图像和视频中的文字获取技术

王 勇[1),2)] 郑 辉[1)] 胡德文[2)]

[1)](西南电子电信技术研究所现代信号处理国家重点实验室,成都 610041)

[2)](国防科技大学机电工程与自动化学院,长沙 410073)

摘 要 许多图像都包含丰富的文字信息,如用作网页设计的以图像形式存在的标语和视频图像中的字幕。这些文字的自动检测、分割、提取和识别,对图像高层语义内容的自动理解、索引和检索非常有价值,因此引起国内外众多学者的研究兴趣。为使人们对该领域有一个系统的了解,并使该领域研究人员有所借鉴,在对目前国内外图像和视频中文字获取技术相关文献综合理解的基础上,综述了该领域的发展现状,同时从文字检测、抽取和文字识别两个方面,重点讨论了其主要的技术方法及应用优缺点,并结合当前面临的问题,指出今后可进一步研究的方向。

关键词 图像 视频 文字 检测 抽取 分割 识别 基于内容 图像检索 视频检索

中图法分类号: TP391.4 **文献标识码**: A **文章编号**: 1006-8961(2004)05-0532-07

The Technology to Acquire Text in Image and Vedio

WANG Yong[1),2)], ZHENG Hui[1)], HU De-wen[2)]

[1)](*National Key Laboratory of Modern Signal Processing, Southwest Institute of Electron & Telecom. Techniques, Chengdu* 610041)

[2)](*College of Mechatronics and Automation, National University of Defence Technology, Changsha* 410073)

Abstract Many images contain abundant text, such as text in banner used for page design on web pages and text in video. If these text occurrences could be detected, segmented, extracted and recognized automatically, they would be a very valuable source of high-level semantics for image indexing and retrieval. So many international researchers pay more and more attention to acquiring text in images and videos. But now national researchers enter into this field. In order to make people know the academic area more systemically and researchers look up references more conveniently, this paper gives an overview of state of the art of text acquirement research in images and videos. Firstly, this paper discusses the current development of the area based on understand and analysis of related recent papers . Then, from the two aspects of text detection and extraction and text recognition, typical techniques and approaches are discussed mostly, as well as their merits and shortcomings, such as techniques and approaches based on edge, texture, color region, machine learning, video multi-frames and OCR. Finally, with the present problems in this area, the paper give some work and open issues that can be researched more in the future.

Keyword computer information manage system, image, video, text detection, text extraction, text segmentation, text recognition, content-based, image indexing, video indexing

1 引 言

随着计算机技术、多媒体技术以及通讯技术的飞速发展,以图像、声音和视频为主的多媒体信息迅速成为信息交流与服务的主流,在 Internet 上,纯文本页面正逐渐被加入多幅图像,以使页面更吸引人,相当数量的文字信息正越来越多地以图像形式出现,同时,本来纯粹收藏文字资料的图书馆也正在不断地把图像、视频和声音剪辑加入收藏,数字图像和视频图书馆正在兴起,它们都需要一个自动的方法去有效地索引和检索多媒体内容。由于传统的数据

基金项目:国家专项工程项目(613 项目);国家杰出青年科学基金项目(60225015);高等学校优秀青年教师教学科研奖励计划资助项目

收稿日期:2003-07-29;**改回日期**:2003-12-19

库检索中采用的基于关键词的检索方式已逐渐显得力不从心,不能满足人们的需要,所有现存的搜索引擎都不能完全的索引图像丰富的网页内容,于是基于内容的检索(Content-Based Retrieval,CBR)技术应运而生[1]。图像视频检索技术的研究工作,初期主要集中在分析图像及视频帧的低层次特征上,如色彩、纹理结构等,而近期的工作则集中在更加接近直观内容的分析上,如图中主体的构成关系、主体运动特征、视频节目的伴音分析及字幕识别等,这反映了更加接近图像和视频高层语义信息的趋势[2,3]。

图像中的文字,尤其是网页和视频图像中的文字,是图像高层语义内容的一个重要来源,而在 Internet 上,越来越多的网页文字则以图片的形式出现[4,5];在视频中,字幕信息(如新闻标题、节目内容、播出时间、旁白、节目后的致谢、工作人员名单等等)均包含了丰富的高层语义信息[6,7]。如果这些文字能自动地被检测、分割、识别出来,则对图像高层语义的自动理解、索引和检索是非常有价值的[8,9],例如,在 Carnegie Mellon 大学的 Informedia 项目,其视频中的文字就是一个用来支持海量数字视频库全内容搜索的重要信息源[10]。

另一方面,现有的技术也为提取视频及图像中的文字提供了较成熟的技术基础。在文本检索方面,人们已经作了大量的工作,且通过关键词查找相应内容的技术已经非常成熟。例如,现在通过 Internet 上的文本搜索引擎,如 Google 等,人们已经可以比较方便地通过提交关键词的方法来检索到自己感兴趣的文本内容。在字符识别方面,20 世纪 70 年代兴起的光学字符识别(OCR)技术现已日渐完善,在对文字材料进行扫描并将其转换为计算机能够统一识别和存储的内码方面已取得了较大成功[11]。可以说,视频和图像中的文本信息提取的后期处理技术已经相当成熟,人们研究的重点便放在了如何迅速有效地检测及提取这些文字信息上。

由于图像和视频帧没有实质的区别,所不同的仅是制作标准和内容的侧重以及获取的途径,因此其分析与检索方法相同。视频序列的处理是在单帧图像处理的基础上,增加了时间与运动的概念,所以,本文将图像检索技术作为视频检索技术的基础讨论。

2　相关工作

在 20 世纪 70 年代,随着光学字符识别(OCR)技术的兴起,许多学者就开始进行文档图像中文字提取的研究,到了 90 年代,随着多媒体技术的发展以及对基于内容的多媒体检索的需求,图像和视频中的文字获取又逐渐成为研究热点之一。通常从图像中提取文字都需要首先定位包含文字的图像区域,但文字在字体、大小、对齐方式和排列上变化多端,文字背景复杂,图像分辨率低,而且许多应用场合还要求算法具有一定处理速度,这些都使得从图像中有效地提取文字变得非常困难。很多学者在这方面已作出了有益的探索和尝试。

Zhong 等提出了一个解决方案用来定位复杂图像中的文字[12]。该方案由两种方法融合而成,它们分别基于寻找特定尺寸的单色连接区域和基于文字本身特殊的空间差异,但其主要是针对彩色 CD 封面的扫描图像中的文字定位,还不能直接应用到视频帧图像中。通常,扫描图像信噪比(SNR)是相当高的,而视频图像的信噪比却相当低,这也是视频图像文字分割面临的最大挑战之一。Jain 等介绍了一种适合报纸、网页和通常的图像、视频的文字定位方法[13],该方法对广告图像、网页标题图像、扫描杂志页面以及视频帧都能获得较好的实验结果。作者也指出了小文字字体的识别问题和实例,然而,这些实例更经常出现在视频图像中,也没有定量地给出该实例条件下的性能指标。Smith 等提出了一个在视频帧中检测文字的方法[14],他们提取的文字特征为一个有密集边缘的水平长方框,并利用这个特征去辨识单帧中的文字,但这个方法有尺寸上的局限性,即只有特定字体尺寸范围内的文字才能被检测,并且该方法没有利用相同的文字会出现在连续的多帧中这个特征去进一步增强文字检测性能,也没有提到为 OCR 做准备的文字分割。Wu 等提出了一个分 4 步从图像中检测和抽取文字的系统[15],该系统进行文字检测时,首先,文字被看作一种特殊的纹理,并在 3 个不同尺度上使用 3 个二阶高斯导数来找到可能的文字区域;然后,从水平排列的文字区域中抽取出明显的垂直笔划,再基于经验规则将笔划组成紧凑的长方形框,并且在原始分辨率条件下合并这些长方形框;接着清除这些文字框区域的文字背景,并且二值化图像;最后,对得到的文字框区域重复前两步进行提炼,再对每个文字框图像得到的二值图像通过标准的 OCR 软件进行识别。Wu 等对 35 幅图像进行了实验,结果表明,识别率达到 84%。然而,这个方法也是针对扫描图像的,因为扫描图像比

视频帧有更好的信噪比，所以该方法用于视频帧的效果并不理想。

近来，Sato 等开发了一个针对静态的低分辨率新闻标题的文字分割识别系统[16]。他们首先使用在文献[14]中提出的方法去检测标题文字，然后将被检测到的标题文字块以 4 倍放大，并且利用视频帧的基于时间的最小像素值来对标题文字块进行整合。这种多帧整合方法通常是假设标题比其背景像素更亮。这些约束条件对新闻节目是可以接受的，但是对于通常视频中的字幕就显得太苛刻。Sato 等还介绍了一种新的字符抽取过滤器和一个文字分割识别的整合方法，然而，它只适用于新闻节目和 MPEG-1 视频格式[17]。该方法通过查牛津词典和标题词典来改善单词识别性能，在文字定位检测正确率为 89.6%的条件下，字符识别率达到 83.5%，单词识别率达到 70.1%；而在文字定位检测正确率为 100%的条件下，字符识别率达到 74.8%，单词识别率达到 62.8%。

Lienhart 等先后开发出两个视频中的文字检测、分割和识别系统[18,19]。该两个系统都是利用文字的单色性、相对于背景的高对比度和视频字幕的简单纹理来进行图像分割。为了排除非文字区域，文献[18]中的系统采用基于颜色的分割融合算法，并且只对单帧进行分割和识别，而没有考虑连续的多帧。该系统还采用一个迭代的文字识别算法来识别文字成为 ASCII 码。在文献[19]中，由于考虑了字幕纹理各向异性的扩散，而且文字在其存在的连续多帧内被追踪和整合利用，因而使得该系统在商用的 OCR 识别引擎条件下，识别率达到了 41%～76%。Li Hui-ping 等还使用前向反馈的神经网络来定位视频中的文字[20,21]，即将单帧的高频小波系数作为网络的输入来训练神经网络监视视频中是否有文字字幕出现，如果没有文字出现，那么每次就继续处理特定数量的帧；如果有文字出现，就使用块匹配去追踪简单背景下的文字，但该方法只适用于视频，并且其只能在块水平检测文字，如果文字行不是均匀运动的，则该方法将导致严重问题。

近几年，国内学者也开始关注并积极投身到这个领域中来，但都仅在进行基于内容的多媒体检索的研究，只附带地介绍了图像和视频中的文字获取，并没有进行系统深入的研究，也没有开发出相应可行的实用系统。

3 主要技术方法

通过相关文献的对比阅读可以看出，通常获取图像和视频中文字信息的技术路线是：(1)文字区域检测，即利用文字的一些先验知识来减少计算复杂度，例如一个典型的文本区域可以看成为一个水平矩形区域，其中有很陡峭的边缘，字符与背景之间有显著差异；(2)改善字符区域的图像质量，可先利用双线性插值方法来分别提高图像水平和垂直方向的分辨率，然后用多个连续图像帧综合方法来提高插值后的图像质量；(3)字符抽取，即对字符区域的图像进行二值化处理，分割出每个字符；(4)字符识别，即采用目前比较成熟的印刷体字符识别技术来进行识别。

3.1 文字检测和抽取

3.1.1 基于边缘的方法

基于边缘的方法是通过寻找垂直边缘来检测文字。因为文字笔画丰富，且使得文字所在图像区域的边缘非常丰富，所以该方法首先检测出图像的边缘，然后通过平滑滤波或形态学膨胀的方法来将边缘连接成为文字块，再使用一些启发式规则来对文字块进行进一步筛选。文献[21]就先用一个 3×3 的水平差分过滤器来获得垂直边界，然后用平滑过滤器来使分离的文字部分相连，并排除多余碎片，再利用一些文字行的特征(如大于 70 像素、45%以上的填满率、横纵比率大于 0.75 等)来查找文字区域。在 MIPS R4499 200MHz 计算机上用该方法处理一幅 352×240 的图像大约需 0.8 s。还有文献[6,22,23]也是利用图像的垂直边缘检测来定位文字区域。

虽然基于边缘的方法可以达到快速检测文字的效果，但该方法不能适应图像背景的复杂变化，检测错误率较高。

3.1.2 基于纹理的方法

基于纹理的方法是利用纹理特征去决定一个像素点或像素块是否属于文字。由于字符通常由许多较细笔划组成，因此存在笔划的区域通常也是全图纹理较丰富的区域，实现对纹理的寻找即可以寻找到字符的区域。Wu 等提出了一种基于 K-means 的算法[15,22,24]去识别文字像素，该方法在 3 个标度下使用了 9 个高斯二阶导数。Li Hui-ping 等使用神经网络在 Haar 小波解析特征空间去抽取文字块[25]。Zhong 和 Jain 在文献[12,13]中提出一种方法，该

方法综合分析了空间差异(纹理特征)和连通区域。基于纹理的方法虽能检测复杂背景下的文字,但其计算非常耗时[25],并且文字精确定位的稳定性也不好[12]。此外,还有通过 Gabor 滤波[26],空间方差分析等通过分析纹理来提取文字区域的方法。

基于纹理的方法虽具有一定的通用性,但这类方法对于文本的字体和风格比较敏感,存在着定位不准和算法复杂度高的缺点,而且为了提取纹理信息,有时必须通过对全图进行微分运算来寻找微分结果较大的区域。由于对全图进行微分运算速度比较慢,有时还需消耗一定的资源,因此该方法的效率比较低。在利用纹理信息进行分割的时候,还需注意防止全图高频噪声的干扰,否则将严重影响分割算法的灵敏度。

3.1.3　基于区域的方法

基于区域的方法是把字符作为满足特定启发式规则的单色区域来检测。假设每个字符的像素都有相似的颜色,那么,用图像分割的方法[19~21,27,28]或颜色聚类的方法[14]或连通区分析技术[29]即可把字符从背景中分割出来;然后再使用一些简单的启发式规则,如区域的尺寸和长宽比或者基线等来对分割得到的区域进行进一步筛选即可得到字符。在文献[18]中的文字定位算法就是基于连通区域的分析,需要文字或其背景是单色的。基于区域的方法不仅能识别人工的字幕,也能分割出图像背景中的文字。然而,由于图像和视频帧中文字并不总是单色的,故这种方法对于复杂背景图像和视频来说,其鲁棒性较差。由此可见,基于区域的方法虽然具有很高的处理速度和定位精度,但是其只适用于二值图像,而不适用于彩色和灰度图像[29]。

3.1.4　基于学习的方法

对视频流字幕进行定位面临很多困难,如,(1)视频流字幕的大小尺寸经常发生变化,且在同一场景视频图像序列中,大尺寸和小尺寸字幕会同时出现;(2)视频流字幕字体呈现多样性,不同语种的字符形式不一样,即使对同一语种来说,也存在形状多样的字符;(3)由于视频字幕和视频背景颜色都是多变的,且字幕是嵌入视频背景中的,因此字幕色彩信息是不可预测和复杂的;(4)在有些情况下字幕会进行左右平移或上下垂直移动,因此在定位提取时要考虑字幕的运动状态。

由此可见,视频字幕的定位不能只考虑字幕本身固有特征,还应该考虑利用一种学习机制去处理这些多变因素。事实上,视频字幕定位就是构造一个学习机,使之能实现两类模式分类,即在视频中对字幕与非字幕进行分类,但由于现有视频字幕定位分类技术只考虑到特征提取因素,而没有考虑分类机制和如何保证不产生过学习问题,因此只有分别考虑这两个因素的分类机制,才能在适宜样本数目的前提下,取得最好分类效果。

庄越挺等提出一种使用 SVM 机制来自动定位提取视频字幕的方案[30],即首先对每幅视频图像按照 $N\times N$ 大小切分成若干图像子块,然后把每个子块分别人工训练标注为字幕和非字幕两类,并通过提取图像的子块特征向量来训练 SVM 分类器。对于测试图像,则首先将其切分成子块,然后应用训练好的 SVM 分类器对其进行判断,最后通过后期处理进行去噪和合成,即可得到字幕提取结果。Chen Da-tong 等也使用了 SVM 来进行图像中文字区的分类[28]。两种方法在样本不是很多的情况下,都实现了较高精度的视频字幕定位提取。

Lienhart 等使用了多层前向反馈神经网络[31],即先通过训练一个前向反馈神经网络来将文字行在一个固定的尺寸和位置上检测出来;然后将所有尺寸和位置上的网络输出整合到一个单一文字图中,作为候选文字行。Chang 等人先将一帧视频图像分成 16×16 像素的小区域,然后用小波神经网络方法来判别每个小区域是否是字符区域[32];Chen Xiang-rong 等利用基于矢量量化的贝叶斯分类器进行非文字候选区图像块的消除[33]。

基于学习的方法作为一种智能识别方法,其虽在相当程度上解决了许多传统方法遇到的困难,但由于其需要事先通过选取样本来对分类学习机进行训练,所以,训练样本集与测试样本集的相似程度就决定了该方法的最终识别效果。

3.1.5　基于视频的多帧平均方法

视频字幕一般具有如下时空特性:

(1) 字幕的存在可能跨越若干帧,甚至若干镜头;

(2) 字幕存在时,尽管不同帧之间变化很大,但是字幕所在区域的亮度或颜色变化不大;

(3) 字幕出现时,字幕对应区域在相继帧间会出现很大亮度或颜色的变化,同样,当字幕消失时,也会产生类似跳变;

这样通过对字幕出现或消失的相邻两帧进行比较,就可以检测得到候选字幕区域;对字幕持续存在

的多帧进行平均，就可以进一步排除一些被错误检测到的非字幕候选区域，而且可以使字幕候选区域的图像质量得到改善和增强[21,34~36]。

由于视频多帧的利用，就需要采用视频结构化方法来有效地选取所需要的若干视频帧。这样，该方法的效能和自动化程度就很大程度上依赖于视频结构化处理的效能。

3.2 文字识别方法

文字识别是一项重要的技术，只有把图像和视频中的文字识别转换为计算机能够统一识别和存储的内码，才能实现对图像和视频的自动理解和标注，再结合其他图像和视频理解技术，才能提高对视频内容的整体理解程度，以实现图像和视频的基于高层语义内容的检索，然而，和普通印刷体光学字符识别不同，图像和视频中的字符识别面临如下两个技术难题：(1)很低的分辨率，即由于电视制式和图像大小的限制，因此字符的分辨率不可能太高；(2)复杂多变的背景，即由于字符通常叠加在复杂背景之上，有时候字符和背景具有类似颜色，有时候字符区域又呈半透明状态，因此不易区分字符和背景，难以实现字符分割。

3.2.1 候选字幕区域的图像二值化

对于字符区域图像质量的改善，文献[22]首先利用双线性插值方法，将水平和垂直方向的像素密度分别提高 2 或 4 倍，然后用多个连续视频图像帧综合方法来提高插值后的图像质量。一般情况下，由于连续视频图像中字符亮度不变而背景变化，因此可以利用这个特点综合几帧连续图像，以提高字符区域的图像质量[34]，然后再通过对字符区域的图像进行二值化处理来分割出每个字符。

3.2.2 文字识别

目前，20 世纪 70 年代兴起的光学字符识别(OCR)技术已日渐完善。由于其在对文字材料进行扫描并将其转换为计算机能够统一识别和存储的内码方面已取得了较大成功，所以，绝大多数的相关文献和系统[37,38]都使用了已商业化的 OCR 软件模块来进行文字识别的内码转换，但由于文本区域的低分辨率和复杂背景，因此视频文字二值化后的效果与印刷文字的扫描二值化结果相差较大，且由于现有的 OCR 系统主要是针对扫描印刷文字的，所以识别效果不够理想。目前已有一些系统使用查词典的方法来提高识别效果，并取得了一些效果，但总的来说，文字的识别转换，尤其是低分辨率文字图像识别转换，还有待进一步的研究，以提高识别正确率。

4 存在的问题及可进一步研究的方向

通过以上对图像和视频中文字获取技术以及现有的实验原型系统的分析可以看出，目前图像和视频中的文字获取所面临的最大困难无外乎以下几个方面：①图像分辨率低，图像质量差；②文字图像的背景复杂；③文字的尺寸、字体、颜色、排列方式和运动方式多变。在这种条件下，现有方法仍存在以下需要在今后的研究中继续努力的问题：

(1)由于现有许多方法对阈值的依赖非常强，如边缘检测后的字幕行判别、文字提取时的图像二值化等，因此自适应的、有效的、智能的阈值选取方法需要进一步研究。

(2)虽然随着图像和视频的压缩编码技术以及高清晰数字电视的发展，图像和视频的质量将得到比较大的改善，但是图像质量的增强，尤其是文字图像的增强，仍需要进一步研究。

(3)现代图像和视频的制作水平日新月异，为了吸引人们的注意力，其文字特效的多变、背景的变化越来越复杂，在这样的复杂变化中，智能的机器学习方法在文字识别中的有效应用，将是以后研究的重点。

(4)不难发现，目前绝大多数的系统都采用现有的 OCR 技术来进行图像文字到 ASCII 码文字的转换，虽然 OCR 从 20 世纪 70 年代兴起到现在发展已经比较成熟，但由于其主要针对的是扫描的高分辨率图像和不包含字幕制作特效的手写或印刷体文字，所以对图像和视频中文字的识别效果并不理想，针对图像和视频文字的字符识别系统也将是以后研究的重点。

(5)在有效的识别后处理过程中，尽管已有学者应用查词典和语音识别的方法来对图像和视频中文字的获取结果进行修正，并取得了一定的效果，但随着人类对自身认识(如人类的联想思考、视觉、心理等)的不断提高，后处理过程将得到进一步发展。

(6)目前许多系统还处于实验原型的阶段，系统的实用化还需要进一步的研究，如多种方法的综合运用、合适的计算复杂度、系统处理过程的自动化程度等。

5 结　论

随着计算机技术、多媒体技术以及通讯技术的飞速发展,信息量急剧增长,相当多的新增信息都是以数字形式存在着,它们不仅包括文字和声音,更多的是图形、图像和视频等视觉信息,因此,如何管理和检索海量的图像和视频数据已经成为全球学术界和工业界一个富有挑战性的热门话题之一。在此形势下,基于内容的多媒体检索的研究和被称为多媒体内容描述接口的国际标准 MPEG-7 的制定也就引起了人们广泛的关注,而图像中的文字,尤其是网页和视频图像中的文字,就是其一个高层语义的来源。实践证明,这些文字对图像和视频的高层语义索引和检索是非常有价值的,虽然人眼可以轻而易举地识别出这些文字,但对于计算机来说,这些文字的获取还面临很大困难。正如本文所综述的,现有的许多工作对这些难题都做了有益的尝试和得到一定程度上的解决,但其结果离实用化还有相当的距离,需要国内外学者的继续努力。

参 考 文 献

1 卢汉清,孔维新,廖明等. 基于内容的视频信号与图像库检索中的图像技术[J]. 自动化学报, 2001,**27**(1):56～59.

2 庄越挺,潘云鹤,芮勇等. 基于内容的图像检索综述[J]. 模式识别与人工智能, 1999,**12**(2):170～177.

3 王惠锋,孙正兴,王箭. 语义图像检索研究进展[J]. 计算机研究与发展, 2002,**39**(5):513～523.

4 Zhou J, Lopresti D. Extracting text from WWW images[A]. In: Fourth International Conference on Document Analysis and Recognition(ICDAR)[C], Ulm, Germany, 1997,**1**:248～252.

5 Lopresti D, Zhou J. Locating and recognizing text in WWW images[J]. Information Retrieval, 2000,**2**(2):177～206.

6 Hori O. A video text extraction method for character recognition [A]. In: Proceedings of 5th International Conference Document Analysis and Recognition (ICDAR1999)[C], Bangalore, India, 1999:25～28.

7 Lienhart R. Automatic text recognition in digital videos[A]. In: Proceedings of SPIE, Image and Video Processing IV[C], San Jose, CA, USA, 1996,**2666**:180～188.

8 Lienhart R. Indexing and retrieval of digital video sequences based on automatic text recognition[A]. In: Proceedings of 4th ACM International Multimedia Conference[C], Boston, MA, USA, 1996:212～216.

9 Pfeiffer S, Lienhart R, Fischer S, *et al*. Abstracting digital movies automatically[J]. Journal Vision Communication. Image Represent, 1996,**7**(4):345～353.

10 Wactlar H D, Christel M G, Gong Y, *et al*. Lessons learned from building a terabyte digital video library [J]. IEEE Computer, 1999,**32**(2):66～73.

11 Mori S, Suen C Y, Yamamoto K. Historical review of OCR research and development[J]. In: Proceedings of IEEE, 1992, **80**(7):1029～1058.

12 Zhong Y, Karu K, Jain A K. Locating text in complex color images [J]. Pattern Recognition, 1995,**28**(10):1523～1536.

13 Jain A K, Yu B. Automatic text location in images and video frames[J]. Pattern Recognition, 1998,**31**(12):2055～2076.

14 Smith M A, Kanade T. Video skimming for quick browsing based on audio and image characterization [R]. Technology Report CMU-CS-95-186, Carnegie Mellon University, Pittsburgh, PA, USA, July 1995.

15 Wu V, Manmatha R, Riseman E M. Finding text in images [A]. In: Proceedings of 2nd ACM International Conference Digital Libraries[C]. Philadelphia, PA, USA, 1997:23～26.

16 Sato T, Kanade T, Hughes E, *et al*. Video OCR: Indexing digital news libraries by recognition of superimposed caption [J]. Multimedia Systems, 1999,**7**(5):385～395.

17 Sato T, Kanade T, Hughes E K, *et al*. Video OCR for digital news archives [A]. In: Proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database (CAVID'98)[C], Bombay, India, 1998:52～60.

18 Lienhart R. Automatic text recognition for video indexing[A]. In: Proceedings of ACM Multimedia 96[C], Boston, MA, USA, 1996:11～20.

19 Lienhart R, Effelsberg W. Automatic text segmentation and text recognition for video indexing [J]. Multimedia Systems, 2000,**8**(2):69～81.

20 Li Hui-ping, Doermann D, Kia O. Automatic text detection and tracking in digital video [J]. IEEE Transactions Image Processing, 2000,**9**(1):147～156.

21 Li Hui-ping, Kia O, Doermann D. Text enhancement in digital videos[A]. In: Proceedings of SPIE99-Document Recognition and Retrieval[C], San Jose, CA, USA, January 1999:1～8.

22 Wu V, Manmatha R, Riseman E. TextFinder: An automatic system to detect and recognize text in images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1999,**21**(11):1224～1229.

23 Hua Xian-sheng, Chen Xiang-rong, Liu Wen-yin, *et al*. Automatic location of text in video frames [A]. In: 3rd International Workshop on Multimedia Information Retrieval (MIR2001)[C], Ottawa, Canada, 2001:126～129.

24 Wu V, Manmatha R. Document image clean ～ up and binarization [A]. In: Proceedings of SPIE Symposium on Electronic Imaging1998 [C]. San Jose, CA, USA, January 1998:263～273.

25 Li Hui-ping, Doermann D, Kia O. Automatic text detection and tracking in digital video[R]. LAMP Technology Report 028, Maryland University, USA, 1998.

26 Jain A K, Sushil Bhattacharjee. Text segmentation using gabor filters for automatic document processing[J]. Machine Vision and Applications, 1992, **5**(3): 169～184.

27 Wernicke A. Text localization and text segmentation in images, videos and web pages [D]. M. S. thesis, University of Mannheim, Mannheim, Germany, Mar. 2000.

28 Chen Da-tong, Bourland Herve, Thiran Jean-Philippe. Text identification in complex background using SVM [A]. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition2001 [C], Kauai Marriott, Hawaii, USA, 2001, **2**: 621～626.

29 Tan C L. Text extraction using pyramid [J]. Pattern Recognition, 1998, **31**(1): 63～72.

30 庄越挺,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.

31 Lienhart R, Wernicke Axel. Localizing and segmenting text in images and videos [J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, **12**(4): 256～268.

32 Chang S F, Chen W, Meng H J, *et al*. VideoQ-an automatic content-based video search system using visual cues [A]. In: Proceedings of ACM Multimedia Conference[C], Seattle, WA, USA, 1997: 147～151.

33 Chen Xiang-rong, Zhang Hong-jiang. Text area detection from video frames [A]. In: 2nd IEEE Pacific-Rim Conference on Multimedia (PCM 2001)[C], Beijing, China, 2001: 222～228.

34 Li Hui-ping, Doermann D. Text enhancement in digital video using multiple frame integration[A]. In: Proceedings of ACM Multimedia 1999[C], Orlando FL, USA, 1999: 19～22.

35 Hua Xian-sheng, Yin Pei, Zhang Hong-jiang. Efficient video text recognition using multiple frame integration[A]. In: 2002 International Conference on Image Processing (ICIP2002)[C], Rochester, New York, USA, 2002: 214～217.

36 Tang Xiao-ou, Luo Bo, Gao Xin-bo, *et al*. Video text extraction using temporal feature vectors [A]. In: Proceedings of IEEE International Conference on Multimedia and Expo(ICME2002) [C], Lausanne, Switzerland, 2002: 168～171.

37 Ohya J, Shio A, Aksmatsu S. Recognition characters in scene images [J]. IEEE Transactions Pattern Analysis and Machine Intelligence, 1994, **16**(2): 214～220.

38 Xi Jie, Hua Xian-sheng, Chen Xiang-rong. A video text detection and recognition system [A]. In: IEEE International Conference on Multimedia and Expo (ICME 2001)[C], Waseda University, Tokyo, Japan, August 22-25, 2001: 1080～1083.

王 勇 1976 年生,先后获国防科技大学自动控制系工学学士学位、机电工程与自动化学院工学硕士学位,现为国防科技大学与西南电子电信研究所联合培养博士生。主要研究方向为智能图像识别、基于内容的多媒体检索。

郑 辉 1957 年生,毕业于南京东南大学无线电系,2000 年赴美国夏威夷大学做访问学者,现为西南电子电信技术研究所总工、教授级高工、博士生导师、IEEE 高级会员。主要研究方向为多媒体通信、盲信号处理、移动通信技术。

胡德文 1963 年生,先后获西安交通大学硕士、国防科技大学工学博士学位,1995～1996 年赴英国 Sheffield 大学做高级访问学者,现为国防科技大学机电工程与自动化学院教授、博士生导师。主要从事图像处理、神经网络、系统辨识、脑科学等研究。