计算机科学2004Vol.31№.12

一种汉字笔画提取模型的改进和实现

陈 睿 唐 雁
(西南师范大学计算机与信息科学学院 重庆400715)

摘 要 本文对一种基于笔画段分割和组合的汉字笔画提取模型,进行了细化和改进,以提高笔画提取的正确率和速度,并给出了其实现方法。
关键词 笔画提取,分叉点区域,笔画段,PBOD曲线

An Improved Stroke Extraction Model for Chinese Character and its Implementation

CHEN Rui TANG Yan
(Department of Computer Science, Southwest China Normal University, Chongqing 400715)

Abstract The stroke extraction model based on stroke segmentation and combination is an effective model based on local information applied in off-line recognition system for Chinese characters. This paper proposes an improved stroke extraction moel based on stroke segmentation and combinatin to deal with the complexity and the specialty of the strokes of Chinese character. The experiments show that the new model is more reliable and has been speeded up.
Keywords Stroke extraction, Region decomposition, Stroke segmentation, PBOD curve

1 基于笔画段分割和组合的汉字笔画提取模型

汉字笔画提取有从汉字骨架图二值图、轮廓图中提取笔画等技术。这些技术都有各自的适用范围和优缺点。其中,二值图技术能够利用更多的字符像素信息,使提取结果达到更高的正确率,但缺点是时间复杂度较高。在此类技术中,存在一种基于笔画段分割和组合的汉字笔画提取模型[1]。这种模型是在二值图中进行运算,利用分叉点区域提取及笔画段分割和组合的思想,达到笔画提取的目的。它主要包含以下几个步骤:

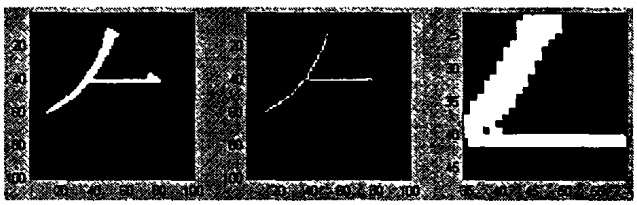
1. 分叉点区域提取和笔画段分割
2. 笔画段组合
   2.1单分叉点区域提取
   2.2双分叉点区域提取
   2.3三分叉点区域提取
3. 笔画修正

2 对笔画提取模型的几点改进

2.1 笔画段组合标准的改进

在文[1]模型中,对笔画段组合正确与否的判断采用了两种标准。设分叉点区域的中心为 p,第一种标准是在计算每个笔画段组合图像中 p 的PBOD曲线后,看该曲线中是否只包含两个波峰,且这两个波峰相距是否接近180度;第二种标准是计算每个笔画段组合图像中 p 的BBOD曲线后,看该曲线是否只有一个波峰。

实验证明第二种标准存在问题。如图1所示,a和b分别代表分割前的笔画和其骨架图,c代表笔画段组合的一种情况,即上边和右边的笔画段组合。p 在c中用黑点标出。很明显,该组合不正确,却满足第二种标准,其原因为c图中 p 点与横向的笔画段错开。它不满足第一种标准,所以只采用第一种标准。实验中波峰相距接近180度定义为150~210度。



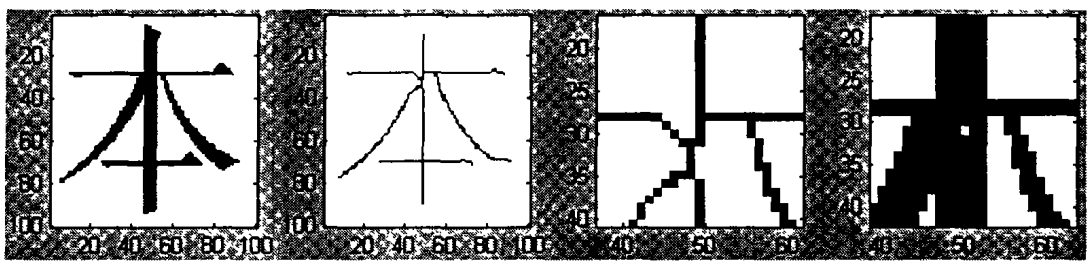
a b c

图1 第二种标准出错的情况

2.2 相邻分叉点合并与否的讨论

在原模型中,若骨架图中存在几个相距很近(小于某个阈值)的分叉点,就将其合并为它们的重心。这种做法能大大降低运算量并快速提取分叉点区域,在多数情况下也能正确地进行笔画段分割与组合,但在某些情况下会发生问题。

如图2所示,a和b分别代表二值图和骨架图[2],c是b的部分放大图。根据相交数定义,c中明显有4个相距很近的分叉点。图d中的白点代表这4个分叉点的重心,而如果仅用该点作为分叉点区域的中心进行笔画段分割与组合,则无法提



a b c d

图2 相邻分叉点合并前后的情况

取出该点上方的笔画"横",其原因为该点与"横"的两个笔画段错开。

对此类问题的解决办法是,先用分叉点的重心来进行笔画段组合,若结果图对应的骨架图中仍有几个相距很近的分叉点,则分别将其作为分叉点中心进行笔画段组合。

2.3 PBOD 曲线波峰数和位置确定方法的改进

PBOD 曲线[3]的获得:用长为3的跨度在360度的范围内取121个采样点,分别计算这些点对应的 PBOD 值,得到一个长度为121的数组,将其归一化后用这些值画出 PBOD 曲线。

在原模型中,波峰数及位置的确定方法如下:首先确定一个波峰的阈值(如0.4),然后数出曲线中从小于到大于该阈值的正跳变的数目,将其作为波峰数,同时将正跳变出现的位置作为波峰的位置。

这种方法虽然简单,但存在一些问题。1. 正跳变的数目不一定就是波峰的数目,若两波峰间波谷的值大于阈值,则这两个波峰只计数为1。2. 两个相距很近(小于10度)的波峰实际上应属于同一个笔画段。3. 波峰的位置不应该是跳变出现的位置。4. 横跨曲线首尾的波峰应只计数1次。

求 p 点 PBOD 曲线的算法

Step1:初始化曲线向量 f,f 的维数为121,表示从0°到360°以3°为间隔进行采样;

Step2:将角度集合 s 定义为[0°,3°,6°,…,360°];

Step3:依次取出 s 中的各个角度,计算角度 $s(i)$ 方向上值为1的点的连续长度 $f(i)$,这些点的横纵坐标的计算公式为 $x=L\times\sin(s)$ 和 $y=L\times\cos(s)$,其中 L 为该点到 p 点的距离;

Step4:对向量 f 进行归一化,即 $f=f/\max(f)$;

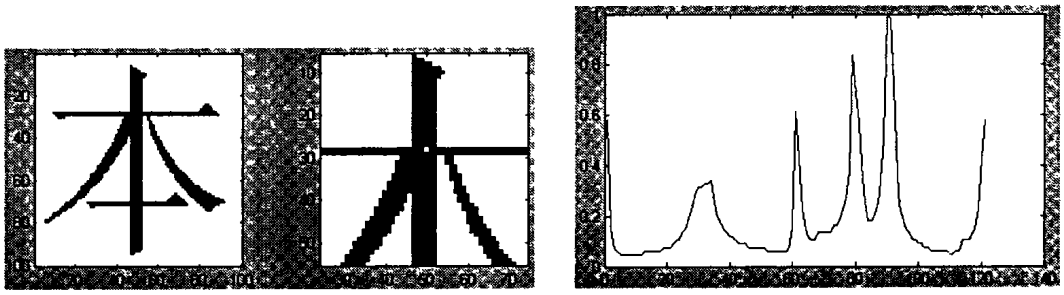
Step5:依次取出 f 中的各个点 $f(i)$,计算它的四个相邻点 $n1$,$n2$,$n3$,$n4$,对 f 中的首尾端点当作相邻点考虑;

Step6:如果 $f(i)$ 小于它的四个相邻点中的某一个,转 Step5;

Step7:依次取出 f 中的相邻的两个波峰,如果它们之间的波谷值大于阈值,或它们的距离<10,则合并为一个波峰,位置在二者中点,强度为二者平均值;

Step8:取出 f 中的首尾两个波峰,如果它们的距离<10,则合并为一个波峰。

如图3所示,a 代表字符图像,放大图中的白点是计算的对象。b 代表该点的 PBOD 曲线,实验证明新方法能解决原方法中的问题。



a　　b

图3　字符像素及其 PBOD 曲线

2.4 多分叉点区域提取的实现方法的改进

在原模型中,对单分叉点区域提取不能将笔画段分割为≥3个的情况,采取了双分叉点和三分叉点区域提取的方法,即同时提取两个或三个分叉点区域。对多分叉点区域提取采用向下层(即单分叉点区域提取)调用的方法实现。

三分叉点区域提取算法

Step1:将字符二值图导入二维数组 I;

Step2:清空结果三维数组 N,其中前两维存放图像内容,第三维存放图像序号;

Step3:计算 I 的骨架图 K;

Step4:找出 K 中的所有分叉点,将它们存放在集合 F 中,清空三维数组 *outs*;

Step5:依次取出 F 中的两个分叉点 $F(i)$ 和 $F(j)$,分别计算它们对应的分叉点区域 $F1$ 和 $F2$;

Step6:计算字符图像 I 减去 $F1$ 和 $F2$ 后的二值图 J,$J=I-F1-F2$;

Step7:对 J 进行单分叉点区域提取,结果存入 *outs*;

Step8:如果 *outs* 中笔画段数小于3,转 Step5;

Step9:如果 *outs* 和 $F1$ 相邻,则对二者进行或运算后,再进行笔画修正;

Step10:如果 *outs* 和 $F2$ 相邻,则对二者进行或运算后,再进行笔画修正;

Step11:将 *outs* 存入 T 中,清空 *outs*1,对 T 分别以 $F1$ 和 $F2$ 为分叉点区域,$F(i)$ 和 $F(j)$ 为分叉点进行笔画段组合,组合结果存入 *outs*1;

Step12:去掉 *outs* 中已经组合的笔画段,合并 *outs* 和 *outs*1,将结果存入 N 中,转 Step5;

Step13:算法结束。

3 汉字笔画提取模型的实现

输入:一幅二值字符图像,将其大小调整为100×100,存储为一个三维数组,前两维代表该图像,第三维大小为1。输出:一组笔画图像,存储为一个三维数组,前两维代表一幅笔画图像,第三维代表该笔画图像的序号。

总体流程:将输入图像存储为三维数组 *strokes*,找出该图像中的所有笔画段,将 *strokes* 拓展为一组图像,每幅图像存储一个笔画段,大小都为100×100。依次取出 *strokes* 中的每幅图像进行如下处理(*P*):找出它的骨架图 K,找出 K 中的所有分叉点 *points*(二维数组),依次取出 *points* 中的每个点进行分叉点区域提取和笔画段分割,其中多分叉点区域提取用向下层调用的方法实现。其中每提取出一个笔画,都要先对其进行笔画修正,然后在 *strokes* 数组中增加这幅笔画图像(100×100),同时原图中的剩余部分(不能合并的部分)也增加为 *strokes* 中的一幅图像。对 *strokes* 的 *P* 处理一直进行下去,结束条件为 *strokes* 中的每幅图像的骨架图中都不含分叉点。

结论　本文对一种基于笔画段分割和组合的汉字笔画提取模型进行了细化和改进,提高了笔画提取的正确率和速度,并给出了它的实现方法。实验证明,这种改进和实现方法非常有效。

参 考 文 献

1　陈睿,唐雁. 基于笔画段分割和组合的汉字笔画提取模型. 计算机科学,2003(10)

2　Zhang T Y, Suen C Y. A Fast Parallel Algorithm For Thinning Digital Patterns. Communications of the ACM, 1984, 27(3)

3　Cao R, Tan C L. A Model of Stroke Extraction from Chinese Character Images