一种脱机手写体汉字识别的容错编码方法研究

王建平 赵丽欣 王金玲
(合肥工业大学电气及自动化工程学院,合肥 230009)

摘 要 手写体汉字识别是字符识别领域中的难点。为了使机器识别汉字适应于手写体汉字的变形等因素,基于人类认识汉字的容错机理,提出了一种用于机器识字的汉字容错编码方法,以提高手写体汉字识别率。该编码方法首先对横竖撇捺笔划形态给出了模糊化表示;然后定义了仿人拆字的字元集,并给出了易混淆笔划字元的多归类容错编码;接着给出了笔划字元的顺序判断规则和归结了 36 类简单常用字的部首子结构,并给出冗余的容错编码;进而建立了仿人构字的汉字编码规则和具有容错性的多模板字典,并对《新华字典》中收录的10 000余个单字汉字进行了标准编码,重码率为 0.48%;最后对 HCCORG 和 NKIM 手写体汉字库中的 100 个手写体汉字进行了仿真识别,识别正确率为 96%。试验结果表明,这种编码方法可生成多模板字典,不仅对手写体汉字变形具有较好的容错性,且重码率和误识率较低。

关键词 脱机手写体汉字识别 容错编码 字元集 笔划顺序 子结构
中图法分类号: TP391.43 **文献标识码**: A **文章编号**: 1006-8961(2007)12-2169-10

A Study of Chinese Characters Code of Bearable Mistakes Method for Off-line Handwritten Chinese Characters Recognition

WANG Jian-ping, ZHAO Li-xin, WANG Jin-ling
(*School of Electric Engineering and Automation, Hefei University of Technology, Hefei* 230009)

Abstract Handwritten Chinese characters recognition is the difficulty of character recognition. Based on the mechanism of apery imitation, a kind of Chinese characters codes for computer cognition is presented in this paper to apply to the deformation factors of handwritten Chinese characters and to improve the recognition rate of Chinese characters. The configurations of horizontal stroke, upright stroke, left-falling stroke and right-falling stroke are defined in a fuzzy way. Elements groups of Chinese characters are made for machine cognition. Bearable mistakes codes of various categories are given to the elements which are easily confused. Rules for judging stroke sequence are given. 36 kinds of subsidiary configurations codes and bearable mistakes codes are constructed. The code principles and multi-template dictionary of Chinese characters which agree with apery imitation are established. 10 000 Chinese characters in Xin Hua Dictionary are standardized coded, the rate of repeated codes of which is 0.48%. After testing the recognition on 100 handwritten Chinese characters in the handwritten Chinese character library of HCCORG and NKIM, the recognition rate is 96%. Emulational experimental results show that this kind of coding applies to the deformation of handwritten Chinese characters well and the rates of repeated codes and wrong codes are low.

Keywords off-line handwritten Chinese characters recognition, bearable mistakes code, elements groups, stroke sequence, subsidiary configurations

1 引 言

手写体汉字的识别是字符识别领域中的难点[1]。在脱机手写体汉字识别中,由于笔画复杂、模式类别多[2]以及缺少笔划和笔顺信息,特别是由于不同人汉字书写风格的差异造成的手写体汉字变形很大,使得属于同一汉字类别的不同样本之间的差异较大,因此产生一个具有代表性的识别字典是补偿手写体汉字变形、提高手写体汉字识别率的有

收稿日期:2005-12-02;**改回日期**:2006-05-20

第一作者简介:王建平(1955~),男,教授,博士。主要研究方向为智能测控技术、机器视觉与图像识别系统等。发表学术论文 50 多篇,出版学术著作 2 本。E-mail:wangww@mail.hf.ah.cn

效途径。通常识别字典可以分成单模板字典和多模板字典两大类,前者指每个类别仅有一个参考模板(也可称为代表元)存放在识别字典中;后者指每类有一个或一个以上的参考模板。由于手写体汉字变形的存在,使单模板字典显得不足,因此有必要生成多模板字典[3]。

现有能比较完整地表征汉字的编码方法都是人类用于将汉字输入计算机的方法,如,五笔字型码、太极码、表形码[4]等。而用于机器识别的能较完整表征手写体汉字的,并可以生成多模板字典的编码方法还较少,是一个值得研究的方向。

本文模仿人认识汉字的过程机理和容错机制,首先对横竖撇捺笔划形态给出模糊化定义,然后制定了仿人拆字的图像汉字容错字元集,同时选取 36 类汉字子结构编码,并给出了易变形的容错编码,进而建立了多模板汉字构字的字元编码规则和字典。该编码方法形成了多码一字的关系,并对易重码和误码的字做了特定的区分。实验和仿真表明,本文编码方法能很好地表征和区分汉字集,不但字元提取稳定,且编码重码率低;同时对手写体汉字的变形等具有较好的容错性,且误识率较低。

2 汉字的特征分析和选取

由汉字国标一、二级字库(GB2312-80)的6 763个汉字的统计结果表明[5],包含横笔划的汉字占 99.8%,包含竖笔划的汉字占 99.85%,包含撇笔划的占 93.5%,包含捺笔划的占 76.5%,4 种笔划在汉字中出现的频率为:横笔划占 39.51%,竖笔划占 33.94%,撇笔划占 16.77%,捺笔划占 9.78%。由以上分析结果可知,以横竖撇捺笔划和少量自规定形状为字元,再结合拓扑结构特征,就能够完整地表征汉字集的特征。因此,把图像汉字转化为由横、竖、撇、捺等基本笔划在不同位置组成的图形,再根据汉字笔划形态、顺序、数量、位置关系信息即可描述和表征每个汉字特征。

2.1 汉字的笔划特征选择

众所周知,汉字有横、竖、撇、捺、折 5 种基本笔划。在待识别的手写体汉字样本中,由于每个人的书写习惯不同,同一种汉字的横、竖、撇、捺笔划形态可能会有变形,故采用撇捺笔划和横竖笔划之间的笔划形态类型之间相互包含的模糊化表示方法,以减少拒识和提高容错性。

本文建立的用于描述横、竖、撇、捺笔划形态类型的隶属度函数如下:

横:

$$\mu H_k(\theta)=1-\mathrm{abs}\left(\frac{\theta}{30^\circ}\right)\quad -30^\circ\leqslant\theta\leqslant 30^\circ \tag{1}$$

竖:

$$\mu V_k(\theta)=1-\mathrm{abs}\left(\frac{\theta-90^\circ}{10^\circ}\right)\quad 80^\circ\leqslant\theta\leqslant 90^\circ \tag{2}$$

$$\mu V_k(\theta)=1-\mathrm{abs}\left(\frac{\theta+90^\circ}{10^\circ}\right)\quad -90^\circ\leqslant\theta\leqslant -80^\circ \tag{3}$$

撇:

$$\mu P_k(\theta)=1-\mathrm{abs}\left(\frac{\theta+45^\circ}{45^\circ}\right)\quad -90^\circ<\theta<0^\circ \tag{4}$$

捺:

$$\mu N_k(\theta)=1-\mathrm{abs}\left(\frac{\theta-45^\circ}{45^\circ}\right)\quad 0^\circ<\theta<90^\circ \tag{5}$$

其中,$\mu H_k(\theta)$,$\mu V_k(\theta)$,$\mu P_k(\theta)$,$\mu N_k(\theta)$分别表示第 k 个笔划隶属于横、竖、撇、捺笔划的隶属度函数,abs 表示绝对值函数。设笔划的起点坐标为(i_s, j_s)(下角 s 代表 start),终点坐标为(i_e, j_e)(下角 e 代表 end),则 θ 可定义如下:$\theta=\arctan\frac{j_s-j_e}{i_e-i_s}(i_s\neq i_e)$ $-90^\circ<\theta<90^\circ$;$i_s=i_e$;且 $j_s<j_e$ 时 $\theta=90^\circ$;$i_s=i_e$ 且 $j_s>j_e$ 时 $\theta=-90^\circ$。

若一个笔划具有两种笔形隶属度,则按隶属度大小对笔形分别编码。

由于“折”作为独立笔形不易被机器直接提取,因此,从机器认知角度可把“折”看作是笔划组合,可用横竖撇捺来定义。通过对《小学生字典》袖珍本中 28 种基本构字笔划的研究分析,为保证基本的构字需要,可将“折”分为以下 5 类:

(1) 横折 横笔划的终点与竖[左竖勾]或撇的起点相接。横起笔向下转折 

(2) 竖折 竖笔划的终点与勾或横或横勾的起点相接。竖起笔向右转折

(3) 撇折 撇笔划终点与点或捺或横的起点相接。撇起笔向右转折 

(4) 斜勾 捺笔划终点与点或横勾的起点相接。捺起笔向右上转折 

(5) 弯勾 捺笔画终点与左竖勾的起点相接。捺起笔向左弯

为适应于不同字体,规定:若折笔划的起笔笔划

的终点与另一笔划的起点之间像素点的距离不超过阈值 T,则认为这两笔划的一个终点与另一个起点重合,即构成折笔划。

2.2　汉字的字型特征选择与代码

汉字字型的划分是基于对汉字整体结构的认识[4],而且无论对手写体或印刷体汉字,字型都是一项稳定的特征。为了更明细地划分汉字字型,本文采用了两级划分法。

汉字字型分为:左右型、上下型、杂合型 3 类。

两级划分法为:若整体汉字字型为杂合型,则不再区分,若为合体字,则再分别判断左、右(或上、下)每一部分的字型信息,这两部分又按 3 种类型划分。对合体字第 1 级进行整体划分时,按从左到右(从上到下)划分,最左部分(上部)为左(上)部,其余为右(下)部。具体划分及代码如表 1所示。

表 1　汉字字型代码表

Tab.1　Codes of Chinese Character form

字型特征名称	一级字型划分	二级字型划分		字型特征代码
		左部分字型	右部分字型	
整体左右型	左右型	杂合型	杂合型	133
		杂合型	左右型	131
		杂合型	上下型	132
		上下型	杂合型	123
		上下型	左右型	121
		上下型	上下型	122
		上部分字型	下部分字型	
整体上下型	上下型	杂合型	杂合型	233
		杂合型	左右型	231
		杂合型	上下型	232
		左右型	杂合型	213
		左右型	左右型	211
		左右型	上下型	212
整体杂合型	杂合型	不可分		300

例如:“神”的字型代码记为“133”;“侧”字型代码记为“131”。“意”字型代码记为“232”;“想”字型代码记为“213”。“困”、“秉”等杂合型汉字代码记为“300”。

2.3　汉字字元选取与代码

为了完全表征汉字笔划组成,并降低计算机提取字元的难度,选取的字元及笔划字元代码如表 2所示,其他结构均可看作这些字元的有机组合。

表 2　笔划字元代码表

Tab.2　Codes of stroke elements

笔划类型	代码	编码名称	基本笔划	单笔、组合笔划示例	定义
单笔划	1	横	一	一 ㇐ 𠂉	横、横撇、横点
	2	竖	丨	丨 丨 亅 丿	竖、左竖钩、右竖勾、竖撇
	3	撇	丿	𠂉 丿 丿	撇(横撇、竖撇、普通撇)、提
	4	捺	丶	丶 丶 乀	点、捺、斜勾
	5	横折	𠃌	𠃌 𠃌 𠃌	横折
笔划组合	6	竖折	𠃍	𠃍 𠃍 𠃍 丨 丨	竖折、竖、左竖勾
	7	撇折	𠃋	𠃋 𠃋	撇折
	8	斜勾	乀	乀 乀 乀	斜勾、捺
	9	弯勾	𠄎	𠄎	弯勾

2.4 汉字字元顺序特征选取

2.4.1 单笔划字元横、竖、撇、捺的字元优先顺序判断算法[6]：

单笔划可用笔划的起点和终点坐标来表示，笔划 $S_1=[(X_{11},Y_{11}),(X_{12},Y_{12})/X_{11}\leqslant X_{12}]$ 与笔划 $S_2=[(X_{21},Y_{21}),(X_{22},Y_{22})/X_{21}\leqslant X_{22}]$ 的笔划字元顺序可以分为相交和不相交两种情况来判断。若相交，则按横、竖、撇、捺的次序产生相应笔划字元顺序；若不相交，则其笔划字元顺序先后的判断可以依据笔划字元的中点位置进行判定，其判定公式为

$$\begin{aligned}&\text{当 } X_{21}+X_{22}-X_{11}-X_{12}\neq 0 \quad \text{则 } -1<\frac{Y_{21}+Y_{22}-Y_{11}-Y_{12}}{X_{21}+X_{22}-X_{11}-X_{12}}\leqslant 1\\&\text{当 } X_{21}+X_{22}-X_{11}-X_{12}=0 \quad \text{则 } Y_{22}+Y_{21}-Y_{12}-Y_{11}<0\end{aligned} \tag{6}$$

若满足式(6)的不等式条件，则可认为笔划字元 S_1 的笔顺要先于笔划字元 S_2 的笔顺。

2.4.2 折笔划字元优先顺序判断算法

折笔划字元以它的起笔笔划与其他字元按2.4.1节中的方法判断笔划字元的笔顺。

2.4.3 汉字字元的顺序选取规则

依照2.2节的汉字字型特征的划分结果，依次对每部分结构选取笔划字元的顺序。

(1) 第1笔字元 取这部分汉字图像中最高点(如有几个点，取最左边的)所在字元为第1笔字元，若有两个字元，则按2.4.1节和2.4.2节中的方法比较笔划字元顺序，笔划字元笔顺优先的为第1笔字元。

(2) 第2笔字元 与第1笔字元相交或相连(包括相切和相接)的所有笔划字元中，如果有和第1笔字元共最高最左点的，则为第2笔字元，若没有，则按2.4.1节和2.4.2节中的方法比较笔划字元的笔顺，笔划字元笔顺最优先的为第2笔字元；若没有与第1笔字元相交或相连的字元，则除第1笔字元外，取这部分汉字图像中最高点(如有几个点，取最左边的)所在字元为第2笔字元，若有两个字元，则按2.4.1节和2.4.2节中的方法比较笔划字元顺序，笔划字元顺序优先的为第2笔字元。

(3) 第3笔字元 第3笔字元的确定方法同第2笔字元，从与第2笔字元相交或相连的字元中选取，或取除第1，2笔字元外汉字图像中最高最左点所在字元。

(4)第4笔字元及以后的笔划字元顺序的选取依次类推。

2.5 汉字统计特征选取

对手写体或印刷体汉字而言，由于汉字笔划相交点是一个稳定的特征，因此选取汉字的交点数量是一个稳定的统计特征。由于汉字中的横、竖笔划出现频率最高，因此可选取横、竖笔划的数量作为汉字统计特征。同时，由于机器对汉字横、竖笔划和相交点提取准确稳定，因此用这3种汉字的全局统计特征首先对汉字集进行粗分类，这不仅可提高识别速度，而且具有较强的鲁棒性。

2.6 汉字的子结构选取与编码

汉字存在多种结构稳定的部首或字根，称之为子结构[7]。通过对《新华字典》部首检字表的部首目录中的所有部首进行研究分析，归结出36类易于机器识别的汉字常用子结构，具体如表3所示。

表3 子结构的特征判断和字元构成表

Tab.3 Characteristic judgement and element construct of subsidiary configurations

序号	名称	交点数	笔划码	容错码1	容错码2	容错码3	容错码4	容错码5
a000	彡(左、右)	0	3330	1330	3300			
a001	阝(左、右)	0	2590	2544	2543			
a002	大(上、下)	1	3140	2140				
a003	巾(左、下)	1	2520	2122	2250	2212		
a004	卩(右、下)	0	2500	2120				
a005	女(左、下)	2	7130	3134				
a006	刂(右)	0	2200	2600	2240			
a007	讠(左)	0	4530	4160	4510	4120	4500	1530
a008	礻或衤(左)	0	4527	4524	4132	1527	1524	1132
a009	氵或冫(左)	0	4430	4440	4410	1130	4130	1430
			4300	4400	4100	1300	1100	1110
a010	犭或亻(左)	0	3320	1320	3200	1200	1120	
a011	纟(左)	0	7730	7710	7313	7311	3171	3173

续表

序号	名称	交点数	笔划码	容错码 1	容错码 2	容错码 3	容错码 4	容错码 5
a012	𠂉(左)	0	3560	3520	3120	3160		
a013	𠂆(左)	1	3116	3112				
a014	扌(左)	2	2130	2110	6130	6110		
a015	忄(左)	0	2240	2340	2440			
a016	犭(左)	1	9330	4323	3930	3423		
a017	廾(下)	2	3120	2123	2120	2130		
a018	灬(下)	0	3444	4444	4440	3440		
a019	艹(上)	2	2120	4120	4130	2130	3120	3140
a020	宀或冖(上)	0	4520	4540	4530	5200	5400	5300
a021	亠(上)	0	4100	1100				
a022	厶	0	7400	3140				
a023	夂或夊	1	3134	3540				
a024	十	1	2100	2300				
a025	人	0	3400	2340	2430			
a026	土或士	1	2110	2130	3110	3130		
a027	牛	2	2131	2133	3121	3123		
a028	王	1	1211	1213	1311	1313		
a029	木	1	2134	2334				
a030	力	1	3500	2530	2500	3120	2123	3540
a031	山	0	2620	2122	2322			
a032	贝	0	2534	5234	2123	1223		
a033	工	0	1210	1230	3210	3230		
a034	口	0	2510	5210	1212	2121	2131	1231
a035	子	1	5210	5230	1321	1323		
a036	又	1	5400	1340	4500	4130		

根据汉字字型的特征划分,对各部分的位置、字元顺序和相交点特征进行判断,若所取的笔划字元数不大于 4,则将其与子结构的特征((位置 and 交点数 and 笔划码)or(位置 and 交点数 and 容错码))相比较,若与某种子结构的特征码相符合,则认为待识别汉字具有此种子结构,并以子结构的序号为构字编码。

3　汉字编码规则

本文编码为:汉字编码等于统计特征码(3 个码)加字型码(3 个码)加笔划码(16 个码)

3.1　汉字统计特征码编码规则

整个汉字编码的前 3 个码为汉字统计特征码,编码规则如表 4 所示。

为使提取的横、竖笔划数量统计特征稳定可靠,只对笔划长度大于 $W/5$(W 是图像汉字的宽度)的横、竖笔划数量进行统计。若相交点、横笔划、竖笔划任何一项的数量超过 9,则记为 9。如,“齾”的统计特征码为 199。该字的横笔划数量虽超过 9,但仍记为“9”。

表 4　汉字统计特征码编码规则表[8]

Tab. 4　The coding rules of statistics characteristic codes[8]

编码位置	首码	第 2 码	第 3 码
编码含义	相交点数量	横笔划数量	竖笔划数量
编码取值	0 ~ 9	0 ~ 9	0 ~ 9

3.2　汉字字型码编码规则

整个汉字编码的第 4 ~ 第 6 个码是字型码,编码规则见 2.2 节的表 1。

3.3　汉字笔划码编码规则

汉字笔划码编码规则如下:

(1) 按 2.4 节规定的笔划字元顺序取码,字元以表 2 中的 9 种代码表示;

(2) 折笔划字元优先独立编码;

(3) 笔形不重复取码(即取过码的笔划字元不再取码);

(4) 按汉字字型特征取码(共 16 个码),其规则如下:

1) 整体杂合型(300 型)笔划码编码规则:按规则(1)~规则(4)的规则取汉字前 10 个码,不足码补 0。后边 6 个码补 0。若杂合型汉字的笔划字元数量小于等于 4,则检测前 4 个码是否与某种子结构的特征((位置 and 交点数 and 笔划码)or(位置 and 交点数 and 容错码))相符合,若有特征相符的,则这 4 个码位置以子结构序号码编码;若没有特征相符的,则保留原取码。

2) 整体左右型(133、121、122、123、131、132 型)笔划码编码规则为:①整体左和右部分分别对应前 8 位和后 8 位笔划字元码;②若左和右任何一部分为杂合型,则按字元顺序取前 4 个码,不足补 0,并在余下 4 个码补 0;③若可以再分,则每一再分的部分再分别按字元顺序取前 4 个码,不足补 0。然后检测每部分的前 4 个码是否有与子结构特征((位置 and 交点数 and 笔划码)or(位置 and 交点数 and 容错码))相符合,若有,则这部分笔划码以子结构序号码编码;若没有,则保留原取码。

3) 整体上下型(233、211、212、213、231、232 型)笔划码编码规则为:①整体上和下部分分别对应前 8 位和后 8 位笔划字元码;②若上和下任何一部分为杂合型,则按基元顺序取前 4 个码,不足补 0,在余下 4 个码补 0;③若可以再分,则每一再分的部分再分别按字元顺序取前 4 个码,不足补 0。然后检测每部分的前 4 码是否有与子结构特征((位置 and 交点数 and 笔划码)or(位置 and 交点数 and 容错码))相符合,若有,则这部分笔划码以子结构序号码编码;若没有,则保留原取码。按照本文容错编码规则,即可得到汉字的容错编码,表 5 给出了汉字“组”、“件”、“鬃”的容错编码举例。

表 5 汉字容错编码举例

Tab. 5 Examples of bearable mistakes codes for Chinese characters

汉字	组		件		鬃	
笔划码	纟	且	亻	牛	髟	宗
	77300000	25110000	32000000	21310000	21113330	45201123
	77100000		12000000	21330000	21111330	45401123
	73130000		33200000	31210000	21113300	45301123
	73110000		13200000	31230000		52001123
	31730000					54001123
	31710000					53001123
	a012000025110000		a0100000a0270000		2111a000a0201123	
统计码	062		222		082	
字型码	133		133		212	
编码	062133a012000025110000		222133a0100000a0270000		0822122111a000a0201123	

4 重码与容错码的处理

4.1 重码字的分析与处理

基于笔划的编码方法易对笔划形态、数目和笔顺及字型等完全一致的汉字造成重码,但增加汉字结构中的笔划交点特征可进一步减少汉字编码的重码率。而采用 22bit 汉字编码,则理论上可将汉字分为 $10^3 \times 4^3 \times 10^{16}$ 类,这远远大于实际应用的汉字数量(GB2312-80 中6 763个汉字),字元代码表中共有 25 种笔形,同一种代码最多有 5 种笔形,也就是说,取一个码的字元码,其最大重码概率为 1/5,取两个码的重码概率为 1/5 × (1/5) = 1/25,……,这样依次每多取一码重码概率就比前面的减少 4/5,本文编码字元最多取 16 个码,因此从理论上说,汉字重码数量很少,仅笔划形状、数目和笔顺一致的字形相近的简单汉字存在重码的可能性。

重码汉字又分为编码重码和容错码重码两种。其中,编码重码,即一码多字重码(如“土”和“士”);容错重码,由于在字元代码中,为了对手写体汉字进行更好的容错,使易混淆的字元间相互包含(见表 2),因而造成一字多码之间的重码(如“天”和“夭”)。

编码重码汉字有:“己、已、巳”;“末、未”;“日、曰”等。

容错重码汉字有:“天、夭”;“干、于,千”;“刁、

刀”等。

对重码汉字,可采用借助于笔划间的对比关系加以区分。其对比关系特征为

$$\boldsymbol{T} = \{(S_i, S_j, r)/r \in \boldsymbol{R}\} \tag{7}$$

其中,(S_i, S_j, r)代表汉字第 i 个笔划字元 S_i 和第 j 个笔划字元 S_j 存在着 r 的对比关系;$\boldsymbol{R}$ 为两个字元之间的对比关系集合(如长于、短于、高于、低于等)。这样就能将笔划形状、数目和笔顺一致的汉字区分开来。

例如,“未”、“末”按本文的编码都为2212113400000000000,对其建立的对比关系区分如下:

$\boldsymbol{T}_{未} = (S_2, S_3, 短于)$;$\boldsymbol{T}_{未} = (S_2, S_3, 长于)$

通过汉字字元的对比关系就可以把笔划形状、数目和笔顺一致的编码重码字辨别出来。

在表 3 子结构的特征(位置 and 字元顺序码 and 相交点)编码中,不存在重码。

4.2　冗余容错编码机制

鉴于机器识别汉字笔划字元可能存在不完整和不一致性,因此为提高机器认字的识别率和正确率,必须采用模仿人的容错机制构建用于机器识别的汉字冗余编码。构建步理如下:

(1) 根据汉字笔划类型统计分析结果,归结出机器能够认知的 9 类笔划字元,同时给出了笔划字元的编码(代码),并针对这 9 种类型笔划字元在提取时可能混淆的笔划,给出了多归类容错编码(笔划编码时有相互包含容错的关系)。具体如表 2 所示。

例如,“重”字编码为 5633215211110000000,容错码为 5731215211110000000。“重”字第 1 笔字元横撇按表 2 中的定义,既可以按照撇编码,也可以按横编码。

(2) 针对汉字国标字库(GB2312-80)存在多种结构稳定的部首或字根,归结出了 36 类易于机器识别的常用子结构,还规定了子结构的编码(序号码),并针对这些子结构提取时可能存在的不完整和不一致性,给出了多个容错码。具体如表 3 所示。

(3) 针对复杂汉字的笔划密集区的笔划尤其是撇捺笔划提取时可能出现不稳定性问题,仿照人认识汉字时的容错性,按编码规则中规定的取码顺序和数量,复杂汉字的笔划密集区的撇捺笔划基本不参与编码,从而提高了编码识别时的鲁棒性。

(4) 根据汉字的横、竖笔划和交点出现的频率高,且提取比较稳定的特点,给出了汉字整体统计特征编码(前 3 个码)。这种编码方法优点在于出现拒识、误识时,可根据这 3 种特征的提取误差不会太大,且允许先把粗分类的范围扩大到分别给这 3 类码数值减 2 ~ 加 2 之间(在保证 3 个码中的每一码数值大于等于 0 的前提下),再进行识别的容错编码,进而实现容错识别。

本文编码方法优点是:一个汉字有多个编码,较少出现一个编码多个汉字情况,不仅具有机器识字的容错性,且易于实现。

5　识别结果置信度评价

5.1　手写体汉字与标准汉字相似度的定义

对待识别手写体汉字进行编码,和标准汉字编码相比较。将待识别手写体汉字编码的第 i 个位置码记为 $\hat{M}_i (i = 1, 2, \cdots, 22)$,将标准汉字库的汉字编码第 i 个位置码记为 $M_i (i = 1, 2, \cdots, 22)$,待识别手写体汉字编码和标准汉字编码的整体能匹配的码的个数计数变量为 sum,初始 $sum = 0$,若 $\hat{M}_i = M_i$,则 $sum = sum + 1$;若 $\hat{M}_i \neq M_i$,则 $sum = sum + 0$。因此可定义待识别手写体汉字与被匹配标准汉字的整体相似度 $D_z = sum/22$。

设手写体汉字从 $\hat{M}_7$ 到 $\hat{M}_{22}$ 的字型笔划码部分不为 0 的码数量为 $N (1 \leqslant N \leqslant 16)$,16 位笔划码的匹配个数的计数变量为 sm,初始 $sm = 0$,将这 N 个不为 0 的码与其相对位置的标准汉字码进行比对匹配,若相同,则 $sm = sm + 1$;若不同,则 $sm = sm + 0$。因此可定义待识别手写体汉字与被匹配标准汉字的笔划字元相似度为 $D_b = sm/N$。

5.2　识别结果置信度

在进行手写体汉字与被匹配汉字的识别时,首先在前 3 个码(粗分类)是 $\hat{M}_1\hat{M}_2\hat{M}_3$ 的汉字范围内进行匹配,并取 D_z 值最大的作为识别结果。当识别结果满足条件 1:$D_b \geqslant 0.6 \&\& D_z \geqslant 18/22$ 时,则认为识别结果可信,即待识别汉字为被匹配的汉字。当识别结果不满足 $D_b \geqslant 0.6 \&\& D_z \geqslant 18/22$,则认为识别结果不可信,此时将粗分类的前 3 个码范围扩大到满足 $(\hat{M}_1 - 2)(\hat{M}_2 - 2)(\hat{M}_3 - 2) \sim (\hat{M}_1 + 2)(\hat{M}_2 + 2)(\hat{M}_3 + 2)$ 进行匹配,并取 D_z 值最大的作为识别结果。当二次识别结果满足 $D_b \geqslant 0.6 \&\& D_z \geqslant 16/22$ 时,则认为识别结果可信,待识别汉字为被匹配的汉字。当二次识别结果不满足 $D_b \geqslant 0.6 \&\& D_z \geqslant 16/22$ 时,则拒识。(前 3 个码粗分类范围扩大时,每一码

必须满足 0≤码值≤9)

6 实验结果

6.1 对标准样本汉字的编码结果

采用本文编码方法,对《新华字典》中收录的 10 000余个单字汉字进行了标准编码,编码结果如表 6 所示。

表 6 重码率

Tab.6 The rates of repeated codes

选取的编码样本数	重码样本数	重码率(%)
10 000	48	0.48

重码汉字:(a)干、千、于;(b)天、夭;(c)末、未;(d)壬、王;(e)己、已、巳;(f)日、曰;(g)犬、太;(h)刁、刀;(i)子、孑;(j)目、且;(k)冶、治;(l)住、往;(m)仿、彷;(n)侍、待;(o)汗、汙;(p)纤、纤;(q)竿、竽;(r)芊、芋;(s)汩、汨;(t)泪、泪;(u)洗、冼;(v)袄、祆;(w)叼、叨。

利用 4.1 节重码汉字处理方法,引入笔划的对比关系后,可以区别的重码字为:(c)末、未;(e)己、已、巳;(f)日、曰;(g)犬、太;(j)目、且;(s)汩、汨;(t)泪、泪。

其他重码汉字按照本文编码方法无法区分。

重码汉字类型分析:

(1) 为了更好的容错,本文将手写时可能混淆的字元之间有相互包含,或相似的汉字字元归为一类,如将竖"丨"和左竖勾"亅"都归为竖编码;横撇折"㇇"和横竖折"㇕"都归为横折编码等。这样在有些非常相近的汉字只有这一点细微区别时,则会造成重码。如"干"和"于"。

(2) 为了适应手写汉字的不规范和提取时的可能失误,本文将形状相近的部首归为一类子结构编码,如"彳"和"亻"按同一种子结构编码等。这样如果形近汉字只有在这种子结构上的一点区别时,则会造成重码。如"往"和"住"。

6.2 对手写体汉字的识别结果

本文对 HCCORG 和 NKIM 手写体字库中的 100 个手写体汉字(细化后的)进行了仿真识别,识别结果如表 7 所示。

表 7 识别率

Tab.7 The rates of recognition

识别汉字样本数	误识汉字数	拒识汉字数	正确识别率(%)
100	1	3	96

(1) 试验结果误识汉字类型分析如下:

本文对整体杂合型汉字的笔划码取前 10 个码,对于笔划字元数目较多的形近杂合型汉字,其笔划码的前 10 个码可能相同,区别仅在整体前 3 个码的统计特征。如果标准统计特征码相差很小,而在提取时又出现误差,则有可能造成误识。

(2) 试验结果拒识汉字类型分析如下:

① 本文笔划码的取码是基于字型的,由于手写字体的不规范性,因此对于有些连笔或断笔较多的汉字,可能会造成判断字型的错误,这时会造成拒识。

② 为了方便计算机识别笔划字元顺序,同时提高识别笔顺的速度,本文规定第 1 笔字元取分型后汉字图像的最高最左点所在字元,由于手写汉字不规范,因此对不是子结构的汉字(或汉字部分)可能会造成第 1 笔字元识别错误,这种情况也可能造成拒识。

其中"皇"和"法"两汉字的试验结果如下:

折定义中的阈值 T 取 $0.1\min\{w,h\}$,w 是图像汉字的宽度,h 是图像汉字的高度。

对"皇"字的分型结果是 233,提取的"皇"字上部分(图 1(b))的特征如表 8 所示。

(a) 原图 (b) 整体上下型上图 (c) 整体上下型下图

图 1 "皇"汉字图像

Fig.1 Image of Chinese character "皇"

表 8 "皇"字上部分特征

Tab.8 The upside characteristic of Chinese character "皇"

笔划号	笔划端点坐标	笔形隶属度	最高最左点坐标
①	(32,28)(50,28)	$\mu H_1(\theta)=1$	
②	(35,33)(47,33)	$\mu H_2(\theta)=1$	
③	(30,27)(39,23)	$\mu P_3(\theta)=0.5325$ $\mu H_3(\theta)=0.2013$	
④	(33,40)(48,40)	$\mu H_4(\theta)=1$	(39,23)
⑤	(32,28)(33,42)	$\mu V_5(\theta)=0.5914$ $\mu N_5(\theta)=0.0908$	
⑥	(48,40)(50,28)	$\mu P_6(\theta)=0.2103$ $\mu V_6(\theta)=0.0538$	

如表 8 所示,由端点坐标和隶属度可知,笔划①和⑥构成折笔划“横折”。由最高最左点坐标可知,“皇”字第 1 笔字元为③,通过搜索图像,并根据本文规定的笔顺判断规则,带入端点坐标计算得到的笔划字元顺序为③ > ① + ⑥ > ⑤ > ④ > ②。“皇”汉字图像的上部分(图 1(b))的特征编码如下:统计特征码为 031 ,笔划码为 35210000。

对照子结构特征,“皇”字的上部分(图 1(b))的特征与所有子结构特征不相符。

提取的“皇”字的下部分(图 1(c))特征如表 9 所示。

表 9　“皇”字下部分特征

Tab. 9　The downside characteristic of Chinese character “皇”

笔划号	笔划端点坐标	笔形隶属度	最高最左点坐标
①	(38,51)(50,48)	$\mu H_1(\theta)=0.5321$ $\mu P_1(\theta)=0.3119$	(50,48)
②	(38,58)(51,55)	$\mu H_2(\theta)=0.5668$ $\mu P_2(\theta)=0.2888$	
③	(33,68)(64,63)	$\mu H_3(\theta)=0.6946$ $\mu P_3(\theta)=0.2036$	
④	(45,50)(45,64)	$\mu V_4(\theta)=1$	

由端点坐标和隶属度可知,“皇”字的下部分无折笔划。由最高最左点坐标可知,“皇”字的第 1 笔字元为①,通过搜索图像,再根据本文规定的笔顺判断规则,带入端点坐标计算得到的笔划字元顺序为① > ④ > ② > ③。

“皇”汉字图像下部分(图 1(c))的特征编码如下:统计特征码为 131 笔划码为 12110000。

对照子结构特征,由于“皇”字的下部分(图 1(c))的特征与子结构 a028 的特征相符,因此可用子结构序号编码。编码如下:统计特征码为 131,笔划码为 a0280000。

因此待识别“皇”字的编码为 16223335210000a0280000。

依据本文编码规则,“皇”字的标准编码为 16323332510000a0280000 或 16323335210000a0280000。

对“法”字的分型结果是 133,提取的“法”字左部分(图 2(b))的特征如表 10 所示。

由端点坐标和隶属度可知,“法”字的左部分无折笔划。

由最高最左点坐标可知,“法”字的左部分第 1 笔字元为①,通过搜索图像,再根据本文规定的笔顺判断规则,带入端点坐标计算得到的笔划字元顺序为① > ② > ③。

(a) 原图　(b) 整体左右型左图　(c) 整体左右型右图

图 2　“法”汉字图像

Fig. 2　Image of Chinese character “法”

表 10　“法”字左部分特征

Tab. 10　The left part characteristic of Chinese character “法”

笔划号	笔划端点坐标	笔形隶属度	最高最左点坐标
①	(42,35)(47,38)	$\mu N_1(\theta)=0.6881$	(42,35)
②	(47,44)(39,45)	$\mu H_2(\theta)=0.7625$ $\mu P_2(\theta)=0.1583$	(47,44)
③	(42,55)(46,49)	$\mu P_3(\theta)=0.7487$	(46,49)

“法”汉字图像的左部分(图 2(b))的特征编码如下:统计特征码为 010,笔划码为 41300000。

对照子结构特征,由于“法”字左部分(图2(b))的特征与子结构 a009 特征相符,因此可用子结构序号编码。编码如下:统计特征码为 010, 笔划码为 a0090000。

提取的“法”字右部分(图 2(c))的特征如表 11 所示。

表 11　“法”字右部分特征

Tab. 11　The right part characteristic of Chinese character “法”

笔划号	笔划端点坐标	笔形隶属度	最高最左点坐标
①	(53,41)(61,39)	$\mu H_1(\theta)=0.5321$ $\mu P_1(\theta)=0.3119$	(57,31)
②	(53,48)(64,45)	$\mu H_2(\theta)=0.4915$ $\mu P_2(\theta)=0.3390$	
③	(53,56)(62,53)	$\mu P_3(\theta)=0.4097$ $\mu H_3(\theta)=0.3855$	
④	(57,31)(58,44)	$\mu V_4(\theta)=0.7068$ $\mu N_4(\theta)=0.0977$	
⑤	(52,57)(56,48)	$\mu P_5(\theta)=0.5325$	
⑥	(60,49)(64,58)	$\mu N_6(\theta)=0.5325$	

由表 11 端点坐标和隶属度可知,“法”字右部分无折笔划。

由最高最左点坐标可知,“法”字右部分的第 1 笔字元为④,通过搜索图像,再根据本文规定的笔顺判断规则,带入端点坐标计算得到的笔划字元顺序为④ > ① > ② > ⑤。

“法”汉字图像的右部分(图 2(c))的特征编码如下:

统计特征码为 121,笔划码为 21130000。

对照子结构特征,“法”字的右部分(图 2(c))特征与所有子结构的特征不相符。

因此待识别“法”字的编码为 131133a009000021130000。

依据本文编码规则,“法”字标准编码为 131133a009000021170000。

7 结 论

本文模仿人的容错机制构建了用于机器识别的汉字冗余编码,提出了一种适于机器识字的汉字容错编码方法。该方法先采用笔划形态的模糊隶属度表示、字元代码容错、子结构容错等多级容错方式,使编码识别可以适应手写体笔划的变形和字体的大小变化,如实验中手写体汉字“法”的“氵”部分第 2 笔点写成横;然后按子结构的构成表,就可以容错识别;手写体汉字“皇”的“白”部分的“横折”本应由横和竖相接构成,但手写成横与撇构成,也不影响识别结果。本文对文献[6]中笔顺进行了改进,即利用了笔划构字的拓扑特征,使笔顺的提取不受笔划长短、位置和倾斜角度的影响,取码位数的限定,也允许复杂汉字的笔划密集区撇捺笔划增笔和减笔,同时创建了多模板字典,基本模仿了人认识汉字的机理和容错机制。实验结果表明,该编码方法能很好地表征和区分汉字集;对手写体汉字识别取得了较好的效果,解决了大部分相似汉字的区分问题。但用到的中间特征,如字型提取、笔顺中第 1 笔的确定,还有待于进一步研究。

参考文献(References)

1 Hildebrandt T, Liu W. Optical recognition of handwritten Chinese Characters since 1980[J]. Pattern Recognition, 1993, **26**(2): 205 ~ 225.

2 Zhang Rui, Ding Xiao-qing, Fang Chi. New method of optimal sampling features for offline handwritten Chinese Character recognition [J]. Journal of Image and Graphics, 2002, **7**(2): 176 ~ 180. [张睿,丁晓青,方驰. 脱机手写汉字识别的最优采样特征新方法[J]. 中国图象图形学报, 2002, **7**(2): 176 ~ 180.]

3 Chen You-bin, Ding Xiao-qing, Wu You-shou. Non-specific Person Offline handwritten Chinese Character Recognitiong [EB/OL]. http://www.chinaocr.net/show_hdr.php? xname = TVKUIV0&xpos = 6&dname = , 2005-06-20. [陈友斌,丁晓青,吴佑寿. 非特定人脱机手写汉字识别[EB/OL]. http://www.chinaocr.net/show_hdr.php? xname = TVKUIV0&xpos = 6&dname = , 2005-06-20.]

4 Yang Sen, Zhang Wei-lin, Chen Huai-min, *et al*. Chinese Character input code dictionary on computer[M]. Hefei: University of Science and Technology of China Publishing Company, 1995: 583 ~ 673. [杨森,张伟林,陈淮民等. 计算机汉字输入编码字典[M]. 合肥:中国科学技术大学出版社,1995: 583 ~ 673.]

5 Bian Zhao-qi, Zhang Xue-gong, *et al*. Pattern Recognition [M]. Beijing: Tsinghua University Publishing Company, 2000: 315 ~ 329. [边肇祺,张学工等. 模式识别[M]. 北京:清华大学出版社, 2000: 315 ~ 329.]

6 Chen Zhi-ping, Lin Ya-ping, Li Jun-yi. A recognition algorithm of Chinese character based on stroke segment and order[J]. Journal of Hunan University, 2000, **27**(4): 103 ~ 104. [陈治平,林亚平,李军义. 基于笔划和笔顺的汉字识别算法[J]. 湖南大学学报, 2000, **27**(4): 103 ~ 104.]

7 Wang Zhu-lin. The Study on the Mechanism of Humanoid Recognition Characteristic for Image Character [D]. Hefei: Hefei University of Technology, 2005: 35 ~ 41. [王竹林. 图像字符的仿人认知特征的机理研究[D]. 合肥:合肥工业大学, 2005: 35 ~ 41.]

8 Qian Zi-tuo. Research on Chinese Character Image Recognition[D]. Hefei: Hefei University of Technology, 2005: 34 ~ 38. [钱自拓. 汉字图像识别研究[D]. 合肥:合肥工业大学, 2005: 34 ~ 38.]