

一种脱机手写汉字书写顺序恢复模型

曹忠升 苏哲文 王元珍

(华中科技大学计算机科学与技术学院, 武汉 430074)

摘要 书写顺序恢复是从静态文本图像中提取动态的字符书写顺序信息,将2维的图像转换为1维的书写位置的时间序列的过程。为了对手写汉字进行书写顺序提取,提出了一种脱机手写汉字书写顺序的恢复模型。该模型首先将汉字分为整字、部件、子部件和笔画4个层次;然后利用4种拆分操作将整字拆分为部件,再将部件拆分为子部件;最后通过定义一组拆分关系与子部件偏序关系之间的对应规则来得到子部件的全序关系。而将子部件作为最基本的恢复单位,其书写顺序可通过对笔画和交叉笔画对进行分类来得到。实验表明,该模型提出的汉字书写顺序恢复方法的恢复结果具有较高的准确率,且处理速度达到了6.9字/s。

关键词 笔迹鉴别 层次模型 时序信息 笔画恢复

中图法分类号: TP391.43 文献标识码: A 文章编号: 1006-8961(2009)10-2074-08

A Model for Recovering Writing Order from Offline Handwritten Chinese Character

CAO Zhong-sheng, SU Zhe-wen, WANG Yuan-zhen

(College of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074)

Abstract The objective of recovering writing order is to extract dynamic handwriting information from static text images, which can be seen as to convert a 2-dimensional image into sequences of 1-dimensional vectors of pen tip positions along the time axis. This paper proposes a model for recovering writing order from offline handwritten Chinese character. In this model, a 4-layer hierarchy is presented to model each Chinese character, where character, component, subcomponent and stroke are located at each layer, respectively. Characters are decomposed into components and each component is decomposed to subcomponents in turn by four decomposing operators. The totally-ordered relations between subcomponents are retrieved by defining the corresponding rules between decomposed relations and a poset of subcomponents. Subcomponents are the basic recovering primitives in this model, whose writing orders are recovered by classifying strokes and pairs of crossing strokes. Experimental results show that the proposed method is effective and accurate.

Keywords writer identification, hierarchical model, temporal information, stroke recovery

1 引言

手写汉字笔迹鉴别的目的是为了区别不同的书写者,相对于文字识别来说,其提取的特征需要强调不同的书写者之间的差异性。根据输入设备的差异,笔迹鉴别可以分为联机鉴别和脱机鉴别两大类^[1]。相对于联机的动态书写序列来说,脱机的文本图像丢失了书写过程的时序信息,但是时序信息

却能在很大程度上反映书写者的个体特征。

书写顺序恢复就是指从静态文本图像中提取动态的书写顺序信息,即将2维的文本图像转换为1维的书写过程中,笔尖移动位置的时间序列^[2]。目前,关于书写顺序恢复方法的文献^[1-6]主要可以分为基于图搜索、基于模板匹配和基于书写规则3种。其中基于图搜索的恢复方法^[1-4]通常是利用图模型来对细化后的输入图像进行建模,从而将书写顺序恢复转换为在图上搜索一条(单笔画)或多条(多笔

收稿日期:2007-12-04; 改回日期:2008-09-11

第一作者简介:曹忠升(1965~),男,副教授。1993年于华中科技大学控制科学与工程系获得生物医学工程专业博士学位。主要研究方向为多媒体数据库、空间数据库、移动数据库、数据仓库和数据开采。E-mail:caozhongsheng@163.com

画)路径,使该路径满足目标函数的最大化或最小化的问题^[1]。该方法的好处是简单、易于实现,但是搜索路径的过程则需要花费较大的计算代价,并且目标函数的设计一般是基于最小能量代价的准则(minimum energy principle)^[1],该准则适用于数字和西字母的书写顺序恢复,但是用于汉字的书写顺序恢复却并不准确。

基于模板匹配的方法则是首先对每个笔画或字符建立一个模板用来描述其形状,然后通过比较输入图像和模板来达到书写顺序恢复的目的^[6]。该方法的优点是适用范围广、准确率高,但是在比较输入图像和模板的过程中,需要搜索每个模板笔画的最佳匹配路径,其计算代价甚至超过了基于图搜索的方法,并且对于字符数量较多的文字来说,建立一个全面、通用的模板库比较困难。

基于书写规则的恢复方法是首先利用字符的结构特点来有效地表达笔画和笔画之间的关系,然后通过设计一组规则来判断笔画之间的先后顺序。这种方法避免了搜索造成的组合爆炸,其缺点是不易设计通用的规则,以适应书写风格的变化。除此之外,文献[7]提出了一种统计模型,用于脱机签名的书写顺序的恢复。该模型利用联机动态信息进行训练,虽然能够适应不同的书写风格,但是不适用于非连笔的文本。

在上述方法中,大多数能够比较有效地用于数字^[3]、西文^[1-2]或阿拉伯文^[4]的书写顺序恢复,但是,对于手写汉字来说,由于其字量多、结构复杂、书写风格变化大,因此目前还没有一种有效的方法用于手写汉字的书写顺序恢复。

与其他文字不同,汉字是一种非常强调结构性、层次性和规则性的文字,这可以通过层次结构和一组形成规则来反映;而另一方面,由于不同书写者的书写风格的差异又带来书写过程的随机性和任意性,这可以利用统计的方法来解决。因此,本文利用层次模型来表示手写汉字,并结合基于规则和基于统计的方法,提出了一种脱机手写汉字书写顺序的恢复模型。该模型先将汉字分为整字、部件、子部件和笔画4个层次;然后采用上下、左右、包围和间架4种拆分操作将整字拆分为部件,再将部件拆分为子部件,从而构成上下、左右、左中右、包围和间架5种拆分关系;接着,定义了一组拆分关系与子部件偏序关系之间的对应规则,用于该偏序集的拓扑排序,从而得到子部件全序关系。而子部件作为最基本的

恢复单位,其书写顺序则是通过对笔画和交叉笔画对进行分类来得到。本文是笔画提取的后续工作,为了保证本文的方法不会受到笔画提取算法的影响,本文假定笔画已正确提取。

2 汉字的层次模型

众所周知,汉字的结构具有很强的层次性,在汉字识别中,普遍采用层次拆分的方法进行结构分析^[8-9]。在文献[8]中,Shi等人先将汉字拆分为一个部件的集合,然后通过部件匹配和部件组合来识别汉字;Kang和Kim在文献[9]中提出了一个层次随机关系模型,用于韩文的识别。从结构上,汉字通常可以分为整字、部件和笔画3个层次^[8-9],其中笔画是汉字结构的最小单位。笔画之间的相互关系是分析汉字结构的基础,可分为相离(ap),相接(tc)和相交(cr)3种,其中相离是指笔画之间相互分离,如“川”字,相接是指笔画直接接触但不相交,如“上”字;相交是指笔画与笔画相交,如“十”字。

鉴于恢复书写顺序的目的,本文采用4层层次模型来对汉字进行建模(如图1所示)。

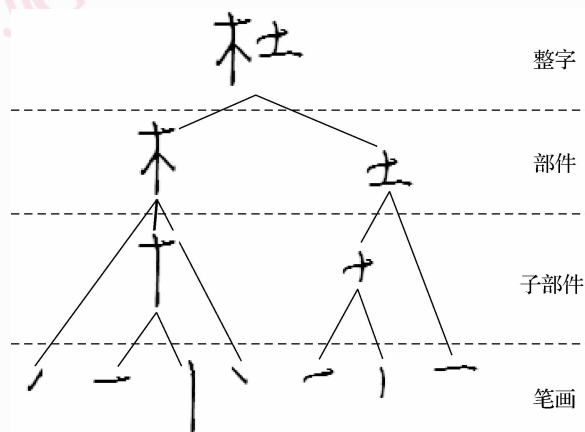


图1 手写汉字“杜”的层次模型

Fig. 1 The hierarchical model for handwritten Chinese characters “Du”

在分析模型之前,先给出以下几个定义:

定义1 部件是一个笔画的集合,是一个独立的8连通区域。一些文献把部件定义为具有独立功能的字根或字元^[8-11],而在本文中,仅把部件作为一个连通区域。一个连通区域中可能包含多个字根,一个字根也可能被划分到多个连通区域中。如图2(b)所示,手写汉字“铜”分别由 c_1, c_2, c_3, c_4 和

c_5 这 5 个部件构成。

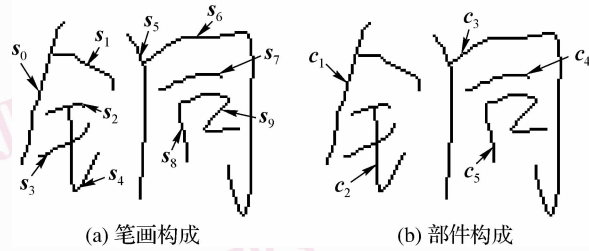


图 2 汉字“铜”的层次结构

Fig. 2 The hierarchical structure of Chinese character “Tong”

定义 2 子部件是一个笔画集合,是组成部件的单位。对于任意笔画集合 b_i ,若 b_i 为子部件,则 b_i 必满足以下条件:

(1) $|b_i| = 1$; 或者,

(2) $|b_i| > 1$,且对于任意笔画 $s_j \in b_i, s_k \in b_i (j \neq k)$,存在 $s_0, s_1, \dots, s_m \in b_i (m \geq 1)$,使得 $s_0 = s_j, s_m = s_k$ 并且 $s_p \text{ cr } s_{p+1} (0 \leq p < m)$ 。如图 2(a) 所示,笔画 s_3 和 s_4 构成了一个子部件,其余每个笔画构成一个子部件。

若采用无向图对字符图像进行建模,将字符上的每条笔画对应为图上的一个节点,同时笔画间的相交关系对应为图上的边,那么每个子部件则可以看作是該无向图上的一个极大连通子图。同样地,若将笔画间的相交和相接关系对应图上的边,那么该无向图上的每个极大连通子图则为一个部件。

定义 3 对于两个任意笔画 s_0 和 s_1 ,且 $s_0 \text{ cr } s_1$,如果在书写顺序上 s_0 先于 s_1 ,那么 s_1 是 s_0 的主交笔画, s_0 是 s_1 的被交笔画。

模型的第 1 层是整字,其由第 2 层的部件组成,而且同一个部件中笔画的书写顺序通常都是连续的。一般地,整字是由一个或多个部件通过一定的规则组合而成,而部件又是由一个或多个的子部件组合而成,汉字的组合规则反映了汉字的书写顺序。模型将整字从大到小逐层拆分为子部件,先通过定义拆分关系和子部件偏序关系之间的对应规则来得到子部件的一个偏序集,然后对偏序集进行拓扑排序,从而得到子部件的一个全序。

模型的第 3 层是子部件,子部件是由单个笔画或多个笔画相交构成,是本文模型中最基本的书写顺序恢复单位。本文采用统计的方法来构造分类器,先通过选取一定的训练数据产生分类规则;然后通过对相交笔画进行分类来得到笔画的一个偏序

集;最后采用图模型对子部件进行建模,并通过将图模型转换为拓扑有序来得到子部件中的笔画全序关系。

模型的第 4 层是笔画,是汉字结构的基本组成成分。笔画的书写顺序恢复是从笔画的两个端点中确定起始端点和终止端点,这样就把问题分为两类,以便于采用分类的方法解决。

3 拆分操作

Sun 等人在文献[10]中提出了一种汉字数学表达方法,将汉字的部件作为操作数,并将定义的 6 种位置关系作为运算符,极大地简化了汉字的信息处理。在此基础上,本文定义了 4 种汉字拆分操作及其优先级。通过拆分形成的部件或子部件之间的 5 种拆分关系如表 1 所示。

表 1 拆分操作与拆分关系

Tab. 1 Decomposing operators and decomposed relations

拆分操作	优先级	拆分关系	例子
间架	高	间架(三元)	坐
包围	中	包围(二元)	回
左右	低	左右(二元)	张
		左中右(三元)	挪
上下	低	上下(二元)	台

3.1 包围拆分

对于给定的部件或子部件集合 C_i ,存在非空真子集 $C_0^{(i)} \subset C_i, C_1^{(i)} = C_i - C_0^{(i)}$,若满足以下条件,则 C_i 可包围拆分为 $C_0^{(i)}$ 和 $C_1^{(i)}$,表示为 $\text{SR}(C_i, C_1^{(i)}, C_0^{(i)})$:

(1) $\frac{A_1}{A_0} > T_{\text{area}}$,其中, A_0 为 $C_0^{(i)}$ 的外接矩形面积, A_1 为 $C_0^{(i)}$ 的外接矩形和 $C_1^{(i)}$ 的外接矩形重叠部分的面积, T_{area} 为面积阈值 ($0.5 < T_{\text{area}} \leq 1$);

(2) 不存在非空子集 $C_{10}^{(i)} \subset C_1^{(i)}, C_{11}^{(i)} \subset C_1^{(i)}$,

$C_{10}^{(i)} \cap C_{11}^{(i)} = \emptyset$,使 $\frac{A_{11}}{A_{10}} > T_{\text{area}}$,其中 A_{10} 为 $C_{10}^{(i)}$ 的外接矩形面积, A_{11} 为 $C_{10}^{(i)}$ 的外接矩形和 $C_{11}^{(i)}$ 的外接矩形重叠部分的面积;

(3) 不存在非空真子集 $C_{10}^{(i)} \subset C_1^{(i)}$,使 $\frac{A_{10}}{A_0} > T_{\text{area}}$,其中 A_{10} 为 $C_{10}^{(i)}$ 的外接矩形和 $C_{10}^{(i)}$ 的外接矩形重叠部分的面积。

同时 $C_0^{(i)}$ 和 $C_1^{(i)}$ 构成包围关系,可表示为 $C_1^{(i)} \text{ sr } C_0^{(i)}$ 。如图 2(b) 所示, $\{C_3\} \text{ sr } \{C_4, C_5\}$ 。

根据 $C_0^{(i)}$ 和 $C_1^{(i)}$ 的相对位置,本文将包围结构分为如表 2 所示的 9 种类型(其中右三包围结构在汉字的结构中并不存在,本文增加这种类型是为了分析间架结构),包围结构类型的判断过程如下:

设 $C_1^{(i)}$ sr $C_0^{(i)}$, $C_1^{(i)}$ 和 $C_0^{(i)}$ 的外接矩形如图 3(a)所示, A, B, C, D 分别为形成 $C_1^{(i)}$ 的外接矩形的 4 个边界区域(如图 3(b)所示)。对每个包围结构分配一个 4 bits 的二进制串,每一位对应一个边界区域。对于 A, C 区域来说,若 $C_1^{(i)}$ 中的笔画像素在该区域中的点数大于矩形宽度,则置 1,否则,置 0;对于 B, D 区域来说,若 $C_1^{(i)}$ 中的笔画像素在该区域中的点数大于矩形高度,则置 1,否则,置 0。由二进制串就可得到包围结构的类型,其对应关系见表 2。

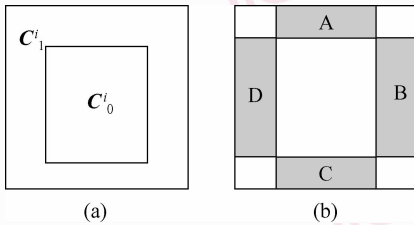


图 3 $C_0^{(i)}$ 和 $C_1^{(i)}$ 的相对位置

Fig. 3 The relative positions of $C_0^{(i)}$ and $C_1^{(i)}$

表 2 包围类型与二进制串的对应关系

Tab. 2 The correspondence between the surround type and the binary string

边界区域 二进制串 ABCD	包围类型	边界区域 二进制串 ABCD	包围类型
		1110	右三
1100	上右(如“句”)	1011	左三(如“匡”)
0110	下右(如“头”)	1101	上三(如“同”)
1001	上左(如“厄”)	0111	下三(如“凶”)
0011	下左(如“建”)	1111	全包围(如“回”)

3.2 间架拆分

对于给定的部件或子部件集合 C_i ,若存在非空真子集 $C_0^{(i)}$, $C_1^{(i)}$ 和 $C_2^{(i)}$, $C_0^{(i)} \cup C_1^{(i)} \cup C_2^{(i)} = C_i$, $C_0^{(i)} \cap C_1^{(i)} = \emptyset$, $C_0^{(i)} \cap C_2^{(i)} = \emptyset$, $C_1^{(i)} \cap C_2^{(i)} = \emptyset$ (不妨设 $C_0^{(i)}$ 的质心在 $C_2^{(i)}$ 的质心的左侧),且满足以下条件,则 C_i 可间架拆分为 $C_0^{(i)}$, $C_1^{(i)}$ 和 $C_2^{(i)}$,并表示为 $SP(C_i, C_0^{(i)}, C_1^{(i)}, C_2^{(i)})$:

(1) $C_1^{(i)}$ sr $C_0^{(i)}$ 为下左包围, $C_1^{(i)}$ sr $C_2^{(i)}$ 为下右包围;或者,

(2) $C_1^{(i)}$ sr $C_0^{(i)}$ 为左三包围, $C_1^{(i)}$ sr $C_2^{(i)}$ 为右三包围;或者,

(3) $C_1^{(i)}$ sr $C_0^{(i)}$ 为上左包围, $C_1^{(i)}$ sr $C_2^{(i)}$ 为上右包围。

同时 $C_0^{(i)}$, $C_1^{(i)}$ 和 $C_2^{(i)}$ 构成间架关系,可表示为 $C_1^{(i)}$ sp ($C_0^{(i)}, C_2^{(i)}$)。

3.3 上下或左右拆分

对于给定的部件或子部件集合 C_i ($|C_i| > 1$),令 $C_p^{(i)}$ 表示 C_i 的任意一个非空真子集, $C_p^{(i)}$ 的质心坐标表示为 $(x_c(C_p^{(i)}), y_c(C_p^{(i)}))$,并且

$$\Delta x_c(C_p^{(i)}) = |x_c(C_p^{(i)}) - x_c(C_i - C_p^{(i)})| \quad (1)$$

$$\Delta y_c(C_p^{(i)}) = |y_c(C_p^{(i)}) - y_c(C_i - C_p^{(i)})| \quad (2)$$

则存在 C_i 的两个非空真子集 $C_j^{(i)}$ 和 $C_k^{(i)}$,并满足

$$\Delta x_c(C_j^{(i)}) = \max_{C_j^{(i)} \subset C_i} (\Delta x_c(C_p^{(i)})) \quad (3)$$

$$\Delta y_c(C_k^{(i)}) = \max_{C_k^{(i)} \subset C_i} (\Delta y_c(C_p^{(i)})) \quad (4)$$

如果 $\Delta x_c(C_j^{(i)}) > \Delta y_c(C_k^{(i)})$,则 C_i 可左右拆分为 $C_j^{(i)}$ 和 $C_i - C_j^{(i)}$,可表示为 $LR(C_i, C_j^{(i)}, C_i - C_j^{(i)})$;

(1) 若 $x_c(C_j^{(i)}) < x_c(C_i - C_j^{(i)})$,则 $C_j^{(i)}$ 和 $C_i - C_j^{(i)}$ 构成左右关系,可表示为 $C_j^{(i)}$ lr $C_i - C_j^{(i)}$;

(2) 否则, $C_i - C_j^{(i)}$ lr $C_j^{(i)}$ 。

否则, C_i 上下拆分为 $C_k^{(i)}$ 和 $C_i - C_k^{(i)}$,可表示为 $UD(C_i, C_k^{(i)}, C_i - C_k^{(i)})$;

(3) 若 $y_c(C_k^{(i)}) < y_c(C_i - C_k^{(i)})$,则 $C_k^{(i)}$ 和 $C_i - C_k^{(i)}$ 构成上下关系,可表示为 $C_k^{(i)}$ ud $C_i - C_k^{(i)}$;

(4) 否则, $C_i - C_k^{(i)}$ ud $C_k^{(i)}$ 。

如图 2(a)所示, $\{C_1, C_2\}$ lr $\{C_3, C_4, C_5\}$, $\{C_1\}$ ud $\{C_2\}$ 。

定义 4 给定部件或子部件集合 C_0 和 C_1 ,且 C_0 lr C_1 :

(1) 若 $LR(C_0, C_2, C_3)$ 且 C_1 不满足左右拆分条件,则 C_2, C_3 和 C_1 构成左中右并列的三元关系,可表示为 $lmr(C_2, C_3, C_1)$;

(2) 同样地,若 $LR(C_1, C_2, C_3)$,且 C_0 不满足左右拆分,则 $lmr(C_0, C_2, C_3)$;

4 汉字拆分

根据拆分对象,汉字拆分可以分为整字拆分和部件拆分两个部分。拆分时,利用拆分操作,先将整字拆分为部件,再将部件拆分为子部件。

整字拆分是指根据汉字的结构关系,将整字拆分

为部件的递归过程,其中主要是相离拆分。如图 2(b)所示,手写汉字“铜”由 c_1, c_2, c_3, c_4 和 c_5 5 个部件组成,其整字拆分过程如图 4 所示。经过整字拆分,整字可以表达为部件和拆分关系的树形结构,其中叶子节点为部件,非叶子节点为拆分关系(如图 5 所示)。

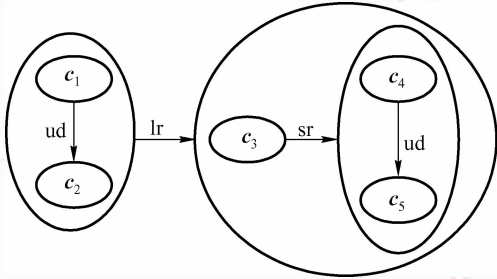


图 4 汉字“铜”的整字拆分过程

Fig. 4 The decomposing process of Chinese character “Tong”

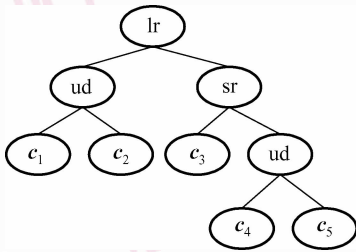


图 5 汉字“铜”的树形表达

Fig. 5 The tree structure representation of Chinese character “Tong”

部件拆分是指根据子部件间的结构关系,将部件拆分为子部件的递归过程,其中主要是相接拆分。如图 6(a)所示,手写汉字“用”共有 s_0, s_1, s_2, s_3 和 s_4 5 个笔画,这 5 个笔画形成了 $\{s_0\}, \{s_2\}$ 和 $\{s_1, s_3, s_4\}$ 3 个子部件,分别用 b_0, b_1 和 b_2 表示。

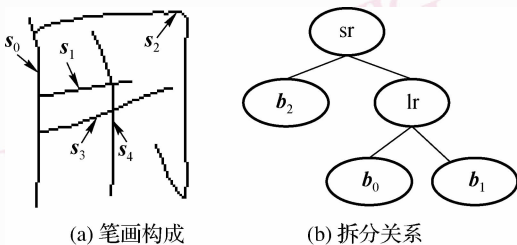


图 6 汉字“用”的部件拆分

Fig. 6 The component decomposition of Chinese character “Yong”

应用上述 4 种拆分操作,就可以对 $\{b_0, b_1, b_2\}$ 做如下拆分:

$$(1) SR(\{b_0, b_1, b_2\}, \{b_0, b_1\}, \{b_2\});$$

$$(2) LR(\{b_0, b_1\}, \{b_0\}, \{b_1\});$$

b_0, b_1 和 b_2 之间的拆分关系如图 6(b)所示。

5 书写顺序恢复

根据模型的层次结构,本文将书写顺序的恢复分为笔画的书写顺序恢复、子部件的笔画全序关系恢复以及整字的子部件全序关系恢复 3 个部分。

5.1 笔画的书写顺序恢复

标准汉字中的笔画共有 26 种,其中某些笔画之间的差异很小,可以将其视为一种类型,文献[12]将汉字笔画分为 17 种类型。尽管大多数汉字笔画符合“从左到右、从上到下”的基本书写规则,但是也存在像“撇”和“提”这两种骨架接近,但书写顺序相反的笔画。

事实上,笔画的书写顺序恢复只需要从笔画的两个端点中确定起始点和终止点即可实现笔画书写顺序的恢复。这样先把问题分为两类,然后就可比较容易地采用分类的方法提取笔画书写顺序。

本文采用构造贝叶斯分类器的方法来判断笔画的书写顺序。对于任意的笔画 s_i ,设 s_i 由 $n+1$ 个点构成,分别为 $(p_0^{(i)}, p_1^{(i)}, \dots, p_n^{(i)})$,其中点 $p_0^{(i)}$ 和 $p_n^{(i)}$ 为笔画 s_i 的端点。在前期的笔画提取工作中,由于是采用跟踪笔画轮廓两侧对应点的方法来细化笔画,因此很容易得到笔画上每一点的笔画宽度。设点 $p_k^{(i)}$ ($0 \leq k \leq n$) 的坐标为 $(x_{p_k^{(i)}}, y_{p_k^{(i)}})$,其对应的笔画宽度为 $w_{k,i}$ 。

本文用一个 3 维的特征向量 $H = \{H_0, H_1, H_2\}$ 来描述 s_i ,其中 H_0, H_1 和 H_2 分别是对特征属性 h_0, h_1 和 h_2 的度量, h_0 为 $x_{p_n^{(i)}} - x_{p_0^{(i)}}$, h_1 为 $y_{p_n^{(i)}} - y_{p_0^{(i)}}$, h_2 为

$$\frac{1}{(n+1)/2} \sum_{j=\lfloor (n+1)/2 \rfloor}^n w_{j,i} - \frac{1}{(n+1)/2} \sum_{j=0}^{\lfloor (n+1)/2 \rfloor - 1} w_{j,i}$$

而分类器的类别属性为 {正序, 逆序}, 分别表示其书写顺序是从点 $p_0^{(i)}$ 到点 $p_n^{(i)}$ 和从点 $p_n^{(i)}$ 到点 $p_0^{(i)}$ 。

5.2 子部件的笔画全序关系

对于任意的子部件 b_i ,可用一个无向图 $G = (V, E)$ 对其进行建模,其中 V 是 G 中所有节点的集合,而任意的节点 $v \in V$ 则对应了 b_i 中一条笔画; E 是 G 中所有边的集合,对于任意的节点 $v_j, v_k \in V$,若其对应的笔画存在相交关系,则有 $(v_j, v_k) \in E$ 。由子部件的定义可知, G 必是连通的。

定义 5 对于任意的笔画 s_0 和 s_1 ($s_0 \neq s_1$),若

s_0 的书写顺序先于 s_1 , 则 s_0 和 s_1 存在偏序关系, 记为 $s_0 > s_1$; 否则, $s_1 > s_0$ 。

定义 6 对于任意的笔画 s_0 和 s_1 ($s_0 \neq s_1$), 若在书写顺序上, s_0 的下一条笔画为 s_1 , 则 s_0 和 s_1 存在全序关系, 记为 $s_0 \rightarrow s_1$ 。

定义 7 笔段是笔画上无明显曲率变化的线段, 笔画是由笔段串联构成。通过对笔画上的像素点进行断点检测^[13], 即可以得到笔画上的所有笔段。

定义 8 对于任意的子部件 b_0 和 b_1 ($b_0 \neq b_1$), 若 $\forall s_i^0 \in b_0, \forall s_j^1 \in b_1$, 有 $s_i^0 > s_j^1$, 那么 b_0 和 b_1 就存在偏序关系, 可表示为 $b_0 > b_1$ 。

为了获得子部件中笔画全序关系, 本文首先对相交笔画对进行分析, 以确定其中的主交笔画和被交笔画。同样地, 可采用构造分类器的方法对每个相交笔画对进行分类。

对于 $\forall e_i = (v_j, v_k) \in E$, 设与节点 v_j 和 v_k 对应的笔画分别为 s_j 和 s_k , s_j 和 s_k 的交点所在的笔段分别为 g_j 和 g_k 。按 s_j 和 s_k 的书写顺序, 不妨设 g_j 和 g_k 的前一个笔段和后一个笔段分别为 g_{j-1}, g_{j+1} 和 g_{k-1}, g_{k+1} 。通过将 $g_{j-1}, g_j, g_{j+1}, g_{k-1}, g_k$ 和 g_{k+1} 的方向向量作为特征, 构造一个 6 维的特征向量, 其类别属性为 $\{s_j > s_k, s_k > s_j\}$, 即可构造一个贝叶斯分类器。

通过相交笔画对的分类, 就可以得到一个笔画的偏序集, 但是该偏序集并不一定拓扑有序。为了得到笔画的全序关系, 还需要进行如下处理:

(1) 将 G 转换为有向图。将偏序集中的偏序关系应用到 E 中的每条边上, 并将分类的概率作为边的权重。

(2) 将 G 转换为有向无环图。若 G 中有环存在, 说明偏序集中出现相悖, 则将环中权重最小的边去掉。

(3) 将 G 转换为单侧连通。若在 G 上存在路径 $P_0 = (v_0^0, v_1^0, \dots, v_n^0), P_1 = (v_0^1, v_1^1, \dots, v_m^1)$, 使得节点 $v_i^0 \neq v_j^1$ ($0 \leq i \leq n, 0 \leq j \leq m$), v_i^0 和 v_j^1 不可达, 且 P_0 和 P_1 是满足该条件的最长路径, 那么与 $\{v_0^0, v_1^0, \dots, v_n^0, v_1^1, \dots, v_m^1\}$ 对应的笔画就构成了两个部件或子部件; 然后利用拆分操作对其进行进一步的拆分, 由拆分关系就可以得到节点 v_i^0 和 v_j^1 的偏序关系 (对应关系见 5.3 节); 最后将该偏序关系应用到 G 中。

基于汉字的书写常识, 可以假定任意两条笔画间的交点数都小于 2。由于 G 为有向无环图, 且单侧连通, 故 G 是拓扑有序的。

5.3 整字的子部件全序关系

经过层次拆分, 整字可以由子部件和 5 种拆分

关系表示为树形结构 (如图 7 所示)。基于汉字的书写规则, 本文定义了以下 5 种拆分关系和子部件偏序关系的对应规则:

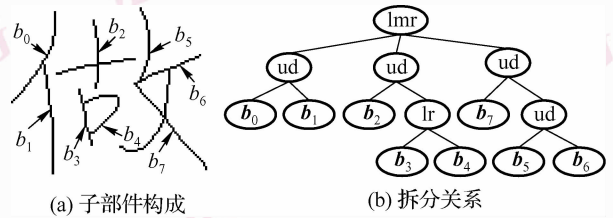


图 7 汉字“做”的子部件的拆分关系

Fig. 7 The composed relations of sub-components of Chinese character “Zuo”

规则 1 对于任意的子部件集合 B_0 和 B_1 , 若 B_0 ud B_1 , 那么, 对于 $\forall b_i^0 \in B_0, \forall b_j^1 \in B_1$, 则有 $b_i^0 > b_j^1$ 。

规则 2 对于任意的子部件集合 B_0 和 B_1 , 若 B_0 lr B_1 , 那么, 对于 $\forall b_i^0 \in B_0, \forall b_j^1 \in B_1$, 则有 $b_i^0 > b_j^1$ 。

规则 3 对于任意的子部件集合 B_0, B_1 和 B_2 , 有 $lmr(B_0, B_1, B_2)$, 设 (g_0, g_1, \dots, g_n) 为 B_1 上的笔段序列, β_i 为笔段 g_i 的方向向量 ($1 \leq i \leq n$),

(1) 如果存在 m ($0 < m < n$) 个笔段, 使得 $\beta_j \in (\pi/2 + T_\beta, 3\pi/2 - T_\beta)$, 并且 $|B_0| \leq 2$ 和 $|B_2| \leq 2$ (其中, $n - m \leq j \leq n, 0 < T_\beta < \pi/2$ 是方向向量 β_i 的阈值, 在本文的实验中, $T_\beta = 0.4$), 则对于 $\forall b_i^0 \in B_0, \forall b_j^1 \in B_1, \forall b_k^2 \in B_2$, 有 $b_j^1 > b_i^0 > b_k^2$;

(2) 否则, $b_i^0 > b_j^1 > b_k^2$ 。

规则 4 对于任意的子部件集合 B_0 和 B_1 , 有 B_0 sr B_1 , 那么

(1) 若包围的类型为上右、上左、下左、上三, 则对于 $\forall b_i^0 \in B_0, \forall b_j^1 \in B_1$, 有 $b_i^0 > b_j^1$;

(2) 若包围的类型为下右或下三, 则 $b_j^1 > b_i^0$;

若将 B_0 分为 3 个子集 B_0^0, B_1^0 和 B_2^0 , 则对于 $\forall b \in B_0$, 有

① 如果 b 的质心在 C 区域中 (如图 3 (b) 所示), 则 $b \in B_0^0$;

② 如果 b 的质心在 D 区域中, 则 $b \in B_1^0$;

③ 否则, $b \in B_2^0$ 。

(3) 若包围类型为左三, 则对于 $\forall b_i^0 \in B_0^0 \cup B_1^0, \forall b_j^0 \in B_2^0, \forall b_k^1 \in B_1$, 有 $b_j^0 > b_k^1 > b_i^0$;

(4) 若包围类型为全包围, 则对于 $\forall b_i^0 \in B_0^0, \forall b_j^0 \in B_1^0 \cup B_2^0, \forall b_k^1 \in B_1$, 有 $b_j^0 > b_k^1 > b_i^0$ 。

规则 5 对于任意的子部件集合 B_0, B_1 和 B_2 , 有 $B_0 \text{ sp}(B_1, B_2)$, 那么

(1) 若 $B_0 \text{ sr } B_1$ 为下左包围, $B_0 \text{ sr } B_2$ 为下右包围, 则对于 $\forall b_i^0 \in B_0, \forall b_j^1 \in B_1, \forall b_k^2 \in B_2$, 有 $b_j^1 > b_i^0 > b_k^2$;

(2) 若 $B_0 \text{ sr } B_1$ 为上左包围, $B_0 \text{ sr } B_2$ 为上右包围, 则 $b_i^0 > b_j^1 > b_k^2$;

(3) 若 $B_0 \text{ sr } B_1$ 为左三包围, $B_0 \text{ sr } B_2$ 为右三包围, 则对于 $\forall b_i^0 \in B_0, \forall b_j^0 \in B_0, \forall b_k^1 \in B_1, \forall b_p^2 \in B_2$, 有 $b_j^0 > b_k^1 > b_p^2 > b_i^0$.

(其中, B_0^0 和 B_1^0 是 B_0 的两个非空真子集, 对于 $\forall b \in B_0$, 如果 b 的质心在 C 区域中, 则 $b \in B_0^0$, 否则, $b \in B_1^0$)

由此就可以得到子部件的一个偏序集, 由整字的树形结构可知, 由该偏序集必可以得到子部件的一个全序。图 7 中子部件的全序关系为 $b_0 \rightarrow b_1 \rightarrow b_2 \rightarrow b_3 \rightarrow b_4 \rightarrow b_5 \rightarrow b_6 \rightarrow b_7$ 。

6 实验结果

为了验证本文恢复模型的有效性, 笔者收集了来自 20 个书写者的 200 幅手写汉字图像用于本文

的实验。每个样本图像的大小为 100×100 pixels。实验平台为 Core 2 1.66 GHz, 1 GB 内存的微机。根据本文模型的层次结构, 可将实验分为 4 个部分分别用于测试各层的恢复性能。为了保证各层测试的独立性, 可假定在下一层的结果正确的前提下, 再进行上一层的测试。

用于实验的 200 幅手写汉字图像中, 共包含笔画 1 564 条, 交叉笔画对 324 个。在笔画书写顺序恢复的实验中, 本文选取了其中 500 条笔画用于该层的贝叶斯分类器的训练, 其余的 1 064 条笔画用于测试。同样地, 在子部件书写顺序恢复的实验中, 则选取了 100 个交叉笔画对用于该层贝叶斯分类器的训练, 其余的 224 个交叉笔画对用于测试。各层的书写顺序恢复的准确率比较如表 3 所示, 平均处理速度为 6.9 字/s。

表 3 各层书写顺序恢复准确率比较

Tab. 3 Comparison of the percentage of successful recovery among four layers of the proposed model

	笔画	子部件	部件	整字
准确率 (%)	98.97	94.2	97.5	98

图 8(a) 为分别来自不同书写者的 3 个样本书

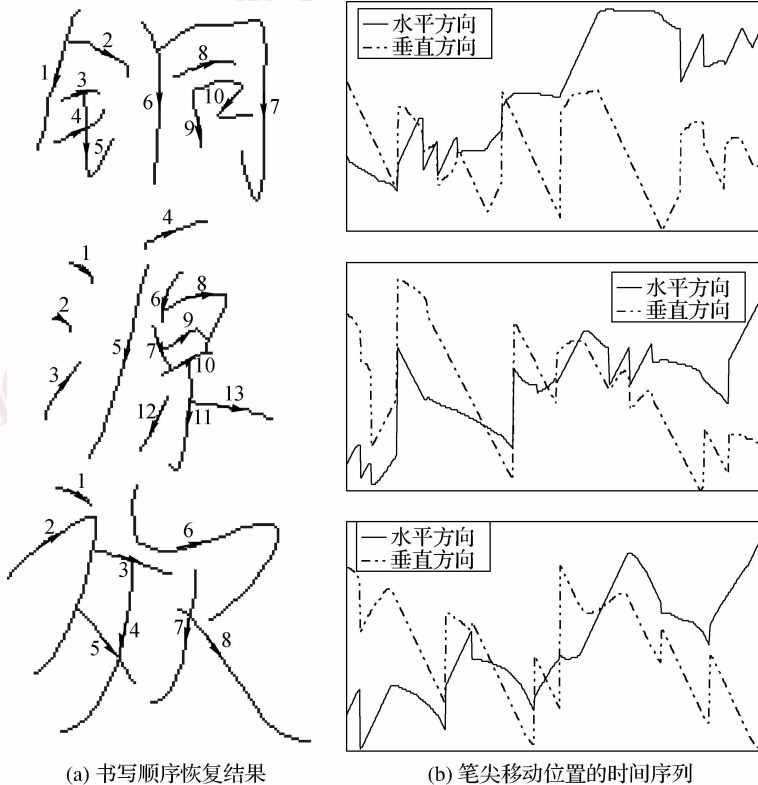


图 8 实验结果

Fig. 8 The experimental results

写顺序的恢复结果,其书写顺序如图中箭头和数字所示。图8(b)显示了对应的书写顺序产生的笔尖移动位置的时间序列,其中实线表示水平方向上书写位置的时间序列,而虚线则表示垂直方向上书写位置的时间序列。

7 结 论

本文提出了一种脱机手写汉字书写顺序恢复模型,该模型将汉字分为整字、部件、子部件和笔画4个层次。恢复时,先采用上下、左右、包围和间架4种拆分操作将整字拆分为部件,再将部件拆分为子部件,形成上下、左右、左中右、包围和间架这5种拆分关系;然后通过定义一组拆分关系与子部件偏序关系的对应规则来得到子部件的全序关系,并通过对笔画和交叉笔画对进行分类来实现子部件的书写顺序恢复。

实验表明,本文提出的汉字书写顺序恢复模型方法对大多数手写汉字样本的书写顺序恢复具有较高的准确率和处理速度,但是在一些连笔较多的汉字(如签名)的测试中也出现了一些错误,其主要是由于笔画的连续造成了汉字原有结构的破坏,这将在以后的工作中对此进行改进。

参考文献 (References)

- 1 Qiao Yu, Yasuhara Makoto. Recovering drawing order from offline handwritten image using direction context and optimal Euler path [A]. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing[C], Toulouse, France, 2006: 765-768.
- 2 Qiao Yu, Yasuhara Makoto. Recovering dynamic information from static handwritten images [A]. In: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition[C], Tokyo, Japan, 2004: 118-123.
- 3 Li Guo-hong, Shi Peng-fei. An approach to the reconstruction of handwriting traces for off-line handwritten numerals[J]. Journal of Shanghai Jiao tong University, 2005, 39(4): 561-564. [李国宏,

- 施鹏飞. 离线手写体数字笔迹重构方法[J]. 上海交通大学学报. 2005, 39(4): 561-564.]
- 4 Baati A E, Charfi M, Alimi A M, *et al.* Recovery of temporal information from off-line Arabic handwritten[A]. In: Proceedings of the 3rd ACS/IEEE International Conference on Computer Systems and Applications[C], Cairo, Egypt, 2005: 127-132.
- 5 Lee Sukhan, Pan J C. Offline tracing and representation of signatures [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1992, 22(4): 755-771.
- 6 Qiao Yu, Yasuhara Makoto. Recover writing trajectory from multiple stroked image using bidirectional dynamic search[A]. In: Proceedings of the 18th International Conference on Pattern Recognition[C], Hong Kong, China, 2006: 970-973.
- 7 Lau Kai-kwong, Yuen Pong-chi, Tang Yuan-yan. Universal writing model for recovery of writing sequence of static handwriting images [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2005, 19(5): 1-27.
- 8 Shi Da-ming, Damper R I, Gunn S R. Offline handwritten Chinese character recognition by radical decomposition[J]. ACM Transactions on Asian Language Information Processing, 2003, 2(1): 27-28.
- 9 Kang Kyung-won, Kim J H. Utilization of hierarchical, stochastic relationship modeling for Hangeul character recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004: 26(9): 1185-1196.
- 10 Sun Xin-ming, Chen Huo-wang, Yang Li-hua, *et al.* Mathematical representation of a Chinese character and its applications [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2002, 16(8): 735-748.
- 11 Zhang Wen-yin, Sun Xing-ming, Zeng Zhen-bing, *et al.* Automatic generation of mathematical expression of Chinese characters [J]. Journal of Computer Research and Development, 2004, 41(5): 848-852. [张问银, 孙星明, 曾振柄等. 汉字数学表达式的自动生成[J]. 计算机研究与发展, 2004, 41(5): 848-852.]
- 12 Liu Xia-bi, Jia Yun-de. Chinese character structure models for handwritten Chinese character recognition [J]. Journal of Beijing Institute of Technology. 2003, 23(3): 322-325. [刘峡壁, 贾云得. 用于手写体汉字识别的汉字结构模型[J]. 北京理工大学学报, 2003, 23(3): 322-325.]
- 13 Han J H. Detection of convex and concave discontinuous points in a plane curve[A]. In: Proceedings of the 3rd International Conference on Computer Vision[C], Osaka, Japan, 1990: 71-74.