中 文 信 息 学 报
第 18 卷 第 5 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.5

文章编号:1003 - 0077(2004)05 - 0042 - 06

自然场景文本定位

欧文武,朱军民,刘昌平

(中国科学院自动化研究所 文字识别工程中心,北京 100080)

摘要:随着自然场景文本识别研究的不断深入,建立标准的场景文本图像库和了解该领域的研究现状变得越来越重要。为此,2003 年国际文档分析和识别大会专门建立了一个这样的图像库,并组织了自然场景文本识别比赛,我们参加了其中的自然场景文本定位分赛。本文对我们参加这次比赛的算法做了介绍并给出了比赛结果,在文章最后,对参赛算法做了比较,指出了场景文本定位的发展现状。

关键词:人工智能;模式识别;文本定位;边缘密度;字符识别;图像处理

中图分类号:TP391.41 **文献标识码**:A

Text Location in Natural Scene

OU Wen-wu, ZHU Jun-min, LIU Chang-ping

(Character Recognition Center, Institute of Automation Chinese Academy of Science, Beijing 100080, China)

Abstract: With the rapid growth of research on text recognition in natural scene, it turns to be urgent to understand the development situation of this art and to establish common benchmark datasets. So the organizers of international conference on document analysis and recognition 2003 develop a dataset on this art specially and organize the robust reading competition, and we take part in the sub-competition of robust text location. In this paper we shall introduce our algorithm on this competition and give the competition result; in the end of paper we give the compare of each entries' algorithms and point out the development situation of robust text location presently.

Key words: artificial intelligence; pattern recognition; text location; edge intensity; character recognition; image processing

1 引言

经过多年的研究发展,OCR(Optical Character Recognition)技术已能高速、准确地处理一般文档,将大量的印刷、手写文档转为电子文档。然而,传统的 OCR 技术只能识别分辨率高且背景简单的扫描图像,而许多文本图像是很难用扫描方式得到的,比如广告牌、车牌、碑文等自然场景中的文本。虽然已有许多学者对自然场景的文本识别做了深入的研究,但就目前的发展状况来看,自然场景文本识别远不及人们所期望的:肉眼一样的阅读速度和准确性,并且该领域还没有公认的标准库。为了更好地了解和推动自然场景文本识别,ICDAR'2003(7th International Conference on Document Analysis and Recognition)组织者严格按照 2000 年指纹识别比赛(Finger Verification 2000 competition[1])的程序和规则,组织了 2003 年自然场景文本识别比赛(ICDAR'2003 Robust Reading Competition)。比赛的训练样本库、评分标准、比赛规则等都在网上

收稿日期:2004 - 02 - 24

基金项目:国家 863 计划资助项目(2001AA114130)

作者简介:欧文武(1977—),男,湖南人,硕士研究生,研究方向是数字图像处理,模式识别.

公布,参赛者在最后期限前按要求完成程序,由组织者用测试样本来评测参赛程序。比赛被分为三个分赛:文本定位(Robust Text Location)、字符识别(Character Recognition)和单词识别(Word Recognition)。我们参加了其中的文本定位比赛,获得第二名。

自然场景文本识别对于信息检索、数字图书馆、网页检索和智能交通等领域有重要的意义。虽然 OCR 技术经过多年的发展,已经达到实用的要求,很多公司推出了这方面的商业软件包,但是自然场景文本识别远没达到人们的期望,其中一个重要原因就是自然场景文本定位在很大程度上限制了系统的识别结果。文本定位就是找出图像中文本所在的位置或刚好包围文本的矩形区域,是文本识别非常关键的一步,文本定位好坏直接影响到整个识别系统的结果。文本定位方法大致可以分为以下三类:1. 基于区域的文本定位方法,这种方法假定同一区域的字符颜色相近,并且可和背景颜色区分;2. 基于纹理的方法,通过图像的纹理特征区分文本区域和非文本区域;3. 基于边缘的方法,利用边缘或边缘密度找出字符的位置。文本定位通常要用到先验知识或规则,即文本区域的特征。文本区域的特征包括:字符笔画边缘分布紧凑并且以水平和垂直为主,连接笔画边缘会得到比较规则的矩形区域;文本区域的连通域高度基本一致、分布比较均匀;包含文本的矩形区域满足一定的高度、宽度限制等。合理地利用这些先验知识或规则能够帮助我们正确区分文本和非文本区域,提高定位精度。

2 算法描述

我们是通过笔画边缘特征得到文本区域的,笔画边缘具有以下特点:

a. 笔画边缘比较明显,因为字符笔画颜色和背景颜色一般有较大的差异;

b. 笔画边缘比较规则,一般以水平和垂直为主,垂直笔画边缘有相近的高度;

c. 笔画边缘在空间上表现出粘连性,连接相邻的笔画边缘会得到规则的矩形区域;

算法如图 1 所示,由金字塔分解、边缘提取、形态学运算、先验知识限制和结果合成五部分组成,在文章接下来的部分,我们将分别介绍。

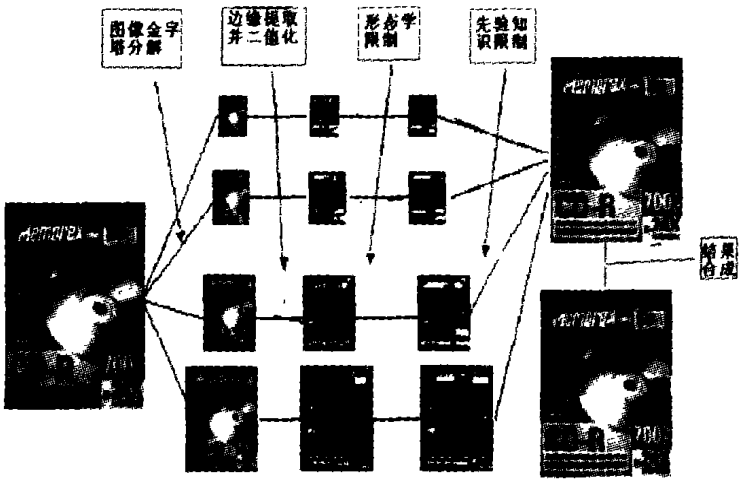


图 1 算法结构图

2.1 金字塔分解

自然场景图像上字符的大小变化范围很大,如 ICDAR 提供的图像上字符高度从 10 到 1200 个象素不等,有的单一字符占到整幅图像面积的 50% 以上,而有的不到 0.1%。目前,几乎所有的文本定位算法都对字符大小很敏感,为了能够找出大小不一的文本区域,我们采用金字塔分解的方法:将图像分解为原分辨率的 1/1,1/2,1/4,1/9 四幅子图,对每幅子图分别采用

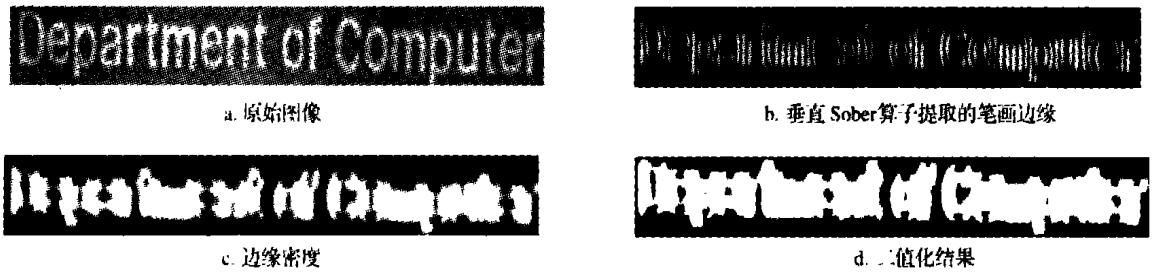
相同的文本定位算法,然后将不同子图上检测到的文本区域放大到原始图像大小,最后综合每幅子图的定位结果就可以找出大小不同的文本区域。如图 1 中,小的字符在底层子图上被检测到,而在高层的子图上找到了较大的字符,最后的定位结果中包含了不同大小的文本区域。

2.2 边缘提取和二值化

前面我们已经提到字符笔画边缘明显、分布紧凑和以垂直、水平分布为主等。文中我们采用垂直 Sober 算子提取笔画边缘(如图 2.b),然后用边缘密度连接笔画边缘形成矩形区域(如图 2.c)。边缘密度 $EI(x,y)$如公式(1)所示,其中 $E(x,y)$表示由垂直 Sober 算子提取的边缘,w 表示水平窗口大小,窗口的选取与字符的高度和稀疏程度密切相关。

$$EI(x,y) = \sum_{i=-w/2}^{i=w/2} E(x+i,y) \tag{1}$$

本文中,我们选取窗口宽度为 9 个象素,试验表明对高度小于 60 的文本效果较好。边缘密度图像规一化后,采用双阈值的 Ostu 全局二值化方法[2,3],二值化边缘密度图像,如图 2.d。



a. 原始图像　　b. 垂直 Sober算子提取的笔画边缘

c. 边缘密度　　d. 二值化结果

图 2　字符笔画空间的粘连

2.3 形态学限制

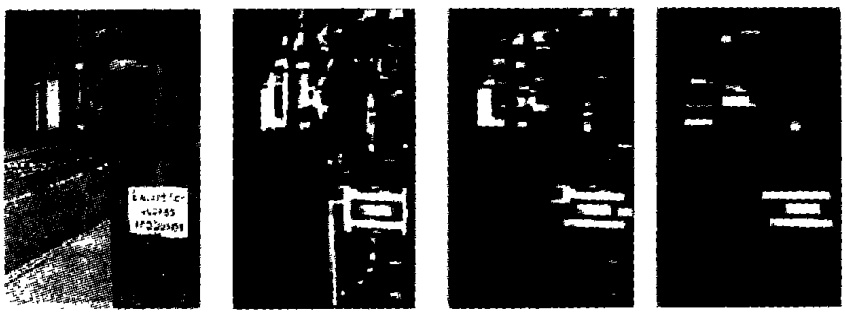
形态学运算能够对图像上的物体做形状等方面的限制,常用于目标检测等。在本文中,我们用形态学运算检测二值边缘密度图像上的矩形区域,这些矩形区域通常代表了文本位置。形态学运算包括以下两步:

step 1:对二值化的边缘密度图像做 7 个象素宽度的水平闭运算,连接字符笔画形成矩形区域,再做 15 个象素宽度的水平开运算,去除孤立的背景(如图 3.c);

step 2:求 step 1 形态学运算后图像的连通域,对每个连通域作 $w_{dilation}$ 宽度的膨胀运算,和 $w_{erosion}$ 宽度腐蚀运算(如图 3.d)。

$w_{dilation}$ 和 $w_{erosion}$ 的定义如公式(2),此处 *width*、*height* 分别为对应连通域的宽度、高度。

$$\begin{aligned} w_{dilation} &= \min(height,\ width/8) \\ w_{erosion} &= width/4 \end{aligned} \tag{2}$$



(a) 原始图像　(b) 二值边缘密度图像　(c) 第一次形态学运算结果　(d) 第二次形态学运算结果

图 3　形态学限制

从图3可以看出,第一步形态学运算去除了部分背景区域,并且将原来相连的背景和文本分离开来;第二步形态学运算后图像上只剩下部分规则的矩形区域。通过以上两步形态学运算得到了我们想要的矩形区域,图3.d右下的矩形区域准确地代表了文本位置。

2.4 先验知识限制

虽然形态学限制可以检测到文本区域,但从图3可以看出仅形态学限制是不够的,图中还有不包含文本的矩形区域。我们对形态学运算后的二值图像做连通域分析,如果连通域宽度、高度、连通域内白点象素和黑点象素的比率等满足公式(3)所示条件,则认为该连通域为备选文本区域,否则丢弃。

$$\frac{width}{height} > r \wedge \frac{white\cdots dots}{black\cdots dots} > t \tag{3}$$

此处 $width$, $height$ 分别为连通域的宽度和高度,本文中 $r=1, t=2.3$。

为了根据先验知识判断备选文本区域是否包含文本,首先我们要解决的问题是字符和背景灰度值的判断。因为我们在二值化备选文本区域时要知道是小于二值化阈值的象素作为背景,还是大于二值化阈值的象素是背景。为此我们按公式(4)计算备选区域顶部的灰度 L_{top}、底部灰度 L_{bottom} 和中部灰度 L_{middle},如果它们满足条件(5),则有:如果 $L_{top} > L_{middle}$,大于二值化阈值的为背景,反之则反;如果它们不满足条件(5)则根据备选区域内字符笔画象素应该少于背景象素来区分背景和字符灰度。

$$\begin{aligned} L_{top} &= \sum_{i=0}^{i<Width} I(i + Left, Top - 4) \\ L_{bottom} &= \sum_{i=0}^{i<Width} I(i + Left, Bottom + 4) \\ L_{middle} &= \sum_{i=0}^{i<Width} I(i + Left, (Bottom + Top)/2) \end{aligned} \tag{4}$$

此处 $I(x, y)$ 表示图像在 (x, y) 点的灰度值,$Left$, $Bottom$, Top, $Width$ 分别表示备选文本区域的左边横坐标,下面和上面纵坐标及宽度。

$$|L_{top} - L_{bottom}| < (|L_{top} - L_{middle}| + |L_{bottom} - L_{middle}|) \times 1/3 \tag{5}$$

我们从以下三个方面分别对备选文本区域做条件限制:

条件一:颜色空间可分

字符和背景一般达到肉眼能区分的程度,即文本区域内字符和背景在颜色空间上可分,并且字符象素应该占到整个文本区域象素的一定比率。我们统计备选文本区域的灰度直方图,用R J Whatmough在文献[4]中提出的模糊二值化方法,求二值化阈值 T。如果二值化阈值 T 不满足条件(6),则认为该区域不包含字符,被丢弃。文中 $a=0.2, b=0.8$。

$$\frac{Sum(p > T)}{Sum(p)} \in (a, b) \tag{6}$$

此处 $Sum(p > T)$ 表示大于阈值 T 的象素个数,$Sum(p)$ 表示矩形区域内象素个数。

条件二:连通域分布

二值化备选文本区域,做连通域分析,如图4.b中白色矩形框为连通域分析结果。因为文本区域连通域分析得到的矩形框通常代表了文本区域的字符位置,并且同一文本区域的字符应该在字符高度、宽度、字符间距等方面有一定的限制,所以我们可以用矩形框的分布来判断备选区域内是否包含文本。

a. 原始图像　　b. 二值化后连通域分析结果

图4　文本区域的连通域分布特征，白色矩形框为连通域分析的结果

因为文本区域的矩形框通常刚好包围一个或多个字符，首先我们对矩形框宽度 *width*、高度 *height*、白点象素和黑点象素比率 r 等做条件(7)所示限制，如果矩形框不满足条件(7)，则认为该矩形框内不包含字符。

$$r\in(0.2,0.8)\vee r\geqslant 0.8\wedge width>0.5\times Height\wedge height\times 2<width\vee r\geqslant 0.8\wedge height>0.5\times Height\wedge height>2\times width \tag{7}$$

此处 *Height* 为备选文本区域高度

统计所有的备选文本区域，保存到列表 *List*(i)，按下面的算法对备选区域做条件判断：

Step 1：从列表 *List*(i)中读入一个备选区域，如果 *List*(i)为空，则算法结束；

Step 2：从左到右依次统计备选区域内满足条件(7)的所有矩形框，顺序保存到列表 *CCList*(j)；

Step 3：如果列表 *CCList*(j)内相邻的两矩形框水平距离大于两倍备选区域的高度，则：

Step 3.1：备选区域分为左右两部分，如果右部分满足条件(3)，则作为新的备选区域加入列表 *List*(i)，否则丢弃；

Step 3.2：如果左部分满足条件(3)，则执行 *step* 4，否则丢弃左部分，执行 *step* 1；

Step 4：如果备选区域内所有满足条件(7)的矩形框的宽度之和大于备选区域宽度的一半，则保留该区域，否则丢弃；执行 *step* 1。

条件三：投影分析

文本区域向 x 轴的投影曲线有明显的波峰和波谷，波峰对应字符的笔画，波谷对应字符间隙，而非文本区域向 x 轴的投影曲线相对平滑，没有明显的波峰或波峰的数量较少。我们采用投影曲线的波峰数量 *Num* 和曲线方差 *Variance* 作为判断标准，分别对它们做条件限制。因为在试验中我们发现：曲线方差大小往往与投影区域高度的平方有正比关系，所以我们对曲线方差做了相应的调整，如公式(8)所示。文中如果 *Variance* 小于 0.05 或 *Num* 小于 5，则认为该区域不包含文字。

$$Variance=\frac{1}{Width\times Height^2}\sum_{i=Left}^{i<Right}(Sun(i)-mean)^2$$
$$mean=\frac{1}{Width}\sum_{i=Left}^{i<Right}Sum(i) \tag{8}$$

这里，$Sum(i)$ 表示投影曲线在 i 点对应的值；*Width*，*Height*，*Left*，*Right* 分别为备选文本区域的宽度、高度、左侧和右侧横坐标。

2.5　结果合成

因为通常情况下同一文本区域会在不同的子图上检测到，而背景一般不具备这种特征，所以为了得到最终的文本定位结果我们用投票的方式决定某一区域是否为文本区域：如果某一区域在两幅以上子图同时被检测到，则认为该区域为文本区域，否则丢弃。由于不同子图得到的文本矩形区域经常相互交错、覆盖，我们采用区域合并的方法，合并重叠的文本区域。合并条件如(9)所示，如果两个相交文本区域 R_1 和 R_2 满足条件(9)，则这两个区域合并为一个区域，本文中 $p=0.8, q=0.2, s=10$。

$$\frac{A(R_1 \cap R_2)}{\min(A(R_1),A(R_2))} > p \vee \frac{A(R_1 \cap R_2)}{\min(A(R_1),A(R_2))} > q \wedge \frac{\max(A(R_1),A(R_2))}{\min(A(R_1),A(R_2))} > s \quad (9)$$

其中 $A(R)$表示矩形区域 R 的面积，$R_1 \cap R_2$ 表示 R_1, R_2 相交区域

3 评测方法和比赛结果

对文本定位结果的评测目前有多种方法。有基于目的的方法，即用自然场景中识别正确的字符比率作为定位的评价标准，但是由于定位结果的稍微偏差可能会导致识别结果截然不同并且字符的识别还与字符串的切分及二值化方法有关，所以这种方法不能全面反映定位结果；有基于区域数量的方法，即用检测到的矩形区域数量作为评价标准，但是检测到的矩形区域经常与实际的文本区域有或多或少的偏差，比如检测到的区域可能小于实际文本区域而遗漏部分字符或大于实际区域而包含过多背景等，所以这种评价方法不够精确。为了更加全面地反映定位结果，这次比赛组织者用区域匹配的方法评测比赛算法，如式(10)所示。在(10)式中，$m(r, R)$是通过面积匹配得到的，对文本定位的精确性提出了很高的要求；p'表示准确率，p'越高误检率越低，即检测到的非文本区域越少；r'表示召回率，r'越高遗漏率越低，即没有被检测到的文本区域越少。

表 1 ICDAR'03 自然场景文本定位比赛结果

参赛者	准确性	召回率	综合结果	时间开销
Ashida	0.55	0.46	0.50	8.7 秒
HWDavid	0.44	0.46	0.45	0.3 秒
Wolf	0.30	0.44	0.35	17.0 秒
Todoran	0.19	0.18	0.18	0.3 秒
Full	0.1	0.06	0.08	0.2 秒

$$p' = \frac{\sum_{r_E \in E} m(r_E, T)}{|E|}, r' = \frac{\sum_{r_T \in T} m(r_T, E)}{|T|}$$

$$\text{where} \quad (10)$$

$$m(r,R) = \max(m_p(r,r') \mid r' \in R);$$

$$m_p(r_1,r_2) = \frac{2a(r_1 \cap r_2)}{a(r_1) + a(r_2)}$$

此处，$a(r)$表示面积，E 表示检测到的文本区域集合，T 表示实际文本区域集合，$|R|$表示集合内元素个数。

综合评测结果 f 如公式(11)表示，f 综合考虑了文本定位的精确性、准确率和召回率。文中 $a = 0.5$。

$$f = \frac{1}{a/p' + (1-a)/r'} \quad (11)$$

表 1 中给出了部分参赛算法的评测结果，其中 $t(s)$表示算法的平均时间(单位：秒)，HWDavid 为我们的算法。

4 结论

由于我们的算法是基于边缘密度的，对于某些边缘不明显的文本，结果不太理想，相比表(1)中的 Ashida 采用颜色约减(Color Reduction)的方法可以检测到连肉眼也难以识别的字符，但有时却检测不到我们算法能够检测到的文本区域。不同的文本定位方法，对不同类型图像的定位结果差异很大，同一种方法对同一种图像有时也表现的很不一致。比赛组织者按照一

(下转第 63 页)

ceedings 4th ISCA Tutorial and Research Workshop on Speech Synthesis[C], Perthshire Scotland, August 29th - September 1st, 2001.

[6] 牛正雨,柴佩琪. 基于边界点词性特征统计的韵律短语切分[J]. 中文信息学报, 2001, 15(5):19 - 25.

[7] 赵晟,陶建华,蔡莲红. 基于规则学习的韵律结构预测[J]. 中文信息学报, 2002, 16(5):30 - 37.

[8] Adam L. Berger, Stephen A. Della Pietra, Vincent J. Della Pietra. A maximum entropy approach to natural language processing[J]. Computational Linguistics 1996, 23(4): 597 - 618.

[9] Adwait Ratnaparkhi. A Maximum Entropy Part-Of-Speech Tagger[A]. Proceedings of the Empirical Methods in Natural Language Processing Conference[C], May 17 - 18, 1996.

[10] 周雅倩,郭以昆,黄萱菁,吴立德. 基于最大熵方法的中英文基本名词短语识别[J]. 计算机研究与发展,2003, 40(3):440 - 446.

[11] Hanna Wallach. Efficient training of conditional random fields[D]. Master's thesis, University of Edinburgh, 2002.

[12] Adwait Ratnaparkhi. (1998). Maximum Entropy Models for Natural Language Ambiguity Resolution[D]. Ph.D. Dissertation. University of Pennsylvania, 1998.

(上接第 47 页)

定的权系数用投票的方法集成了表 1 中的前四种方法,集成算法明显优于单一算法,p, r, f 分别为 0.53,0.53,0.53,这说明多种文本定位方法的集成是以后文本定位的一个研究方向。

另外从参赛算法我们也看到一些共同的特征:参赛者都用金字塔分解的方法解决字符大小不一的问题;参赛算法都表现出对图像的缩放敏感,同一种算法对同一幅图像在不同的缩放尺度下可能返回不同的结果;几乎所有算法都对光照敏感;算法的速度差异很大,比如我们算法的速度几乎是 Ashida 的 29 倍,Wolf 的 57 倍。

参 考 文 献:

[1] D. Maio, D Maltoni, R Cappelli, J Wayman, and A Jain. Fvc2000: Fingerprint verification competition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, (24): 402 - 412.

[2] Christian Wolf and Jean-Michel Jolion. Extraction and Recognition of Artificial Text in Multimedia Documents[R]. Technical Report RFV-RR-2002, 2002.

[3] N Ostu. A threshold selection method from grey-level histogram[J]. IEEE Transaction on System, Man and Cybernetics, 1979, 9(1): 62 - 66.

[4] R J Whatmough. Automatic threshold selection from a histogram using the exponential hull[J]. Graphical Models and Image Processing, 1991, (53):592 - 600.

[5] Simon M. Lucas, A Panaretos, L Sosa, A Tang, S Wong, R Young. ICDAR 2003 Robust Reading Competitions [A]. 7th International Conference on Document Analysis and Recognition (ICDAR 2003)[C],2003,(2):682 - 687.

[6] 陈又新,刘长松,丁晓青.复杂彩色文本图像中字符的提取[J].中文信息学报.2003,17(5):55 - 60.