

一种手写汉字拓扑图表示及其动态获取*)

周昌乐

(杭州大学计算机科学系 杭州 310028)

TP391.41

摘 要 In this paper, we give out a topo-graph representation and its computation for recognizing handwritten Chinese characters. The representation describes not only the topological features, but also the statistical features in the handwritten Chinese characters' shape. By applying the topo-graph representation to practical recognition system, we have obtained quite excellent results.

关键词 HOCR, representation method, feature abstracting, topo-graph

手写汉字的机器识别,属于图像模式分类问题。所谓图像模式分类,指的是把一定范围内的图像,分成预先确定的类别,然后再去对给定范围内的图像进行识别分类。显然,这其中预先确定类属性特征、类标准模板以及具体分类识别策略都是图像模式识别的关键问题。对于我们具体的手写汉字识别问题,由于单字结构分析和基元形态分析一起可以给出汉字形体的完整描述,而统计形态分析却难以满足大规模汉字字集的集群性和分离性要求。所以,从根本上来讲,要更好地解决手写汉字的机器识别问题,首先必须寻找出能有效描述手写汉字形体的结构表示方法^[1-5]。所谓结构表示方法,就是基于汉字结构分析的一种汉字形体表示方法,其共同点就是把汉字形体看作是由基元(要点、笔段、笔划)、构件(元件、部件)和单字分层次按一定组合方式(图式)构成的一个整体,然后采用一种能反映这些描述规律的数据结构来表示汉字。

最直观的结构表示方法是笔划有序列的表示方法。这种表示方法主要是基于汉字笔划书写的有序性之上的。因为除了笔划本身,笔划序列信息也是一个有用的结构特性,所以许多研究工作都是建立在笔划序列的基础上来进行的^[1-5]。笔划有序列表示方法通过将构成汉字的笔划基元,按先上后下,先左后右,先部件后笔划的原则排列起来,必要时加上特定的关系符和界符,形成笔划和笔划间关系的有序列,就可将二维汉字结构线性化,从而达到对汉字形体的符号化描述。笔划有序列表示法的特点是采用基元种类较少,但已足够表达汉字全集,并且具有较好的容变能力。另外,用笔划有序列来描述汉字,对于建立识别字库而言,时空代价也比较低,有利于大

字集的识别实现。缺点是日前尚缺少有效提取笔划并确定其次序关系的方法和技术。

更一般地,为了表示每一汉字模式的层次结构,可以把汉字描述的结构方法类比语言的语法。由基元、构件以不同方式构成汉字的过程恰似由字构成词,由词构成句子的过程。描述汉字结构的文法包括两个部分,即汉字基元和对基元合成的操作规则。根据所采用操作规则的不同,^[6]及描述汉字的出发点不同,往往可以用到不同的文法^[9-17]。在手写汉字识别中常用的文法有符号文法、属性文法、随机文法和构形文法。

作为二维属性文法的更一般化推广,目前汉字的属性关系图描述方法已成为一种较有广泛影响的方法^[18-21]。在这种方法中,摆脱文法的形式束缚,直接采用图结构来描述汉字构成成员及其关系,这样能够更好地描述汉字笔划之间的多种高维关系,并且由于属性的引入,又能很好地反映汉字形体的非结构特性和基元之间的约束关系。比如象笔划长度、位置以及优先关系等等,不但使描述更加直观,也降低了处理的复杂性。

实际情况中,由于手写笔段并非是精确稳定的,所以笔划竖横撇捺的确认也较为模糊,常需引入相关度刻画,将属性关系图改进为相关属性关系图,或模糊属性关系图,以有利于实际手写汉字识别问题的解决。

作为对汉字属性关系图的一种改造,针对我们对全部汉字所作的结构和形态分析,考虑到汉字结构的多层次性,我们提出了一种汉字形体的拓扑图描述方法,通过多层次汉字拓扑结构的刻画来给出汉字形体的完整描述。

* 本课题为浙江省科技计划项目的一部分并得到中科院自动化所模式识别国家实验室的赞助。周昌乐,副教授,副所长。

94

一、汉字拓扑图描述

首先,我们按照汉字层次结构分析中诸种构件的构成次序和关系,来规定其相应的阶次和关系名称,如表1所示,并约定低阶构件均可作为高阶构件运用,反之则不然。然后,不失一般性,我们设 a 为一个 k 阶构件, $P(i)$ 为构成 a 的一个 i 阶($i \leq k$)构件集, $S(i)$ 为 $P(i)$ 中构件之间的 i 阶构成关系集,则我们称以 $P(i)$ 为节点集, $S(i)$ 为联边集的无向图:

$$G(a, k, i) = (P(i), S(i), f_p^{(i)}, f_s^{(i)}, f_d^{(i)})$$

为 a 的一个 k 类 i 阶拓扑图,其中 $f_p^{(i)}: P(i) \rightarrow D(i)$ (i 阶构件属性生成函数); $f_s^{(i)}: S(i) \rightarrow R$ (i 阶关系权值函数); $f_d^{(i)}: P(i) \rightarrow X^2$ (i 阶构件位置分布函数)。这里 $D(i)$ 为 i 阶构件属性集, R 为实数集, X^2 为二维欧氏空间。

表1 构件阶次及构成关系表

构件名	阶次	构成关系
要点	0	共线关系
笔段	1	相交关系
交件	2	相连关系
连件	3	相接关系
接件	4	相触关系
散件	5	共字关系
本字	6	

不难看出,对于给定汉字,6类0阶拓扑图就是对构成汉字要点属性、位置分布及其关联关系的一种整体描述;而6类1阶拓扑图则是对构成汉字笔段属性、位置分布及其关联关系的一种整体描述。特别是由于同一汉字不同阶次拓扑图之间的内在联系,使得给定汉字的诸阶次拓扑图,不但给出了汉字结构和形态的全部层次描述,而且从低阶到高阶,拓扑图的构造也具有连续性,变得相当容易。实际上,如果设 $G(a, i, j) = (P(j), S(j), f_p^{(j)}, f_s^{(j)}, f_d^{(j)})$ 为一个 i 类 j 阶拓扑图, $G_k(a_k, j, i) = (P_k(i), S_k(i), f_p^{(i)}, f_s^{(i)}, f_d^{(i)})$ 为 $n = |P(j)|$ 个 j 类 i 阶拓扑图, $a_k \in P(j)$,则由于 f_p, f_s, f_d 在相同域上的可迭加性,故:

$$G(a, i, j) = \bigcup_{k=1}^n G_k = (\bigcup_{k=1}^n P_k(i), \bigcup_{k=1}^n S_k(i), f_p^{(i)}, f_s^{(i)}, f_d^{(i)})$$

即为 a 的一个 i 类 j 阶拓扑图。其中 $f_p^{(j)}, f_s^{(j)}, f_d^{(j)}$ 分别为各对应子图中相同函数的迭加函数。

拓扑图不但对汉字结构和形态的描述全面、构造方便,而且也方便对其各种拓扑指数、基元属性以及组合图式的求取。比如通过各阶属性生成函数,我们可以方便地获取基元属性取值。至于汉字构成图式,正好是各阶位置分布函数的近似描述。最后对于

诸种拓扑指数,我们也有如下对应关系:

(1)各类要点可以根据0阶图诸节点的出度及其位置接近关系来确定,从而不难统计各类要点的个数;

(2)笔段数即为1阶图的节点数 $|P(1)|$;

(3)交件数为 $|P(2) - P(1)|$;

(4)非交笔段数为 $|P(2) - (P(2) - P(1))|$;

(5)相交笔段数为 $|P(1)| - |P(2) - (P(2) - P(1))|$;

(6)连件数为 $|P(3) - P(2)|$;

(7)接件数为 $|P(4) - P(3)|$;

(8)散件数为 $|P(5) - P(4)|$;

如此等等。

总之,汉字拓扑图,保留了汉字全部拓扑结构信息和基元形态信息,所以根据汉字拓扑图可以给出汉字形体的完整描述。

二、表示结构的获取生成

输入的汉字形体,经过适当的阈值处理,我们可以获得一幅二值灰度图像。那么不管采用哪种结构表示方法,识别汉字的首要问题都是要通过对二值灰度图像的检测和分析,寻找要点和笔段,从而可以进一步生成所选定的表示结构。对于拓扑图而言,归根结底,就是要获取生成输入汉字的各阶次拓扑图。

生成拓扑图,就是根据从二值汉字图像中检测到的要点集、线段集和笔段集及其有关附带数据,来依次生成汉字的各阶拓扑图,因为不涉及到图像本身,所以生成过程是相对独立的确定性过程。

首先,对于0阶拓扑图,节点集即为要点集,联边集则为线段集,所以有 $P(0) = P, S(0) = S$ 。至于要点属性生成函数 $f_p^{(0)}$ 可令为要点类型函数,即 $f_p^{(0)}(p)$ 表示 p 的类型;联边权值函数 $f_s^{(0)}$ 可令为线段点列 $p_1 d_1 d_2 \dots d_n p_2$ 中的最薄厚度,即:

$$f_s^{(0)}(s) = \min \{ \Delta d_i \}$$

其中 Δd_i 为 d_i 点处的厚度。而位置分布函数 $f_d^{(0)}$ 取值为要点坐标:

$$f_d^{(0)}(p) = \langle x(p), y(p) \rangle$$

其中 $x(p), y(p)$ 分别表示点 p 的横坐标和纵坐标。

其次,对于1阶拓扑图,节点集为笔段集 L ,可令 $S(1) = L$,不失一般性,对于 L 中的元素,可表示为:

$$l = s^{(1)} s^{(2)} \dots s^{(n)}$$

其中 $s^{(i)} = p^{(i)} d_1^{(i)} d_2^{(i)} \dots d_n^{(i)} p^{(i+1)}$ 为第 i 个线段点列,于是笔段的属性函数 $f_p^{(1)}$ 可定义为五元组:

$$f_p^{(1)}(l) = \langle \text{长度, 厚度, 斜度, 曲度, 尖度} \rangle$$

其中:

$$\text{长度} = \sum_{i=1}^n (\text{dist}(p^{(i)}, d_i^{(i)}) + \sum_{j=1}^{m_i} \text{dist}(d_i^{(i)}, p^{(i+j)}))$$

$$\text{厚度} = \frac{(\sum_{i=1}^{n+1} \Delta p^{(i)}) + (\sum_{i=1}^n \sum_{j=1}^{m_i} \Delta d_i^{(i)})}{n+1 + \sum_{i=1}^n m_i}$$

斜度 = $p^{(1)}p^{(n+1)}$ 向量与水平轴的夹角

曲度 = $p^{(1)}$ 与 $p^{(n+1)}$ 连线两侧垂直偏离最大的点距

$$\text{尖度} = \begin{cases} \Delta p^{(1)} / \Delta p^{(n+1)} & \Delta p^{(1)} < \text{厚度} < \Delta p^{(n+1)} \text{ 或反之} \\ 1 & \text{其它} \end{cases}$$

而笔段的位置分布函数 $f_p^{(1)}$ 可定义为:

$f_p^{(1)}(l) = l$ 的形心坐标 = (x_c, y_c) , 其中:

$$x_c = \frac{\sum_{i=1}^{n+1} x(p^{(i)}) + (\sum_{i=1}^n \sum_{j=1}^{m_i} x(d_i^{(i)}))}{n+1 + \sum_{i=1}^n m_i}$$

$$y_c = \frac{\sum_{i=1}^{n+1} y(p^{(i)}) + (\sum_{i=1}^n \sum_{j=1}^{m_i} y(d_i^{(i)}))}{n+1 + \sum_{i=1}^n m_i}$$

这里 $x(p)$ 和 $y(p)$ 分别表示 p 的横坐标和纵坐标。

对于 1 阶拓扑图的联边集 $S(1)$, 则可根据共线数大于 1 的要点所处的复合笔段集来两两组对形成。设共线数大于 1 的要点集为 $A = \{a_i | i=1, 2, \dots, n\}$, 每个 a_i 穿过的复合笔段集分别为 $L(a_i) = \{l_j^{(i)} | j=1, 2, \dots, m_i\}$, 则每个 a_i 的笔段相交关系集为:

$$S(a_i) = \{(l_{j_1}^{(i)}, l_{j_2}^{(i)}) | j_1 \neq j_2 \wedge l_{j_1}^{(i)} \in L(a_i) \wedge l_{j_2}^{(i)} \in L(a_i)\}$$

于是 1 阶拓扑图的联边集 $S^{(1)}$ 就为 $S^{(1)} = \bigcup_{i=1}^n S(a_i)$ 。而 $S^{(1)}$ 上的权值函数可定义为两笔段相交关系的可信度, 令为

$$f_s^{(1)}((l_{j_1}^{(i)}, l_{j_2}^{(i)})) = h^{(1)}(b_{j_1}^{(i)}, b_{j_2}^{(i)}, b_i)$$

其中 $b_{j_1}^{(i)}, b_{j_2}^{(i)}$ 分别为形成 $l_{j_1}^{(i)}$ 和 $l_{j_2}^{(i)}$ 时在 a_i 处的两线段夹角, b_i 为要点 a_i 的确认度, $h^{(1)}()$ 为采用的可信度计算公式。

在 1 阶拓扑图的基础上, 我们很容易构造出 2 阶拓扑图的节点集, 只要按连边子图, 将 1 阶拓扑图中的节点划分为自然小集团, $X_i (i=1, 2, \dots, m)$, 每个小集团代表 2 阶拓扑图中的一个节点, 并定义每个这样的一个节点的属性函数 $f_p^{(2)}$ 为笔段小集团中所含的交点数和笔段数, 而节点的位置分布函数为笔段小集团的形心坐标。所以关键是要根据折点(包括含折点的复点)来寻找笔段小集团之间的相连关系, 以便形成 2 阶拓扑图的联边集。

设所有的折点(包括含折点的复点)集合为 $A = \{a_i | i=1, 2, \dots, n\}$, 每个折点 a_i 的二条笔段为 l_i' 和 l_i'' , 令 $L' = \{l_i' | i=1, 2, \dots, n\}$; $L'' = \{l_i'' | i=1, 2, \dots, n\}$, 则对于每个小集团 X_j , 分别令 $X_j' = L' \cap X_j$; $X_j'' = L'' \cap X_j$ 。

如果 $X_j' \neq \Phi$ 则用 $L''(X_j') = \{l_i'' | l_i' \in X_j'\}$ 去找出所有与之相交不为空的小集团 $X_k (k \neq j), k = j_1, j_2, \dots, j_k$, 形成连边子集:

$$S'(X_j) = \{(X_j, X_k) | k = j_1, j_2, \dots, j_k\}$$

同样如果 $X_j'' \neq \Phi$ 则用 $L'(X_j'') = \{l_i' | l_i'' \in X_j''\}$ 去形成连边子集 $S''(X_j)$ 。最后在不计有序对次序的前提下, 便可得 2 阶拓扑图的连边集为:

$$S(2) = \bigcup_{j=1}^m (S'(X_j) \cup S''(X_j))$$

并定义其上的权值函数为:

$$f_s^{(2)}((X_j, X_k)) = h^{(2)}(b_{jk})$$

其中 b_{jk} 为连接 X_j 和 X_k 二个小集团折点二线段夹角均值, $h^{(2)}$ 为采用的可信度计算公式。

类似于 2 阶拓扑图的构造, 对于 3 阶和 4 阶拓扑图, 均可建立在 2 阶和 3 阶拓扑图上交件和连件小集团的寻找及相接和相触关系的构建来形成。不同的是此时依据的要点集分别是接点(含复合接点)集和触点集, 节点属性函数值则分别是折点数及交件数和接点数及连件数, 可信度计算公式 $h^{(3)}$ 和 $h^{(4)}$ 的参数分别为连接二个小集团接点和触点的可信度均值, 而位置分布函数为小集团的形心坐标。

最后, 对于 5 阶拓扑图, 则仅需通过 4 阶拓扑图的连通子图来形成接件小集团, 作为 5 阶拓扑图的节点集, 并定义其上的属性函数为触点个数和接件数, 位置分布函数同样为小集团的形心坐标。鉴于触点的稳定性最差, 所以一般对 5 阶拓扑图的生成不作考虑, 可以缺省。

总之, 通过要点寻找、线段合并、笔段形成和要点确认, 我们可以生成输入汉字图像的各阶拓扑图, 为单字识别的结构匹配提供了最初始的输入数据。

结论 在本文中给出的拓扑图表示法动态获取算法, 已用于具体的实用化手写汉字识别系统的构建中, 实验表明, 这种表示方法的简明有效, 使系统的识别率和识别速度都有明显的提高, 特别是根据我们方法替换测试, 从文[22, 23]中的 80% 提高到 95% 以上^[24, 25], 有 5% 左右得益于这种新的表示方法。这说明手写汉字拓扑图表示法是一种较为理想的方法。(参考文献共 25 篇略)