基于连通域的汉字切分技术研究

陈 艳[1,2]，孙羽菲[1,2]，张玉志[1]

（1. 中国科学院 计算技术研究所，北京 100080；2. 中国科学院 研究生院，北京 100039）

摘 要：字符切分技术已经成为汉字识别系统设计中的关键问题，对于质量较差的文本图像，用灰度图像取代传统的二值化黑白图像能够取得更好的切分效果，基于连通域的切分算法能够对灰度图像进行较好的切分，基于连通域的汉字切分算法能有效地对文本图像中汉字字符部件进行合并及对粘连字符进行分割。

关键词：灰度图像；连通域；粘连字符切分；合并

中图法分类号：TP391.41 **文献标识码**：A **文章编号**：1001-3695(2005)06-0246-03

Chinese Character Segmentation Technique Based on Interconnected Domain

CHEN Yan[1,2], SUN Yu-fei[1,2], ZHANG Yu-zhi[1]

(1. *Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080, *China*; 2. *School of Graduate, Chinese Academy of Sciences, Beijing* 100039, *China*)

Abstract: Segmentation technique has become the bottleneck for the development of character recognition system. For low-quality document image, grayscale image is a better segmentation source than monochrome image. This paper presents a novel segmentation method, which is based on the concept of interconnected domain. Experiment results showed that this method can achieve better results than traditional methods.

Key words: Grayscale Image; Interconnected Domain; Character Segmentation; Merged

1 引言

字符识别技术经过几十年的发展，已经取得了长足的进步。目前某些识别系统对中等质量印刷体样本的识别率已达到了99%甚至更高。在这种情况下，字符切分已经成为字符识别中的关键问题。实践表明，文字识别系统的识别率与切分技术密切相关，错误切分直接导致错误识别。

目前的字符切分方法主要可分为三种[3]：①基于结构分析的切分，即从图像特征中寻找字符切分的规则；②以识别为基础的分割，该方法效果好，但比较耗时，实际应用较少；③整体切分策略，即系统将字符串作为一个整体进行词识别而不是字识别。这些切分方法大都以二值化之后的文本图像为处理对象。目前已投入使用的字符识别系统对印刷质量较好的文本图像能取得比较好的切分效果，但对于年代久远的报刊杂志的识别，图像本身质量较差，且文本图像并非只包含一种字体、一种字号，背景色时常变化，文字排版也非固定格式，使用二值化后的图像很难取得良好的切分效果。灰度图像比黑白图像记录了更多的图像信息，因而也能更准确地判断出切分的位置，因此利用灰度图像进行字符切分是提高字符识别系统识别率的有效途径，而这也对切分算法提出了新的要求。

2 分级连通域

利用灰度文本图像代替黑白图像能够提高字符的正确切分率，而采用分级连通域的概念则为灰度图像的切分打下了基础。为了既保证切分的效率，又保证不丢失有用的信息，需要对图像中的灰度值进行分级。灰度值分级通常采用直方图均衡法，这种分级不改变像素的灰度值，只对每个像素设置一个级别。根据灰度分级的结果把图像处理为分级的连通域，根据分级的连通域对图像进行粗切分，最后进行连通域之间的合并和粘连字符连通域的切分。

定义1 一幅图 $f(x,y)$（x，y 表示图像的长、宽）由若干个（设为 n 个）连通域 $C^{(k)}$（$k=1,\cdots,n$）组成，记为

$$f(x,y)=\bigcup_{\forall k\in(1,2,\cdots,n)}C^{(k)}$$

定义2 连通域 C 指由若干个（设为 m 个）像素 $a^{(k)}$（$k=1,\cdots,m$）组成的集合，且满足：

对 $\forall a^{(i)}a^{(j)}(a^{(i)}\in C,a^{(j)}\in C,i\neq j)$，它们之间存在一条通路 L，且 $\forall q(q\in L)$ 有 $q\in C$。

定义3 连通域 C 具有一定的级别，记为 $grade(C)$，其级别由它所包含的像素的最高灰度级确定如下：

$grade(C)=\max(grade(a^{(1)}),grade(a^{(2)}),\cdots,grade(a^{(m)}))$，其中 $grade(a^{(k)})$ 为像素 $a^{(k)}$ 的灰度级。

按照以上定义生成的连通域，其内部可以包含若干不同级别灰度的像素，这些像素彼此之间都相互连通。可以看到，某个连通域的子集仍然可能是一个连通域，引入相邻连通域的概念如下。

定义4 设 A,B 为连通域，若 A,B 满足以下条件则为相邻连通域，若

(1) $grade(A)=grade(B)$

(2) 存在两像素点 a,b，满足下列条件：

收稿日期：2004-03-28；**修返日期**：2004-07-28

①$a \in A, b \in B$;②a, b 两点"8 连通"。

定义5　对于灰度文本图像,若其灰度级分为 $G_1 \sim G_n$,可以找到一个灰度级 $G_p(1 \leq p \leq n)$,如果将所有低于 G_p 的灰度级看作是相同的,则大部分的 G_p 级连通域的外接矩形都能正确表示单个字符的字符域。这样的灰度级别称之为主灰度级别(如图1中的主灰度级别为 G_2)。

采用适当的方法确定了文本图像的主灰度级别之后,就能对图像中的字符进行切分。在对文本图像灰度分级后,生成连通域时需要合并相邻连通域。在本文中,如未加说明,所述连通域都是指不再具有相邻连通域,即不能再参与进一步合并的连通域。以图1为例说明连通域的生成。当通过直方图均衡把图像中的像素划分为 $G_1 \sim G_n$ 个灰度级别之后,相应地,文本图像中的连通域也分为 $G_1 \sim G_n$ 个级别。该文本图像分为四个灰度级别,分别用黑、深灰、浅灰和白色表示。黑色像素表示 G_1 级别的像素点,深灰的像素表示 G_2 级别的像素点,浅灰色像素表示 G_3 级别的像素点。相对应地,在图1中可以看到 $G_1 \sim G_3$ 作为主灰度级别时的连通域(G_4 级连通域即为整幅图像;为便于观察,G_3 级连通域的边界用深色线条标出)。

(a)分级后文本图像 (b)G_1 作为主灰度级 (c)G_2 作为主灰度级 (d)G_3 作为主灰度级

图1　文本图像中的连通域

3　汉字字形特点的连通域表示

从构成上讲,汉字是由笔画(点横竖撇捺等)、偏旁部首构成的。文献[1]认为,汉字由部件构成,组成汉字的部件并不一定具有音和义。从存在形式上看,它是一个独立的书写单位,不管笔画多么复杂,凡是笔画连在一起的,都作为一个部件看待。分析汉字中部件与部件的关系,可得到如图2所示的九种情况。

(a)　(b)　(c)　(d)　(e)　(f)　(g)　(h)　(i)

图2　汉字中部件的关系

图2(a)表示上下结构的汉字,如吕、昌等;图2(b)表示左右结构的汉字,如朋、明等;图2(c)为包含结构的汉字,如回等;图2(d)~图2(i)都是半包围结构的汉字,如庆、犬、句、太、区、凶等。对于灰度文本图像,汉字的部件是由一个或多个连通域组成的。在图像质量较差时,如果出现断笔的情况,一个部件可能被分割成多个连通域。而在字符之间出现粘连时,一个连通域又可能包含多个部件。

以连通域表示汉字的字形特点,可将连通域的外接矩形作为连通域的特征,汉字部件的上述九种关系对于连通域来说成为三种情况:上下关系、左右关系和重叠关系(包括包含关系)。这样就能大大简化切分时对汉字字形的判断分析过程。

图3是两个连通域 $C^{(i)}, C^{(j)}$ 存在重叠关系。对它们的外接矩形区域进行分析,(L_i, U_i),(R_i, D_i) 分别记录了 $C^{(i)}$ 连通域的外接矩形的左上角和右下角的坐标。定义以下标记:

连通域的宽 $Width(C^{(i)}) = R_i - L_i$;

连通域的高 $Height(C^{(i)}) = D_i - U_i$;

连通域的宽高比 $Ratio(C^{(i)}) = \dfrac{Width(C^{(i)})}{Height(C^{(i)})}$;

连通域的面积 $Square(C^{(i)}) = Width(C^{(i)}) \times Height(C^{(i)})$;

合并后的宽度 $Uwidth(C^{(i)}, C^{(j)}) = \max(R_i, R_j) - \min(L_i, L_j)$;

合并后的高度 $Uheight(C^{(i)}, C^{(j)}) = \max(D_i, D_j) - \min(U_i, U_j)$;

合并后的宽高比 $URatio(C^{(i)}, C^{(j)}) = Uwidth(C^{(i)}, C^{(j)}) / Uheight(C^{(i)}, C^{(j)})$;

重叠的宽度 $Owidth(C^{(i)}, C^{(j)}) = Width(C^{(i)}) + Width(C^{(j)}) - Uwidth(C^{(i)}, C^{(j)})$;

重叠的高度 $Oheight(C^{(i)}, C^{(j)}) = Height(C^{(i)}) + Height(C^{(j)}) - Uheight(C^{(i)}, C^{(j)})$;

重叠的面积 $Osquare(C^{(i)}, C^{(j)}) = Owidth(C^{(i)}, C^{(j)}) \times Oheight(C^{(i)}, C^{(j)})$。

以上所有的长、宽以及面积的单位都用像素点个数表示。通过对以上的定义,就可以对连通域的重叠关系进行定量化分析,从而得到正确的切分策略。

4　部件的组合与粘连字符/背景的分割

假设已对文本图像进行了灰度值分级,且已找到了图像的主灰度中 G_p,图像中级别为 G_p 的连通域有 M 个,表示为 $C^{(k)}$ $(k = 1, \cdots, M)$。此时这 M 个连通域中,可能有部分连通域所表达的并非完整字符,而是字符部件。因此需要对切分出来的字符部件进行组合。基于连通域的部件组合算法步骤如下:

(1)初始化 flag = 0;

(2)计算平均字宽、字高:$Awidth, Aheight$;

(3)对所有的 $C^{(i)}, C^{(j)}$ $(i \neq j$,且 $0 \leq i, j \leq M)$,

①if($\rho \leq Owidth(C^{(i)}, C^{(j)}) \leq 0$) or($\rho \leq Oheight(C^{(i)}, C^{(j)}) \leq 0$)
//此时两连通域不重叠,但相距较近

then if ($URatio(C^{(i)}, C^{(j)}) \approx 1$) and ($UWide(C^{(i)}, C^{(j)}) \approx AWide$)
合并 $C^{(i)}, C^{(j)}$,flag 置为1　//此时认为两连通域属于同一个字符

else 两连通域属于不同的字符域

②if($OWide(C^{(i)}, C^{(j)}) \leq \rho$) or($OHigh(C^{(i)}, C^{(j)}) \leq \rho$)
//两连通域不重叠且相距较远

then 两连通域属于不同字符域

③if($0 \leq OWide(C^{(i)}, C^{(j)})$) and($0 \leq OHigh(C^{(i)}, C^{(j)})$)
//两连通域重叠;

then if $OSquare(C^{(i)}, C^{(j)}) \geq \lambda \times \min(Square(C^{(i)}), Square(C^{(j)}))$
合并 $C^{(i)}, C^{(j)}$,flag 置为1　//此时认为两连通域属于同一个字符

(4)如果 flag 为1,则返回(1),否则退出。

以上算法中,ρ 与 λ 为常数因子,ρ 为小于0的负整数,λ 的值介于0与1之间。这两个常数因子需要由实验方法获得。

由于文本图像的过渡空间的灰度的渐进性,图像背景的复杂性会造成多个字符在一个连通域中,这时连通域外接矩形包含的则是多个字符组成的字符域。另外的情况是连通域中不仅包括了字符域也包括了背景域。这时就需要对连通域进行再次分割。如果连通域 $C^{(k)}$ 满足下列条件之一,则认为此连通域需要进一步的切分:

①$Ratio(C^{(k)}) >> 1$

②$Square(C^{(k)}) \approx Num(C^{(k)})$

//$Num(C^{(k)})$ 表示该连通域含像素点的个数

若满足条件①则对连通域采用改进的滴水算法对其进行切割,改进的滴水算法见文献[2]。切割的规则如图4所示。

若满足条件②则考虑返回上一级的连通域进行切分，即将该连通域中级别小于等于 G_{p-1} 的连通域作为可切分的单字字符域。

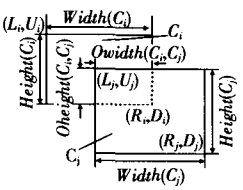

图 3 连通域外接矩形的重叠关系

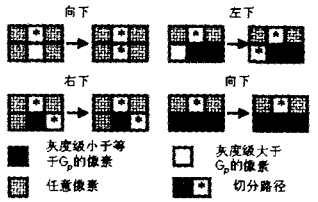

图 4 改进的滴水算法

5 实验结果

为验证基于连通域的切分算法的可行性及正确率，选取某报一个月的扫描灰度图像作为实验样本。在混合文本及复杂背景的情况下，粘连字符的切分正确率达到 95% 以上。部件合并的正确性达到 92% 以上。图 5 为基于连通域的切分算法和常规切分算法效果的对比图，深色线条包围的区域表示字符域。图 5 中的汉字包含了上下、左右、包含、重合等多种情况。可以看到，基于连通域的切分算法处理这些字符都取得了很好的效果，并且在中英文混排的情况下也不影响部件的正确合并。

(a)合并前的效果 (b)合并后的效果

图 5 连通域合并前与合并后的效果对比

图 6 中文本图像的连通域满足第 4 节中的再切分条件(1)，即连通域的宽高比远大于 1，因此需要进一步切割的实例。在此例中应用改进的滴水算法对连通域进行处理，得到了较好的切分效果。

图 6 对宽高比较大的连通域采用滴水算法切分

图 7 中的文本图像背景色不均匀，在图像分级处理后，得到级别为 G_p 的连通域的外接矩形，如图 7 中初次切分结果所示。可以看到，其中“试用手记”四个字符及其背景域属于同一个连通域，被认为构成同一个字符。但此时该连通域符合第 4 节中的再切分条件(2)，判断连通域需要进行进一步的划分，则返回到 G_{p-1} 级连通域把背景域与字符域区别开，则能正确地切分出字符。

(a)文本图像 (b)初次切分 (c) 采用新的主灰度级别再次切分

图 7 对于背景色变化的文本图像进行切分

6 结论

灰度图像比黑白图像更适用于低质量文本图像的切分。通过灰度值分级可以在文本图像中构造分级连通域，以分级连通域的概念为基础设计的切分算法，可以有效地对字符的部件进行组合并对粘连字符进行分割。实验结果表明该切分方法收到了比传统切分方法更好的切分效果。如何准确确定灰度图像中的主灰度级别以及滴水算法中起始点都是未来需要进一步研究的问题。

参考文献：

[1] 傅永和. 汉字结构和构造成分的研究[A]. 现代汉语用字信息分析[M]. 上海：上海教育出版社，1993.

[2] Punnoose J. An Improved Segmentation Module for Identification of Handwritten Numerals[D]. Boston: Department of Electrical Engineering and Computer Science, M. I. T, 1999.

[3] Casey R G, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(19): 690-709.

[4] Lu Y. Machine Printed Character Segmentation 2 an Overview[J]. Pattern Recognition, 1995, 28(1): 67-80.

[5] 韩布新. 部件组合——潜在的汉字结构层次[J]. 中文信息学报，1995, 9(3): 27-32.

作者简介：

陈艳(1980-)，女，硕士研究生，主要研究方向为字符识别、切分技术；孙羽菲(1976-)，女，博士研究生，主要研究方向为模式识别、字符识别、图像处理；张玉志(1964-)，男，博士生副导师，主要研究方向为人工智能、模式识别。

(上接第 237 页)

[6] Joshua B Tenenbaum, Vin de Silva, John C Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction [J]. Science, 2000, 290(5500): 2319-2323.

[7] Sam T Roweis, *et al*. Nonlinear Dimensionality Reduction by Locally Linear Embedding[J]. Science, 2000, 290(5500), 2323-2326.

[8] 彭辉，张长水. 基于 K-L 变换的人脸自动识别方法[J]. 清华大学学报(自然科学版)，1997, 37(3): 67-70.

[9] Vytautas PerlibaKas. Distance Measures for PCA-based Face Recognition[J]. Pattern Recognition Letters, 2004, 25(6): 711-724.

[10] Guoqiang Peter Zhang. Neural Networks for Classification: A Survey [J]. IEEE Transaction on System, Man, and Cybernetics: Application and Reviews, 2000, 30(4): 451-462.

[11] Sheng Ma, Chuanyi Ji. Performance and Efficiency: Recent Advances in Supervised Learning[J]. Proceeding of the IEEE, 1999, 87(9): 1519-1535.

[12] Richard O Duda, Perter E Hart, David G Stork. Pattern Classification(2 Edition)[M]. John Wiley, 2001.

作者简介：

黎奎，硕士研究生，研究方向为图像处理、人脸识别；宋宇，硕士研究生，研究方向为图像处理、人脸检测；邓建奇，硕士研究生，研究方向为图像处理；刘民，硕士研究生，研究方向为图像处理；陈忠林，博士生，研究方向为图像处理、人脸检测；周激流，教授，博士研究生导师，博士，研究方向为图像处理、模式识别和人工智能。