基于笔画段分割和组合的汉字笔画提取模型

陈 睿 唐 雁 邱玉辉
（西南师范大学计算机与信息科学学院 重庆400715）

A Stroke Extraction Model for Chinese Character

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

Abstract The offline recognition of Chinese Characters is a very important research field in OCR. In this paper, the authors propose an effective stroke extraction model for Chinese Character based on Stroke Segmentation and Combination. Comparing with the existed models, the experiments prove that it could improve the precision for the stroke extraction and reduce the computation complexity considerably.

Keywords OCR, Chinese character recognition, Stroke extraction, Region decomposition, Stroke segmentation

1. 介绍

在过去的几十年中，学术界提出了大量的汉字手写体离线识别技术。这些技术主要可以分为两类，一类是基于整体形态的识别技术，如中心投影变换等；另一类是基于笔画和字根的识别技术。由于汉字的结构复杂，整体形变极易出现，因此，基于笔画和字根的汉字识别技术十分重要。此类技术的关键问题是笔画的分割、提取和匹配。本文在研究已存在的几种典型的笔画提取模型的基础上，提出了一种新型的基于笔画段分割和组合的汉字笔画提取模型。

在 Hung-Hsin Chang 和 Hong Yan 提出的汉字手写体笔画提取技术[1]中，首先将字符扫描成一幅灰度图像，然后根据欧拉距离变换(EDT)将其转换为一幅三维图像，再分析这幅图像中的形态特征，提取出“山峰”、“山脊”和“马鞍”等特征区域。根据这些特征区域对笔画分割进行预处理。在笔画分割前进行了5个步骤：线段近似、星型连接分析、X 型连接分析、T 和 V 连接分析、删除多余笔画。笔画分割中包含合并、分割、线性笔画连接和方向法则4个步骤。这里的每个步骤都要经过复杂的数学计算，从笔画提取的结果来看，结果图与原字符的骨架图非常相似，不同之处是提取出了笔画的方向。

在 Feng Lin 和 Xiaoou Tang 提出的汉字手写体笔画提取技术[2]中，首先提取出字符的骨架图，根据每个字符像素的相交数(crossing number)找出端点、连接点和分叉点，对错误和多余的分叉点建立了一个处理规则。然后根据每个分叉点所连接的笔画的方向建立了一个笔画关系图，应用一个双向连接规则进行笔画提取。由于该技术完全建立在骨架图的基础上，所以该技术能否适应复杂字符的形变还值得研究。

Ruini Cao 和 Chew Lim Tan 提出了一种基于二值图像素区域分解的笔画提取模型[3]。由于这种模型的研究对象是宽笔画字符二值图像，所以它能够充分利用笔画的宽度、连续性和曲率变化等特征。其不足之处在于要对每个字符像素进行运算，时间耗费较高。我们在采用该模型区域分解原理的基础上，提出了一种更有效的基于笔画段分割和组合的汉字笔画提取模型。这里先介绍一下基于二值图像素区域分解的笔画提取模型。

2. 基于二值图像素区域分解的笔画提取模型

这种模型[3]的步骤是：首先根据计算二值图中每个字符像素的点到边界的方向距离(PBOD)将这些像素分为3类：普通点区域、端点区域和分叉点区域。再计算每个字符像素的边界到边界的方向距离(BBOD)找出其切线方向，以此为第三维坐标将每个字符像素映射到一个三维 ρ 空间，利用字符像素在该三维 ρ 空间的分布对相交于分叉点的笔画进行分离。由于字符像素间切线方向变化缓慢，所以该模型提取出的笔画较流畅。下面首先介绍一下二值图像素区域分解。

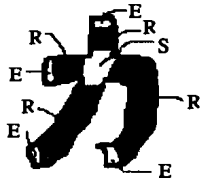


图1 二值图像素区域分解

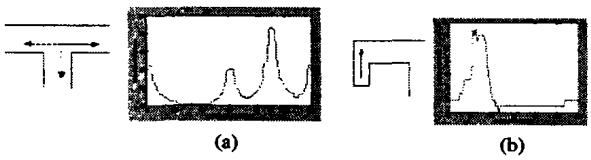


图2 PBOD 的分布曲线

在图1中，E 所指的区域代表笔画端点区域(end point region)，R 所指的区域代表笔画常规点区域(regular point region)，S 所指的区域代表笔画分叉点区域(singular point region)。这里对每个像素进行区域分解的依据就是 PBOD，即点到边界的方向距离(point-to-boundary orientation distance)。对每个像素从0到360度计算每个方向上的 PBOD，再画出 PBOD 的分布曲线。

图2中的 a 和 b 分别代表了一个分叉点和一个端点的 PBOD 分布曲线。这里的角度分辨率为3度。从图中很容易看

陈 睿 硕士生，研究方向：模式识别。唐 雁 副教授，研究方向：模式识别，人工智能。邱玉辉 教授，博导，研究方向：人工智能，模式识别。

出，曲线中波峰的数目就是分类的依据。波峰数为1就是端点，为2就是常规点，大于2就是分叉点。显然，这里的波峰的阈值选择是一个问题。

接下来将分叉点区域映射到一个三维 ρ 空间，通过计算每个字符像素的边界到边界的方向距离(BBOD)找出其切线方向，然后利用切线方向对相交于分叉点的笔画进行分离。

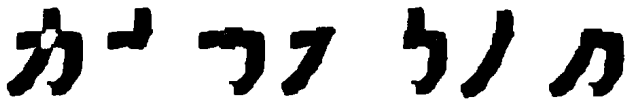
实验证明，用该模型进行笔画提取十分有效。但它也存在一些问题，首先，要计算所有字符像素的 PBOD 曲线才能找出分叉点区域。其次，在 ρ 空间映射时要计算每个字符像素的 BBOD 曲线，时间耗费太高。我们在采用该模型区域分解原理的基础上，提出了一种更有效的基于笔画段分割和组合的汉字笔画提取模型。

3. 基于笔画段分割和组合的汉字笔画提取模型

本模型的大体步骤是：

第一步，分叉点区域提取和笔画段分割。先找出字符图像中的分叉点区域，即图1中的 S 区域。再将第一个分叉点区域去掉，以图1中的“力”字为例，就得到了四个独立的笔画段，如图3a 所示。

第二步，笔画段组合。从这4个笔画段中取出两个与分叉点区域组合，就能够产生笔画提取的效果。在图3中，笔画段组合情况共有 C(4,2)＝6种，其中正确的是(c)和(f)。设分叉点区域的中心为 p，则笔画段组合正确与否的判断标准有两个。第一个标准是计算每个笔画段组合图像中 p 的 PBOD，看其中是否只包含两个波峰，且这两个波峰相距是否接近180度。第二个标准是计算每个笔画段组合图像中 p 的 BBOD，看其是否只有一个波峰。我们只保留笔画段组合正确的情况。



(a) (b) (c) (d) (e) (f) (g)

图3 笔画段分割和组合的情况

第三步，笔画修正。由于分叉点区域可能不规则，所以提取出的笔画可能会有微小的偏差。通过计算提取出的笔画所对应的分叉点区域中每个点的 BBOD 曲线的最大值，删除其中值小于某个阈值的点。下面来看一下这3个步骤的细节。

3.1 分叉点区域提取和笔画段分割

我们认为，并不需要对二值图中的所有字符像素计算其 PBOD 后才能找出分叉点区域，事实上，二值图中的分叉点区域只可能出现在其对应的骨架图的分叉点的一个二维邻域内。这里提取骨架图的细化算法我们采用的是文[4]中的并行算法，因为它能够很好地保留笔画的连贯性。

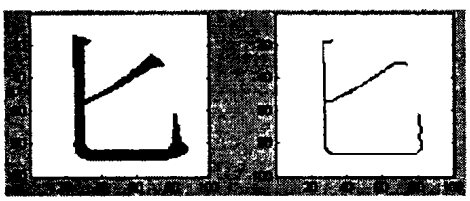


图4 分叉点区域快速提取

首先我们将字符图像规范化为100×100，再提取骨架图。对于骨架图中的分叉点的提取，文[2]中给出了方法，即计算每个字符像素的相交数 $N_c(p)$(crossing number)，

$$N_c(p)=\frac{1}{2}\sum_{i=1}^{8}|x_{i+1}-x_i| \quad \begin{array}{|c|c|c|}\hline x_4 & x_3 & x_2 \\ \hline x_5 & p & x_1 \\ \hline x_6 & x_7 & x_8 \\ \hline \end{array} \quad =x_9$$

其中，$x_i(i=1,\cdots,9)$是像素 p 的邻接点，并且 $x_1=x_9$。若 $N_c(p)>2$，则 p 为分叉点。这样我们就可以找到图4骨架图中的分叉点(54,28)，然后计算图4二值图中点(54,28)的一个半径为4的二维邻域中的点 PBOD 或 BBOD，就可以将分叉点区域提取出来。在实际情况中，骨架图中可能会有几个相距很近的分叉点。我们就将这些分叉点的重心作为新的分叉点。一旦第一个分叉点区域提取出来后，原字符就分为了几个笔画段，这里的笔画段是指原字符图像减去分叉点区域后能够4-连接的部分，如图5所示。

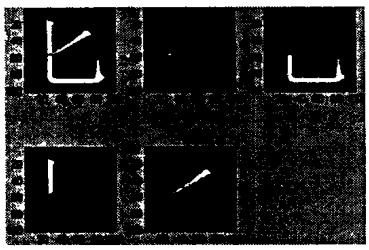


图5 分叉点区域提取和笔画段分割

显然，这里的邻域大小的选择是一个问题，选择过大会计算一些不必要的点，过小会造成笔画段不能完全分割。解决办法有两种，一种是若预先知道笔画的平均宽度，就可以用这个值的一半多一点作为邻域半径。另一种是先用一个较小的值作半径，如上述的“4”，若能成功分割则罢，否则再加大半径，直到能分割成3个或3个以上的笔画段为止。这里的 PBOD 或 BBOD 的峰值的阈值选择也是一个问题，选择过大会漏掉一些小的笔画段，过小会导致分叉点区域过大或者不规则，一般情况下我们对归一化的 PBOD 曲线用0.5作为峰值的阈值。然后我们通过4-连接的方式进行笔画段的分割，这种方式显然比8-连接的分割效果好。

3.2 笔画段组合

在前述的工作中，我们已经将原字符图像分割成分叉点区域和几个笔画段并分别存放成一幅二值图像。笔画段组合就是将两个笔画段图像和分叉点区域图像进行或运算得到一个新的二值图像，再计算分叉点的 PBOD 或 BBOD，看是否满足前述的标准。即看 PBOD 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(在我们的实验中取160～200度这个范围)，或者看 BBOD 曲线中是否只有一个波峰。

3.2.1 *单分叉点区域提取* 一般情况下一个分叉点区域能将字符图像分割成3个或以上的笔画段，如图5所示。再将笔画段按前述标准进行组合就能提取出正确的笔画，如图6所示。

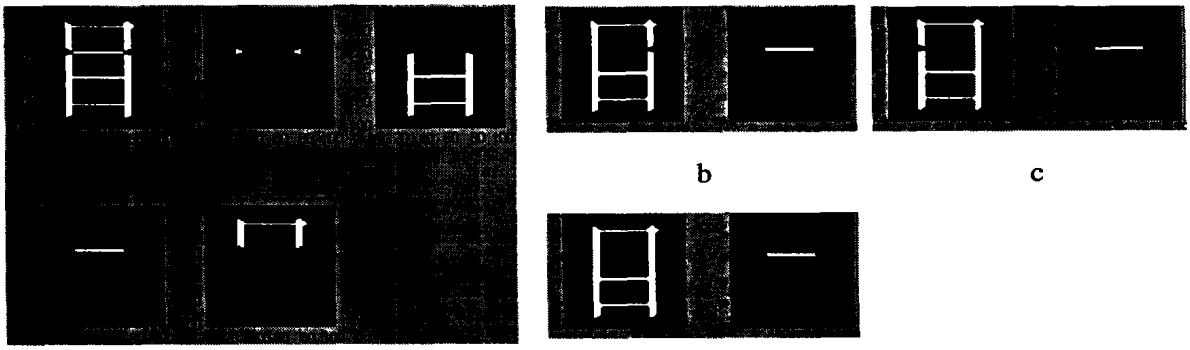
图6 “匕”的笔画提取结果

图7 “目”的二值图和骨架图

图8 错误的分叉点区域组合

3.2.2 双分叉点区域提取 某些情况下一个分叉点区域只能将字符图像分割成少于3个笔画段，如图7所示。

从图7可以看出，“目”字有4个分叉点区域(由于靠近下边界的两个分叉点对应很短的笔画段，所以将它们忽略)，但无论按哪个分叉点区域进行笔画分割，我们都只能得到一个笔画段，再对笔画段组合就毫无意义。这时我们必须考虑分叉点区域的组合，即同时提取两个分叉点区域进行笔画分割。如图8所示，若我们同时对左边两个分叉点区域进行提取，这样笔画段分割结果为3段。但每个分叉点区域都只对应两个笔画段，这样再对笔画段组合同样毫无意义。这时我们考虑分叉点区域的下一种组合，即同时提取上边的两个分叉点区域，这样就把字符图像分成了3个笔画段，如图9a 所示。其中，每个分叉点区域都对应同样的3个笔画段。图9b 和 c 分别代表按左右两个分叉点区域进行笔画段组合的结果，b 的左边和 c 的左边非常相似，而右边则完全相等。这样，我们就对这两个结果进行或运算，得到如图9d 所示的结果。当然，这个结果的左边部分由于包含分叉点区域，可以继续递归前面的步骤进行笔画分割和组合。这样就能够提取出正确的笔画。



a b c d

图9 同时提取两个分叉点区域的笔画分割和组合

3.2.3 三分叉点区域提取 某些特殊情况下双分叉点区域提取也不能在分割后让每个分叉点区域对应3个或以上的笔画段，这时就必须对3个分叉点区域同时提取。如“申”字，共有5个分叉点区域。图10a 显示了一种错误的三分叉点区域提取组合，因为“申”字中心的那个分叉点区域对应的3个笔画段，无论怎样组合都不满足前述的组合标准，即 BBOD 曲线总有两个波峰。图10b 显示了一种正确的三分叉点区域提取组合，图10c～f 显示了笔画段组合的过程，这个过程和图9类似。

a b c d e f

图10 同时提取3个分叉点区域的笔画分割和组合

3.3 笔画修正

由于分叉点区域可能不规则，所以提取出的笔画可能会有微小的偏差，如10a 所示。

a b c

图11 笔画修正

为了解决这个问题，我们首先找出笔画所对应的分叉点区域，如图11(b)所示，设该区域点数为 m，再计算图11a 对应区域中每个点的 BBOD 曲线的最大值，即 $B(p_i)$。设 $M=\max_{i=1}^{m}B(p_i)$，$T=0.7\times M$。实验中我们用 T 作为阈值，该区域中 $B(p_i)<T$ 的点都被删掉。图11c 表示处理的结果，可以看出这种方法非常有效。

4. 对比实验

我们选取了3个有代表性的宋体汉字和一个手写体汉字作为实验对象，先将每个汉字扫描成一幅二值图像，再将其规范化成100×100的大小后进行笔画提取实验。

图12显示了两个宋体字的笔画提取的对比实验，3行笔画从上到下是分别按文[5]和文[3]中的模型及本模型提取。图

13显示了一个手写体字的笔画提取的对比实验。

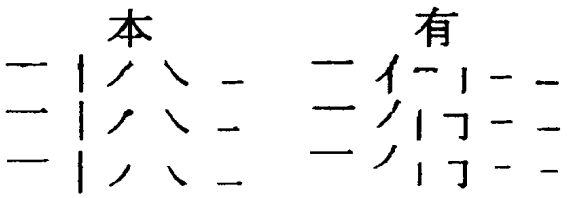


图12 对比实验

用[3]中模型提取的笔画　　用本模型提取的笔画

图13 对比实验

从图12和13可以看出，用本模型进行笔画提取的实验结果不比用文[3]中的模型的结果差，且在某些笔画的提取效果上更好。而本模型和文[3]中模型的提取效果都比文[5]中模型的要好。同时本模型比文[3]中模型的时间耗费大大降低，因为本模型只考虑分叉点区域的点，而文[3]中的模型要考虑字符的所有点，如表1所示。

结论 本文在研究已存在的几种笔画提取模型的基础上，提出了一种新型的基于笔画段分割和组合的汉字笔画提取模型。该模型基于文[3]中宽笔画二值字符图像区域分解的原理，并结合了笔画段分割与组合的新技术。实验证明，与已存在的几种模型相比，该模型在保证笔画提取准确性的前提下，能大大降低时间耗费。从表1可以看出，本模型与文[3]中的模型的时间耗费比都低于20%。同时，该模型应用于汉字印刷体和手写体字符笔画提取都能够达到很好的效果。

表1 时间耗费对比(时间单位为单个像素的PBOD曲线计算时间)

模型	步骤 / 字符	本	有	膀	力
文[3]中的模型	计算所有像素的PBOD曲线	1344	1337	2177	723
	计算所有像素的BBOD曲线	1344	1337	2177	723
	总计	2688	2674	4354	1446
本模型	计算分叉点区域像素的BBOD曲线	204	184	183	85
	笔画段4-连接选择与组合	48	80	96	12
	笔画修正	204	184	183	85
	总计	456	448	464	182
本模型与文[3]中的模型的时间耗费比		16.96%	16.75%	10.66%	12.59%

参考文献

1 Chang H-H, Yan Hong. Analyse of Stroke Structures of Handwritten Chinese Characters. IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS, 1999, 29(1)

2 Lin Feng, Tang Xiaoou. Off-line Handwritten Chinese Character Stroke Extraction

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

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

5 Chen Y S, Hsu W H. An interpretive model of line continuation in human visual perception. pattern recognition, 1989, 22: 619～639

(上接第63页)

联规则的挖掘。因为分类规则与关联规则的不同之一在于规则的结果中(then…)属性的个数。在关联规则中，结果属性可多于一个[11]。该算法还可用于医学临床数据的分析、银行数据的分析、人口数据分析、股市的预测、生物数据的分析、化学实验数据分析、最新的SARS病例数据分析等等。难点在于，对于数量类型的属性需要选择合适子范围来对其进行离散化，这常常需要一些领域知识。

参考文献

1 邢乃宁，孙志挥. 一种基于粗集理论的分类规则挖掘的实现方法. 计算机应用，2001，21(12)：29～31

2 Pawlak Z. Rough set approach to knowledge-based decision support. European Journal of Operational Research, 1997(99): 48～57

3 Pawlak Z. Rough sets and intelligent data analysis. Information sciences, 2002(147): 1～12

4 王国胤. Rough集理论与知识获取. 西安交通大学出版社，2001

5 曾黄麟. 粗糙集理论及其应用—关于数据推理的新方法. 重庆大学出版社，1996

6 张文修，吴伟志，梁吉业等编著. 粗糙集理论与方法. 科学出版社，2003

7 Pawlak Z. Rough classification. Int. J. Human-Computer Studies, 1999(51): 369～383

8 [美]C. L. Liu 著，刘振宏译. 离散数学基础. 人民邮电出版社，1982

9 Walczak B, Massart D L. Rough sets theory, Chemometrics and Intelligent Laboratory System, 1999 (47): 1～16

10 Ziarko W, Golan R, Edwards D. An Application of Datalogic/R Knowledge Discovery Tool to Identify Strong predictive Rules in Stoc Market Data. AAAI-93 Workshop on Knowledge Discovery in Databases, Washington: DC, 1993

11 Freitas A A. A Survey of Evolutionary Algorithm for Data Mining And Knowledge Discovery. http://www.ppgia.pucpr.br～alex