

粘连字符切分综述^{*}

魏湘辉¹ 马少平²

(中国科学院软件所 北京100080)¹

(清华大学计算机科学与技术系 智能技术与系统国家重点实验室 北京100084)²

摘要 字符切分是字符识别系统的一个重要环节。粘连则是指字符图像间有笔画接触的情况,它普遍存在于手写和低质量的印刷文档中。然而普通切分算法很难处理粘连字符。本文综述了主要的粘连字符切分算法以及切分路径的评价方法,在详细介绍算法流程的基础上,对每种算法的优缺点以及粘连字符切分的未来工作进行了探讨。

关键词 字符切分,粘连字符切分,字符识别

A Survey of Touching Character Segmentation Algorithms

WEI Xiang-Hui¹ MA Shao-Ping²

(Institute of Software, CAS Beijing 100080)¹

(Dept. of Computer Science, Tsinghua University State Key Laboratory of Intelligent Technology and Systems, Beijing 100084)²

Abstract Character segmentation is a key step of a character recognition system. In this step it is most difficult to separate connected characters or touching characters, which occur frequently in handwritten documents and printed ones with low quality. This paper focuses on segmentation of touching characters. It surveys 5 approaches that construct segmentation paths or candidates and 3 evaluation functions that are used to select the best path. Based on the discussion of the advantage and disadvantage of each approach, some promising directions are proposed.

Keywords Character segmentation, Touching character segmentation, Character recognition

1 前言

字符识别通常要经过版面分析和光学字符识别(OCR)两个步骤。版面分析输出的文字图像一般包含多个字符,而现有的OCR系统的识别对象多是单个字符,因此,从多字符图像中切分出单字就成为衔接版面分析和OCR系统的一个关键步骤。切分算法的目标就是从多字符图像中切出一系列子图像,其中每个子图像都包含一个独立完整的字符。图1显示了典型的字符识别流程。

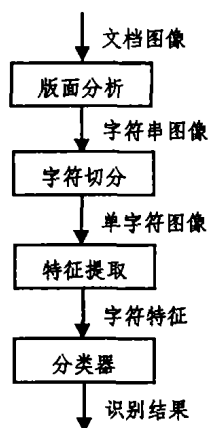


图1 字符识别流程图

在多字符图像中,粘连是指字符间笔画相互接触的情况。在手写体中,字符粘连主要是由于手写风格造成的;在印刷体

中,则多是由印刷质量引起。对于粘连字符,普通的切分算法,如基于联通单元的算法、基于投影分析的算法等,很难取得满意的切分效果,必须使用有针对性的切分算法。本文从大量文献中归纳出五种粘连字符切分算法和三种切分路径的评价方法,并在描述算法流程的基础上,讨论了每种算法的优缺点以及粘连字符切分的未来发展方向。

本文的第2节介绍五种粘连字符切分算法,第3节介绍三种切分路径的评价方法,第4节讨论了现有算法中的一些问题以及以后的研究方向。

2 切分算法

2.1 基于结构特征的算法

利用字符结构上的特征点来进行切分的一类算法被称为基于结构特征的算法。所谓特征点是指字符轮廓上斜率变化剧烈的点和局部极值点。

在文[1]中,特征点被具体化为重要轮廓点(Significant contour points, SCP),包含了拐角点、局部极小点、局部极大点和延伸凹角处的直笔画得到的潜在路径退出点。算法从这些特征点中选择一对作为切分路径的进入点和退出点,选择时考虑的因素有:

- 局部极小-极大点对优先
- 拐角点优先
- 点间的距离
- 轮廓变化的剧烈程度
- 极小极大点对间的水平距离

^{*} 本课题受到自然科学基金项目(No. 60223004)和国家863高科技项目(No. 2001AA114082)资助。魏湘辉 硕士,主要研究方向为字符切分。马少平 教授,博士生导师,主要研究方向为汉字识别,信息检索。

- 一个极值点到另一个极值点的距离
- 离图像左边远的距离

文[2]中使用的 Min-Max 算法则主要考虑极小极大点对。该算法将粘连字符分为上下轮廓,再找出上轮廓的极小点和下轮廓的极大点,最后选择其中一对用直线连接作为切分路径。选择时主要考虑的因素有两点间的距离、到图像中心的距离以及沿切分路径颜色变化的次数。

2.2 滴水算法

论文[3]中的滴水算法通过模拟水滴在字符粘连处滚落来构造切分路径。它从当前位置出发,根据一系列规则(如图2所示),在相邻位置搜索路径的下一个点。滴水算法中的一个关键问题是如何选择水滴滚落的起始位置。文[3]通过从上至下,从左至右扫描图像,选择同时满足以下两个条件的点作为起始点:

- 左邻接点为黑像素点
- 右边存在黑像素点

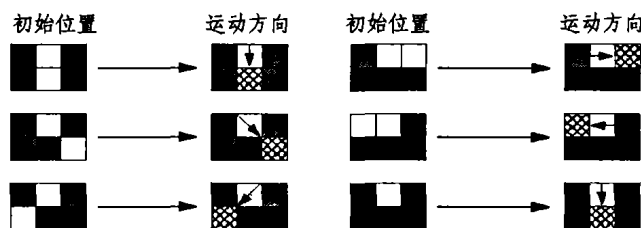


图2 滴水算法的搜索规则

2.3 基于细化的算法

文[5]使用了一种基于背景细化的切分算法。作者观察到细化后的背景能够提供构造切分路径的有用信息,即背景上的特征点往往指向粘连位置或是可以成为切分路径的起始点。该算法分为如下几个步骤:

- 细化背景;
- 在细化后的背景上提取特征点,包括叉点、端点和角点;
- 构造候选切分路径,即通过搜索来连接背景上的一些特征点,从而构成从图像的顶部延伸到底部的所有路径;
- 对候选路径进行评价,选择最佳切分路径。

文[6]进一步改进了上述算法,同时使用了前景和背景细化后提供的信息,并且考虑到字符多处粘连的情况,在切分路径构造算法中给出了相应的处理。

2.4 储水区算法

文[4]使用了储水区(Water Reservoir)的概念来切分粘连字符。储水区被定义为从粘连字符上方和下方倒入水而形成的储水区域,并根据其位置分为:上储水区、中储水区和底储水区。储水区算法可以根据储水区域的大小和位置确定字符的粘连位置,并通过分析储水区的边界、粘连位置和粘连字符的形态特征来确定最佳切分点,结合结构形态特征构造出切分路径。

2.5 基元合并算法

基元合并算法首先将字符图像分解为基元,然后根据合并策略逐步合并基元,最终得到切分结果。

文[7]使用笔划作为字符的基元,笔划的外接矩形框被定义为笔划框。对笔划框进行合并的过程分为两个阶段:第一个阶段是基于知识的合并;第二个阶段则定义了笔划框的不相似度函数,然后采用动态规划得出使总体不相似度最小的合

并方案,所得结果即为最终切分结果。

由于使用了基元而不是单纯的连通区域,基于自底向上策略的切分算法在一定程度上能够解决粘连字符的切分问题。该算法存在的问题是合并策略并不能完全保证基元合并到正确的字符中去,因此错误的切分也难以避免。

3 切分路径评价方法

由于字符粘连情况的复杂性,一次性得出最佳切分路径极为困难。因此构造多条候选路径,然后从中选择出最佳路径,成为多数算法的选择。对于候选路径的评价可以分为基于识别和基于特征这两类方法。

基于识别的方法将候选路径切分的子图像送入分类器,以分类器的信任值作为评价标准。基于识别的方法依赖于分类器的稳健性,且效率不高。

为了进一步提高评价的正确率,很多系统使用基于特征的方法。这类方法首先提取与路径相关的特征向量,再根据相应的评价手段对特征向量进行排序,从而得出最佳路径。基于特征的方法可以单独使用,也可以与基于识别的方法联合使用,进一步提高切分正确率。以下介绍了三种基于特征的切分路径评价方法。

3.1 决策树

文[5]使用的是基于决策树的评价方法。首先提取了以下特征:

- F_1 : 切分后两部分的宽度比
- F_2 : 切分后两部分的高度比
- F_3 : 切分后左半部分的宽高比
- F_4 : 切分后右半部分的宽高比
- F_5 : 两部分水平重叠部分的与切分后两部分最小宽度的比值
- F_6 : 路径上黑像素数目与图像宽度的比值
- F_7 : 路径中心与图像中心坐标的比值
- F_8 : 切分后两部分黑像素比值

于是,每条候选切分路径对应一个8维的特征向量。对训练集中的每幅图像,人工指定最佳路径,并使用决策树学习,得出评价路径好坏的规则。在定义规则相似度的基础上,该方法以候选路径与规则的相似度作为最终的评价标准。

3.2 混合高斯概率函数

文[6]使用混合高斯概率函数,使用的特征与文[5]相同。在训练集上用 MKM(Modified K-means)算法对特征向量进行聚类,并根据以下公式计算特征向量 v_n 的混合高斯概率值。

$$P(v_n) = \sum_{m=1}^M \frac{c_m}{(2\pi)^{D/2} \sqrt{\prod_{d=1}^D \sigma_{m,d}^2}} \exp\left(-\sum_{d=1}^D \frac{(v_{n,d} - \mu_{m,d})^2}{2\sigma_{m,d}^2}\right)$$

其中:

- D 是特征向量的维数,文[6]中 $D=8$
- $v_{n,d}$ 是特征向量 v_n 的第 d 维
- $\mu_{m,d}$ 是第 m 类均值向量的第 d 维
- $\sigma_{m,d}$ 是第 m 类标准差向量的第 d 维

在测试集上,以混合高斯概率值最大的候选路径为最优路径。

3.3 贝叶斯置信网络

文[10]对训练集上的特征向量用贝叶斯置信网络来学习。贝叶斯置信网络对特征向量 $(X_1, X_2, \dots, X_n)$ 的联合概率可以给出以下的计算公式:

$$P_B(X_1, X_2, \dots, X_N) = \prod_{i=1}^N P(x_i | \text{parents}(X_i))$$

用这个联合概率值对所有候选路径进行排序, 概率值最大的为最优路径。

4 分析和展望

本文综述了粘连字符的切分算法以及切分路径评价方法。

在现有的切分算法中, 基于细化的算法具有较好的适应性, 而且可以处理复杂的粘连情况, 如接触面倾斜、字符间相互重叠等, 对于复杂的汉字粘连, 也能进行有效的处理^[10]。

在现有的切分评价方法中, 基于识别的方法依赖于具体的识别器, 识别器的错误率决定了评价的错误率。基于特征的方法是相对独立的一类方法, 能获得较高的正确率和效率。对复杂字符(如汉字)的切分, 如能结合评价、识别和上下文关系, 将进一步减少切分错误。

参考文献

- 1 Strathy N W, Suen C Y, Kryzak A. Segmentation of handwritten digits using contour features. In: Second Int. Conf. Document Anal. Recognition, Oct. 1993. 577~580
- 2 Dimauro G, Impedovo S, Pirlo G, Salzo A. Automatic Bankcheck processing: A New Engineered System. International Journal of Pattern Recognition and Artificial Intelligence, 1997, 11(4): 467~504
- 3 Congedo G, Dimauro G, Impedovo S, Pirlo G. Segmentation of Numeric strings. In: Proc. of Third Int. Conf. on Document Analysis and Recognition, Montreal, Aug. 1995. 1038~1041

(上接第195页)

发现的效率。分类中的每个注册节点之间是对等、自适应、自管理的, 没有严格的互相依赖关系。实验结果表明, 本模型具有良好的可扩展性。

参考文献

- 1 Schmidt C, Parashar M. A Peer-to-Peer Approach to Web Service Discovery, submitted for publication, February 2003. <http://citeseer.nj.nec.com/schmidt03peertopeer.html>
- 2 Schlosser M, Sintek M, Decker S, Nejdl W. A Scalable and Ontology-Based P2P Infrastructure for Semantic Web Services. Second Intl. Conf. on Peer-to-Peer Computing (P2P'02), Linköping, Sweden, Sept. 2002
- 3 Hoschek W. Peer-to-Peer Grid Databases for Web Service Discovery. Grid Computing: Making the Global Infrastructure a Reality. Nov. 2002
- 4 Verma K, et al. METEOR-S WSDI: A Scalable Infrastructure of Registries for Semantic Publication and Discovery of Web Services. <http://lsdis.cs.uga.edu/lib/download/MWSDI.pdf>, 2002
- 5 Stoica, Morris R, Karger D, Kaashoek M F, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for internet applications. Submission to ACM SIGCOMM, 2001
- 6 Druschel P, ARowstron A. Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems. Submission

- 4 Pal U, Belad A, Choisy C. Touching numeral segmentation using water reservoir concept. Pattern Recognition Letters, 2003, 24: 261~272
- 5 Lu Zhongkang, et al. Zheru Chi, Wan-Chi Siu, Pengfei Shi. A background-thinning-based approach for separating and recognizing connected handwritten digit strings. Pattern Recognition, 1999, 132: 921~933
- 6 Chen Y-K, Wang J-F. Segmentation of Single- or Multiple-Touching Handwritten Numeral String Using Background and Foreground Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(11)
- 7 Tseng L Y, Chen R C. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. Pattern Recognition Letters, 1998, 19: 963~973
- 8 Lu Y, Shridhar M. Character Segmentation in Handwritten Words-An Overview. Pattern Recognition, 1996, 29(1): 77~96
- 9 Casey R G, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(7)
- 10 Maragoudakis M, Kavallieratou E, Fakotakis N, Kokkinakis G. How Conditional Independence Assumption Affects Handwritten Character Segmentation. ICDAR 2001
- 11 Lu Y, Shridhar M. Character Segmentation in Handwritten Words-An Overview. Pattern Recognition, 1996, 29(1): 77~96
- 12 Casey R G, Lecolinet E. A Survey of Methods and Strategies in Character Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18(7)

to ACM SIGCOMM, 2001

- 7 Zhao B Y, Kubiawicz J, Joseph A D. Tapestry: An Infrastructure for Fault-tolerant Widearea Location and Routing. Technical Report No. UCB/CSD-01-1141, Computer Science Division, University of California at Berkeley, April 2001
- 8 Sivashanmugam K, Verma K, Mulye R, Zhong Z, Sheth A. Speed-R: Semantic P2P Environment for diverse Web Service Registries
- 9 Universal Description, Discovery and Integration: UDDI Technical White paper, September 2000
- 10 Gnutella website. www.gnutella.com
- 11 Ratnasamy S, Francis P, Handley M, Karp R, Shenker S. A Scalable Content-Addressable Network. In: Proc. of ACM SIGCOMM, 2001
- 12 Nejdl W, et al. Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks. In: Proc. of the Intl. World Wide Web Conf. (Budapest, Hungary, May 2003)
- 13 Gandon F. Ontology Engineering: A Survey and a Return on Experience, Institut de Recherche en Informatique et Automatique (INRIA), March 2002, ISSN 0249-6399. <ftp://ftp-sop.inria.fr/acacia/fgandon/FabienGandon-RR4396.pdf>
- 14 El-Ansary S, Alima L O, Brand P, Haridi S. Efficient Broadcast in Structured P2P Networks. In: 2nd Intl. Workshop on Peer-to-Peer Systems (IPTPS '03)