ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.3, March 2006, pp.403−409
DOI: 10.1360/jos170403

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

MST 在手写汉字切分中的应用*

韩 勇[1+], 须 德[1], 戴国忠[2]

[1](北京交通大学 计算机与信息技术学院,北京 100044)

[2](中国科学院 软件研究所 人机交互技术与智能信息处理实验室,北京 100080)

Using MST in Handwritten Chinese Characters Segmentation

HAN Yong[1+], XU De[1], DAI Guo-Zhong[2]

[1](School of Computer & Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2](Laboratory of Human-Computer Interaction and Intelligent Information Processing, Institute of Software, The Chinese Academy of Sciences, Beijing 100080, China)

+ Corresponding author: Phn: +86-10-62645407, E-mail: hanyong_sc@hotmail.com, http://www.njtu.edu.cn

Han Y, Xu D, Dai GZ. Using MST in handwritten Chinese characters segmentation. *Journal of Software*, 2006,17(3):403–409. http://www.jos.org.cn/1000-9825/17/403.htm

Abstract: Handwritten Chinese characters segmentation is to process strokes based on its spatial relations to form character elements for recognition. This paper introduces a method to segment Chinese characters according to the topological relations of Chinese component and minimal span tree of strokes. The experiment shows that this new method can achieve good performance. The accuracy of segmentation is over 91.6%.

Key words: character segmentation; handwritten Chinese character; component structure; online recognition

摘 要: 手写汉字切分是根据输入笔迹的空间位置关系进行汉字部件的合并切分,形成完整的汉字笔划以便进行识别处理.综合利用了汉字部件的结构位置关系和笔划的空间位置关系,根据笔划的最小生成树(minimal spanning tree,简称 MST)对联机连续手写输入汉字进行切分,取得了较好的切分结果.切分的准确率超过 91.6%.

关键词: 字符切分;手写汉字;部件结构;联机识别

中图法分类号: TP391 文献标识码: A

自从 1966 年 IBM 公司的 Casey 和 Nagy 发表了关于汉字识别的文章以来[1],汉字识别技术经过几十年的发展,已经取得了长足的进展.目前,市场上已经有了比较成熟的联机手写汉字识别和离线印刷体汉字识别的商业产品.但是在联机手写汉字识别中,文字是单个输入的,下一个文字的输入在上一个输入文字的识别完成之后才能进行,这不符合人们日常生活中使用纸笔连续输入的交互方式,同时又降低了输入的效率.解决这个问题,需要正确切分连续手写汉字.由于每个人的输入习惯不同,书写的文字不可能像印刷体文字那么工整、规范,所以需要根据笔划的结构位置信息作切分.而对于手写体的正确切分是正确识别的前提[2].因此,汉字切分的研究

* Supported by the National Natural Science Foundation of China under Grant Nos.60033020, 60373030 (国家自然科学基金); the National Grand Fundamental Research 973 Program of China under Grant No.2002CB312103 (国家重点基础研究发展规划(973)); the Foundation of Beijing Jiaotong University under Grant No.2004SM013 (北京交通大学基金)

Received 2004-05-11; Accepted 2005-01-06

对于手写汉字输入系统的推广应用具有重要意义.

脱机手写汉字识别系统识别的对象是二维图像点阵,而联机手写汉字识别系统识别的对象是表示成一系列坐标点的手写汉字,这些坐标点是对书写时笔尖运动的轨迹进行时域采样得到的.因此,联机识别更加容易从中获得对于汉字识别十分重要的结构信息.字符的切分有 4 种主要的方法:一是基于结构分析的切分,即把复杂的结构分解为简单的子模式[1−4];第 2 种方法避免切分,利用预先定义的框格来进行切分;第 3 种采用了先切分再按规则组合的切分方法[4−7];第 4 种方法是整体切分策略,即系统将字符串作为一个整体进行词识别而不是单字识别[5−7].近年来,中西文系统在印刷体的识别方面取得了较大的进展,但是由于手写体的随意性和不规范性,在印刷体汉字识别中采用的基于投影的切分方法在手写体切分中效果不是很好.

西文的手写体切分采用了雨滴法(DF)[7]、BP 神经网络[8]以及连通性规则,将连笔输入的单词切分为单个的字母.但是中文字符的切分与英文字母、数字的切分不同:中文字符的笔划多、结构复杂、图像可能包含多个连通元,而且中文字符可由多个部件构成(有的部件自身也是中文字符).因此,中文字符切分除了区分相邻汉字的笔划以外,还要将构成单个汉字的多个笔划合并到一起.在中文字符切分中使用雨滴法(DF)的难点在于,如何准确找到雨滴的下落起始点,而且对于左右结构的汉字容易产生“过切分”.所以,在西文中使用的切分方法不一定适合中文字符的切分.但是,中文的部件之间的位置关系仅有有限的几种[4],而且字与字之间的间距与汉字内部部件之间的间距有明显差异[9].因此,我们可以尝试根据部件的位置关系规则以及中文近似方块字的字体特征进行部件合并.另外,可以注意到,汉字的笔划部件都集中在字符中心位置附近,因此,我们考虑利用最小生成树建立起笔划之间的关系[10].文献[4]利用汉字的部件结构知识以及部件间的距离进行切分,该方法只是考虑了部件与临近部件的关系,没有考虑全局部件之间的位置层次关系.文献[9]先采用若干字间距阈值进行连续手写中文分割,获得多个分割结果,然后根据字间距方差从中选取最佳两组结果,在不提高字间距方差的前提下,合并邻近的候选单字,分裂较宽的候选单字,最后利用识别结果提取单字[11].该方法没有利用到汉字的部件结构等知识.本文的方法不仅利用了笔划的空间信息和笔划间的位置关系,而且还利用了最小生成树,考虑了全体待切分笔划间的位置关系,从全局的观点对笔划进行切分.实验结果表明,利用最小生成树建立笔划的联系以后,相邻汉字之间相连的笔段只有一条,在部件合并时,沿着最小生成树进行,可以使部件合并工作得以简化.该方法综合利用了汉字的结构特点以及各部件的空间位置、距离等信息,取得了较好的切分结果.实验表明,切分的准确率超过 91.6%.

1 算法描述

1.1 算法思路

笔划是指在书写文字时,从落笔到抬笔之间笔尖所描绘的轨迹[1].笔划是组成汉字的基本单位.在本文中,认为笔划是部件的基本组成单位,笔划的外接矩形是部件的位置和形状信息,部件与部件的组合构成新的部件或构成最终的文字.两个部件之间有如图 1 所示的关系[3,4].

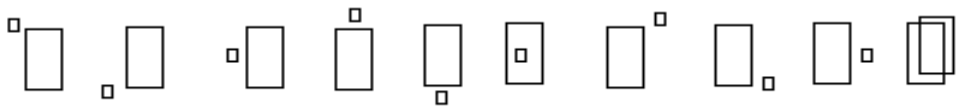

Fig.1 Spatial relations of components

图 1 部件间的位置关系

在考虑单行文字切分时,首先以单个笔划为部件开始合并工作,从最左边的部件开始生成所有笔划中点距离的最小生成树.如图 2 所示,每个汉字与下一个字之间只有一条树枝相连.因此,在进行文字切分时可以减少笔划位置关系的比较判断数目.只需判断与当前部件最接近的笔划间的位置关系.

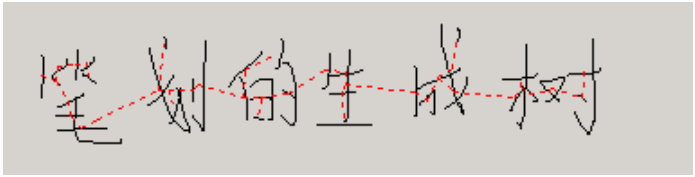

Fig.2 The MST composing by the central points of handwriting strokes (the dashed is the branches of MST)

图 2 汉字笔划中点的最小生成树(虚线为笔划的最小生成树的树枝)

由于输入的一行汉字笔划数目较多,所以计算整个笔划的最小生成树、存储和计算的代价较大.根据汉字的笔划数目特点,我们只是在当前笔划临近的 N 个笔划中寻找最近的笔划,提高了切分的速度和效率,同时不影响最终的结果.

在进行部件合并时,如果笔划与部件处于上下、包含、交叉或位于当前部件的左边时,则将笔划合并到当前部件中,同时更新当前部件的外接矩形.与此同时,判断该笔划最小生成树中临近结点与当前部件的位置关系,进行相应的操作.

1.2 连笔与倒笔输入的处理

在汉字输入中,通过对用户的输入习惯研究发现,单个汉字内出现笔划连笔输入的情况较多,两个汉字之间的笔划连笔输入较少,而且同时考虑这两种连笔输入情况的处理比较复杂,所以我们目前没有研究汉字间的连笔输入.单个字内部的连笔输入,可以认为连笔笔划是按照用户意图,将连笔笔划中所包括的几个单笔笔划合并后,作为一个整体部件进行处理.因此,可将连笔笔划视同其他单笔笔划一样进行处理,根据笔划与部件之间的位置关系进行部件合并.对于倒笔输入,不论字内倒笔还是字间倒笔,切分都是先对一行内所有笔划按中点 X 坐标排序后再进行切分.因此,输入的先后顺序不影响切分结果,切分只是利用了笔划的空间信息,而没有利用时序信息.

1.3 切分算法的形式化描述

- 定义笔划部件间的位置关系:

$$P::=p_{LU}|p_{LD}|p_L|p_U|p_D|p_C|p_{RU}|p_{RD}|p_R|p_I.$$

其中,下标的含义分别为:LU(左上),LD(左下),L(左),U(上),D(下),C(包含),RU(右上),RD(右下),R(右),I(交叉).

- 部件定义

部件可以由单个笔划构成,部件与部件(笔划)合并的结果仍然是部件.定义 *Component* 表示部件,*Stroke* 表示笔划,则有如下描述:

$$Component::=(Component+Compontnt)|(Compontnt+Stroke)|Stroke.$$

部件的外接矩形是部件的几何形状和空间信息.设定部件的外接矩形的左上角坐标为($left$,top),右下角坐标为($right$,$bottom$),则可以定义笔划(部件)的中心点坐标为 $\left(\frac{left+right}{2},\frac{top+bottom}{2}\right)$.

- 定义符号

Merge(w,s)表示将部件 s 合并到部件 w 中.

MX(s)表示笔划 S 中点的 x 坐标.

M(w)表示部件 w 的中点.

E(p_1,p_2)表示以 p_1,p_2 为顶点建立一条直接通路.

Distance(p_1,p_2)表示点 p_1,p_2 之间的距离.

MakeMST(S)表示生成笔划集合 S 的最小生成树.

CON(*s*)表示最小生成树中与笔划 *S* 有边相连的其他笔划.

CountCON(*s*)表示最小生成树中与笔划 *s* 相连的边的个数.

RemoveCON(*s*)表示删除最小生成树中笔划 *s* 与其他笔划相连的边.

AddTail(*List*,*s*)表示将笔划 *s* 加到链表 *List* 的尾部.

GetHead(*List*)表示获取链表 *List* 的头节点.

Remove(*List*,*s*)表示将笔划 *S* 从链表 *List* 中删除.

Count(*List*)表示链表 *List* 中的节点个数.

NewWord(*w*)表示生成新部件 *w*,*w*=∅.

AddWord(*W*,*w*)表示向部件集合 *W* 中添加部件 *w*.

RemoveWord(*W*,*w*)表示从部件集合 *W* 中删除部件 *w*.

$R(w_1,w_2)$表示部件 w_1 相对于部件 w_2 的位置关系.

Height(*w*)表示部件 *w* 的高度.

Width(*w*)表示部件 *w* 的宽度.

ES(*e*)表示边 *e* 所连接的笔划.

$G(w_1,w_2)$表示部件 w_1 和 w_2 间的间距.

Aarctg(w_1,w_2)表示部件 w_1 和 w_2 合并后的新部件高度与宽度比的反正切值.

- 笔划集合 *Y* 切分算法描述

1. 定义空链表 *StrokeBuffer*,部件集合 *W*=∅;

对于输入笔划集合 $Y=\{y_1,y_2,y_3,\ldots,y_n\}$,存在按笔划中点 *x* 坐标升序排列的笔划集合 $S=\{s_1,s_2,s_3,\ldots,s_n\}$,$\forall y_i,\exists s_j((s_j=y_i)\wedge(MX(s_j)\le MX(s_{j+1}))),(1\le j\le n-1)$.然后对于任一 $s_i\in S$,取 $s_k\in S(i+1\le k\le i+N)$,由 $E(M(s_i),M(s_k))$构建图 *G*.

2. 对第 1 步生成的图 *G*,设连接笔划 s_j 和 s_k 的边为 e_i,e_i 的权值 $Value(e_i)=Distance(M(s_j),M(s_k))$,其中 *i* 为连接笔划 s_j 和 s_k 的边的边号.生成笔划集合 *S* 的最小生成树为

$$MakeMST(S)\Leftrightarrow\left(\bigcup_{j=1}^{n-1}ES(e_j)=S\right)\wedge\min\left(\sum_{j=1}^{n-1}Value(e_i)\right).$$

3. *AddTail*(*StrokeBuffer*,s_1).

4. If (*Count*(*StrokeBuffer*)==0) goto step 10.

5. 令 *s*=*GetHead*(*StrokeBuffer*),*NewWord*(w_j)∧*Merge*(w_j,*s*)∧*AddWord*(*W*,w_j)∧*Remove*(*StrokeBuffer*,*s*)∧
AddTail(*StrokeBuffer*,*CON*(*s*))∧*RemoveCON*(*s*).

6. 令标记变量 *hasMerge*=*false*,{(∀s_i)(($s_i\in$*StrokeBuffer*)∧(*R*(s_i,w_j)∈{$p_{LU}|p_{LD}|p_L|p_U|p_D|p_C|p_I$ }))⇒
(*hasMerge*=true∧*Merge*(w_j,s_i)∧*Remove*(*StrokeBuffer*,s_i)∧*AddTail*(*StrokeBuffer*,*CON*(s_i))∧
RemoveCON(s_i))}|{(∀s_i)(($s_i\in$*StrokeBuffer*)∧(*R*(s_i,w_j)∈{$p_R|p_{RU}|p_{RD}$})∧($\frac{Height(w_j)}{Height(s_i)}<\varepsilon$))⇒(*hasMerge*=
true∧*Merge*(w_j,s_i)∧*Remove*(*StrokeBuffer*,s_i)∧*AddTail*(*StrokeBuffer*,*CON*(s_i))∧*RemoveCON*(s_i))}.

7. If (*hasMerge*==true) goto step 6.

8. (*hasMerge*==false∧*Count*(*StrokeBuffer*)>1)⇒∃s_k(($s_k\in$*StrokeBuffer*)∧Max(*Distance*(*M*(s_k),*M*(w_j)))∧
CON(s_k)>0) then (∀s_f)($s_f\in$*StrokeBuffer*∧$s_f\neq s_k$)⇒*Merge*(w_j,s_f)∧(*Remove*(*StrokeBuffer*,s_f)∧*AddTail*
(*StrokeBuffer*,*CON*(s_f))∧*RemoveCON*(s_f)).

9. Goto step 4

10. 计算现有部件的平均高度 H_{ave},平均宽度 W_{ave} 和平均间距 G_{ave},令 *hasMerge*=false.

11.(∀$w_k|w_k\in W$),(*Height*(w_k)<αH_{ave})∧(*Width*(w_k)<βW_{ave})∧(∃$w_m|w_m\in W$)(*G*(w_k,w_m)<γG_{ave})∧(*Aarctg*(w_k,w_m)>ζ))⇒
Merge(w_k,w_m)∧*hasMerge*=true∧*RemoveWord*(*W*,w_m),($\alpha,\beta,\gamma,\zeta$为常数).

12. If (*hasMerge*==false),stop; else goto step 10.

以上讨论中使用到的常数$\varepsilon,\alpha,\beta,\gamma,\zeta$是通过对用户的输入样本进行学习处理以后得到的常量.这些参数随用户输入习惯的不同而取不同的值.

2 实验结果

使用以上算法分别对中文和数字符号进行了切分实验,对算法中各参数的取值是:N=12,α=0.58,β=0.5,γ=0.5,ζ=0.7,ε=0.25.切分结果如图 3~图 9 所示.

Fig.3 Chinese handwriting strokes

图 3 输入的汉字笔划

(a) Chinese handwriting strokes with branches of MST (a) 汉字笔划及其最小生成树的树枝

(b) The central points of handwriting strokes with branches of MST (b) 笔划中点与最小生成树的树枝

Fig.4 Chinese handwriting strokes with branches of MST composing by the central points of strokes

图 4 汉字笔划和笔划中点构成的最小生成树

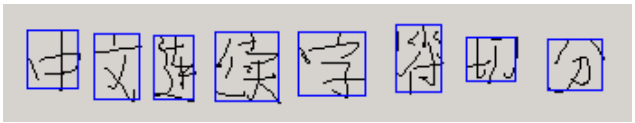

Fig.5 Chinese handwriting segmentation results

(the periphery rectangle is the outer boundary of character)

图 5 汉字切分结果(最外围的矩形框为字符的外包围框)

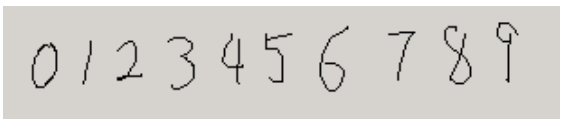

Fig.6 Numerical handwriting strokes

图 6 输入的数字笔划

(a) Numerical handwriting strokes with branches of MST (a) 数字笔划及其最小生成树的树枝

(b) The central points of handwriting strokes with branches of MST (b) 数字笔划中点与最小生成树的树枝

Fig.7 Numerical handwriting strokes with branches of MST composing by the central points of strokes

图 7 数字笔划和笔划中点构成的最小生成树

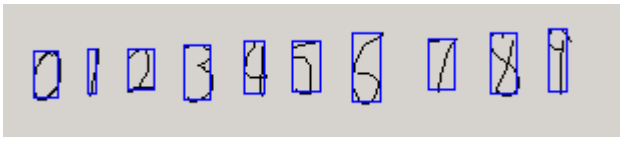

Fig.8 Numerical handwriting segmentation results
(the periphery rectangle is the outer boundary of number)
图 8 数字切分结果(最外围的矩形框为字符的外包围框)

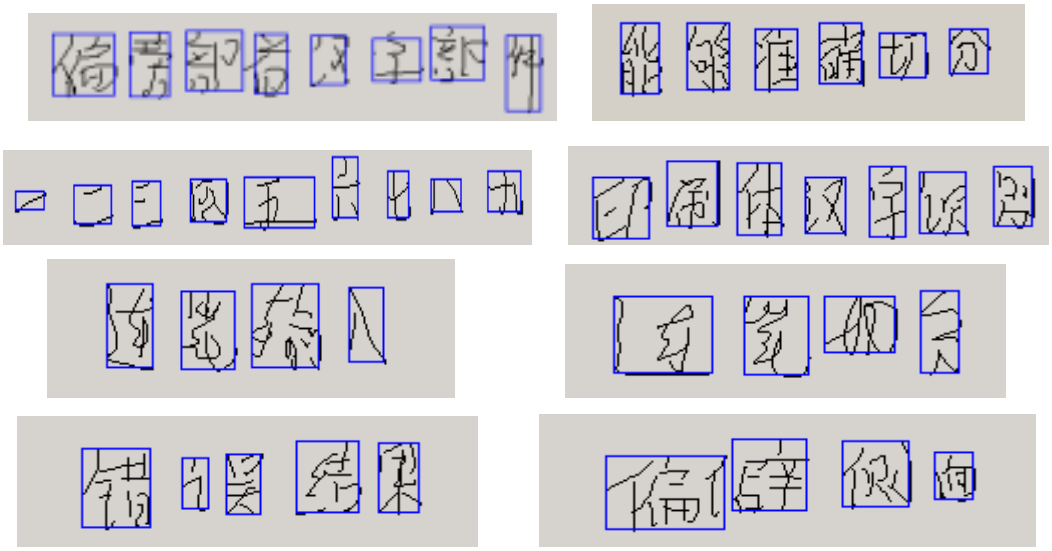

Fig.9 Some other handwriting segmentation results
图 9 其他实验切分结果

对本文中的汉字切分方法,我们进行了实验统计:对于连续输入的 500 个汉字,完全由单笔进行的输入,出现了切分错误 23 处,影响到 42 个汉字,切分准确率为 91.6%;由单笔和连笔混合输入测到的切分准确率为 95.2%;完全由连笔输入的切分准确率为 98%.切分错误主要出现在 step 11:对于左右结构的合并,如果合并条件过于宽松,就会把多个汉字合并到一个字中去;反之,就可能把一个汉字切分成 2 个甚至 3 个汉字(如图 9 中最后两个切分结果).因此,在汉字左右结构的合并中,仅仅利用空间位置等信息是不够的.

3 结 论

本文利用最小生成树和汉字的部件结构特征对手写体汉字输入进行了切分,取得了较好的结果,尤其是对上下结构和包围结构汉字的切分准确率较高,切分错误主要出现在对左右结构的汉字切分中.根据目前的研究结果,我们感觉到:在对汉字进行切分时,仅仅利用空间信息是不够的,最好还能结合汉字识别算法对手写笔划和识别结果的相似性度量以及上下文语义,对相似性较小的切分结果进行修正,重新切分后再进行识别,以便获得更为圆满的结果.

致谢 本文的部分工作源于作者在中国科学院软件研究所人机交互及智能工程实验室学习和实习期间所参加的课题.在此,特别要感谢戴国忠研究员、陈由迪研究员、栗阳博士以及实验室全体同学和员工对我们的指导和帮助.

References:

[1] Wu YS. Chinese Character Recognition-Principle, Method and Realization. Beijing: Higher Education Press,1992 (in Chinese).

[2] Breuel TM. Representation and metrics for off-line handwriting segmentation. In: Proc. of the 8th Int'l Workshop on Frontiers in Handwriting Recognition. 2002. 821−826. http://citeseer.ist.psu.edu/breuel02representations.html

[3] Han BX. Combination of Chinese character constituents-A latent structural unit. Journal of Chinese Information Processing, 1995,9(3):27−32 (in Chinese with English abstract).

[4] Lu Y. Segmentation of Free-format Handwritten Chinese Characters Based on Structure Features of Characters. Acta Electronica Sinica, 2000,128(5):102−104 (in Chinese with English abstract).

[5] Casey RG. A survey of methods and strategies in character segmentation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1996,18(7):690−706.

[6] Breuel TM. Segmentation of handprinted letter strings using a dynamic programming algorithm. In: Proc. of the 6th Int'l Conf. on Document Analysis and Recognition. 2001. 821−826. http://citeseer.ist.psu.edu/breuel01segmentation.html

[7] Punnoose J. An improved segmentation module for identification of handwritten numerals [MS. Thesis]. Massachusetts Institute of Technology, 1999.

[8] Verma B, Blumenstein M. Recent achievements in off-line handwriting recognition systems. In: Proc. of the Int'l Conf. on Computational Intelligence and Multimedia Applications. 1998. 27−33. http://citeseer.ist.psu.edu/130825.html

[9] Chen H. Segmentation and recognition of continuous handwriting Chinese text. In: Proc. of the Int'l Conf. on Computer Processing of Oriental Languages. 1997. 630−633. http://citeseer.ist.psu.edu/hong97segmentation.html

[10] Nicholas E. Recognition of handwritten mathematical expressions [MS. Thesis]. Massachusetts Institute of Technology, 1999.

[11] Zhang XW. Adaptive Character Extraction from Continuous HandWriting Chinese Text Based on Multilevel Constrains. 2003 (in Chinese with English abstract). http://iel.iscas.ac.cn/lab_achive/lab_achive_publications_2003.htm

附中文参考文献:

[1] 吴佑寿.汉字识别——原理、方法与实现.北京:高等教育出版社,1992.

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

[4] 吕岳.基于汉字结构特征的自由格式手写体汉字切分.电子学报,2000,128(5):102−104.

[11] 张习文.基于多层次信息的连续手写中文的自适应分割方法.2003.http://iel.iscas.ac.cn/lab_achive/lab_achive_publications_2003.htm

韩勇(1974－),男,四川崇州人,博士生,主要研究领域为人机交互.

戴国忠(1944－),男,研究员,博士生导师,CCF 高级会员,主要研究领域为人机交互,软件工程.

须德(1944－),男,教授,博士生导师,主要研究领域为信息检索,数据库,多媒体信息处理.