一种基于多帧融合的视频文字信息提取方法

庞琳[1,2,3]，刘安安[1,2,]，宋砚[1,2,3]，李锦涛[1,2]，张勇东[1,2]，唐胜[1,2]

(1. 中国科学院计算技术研究所虚拟现实技术实验室，北京，100190；
2. 中国科学院计算技术研究所 智能信息处理重点实验室，北京，100190；
3. 中国科学院研究生院，北京， 100049)

摘 要： 视频文字信息的提取对于基于内容的视频分析具有重要意义。本文提出了一种综合图像边缘信息和视频时序信息的视频文字信息提取方法,通过多尺度文字定位,多帧文字区域的跟踪和增强以及自适应的文字区域分割,实现文字信息提取。实验表明,该方法具有较高的准确性,并对复杂背景和文字大小的变化具有较强的鲁棒性。

关键词： 视频文字信息提取 ;多尺度 ； 多帧融合 ； 均值迁移

Video Text Extraction based on Multiple Frames Integration

Lin Pang[1,2,3] , Anan Liu[1,2] , Yan Song[1,2,3,] , Jintao Li[1,2],

Yongdong Zhang[1, 2] , Sheng Tang[1, 2]

(1 Virtual Reality Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190)

(2 Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190)

(3 Graduate School of the Chinese Academy of Sciences, 100049)

Abstract: Video text extraction plays an important role in content based video analysis. This paper presents a novel method for video text extraction based on both edge and temporal characteristics. Texts are extracted by multiple scale text localization, text area tracking and integration in consecutive image frames and adaptive text image segmentation. Experimental results show that this method can improve the text extraction ratio especially for images with complex background and multiple scale text.

key words: Video Text Extraction; Multiple Scale; Multiple Frames Integration ; Mean Shift

1 引言

随着信息技术及互联网技术的飞速发展,网络信息成为一种人们熟知的信息来源。同时,Internet 的开放性也带来了巨大的安全隐患,网络安全成为关系到国家与社会安全的一项重要问题。如何对互联网上的海量信息进行有效的分析和管理，提高信息识别的能力，成为目前亟待解决的问题。由于视频中的文字信息对视频内容有较强的描述作用，且具有提取容易、直接表征视频高层语义的特点，因此视频文字信息提取对于视频内容分析和检索具有重要意义。

视频中文字提取主要包括文字区域检测、分割、识别三个步骤[1]。Zhong 等[2] 通过使用连通成分定位复杂背景中的文字, 但是，当文字出现不同颜色时该方法检测结果不理想。Chen 等[3]使用 Canny 边缘算子检测边缘,然后将边缘通过形态学闭合运算形成文字块,并使用 SVM 分类器检测文字区域, 但是该方法对于多尺度的文字区域不具有通用性。Lienhart 等[4]利用文字的单色性与背景的高对比度和视频字幕的简单纹理来进行图

基金资助：中国国家基础研究项目(973 项目,2007CB311100)，中国国家高科技和研究发展项目(863 项目，2007AA01Z416),北京市科技新星项目(2007B071)

作者简介：庞琳（1985-)，女，山东聊城，博士在读; Email: panglin@ict.ac.cn

像分割，但该方法受噪声和复杂背景的影响较大。可见，尽管已有大量研究人员从事该方面的研究，但是，由于图像背景复杂，文字的字体、尺寸、颜色的多样等使得目前还没有通用且理想的视频文字信息提取方法。

针对上述问题，我们提出了一种综合图像边缘信息和视频时序信息的多尺度视频文字信息提取的方法。首先通过边缘能量及连通分析对视频图像进行多尺度的文字区域检测，并使用边缘的水平和垂直投影图分割对文字行进行准确定位；其次使用基于颜色直方图分布的 Mean Shift [5]方法进行连续多帧文字区域的跟踪和融合；再次使用 Niblack[6]方法对增强的文字区域进行自适应的文字分割；最后通过 OCR 软件进行文字识别。本文主要包括如下内容：第 2 节详细介绍了视频文字信息提取方法；第 3 节中介绍了实验的结果；第 4 节对本文进行总结。

2 视频文字信息提取

我们提出的基于多帧融合的视频文字信息提取方法包含以下步骤，如图 1 所示。下面分别对各个部分进行详细介绍。

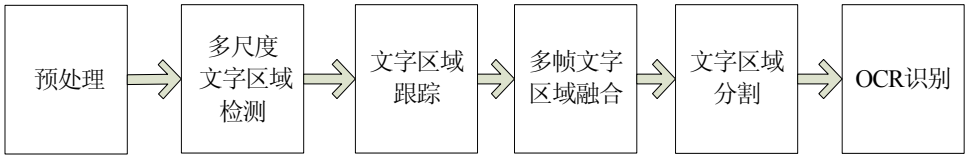

图 1 视频文字信息提取方法流程图

Fig.1 Flow Chart of Multiple frames Intergration based Video Text Extration

2.1 多尺度文字区域检测

2.1.1 图像金字塔变换

由于视频图像中包含的字体大小不一致，过小的字体往往被漏检，而过大的字体往往被部分定位，造成对不同大小的文字在检测准确率上的偏差。因此为了对多尺度文字进行准确检测，我们采用图像金字塔的方法对视频图像进行尺度变换操作。将原视频图像的尺度记为 W×H，我们采用 Shannon 插值方法[7]将其分别变换为尺寸为 2W×2H 的图像和 0.5W×0.5H 的图像，由这三个尺度的图像构成了图像金字塔。

2.1.2 文字区域检测

在视频图像帧中，文字区域通常包含有丰富的水平和垂直方向边缘，因此我们对三个尺度的图像分别使用 Sobel算子提取边缘，采用最小误差法自适应选择阈值，对边缘图像进行二值化。经过阈值过滤后，部分非文字区域的边缘被去除掉。

然后对边缘二值化图建立标记图，统计边缘图中边缘象素点的分布状况。我们使用一个大小为 4*4 的滑动窗口对边缘图进行扫描，统计窗口内上、下、左、右四个 2*2 子窗口内的边缘象素个数 n_a，n_b，n_c，n_d。计算表达每个窗口边缘象素分布散度的值 n

$$n = n_a \times n_b \times n_c \times n_d \tag{1}$$

在标记图中，每个点对应上述一个 4*4 大小的窗口，当 n>0 时，其象素值 d 为 1，即只有周围的边缘点分散分布的块才会被标记成 1，而对应背景中物体边缘的区域由于边缘分布集中被去除。对标记图进行腐蚀处理并去除孤立点，并使用序贯法分析其中的连通区域，对每个连通区域获得其外接矩形，并根据经验性规则把重叠或很接近的矩形块进行合并。最后将标记图中检测到的矩形区域还原到原图像中，即定位到文字区域。

由于得到的区域可能包含多行和多列，通过将文字区域边缘图进行水平方向和垂直方向的投影中寻找极小值点可以得到各行和列的分割点进行文字区域的精确定位。

对图像金字塔中三种尺度的图像分别进行文本区域检测后，将检测结果进行“或”操作，就得到了多尺度融合后的文字区域。如图 3 所示是多尺度检测的过程。

图 2 单帧文字区域检测图

Fig.2 Text area detection for single frame

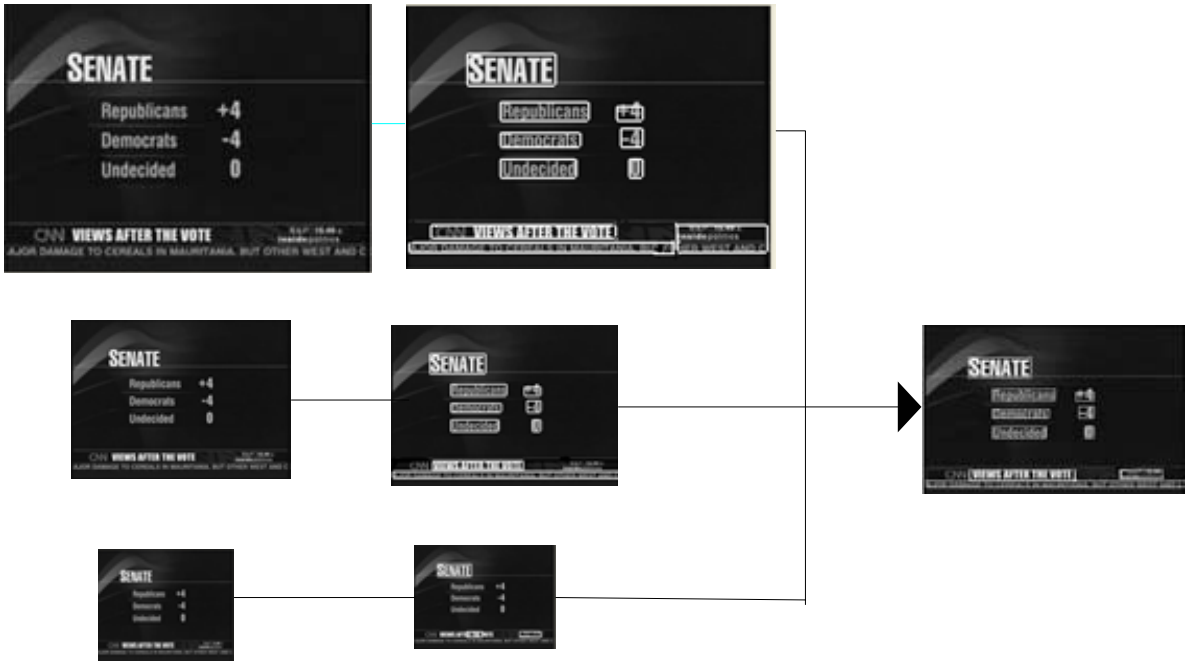

图 3 文字区域多帧融合

Fig.3 Text area Integration for multiple frames

2.2 多帧文字区域增强及文字分割

2.2.1 多帧文字区域增强

视频中文字区域一般具有如下特性：(1)视频连续帧往往包含相同的文字；(2)尽管不同帧之间视频内容可能存在较大变化,但是文字所在区域的亮度或颜色变化不大。有效地利用视频文字的时序特性来进行文字跟踪并增强文字区域的对比度,有利于文字信息的提取。

由于视频图像中文字是静止的,而大部分背景物体是运动的,所以对于各帧相同位置的像素点,若它属于背景,则变化较大,若属于文字,则变化较小。通常文字像素的亮度高于背景物体,所以将具有相同文字的多帧图像相同位置的象素点取最小值可以增强文字与背景的对比度。

我们使用均值迁移跟踪算法（Mean Shift 算法）[5]通过目标和当前搜索窗口中的候选目标的核函数加权直方图分布建立模型,然后以两个模型的相似性最大作为判断准则,使搜索窗口移向目标的真实位置。

对需跟踪的运动目标图像,根据映射 $b: R \to \{1,...,m\}$ 把相应位置处像素颜色进行 m 级量化,建立相应的颜色直方图模型,则目标颜色分布可以表示为

$$q_u = c\sum_{i=1}^{n} k\left(\left\|\frac{X_0 - X_i}{h}\right\|^2\right)\delta\left[b(X_i) - u\right] \tag{2}$$

其中 X_0 是窗口（共包含 n 个象素）中心象素的坐标, X_i 是窗口内第 i 个象素的坐标, $k(\|x\|^2)$ 是核函数,b

表示核函数窗宽,即跟踪窗口的半径；δ 表示 Kronecker delta 函数。C 为归一化常数。通过约束条件 $\sum_{i=1}^{m} q_i = 1$ 即可求得。

为了度量两者之间的相似性,我们定义目标模型和候选目标模型之间的相似性函数

$$\rho(p(y),q)=\sum_{u=1}^{m}\sqrt{p_u(y)q_u} \tag{3}$$

其中 $\rho(p(y_0),q)$ 表示目标模型q与候选目标模型p的特征概率密度分布的Bhattacharyya系数, p_u和q_u分别表示两个直方图模型中对应分量的值。

在当前帧中,以前一帧文字区域的位置作为当前搜索窗口的初始位置,设窗口中心为y_0,根据相似性函数判断,在y_0邻域内寻找局部最优的目标位置y_1。 ρ 最大时,均值迁移向量m_s为

$$m_s=y_1-y_0=\frac{\sum_{i=1}^{n} x_i \omega k\left(\left\|\frac{y_0-x_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} \omega k\left(\left\|\frac{y_0-x_i}{h}\right\|^2\right)}-y_0 \tag{4}$$

均值迁移算法就是反复不断的把搜索窗口按照均值迁移矢量制定方向进行移动以至收敛于某个极值,此时窗口中心位置对应于运动目标的中心,实现了目标的跟踪。由于在视频相邻帧间相同文字区域出现的位置一般没有过大偏移,故迭代很少的次数即可达到收敛,从而有效的实现了快速跟踪。如若有限次内不能收敛则说明当前帧中文字区域有所变化。于是从当前帧开始重新进行文字区域的检测,并重新进行跟踪。

图 4　连续多帧跟踪

Fig.4 Multiple frames tracking

图 5　多帧融合

Fig.5 Multiple frame Integration

2.2.2 文字区域分割

经过多帧增强后的图中,文字区域中文字与背景的对比度得到提高,并减少了噪声。接下来我们使用一种自适应的 Niblack[6]局部阈值法对文字区域进行二值化。对文字区域的每个点,计算其 15×15 邻域内其余点的亮度均值和方差,根据均值和方差确定二值化的阈值,使用此局部阈值决定当前点是否属于文字。

对像素点(x,y), 其局部阈值为

$$T(x,y)=m(x,y)+k\times s(x,y) \tag{5}$$

其中 m(x,y)和 s(x,y)分别为邻域内其余点的亮度均值和方差, k 是用来调整字符目标边界的参数, 本文实验中选用 k=-0.2 时达到最好的分割效果。

(a)单帧图像的分割结果　　(b)多帧融合后图像的分割结果

图 6　多帧增强对分割结果的优化

Fig.6 Enhancement for Segmentation of the Multiple frame Integration

3 实验

实验中, 我们选用了国际视频检索评测TRECVID 2005的镜头边界检测任务数据集里的部分视频作为测试数据, 其中包含CNN,NBC,CCTV,NTDTV四个电视台的100分钟的节目, 共包括3734帧文字图像（尺度为352*240)。我们使用汉王OCR软件HWSDK1.2从二值化后的图像中识别文字。对于文字识别的评价标准, 我们采用字符识别率和准确率, 其中识别率是识别正确的字符数与字符区域所含字符总数之比, 准确率是识别正确的字符数与识出的字符总数之比。我们对字符提取的结果计算识别率和准确率, 并与文献[8]中的方法进行比较, 如图7所示。

从图7 中可以得到如下结论：

首先，由于我们的测试视频具有背景复杂和文字尺度变化大等特点，使用文献[8]中的方法进行文字识别得到的英文字符识别率和准确率分别为90.15%和91.22%，而中文字符的识别率和准确率仅为62.40%和73.80%，对中文的识别能力远远低于英文。而从图7中可以看到我们的方法在仅加多尺度检测时相比文献[8]中的方法对于不同语言的字符识别率和准确率已经获得了较大的提高，而再增加多帧增强后，相比文献[8]的方法，英文字符的识别率和准确率分别提高了3.39%和3.22%, 而中文字符的识别率和准确率更分别提高了17.89%和9.4%，对识别效果得到更进一步的改善。

另外，我们的方法对于文字识别尤其是中文字符识别效果的提高是非常显著的，这是由于我们综合使用了空间和时间信息，增强了图像的对比度，有效排除了背景干扰，因而对复杂背景，文字尺度和语言多样化的视频文字识别具有较好的鲁棒性。

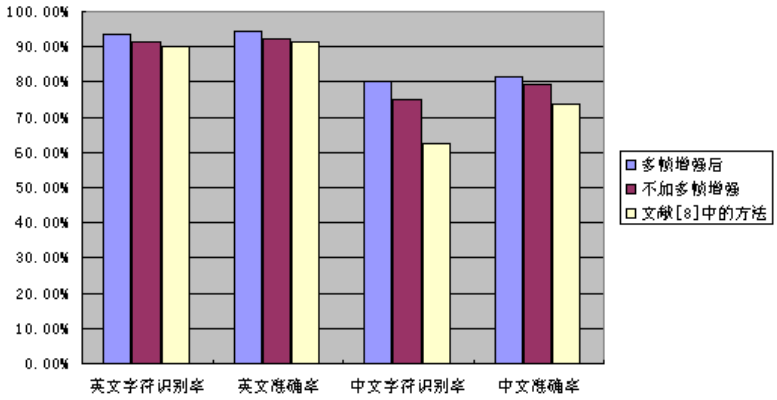

图 7　实验结果对比图

Fig.7 Experimental results

4 总结

在本文中, 我们提出了一种综合图像边缘信息以及视频时序信息的视频文字信息提取方法。通过对视频图像进行多尺度的文字区域检测, 基于颜色直方图分布的Mean Shift方法对时序上连续多帧相同文字区域进行跟踪和融合，以及自适应的文字区域分割, 实现了文字信息提取。实验表明，由于融合了多帧信息进行文字区域增强, 能够弥补对单帧图像进行处理的不足, 有效实现了文字对象的分割效果, 从而提高了字符识别率，因此该方法具有较高的准确性, 并对复杂背景和文字大小的变化具有较强的鲁棒性。

参考文献：

[1] K Jung, K In Kim, A K. Jain. Text Information Extraction in Images and Video: A Survey. Pattern Recognition, 37 (5):977-997,2004

[2] Y.Zhong, K.Karu, and A.K.Jain. Locating text in complex color images. Pattern Recognition. 28(10),1523-1535,1995

[3] D.T.Chen, H.Bourlard, J-P.Thiran.Text identification in complex background using SVM. Computer Vision and Pattern Recognition, 2001

[4] R Lienhart, W Effelsberg. Automatic text segmentation and text recognition for video indexing , Multimedia System, ,8:69-81, 2000

[5] D Comaniciu, V Ramesh, P Meer. Kernel-based object tracking . IEEE Transactions on Pattern Analysis and Machine Intelligence, 25(5):564-577,2003

[6] Niblack W. An introduction to image processing. Prentice Hall ,Englewood Cliffs,1986

[7] Huiping Li, Omid Kia, David Doermann .Text Enhancement in Digital Video. Proceedings of SPIE Document Recognition and Retrieval VI,1999

[8] M.R.Lyu, J.Q.Song,and M.Cai, A comprehensive method for multilingual video text detection, localization, and extraction. IEEE Transactions on Circuit and System on Video Technology, 15(2), 2005