中 文 信 息 学 报
第7卷 第3期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.7 No.3

手写印刷体汉字的笔段抽取及偏旁识别

胡家忠

(武汉工业大学)

【摘要】本文采用对汉字点阵图象进行方向变换的方法抽取汉字的笔段，采用结构分析的方法识别分布于汉字四周的偏旁。对国标一级汉字中的99类偏旁计一万余字进行了偏旁抽取试验，当候选偏旁数<5时，累计正确候补率>96%。

一、前　言

手写印刷体汉字识别是模式识别中的一个重要课题。由于汉字的字数多、结构复杂，为了高速地识别每一个扫描输入的汉字，笔者认为采用三级分类识别的方法较好。即首先按分布于汉字四周局部区域的偏旁进行第一级粗分类，接着按四周外围特征进行第二级细分类，最后进行逐字详细识别。

我们知道，汉字的结构虽然复杂，但汉字都是由直线线段构成的，而且这些线段具有横、竖、撇、捺四个方向。多数汉字（约占85%）具有偏旁，这些偏旁分布于汉字四周的局部区域，并且由少数笔段组成。相对汉字来说，它们的种类少，结构简单，因此，在进行逐字识别前，如果能先识别它们的偏旁，无疑是进行汉字识别的一条捷径。采用传统的模板匹配方法，由于书写者书写风格各异，字形变化大，再加上偏旁的笔划少，它们之间的区别往往只有一笔之差，为了保证分类的精度，不得不采用多个模板，选用多个侯补偏旁的办法来满足后级分类的需要，这样反过来又导至分类速度下降，达不到分类既快又准的要求。

正因为偏旁分布于汉字四周的外围部分，笔段比较容易提取，它们之间的区别往往在于少数笔划及其位置分布，如果能够正确地提取分布于汉字四周的笔段，则采用结构分析的方法就能较好地识别汉字的偏旁。本文介绍将输入汉字的点阵图象进行方向变换的方法，抽取汉字的笔段。利用对各种偏旁的先验知识，按照线段的长度、方向及其相互位置关系进行分析判断，达到正确抽取汉字偏旁的目的。这种方法的优点是能够保存原始图象的特征，不因细线化、折线化而丢失原始特征，从原理上说能够抽取任意方向的由直线段组成的任意形状的图形。

①本文1993年5月29日收到

二、方向线段的抽取

我们将输入待识汉字正规化后，按图 1 所示的步骤分步实现方向线段的抽取。

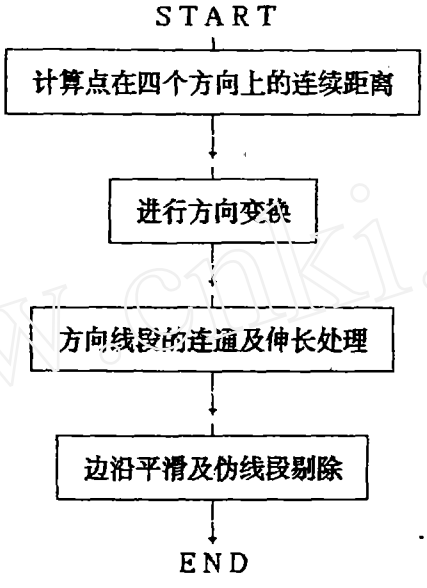


图 1　方向变换流程图

图 1　方向变换流程图

2.1　方向线段的抽取方法

步骤 1　点的方向距离的测定

我们以二进制代码 0001、0010、0100、1000 表示横、竖、撇、捺四个方向对于汉字图象上的黑点（x、y）按图 2 所示的四个方向进行延伸，分别计算连续黑点的点数，称为点（x、y）在四个方向的距离，取其中最大的距离方向赋予点（x、y），这样逐点变换即生成如图 3 所示的方向画面。

步骤 2　方向线段的抽取

在进行图象点的方向变换时，如果两条线段在某处相交，由于交点处的点只被赋予其中较长一条线的方向，结果另一条本来是连续的线却被拦腰截成两段，如图 3 所示，上面的一条横线被竖分成了两段，对于这种情况应进行线段的连通处理。相反，假如横线长于竖线，则竖线就应进行伸长处理了。进行了这样的处理后，横竖相交的地方的方向码就变成 3 了。

步骤 3　边沿平滑及伪线段的剔除

按照上面的变换方法生成的方向画面一般来说是不干净的。如图 3 所示，在方向线段的边沿，有其它方向的点存在，有的地方甚至比较严重，例如图 3 中永字上面的一点，就

有方向码 1 及方向码 2 相混杂，图 3 中捺的方向码 8 被方向码 1 所包围。这些情况在多笔划的汉字中尤其严重，甚至会产生串笔现象，即产生所谓的伪线段（在实际汉字中不存在的笔段），这也是使用这种方法抽取笔段最困难的地方。解决诸如此类问题的方法是：首先进行边沿平滑，可按垂直于该方向线段的方向进行扫描，当异方向的点数小于规定的阈值时则将其矫正之。其次解决笔划的混杂问题，具体做法是求混杂笔划的外接矩形框，根据外接矩形框的长／宽最后决定该笔划的方向。为了剔除伪线段，首先需要发现伪线段，伪线段的共同特点是此类线段没有端点存在，若将其剔除，不会改变原图象其它线段间的接续关系。图 3 中永字经过这些处理后得到如图 4 所示的图象。

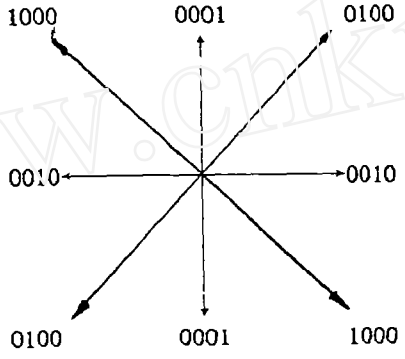


图 2　笔段方向代码

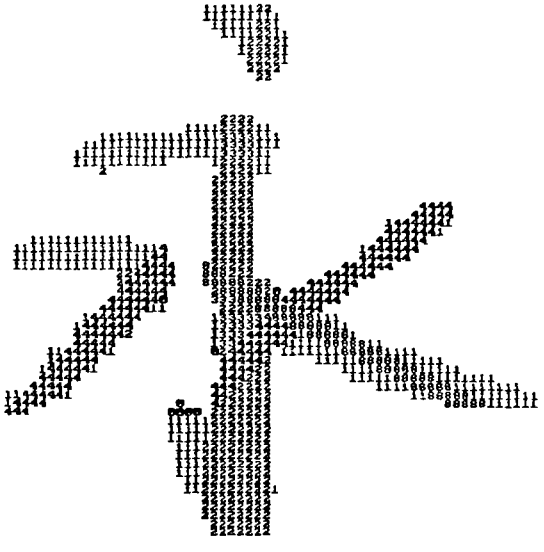


图 3　方向画面

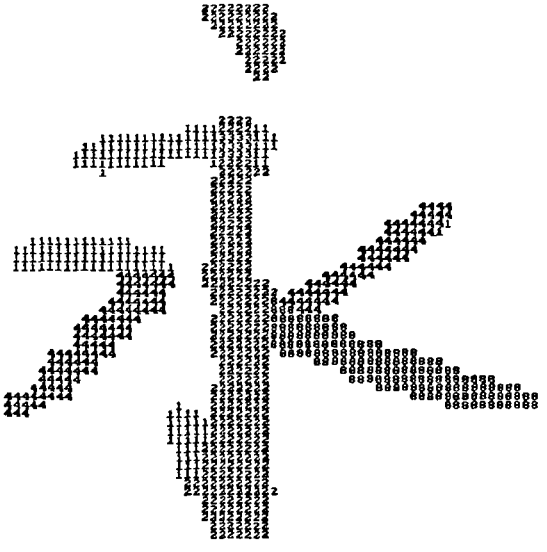


图 4　经处理后的方向画面

2.2　高速方向变换算法

步骤 1　从左→右，从上→下逐行逐列扫描点阵图象．当连续黑点数超过规定阈值时，直接将该线段的点变换成该方向．并逐点记录变换的距离值．

步骤 2　在逐行逐列扫描图象时，记录该行该列的最大最小 x 座标及 y 座标点，仅对这些点进行四方向变换，按最大方向赋予所有该方向的点，并比较以前曾赋予的值，当大于该点以前赋予的值时，按现在的值予以改正，否则不再赋新方向值．

图 3 就是按照上述步骤生成的．

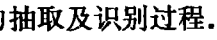
三、偏旁的识别

我们在正确地抽取汉字四周的笔划特征的基础上，还做了以下两项基本工作：

(1) 抽取长横长竖笔划，建立长横、长竖笔划分布表．

(2) 从左上角开始，按反时针方向建立外围笔划排列顺序表．

利用这两张表，根据偏旁所在区域、偏旁的笔划方向，以及相互位置关系的先验知识，将具有同样主笔道特征的偏旁分成一组，所谓主笔道是指该笔道在该偏旁中最为稳定．然后按照规定的顺序，逐一进行分析判断，达到区分近似偏旁的目的．

下面仅以左偏旁中具有长竖笔道的 10 个偏旁  为例，说明偏旁的抽取及识别过程．

如图 5 所示，在长横长竖表中查得第一长竖笔划后，首先判别长竖与左边框间有无笔划存在．当有笔划存在时，接着判有与长笔划相交的笔划否，如果无交点存在，则直接去判别忄与氵偏旁；如果有交点存在，则判是否为扌偏旁，否则以横竖相交为中心，依次判断上、下、左、左上、左下、右下各区域是否存在各偏旁所特有的笔划，达到偏旁识别的目的．

四、结果及其评价

我们利用中科院自动化所提供的汉字样张任意取 3 份共一万余字进行了偏旁抽取试验，结果当按结构分析方法对偏旁进行判断时，累计正确分类率≥96%，而混入其它偏旁的数目最多不超过 5 类，一般仅为 3 类．而采用模板匹配时，侯补偏旁数要取到 10 位，才能保证 90%以上的正确分类率．实践表明将这种方向变换用于抽取汉字的偏旁是成功的．

产生这样结果的原因是显而易见的．因为当采用匹配方法时，匹配距离主要由主笔道决定，而能够区分偏旁的笔道往往为短笔道，它们对距离的影响不大，所以为了取得较高的分类率，必须取多个偏旁，这样当然使混入其它偏旁的概率增加．而采用结构判断的方法却恰好解决了匹配方法所带来的问题．

当然，这种方法也是有局限性的，当笔划增加时，判断条件的变化呈几何级数增长，同时在提取内部笔划时，往往会产生一些令人头痛的伪笔划，完全消除它们有一定的困难．我们认为如果能将这种方法与动态有序弹性匹配方法有机地结合起来，将有可能提高第二级细分类的准确率．特别是在第三级进行逐字判断时，在判别相似字的局部差异时，它的优点会再次表现出来．

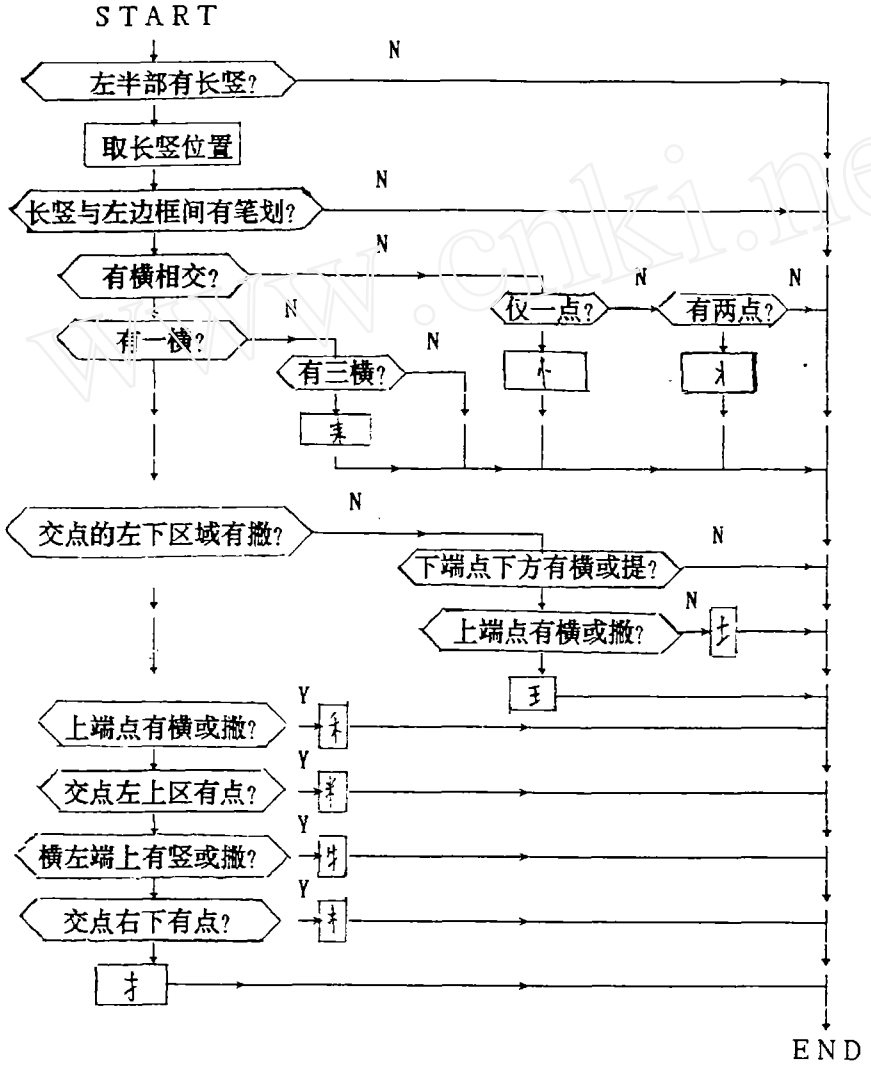


图 5　具有长竖线段的十个左偏旁的判别子程序流程图

(致谢：本研究使用的试验样张由中科院自动化所提供，本项目得到＜863＞的资助，谨致谢意。)

参 考 文 献

[1] 馬場口 登等 構造的セグソント整合による手書き漢字部分パターンの抽出と同定について電子通信学会論文誌 85/3 Vol.J68D No. 3.

[2] 胡家忠等 用动态有序弹性匹配方法识别手写印刷体汉字 武汉工业大学学报 1993.Vol.14.No.4.

Abstract

In the paper, a method for stroke extraction of Chinese characters is presented by means of direction transformation of dot matrix of Chinese characters. With structure analysis, the peripheral rideals are effectively recognized. For 99 different radicals in national standard level 1 Chinese characters, an experiment shows that the correct recognition rate is greater than 96% when candidate number amounts to three.