Visualization of Extracted Digital Ink Text Based on Reliability

Hao Bai

Beijing Language and Culture University, College of Advanced Chinese Training, Beijing, China

baihao@blcu.edu.cn

Keywords: Digital ink; Character extraction; Visualization; Reliability

Abstract. The result of segmented digital ink text contains three levels: characters, lines and paragraphs. Characters make up a significant percentage in the result and the situations of them are always complex. Automatic methods hardly provide completely correct result of extracted characters. So the result needs to be modified by human-computer interactive operation. Optimized visualization could improve the efficiency of modification. Therefore oriented to the modification of the errors of extracted characters, according to the adjacent and overlapping among characters, the paper proposes an improved way of visualization. The approach firstly makes assessment of extracted characters and then visualized them based on diffident reliability of correction. Testing on many sorts of extracted characters from digital ink text in Chinese, the approach is effective.

基于可信度的单字提取结果可视化方法

白浩

北京语言大学 汉语进修学院，中国 北京 100083

baihao@blcu.edu.cn

摘要：数字墨水文本的分割结果包含单字、文本行和段落三个层次对象，单字在其中占有较大比例，而且情况复杂。使用自动的提取方法难以提供完全正确的结果，这时需要进行人机交互校正单字提取结果。优化的可视化方法可以在人机交互时大大提高校正效率。因此，面向交互校正错误的单字提取结果，针对单字结果间的邻近和重叠等情况，给出了一种改进的可视化方法。该方法先对单字提取结果作可信度评价，将可信度低的结果区分性可视化。对多种数字墨水文本的单字提取结果进行可视化表示，取得了较好的效果。

关键词：数字墨水；单字提取；可视化；可信度

1. 引言

人们使用手写板、数码笔[1]等采集设备可以获取中文数字墨水文本并对其进行分割处理，能够提取单字、文本行和段落三个层次对象[2]。基于分割结果，不仅可以进行对象的修改和排版，还可以进行识别文字，从而输入文字处理软件进行后续处理。

在中文数字墨水文本中，单字占有很大比例，包括多种类型，汉字、标点、数字、字母、英文单词等。现有自动分割方法难以提供完全正确的单字提取结果，所以必须进行人机交互校正。优化的可视化方法可以在人机交互时大大提高校正效率。针对单字提取结果，已有多种可视化方法，包括针对汉字的正放最小外接矩形[2]、针对英文单词的斜放最小外接矩形[3]、连线[4]、颜色[5]以及针对单字结果间邻近和重叠等复杂情况的自适应可视化方法[6]，但现有的可视化方法仍然无法满足以汉字为主的中文数字墨水文本的单字提取结果的可视化要求。

一些提取错误的单字间没有发生重叠，甚至字间距较小的情形，此时采用基于重叠的自适应可视化方法[6]时，这些分割错误的单字会以正放矩形框的可视化方法表示，没有进行有效的区分，从而并没有在人机交互式校正时降低用户的认知负担。然而，分割错误中经常有明显的单字宽高比异常或者笔画过多等现象，针对这些单字，可以依据提取结果的可信度进行可

视化表示。

2. 可信度评价

在用户校正时需要利用自身的认知能力对每个单字的正误进行区分，使得校正效率大大降低。有三种类型的分割结果可认为可信度较低：单字宽高比过小（疑似过分割错误）、单字宽高比过大（疑似欠分割错误）和单字笔画数过多（疑似错分割）。

当单字的宽高比过小（<0.5）时，使用自动分割的方法无法区分该单字是出现了过分割错误还是正常的单字，如图 1 所中虚线框所示。此时用户只有结合上下文语义知识才能分辨，但可认为其为分割正确的单字的可信度较低，将其进行区分表示，引起用户的注意，可明显降低其认知负担。

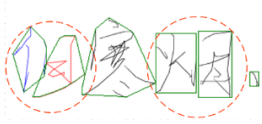

图 1 疑似过分割错误

图 2 疑似欠分割错误

图 3 疑似错分割

单字的宽高比过大（<1.5）时，宽高比超出了正常值，但是自动分割的方法无法区分该单字是出现了欠分割错误还是正常的单字。如图 2 中所示，可认为将“凝”分割为正确单字的可信度较低，将其区分表示，引起用户的注意，可明显降低其认知负担。

错分割的情况往往是只能结合上下文的用户才能分辨，所以自动辨别错分割的情况往往也是很困难的。但常用汉字的平均笔画数为 9.17[7]，如图 3 所示，有时错分割的单字会出现笔画数过多（>10）等情况，这种情形下，可认为其为分割正确的单字的可信度较低，进行区分性表示，从而在人机交互式校正的时候引起用户的注意，降低其认知负担。

3. 基于可信度的可视化算法

分割结果的自适应可视化是为了降低用户的认知负担，从而使用户快速找到错误或者可信度低的单字，通过自身的语义知识进行辨别，然后对分割错误进行人机交互式校正。其中，当找到可信度低的单字时，算法采用将单字内所有笔画加粗的可视化方法，达到引起用户注意的效果，同时也不会与基于重叠的可视化方法[6]发生冲突。该算法的具体步骤为：步骤一，在已有可视化结果[6]的基础上对单字的可信度进行评估；步骤二，若可信度低，则将单字内所有笔画加粗。

其中，当单字的宽高比过小（<0.5）时，疑似欠提取错误，如图 4 所示基于重叠的自适应可视化方法效果图，图 5 中所示为基于可信度的可视化方法的效果图。当单字的宽高比过大（>1.5）时，疑似过提取错误，如图 6 所示基于重叠的自适应可视化方法效果图，图 7 中所示为基于可信度的可视化方法的效果图。当单字笔画数过多（>10）时，疑似错提取，如图 8 所示基于重叠的自适应可视化方法效果图，图 9 所示为基于可信度的可视化方法的效果图。

图 4 基于重叠的自适应可视化方法

图 5 基于可信度的可视化方法

图 6 基于重叠的自适应可视化方法

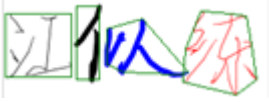

图 7 基于可信度的可视化方法

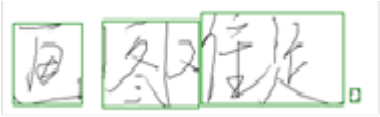

图 8 基于重叠的自适应可视化方法

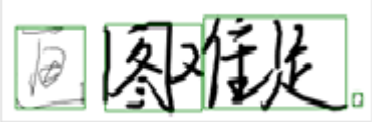

图 9 基于可信度的可视化方法

4. 性能评价

基于上述所提出的方法，本文采用了 Microsoft visual studio 2005 开发平台，使用 C#编程语言开发了一个原型系统。该系统运行于装有 windows xp sp3 操作系统的 PC 上。下面根据大量的中文数字墨水文本的测试结果进行定量分析来对本文所提出的方法做性能评价。

为了验证本节方法的有效性，本文提出了以下三个评价指标：

区分个数：即区分可视化的单字总个数；

有效区分率，即区分可视化的单字为错误单字的个数与区分个数的比值；

漏查率，即未被区分可视化的单字为错误单字的个数与错误单字总个数的比值；

本文的中文数字墨水实验数据来自 6 个大学本科生书写的宋词片段，采用瑞典 Anoto 公司生产的数码笔[1]进行采集，以 MS Tablet PC SDK[8]开发的原型系统对数据进行分割和绘制，可视化的统计结果如表 1 所示，实验结果表明：（1）基于可信度的可视化方法有效区分率并不高，漏查率较高；（2）漏查率与分割正确率成反比例线性变化；（3）该方法有效区分率虽然不高，但被试用户反馈当校正分割错误时该方法是有效的。

表 1 基于可信度的可视化方法统计表

数据	字数	区分个数	有效区分率	漏查率
001(a)	100	22	36.36%	57.89%
002(b)	114	41	17.07%	56.25%
097(c)	236	77	15.58%	47.83%
036(d)	285	38	28.95%	72.50%
038(e)	251	31	35.48%	73.81%
119(f)	329	123	19.51%	31.43%

5. 结语

本文针对现有的自适应可视化方法做出改进，提出了一种基于可信度的单字提取结果的可视化方法，该方法先对单字提取结果作可信度评价，将可信度低的结果区分性可视化。实验数据表明，这种可视化方法在提高交互式校正效率时是有效的。

6. 致谢

本文得到了北京语言大学项目（中央高校基本科研业务费专项资金资助）资助，编号为 16YJ080201。

参考文献

[1] 瑞典Anoto公司[OL], http://www.anoto.com.

[2] Xi-Wen Zhang, Yong-Gang Fu, Kun Zhang.Adaptive Correction of Errors from Recognized Chinese Ink Texts Based on Context[C]. Proceedings of 2009 International Conference on Information Technology and Computer Science, 1, 2009:314-320.

[3] T. Artières. Poorly structured handwritten documents segmentation using continuous probabilistic feature grammars[C]. Workshop on Document Layout Interpretation and its Application 2003 (DLIA-3 2003):5-8.

[4] L. Likforman-Sulem, A. Zahour, B. Taconet. Text line segmentation of historical documents: a survey [J]. International J. Documents Analysis and Recognition, 9, 2007:123-138.

[5] A. Bhaskarabhatla, S. Madhvanath, M. Kumar, et al. Representation and Annotation of Online Handwritten Data[C].Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9), Tokyo, Japan, October 2004:136-141.

[6] 白浩、张习文、付永刚, 等. 数字墨水中单字提取结果的自适应可视化方法[J]. 计算机工程与应用, 2012(15):153-158.

[7] 祝莲、王晨晓、贺极苍等. 中文字体大小、笔画数和对比度对阅读速度的影响[J]. 眼视光学杂志. 2008年2月:96-99.

[8] Yang Li, Zhiwei Guan, Hongan Wang, Guozhong Dai, Xiangshi Ren. Structuralizing freeform notes by implicit sketch understanding[C]. Proceedings of AAAI Spring Symposium on Sketch Understanding. Palo Alto, California, 2002:113-117.

[9] Microsoft Windows XP Tablet PC Edition Software Development Kit 1.7. [CP/OL]http://www.microsoft.com/downloads/details.aspx?familyid=b46d4b83-a821-40bc-aa85-c9ee3d6e9699&displaylang=en.

Acknowledgement

This research was financially supported by Science Foundation of Beijing Language and Culture University (supported by “the Fundamental Research Funds for the Central Universities”) (no. 16YJ080201)

References

[1] Anoto, Inc**.** [OL], http://www.anoto.com.

[2] Xi-Wen Zhang, Yong-Gang Fu, Kun Zhang. Adaptive Correction of Errors from Recognized Chinese Ink Texts Based on Context[C]. Proceedings of 2009 International Conference on Information Technology and Computer Science, 1, 2009:314-320.

[3] T. Artières. Poorly structured handwritten documents segmentation using continuous probabilistic feature grammars[C]. Workshop on Document Layout Interpretation and its Application 2003 (DLIA-3 2003):5-8.

[4] L. Likforman-Sulem, A. Zahour, B. Taconet. Text line segmentation of historical documents: a survey [J]. International J. Documents Analysis and Recognition, 9, 2007:123-138.

[5] A. Bhaskarabhatla, S. Madhvanath, M. Kumar, et al. Representation and Annotation of Online Handwritten Data [C]. Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition (IWFHR-9), Tokyo, Japan, October 2004:136-141.

[6] BAI Hao, ZHANG Xiwen, FU Yonggang, et al. Adaptive visualization of extracted digital ink characters in Chinese[J]. Computer Engineering and Applications, 2012, 48(15): 153-158.

[7] ZHU Lian, WANG Chenxiao, HE Jicang, et al. The effects of font,stroke and contrast on the reading speed of Chinese characters [J]. CHINESE JOURNAL OF OPTOMETRY & OPHTHALMOLOGY. 2008, 10(2):96-99.

[8] Yang Li, Zhiwei Guan, Hongan Wang, Guozhong Dai, Xiangshi Ren. Structuralizing freeform notes by implicit sketch understanding[C]. Proceedings of AAAI Spring Symposium on Sketch Understanding. Palo Alto, California, 2002:113-117.

[9] Microsoft Windows XP Tablet PC Edition Software Development Kit 1.7. [CP/OL]http://www.microsoft.com/downloads/details.aspx?familyid=b46d4b83-a821-40bc-aa85-c9ee3d6e9699&displaylang=en.

作者简介：白浩（1984—），男，北京，初级，模式识别、计算机图形学，baihao@blcu.edu.cn。