换，利用图像空间和 Hough 参数空间的点——线对偶性，把图像空间的检测问题转换到参数空间。通过在参数空间里进行简单的累加统计，然后在参数空间，寻找累加器峰值的方法来检测直线。通常，直线的方程可以用 $y=kx+b$ 来表示，其中 k 和 b 是参数。过某一点 (x_0,y_0) 的所有直线的参数都会满足方程 $y_0=kx_0+b$，即点 (x_0,y_0) 确定了一族直线。方程 $y_0=kx_0+b$ 在参数 k-b 平面上是一条直线。这样图像 x-y 平面上的一个前景像素点就对应到参数平面上的一条直线。

在实际应用中，$y=kx+b$ 形式的直线方程没有办法表示 $x=c$ 形式的直线，所以，通常采用参数方程：

$$\rho=x\cos(\theta)+y\sin(\theta) \tag{4.7}$$

这样，如图 4.2.8 所示，图像平面上的一个点就对应参数 $\rho-\theta$ 平面上的一条曲线，同一条直线上的点对应在 $\rho-\theta$ 平面上的曲线将相交于一点。



(a) x,y 平面 (b) ρ,θ 平面

图 4.2.8 Hough 变换

基于 Hough 变换的非边缘点扫除算法如下：

Step1. 将经过水平线扫除的边缘图进行 Hough 变换，得到一系列 (r,θ) 对；

Step2. 找到包含像素点最多的 n 组 (r,θ) 对，得到一个 $n\times2$ 的矩阵；

Step3. 在边缘图上扫描下一个边界点；

Step4. 判断该边界点是否属于 n 组 (r,θ) 对中的一组；

Step5. 如果属于，则转到 Step3；

Step6. 如果不属于，将该点置为非边界，转到 Step3；

Step7. 重复上述步骤，直到所有的边界点都扫描完毕。

经过上面的处理，大部分的非主题字幕边界处的斜线都被扫除了，如图 4.2.9 所示，它是图 4.2.7 后继处理的结果。

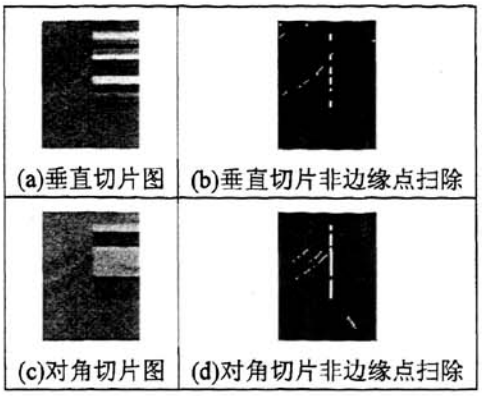


图 4.2.9 非边界处斜线的扫除图

经过如上处理，一部分非边界处的斜线被扫除了，如图 4.2.9 所示，即将建立的时空切片转化为主题字幕边界比较突出的边缘时空切片了。

4.2.3 能量模型

得到边缘时空切片后的工作就是，将边缘时空切片分为若干个区域，每一个区域对应镜头中的主题字幕或非主题字幕区域，而区域交界处就是主题字幕边界的地方。因此，需要从边缘切片中提取能够体现出区域一致性的特征，同时也需要检测区域间的边界。这里建立一个能量模型，用于提取边缘切片的纹理特征，同时获取一致区域的边界信息的模型。区域边界，可以理解为，是出现在纹理特征违反了区域内容一致性的地方。因此，通过对区域边界的检测，刻画出区域边界的范围，就可以实现对主题字幕区域边界帧检测的目的了。

4.2.3.1 计算切片的纹理特征

下面，分别对 V、D 两种切片的 RGBY 分量做分析，在颜色域和亮度域中分别提取其中的纹理特征。

将 V、D 切片在 RGBY 空间中的切片记作为：

$$V=[V_r,V_g,V_b,V_y],\quad D=[D_r,D_g,D_b,D_y] \tag{4.8}$$

分别将其转化为边缘时空切片，然后对 RGB 颜色分量沿着 θ 方向作一阶高斯(Gaussian)导数的卷积滤波，以提取颜色域中包含的纹理特征信息(以 V 切片为例，其余类似)：

$E^{V_i}_{\sigma,\theta}$ 为切片 V_i 在颜色域中的纹理特征信息。时空切片的彩色边缘信息可以通过式 4.9 获得：

$$E^{V_m}_{\sigma,\theta}=\bar{G}'_{\sigma,\theta}*V_m \tag{4.9}$$

其中*表示卷积运算符，$m\in\{r,g,b\}$，$\bar{G}'_{\sigma,\theta}$ 为沿着 θ 方向的一阶高斯导数，表达式如下：

$$\bar{G}'_{\sigma,\theta}(i,j)=-\frac{i}{\sigma^2}\bar{G}_{\sigma,\theta}(i,j),\quad \bar{G}_{\sigma,\theta}(i,j)=\bar{G}_{\sigma}(i',j') \tag{4.10}$$

$$i'=i\cos\theta+j\sin\theta \quad , \quad j'=-i\sin\theta+j\cos\theta \tag{4.11}$$

$$\bar{G}_{\sigma}(i,j)=\exp\{-\frac{(i^2+j^2)}{2\sigma^2}\} \tag{4.12}$$

时空切片亮度域中的纹理特征信息$T_{\sigma_i,\sigma_j,\theta}$用 Gabor 滤波器$\hat{G}_{\sigma_i,\sigma_j,\theta}(i,j)$来提取，以 V 切片为例：

$$T_{\sigma_i,\sigma_j,\theta}=\hat{G}_{\sigma_i,\sigma_j,\theta}*V_y \tag{4.13}$$

$$\hat{G}_{\sigma_i,\sigma_j,\theta}(i,j)=\hat{G}_{\sigma_i,\sigma_j}(i',j') \tag{4.14}$$

$$\hat{G}_{\sigma_i,\sigma_j}(i,j)=(\frac{1}{2\pi\sigma_i\sigma_j})\exp\{-\frac{1}{2}(\frac{i^2}{\sigma_i^2}+\frac{j^2}{\sigma_j^2})\}\exp\{2\pi JWi\} \tag{4.15}$$

其中，$J=\sqrt{-1}$，W表示径向中心频率。

以上分析可知，可以用一个 12 维的特征分量来表征每个像素点的颜色纹理特征，以 V 切片为例，V 上一个像素的纹理特征向量表达式如下：

$$[E^{V_r}_{\sigma,\theta}(i,j),E^{V_s}_{\sigma,\theta}(i,j),E^{V_b}_{\sigma,\theta}(i,j),T_{\sigma_x,\sigma_y,\theta}(i,j)] \qquad \text{其中}\theta=\{0^\circ,45^\circ,-45^\circ\}$$

4.2.3.2 能量模型分析

记$V(i,t),D(i,t)$中的某个像素点：$x_i=(v_i,d_i)$属于字幕区域边界上的点的概率为：

$$\begin{aligned}p(x_i\in\xi\,|\,V,D)&=p(x_i\in\xi\,|\,V_N,D_N)\\&=p(x_i\in\xi\,|\,V_N)p(x_i\in\xi\,|\,D_N)\end{aligned} \tag{4.16}$$

这里，假设V_N,D_N相互独立，其中V_N,D_N分别为像素x_i的 3*3 邻域系，将在下一节中详细介绍。

根据 Markov-Gibbs 等式，可以假设$p(\eta_i),\eta_i\in\{v_i,d_i\}$服从 Gibbs 分布，即：

$$p(\eta_i)=\frac{1}{\lambda}\exp\{-U(\eta_i)\} \tag{4.17}$$

其中，λ为归一化常数，$U(\eta_i)$是邻域系统决定的能量函数。

将式 4.17 代入式 4.16，并对等式两边分别求对数，得到如下公式：

$$L(x\in\xi)\propto-\sum_{i=1}^{N}\{U(v_i)+U(d_i)\} \tag{4.18}$$

其中，$L(x\in\xi)=\sum_{i=1}^{N}\log\{p(x_i\in\xi\,|\,V,D)\}$。可以看出，像素点 x 属于边界区域的概率时刻 t 所有像素能量总和成正比。

能量函数$U(\eta_i)$表达式如式 4.19，可知，能量函数包含有四个元素，分别是(r,g,b)

颜色域和 y 亮度域中切片的纹理特征。

$$U(\eta_i)=\{U^r(\eta_i),U^g(\eta_i),U^b(\eta_i),U^y(\eta_i)\} \tag{4.19}$$

4.2.3.3 边缘像素能量函数计算

像素点 x_i 的能量是根据它的邻域系统来计算的，如图 4.2.10 所示。在 4.2.3.1 节中提及，考虑 $\theta=\{0^\circ,45^\circ,-45^\circ\}$ 这三种情况，可以定义像素 x_i 的 8 邻域结构单元 $C=\{C_1,C_2,\ldots,C_8\}$，如图 4.2.11，可以看出，$\{C_1,C_2\}$ 描述了垂直和水平边界线，$\{C_3,C_5,C_7\}$ 描述了 -45° 负斜率边界线，$\{C_4,C_6,C_8\}$ 描述了 45° 正斜率边界线。

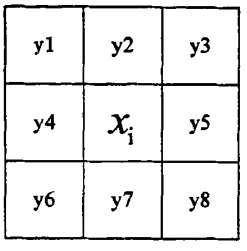


图 4.2.10 像素 x_i 的 8-邻域系统

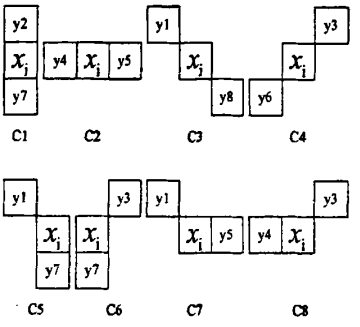


图 4.2.11 像素 x_i 的结构单元

根据领域系统的特点，以 V 切片的 r 分量为例，可将能量函数 $U^r(v_i)$ 定义如下：

$$U^r(v_i)=3\Gamma^r_{c_1}(v_i)-\Gamma^r_{c_2}(v_i)-\Gamma^r_{c'}(v_i)-\Gamma^r_{c''}(v_i) \tag{4.20}$$

其中 $\Gamma^r_{c'}(v_i)=\min_{c\in\{c_3,c_5,c_7\}}\Gamma^r_c(v_i)$，$\Gamma^r_{c''}(v_i)=\min_{c\in\{c_4,c_6,c_8\}}\Gamma^r_c(v_i)$，$\Gamma^r_{c_j}(v_i)$ 为 v_i 邻域结构元 C_j 的能量的 r 分量，v_i 为 $V_r(i,j)$ 上的像素点。

令 $\eta_1=V_r(i_1,j_1),\eta_2=V_r(i_2,j_2)$ 为 C_j 中 v_i 的两个邻点，即由 $\{\eta_1,v_i,\eta_2\}$ 构成一个邻域结构元 C_j。定义 v_i 邻域系结构元 C_j 的能量的 r 分量 $\Gamma^r_{c_j}(v_i)$ 为：

$$\Gamma^r_{c_j}(v_i)=\left|E^{V_r}_{\sigma,\theta}(i,j)-E^{V_r}_{\sigma,\theta}(i_1,j_1)\right|+\left|E^{V_r}_{\sigma,\theta}(i,j)-E^{V_r}_{\sigma,\theta}(i_2,j_2)\right| \tag{4.21}$$

类似的，y 亮度域的纹理信息 $\Gamma^y_{c_j}(v_i)$ 计算公式如下：

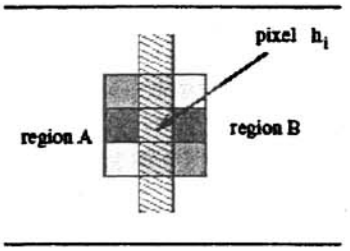
$$\Gamma^y_{c_j}(v_i)=\left|T^{V_y}_{\sigma,\theta}(i,j)-T^{V_y}_{\sigma,\theta}(i_1,j_1)\right|+\left|T^{V_y}_{\sigma,\theta}(i,j)-T^{V_y}_{\sigma,\theta}(i_2,j_2)\right| \tag{4.22}$$

公式 4.21 和 4.22 中：

$$\begin{aligned}&if\quad j\in\{1,2\},\quad then\quad \theta=0^\circ\\&if\quad j\in\{4,6,8\},then\quad \theta=45^\circ\\&if\quad j\in\{3,5,7\},then\quad \theta=-45^\circ\end{aligned}$$

求出所有点的能量之后，可以根据点的能量大小，判定点属于边界区域的概率。图 4.2.12 说明了能量函数(4.20)的直观意义，若 P 像素为边界区域，则其将取得局部最小的

能量。



表示镜头的边界区域，不同的颜色表示能量值的大小，蓝色最小

图 4.2.12 能量函数的意义

能量函数主要完成两个功能：其一，当某个结构单元跨越两个不同区域时，能量Γ_{c_j}的值比较大；其二，当结构单元处在同一区域或区域边界上时，能量Γ_{c_j}的值相对较小。

最终定义V切片能量函数$U(v_i)$为：

$$U(v_i)=\min_{m\in\{r,g,b\}} U^m(v_i)+U^y(v_i) \tag{4.23}$$

与直接进行能量计算相比，经过边缘提取后的时空切片，其能量模型表现为更加集中的边界区域，以及更加明显的能量差异，将在 4.2.4.1 中进行具体的比较。

4.2.4 主题字幕帧的检测

4.2.4.1 主题字幕区边界帧的检测

新闻视频节目中，主题字幕呈水平排列，时空切片纹理比较简单，边界线在 V、D 切片上都表现为一条垂直的直线，如图 4.2.13 所示，边界能量比较集中，因此检测比较容易，准确率也比较高。

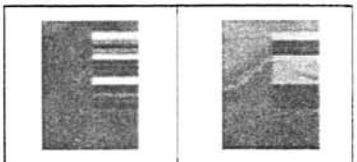


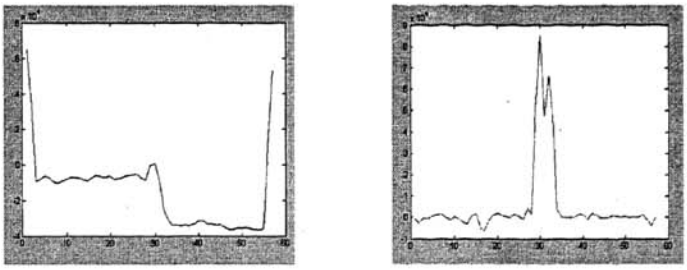
图 4.2.13 主题字幕切片图

在 4.24 式中，令区域边界$\xi=caption$，那么可以得到：

$$L(x\in caption)=-\{\sum_{i=1}^{N}U(v_i)+\sum_{i=1}^{N}U(d_i)\} \tag{4.24}$$

从上式可知，处于边界区域上的点能量比较低，只要寻找局部$L(x\in caption)$的能量最小值，主题字幕边界帧的位置就可以确定了。但是，通过寻找能量比预定阈值低的确定边界法，通常得不到正确的检测结果，因为很难找到一个可以同时避免错检和漏检的阈值。这里采用滤波技术来增大局部最大值，来检测主题字幕边界帧。

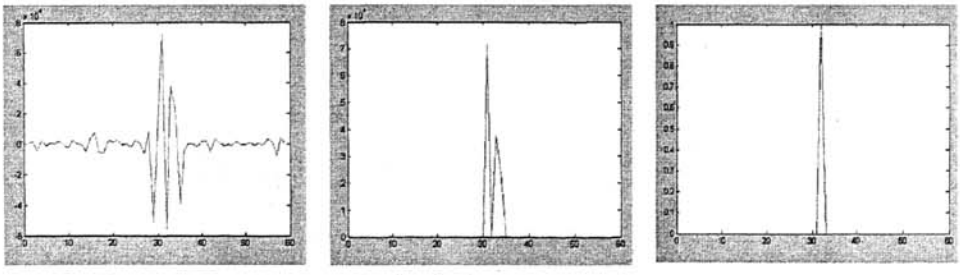
根据图 4.2.13 所示切片，直接计算其能量得到如图 4.2.14(a)所示的能量函数，而先对其进行纹理预处理，变换成边缘切片后，再根据 4.24 式得到的检测函数如图 4.2.14 (b)所示：



(a) 直接使用能量模型计算的检测函数 (b) 对边缘切片计算的检测函数

图 4.2.14 主题字幕帧边界检测函数图

根据能量模型可知，跨越了边界的结构元，其能量大于正好位于边界上的结构元，由此可以推断，在字幕帧的边界处的点的能量具有“大小大”的特点。从图 4.2.14(b)中可以看出，在 30-35 之间存在一个局部最小点，由此推断，字幕帧的边界发生在此点附近。为了放大这个特点，采用 $F=[-1 \quad 2 \quad -1]$ 的滤波器对函数进行卷积滤波，即对检测函数向两边求偏导，滤波的情况如图 4.2.15(a)所示。



(a)滤波之后的检测函数 (b)阈值运算之后的检测函数 (c) 判定结果

图 4.2.15 对检测函数的处理

由图 4.2.15(a)可知，滤波之后的检测函数，“大小大”特征更加明显了。再进行阈值运算，即将大于阈值 λ 的值留下，小于阈值 λ 的值置 0，可以得到的一系列“0”值和“大”值。设滤波之后的检测函数(图 4.2.15(a))为 *Caption*，这里设定阈值 λ 为：

$$\lambda = k * \sum_{i=1}^{T} | Caption(i) | / T \tag{4.25}$$

经过阈值运算之后的检测函数如图 4.2.15(b)，只要判断在 0 值的左右两帧范围内是否有很大的值，就可以找到主题字幕边界帧的位置了，最终判定的结果如图 4.2.15(c)，可知主题字幕边界帧在横坐标的 32 处。

4.2.4.2 采用直方图法进行定位和校验

字幕边界帧分为字幕起始帧和消失帧。以字幕消失帧为例，希望检测到消失帧，但

是有时候检测到的，可能是下一非字幕的第一帧。这里采用直方图对比的方法，在寻找到的边界点前后移动一两个像素点，判断直方图之差，当满足前点与后点差异超过阈值，而前点和自己差异较小的时候，就认为检测到的点处于字幕消失帧。

图 4.2.14(b)的效果令人满意，但有时摄像机运动、视频背景，也会在切片中产生小段类似字幕边界的垂直直线，容易造成误判。可以通过提高阈值的方法来解决，但是可能会漏判字幕边界。同理，可以采用直方图对比法，对检测到的边界进行校验。通过比较检测到的边界点，对比原视频下部 3/10 处两边直方图间的差异，可以消除一些误判。

4.2.4.3 字幕起始帧和消失帧的判别

经过 4.2.4.2 节直方图法的定位，可以得到字幕的边界帧，分为两张情况：第一种情况是，字幕起始帧的前一帧、字幕起始帧，如图 4.2.16(a,b)；第二种情况是字幕消失帧的前一帧、字幕消失帧，如图 4.2.16(c,d)。对这两种情况不加判别的话，如果出现漏检的情况，可能导致其后主题字幕帧的判定错误。例如主题字幕序列为$\{C_1, C_2, ..., C_n\}$，若C_i的字幕消失帧漏检，则接着检测出来的C_{i+1}字幕起始帧将被误判为C_i的字幕消失帧，从而导致后面一系列误判。

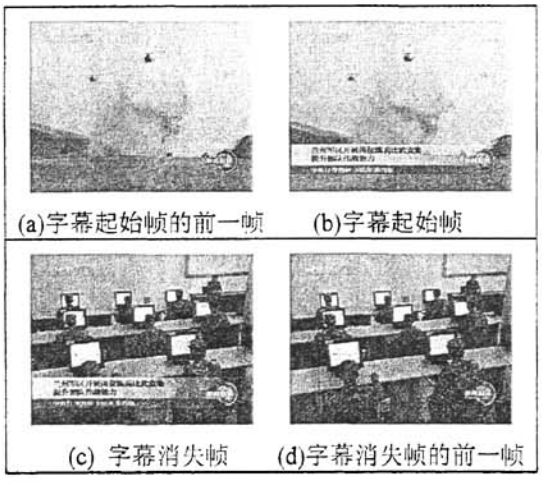


图 4.2.16 字幕帧边界

如图 4.2.16(a,b)、(c,d)这两种情况，边界帧两幅图的区别，主要在于是否包含字幕。如何判别是第一种情况，还是第二种情况，可以从字幕的纹理特点来考虑。这里提出，基于支持向量机的边界帧判别算法，将基于小波变换和支持向量机(SVM)的字幕区提取算法用于边界帧判定，在 4.3 节中，将详细介绍该字幕区提取算法。

边界帧判别具体算法是，用训练好的支持向量机模型，对处于边界的两个图像帧进行标记，即将图像帧划分为 N*N 的子块，将子块标记为字幕块还是非字幕块，字幕块标记为+1，非字幕块标记为-1，统计字幕块总个数，即+1 的个数。如图 4.2.16，可以直观地看出(b)、(c)包含字幕，(a)、(d)不包含字幕。通过支持向量机的分类，可以得到左右边界帧的字幕块总数，哪一帧的字幕块总数多，哪一帧是字幕帧。如果左边界帧是字

幕帧，则该边界是字幕消失帧边界，反之，则是字幕起始帧，从而可以将字幕帧标注为字幕起始帧或字幕消失帧。字幕起始帧、消失帧判别算法流程如图 4.2.17。对经过支持向量机模型标记后的字幕帧，再进行连通性分析，可以提取出最终的字幕区。

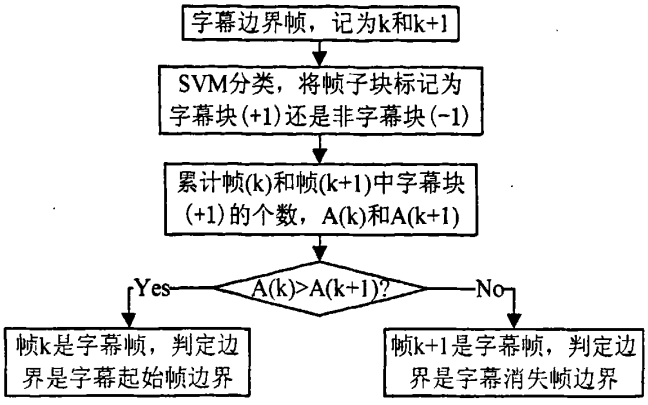


图 4.2.17 字幕起始帧和消失帧的判别流程图

4.2.4.4 主题字幕帧检测流程图

根据以上分析，新闻主题字幕起始帧、消失帧的检测算法流程如图 4.2.18 所示：

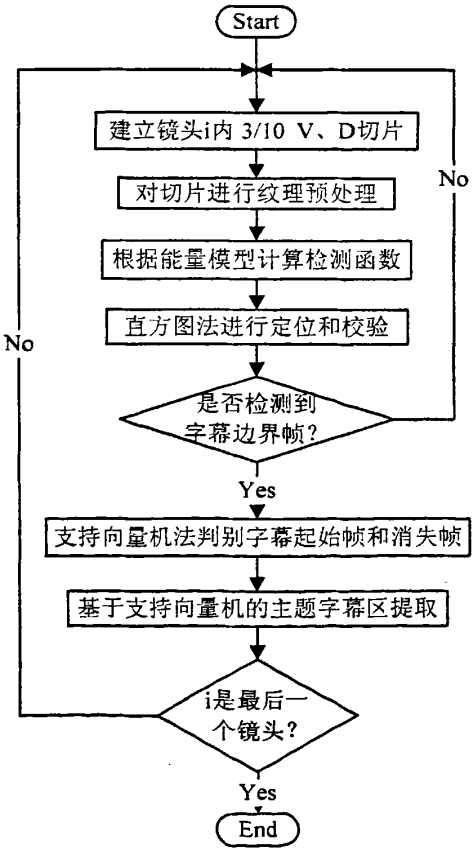


图 4.2.18 主题字幕边界帧流程图

4.2.5 实验结果

字幕帧提取的界面如图 4.2.19，对视频进行镜头分段后，可以进行字幕帧提取。选取的素材是长达三十分钟的新闻联播节目“20051207”，当按下“字幕帧提取”按钮后，界面下半部分显示的是，新闻视频的主题字幕起始帧和消失帧。通过浏览主题字幕帧，也可以知道该新闻联播中的新闻故事内容。

4.2.19 主题字幕帧提取界面

第五章中根据主题字幕的起始帧，进行新闻故事分段，因此这里选择起始帧作为评价依据。采用四天 CCTV-1 的新闻联播作为测试数据，为了衡量所提方法的性能，定义如下两个评价指标：

$$查全率=\frac{正确检测的主题字幕起始帧个数}{正确检测的主题字幕起始帧个数+漏检个数}$$

$$查准率=\frac{正确检测主题字幕起始帧数}{正确检测主题字幕起始帧数+误判个数}$$

表 4.2.1 给出了实验结果。由表可知，实验结果表明，该方法对于主题字幕帧的检测具有较高的查准率和查全率，由于采用直方图法进行校验，误检较少，从而查准率达到 98%以上。

表 4.2.1 字幕起始帧帧检测实验图

类别	实际的字幕起始帧个数	正确检测的起始帧个数	误检个数	漏检个数	查全率	查准率
新闻联播 1	26	25	0	1	96.15%	100%
新闻联播 2	24	22	1	2	91.67%	95.65%
新闻联播 3	21	18	0	3	85.71%	100%
新闻联播 4	30	28	0	2	93.33%	100%
总计	101	93	1	8	92.08%	98.94%

4.3 主题字幕区提取

视频字幕区的提取，是将视频序列中人们感兴趣文字语义信息部分，从复杂的视频背景中分离出来，如果能够把字幕信息提取出来，就可以达到基于语义内容的视频检索和分类。由于主题字幕一般采用与背景有着强烈对比度的颜色，因此字幕区域表现出比其他区域更高的空间频率。小波变换具有多尺度多分辨率的特性，可以通过对图像进行小波变换，提取特征量。本文采用基于小波变换和支持向量机的字幕区提取方法[42]，该方法采用小波变换提取字幕的纹理特征，再利用支持向量机方法进行分类，将视频帧分为主题字幕区和非主题字幕区，从而实现主题字幕区的提取。实验表明在复杂多变的视频背景下，仍能保持较高的正确率。

4.3.1 基于小波变换的特征提取

4.3.1.1 小波变换的概念

小波变换是近年来得到广泛应用的数学工具，与傅立叶(Fourier)变换、窗口傅立叶变换(Gabor 变换)相比，小波变换是空间(时间)和频率的局部变换，因此，能有效地从信号中提取信息。小波变换的最大特点是能够对信号进行显微镜式的观察，针对图像信号，高通滤波可检测图像的边缘信息，而低通滤波则可提供丰富的图像纹理特征。

小波变换的定义[50]，是把某一被称为基本小波(也叫母小波 mother wavelet)的函数 $\psi(t)$ 做位移 τ 后，再在不同尺度 a 下与待分析的信号 $x(t)$ 做内积：

$$WT_x(a,\tau)=\frac{1}{\sqrt{a}}\int_{-\infty}^{+\infty}x(t)\psi^*(\frac{t-\tau}{a})dt, \qquad a>0 \tag{4.26}$$

等效的频域表示是：

$$WT_x(a,\tau)=\frac{\sqrt{a}}{2\pi}\int_{-\infty}^{+\infty}X(\omega)\Psi^*(a\omega)e^{+j\omega\tau}d\omega \tag{4.27}$$

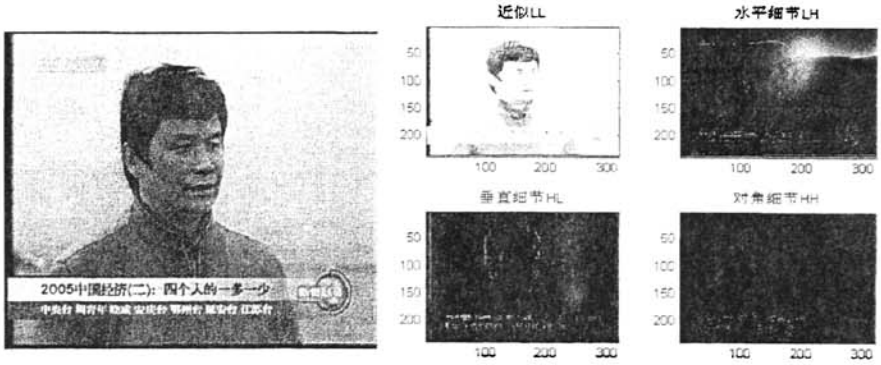
式中 $X(\omega)$ 和 $\Psi(\omega)$ 分别是 $x(t)$ 和 $\psi(t)$ 的傅立叶变换。

小波变换有以下特点：

(1) 有多分辨率(multi-resolution)，也叫多尺度(multi-scale)的特点，可以由粗及细地逐步观察信号。

(2) 可以看成用基本频率特性为 $\Psi(\omega)$ 的带通滤波器，在不同尺度 a 下对信号做滤波。由傅立叶变换的尺度特性可知，这组滤波器具有品质因数恒定，即相对窄带(带宽与中心频率之比)恒定的特点，a 越大相对频率越低。

(3) 适当地选择基小波，使 $\psi(t)$ 在时域上为有限支撑，$\Psi(\omega)$ 在频域上也比较集中，是可以使 WT 在时、频域都具有表征信号局部特征的能力，因此有利于检测信号的瞬态或奇异点。



(a) 原图像 (b) 一级小波分解图

图 4.3.1 视频图像一级小波分解图

如图 4.3.1 所示，图(a)为原图，图(b)为一级小波分解图，从图(b)中可以发现，在三个高频子带(LH、HL、HH)中，字幕区域表现非常明显，由于小波的局部显微特性，小波系数大的地方总是出现在图像的边缘部分，亦即小波分解后的细节分量中有能较好地体现文本位置的信息。本文选择 Haar 小波，因为 Haar 小波在检测边缘信息方面具有良好的性能，并且 Haar 小波计算简单，可用掩模运算来实现。Haar 小波的尺度函数和小波函数可分别描述为：

$$\phi(x)=\sum_{k\in Z}p_k\phi(2x-k)=\phi(2x)+\phi(2x-1) \tag{4.28}$$

$$W_H(x)=\sum_{k\in Z}q_k\phi(2x-k)=\phi(2x)-\phi(2x-1) \tag{4.29}$$

其中 $\phi(x)=\begin{cases}1 & for\ 0\le x<1\\ 0 & otherwise\end{cases}$，对于两尺度有序列 p_k：$p_0=p_1=1,\ p_i=0, i\ge 2$，同样对于 q_k：$q_0=q_1=1,\ q_i=0, i\ge 2$。

对于一幅图像 I：

$$I(x,y)=\begin{bmatrix} i_{0,0} & i_{0,1} & \cdots & i_{0,2N-1} \\ i_{1,0} & i_{1,1} & \cdots & i_{0,2N-1} \\ \vdots & \vdots & \vdots & \vdots \\ i_{2N-1,0} & i_{2N-1,1} & \cdots & i_{2N-1,2N-1} \end{bmatrix}_{2N\times 2N} \tag{4.30}$$

其二维 Haar 小波变换可用如下 Mallat 算法实现：

$$LL_{x,y}=\frac{1}{4}\sum_{k_1,k_2=0}^{1}p_{k_1}p_{k_2}i_{k_1+2x,k_2+2y}=\frac{1}{4}(i_{2x,2y}+i_{2x,2y+1}+i_{2x+1,y}+i_{2x+1,2y+1}) \tag{4.31}$$

$$LH_{x,y}=\frac{1}{4}\sum_{k_1,k_2=0}^{1}p_{k_1}q_{k_2}i_{k_1+2x,k_2+2y}=\frac{1}{4}(i_{2x,2y}-i_{2x,2y+1}+i_{2x+1,y}-i_{2x+1,2y+1}) \tag{4.32}$$

$$HL_{x,y}=\frac{1}{4}\sum_{k_1,k_2=0}^{1}q_{k_1}p_{k_2}i_{k_1+2x,k_2+2y}=\frac{1}{4}(i_{2x,2y}+i_{2x,2y+1}-i_{2x+1,y}-i_{2x+1,2y+1}) \tag{4.33}$$

$$HH_{x,y}=\frac{1}{4}\sum_{k_1,k_2=0}^{1}q_{k_1}q_{k_2}i_{k_1+2x,k_2+2y}=\frac{1}{4}(i_{2x,2y}-i_{2x,2y+1}-i_{2x+1,y}+i_{2x+1,2y+1}) \tag{4.34}$$

即可对图像 I 的 Haar 小波分解可用如图 4.3.2 所示的 Haar 模板作掩模运算实现，其计算效率显而易见是比较高的。

1	1
1	1

(a)低频分量

1	-1
1	-1

(b)水平细节分量

1	1
-1	-1

(c)垂直细节分量

1	-1
-1	1

(d)对角细节分量

图 4.3.2　Haar 小波模板

4.3.1.2 特征提取

将原始视频帧划分成 N×N 的图像块，N 取 16，每个图像块最多可以进行四次小波分解，但是最后一次分解使得图像每个子带只包含一个像素，几乎不包含有用的特征信息，因此对图像块作三级子波分解，每一级对应 4 个子波分量，即近似分量 LL、水平 LH、垂直 HL 以及对角细节分量 HH，然后对每一个分量计算均方值(m)、二阶中心距(μ_2)、三阶中心矩(μ_3)作为纹理特征，共计 3×4×3＝36 个特征。计算公式如下：

$$m(I)=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}I(i,j) \tag{4.35}$$

$$\mu_2(I)=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j)-M(I))^2 \tag{4.36}$$

$$\mu_3(I)=\frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(I(i,j)-M(I))^3 \tag{4.37}$$

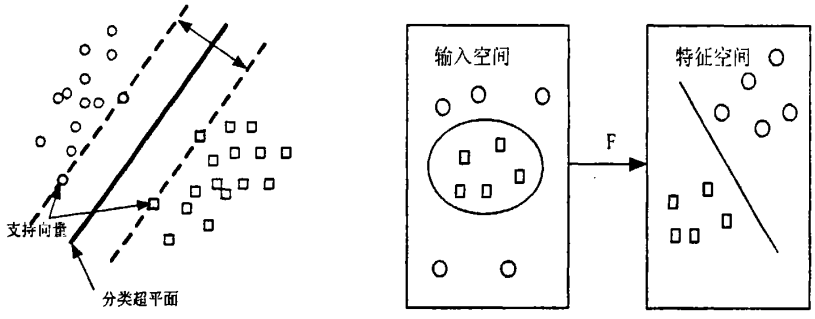
4.3.2 基于支持向量机的字幕区提取

4.3.2.1 支持向量机的概念

对于线性可分的训练样本集$\{(x_i,y_i),i=1,2,\cdots,n,x\in R^n,y\in\{-1,+1\}\}$，$n$维空间中的线性判别函数一般形式为：

$$g(x)=(w\cdot x)+b \tag{4.38}$$

若集合中的所有数据，都可以被分类超平面$(w\cdot x)+b=0$所正确划分，且离超平面(hyper plane)最近的样本点与超平面之间的距离间隔最大，则该分类面就是最优超平面，如图4.3.3(a)所示。



(a) 最优线性超平面 (b) 非线性可分

图 4.3.3 支持向量机

其中过两类样本中离分类面最近的点并且平行于最优分类面的超平面上的训练样本被称为支持向量(support vectors)，支持向量与超平面之间的距离需最大化(即边缘最大化)，一组支持向量可以唯一地确定一个超平面。

在线性可分的情况下，SVM 就是建立一个超平面(hyper plane)使得可分的两类数据到该平面的距离最大。

在现实世界中，很多分类问题都不是线性可分的，也就是说在原来的样本空间中，无法找到一个最优的线性分类函数，这使得支持向量机的应用具有很大的局限性。对于这类非线性可分问题，由于“一个复杂的模式识别分类问题，在高维空间比低维空间更容易线性可分”[61]，SVM首先通过一个非线性映射把原始数据映射到另一个称之为特征空间的新数据集上，使得新数据集在该特征空间上是线性可分的，然后在这个新空间中求最优线性分类面，如4.3.3(b)所示。

SVM 学习机的结构与神经网络类似，分为三层：输入层、隐含层和输出层。输入层接受输入数据，即识别特征。隐含层有两个功能：一是用非线性映射Φ，把输入数据从原始低维空间映射到高维持征空间；二是计算特征向量和支持向量的内积。在实际应用中，这两步是通过构造核函数一步来实现的：

$$K(x,y)=(\Phi(x)\bullet\Phi(y)) \tag{4.39}$$

其中，K、Φ和$\bullet$分别是核函数、高维非线性映射和内积。由于构造的核函数满足Mercer条件，因此在应用中只需考虑核函数K，而不必知道低维向高维的映射函数Φ，即只需指定特定的核函数K，而无需指定原始图像特征到高维特征的映射函数。输出层用于输出分类结果。

采用不同的核函数导致不同的SVM算法，目前常用的核函数主要有三类：

(1) 多项式核函数，$K(x_i,x)=[(x,x_i)+1]^q$，得到的是P阶多项式分类器；

(2) 径向基函数，$K(x_i,x)=\exp(-\|x_i-x\|^2/2\sigma^2)$，得到的是径向基分类器；

(3) sigmoid核函数，$K(x_i,x)=\tanh(v(x,x_i)+c)$得到的是一个两层感知器神经网络。

事实上，SVM方法虽然使变换空间的维数增加很多，但是在求解最优分类面时并没有增加多少计算量，这是因为在一般的情况下，变换空间的内积可以用原样本空间中的变量直接计算得到。本文使用径向基函数作为核函数。

4.3.2.2 支持向量机(SVM)的训练

选取100张来自新闻视频中的视频帧，其中大部分包含主题字幕，将其作为训练样本，对支持向量机进行训练。在实验中将字幕子块定义为+1，非字幕子块定义为−1。输入数据为每个图像子块36个特征值以及对这个子块是否为字幕的标注。

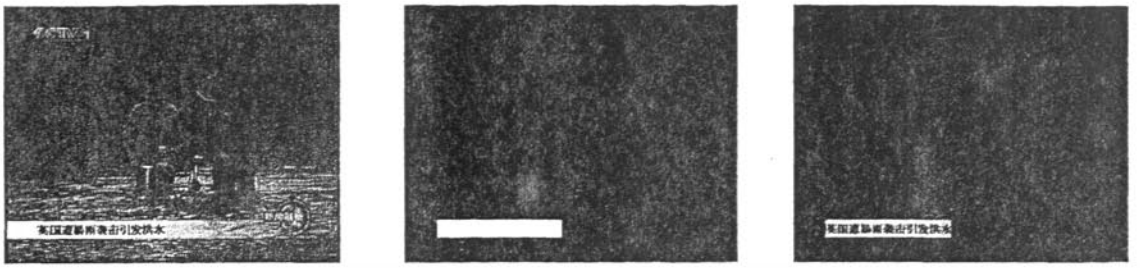
4.3.2.3 字幕区提取

训练好SVM后，使用一个N*N窗口来扫描帧，区分每个窗口是否是字幕。对每一个输入数据，如果输出为正，则该输入块被判定为字幕块，否则为非字幕块。扫描步长取得越大，要处理的窗口数目就越少，计算量越小但检测精度越差，反之步长越小，要处理的窗口数量就越多，检测精度越高但计算量越大。在精度和速度之间权衡，每次将窗口移动4个像素位置。

在检测过程中，如果一个单个的窗口被划分为字幕，所有此窗口中的像素就被标记为字幕，那些没有被任何文本窗口覆盖的窗口标记为非字幕。

4.3.2.4 去噪处理

由于背景图像的复杂性，一些表现出字幕块特性的背景图像被误判为字幕块，这种情况难以避免，通常表现为一些小的孤立区域，必须作进一步处理以尽可能去处被误判的字幕区域。由于新闻视频中的主题字幕，基本上是从左到右、水平聚集排列，其它排列形式很少出现，可以通过对其进行连通性分析，来消除大部分孤立噪声块。检查候选字幕块(i,j)的4连通域，与任何一个其它候选字幕块(x,y)的4连通域是否连通，如果连通则判(i,j)为真字幕块，否则为噪声块加以去除。去除噪声块后，得到的主题字幕区内可能存在空洞；选取3*3的结构算子进行膨胀处理，填补空洞；再用同样的结构算子进行腐蚀处理，保证主题字幕区处理前后一致，得到标记图，如图4.3.4(b)，最后用标记图对原始图像进行掩模运算可以提取主题字幕区，如图4.3.4(c)所示。



(a) 原主题字幕帧 (b) 标记图 (c) 分割好的字幕区域

图4.3.4 主题字幕区提取效果图

4.3.3 主题字幕区提取流程图

根据以上分析，主题字幕区提取的算法流程如图 4.3.5 所示：

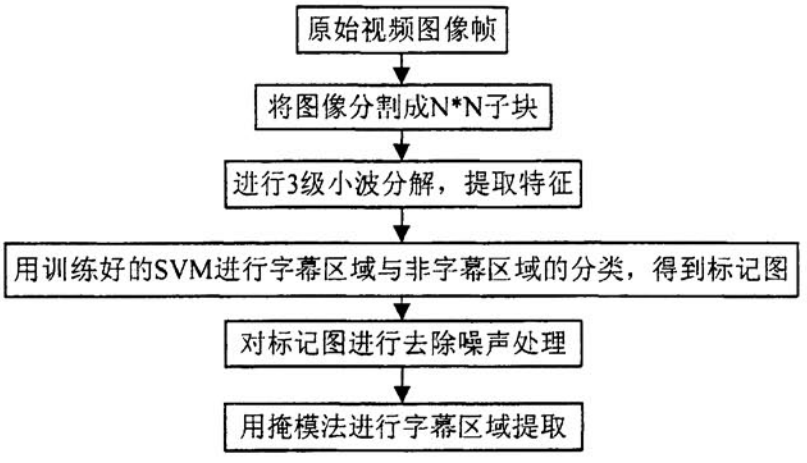


图 4.3.5 主题字幕区提取流程图

4.3.4 实验结果

在Visual C++ 6.0下实现了该算法，并进行了大量的实际测试。以新闻联播视频为素材，随机选取了50幅主题字幕帧作为测试样本，图像尺寸为320×240，把样本分为两组进行测试，测试结果见表4.3.1，为了检测效果，采用查准率和查全率指标来进行评价。漏判的主要是因为字幕笔画比较少，比如“一”字，自身形成孤立块，误认为是噪声块，从而导致漏判；而误判的主要原因是少部分非字幕块具有字幕的纹理特性，导致误判为字幕。

$$\text{查全率}=\frac{\text{准确判定的字符个数}}{\text{准确判定的字符个数}+\text{漏判块数的字符个数}}$$

$$\text{查准率}=\frac{\text{准确判定的字符个数}}{\text{准确判定的字符个数}+\text{误判的字符个数}}$$

表4.3.1 字幕行检测结果

实验	待测的字符个数	准确判定的字符个数	漏判个数	误判个数	查准率	查全率
1	338	325	13	20	96.15%	94.20 %
2	372	350	22	27	94.13%	92.84%

4.4 主题字幕的二值化及识别

在主题字幕区域提取后，分辨率还比较低，必须进行插值放大处理。由于现有的OCR 识别软件只适于二值化的字符，因此还必须做进一步的二值化处理。本节实现主题字幕的插值放大、二值化处理、字符的分割以及 OCR 识别。

4.4.1 字幕的插值

在新闻视频的后期制作过程中，为了使视频中的字幕区域不遮挡视频画面，通常情况下视频字幕的分辨比较低，例如 320× 240 视频帧中平均字符大小仅为 12× 10，因此在将字符送往 OCR 软件识别前，需要对字幕区进行插值放大，提高分辨率。

图像插值的基本思想[7]是：如果要想获得N倍分辨率的图像，则将原图像的水平和垂直方向的相邻相素之间，各拉开N-1个相素的距离。插值在图像的扩大和缩小中有很广的应用，其通常是利用曲线拟合的方法，通过离散的采样点建立一个连续函数，用这个重建的函数便可以求出任意位置的函数值，如图4.4.1所示。

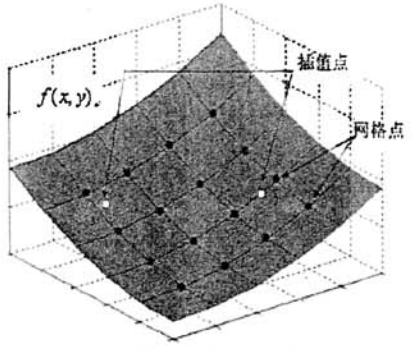


图 4.4.1 图像的插值

对于等间隔离散数据，插值可以表示为：

$$f(x)=\sum_{k=0}^{K-1}C_k h(x-x_l) \tag{4.40}$$

其中 $h(\cdot)$ 为插值核函数，C_k 为权系数。

插值算法的数值精确度及计算量，与插值核密切相关，插值核的设计是插值算法的

核心。最常用的插值方法[7]有：最近邻插值(Nearest neighbor)、双线性插值(Bilinear interpolation)和双三次插值(Bicubic interpolation)。

(1) 最近邻插值(Nearest neighbor)

又称为零阶插值算法，令输出像素的灰度值，等于离它所映射的位置最近的输入像素的灰度值。最近邻插值计算十分简单，许多情况下其结果也可令人接受。然而当图像中包含像素之间灰度级有变化的细微结构时，最近邻插值法会在图像中产生方块或锯齿效应。从计算量的角度来说，最近邻插值是最简单的插值方法，在这种算法中，每一个插值输出像素的值就是在输入图像中与其最临近的采样点的值。算法的数学表达式为：

$$f(x)=f(x_k) \qquad \frac{1}{2}(x_{k-1}+x_k)<x\le\frac{1}{2}(x_k+x_{k+1}) \tag{4.41}$$

最近邻插值及傅立叶谱如图4.2.2所示：

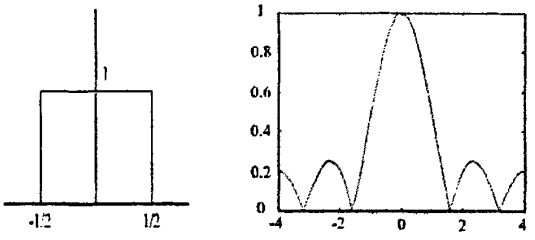


图 4.4.2 最近邻法的插值核及其傅立叶谱

从图4.4.2傅立叶谱上可以看出，它与理想低通滤波器的性质差别很大。用这种方法实现大倍数放大处理时候，在图像中可以明显地看出块状效应。

(2) 双线性插值(Bilinear interpolation)

又称为一阶插值算法，是对最近邻法的一种改进，它是利用映射点在输入图像中的4个邻点的灰度值，对映射点进行插值。因为通过4点确定一个平面是一个过约束问题，所以在一个矩形栅格上进行一阶插值，就需要用到双线性函数。

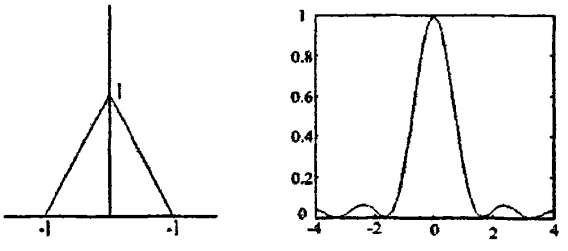


图 4.4.3 线性插值的插值核及其傅立叶谱

从图4.4.3傅立叶谱上可以看出，它的频域特性优于最近邻插值函数，其频谱的旁瓣远小于主瓣，表明它的带阻特性较好。不过，仍有大量高频成分漏入通频带，造成了一定的混叠。此外，通频带在一定程度上被减弱，会使插值后的图像变模糊，从而损失一些细节。

(3) 双三次插值(Bicubic interpolation)

双三次插值的插值核为三次函数，其插值邻域的大小为4×4。它的插值效果比较好，但相应的计算量也较大，三次插值多项式s(u)为4.57式定义。

$$s(u)=\begin{cases}1-2|u|^2+|u|^3 & |u|\le 1\\ 4-8|u|+5|u|^2-|u|^3 & 1\le|u|\le 2\\ 0 & |u|\ge 2\end{cases} \tag{4.42}$$

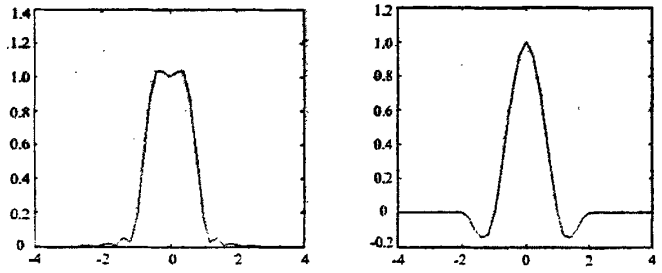


图 4.4.4 三次插值的插值核及其傅立叶谱

双三次插值法的插值核及傅立叶谱如图4.4.4所示。

如图4.4.5可知，最近邻插值的效果最差，双三次插值的效果最好，双线性插值的效果介于两者之间。因此本文采用双三次插值来提高字符的分辨率。

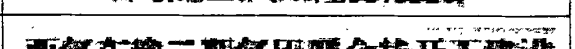
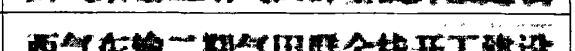
原字幕图像	
1.5倍最近邻插值效果图	
1.5倍双线性插值效果图	
1.5倍双三次插值效果图	

图 4.4.5 三种插值方法效果图

4.4.2 字幕的二值化

二值化在文字图像分析中具有关键作用，利用光学字符识别(OCR)软件进行字符识别时，要求输入的图像是二值图像。二值图像质量直接关系到后续的识别效果，因此如何获得高质量二值图像是值得关注的研究课题。二值化算法，又称阈值算法，目的是找出一个合适的阈值，将图像区域区分为背景和前景两部分。阈值的选取是非常重要，阈值太大则边界变粗，对噪声图像将出现随机的噪声像点，阈值太小则边界不连续甚至消失。

字幕检测的根本目的是为了识别出其中的字符，因此还必须对识别字幕区域二值化，且二值化的质量直接影响识别的效果。通过观察新闻视频主题字幕，发现其背景色大多都比较单一，多为单色，字幕与背景有明显的对比度，本文选用自动单阈值分割(Otsu

算法)进行二值化[30]，Otsu 算法选取出来的阈值非常理性，对各种情况的表现都较为良好，可以说是最稳定的分割。

Otsu算法，又称为大津法，由大津于1979年提出，它是一种最大类间方差法，能够自动选取阈值，它通过寻找一个最大差值来分割图像成两部分。因为方差是图像灰度分布均匀性的一种度量，方差值越大，说明构成图像的两部分差别越大，当部分目标错分为背景或部分背景错分为目标，都会导致两部分差别变小，因此使类间方差最大的分割意味着错分概率最小。算法对输入的灰度图像的直方图进行分析，将直方图分成两个部分，使得两部分之间的距离最大。划分点就是求得的阈值。计算方法如下：

设一副图像的灰度值1~L级，灰度值为i的像素数为n_i，总像素数为N，各灰度值的概率为p_i，公式如下：

$$N=\sum_{i=1}^{L} n_i \quad , \quad p_i=\frac{n_i}{N} \tag{4.43}$$

然后用 K 将其分为两组$C_0=\{1\sim k\}$和$C_1=\{k+1\sim m\}$，C_0和C_1产生的概率分别是ω_0和ω_1，C_0组和C_1组的平均值分别是μ_0和μ_1，公式如下

$$\omega_0=\sum_{i=1}^{k} p_i=\omega(k) \quad , \quad \omega_1=\sum_{i=k+1}^{L} p_i=1-\omega(k) \tag{4.44}$$

$$\mu_0=\sum_{i=1}^{k}\frac{ip_i}{\omega_0}=\frac{\mu(k)}{\omega(k)} \quad , \quad \mu_1=\sum_{i=k+1}^{L}\frac{ip_i}{\omega_1}=\frac{\mu-\mu(k)}{1-\omega(k)} \tag{4.45}$$

其中，$\mu=\sum_{i=1}^{L} ip_i$是整体图像的平均值，$\mu(k)=\sum_{i=1}^{k} ip_i$是阈值为$k$时灰度的平均值，所以全部采样的灰度平均值为$\mu=\omega_0\mu_0+\omega_1\mu_1$，两组间的方差用下式求出：

$$\sigma^2(k)=\omega_0(\mu_0-\mu)^2+\omega_1(\mu_1-\mu)^2 \tag{4.46}$$

从1-L间改变k，求式4.46的最大值的k，即为$\max\sigma^2(k)$的k^*的值，此时k^*值便是求得的阈值。Otsu算法是基于图像像素的灰度值分类，按照类间方差最大的原则获得最佳阈值。该算法简单、快速，具有抗噪性良好等明显优点，可以说是最稳定的分割。由图4.4.6可以看出，Otsu算法基本上不会出现笔画断裂现象，得到的效果图较为清晰。

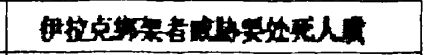
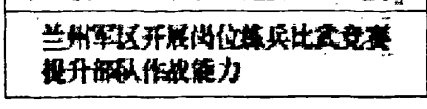
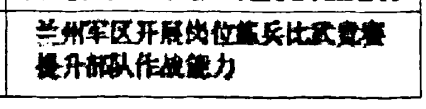
双三次插值后的效果图	Ostu二值化效果图
	
	
	
	

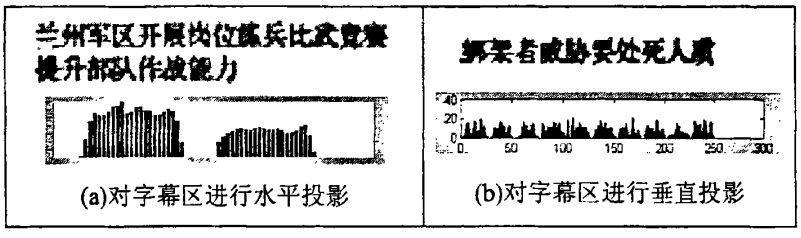
图 4.4.6 Otsu算法后的效果图

4.4.3 字幕的分割

对字幕进行二值化后，下一步就是从字幕区域中把每个字符分割出来，本文采用投影法[40]进行分割，并且对分割后出现的常见问题进行讨论。

4.4.3.1 投影法分割

采用投影法对二值化的主题字幕区域进行字符分割。观察新闻主题字幕可知，主题文字一般都是水平对齐的，因此，对二值化后的边缘图像在水平方向作投影会出现一些很陡的峰值，各峰的宽度对应文字的高度，可得到清楚的行边界。同样，对一行文字区域的图像在垂直方向作投影，相邻两个字符的边界会形成一个波谷，从而可以得到字与字之间的边界。因此，可以对字幕区域先进行水平投影，确定字幕的行边界；再进行垂直投影，确定单个字幕的边界。



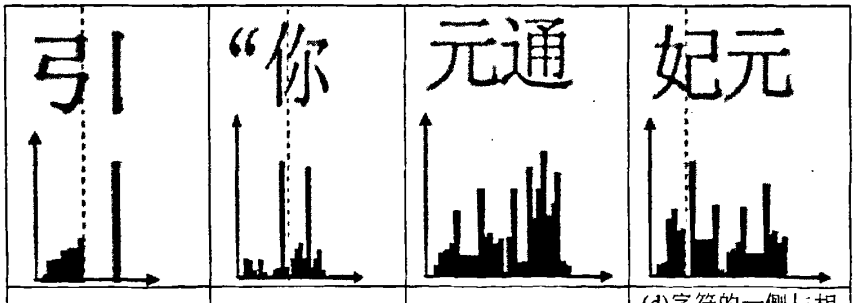
(a)对字幕区进行水平投影　(b)对字幕区进行垂直投影

图 4.4.7 投影分割字符

图 4.4.7 中，图(a)对字幕区进行水平投影，有两个明显的峰值，说明该字幕区域有两行字幕；图(b)对字幕区进行垂直投影，其中的波谷对应相邻字的边界。图(a)所示的结果是令人满意的，由于字幕水平对齐，基本上不会出现字幕行错分的情况；图(b)可以看出“威胁”字符间出现了粘连现象，容易出现字符错分，如果不加处理，会造成后续的 OCR 识别错误。在字幕行中有两类像素：字幕和背景。在每一帧中字幕和背景像素的灰度值常常混杂在一起，二值化后将导致字幕之间通过背景粘连。下面一小节将针对字符错误分割的几种典型情况采取相应的解决方法。

4.4.3.2 错误分割的几种情况

投影法对于以下几种情况[40]，并不能正确地分割字符。



(a)左/右偏旁结构　(b)含有标点符号　(c)相邻字符间粘连　(d)字符的一侧与相邻字符无清晰边界

图 4.4.8 字符分割的几种特殊情况

① 过切分：由于许多汉字由几部分组成，特别是对于一些左/右偏旁结构的汉字，很容易被过切分。如图4.4.8(a)所示，“引”字的垂直方向投影有峰谷出现，这将使“引”被错误地分割为“弓”和“丨”两个字符。

② 一个文字被切分，并与另一个文字粘连：某一个字符是左/右偏旁结构，同时该字符的一侧与相邻字符没有明显清晰的边界。如图 4.4.8(d)所示，“妃”字的“女”与“己”是最低峰谷，而“己”与“元”的边界不清晰，分割结果是“女”单独作为一个字，而“己元”被错误地分割为一个字。

③ 两个文字粘连：造成字符粘连的原因有两点，一是主题文字中含有标点符号，如图 4.4.8(b)所示，标点符号字符的峰值很低，有时会造成标点符号与相邻字符“你”被分割为一个字符；二是相邻字符之间的边界可能并不是很清晰，如图 4.4.8(c)所示，“元”和“通”容易被分割成一个字符；

由于字幕的每个字符的字号都是相同的，绝大多数字符有相似的高度和宽度，因此，从上面的简单分割，可以粗略得到字幕中一般字符的宽度 DW 和高度 DH，以及一般相邻字符之间的间隙为 DG，设简单分割后的某个字符的高度为 DH_i，宽度为 DW_i，与相邻左/右字符间的最大距离为 G_i。下面对上述的特殊情况分别讨论[40]。

(1) 用投影法分割后的字符宽度 $DW_i > 1.2DW$，说明有字符粘连。

- 两个字符粘连

如果垂直方向投影中的所有峰值高度在正常范围内，则在垂直方向投影的最小值处将粘连字符分割为两个字符，如图 4.4.9(a,b)所示。因为字符边界粘连的像素很少。

- 粘连字符中含有标点符号

如果垂直方向投影中有高度低于 $0.4DH$ 的峰，且该粘连字符与相邻左/右字符间的最大距离 $DG_i > 1.1DG$，则在低峰值峰与相邻高峰值峰的峰谷处将粘连字符分割为两个字符，如图 4.4.9(c)所示。因为标点符号字符的高度比一般字符低很多，而且它与相邻字符之间的间距比一般字符要大，

(2) 用投影法分割后的字符宽度 $DW_i < 0.8W$，说明有字符过切分的情况。

将该字符与和它间距小于 $0.8DG$ 的相邻字符合并成一个字符，如图 4.4.9(d,e)所示。因为汉字内偏旁的间隔比汉字间的空隙小。

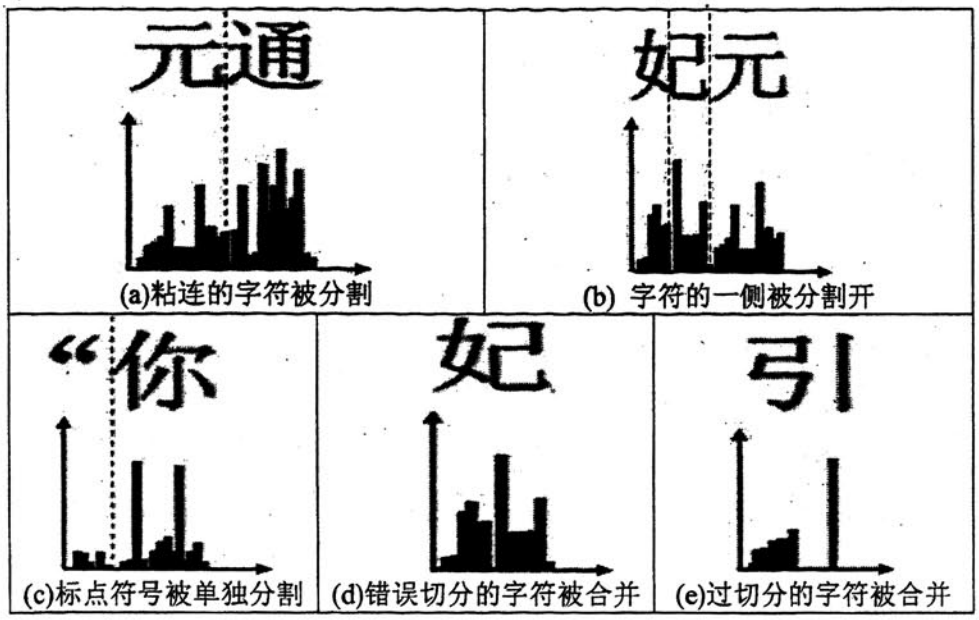


图 4.4.9 字符重新分割结果示例图

4.4.3 字幕的识别

将二值化后的效果图经字符分割后，输入到标准的OCR软件包，得到文字的ASCII码，汉王OCR识别的界面如图4.4.10所示。

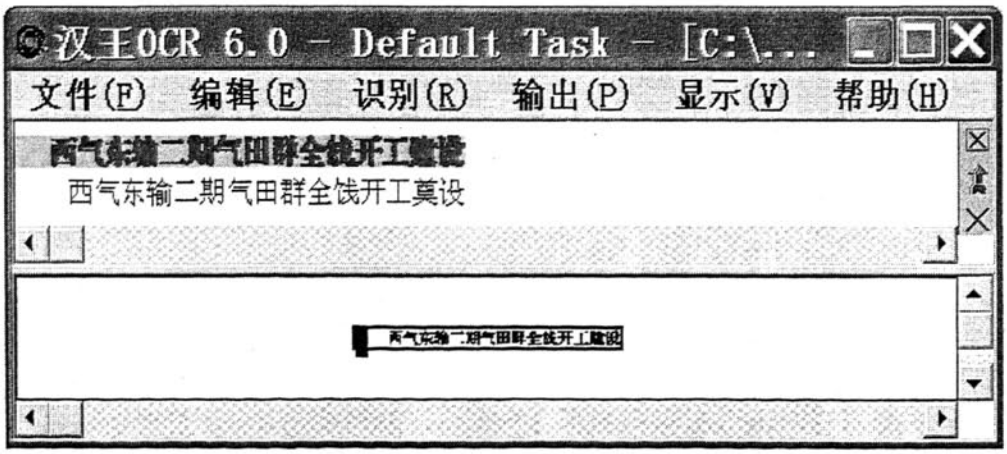


图4.4.10 字幕识别结果

如图4.4.10，其中“线”和“建”这两个字被误识别，例如“建”因为笔画紧凑，有很多横，经过插值放大后，再经过二值化后，字体比较模糊，因此被误识别。本节对提取出来的主题字幕区，进行插值、二值化、字符分割后，得到187个字符，用汉王OCR6.0软件进行识别，得到的实验结果如表4.4.1。实验结果表明，字幕经过本节算法的处理后，具有较高的识别率。

表 4.4.1 OCR识别结果

待识别的字符数	正确识别的个数	错误识别的个数	识别率
187	170	17	91%

4.5 本章小结

本章完成了新闻主题字幕的提取，分为四部分。

第一部分介绍了新闻视频中的字幕分类。

第二部分设计了基于 3/10 时空切片的字幕帧边界检测算法，来检测主题字幕的起始帧和消失帧，实验表明该方法具有较高的查准率和查全率。

第三部分采用小波变换和支持向量机的方法，对主题字幕区进行提取，实验证明能够实现主题字幕的准确定位。

第四部分首先对主题字幕区进行插值放大，提高分辨率；然后采用 Otsu 算法进行二值化处理、投影法分割字符，并对投影法产生的错误分割情况，采取相应的解决方法；最后将分割后的字符用汉王 OCR6.0 软件识别。

5 基于主题字幕提取的新闻故事分段和视频检索

新闻主题字幕的反复出现，是区别于其它视频的一个显著特点，它常常是新闻故事分段的标志。本文第四章中实现了新闻主题字幕的提取，可以根据主题字幕提取的结果，对每个新闻故事的开始和结束位置进行定位，从而将连续的新闻视频数据流分割成一系列单独的新闻故事。对新闻进行分段后，对每个新闻故事片段用主题字幕来标注，为新闻视频建立索引，从而可以完成基于关键字的新闻视频检索。

5.1 基于主题字幕提取的新闻故事分段

新闻视频的结构特征比较明显，其主体内容是一系列新闻故事。准确地定位每个新闻故事的开始和结束位置，是分析和检索新闻视频片段的重要依据。在电视新闻视频中，每个新闻故事都有一个简要的主题，而且75%以上的主题字幕都出现在新闻故事的开始几个镜头；而且在两个连续的新闻故事边界处，必然存在一个持续时间相对较长的静音片段[35]。本节设计了新闻故事分段算法，在第四章主题字幕提取的基础上，通过静音检测和主持人镜头检测，将连续的新闻视频数据流分割成一系列单独的新闻故事。

5.1.1 音频检测

经过对大量新闻节目中音频流的观察发现，它们一般都有比较固定的格式，具有一些显著的特征。新闻开始时，一般是一段音乐、一个静音区间，然后是主持人镜头。在新闻故事变换处，主持人的声音会有一个明显的停顿。由于新闻视频中的音频流具有较高的信噪比，因此，在主持人的声音停顿处音频流中将出现一个静音区间。通过对音乐的检测、以及静音的检测，从而达到辅助新闻故事分段的目的。

5.1.1.1 音频信号特征

短时能量和短时平均过零率，是语音信号中最基本的、也是最重要的时域特征。计算简单且运算量小，广泛应用于语音信号处理的各个领域[57] 。

(1) 短时能量(STE)

短时能量可以用来衡量音频信号的强度。短时能量函数的定义是

$$E_n = \sum_{m=n-N+1}^{n}[x(m)\omega(n-m)]^2 = \sum_{m=n-N+1}^{n}x^2(m)\bullet h(n-m) \tag{5.1}$$

式中，$h(n)=\omega^2(n)$为窗口函数。式(5.2)给出了采样点n处的短时能量。在窗函数$h(n)$的处理下，它等于从$n-N+1$到n的N个采样$x(m)$的平方和。一个简单的窗为矩形窗，它的窗函数定义如下

$$\begin{cases} h(n)=1 & 0\le n\le N-1 \\ h(n)=0 & \text{其他} \end{cases} \tag{5.2}$$

窗的长短，对于能否由短时能量反映语音信号的幅度变化，将起决定性的影响。如果窗选的很长(即 N 很大)，它等效于很窄的低通滤波器，此时 E_n 随时间的变化很小，不能得到平滑的能量函数。因此，应该选择合适的短时窗，短时能量才能反映语音信号快速的幅度变化。

(2) 短时平均过零率(ZCR)

过零，是指时域波形穿过坐标轴，表现在离散信号上就是相邻两个采样值异号。短时过零率，即单位时间内过零发生的次数，其定义如下：

$$Zero(n)=\sum_{m=-\infty}^{\infty}|\,\mathrm{sgn}[x(m)]-\mathrm{sgn}[x(m-1)]\,|\omega(n-m) \tag{5.3}$$

其中

$$\begin{cases} \mathrm{sgn}[x(n)]=1 & x(n)\ge 0 \\ \mathrm{sgn}[x(n)]=-1 & x(n)<0 \end{cases},\quad \begin{cases} \omega(n)=1/(2N) & 0\le n\le N-1 \\ \omega(n)=0 & \text{其他} \end{cases} \tag{5.4}$$

一般说来，短时平均过零率的最主要用处是分辨清音和浊音，有声与无声。

5.1.1.2 静音检测和音乐检测

如图 5.1.1，Signal 是中央一台新闻联播开始的 27 秒音频片段，开始的 24 秒是音乐片段，大约第 25 秒是静音片段，然后是播音片段。STE 表示短时平均能量，ZCR 表示短时平均过零率。由图可以看出，对于音频信号而言，静音片段的短时能量和短时平均过零率，要比非静音片段明显小的多；音乐的短时能量通常较高，且信号频率一般稳定在一定的频率范围内，过零率变化缓慢不表现出突然升高或降落的起伏特性。因此通过分析音频的短时能量和短时平均过零率这两个参数，可以检测出静音片段和音乐片段。

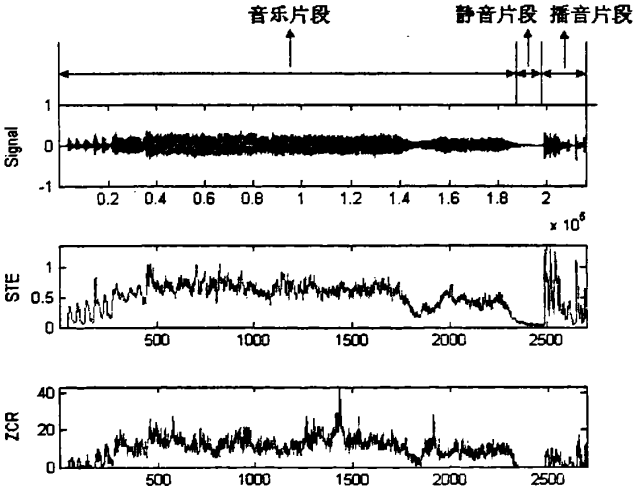


图 5.1.1 音频片段

将连续的音频流按时间关系划分为“音频帧”和“音频片”两个层次，音频帧长度设定为 20ms，帧移设定为 10ms，选择连续 50 个音频帧为一个音频片段。

(1) 静音检测

由于在两个连续的新闻故事边界处，必然存在一个持续时间在 1 秒以上的静音片段，所以，检测这些静音片段，对于新闻故事的分段有着重要的意义。

设一个音频片段音频帧序列为 $F=\{f_i,...,f_j,...,f_k\}$，定义 $f_{ste}^{(j)}$ 为第 j 帧的短时能量，$f_{zcr}^{(j)}$ 为第 j 帧的短时平均过零率。

如果 $f_{ste}^{(j)}<\lambda_{ste}^{(s)}$ 且 $f_{zcr}^{(j)}<\lambda_{zcr}^{(s)}$，则音频帧 f_j 为静音帧，定义静音帧标志为：

$$S(f_j)=\begin{cases}1 & f_j\text{是静音帧}\\ 0 & f_j\text{不是静音帧}\end{cases} \tag{5.5}$$

定义音频片段静音帧比例为：

$$P_S(j)=\left(\sum_{i\le j\le k}S(f_j)\right)\Big/(k-i) \tag{5.6}$$

如果静音帧比例 $P_s(i)\ge 0.8$，则判断该音频片段为静音片段。

(2) 音乐检测

本节开始时曾指出，新闻开始时，一般是一段音乐、一个静音区间，然后是主持人镜头。通过对音乐的检测；可以检测到主持人帧模板。所采用的方法与静音检测相似，不同之处在于阈值的选取。

设一个音频片段 F 中的音频帧序列为 $F=\{f_i,...,f_j,...,f_k\}$，定义 $f_{ste}^{(j)}$ 为第 j 帧的短时能量，$f_{zcr}^{(j)}$ 为第 j 帧的短时平均过零率。

如果 $f_{ste}^{(j)}>\lambda_{ste}^{(m)}$ 且 $f_{zcr}^{(j)}<\lambda_{zcr}^{(m)}$，则音频帧 f_j 为静音帧，定义音乐帧标志为：

$$M(f_j)=\begin{cases}1 & f_j\text{是音乐帧}\\ 0 & f_j\text{不是音乐帧}\end{cases} \tag{5.7}$$

定义音频片段静音帧比例为：

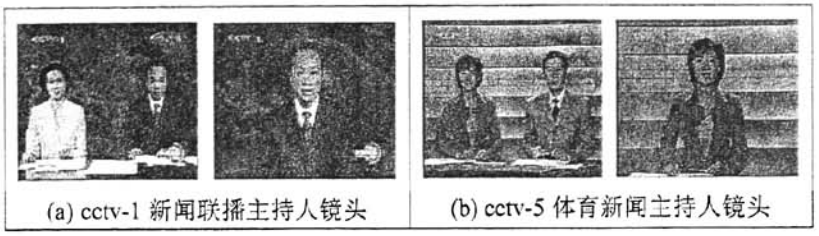
$$P_M(j)=\left(\sum_{i\le j\le k}M(f_j)\right)\Big/(k-i) \tag{5.8}$$

如果音乐帧比例 $P_s(i)\ge 0.8$，则判断该音频片段为音乐片段。

5.1.2 主持人镜头检测

由于主持人镜头是新闻视频中的重要结构特征，因此，主持人镜头的检测始终是新闻视频分析的一个重要方面。许多研究者对这个问题进行过研究和探索，比如：利用模板匹配来进行检测[36]；利用主持人镜头会在整个视频段中反复出现，并以此作为检测的

依据[37]；通过对主持人镜头建立结构模型，利用运动特征和相似匹配来进行检测、提取主持人面部的肤色持征[38]等等。这些方法效果大都不错，但算法都比较复杂，计算量较大。



(a) cctv-1 新闻联播主持人镜头　(b) cctv-5 体育新闻主持人镜头

图 5.1.2　主持人镜头实例图

新闻视频中的主持入镜头，是一类具有鲜明特征的镜头，其一般形式为一个或两个主持人，在固定的演播室背景前进行新闻报道，主持人镜头实例如图 5.1.2。通过观察可以发现，主持人的位置以及字幕、台标和节目标志的出现位置，都有严格的规定，由此可以建立了主持人镜头的空间结构模型(图 5.1.3)。图中，区域 A-D 分别代表主持人、台标、字幕和节目标志所出现的区域。从简化算法和降低计算复杂性方面考虑，本文根据主持人镜头的背景不变性进行检测[39]。从不变的背景出发，通过色矩计算和模板匹配来进行主持人镜头的检测。

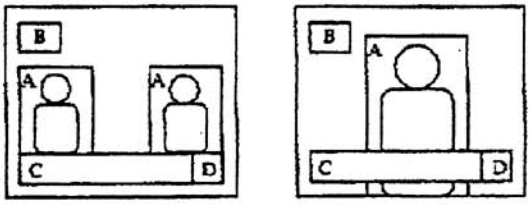


图 5.1.3　主持人模板图

本文的主持人镜头检测方法如下：提取到主持人帧模板，计算如图 5.1.4 各子块的色矩作为模板色矩，通过计算关键帧各子块的色矩向量与模板色矩向量的欧式距离，进行匹配，从而判定关键帧是不是主持人帧，从而判定关键帧所在镜头是不是主持人镜头。在音频特性上，第一个主持人镜头出现之前会有一段音乐过渡，并且从音乐向语音的过渡中间，有一个较长的静音片段。由于音视频具有同步性，检测到静音帧后的第一或第二帧的图像必定是主持人帧，从中可以提取到主持人帧的模板。

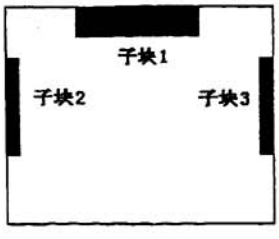


图 5.1.4　主持人模板图

色矩是由 Stricker 和 Orengo 提出的一种简单而有效的颜色特征[14]。它的数学基础是图像中任何的色彩分布均可以用它的矩来表示。由于颜色分布信息主要集中在低阶矩中，这里仅用色彩的一阶矩(mean，均值)、二阶矩(variance，方差)就足以表达图像的颜色分布，其数学表达式为：

$$\mu_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \tag{5.9}$$

$$\sigma_i = (\frac{1}{N}\sum_{j=1}^{N}(p_{ij}-\mu_i)^2)^{\frac{1}{2}} \tag{5.10}$$

其中，p_{ij}表示图像中第j个像素的第i个分量，这里在 HSI 颜色空间进行计算。

主持人镜头具体算法步骤如下：

Step1. 根据音视频的同步性，先是一段音乐，检测到静音帧后的第一或第二帧的图像必定是主持人帧，提取主持人帧的模板。

Step2. 计算主持人帧模板的各子块色矩作为模板色矩。

Step3. 计算各关键帧的各子块色矩，与主持人帧模板的模板色矩进行匹配，确定关键帧是否是主持人帧，从而确定关键帧所在镜头是否是主持人镜头。

实验采用中央一台长达 30 分钟的新闻联播进行主持人镜头检测，共检测出 8 个主持人镜头，无误检，无漏检，查准率和查全率都达到了 100%。

5.1.3 新闻故事分段

作为一种重要的视频类型，新闻视频始终是视频内容分析和检索研究领域的一个重要研究对象。新闻视频的结构特征比较明显，其主体内容是一系列新闻故事单元。准确地定位每个新闻故事单元的开始和结束位置，是分析和检索新闻视频片段的重要依据。本节设计了新闻故事分段算法，综合音频检测(静音和音乐检测)、主持人帧检测和主题字幕文本这几个方面，把连续的新闻视频数据流分割成一系列单独的新闻故事单元。

5.1.3.1 新闻故事特征

通过综合分析新闻视频的视觉特征和音频特征、文字特征，发现新闻故事有如下特征：

(1) 任意一段新闻视频均由一系列新闻故事构成。

(2) 在每一个新闻故事中，总会在特定的区域，以特定的格式出现且仅出现一次反映该故事主题内容的主题字幕。大部分主题字幕都出现在场景的开始几个镜头，即主题字幕离新闻故事的左边界近，因此通过确定所有新闻故事的左边界，确定新闻故事的左右边界。

(3) 在两条相邻主题字幕之间，必然存在一个故事分段处，并且在分段处主持人的声音会有一个明显的停顿，新闻视频中的音频流具有较高的信噪比，因此该处的音频流

中将出现一个持续时间相对较长静音片段，可以通过静音检测将分割两个新闻故事[35]。

(4) 主持人镜头中包含字幕，这类新闻通常都是公告性新闻。例如主持人镜头中主题字幕是“美国总统布什致信朝鲜领导人金正日”，没有具体的新闻故事视频信息，仅仅通过主持人的音频信息播报新闻故事，认为主持人镜头为一个新闻故事。

(5) 按新闻故事中主持人镜头中的有无主题字幕，可将主持人镜头分为两类。对没有主题字幕的主持人镜头，音频是对下一个新闻故事的高度概括，可以视其为普通的主持人镜头，不将其加入下一个新闻故事中；对有主题字幕的主持人镜头，与主题字幕相对应的新闻故事只有主持人播音的镜头，没有详细的新闻故事详情，将主持人镜头分割为一个新闻故事。

5.1.3.2 算法详述

假设一个新闻视频中的新闻故事序列为$\{S_1,S_2,...,S_n\}$，且其对应的主题字幕为$\{C_1,C_2,...,C_n\}$。对于新闻故事$S_i\ (1\le i\le n)$，设其对应的主题字幕C_i所在镜头为L_k。根据主题字幕与主持人镜头的关系，将故事分段分为三种情况：

(1) 主题字幕C_i包含在主持人镜头中，此时主持人镜头代表一个新闻故事，认为新闻故事的左边界是主持人镜头的左边界；

(2) 主题字幕C_i是主持人镜头后的第一个主题字幕，此时主持人镜头中的音频是对新闻故事S_i的概括，认为S_i的左边界是主持人镜头的右边界。

(3) 当主题字幕C_i既不包含在主持人镜头中，也不是主持人镜头后的第一个主题字幕，从镜头L_k开始，向前依次判断每相邻两个镜头切变处是否伴随静音，如果在镜头$L_{(k-j-1)}$到$L_{(k-j)}$的切换处检测到静音片段，则镜头$L_{(k-j)}$的右边界是新闻故事的左边界。

如果所有新闻故事的左边界都确定，则确定了所有新闻故事的左右边界，从而完成了整个新闻视频的新闻故事分段。根据左边界完成新闻故事分段，可能会将主持人镜头误分割到上一个新闻故事中。因此，如果新闻故事S_i右边界是主持人帧，即S_i最后一个镜头是主持人镜头，将主持人镜头从新闻故事中去除，认为新闻故事S_i的右边界是主持人镜头的左边界。

也可以将故事分段分为主题字幕是否包含在主持人镜头中两种情况。对于主持人镜头后的第一个新闻故事，通过主持人镜头的左边界确定新闻故事的右边界，可以减少由非新闻故事边界的静音片段引起的误判。因此本文将新闻分段分为三种情况。分段完毕后，用主题字幕对分割的每个独立新闻故事进行文本标注，为新闻故事建立索引，从而完成基于关键字的新闻视频检索。

新闻故事分段算法的流程图如图5.1.5：

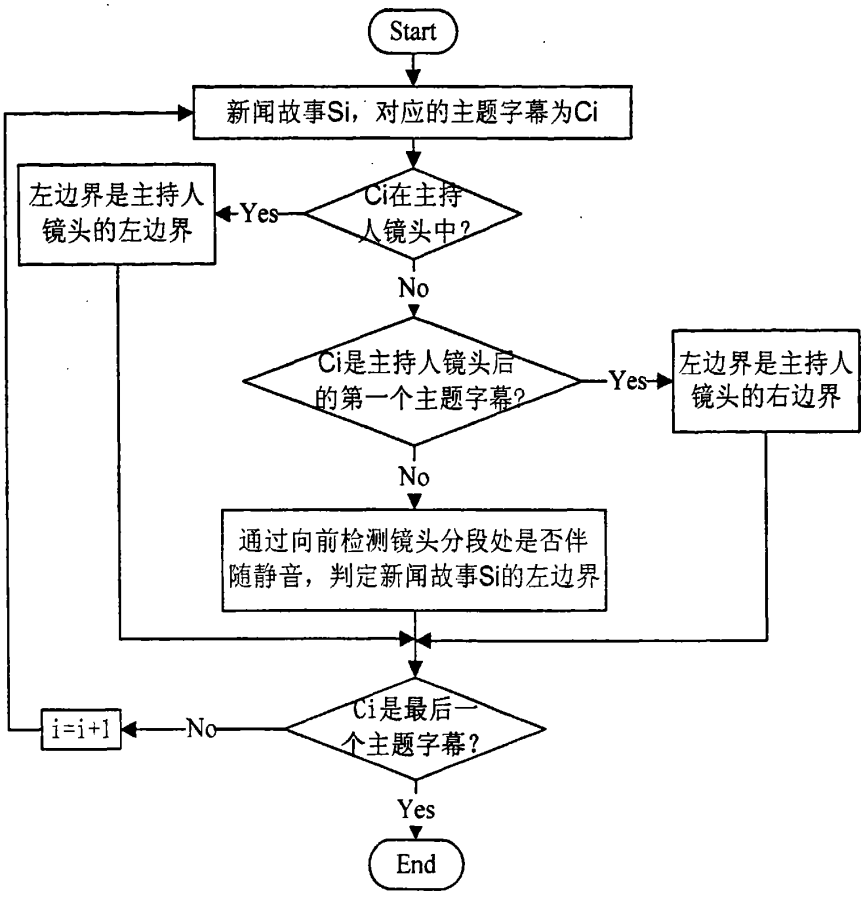


图 5.1.5 新闻故事流程图

5.1.4 实验结果

图 5.1.6 显示了新闻分段界面，在对新闻视频进行镜头分段、关键帧提取和字幕帧提取的基础上，可以进行新闻故事分段。选取的素材是长达三十分钟的新闻联播节目“20051207”，当按下“新闻故事分段”按钮后，弹出新闻故事分段的小对话框，显示该新闻视频的故事总数是 23，界面的下半部分，一行表示一个新闻故事，显示的三张图片，分别表示某新闻故事的起始帧、结束帧和字幕帧。

图 5.1.6 新闻故事分段界面

用本节的新闻故事分段方法，对 4.2 节中的新闻联播素材进行故事分段。新闻故事分段是否正确，将新闻故事的左边界作为评判依据。实验结果如表 5.1.2 所示，四天的新闻联播节目中共包含 101 个新闻故事片段，正确分割的故事片段数目是 88，正确率为 87.13%。其中新闻简讯与其他新闻故事相比，持续时间短，形式比较简单，而且主题字幕起始帧经常处于新闻故事的前两个镜头中，一般不会有其他静音段的干扰，因此简讯的故事分段正确率更高。分段错误有两个原因，其一是在非新闻故事分段处同时出现了镜头变换和静音区间，其二是主题字幕帧的漏检。可知，主题字幕帧检测的结果，直接关系到新闻故事分段的结果，因此主题字幕帧检测的正确率，有待进一步提高。

表 5.1.2 新闻故事分段的实验结果

	新闻联播1	新闻联播2	新闻联播3	新闻联播4
新闻故事的个数	26	24	21	30
正确分割的新闻故事个数	24	20	17	27
正确率	87.13%			

5.2 基于主题字幕提取的新闻视频检索

5.2.1 新闻视频检索系统简介

与一般的视频检索相比，新闻视频的检索有其特殊性。首先，新闻是面向公众服务的，这就决定了新闻检索更强调其宏观的语义，而非低级的视觉特征。此外，系统所提供的检索和服务，应当在很大程度上符合人们在电视等传统媒体上获取新闻的习惯[32]。新闻视频本身具有一种层次化的结构，其中视频文件和视频帧是数字化视频自然具有的结构层次，而新闻故事、镜头需要使用计算机自动或者半自动提取[33]。因此，从视频流的最小单位图像帧出发，逐级归纳、分析和提取视频的物理特征和语义内容，并利用一系列层次化的结构组织方式，可以实现一个初步的内容检索功能，帮助用户快速把握新闻片段的大致内容。图 5.2.1 所示的新闻视频检索系统结构，将视频结构与视频语义综合起来考虑，从而实现了新闻视频的多特征检索。

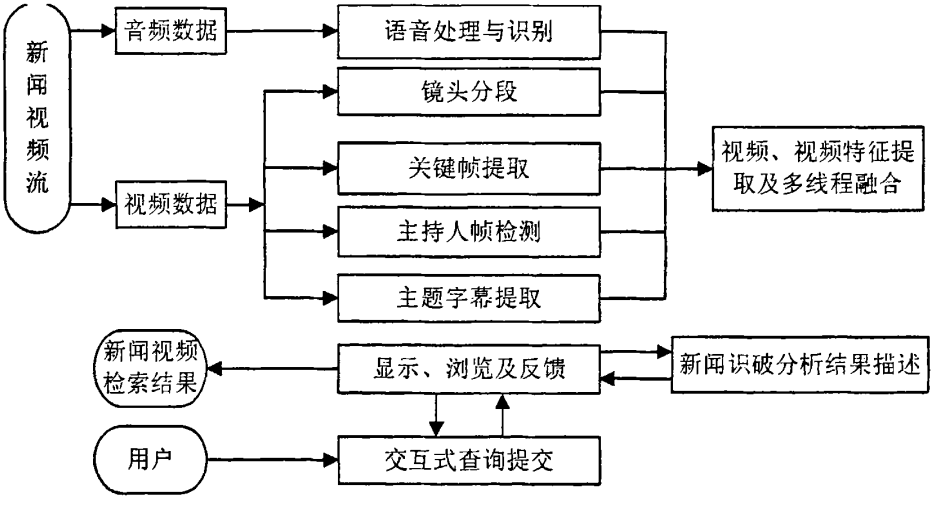


图 5.2.1 新闻视频检索系统结构

由图 5.2.1 可以看到，通过新闻视频的关键帧、主题字幕和主持人帧，都可以对新闻视频进行检索，因为这三种方式具有代表性或总结性。关键帧，是每个镜头中最具有典型性的视频图像，它是用典型的图像信息代表视频镜头；主题字幕是显示在新闻图像中用来概括每个新闻故事的文字信息；主持人帧是对其后相关新闻故事的语音概述。采用这三种信息进行检索，可以形成图像、语音、文本全方位的检索。拥有这几种语义信息，用户可以通过多种途径和方式检索查询感兴趣的新闻内容。图 5.2.2 是对新闻检索方式进行简单的分类。

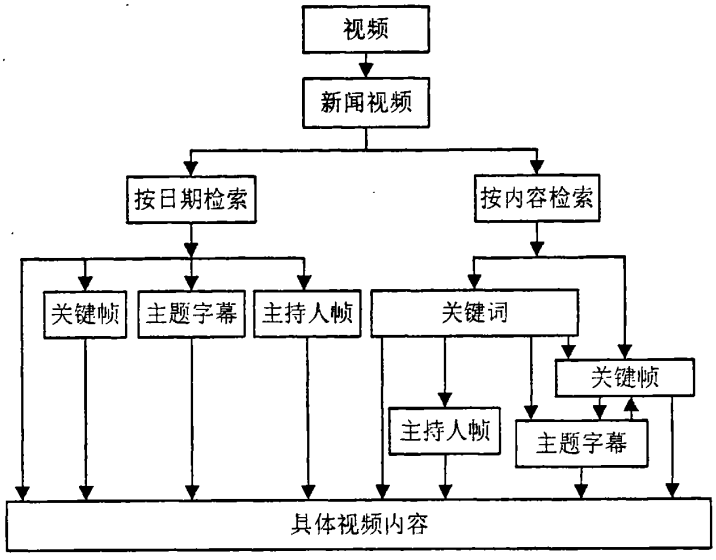


图 5.2.2 新闻视频检索方式示意图

如图 5.2.2，可以看到检索方式大致分成两类：按日期检索和按内容检索。用户查找某日的新闻时，可以按日期检索，还可以选择关键帧、主题字幕和主持人帧三种方式进一步的查询，也可以直接观看该日的具体新闻内容。当用户需要查找含有某种信息的新闻内容时，可以按内容检索，用户可以选择文本查询、或图像查询所对应的关键词或关键帧来查找所需的概括性和代表性的信息，中间可以通过检索带有音频的主持人帧或主题字幕等方式，最终选择到想要获取的信息内容和方式。

比如当需要查找含有“温家宝总理”的信息，可以按关键字查找，查找结果以四种方式呈现：第一种，直接点击搜索到的新闻视频，可逐条观看其具体内容。第二种是相关新闻的主题字幕，可以从文本上看到对应新闻的大概内容，然后可按需要查看具体视频。第三种是主持人帧，可以从主持人帧对应的音频信息了解新闻内容，进而按需查询。最后一种是关键帧，若干关键帧可以用图像的方式呈现某条新闻的大概内容，点击某关键帧即可以进一步展开为所代表的新闻视频。当然，在关键帧、主持人帧和主题字幕之间也可以相互对应查找，从而在各个方面为用户提供检索的便利。

5.2.2 数据库设计

本新闻视频检索系统的数据库名为 NewsVideo，其中包括四个数据表，用来记录视频文件信息和视频处理版块的处理结果，分别是视频文件信息表(VideoTable)、镜头表(ShotTable)、关键帧表(KeyFrameTable)以及新闻故事分段表(NewsStoryTable)。

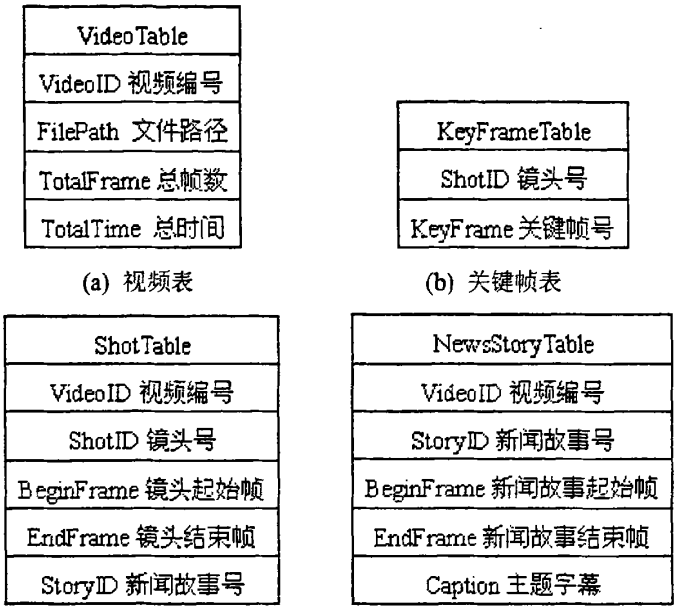
视频表(VideoTable)是视频文件表，存储的是新闻视频文件的相关信息，表项有视频文件名、路径名、总帧数和总时间，如图 5.2.3(a)。

镜头表(ShotTable)是视频文件的分段表，存储的是视频文件经镜头分段产生的物理镜头，这些结构化的数据为后续工作(如主题字幕提取、关键帧提取等)提供了依据。表

项有视频文件名、镜头号、镜头起始帧、镜头结束帧、镜头起始时间、镜头结束时间，以及镜头所对应的新闻故事号。记录镜头的起始和结束时间，可以为后面新闻故事分段的音频检测提供依据，如图 5.2.3(c)。

关键帧表(KeyFrameTable)存储的是镜头表中各个镜头提取的关键帧帧号。表项有镜头号和与镜头号一一对应的关键帧帧号，如图 5.2.3(b)。

新闻故事表(NewsStoryTable)记录的是新闻故事号，以及新闻故事对应的起始帧和结束帧，以及对应的主题字幕，如图 5.2.3(d)。



(a) 视频表 (b) 关键帧表

(c) 镜头表 (d) 新闻故事表

图 5.2.3 数据库表结构

5.2.3 新闻视频检索

视频字幕中包含有丰富的语义信息，可用于对相应视频序列所表达的事件、情节等进行标注，对视频内容的分析与理解有重要的作用。对于新闻视频而言，主题字幕所表达的是新闻故事的中心意思，所有这段新闻的台词都是围绕这段主题字幕进行发挥，识别了这段字幕，就相当于基本把握了这段新闻故事的主要内容，就可以为用户进行视频检索提供有效的索引。

这里，用主题字幕来标注每个故事片段，为新闻视频建立索引，从而可以在检索过程中可以直接利用主题字幕进行检索。提取到的字幕存放在新闻故事表(NewsStoryTable)一个专门的字段中，查询时输入关键词，只要在该字段中检索到相关记录，就可以查看该字幕所属的新闻视频、新闻故事、新闻镜头。

由上节可知，对新闻视频的检索可以分为按日期的检索和按内容的检索，其中按内

容检索又分为基于关键帧、主题字幕等检索方式。本文的检索系统可以实现按日期的检索和按主题字幕的检索。如图 3.2.5(a)，按下“新闻视频检索”，将进入新闻视频检索界面。

(1) 按日期的检索

选择查询方式为日期，输入想要观看的新闻联播的具体日期。如图 5.2.4，输入“20051207”，在新闻视频栏中可以查询到相关视频 20051207。这里，新闻视频编号按照具体日期取名，例如 2005 年 12 月 7 号的新闻联播，取名为“20051207”。新闻故事栏中可以看到具体的新闻故事号，末三位表示新闻的故事数，例如“51207011”，表示第 11 个新闻故事。新闻镜头栏显示该新闻故事所包含的镜头，末三位表示镜头数，如“512070357”，表示该新闻联播的第 357 个镜头。关键帧显示区域，显示新闻故事“51207011”的关键帧。

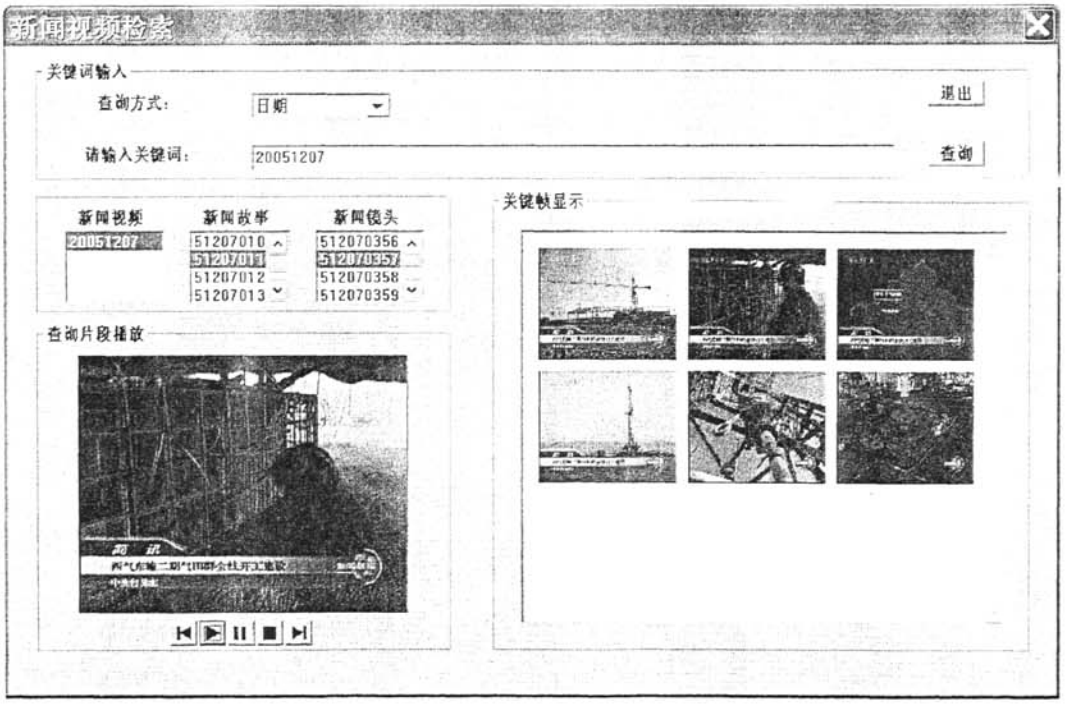


图5.2.4 新闻检索界面(按日期的检索)

(2) 按主题字幕的检索

如图 5.2.5，选择查询方式为主题字幕，通过输入关键词，可以实现相关新闻故事的查询。例如输入“三峡”，可以查到三峡的相关新闻故事：“三峡库区三期清库工程开始启动”。双击新闻故事、新闻镜头，界面左下角的查询片段播放版块中，将显示新闻故事、与新闻故事相对应镜头，可以对其进行播放、暂停等操作。双击新闻故事的同时，界面右下角的关键帧显示模块中，显示该新闻故事的关键帧。

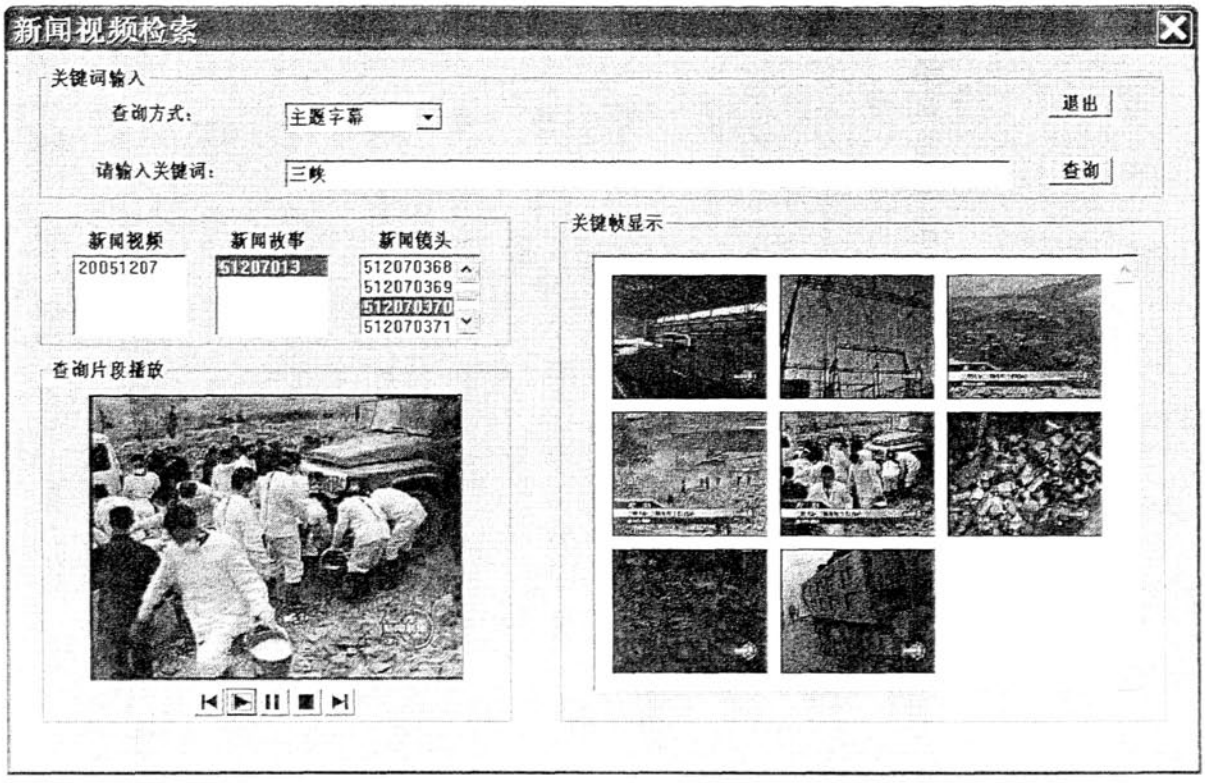


图5.2.5 新闻检索界面(按主题字幕的检索)

5.3 本章小结

本章内容分为两部分。

第一部分设计了新闻故事分段算法，在主题字幕提取的基础上，通过静音检测、主持人镜头检测，把连续的新闻视频分割成一个个的新闻故事，并且用主题字幕来标注每个故事片段，为新闻视频建立索引。

第二部分首先介绍了新闻视频检索的概念，接着介绍了本新闻系统的数据库，最后实现了基于关键字的新闻视频检索，用户可以通过输入关键词，实现对相关日期、相关新闻故事的检索，达到了预期的检索效果。

6 结论与展望

6.1 本文的主要工作

本文主要研究了基于主题字幕提取的新闻视频检索，完成了新闻视频结构化、主题字幕的提取、新闻故事的分段和基于关键字的新闻视频检索。

本文的主要工作体现在以下几个方面：

(1) 针对新闻视频固有特征，改进了双重比较法，将自适应双阈值运用于双重比较法，来检测镜头突变和渐变；根据主题字幕的重要性，提出了基于主题字幕的关键帧算法。

(2) 针对新闻主题字幕的特点，设计了基于 3/10 时空切片的主题字幕帧提取算法，完成了主题字幕帧的提取；采用了基于小波变换和支持向量机的字幕区域提取算法，实现了视频中主题字幕的准确定位；对字幕进行插值放大、二值化、字符分割及 OCR 文字识别。

(3) 针对新闻视频的结构特点，设计了新闻故事分段算法，在主题字幕提取的基础上，根据静音检测、主持人镜头检测，把连续的新闻视频分割成一个个的新闻故事；根据新闻分段结果，以及主题字幕的标注，实现基于日期和主题字幕的新闻视频检索。

6.2 进一步工作的方向

本文虽然完成了新闻视频检索，但仍然属于对基于内容检索的初步探索，同时受作者水平的限制、时间的限制，本文很多尚待完善之处，仍有许多问题值得研究并亟待解决。作者认为进一步的研究可以在以下几个方面展开：

(1) 视频数据的一个很大特点在于数据量巨大，因此对算法的速度以及精确度要求很高。这就需要对许多算法要做大量的优化，以满足实时性的需要，以期达到更加理想的效果。

(2) 本文只完成了对主题字幕的提取；非主题字幕往往也包含着重要的语义信息，例如对一些重要领导人的采访；同时很多新闻视频会出现左右型滚动字幕，其中也包含了大量的新闻事件，对其检测有很重要的意义。因此，下一部的目标是完成各种类型字幕的提取。

(3) 目前系统提供的检索方式是基于关键字的检索，比较单一，下一步考虑提供灵活多样的检索方式，将会极大提高用户的使用积极性。

(4) 本文的研究只针对新闻视频，必然具有一定的局限性。未来的研究将是将各类视频的的检索融合在一个系统中，建立一种具有普适性的视频检索系统，这将会对获取

信息的方式产生深远的影响。

致 谢

首先衷心感谢我尊敬的导师王建宇教授。王建宇教授深厚的学术造诣、严谨的治学态度、一丝不苟的工作作风、宽厚待人的高尚品格、诲人不倦的师长风范，敏锐的学术洞察力都给我留下了深刻的印象，对学生树立正确的治学观念产生了直接而深刻的影响。王老师的言传身教、悉心指导把我带入了科学的殿堂，论文从开题到完成，从理论框架到具体表述，大小环节无不渗透着王老师的心血。可以说，在这近两年来我取得的哪怕一点点的成绩里，都无不浸透着王老师的心血。

感谢教研室的兄弟姐妹们，大家组成了一个融洽的、朝气蓬勃的、催人奋进的学习和生活集体。大家共同创造的团结协作、积极向上的科研环境，让我深受其益。他们是边巧玲硕士、李丹硕士、高珊珊硕士、李奎硕士、朱良峰硕士、王惠文硕士、周秀芬硕士、马飞硕士、曲泽超硕士等。特别感谢范柏超博士和董敏硕士，不会忘记我们共同讨论相互解决问题的美好时光。

感谢我的父母，二十年来一直无微不至地关心我的成长，不时地给予我前进的动力。还要感谢我的男友陈永进，在论文的整个过程中，给予了极大的鼓励与支持，才有了论文的终稿。感谢我的室友两年的时间在生活上关心和照顾，鼓励我不断前进。

难忘在南京理工大学的六年学习生活，学校中博学多才、平易近人的老师，碧草如茵、花香流溢的校园，都给我留下了深刻的印象。“团结、献身、求是、创新”将永远激励我踏踏实实学习，老老实实做人。

向所有关心和帮助过我的人表示最诚挚的感谢！

参考文献

[1] M.Flickner，H.Sawhney，W.Niblack，J.Ashley，Q.Huang，B.Dom，M.Gorkani，Query by Image and Video Content:the QBIC System. IEEE Computer，Septemper 1995.

[2] Marco La，Edoardo Ardizzone，JACOB: Just a Content-Based Query System for Video Database.Proc.ICASSP-96,May 7.10，Atlanta.

[3] 庄越挺等.用语义联想支撑基于内容的视频检索.计算机研究与发展.Vol 36 NO.5

[4] 刘明宝等.复杂背景下的人脸检测与跟踪系统.计算机研究与发展.1997.Vol34.

[5] 庄越挺，潘云鹤，吴飞.网上多媒体信息分析与检索.第 1 版.北京:清华大学出版社，2002.

[6] 章毓晋. 基于内容的视觉信息检索.第 1 版.北京:科学出版社，2003.

[7] 飞思科技产品研发中心 MTALAB 6.5 辅助图像处理 电子工业出版社 2003.6.

[8] 王保雄，余松煜.视频检索中的镜头边界检测.红外与激光工程.2000：Vol 29 NO.5

[9] Zhang h j, Wu jianhua, Zhong di. An integrated system for content-based video retrieval and browsing pattern recognition.1997,30(4):643-657.

[10] A.nagasaka and Y.tanaka. Atutomatic video indexing and full-video search for object appearances, second working conference on visual database systems, IFIP WG2.6, october 1991. 119-133.

[11] Arman F, Hsu A, Chiu M Y. Image processing on compressed video data for large video databases. ACM multimedia, 1993, 267-272.

[12] Shih-Fu Chang and William Chen and Horace J.Meng and Hari Sundaram and Di Zhong, VideoQ: An Automated Content Based Video Search System Using Visual Cues, {ACM }Multimedia, 1997: 313-324.

[13] Atel Nilesh V, Sethi Ishawr K. Video shot detection and characterization for video databases. Pattern recognition, 1997, 30(4):583-592.

[14] Stricker M，Orengo M. Similarity of color images. SPIE Storage and Retrieval for Image and Video Databases III，Feb.1995，2185：381-392.

[15] Kim E Y, Kim K I, Jung K and Kim H j, A video indexing system using character recognition. International Conference on Consumer Electronics, 2000, pp. 358-359.

[16] Gargi U, Crandall D, Antani S, Gandhi T, Keener R and Kasturi, R.A system For automatic text detection in video. Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999, pp. 29-32.

[17] Tang X A, Luo B, XGao X B, Pissaloux using temporal feature vectors. Proceedings

Multimedia and Expo, Vol.1, 2002, pp. 85-88.

[18] Luo B, Tang X 0, Liu J Z and Zhang H j. Video caption detection and extraction using temporal information. Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, Vol.3, 2003, pp. 1723-1728.

[19] 蔡波，周洞汝，胡宏斌.数字视频中字幕检测及提取的研究和实现.计算机辅助设计与图形学学报，Vol.15，No.7，2003，pp.898-903.

[20] Zhong Y, Zhang H J and Jain A K. Automatic Caption Localization in Compressed Video. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, 2000, pp. 385-392.

[21] 何家颖，黎绍发.一种复杂背景图像文字分割算法.模式识别与人工智能，Vol 18, No 2, 2005, pp.148-153.

[22] Gllavat J, Ewerth R and Freislebe B A robust algorithm for text detection in images. Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis, Vol. 2, 2003, pp. 611-616.

[23] Mao W G, Chung F L, Lam K K M and Sun W C. Hybrid Chinese/English text detection in images and video frames. Proceedings of 16th International Conference on Pattern Recognition, Vol. 3, 2002, pp. 1015-1018.

[24] Jain A K and Zhong Y. Page segmentation using texture analysis. Pattern Recognition, Vol.29, 1996, pp. 743-770.

[25] Kim K Z, detection in digital Jung K, Purk S II and Kim H J. Support vector machine-based text video. Pattern Recognition, Vol. 34, 2001, pp.27-529.

[26] Li H P, Doermann D and Kia 0. Automat is text detection and Tracking In digital video. IEEE Transactions on Image Processing, Vol. 9, 2000, pp.147-156.

[27] Zhang D Q. Tseng B L and Chang S F. Accurate overlay text extraction for digital video analysis. Proceedings of International Conference on Information Technology: Research and Education, 2003 pp. 233-237.

[28] Li H, Kia 0 and Doermann D. Text enhancement in digital video. Proceedings of SITE Document Recognition IV, 1999, pp. l-8.

[29] Kwak S, choi y and Chung K. Video caption image enhancement for an efficient character recognition. Proceedings of 15th International Conference on Pattern Recognition, Vol. 2, 2000, pp. 606-609.

[30] 沈淑娟. 基于时空域信息的视频字幕提取算法研究. 西安电子科技大学硕士论文 2004.1.

[31] Zhang H J, Kankanhalli A, Smoilar S W. Automatic Partition of full-motion video,

Multimedia system, 1993:10-28.

[32] 马宇飞，白雪生等.新闻视频中口播帧检测方法的研究.软件学报，12(3):377-382, 2001.

[33] 熊华，老松扬，吴玲琦等.News VideoCAR：一个基于内容的视频节目浏览检索系统.计算机工程，26(11)：73-75，2000.

[34] 高新波. 模糊聚类分析及其应用.西安电子科技大学出版社. 2004，49-62.

[35] 刘华咏, 周洞汝. 一个基于内容的新闻视频浏览和查询系统：NewsBR. 小型微型计算机系统，2004, 25(4): 535-539.

[36] 王润生.图像理解.长沙：国防科技大学出版社,1995

[37] 马宇飞等.新闻视频中的口播帧检测方法的研究.软件学报.2001（3）377-381

[38] 田捷等.实用图像分析与处理技术.北京：电子工业出版社,1995

[39] 徐峻等.新闻视频中主持人镜头识别方法的研究.计算机工程.2002：Vol 28 NO.3

[40] 赵亚琴.基于内容的视频片段检索技术研究. 南京理工大学博士论文. 2007.3，pp. 67-73.

[41] 章东平. 视频文本的提取. 浙江大学博士论文. 2006,5.

[42] 庄挺越，刘骏伟，吴飞等.基于支持向量机的视频字幕自动定位与提取[J].计算机辅助设计与图形学学报，2002，14 (8):750-753,771

[43] 于俊清，周洞汝.基于文字和图像信息提取新闻视频关键帧.计算机工程与应用，2002, 38 (9):83-850

[44] 朱映映，周洞汝. 一种基于视频聚类的关键帧提取方法. 计算机工程，2004，30(4)：12-13,121.

[45] 张继东，陈都. 基于内容的视频检索技术. 电视技术，2002(8)：17-19，23.

[46] 沈帮乐. 计算机图像处理.北京：解放军出版社,1995

[47] 朱曦，林行刚. 视频镜头时域分割方法的研究. 计算机学报，2004，27(8)：1027-103.

[48] 王东辉,朱森良,吴春明.一种用于自动视频分段的 WIPE 转换检测和模式识别方法. 计算机研究与发展,2002,39(2),247－253.

[49] Ngo C W, Pong T C, Chin R T. Detection of gradual transition through temporal slice analysis. Computer Vision and Pattern Recognition,1999. IEEE Computer Society Conference on, 23-25.June 1999, 1.41.

[50] I.Guyon. Applications of neural networks to character recognition. International journal of pattern rocgnition and artificial intelligence, 1991(5):353-382.

[51] J.Hernando. Voice signal processing and representation techniques for speech recognition in noisy environments. Signal processing, 1994(36):393-341.

[52] Shahraray B，Gibbon D C. Automatic Generation of Pictorial Transcripts of Video

retrieval and browsing. Pattern Recognition, 1997(30): 643-648.

[53] Smith J R, Chang S F. single color extraction and image query. In: Proc IEEE Int. Conf. on Image Proc, 1995:80-88.

[54] Wolf W. Key Frame selection by motion analysis. ICASSP 96, 1228-1231.

[55] H.P.Li, D.Doemann, and O.Kia, Text extraction, enhancement and OCR in digital video, in Proc, 3rd IAPR Workshop, Nagoya, Japan, 1998. pp. 363-377.

[56] Hua X S, Yin P and Zhang H J. Efficient video text recognition using multiple frame integration. Proceedings of 2002 International Conference on Image Processing, Vol. 2, 2002, pp. 397-400.

[57] J. T. Foote. An overview of audio information retrieval. Multimedia Systems. 1999.7(1): 2-11

[58] Ngo C W. Analysis of Spatio-temporal Slices for Video Content Representation: PhD Thesis. Hong Kong University of Science and Technology, 2000.

[59] Haritaoglu I. Scene text extraction and translation for handheld devices. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol.2, 2001, pp. 408-413.

[60] Tang X A, Luo B, XGao X B, Pissaloux using temporal feature vectors. Proceedings Multimedia and Expo, Vol.1, 2002, pp. 85-88.

[61] T.M.Cover. Geometrical and statistical properties of systems and linear inequalities with applications in patter recognition. IEEE Trans. On Electronic computers, 1965(3):326-334.