基于字幕文本提取的讲座视频快速浏览技术探讨*

夏玉华[1]　巩海梅[2]
(1.山东大学图书馆，济南 250100；2.山东建筑大学图书馆，济南 250101)

摘要：图书馆中的视频资源越来越丰富，为使读者能够快速地从海量的视频资料中找出想要的视频，并且准确地从视频中定位到想要的段落，文章根据讲座视频字幕位置相对固定的特点，采用帧差法提取字幕文本，通过对字幕文本的识别，将检索词与视频内容联系起来，进而实现读者快速浏览并定位视频段落的目的。

关键词：讲座视频；快速浏览；字幕文本提取
中图分类号：G250.76
DOI：10.3772/j.issn.1673—2286.2014.04.006

近年来，随着计算机技术、多媒体技术和互联网技术的飞速发展，学术讲座越来越普及，讲座视频也应运而生。讲座视频不仅是读者开拓视野、了解学科前沿、提升综合素质的资源，而且给读者提供了学习时间和空间上的自由度，可以随时随地重复观看。在山东大学图书馆多媒体资源中，爱迪科森“网上报告厅”和超星学术视频都收录了大量的讲座视频。这些视频中每个视频包含的内容丰富，时长比较长，对于只关注其中的几个兴趣点的读者来说，如何快速准确地定位所关注的视频内容就成为了高效利用视频资源的关键。虽然现在优酷、搜狐等一些大型视频网站提供了关键帧呈现视频内容的功能，但对于场景相对简单的讲座视频来说，都不能很好地解决问题。目前，山东大学图书馆对讲座视频的介绍仅限于专题名、主讲人、主讲人单位等，如此简单的介绍远不能满足读者对讲座视频内容的检索需求。因此，读者如何快速检索到所需内容的视频实现快速浏览成为亟待解决的问题。

字幕文本是对讲座视频内容准确的描述，通过对字幕文本的提取和解析，可以实现对视频内容贴切的关键词描述。目前，视频的字幕提取算法是国内外多媒体技术领域的研究热点之一。国内主要有基于多示例学习的视频字幕提取算法[1]；基于边缘强度的视频图像字幕提取算法[2]；基于行颜色梯度分析的视频字幕区提取算法[3]；基于边缘和灰度的视频文字提取方法[4]；基于笔画特征的多方法综合视频文本提取算法[5]等。国外主要有基于多层次特征优先级的文本提取算法[6]；基于分层区域的图像模型字幕提取算法[7]；利用图论聚类的视频字幕提取算法[8]等。

1　字幕文本提取算法的需求分析

目前，基于文本检索的技术已经非常成熟。读者在检索图书、期刊、学位论文、会议论文、专利等文献时，可通过题名、摘要、关键词、全文等字段检索所需文献。但在检索视频时，由于在视频帧的低级特征，如颜色、纹理、形状等与其语义特征之间建立准确的对应关系非常困难，因而，往往采取视频标注的方式。比如北京大学图书馆对讲座视频的揭示有题名、主要责任者、内容描述、主题关键词、语种等。其中，“内容描述”是编辑人员对视频内容的一个描述，容易引入个人见解，也就是说不同的人对同一视频的认识和理解可能是不一样的，那么给出来的描述就不一样，这会影响对视频

* 本研究得到国家自然科学基金项目“基于感知哈希和流形降维的视频复制检测技术研究”（编号：61001180）资助。

最本质的描述。视频的字幕文本是一类特殊的文本，它是视频内容的文字呈现，是源于视频本身的不带有任何人主观因素的描述。从这个角度来说，该文本信息可以对视频内容进行可信的有效描述。此外，这种描述由于能准确记录视频内容，可以完成低级特征不能表述的语义表达任务，从而有效建立视频低层特征与高级意义之间的桥梁。同时，字幕在视频中位置相对固定，文字比较突出，技术上实现的难度相对小一些。

2 基于帧差的字幕文本提取算法

2.1 讲座视频字幕文本特征分析

视频文本有两种：场景文本和字幕文本[9]。场景文本就是视频中景物上出现的文字，比如讲座视频中的课件、体育视频中的比分牌等都是在场景内出现并由视频摄制设备记录下来的文本。虽然场景文本在一定程度上反映了视频的内容，但字幕文本才是视频内容的准确表述，不但可以作为视频内容的标注，而且可以据此实现基于内容的检索，定位视频段落。通过对比分析大量的讲座视频，发现其字幕文本具有以下特征。

(1)位置相对固定

字幕文本位置通常在视频帧底部1/4处，且在连续的多帧图像中重复显示，一般只有显示和消失两种变化状态。

(2)字符尺寸、间距均匀且相对固定

为满足讲座视频规范化的要求，字符大小一致，间距均匀无粘连，易于识别。

(3)颜色、亮度与其背景对比明显

讲座视频的字幕文本与背景之间一般保持较高的颜色对比度，如文本通常为白色，亮度较高，其背景通常以蓝色、深红色为主，颜色较深，亮度较低。

总之，讲座视频字幕文本的位置、字符尺寸、颜色、亮度及其运动方向都有很好的稳定性。

2.2 基于帧差的讲座视频字幕文本提取算法

通过对讲座视频字幕文本的特征分析，提出以下三种字幕文本提取算法，以实现基于内容的讲座视频快速浏览。

2.2.1 逐帧字幕文本提取算法

逐帧提取字幕文本算法可以做到对讲座视频内容的完全揭示，是最简单实用的算法。但字幕文本是连续显示的，字数多时，显示的帧数可达到30～40帧；字数少时，显示的帧数也在5～10帧。此外，由于话语停顿，场景转换等原因，字幕帧之间还有无字幕帧。可见，该算法虽然简单但重复计算量大。

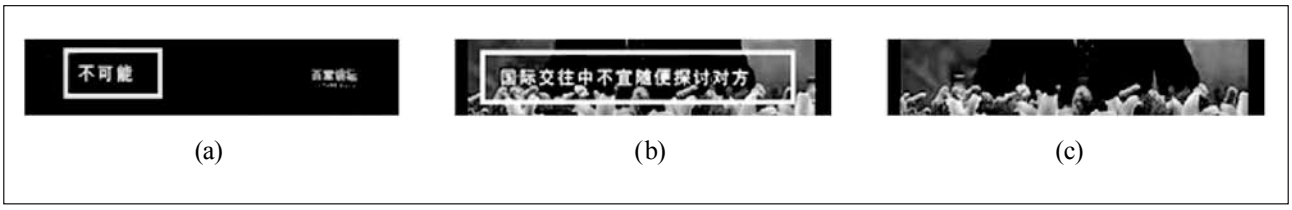
2.2.2 等帧数间隔字幕文本提取算法

由于在讲座视频中，相同的字幕文本是连续多帧重复显示的，因而采取每隔固定数量的视频帧提取一帧进行灰度变换、边缘检测、二值化一系列处理之后提取字幕文本的方法，即等帧数间隔字幕文本提取算法。

图1(a)中的“不可能”，在6帧图像中重复，图4-1(b)中的“国际交往中不易随便探讨对方”在30帧图像中重复，图1(c)无字幕帧在10帧中重复。如图1所示，若间隔帧数为2帧，则图1(a)、(b)、(c)中都有视频帧被重复提取，若间隔帧数为28帧，则会漏掉图1(a)文本帧。可见，帧数间隔大小难以确定。此外，由于镜头切换、话语停顿、场景变换等因素也会影响间隔帧数的选择。

2.2.3 基于帧差的字幕文本提取算法

根据讲座视频字幕文本特征分析和对等帧数间隔



(a) (b) (c)

图1 字幕文本字数不同的有字幕帧和无字幕帧[10]

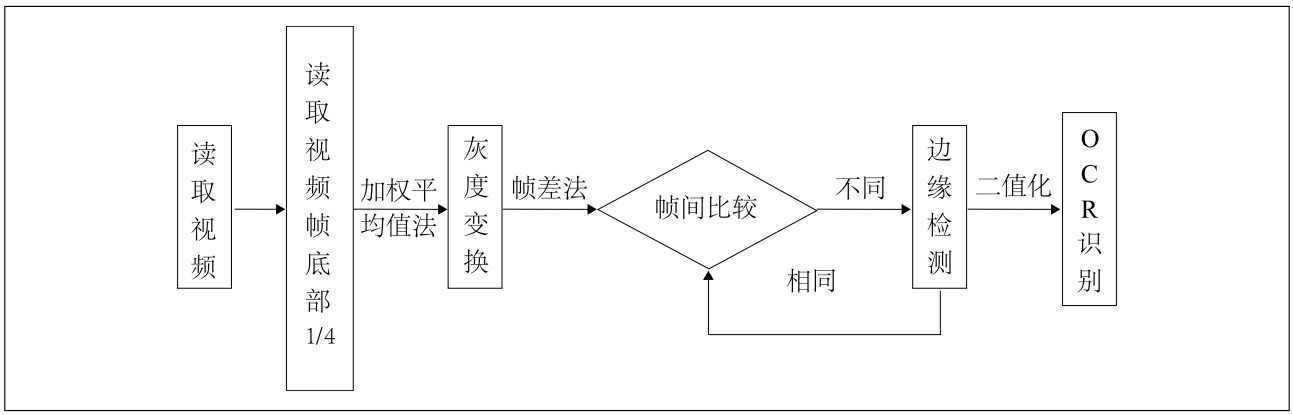


图2 基于帧差的字幕文本提取算法流程图

字幕文本提取算法、逐帧字幕文本提取算法的比较，提出基于帧差的字幕文本提取算法，如图2所示。

(1) 灰度变换

视频帧的灰度变换就是把彩色视频帧转化为黑白颜色图像的过程。读取视频帧底部1/4，按加权平均值法进行灰度变换。

当R=G=B=L时，

$$L(x,y)=0.2989R(x,y)+0.5870G(x,y)+0.1140B(x,y) \quad (1)$$

其中，L(x,y)—像素点(x,y)的灰度值；

R(x,y)—像素点RGB颜色的红色分量；

G(x,y)—像素点RGB颜色的绿色分量；

B(x,y)—像素点RGB颜色的蓝色分量。

(2) 帧差运算

通过对逐帧字幕文本提取算法和等帧数间隔字幕文本提取算法的分析发现，关键是如何过滤掉视频中的重复帧。本文采用帧差法，如图3所示。

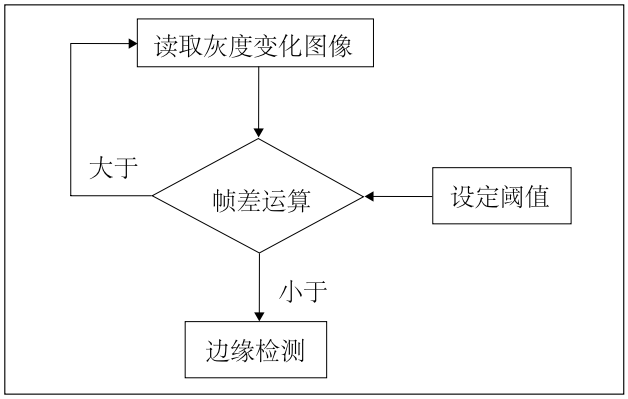


图3 帧差运算

帧差法的基本原理是采用基于像素的时间差分在图像序列相邻两帧通过闭值化来提取图像中的运动区域。讲座视频的字幕文本均在视频帧的底部约1/4范围内，此处环境相对稳定。设定阈值为0.5，那么，如帧差值大于0.5，则把读入视频帧作为参考帧，同时该帧进入边缘检测处理程序。反之，如果帧差值小于0.5，则认为该帧与参考帧相同，删除该帧，如此循环处理。

(3) 边缘检测

图像边缘是图像最基本的特征，如何提取对整个视频场景的识别与理解尤为重要。如图4所示的Sobel算子，(a)、(b)两个卷积核形成了Sobel算子。其中，(a)用于提取水平方向上的边缘，(b)用于提取垂直方向上的边缘。视频帧中的每个像素点都用这两个核做卷积，两卷积核的最大值就是该像素点的输出位。这符合讲座视频字幕文本的空间分布和字符本身的特征，因而可以采用Sobel算子进行图像的边缘检测。

此外，Prewitt算子对灰度渐变和噪声敏感度不高，因此，采用Prewitt算子进行边缘检测也是较佳选择之一。Prewitt算子用卷积模板描述如下：

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

(a) 水平方向的卷积核

-1	0	1
-2	0	2
1	0	1

(b) 垂直方向的卷积核

图4 Sobel算子

$$G(i,j)=|p_x|+|p_y| \tag{2}$$

其中，(i,j)为点$G(i,j)$的像素输出；

$$p_x=\begin{bmatrix}-1 & 0 & 1\\ -1 & 0 & 1\\ -1 & 0 & 1\end{bmatrix}$$ 为水平模板；

$$p_y=\begin{bmatrix}1 & 1 & 1\\ 0 & 0 & 1\\ -1 & -1 & -1\end{bmatrix}$$ 为垂直模板。

视频帧中的每个像素点都用P_x，P_y模板做卷积，其最大值就是该像素点的输出$G(i,j)$。

采用Sobel算子和Prewitt算子的边缘检测结果如图5所示。

(4) 二值化

由于OCR识别软件通常只能识别黑色或者白色背景下的字符，因此还要对边缘图像二值化处理。根据字符和背景在各区域内灰度特性上具有某种均匀性，选定一个阈值来判断哪些像素点是属于字符内部的点，哪些像素点是属于背景的点。二值化图像的质量将直接影响到字幕文本提取的准确度。

本文选用直方图双峰法计算阈值，并对图5中的边缘检测图像进行二值化，结果如图6(b)所示。对于图5，阈值为0.22，若某像素的灰度值小于0.22，则其像素值为0，属于字符内部的点，反之，若某像素的灰度值大于0.22，则其像素值为255，是背景像素点。可见，二值化的关键是阈值的计算。

2.3 仿真实验

利用MATLAB8.0软件实现了基于帧差的字幕文本提取算法。在山东大学图书馆电子资源的爱迪克森“网上报告厅”中任选50个讲座视频中的200段，截取视频段长度为15秒～35秒，总时间约90分钟。图7只展示其中6段视频的实验结果，实验结果见表1。

Sdu001　Sdu002　Sdu003　Sdu004　Sdu005　Sdu006

图7 实验视频

定义评价指标—准确率如式(3)表示。

$$\eta=\frac{m}{n} \tag{3}$$

其中，η—准确率；m—OCR软件正确识别的字幕文本帧数；n—视频段总帧数。

从表1可以看出该算法的准确率都在90%以上，满足了讲座视频基于内容建立索引的需要，为实现基于内容的视频快速浏览提供了技术基础。在验证该算法的

(a)原始灰度图像　(b)Sobel算子结果　(c)Prewitt算子结果

图5 Sobel算子和Prewitt算子边缘检测结果比较

(a)边缘检测图像　(b)二值化图像

图6 二值化结果

表1 基于帧差的字幕文本提取算法实验结果

视频编号	视频长度（帧）	OCR软件正确识别的字幕文本帧数m	视频段字幕文本总帧数n	准确率（%）
Sdu007	890	24	26	92.3
Sdu002	751	26	27	96.3
Sdu003	480	21	23	91.3
Sdu004	461	20	22	90.9
Sdu005	820	22	24	91.7
Sdu006	645	18	19	94.7

实验过程中，讲座视频不同，阈值计算方法不同。一方面因为文本背景复杂，另一方面文本的淡入和淡出也造成了阈值的不稳定。可见，需要在阈值计算时选择适应性更好的方法，同时加入字符检测的方法，而不仅仅是二值化。

3 结语

在对图书馆讲座视频结构特征分析的基础上，充分利用现有的灰度变换、边缘检测、二值化以及阈值计算方法，以帧差的方式提取讲座视频中的字幕文本，通过字幕文本对视频内容进行快速定位。该方法具有实现简单、计算量小、准确率高等优点。对图书馆建立基于内容的讲座视频索引，方便读者根据内容检索所需视频并定位视频段落，实现基于内容的视频快速浏览具有现实意义。

参考文献

[1] 周长建.基于多示例学习的视频字幕提取算法研究[D].哈尔滨:哈尔滨工程大学,2012.

[2] 曹喜信,刘京,杨旭东,等.一种新的视频字幕提取算法(英文)[J].北京大学学报(自然科学),2013(2):197-202.

[3] 李琼.基于行颜色梯度分析的视频字幕区提取算法研究[J].三门峡职业技术学院学报,2013(2):115-118.

[4] 高华.基于边缘和灰度的视频文字提取方法的研究与应用[D].北京:北方工业大学,2011.

[5] 吴智恺.基于笔画特征的多方法综合视频文本提取算法研究[D].上海:上海交通大学,2010.

[6] CHITRAKALA G, MANJULA D. Multi Level Feature Priority algorithm based text extraction from heterogeneous and hybrid textual image [J]. International Journal of Signal and Imaging Systems Engineering, 2009, 2(4): 183-95.

[7] LEON M, VILAPLANA V, GASULL A, et al. Caption text extraction for indexing purposes using a hierarchical region-based image model [C]// Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP 2009), Cairo, Egypt. USA: IEEE, 2009: 1869-72.

[8] CHUN B T, HAN K, LEE J. Caption extraction in videos using graph-theoretic clustering [C]// CALLAOS N, HERNANDEZ-ENCINAS L, YETIM F. 6th World Multiconference on Systemics, Cybernetics and Informatics. Proceedings, Orlando, FL, USA. USA: Int. Inst. Inf. &Syst., 2002: 57-60.

[9] 刘曼曼.基于支持向量机的新闻视频主题式字幕提取[D].天津:天津大学,2007.

[10] 夏玉华.基于高校图书馆学术讲座视频的快速浏览技术研究[D].济南:山东大学,2010.

作者简介

夏玉华，女，1972年生，山东大学图书馆馆员，研究方向：信息与信号处理、学科评价。E-mail：377801915@qq.com。

Quick Browsing Approaches to Lecture Videos Based on Caption Text Extraction Algorithms

XIA YuHua[1] GONG HaiMei[2]

（1. Library, Shandong University, Ji'nan 250100, China; 2. Library, Shandong Jianzhu University, Ji'nan 250101, China）

Abstract: With the enrichment of videos in library, in order to help readers to find the exact video from huge number of videos and locate the required segments in the video is essential. In this paper, we propose a scheme, which utilizes the algorithm of frame difference to extract caption text based on the characteristics of its stationary position and bridge the index words to video content via the extracted caption text. Simulations show that the proposed scheme can help readers to locate the required video segments quickly and effectively.

Keywords: Lecture Videos; Quick Browsing; Caption Text Extraction

（收稿日期：2014-02-06）