基于多信息融合的中文手写地址字符串切分与识别

付 强　丁晓青　蒋 焰
(智能技术与系统国家重点实验室　清华大学电子工程系　北京　100084)

摘　要：该文提出了一种有效的中文手写地址字符串的切分与识别方法。首先，利用笔划提取与笔划合并将字符串图像进行过切分，得到“字根”图像序列；然后综合利用几何信息、识别信息和语义信息挑选最优的“字根”合并路径，得到最优的切分结果及对应的最优识别结果。其中，几何信息是根据当前字符串自身的特点统计得到，因此可适应不同书写风格的字符串。识别信息由单字分类器给出，包括 10 个候选识别结果及其相应的置信度；单字分类器采用 MQDF 分类器。语义信息用基于字的 bi-gram 模型进行描述，模型参数是从包含 18 万条地址数据的数据库中统计得到的。用 3000 个实际的手写地址样本做试验，单字识别正确率达到 88.28%。

关键词：地址识别；字符串切分；手写字符串识别
中图分类号：TP391　　**文献标识码**：A　　**文章编号**：1009-5896(2008)12-2916-05

Segmentation and Recognition Algorithm for Chinese Handwritten Address Character String

Fu Qiang　Ding Xiao-qing　Jiang Yan
(*State Key Laboratory of Intelligent Technology and Systems*, *Department of Electronic Engineering*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: An effective segmentation and recognition method of Chinese handwritten address strings is proposed. Firstly, over-segmentation is applied to character string images by extracting stroke and merging them to obtain “radical” sequences. Next, the optimal segmentation and recognition result is achieved by synthesizing geometric analysis, isolated character classifier and semantic information together. The geometric information is estimated on current character string to adapt to various writing styles of character strings. The isolated character classifier adopts MQDF classifier with ten candidate results and their confidence. The semantic information is described by a character-based bi-gram model, parameters of which are estimated from a database containing 180,000 addresses items. The algorithm is tested on 3,000 actual handwritten address samples and the single-character recognition rate is 88.28%.

Key words: Address recognition; Character string segmentation; Handwritten character string recognition

1　引言

中文手写地址字符串的切分与识别对于模式识别理论研究以及诸如中文邮政自动分拣等实际应用都有重大意义。手写汉字字符串的切分具有相当的难度，这是因为相邻字符笔划常会粘连或交叠，字符大小、间隔等特征变化很大，不同人书写风格差异显著。此外，手写汉字的首选识别率也比较低，这是由于汉字结构复杂，存在大量相似字且手写体汉字字形变化很大。现有的字符串切分识别算法可分为两类：(1)切分与识别是独立过程，即先基于投影、连通域或轮廓等几何特征分析得到字符串切分结果，再进行识别[1,2]；(2)切分与识别是耦合过程，即识别结果会对切分进行反馈，根据识别的好坏来找到最优切分结果[3−5]。第(1)类算法流程简单，时间复杂度和空间复杂度都较低，但识别率较低。第(2)类算法较复杂，但识别率较高。

算法的整体框图如图1所示：先由笔划提取和笔划合并对原字符串图像进行过切分，得到“字根”图像序列。“字根”图像序列根据不同的合并路径可以得到不同的字符图像序列。这样，切分问题就转换为寻找“字根”图像序列最优合并路径的问题。之后，本文算法整合了几何、识别及语义3种不同类型的信息对不同合并路径进行评估，挑出最优的作为最终切分结果，并同时给出其对应的最优识别结果。其中，在估计字符串几何分布置信度时，仅根据当前字符串中字符图像序列自身的一致性来进行评价，因此能够自适应不同书写风格的字符串。本文的第2节将介绍过切分算法，第3节论述确定最优合并路径及识别结果的算法，第4节及第5节分别是试验结果和结论。

2007-06-15 收到，2007-10-16 改回
国家自然科学基金(60472002)和西门子公司合作项目(20030829-24022SI202)资助课题

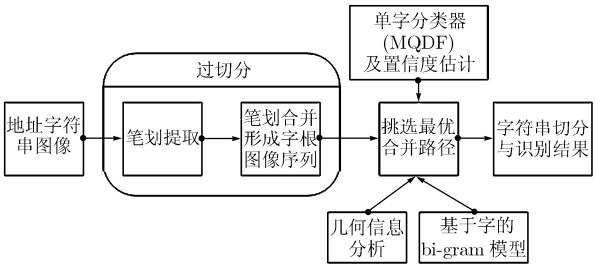


图 1　算法流程图

2　过切分

所谓“字根”，是指汉字中具有紧凑结构的组成部分。过切分的目的是把字符串图像过切分为字根图像序列，要求每个字根图像完全属于某一个字符。在实际的信封图像中，地址字符串相邻字符间常会出现笔划粘连、交叉等情况。传统的基于连通域分析或轮廓分析的算法进行字根提取效果都不理想。现有一些方法是先进行笔划提取，再进行笔划合并，得到字根序列，实现过切分[6−8]。这类方法可较好地解决相邻字符间笔划粘连与交叉等问题，因此本文也采取类似算法。

2.1 笔划提取

笔划主要分为 4 类，即横、竖、撇、捺。利用行方向黑游程的跟踪算法提取笔划。逐行找到图像中所有的黑游程，然后按一定的准则将其合并为笔划。这个准则可以概括为两点：(1)进行合并的游程必须连通，且游程长度差小于一个阈值；(2)游程合并之后所形成区域的方向必须在一定程度上保持恒定，即区域的方向不能有大的转折(笔划都是直线形)。笔划提取算法的细节可参考文献[6]。

2.2 笔划合并为字根

假设地址字符串按照横排方向书写。合并过程如下：设 S_i 表示第 i 个图像块，SR_i 表示它的外接矩形，$\mathrm{SR}_i.\mathrm{left}$ 和 $\mathrm{SR}_i.\mathrm{right}$ 分别表示外接矩形左右边界的横坐标。对图像块序列从左向右进行扫描，若两个块 S_i 和 S_j 符合式(1)，则进行合并。其中，$l_{i,j}$ 表示 SR_i 和 SR_j 在横坐标上的交叠；若 SR_i 和 SR_j 无交叠，则 $l_{i,j}$ 等于 0；T 是一个阈值，此时取 0.95。对此次合并后形成的新图像块序列再次从左向右进行扫描，我们仍用 S_i 表示此新序列的第 i 个图像块，依然根据式(1)对图像块进行合并，不同的只是 T 取 0.7。类似地再次扫描，T 取 0.5。之后再次扫描，T 取 0.3。经过这 4 次合并，笔划序列合并为字根序列。其中 T 的取值是根据试验效果确定的。在第 4 节的算法结果示例中，含有字符串过切分为字根序列的例子，如图 2 中(a-2)及(b-2)所示。

$$\frac{l_{i,j}}{\mathrm{SR}_i.\mathrm{right}-\mathrm{SR}_i.\mathrm{left}}>T \quad \text{或}$$

$$\frac{l_{i,j}}{\mathrm{SR}_j.\mathrm{right}-\mathrm{SR}_j.\mathrm{left}}>T \tag{1}$$

3　基于多信息融合的切分与识别算法

所谓合并路径，是指字根序列合并为字符序列的方式。字符串切分结果，即字符图像序列，可通过字根图像序列按一定的合并路径合并得到。因此字符串的切分问题转化为字根序列合并路径的选择问题。由于所有可能的字根合并路径数量很大，为 $O(2^{N-1})$，其中 N 是字根总个数；为降低运算复杂度，本文先根据简单的评价算法初步选择 L 条合并路径作为候选合并路径(本文取 L=100)，然后再使用本文算法评价每一条候选合并路径，并从中找到最优合并路径及相应的最优识别结果。候选路径的选择方法可参考文献[8,9]。

设过切分后得到 N 个字根图像，RI_i 表示第 i 个字根图像，$1\le i\le N$。S 表示任意一条候选合并路径。用 M^S 表示在合并路径 S 下得到的字符图像个数，$M^S\le N$；用 CI_j^S 表示其中的第 j 个字符图像，C_j^S 表示 CI_j^S 的识别结果，$1\le j\le M^S$。根据最大后验概率准则，最优切分路径及相应的最优识别结果可表述为式(2)。根据条件概率公式，式(2)右边部分可写成式(3)，其中 $p(\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\mid \mathrm{RI}_1,\cdots,\mathrm{RI}_N)$ 的意义为合并路径 S 对应的字符图像序列 $\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S$ 的几何分布置信度；$p(C_1^S,\cdots,C_{M^S}^S\mid \mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S)$ 表示合并路径 S 对应的字符图像序列 $\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S$ 识别为 $C_1^S,\cdots,C_{M^S}^S$ 的识别置信度。下面分别论述如何计算它们。

$$\begin{aligned}&\hat{S},\hat{C}_1^{\hat{S}},\hat{C}_2^{\hat{S}},\ldots,\hat{C}_{M_{\hat{S}}}^{\hat{S}}\\&\quad=\arg\max_{S,C_1^S,\cdots,C_{M^S}^S} p(S,C_1^S,\cdots,C_{M^S}^S\mid \mathrm{RI}_1,\mathrm{RI}_2,\cdots,\mathrm{RI}_N)\end{aligned}\tag{2}$$

$$\begin{aligned}&p(S,C_1^S,\cdots,C_{M^S}^S\mid \mathrm{RI}_1,\mathrm{RI}_2,\cdots,\mathrm{RI}_N)\\&=p(S\mid \mathrm{RI}_1,\cdots,\mathrm{RI}_N)\times p(C_1^S,\cdots,C_{M^S}^S\mid \mathrm{RI}_1,\mathrm{RI}_2,\cdots,\mathrm{RI}_N,S)\\&=p(\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\mid \mathrm{RI}_1,\cdots,\mathrm{RI}_N)\\&\quad\times p(C_1^S,\cdots,C_{M^S}^S\mid \mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S)\end{aligned}\tag{3}$$

3.1 单字分类器及单字识别置信度估计

本文中采取改进的二次分类函数 (Modified Quadratic Discriminant Function，MQDF)作为单字分类器。其原理如下：设各个类别的先验概率相同，且各类样本均为高斯分布，那么根据贝叶斯准则可推导得到二次鉴别函数(Quadratic Discriminant Function，QDF)为最优分类器。对于一个 d 维的特征向量 $\boldsymbol{x}$，可根据式(4)计算得到 QDF 分类器的识别距离[10]。其中 C 是类别总数，$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Psi}_i$ 分别代表类 ω_i 的均值向量和协方差矩阵，$g_i(\boldsymbol{x})$表示 $\boldsymbol{x}$ 到类 ω_i 的距离。$\boldsymbol{\mu}_i$ 和 $\boldsymbol{\Psi}_i$ 均是根据最大似然估计得到的。当样本数量有限时，协方差矩阵的估计误差成为使分类器性能劣化的一个重要因素。为此，Kimura 等提出了 MQDF 来减小协方差矩阵估计误差带来的影响[11]。对 $\boldsymbol{\Psi}_i$ 进行正交分解，并将那些小特征值用一个常量 σ^2 替换，以补偿由于小样本集带来的估计误差，则 MQDF 的识别距离可由式(5)计算得到。其中 λ_{ij} 和 $\boldsymbol{\varphi}_{ij}$ 分别

代表 $\boldsymbol{\Psi}_i$ 的第 j 个特征值(按照从大到小排序)以及这个特征值对应的特征向量。q 代表主轴个数，$q<d$。将特征向量 $\boldsymbol{x}$ 到所有类的识别距离按照升序排列，如式(6)所示，其中 $R_j(\boldsymbol{x})$ 表示 $\boldsymbol{x}$ 的第 j 候选，$g_{R_j(\boldsymbol{x})}(\boldsymbol{x})$表示其对应的识别距离。设定单字分类器给出前 10 个候选识别结果，以便在字符串识别中能应用语言模型。对于前 10 个侯选的识别结果 $R_j(\boldsymbol{x})$，其识别置信度可依据式(7)近似计算得到；对于 10 个候选之外的识别结果，近似认为它的识别置信度为 0。

$$g_i(\boldsymbol{x})=(\boldsymbol{x}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Psi}_i^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_i)+\log\left|\boldsymbol{\Psi}_i\right|,\quad i=1,2,\cdots,C \tag{4}$$

$$g_i(\boldsymbol{x})=\frac{1}{\sigma^2}\left\{\left\|\boldsymbol{x}-\boldsymbol{\mu}_i\right\|^2-\sum_{j=1}^{q}\left(1-\frac{\sigma^2}{\lambda_{ij}}\right)\left[\boldsymbol{\varphi}_{ij}^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu}_i)\right]^2\right\}+\sum_{j=1}^{q}\log\lambda_{ij}+(d-q)\log\sigma^2,\quad i=1,2,\cdots,C \tag{5}$$

$$g_{R_j(\boldsymbol{x})}(\boldsymbol{x})\le g_{R_{j+1}(\boldsymbol{x})}(\boldsymbol{x}),R_j(\boldsymbol{x})\in\{1,2,\cdots,C\},\quad j=1,2,\cdots,C \tag{6}$$

$$p(R_j(\boldsymbol{x})\mid\boldsymbol{x})=\frac{p(R_j(\boldsymbol{x}))\times\exp(-g_{R_j(\boldsymbol{x})}(\boldsymbol{x})/2)}{\sum_{i=1}^{10}p(R_i(\boldsymbol{x}))\times\exp(-g_{R_i(\boldsymbol{x})}(\boldsymbol{x})/2)},\quad j=1,2,\cdots,10 \tag{7}$$

3.2 图像序列几何分布置信度

设字符图像序列的几何分布置信度与其识别信息及语义信息无关，因此可仅由字符图像的几何特征来计算几何分布置信度。以字符图像的外接矩形(x, y, w, r)来表征字符图像的几何信息，其中，x, y, w, r 分别表示字符图像外接矩形的中心点横、纵坐标及外接矩形的宽度和高宽比。则几何分布置信度可由式(8)表达。假设字符图像外接矩形的中心坐标与其形状大小无关，则式(8)可化简为式(9)。再假设 x 与 y 无关，w 与 r 无关，则可得式(10)。由于中文地址均是从左向右按行书写，各字符外接矩形中心的纵坐标基本相同，因此我们忽略纵坐标这一次要因素，进一步得到式(11)。

$$p(\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\mid\mathrm{RI}_1,\cdots,\mathrm{RI}_N)=p\left(\left(x_1^S,y_1^S,w_1^S,r_1^S\right),\cdots,\left(x_{M^S}^S,y_{M^S}^S,w_{M^S}^S,r_{M^S}^S\right)\mid\mathrm{RI}_1,\cdots,\mathrm{RI}_N\right) \tag{8}$$

$$\propto p\left(\left(x_1^S,y_1^S\right),\cdots,\left(x_{M^S}^S,y_{M^S}^S\right)\right)\times p\left(\left(w_1^S,r_1^S\right),\cdots,\left(w_{M^S}^S,r_{M^S}^S\right)\right) \tag{9}$$

$$=p\left(x_1^S,\cdots,x_{M^S}^S\right)\times p\left(y_1^S,\cdots,y_{M^S}^S\right)\times p\left(w_1^S,\cdots,w_{M^S}^S\right)\times p\left(r_1^S,\cdots,r_{M^S}^S\right) \tag{10}$$

$$\approx p\left(x_1^S,\cdots,x_{M^S}^S\right)\times p\left(w_1^S,\cdots,w_{M^S}^S\right)\times p\left(r_1^S,\cdots,r_{M^S}^S\right) \tag{11}$$

设 $d_i^S=x_{i+1}^S-x_i^S$，并假设几何分布置信度与具体的起始坐标无关，则可得到式(12)。假设 d_i^S $(1\le i\le M^S-1)$为独立抽样，并满足高斯分布。这个高斯分布的均值和方差可以根据最大似然准则估计出来，如式(13)和式(14)所示。然后，$p\left(d_1^S,\cdots,d_{M^S-1}^S\right)$ 可根据式(15)计算。对 w 和 r 做类似的推导，可得式(16)和式(17)。最后，根据式(15)−式(17)及式(11)，可以得到合并路径 S 的几何分布置信度为式(18)所示。从推导可以看出，本文算法仅从字符图像序列自身的一致性来评价其几何分布置信度，而对诸如字符宽度、字符宽高比、及字符间距等不作先验约束。这使得字符图像序列几何分布置信度的估计算法可以适应不同的书写风格。

$$p\left(x_1^S,\cdots,x_{M^S}^S\right)=p\left(d_1^S,\cdots,d_{M^S-1}^S\right) \tag{12}$$

$$\mu_d^S=\frac{1}{M^S-1}\sum_{i=1}^{M^S-1}d_i^S \tag{13}$$

$$\sigma_d^S=\sqrt{\frac{1}{M^S-1}\sum_{i=1}^{M^S-1}\left(d_i^S-\mu_d^S\right)^2} \tag{14}$$

$$p\left(d_1^S,\cdots,d_{M^S-1}^S\right)=\prod_{i=1}^{M^S-1}\frac{1}{\sqrt{2\pi}\times\sigma_d^S}\exp\left[-\frac{\left(d_i^S-\mu_d^S\right)^2}{2\times\left(\sigma_d^S\right)^2}\right] \tag{15}$$

$$p\left(w_1^S,\cdots,w_{M^S}^S\right)=\prod_{i=1}^{M^S}\frac{1}{\sqrt{2\pi}\times\sigma_w^S}\exp\left[-\frac{\left(w_i^S-\mu_w^S\right)^2}{2\times\left(\sigma_w^S\right)^2}\right] \tag{16}$$

$$p\left(r_1^S,\cdots,r_{M^S}^S\right)=\prod_{i=1}^{M^S}\frac{1}{\sqrt{2\pi}\times\sigma_r^S}\exp\left[-\frac{\left(r_i^S-\mu_r^S\right)^2}{2\times\left(\sigma_r^S\right)^2}\right] \tag{17}$$

$$\begin{aligned}p\left(S\mid\mathrm{RI}_1,\cdots,\mathrm{RI}_N\right)&=p\left(\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\mid\mathrm{RI}_1,\cdots,\mathrm{RI}_N\right)\\&\propto\left\{\prod_{i=1}^{M^S-1}\frac{1}{\sqrt{2\pi}\times\sigma_d^S}\exp\left[-\frac{(d_i^S-\mu_d^S)^2}{2\times(\sigma_d^S)^2}\right]\right\}\\&\times\left\{\prod_{i=1}^{M^S}\frac{1}{\sqrt{2\pi}\times\sigma_w^S}\exp\left[-\frac{(w_i^S-\mu_w^S)^2}{2\times(\sigma_w^S)^2}\right]\right\}\\&\times\left\{\prod_{i=1}^{M^S}\frac{1}{\sqrt{2\pi}\times\sigma_r^S}\exp\left[-\frac{(r_i^S-\mu_r^S)^2}{2\times(\sigma_r^S)^2}\right]\right\}\end{aligned} \tag{18}$$

3.3 图像序列识别置信度

这节介绍如何计算合并路径 S 对应的识别置信度 $p\left(C_1^S,\cdots,C_{M^S}^S\mid\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right)$。根据条件概率公式，可以推导得到式(19)。本文假设：一个字符图像的识别结果仅仅与它自身的图像以及前一个字符图像的识别结果有关系，这可用式(20)来表示。将式(20)代入式(19)，可得到式(21)。现在再做另外两个假设，如式(22)及式(23)所示。它们的含义是假设字符图像只与当前字符有关，而与之前的字符无关。根据式(22)和式(23)的假设，可以得到式(24)。根据式(21)和式(24)，图像序列的识别置信度可由式(25)计算得到。在式(25)中，$p(C_i^S\mid\mathrm{CI}_i^S)$ 表示将图像 CI_i^S 识别为 C_i^S 的置信度，可根据式(7)计算得到；$p(C_i^S)$ 和 $p(C_i^S\mid C_{i-1}^S)$ 分别表示 bi-gram 模型中字符的先验概率和二元转移概率，它们可通过地址数据库统计得到的。本文中使用的地址数据库包含约 18 万条地址条目。

$$p\left(C_1^S,\cdots,C_{M^S}^S\mid\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right)=p\left(C_1^S\mid\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right)\times\prod_{i=2}^{M^S}p\left(C_i^S\mid C_{i-1}^S,\cdots,C_1^S,\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right) \tag{19}$$

$$p\left(C_i^S \mid C_{i-1}^S,\cdots,C_1^S,\mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right) = p\left(C_i^S \mid C_{i-1}^S,\mathrm{CI}_i^S\right),\quad 2\le i\le M^S \tag{20}$$

$$p\left(C_1^S,\cdots,C_{M^S}^S \mid \mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S\right) = p\left(C_1^S \mid \mathrm{CI}_1^S\right)\times\prod_{i=2}^{M^S} p\left(C_i^S \mid C_{i-1}^S,\mathrm{CI}_i^S\right) \tag{21}$$

$$p(\mathrm{CI}_i^S) = p(\mathrm{CI}_i^S \mid C_{i-1}^S),\quad 2\le i\le M^S \tag{22}$$

$$p(\mathrm{CI}_i^S \mid C_i^S) = p(\mathrm{CI}_i^S \mid C_i^S, C_{i-1}^S),\quad 2\le i\le M^S \tag{23}$$

$$p(C_i^S \mid C_{i-1}^S,\mathrm{CI}_i^S) = \frac{p(C_i^S \mid C_{i-1}^S)\times p(\mathrm{CI}_i^S \mid C_i^S, C_{i-1}^S)}{p(\mathrm{CI}_i^S \mid C_{i-1}^S)} = \frac{p(C_i^S \mid C_{i-1}^S)\times p(C_i^S \mid \mathrm{CI}_i^S)}{p(C_i^S)},\quad 2\le i\le M^S \tag{24}$$

$$p(C_1^S,\cdots,C_{M^S}^S \mid \mathrm{CI}_1^S,\cdots,\mathrm{CI}_{M^S}^S) = \frac{\left[p(C_1^S)\times\prod_{i=2}^{M^S} p(C_i^S \mid C_{i-1}^S)\right]\times\left[\prod_{i=1}^{M^S} p(C_i^S \mid \mathrm{CI}_i^S)\right]}{\prod_{i=1}^{M^S} p(C_i^S)} \tag{25}$$

综上，本文先通过切分路径的初步选择，得到 L 条候选切分路径。然后再对每一条候选切分路径进行精细评估，找出最好的那条切分路径作为最终的切分结果，此最优切分路径对应的识别结果作为最终的识别结果。在精细评估中，综合利用了单字识别信息、几何信息及语言模型，如 3.1 至节 3.3 节所述。具体来说，对于每一条确定的切分路径，都可以根据它得到合并后的字符图像序列。然后应用单字分类器，则每一个字符图像都可得到 10 个候选识别结果及相应的识别置信度。然后应用 Veterbi 算法从候选识别结果中挑出使得式(25)最大的识别结果序列 $C_1,\cdots,C_M$，并将此作为此切分路径对应的识别结果，相应地以式(25)计算得到的值作为此切分路径的识别分数。然后根据式(18)计算此切分路径的几何分数。将式(25)的识别分数与式(18)的几何分数求积，得到此切分路径的总分数。然后从 100 条候选切分路径中找到总分数最高的作为最终结果。当然，我们还可以对各切分路径的总分数进行排序，并输出多个候选切分和识别结果。

4　试验结果及示例

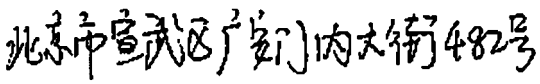
我们实际收集了 3000 个手写中文地址信封，提取其中的收件人地址字符串进行识别。识别错误由两方面原因造成，一方面是切分引起的识别错误，另一方面是切分正确但单字分类器识别错误。本文统计了识别结果的单字识别正确率，切分引起的识别错误率及单字分类器引起的识别错误率，如表 1 所示。进一步分析实验结果可知，有一半以上的切分错误是由地址字符串中所含的数字区域引起的。这是因为数字在大小、间距等方面的特性与汉字有较大区别，数字的大小及间隔往往会小一些。例如数字“1”很窄，易被合并到其它字符中而造成切分错误。对数字区域进行再处理是我们下一步的工作。图 2 给出了一些示例，包括字符串图像、过切分结果、切分结果及识别结果。我们在主频为 3.2G 赫兹，内存 1G 的电脑上运行此算法程序，处理一条字符串平均耗时约 0.9 秒。

表 1　试验结果

地址字符串个数	总字符数	单字识别正确率	切分引起的错误率	分类器引起的错误率	总错误率
3000	40792	88.28 %	7.66 %	4.06 %	11.72%

5　结束语

试验结果证明了本文提出的算法对于中文手写地址字符串的切分与识别是非常有效的。其中鲁棒的过切分算法是整个算法的必要基础。本文采取笔划提取与合并的方法进行过切分，有效地解决了字符交叠、粘连等问题，达到了较满意的效果。综合单字分类器、几何模型及语言模型这 3 种信息共同实现最优切分路径及识别结果的选择是算法的关键，因为充分利用了多种信息，才使算法最终达到了较高的识别



(a-1) 输入手写地址字符串图像

(a-2) 过切分后形成字根图像序列

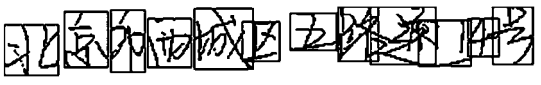
(a-3) 最终的切分结果

北京市宣武区广安门内大街482号

(a-4) 最终的识别结果

(b-1) 输入手写地址字符串图像

(b-2) 过切分后形成字根图像序列



(b-3) 最终的切分结果

北京市西城区五路进口4号

(b-4) 最终的识别结果

图 2　算法结果示例

性能。采用“先根据粗评估挑选候选切分路径集合，再进一步应用精细评估得到最终结果”这一策略是算法运算效率的保障，这是因为精细评估算法的运算复杂度较高，无法针对每一条可能的切分路径都使用。综上，本文提出了一种切实有效的处理中文手写地址字符串的算法。对算法的分析可知，此算法也可应用于一般类型的中文手写字符串的识别，只是应用的语言模型参数不同而已。

参 考 文 献

[1] Chiang C C and Yu S S. An iterative character segmentation method for irregularly formatted Chinese documents. Proceedings of the Optical Character Recognition and Document Analysis, Taiwan, 1996: 61–67.

[2] Lu Y and Shridhar M. Character segmentation in handwritten words-an overview. *Pattern Recognition*, 1996, 29(1): 77–96.

[3] Arica N and Yarman-Vural F T. An overview of character recognition focused on off-line handwriting. *IEEE Trans. on Systems, Man, and Cybernetics-Part C: Applications and Reviews*, 2001, 31(2): 216–233.

[4] Casey R G and Lecolinet E. A survey of methods and strategies in character segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1996, 18(7): 690–706.

[5] Liu C L, Koga M and Fujisawa H. Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2002, 24(11): 1425–1437.

[6] Tseng L Y and Chuang C T. An efficient knowledge based stroke extraction method for multi-font Chinese characters. *Pattern Recognition*, 1992, 25(12): 1445–1458.

[7] Tseng L Y and Chen R C. Segmenting handwritten Chinese characters based on heuristic merging of stroke bounding boxes and dynamic programming. *Pattern Recognition Letters*, 1998, 19(10): 963–973.

[8] 王崾, 丁晓青, 刘长松. 基于笔划合并的手写体信函地址汉字切分识别. 清华大学学报(自然科学版), 2004, 44(4): 498–502.
Wang R, Ding X Q and Liu C S. Handwritten Chinese address segmentation and recognition based on merging strokes. *J of Tsinghua Univ.* (*Sci & Tech*), 2004, 44(4): 498–502.

[9] Fu Q, Ding X Q, and Liu C S, *et al*.. A hiddern Markov model based segmentation and recognition algorithm for Chinese handwritten address character strings. International Conference on Document Analysis and Recognition, Seoul, Korea, 2005: 590–594.

[10] Duda R O, Hart P E and Stork D G. Pattern Classification. Second Edition, New York, John Wiley & Sons Inc, 2000: 36–45.

[11] Kimura F, Takashina K, and Tsuruoka S, *et al*.. Modified quadratic discriminant functions and its application to Chinese character recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1987, 9(1): 149–153.

付 强： 男，1980 年生，博士生，研究方向为模式识别、图像处理、智能图文信息处理.

丁晓青： 女，1939 年生，教授，博士生导师，电子学会高级会员，中国通信学会会士，研究方向为智能图文信息处理、模式识别、图像处理、文字识别、多媒体信息处理以及视频智能检测.

蒋 焰： 男，1980 年生，博士生，研究方向为模式识别、智能图文信息处理.