一种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,安徽 合肥 230009)

摘 要:提出了一种用于机器识字的汉字容错编码方法。该编码采用统计粗分类和结构细分类相结合的方法,定义了易于机器识别的汉字结构字元集,给出了笔划字元的顺序判断规则。构建了37类子结构的编码和冗余容错编码,建立了仿人构字的汉字编码规则和字典。仿真实验表明,这种编码方法易于机器识别,具有容错性,且拒识和误识率较低。

关键词:汉字编码;字元;汉字特征;容错
中图分类号:TP391.43 **文献标识码**:A **文章编号**:1673-629X(2006)11-0067-03

A Study of Chinese Characters Code of Bearable Mistakes Method

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

Abstract: A kind of Chinese characters codes for computer cognition is presented in this paper. This kind of codes adopts the method based on the combination of stat classification in general with fine structure classification. Elements groups of Chinese characters are made for machine cognition. Rules for judging stroke sequence are given. Thirty-seven kinds of subsidiary configurations codes and bearable mistakes codes are constructed. The code principles and dictionary of Chinese characters are established which agree with apery imitation. Emulational experimental results show that it applies to computer cognition with a low rate of repeated codes and wrong codes.

Key words: Chinese characters code; character elements; Chinese characters characteristic; bearable mistakes

0 引 言

汉字识别是模式识别的一个重要分支,属于多类问题,识别方法较多,但就特征而言,主要可以分为两类:统计方法和结构方法[1]。采用统计或结构方法识别汉字各有优缺点[2]。统计方法具有良好的鲁棒性,较好的抗干扰抗噪声的能力,但区分相似字的能力较差;而结构方法区分相似字的能力较强[2],但计算量大,对归属不明确的线段难以表征或易于产生错误编码。

文中采用统计特征和结构特征相结合的方法,模仿人认识汉字过程,建立了仿人构字的汉字字元编码规则和字典。实验和仿真表明,文中编码方法能很好地表征和区分汉字集,字元提取稳定,重码率低,并具有容错性。

1 汉字的特征分析和选取

汉字国标一、二级字库(GB2312-80)的6763个汉字的统计结果表明,包含横笔划的汉字占99.8%,包含竖笔划的占99.85%,包含撇笔划的占93.5%,包含捺笔划的占76.5%[3]。以上结果分析可知,以横竖撇捺笔划和少量自规定形状的字元,再结合拓扑结构特征,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,根据汉字笔划形态、顺序、数量、相互关系信息可描述和表征每个汉字特征。

1.1 汉字的笔划特征选择

汉字有5种基本笔划:横、竖、撇、捺、折。

从机器认知角度可把"折"看作笔划组合,用横竖撇捺来定义。为了突出不同"折"之间的差异性,并保证基本的构字需要,文中将"折"分为5类:

* 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 ㄱ ㄴ ㄱ

* 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 ㅣ ㅣ ㄴ

* 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 ㇇

* 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 ㇏

* 弯勾:捺笔画终点与左竖勾的起点相接。捺起笔向左弯折 ㇆

考虑提取时的可能误差,规定折笔划的起笔笔划的终点与另一笔划的起点之间距离不超过阈值 T,则认为这两笔划的一个终点与另一个起点重合,构成折笔划。

1.2 汉字的字型特征选择

汉字字型的划分是基于对汉字整体结构的认识,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

收稿日期:2006-02-09

作者简介:王建平(1955-),男,河北藁城人,教授,研究方向为智能测控技术、机器视觉与图像识别系统等。

更明细地划分汉字字型，采用了两级划分法。

汉字字型分为：左右型、上下型和杂合型 3 类。

两级划分法为：首先看整体汉字字型，若为杂合型则不再区分；若为合体字，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。合体字第一级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。例如："侧"一级：整体字型为左右型；二级划分：左边"亻"杂合型，右边"则"左右型。

1.3 汉字字元选取

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，选取的字元如表 1 所示，其它结构均可看作这些字元的有机组合。

表 1 笔划字元代码表

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 一 一	横、横撇、横点
	2	竖	丨	丨 丨 亅 丨	竖、左竖钩、右竖勾、竖撇
	3	撇	丿	一 丿 丿 丿	撇（横撇、竖撇、普通撇）、提
	4	捺	丶	丶 丶 丶	点、捺、斜勾
笔划组合	5	横折	乛	乛 乛 乛 乛	横折
	6	竖折	乚	乚 乚 乚 丨 丨	竖折、竖、左竖勾
	7	撇折	𠂉	𠂉 𠂉	撇折
	8	斜勾	㇏	㇏ ㇏ ㇏	斜勾、捺
	9	弯勾	𠃌	𠃌	弯勾

1.4 汉字字元顺序特征选取

①单笔划字元的笔划顺序判断算法[4]：

单笔划 $S_1=\{(X_{11},Y_{11}),(X_{12},Y_{12})\mid X_{11}\leqslant X_{12}\}$ 与笔划 $S_2=\{(X_{21},Y_{21}),(X_{22},Y_{22})\mid X_{21}\leqslant X_{22}\}$ 的笔划字元顺序可以分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，其判定公式为：

$$-1<\frac{Y_{21}+Y_{22}-Y_{11}-Y_{12}}{X_{21}+X_{22}-X_{11}-X_{12}}\leqslant 1$$

（当 $X_{11}+X_{12}-X_{21}-X_{22}\neq 0$）

$$Y_{22}+Y_{21}-Y_{12}-Y_{11}<0 \qquad (1)$$

（当 $X_{11}+X_{12}-X_{21}-X_{22}=0$）

若满足式(1)的不等式条件则可认为笔划字元 S_1 的笔顺要优先于笔划字元 S_2 的笔顺。

②折笔划字元以它的起笔笔划与其他字元按①中方法判断笔划字元的笔顺。

③汉字字元的顺序选取规则：

依照 1.2 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。

第一笔字元：取这部分汉字图像中最高点（如有几个点，取最左边的）所在字元为第一笔，若有两个字元，则按①和②中方法比较笔划字元顺序，顺序优先的为第一笔。

第二笔字元：与第一笔相交或相连（包括相切和相接）的所有笔划字元按①和②中方法比较笔划字元的笔顺，笔划字元笔顺最优先的为第二笔；若没有与第一笔相交或相连的字元，则除第一笔外，取这部分汉字图像中最高点（如有几个点，取最左边的）所在字元为第二笔，若有两个字元，取顺序优先的为第二笔。

第三笔字元：第三笔的确定方法同第二笔，从与第二笔相交或相连的字元中选取，或取除第一、二笔外汉字图像中最高点所在字元。

第四笔及以后的笔划字元顺序选取依次类推。

1.5 汉字统计特征选取

对手写体或印刷体汉字而言，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。汉字中横、竖笔划出现频率最高，同时，机器对汉字横、竖笔划提取准确稳定。用这 3 种汉字的全局统计特征首先对汉字集进行粗分类，可提高识别速度，具有较强的鲁棒性。

1.6 汉字的子结构选取

对汉字的研究分析，汉字存在多种结构稳定的部首或字根，称之为子结构[5]。文中归结出 37 类易于机器识别的常用子结构，具体如表 2 所示。

表 2 子结构的特征判断和字元构成表

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
a000	彡(左、右)	0	3330	1330	3300			
a001	阝(左、右)	0	2590	2544	2543			
a002	大(上、下)	1	3140	2134	2143			
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	2240				
a007	讠(左)	0	4530	4160	4510	4120	4500	
a008	礻或礻(左)	0	4527	4524	4132			
a009	氵或氵(左)	0	4430	4440	4410	4300	4400	4100
a010	彳或亻(左)	0	3320	1320	3200	1200		
a011	纟(左)	0	7730	7710	7313	7311	3171	3173
a012	饣(左)	0	3560	3520	3120	3160		
a013	钅(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			
a018	灬(下)	0	3444	4444	4440	3440		
a019	艹(上)	2	2120	4120	4130	2130		
a020	宀或冖(上)	0	4520	4540	4530	5200	5400	5300
a021	亠(上)	0	4100					
a022	厶	0	7400	3140				
a023	夂或夂	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		

(续表 2)

a028	王	1	1211	1213	1311	1313	
a029	木	1	2134	2334			
a030	力	1	3500	2530	2500	3120	2123
a031	山	0	2620	2122	2322		
a032	贝	0	2534	5234	2123	1223	
a033	工	0	1210	1230	1310	1330	
a034	口	0	2510	5210	1212	2121	
a035	子	1	5210	5230	1321	1323	
a036	又	1	5400	1340	4500	4130	

根据汉字字型特征划分,对各部分的起始笔划字元、字元顺序和相交点进行判断[5],所取的笔划字元数不大于4,将其与子结构的相比较,若与某种子结构的相符合,则认为待识别汉字具有此种子结构[5]。

2　汉字编码规则

文中编码可以用公式表示: 汉字编码 = 统计特征码(3 码) + 字型笔划码(16 码)。

2.1　汉字整体统计特征码编码规则

整个汉字编码的前三码为汉字整体统计特征码,规则如表 3 所示[3]。

表 3　汉字整体统计特征编码规则表

编码位置	首码	二码	三码
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

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 $M/5$(M 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,还记为 9。如:"齉"的统计特征码为 199。

2.2　汉字字型笔划码编码规则

(1)按 1.4 节规定的笔划字元顺序取码。字元以表 1.2 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 码),规则如下:

①整体杂合型笔划码编码规则:按(1)~(4)的规则取汉字前十码,不足码补 0。后边六码补 0。

②整体左右型笔划码编码规则:左和右部分各 8 码。左和右任何一部分为杂和型,则按字元顺序取前四码,不足补 0,在余下四码补 0;若可以再分,每一再分的部分分别按字元顺序取前四码,不足补 0。然后检测每部分的前四码是否有与子结构特征相符合的(交点数 and 笔划码 or 交点数 and 容错码)。若有,则这部分笔划码以子结构序号码编码;若没有,保留原取码。

③整体上下型笔划码编码规则:类同整体左右型。

3　重码与容错码的处理

3.1　重码字的处理

文中编码由于是基于笔划的,因此对笔划形态、数目和笔顺完全一致的汉字,容易造成重码。因此,必须借助于别的特征进行判别。经统计分析,采用 19 位汉字编码,汉字重码数量很少,仅为笔划形状、数目和笔顺一致的简单的汉字。

对重码汉字,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:

$$T = \{(i,j,r) \mid r \in R\} \tag{2}$$

其中:(i,j,r) 代表汉字第 i 个笔划字元和第 j 个笔划字元存在着 r 的对比关系;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)。

例如:"未"、"末"按文中编码都为 2212113400000000000,建立对比关系区分如下:

$T_{未} = \{(2,3,短于)\}$;$T_{末} = \{(2,3,长于)\}$。

利用汉字模板中的字元对比关系检测待识别汉字是否满足其对比关系,若满足可认为备选汉字即为识别的汉字;若不满足,则进行下一个编码相同汉字的比较过程[4]。

"工"、"士"、"干",和"甲"、"由"和"申"等相似字按文中编码方法是可以区分的。

3.2　冗余容错编码机制

鉴于机器识别汉字笔划字元可能存在的不完整和不一致性,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。

(1)归结出 37 类易于机器识别的常用子结构,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。如表 2 所示。

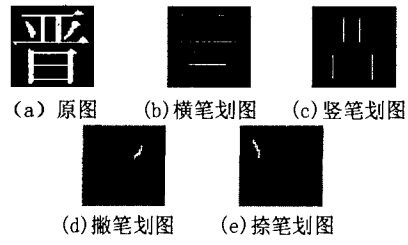
(2)归结出机器能够认知的 9 类笔划字元,并针对这 9 种字元在提取时可能混淆的笔划,给出了多归类容错编码(笔划编码时有相互包含容错的关系)。具体如表 1 所示。

(3)给出了汉字整体统计特征编码(前 3 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,允许分别给这 3 类码数值加 1(或减 1)再识别的容错编码,进而实现容错识别。

文中编码方法优点是:一个汉字有多个编码,但很少出现一个编码多个汉字情况,有易于实现机器识字的容错性。

4　实　验

依据文中编码规则所确定的标准编码和采用文中编码识别算法对汉字图像(32×32 点阵)提取的编码,如图 1 所示。



(a) 原图　(b)横笔划图　(c)竖笔划图

(d)撇笔划图　(e)捺笔划图

图 1　汉字"晉"图像编码提取图

(下转第 72 页)

常见的核函数有：

* 多项式核函数：$K(x,y)=[s(x\cdot y)+c]^d, d\in N$

* RBF(Radial Basis Function) 核函数：$K(x,y)=\exp[-\|x-y\|^2/(2\sigma^2)]$

* Sigmoid 核函数：$K(x,y)=\tanh[s(x,y)+c]$

3 实验与分析

实验中分类模型所用的训练和测试文本来自互联网，按类别存放。每个类别的训练集和测试集中都含有 300 个文本文件。首先采用双向匹配法对文本进行分词，再使用 X^2 统计评估函数对分词所得的特征全集进行特征提取，将文本表示成多维空间向量，最后使用支持向量机分类算法建立文本分类器。评判分类效果一般用查全率(recall)和查准率(precision)表示分类效果，文中综合了查全率和查准率在评判中的因素，引入 F 测度值作为分类效果的评判指标，定义如下：

查全率 = 事实属于此类且被分类正确的文档数 / 属于此类的总文档数；

查准率 = 事实属于此类且被分类正确的文档数 / 被判为此类的文档数；

F 测度值 = 2×查全率×查准率 /(查全率 + 查准率)。

实验结果如表 1 所示。

从表中可以看出，这种采用支持向量机算法建立的分类模型普遍具有较高的 F 测度值，其中以线性 SVM 和多项式核 SVM 这两种分类器的分类效果最为平均，Sigmoid 核 SVM 分类器具有相对最高的单类别 F 测度值，RBF 核 SVM 分类器具有相对最低的单类别 F 测度值。由此可见，SVM 分类器的参数选择对分类效果的影响与训练集和测试集的选择有很大关联，参数选择对分类效果的影响并不具备明显的规律性。

表 1 分类效果实验结果

文本类别	政治	经济	农业	环境	计算机	体育
线性 SVM	92.73%	86.27%	90.47%	89.40%	95.987%	92.23%
多项式核 SVM	92.88%	85.13%	89.25%	88.77%	94.98%	91.58%
RBF 核 SVM	96.94%	91.13%	78.96%	81.86%	93.65%	97%
Sigmoid 核 SVM	96.68%	91.09%	79.77%	84.26%	95.47	97.27%

4 结束语

支持向量机是一种已被证明有效的学习算法，在文本分类领域中有着广泛和深入的应用。文中介绍了中文文本分类实现过程，阐述了线性和非线性支持向量机的原理，设计出了基于向量机的文本分类模型并进行了实验分析。此外，还对不同支持向量机分类器的性能作了比较。实验表明，这种模型在小样本条件下获得了较好的分类效果。

参考文献：

[1] 贾 洞，梁永帧. 基于支持向量机的中文网页自动分类[J]. 计算机工程，2005，31(10)：145－147.

[2] 邱均平，文庭孝，周黎明. 汉语自动分词与内容分析法研究[J]. 情报学报，2005，24(3)：309－315.

[3] 孙国菊，张 杰. 中文文本分类的特征选取评价[J]. 哈尔滨理工大学学报，2005，10(1)：76－78.

[4] 李 亮，刘万春，徐泉清，等. 一种基于支持向量机的专业中文网页分类器[J]. 计算机应用，2004，24(4)：58－61.

[5] Vapnik V N. The Nature of Statistical Learning Theory[M]. New York：Springer－Verlag，2000.

(上接第 69 页)

依据文中编码规则“晋”字标准编码：

0541212000052110000 或 0541212000025110000

根据图 1(b)～(e)，对“晋”字提取编码：

0541212000052110000

5 结 论

文中采用汉字统计特征粗分类和笔划结构细分类相结合的方法，模仿人认识汉字的过程，提出了一种适于机器识字的汉字容错编码方法。

(1)定义了仿人拆字的字元集，给出了易混淆笔划字元的容错编码。

(2)给出了笔划字元的顺序判断规则。

(3)归结了 37 类简单常用的部首为子结构编码，并给出冗余的容错编码。

(4)建立了仿人构字的编码规则。

实验和仿真结果表明，该编码方法的字元易于稳定提取，能很好地表征和区分汉字集。

参考文献：

[1] 朱 辉，杨 扬，颉 斌，等. SVM 在小字符集手写体汉字识别中的应用研究[J]. 微计算机信息，2004，20(4)：74－75.

[2] 陈友斌，丁晓青，吴佑寿. 非特定人脱机手写汉字识别[EB/OL]. 2005－03－20. 中国 OCR 信息网. http://www.chinaocr.net/show_hdr.php?xname=TVKUIV0&xpos=6&dname=.

[3] 钱自拓. 汉字图像识别研究[D]. 合肥：合肥工业大学，2005：34－38.

[4] 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.

[5] 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，2005：34－38.