基于网格的上中下结构汉字的结构识别研究

王素利，皮佑国，梁添才，丘志文

（华南理工大学 自动化科学与工程学院，广东 广州 510641）

摘 要： 采用简易网格，以GB2312规定的二级汉字基本集中的上中下结构汉字作为研究对象，研究了汉字的结构识别。实验表明了方法的有效性，并给出了分析和实验过程。

关键词： 结构；网格；识别

中图分类号： TP391.12 **文献标识码：** A **文章编号：** 1673-9833(2007)06-0098-03

Recognition Research on Structure for Chinese Character of Up-Center-Down Structure Based on Grid

Wang Suli，Pi Youguo，Liang Tiancai，Qiu Zhiwen

（School of Automatic Science and Engineering，South China University of Technology，Guangzhou 510641，China）

Abstract： In view of brief grid, the research object is Chinese character about up-center-down structure in secondary Chinese character basic set within GB2312. It studies recognition for structure of Chinese character. The experiment proves that method is feasible and also put forward the experiment process.

Key words： structure；grid；recognition

0 引言

汉字作为世界上仅存的最古老的表意文字，字形是它的本体[1]，它用形象具体的形式表达一般抽象的内容，集中体现了中华民族的思维方式，是中华民族集体智慧的结晶。

汉字是拼合文字，是由部件（含笔画、偏旁和部首）按一定规则拼合而成的。汉字的部件是由笔画组成的具有组配汉字功能的构字单位[2]（国家语评）。虽然拼音文字也属于拼合文字，但它以首字母为起点顺序排列，字母的竖直位置、形态和大小都不变[3]；而汉字在用部件组成合字时，根据汉字结构的不同，部件的位置、形态和大小都会发生一定的变化，所以汉字的拼合规则要比拼音文字的拼合规则复杂得多。

汉字的一级结构有：左右结构、左中右结构、上下结构、上中下结构、全包围结构、半包围结构、整体结构、品字结构8类。GB2312规定的二级汉字基本集中的汉字6 763个，其中上中下结构的汉字324个。本文以GB2312规定的二级汉字基本集中的上中下结构汉字为研究对象，研究上中下结构汉字的结构识别。

上中下结构汉字：该汉字由上中下3个或3个以上部件组成。按照结构分，上中下结构的汉字由上部、中部、下部组成。上中下结构汉字的结构的识别，要求准确地识别出汉字的上部、中部、下部。

在刚刚开始学习写字的时候，一般会利用田字格(2 × 2网格)和九宫格(3 × 3网格)，如图1所示。田字格和九宫格有助于初学写字的人对于汉字结构的认知，根据网格的这种特点，基于传统的汉字结构的认知机理，本文利用网格实现上中下结构汉字的结构的识别。

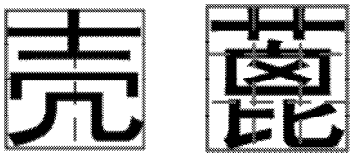

图1 田字格和九宫格

Fig. 1 Tian grid and Nine Gong grid

收稿日期： 2007-08-30

作者简介： 王素利（1979-），女，辽宁锦州人，华南理工大学硕士生，主要研究方向为图像处理；
皮佑国（1953-），男，重庆开县人，华南理工大学教授，博士生导师，主要研究方向为智能控制理论与应用.

1　识别机理

1.1　分析工具

我们采用如图 2 所示的 2 × 2、3 × 3 简易网格，及其扩展的 4 × 4、6 × 6、9 × 9 网格来分析和描述汉字结构[4, 5]。

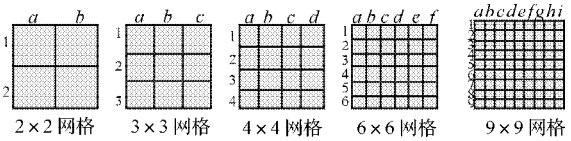

图 2　简易网格

Fig. 2　Simple grid

1.2　识别对象分析

选取 GB2312 二级汉字基本集中的上中下结构汉字作为研究对象，字体为黑体。将其放到网格中，示例如图 3 所示。

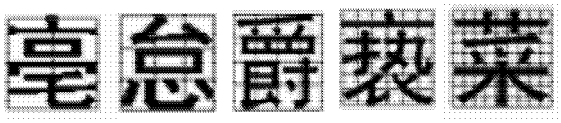

图 3　网格分析示例

Fig. 3　Grid description for structure of Chinese character

上中下结构汉字的上部、中部、下部之间的界限分别在上 1/3，2/3 处；上 1/6，4/6 处；上 1/3，1/2 处；上 3/6，5/6 处；上 1/4，3/4 处；上 1/4，2/3 处；上 1/2，2/3 处；上 1/4，5/9 处；上 1/4，2/4 处；上 1/4，5/6 处；上 1/3，3/4 处；上 1/6，5/6 处；上 1/4，4/9 处；上 1/6，3/6 处；上 1/4，1/3 处；上 1/4，5/6 处。共 16 种情况。

经分析表明：研究对象的上、中、下各部之间位置关系如图 4 所示，分为如下 4 种。

1）独立：汉字各部分开为独立整体。

2）交叠：各部间无接触，但无法简单地用水平分割线分割。

3）粘连：某部在一点或几点与相邻部接触。

4）粘连且交叠：粘连与交叠的情况并存。

a）独立　　b）交叠　　c）粘连　d）粘连且交叠

图 4　各部间的位置关系

Fig. 4　Place relation of every part

汉字各部分的连通性分为：

1）上部（中部或下部）为单连通区域。

2）上部（中部或下部）为非单连通区域，它们都处在网格的同一横向子格中，如“灬”、“巛”“心”。

3）上部（中部或下部）为非单连通区域，它们不处在网格的同一横向子格中，如“彡”、“穴”。

4）存在比较小的连通区域，如“丶”。

2　识别方案

对于识别对象分析中介绍的汉字的上部、中部、下部的位置关系情况有：

独立：可以根据各部的上下位置关系识别出上部、中部、下部。

粘连、粘连且交叠：首先进行去除粘连部分的处理，进行去除粘连处理后的上部、中部、下部之间的位置关系，如果是独立的，可以根据各部的上下位置关系识别出上部、中部、下部。去除粘连部分的处理只能解决那些粘连不太严重的部分，对于粘连严重的部分，则需要对存在粘连的部分进行分割，进行分割时需要考虑以下条件：

1）封闭区域不分割。在候选连通区域中，若有封闭区域，则不做分割。

2）极小值处分割（必要条件）。候选分割处于水平投影的波谷或局部极小值处。

3）横向分割。分割只考虑横向分割情形，不考虑纵向分割。

4）突变地方（必要条件）。

连通区域描述：$C_{\text{contour}}=\{F(x_i,y_i),R_{\text{Rect}},S,\Omega\}$，其中，$F(x_i,y_i)$为连通区域像素集；$S$ 为连通区域面积，即连通区域像素点个数；R_{Rect} 为连通区域的外接矩形，$R_{Rect}=(\min(x_i),\min(y_i),\max(x_i),\max(y_i))$；$\Omega$为外接矩形像素点集，且$\Omega=\{f(x_i,y_i)|\min(x_i)\le x_i\le\max(x_i),\min(y_i)\le y_i\le\max(y_i)\}$。

对于识别对象分析中介绍的汉字的上部、中部、下部的位置关系情况的交叠则需要根据连通体的位置情况进行结构识别。

对于识别对象分析中介绍的汉字的各部内的连通区域情况：即对于第 2、3、4 种情况，首先要对连通区域进行合并，然后再根据各部的连通区域，投影，封闭区域关系等条件综合识别汉字的结构。

根据上中下结构汉字的结构特征设计如下结构识别方案：

1）读入汉字图像。

2）判断图像中是否含有粘连部分，如果有则进行去除粘连部分的处理，如果没有则转到下一步。

3）连通区域标记。对图像中的每个连通区域进行标记。

4）连通区域合并。若上部（中部或下部）为单连通区域，则不必进行合并，直接进入下一步骤。

连通区域关系定义：设 C_1、C_2 为两个连通区域，C_1 的外接矩形为：$R_{\text{Rect}C_1}=\left(\min\left(x_i^1\right),\min\left(y_i^1\right),\max\left(x_i^1\right),\max\left(y_i^1\right)\right)$，外接矩形点集为：$\Omega_{C_1}=\left\{f(x_i,y_i)=1|\min\left(x_i^1\right)\le x_i\le\max\left(x_i^1\right)\right.$，

$\min\left(y_i^1\right) \le y_i \le \max\left(y_i^1\right)\}$。$C_2$ 的外接矩形为：

$R_{\mathrm{Rect}C_2}=\left(\min\left(x_i^2\right),\min\left(y_i^2\right),\max\left(x_i^2\right),\max\left(y_i^2\right)\right)$，外接矩形点集为：$\Omega_{C_2}=\{f(x_i,y_i)=1|\min(x_i^2)\le x_i\le\max(x_i^2),$

$\min\left(y_i^2\right)\le y_i\le\max\left(y_i^2\right)\}$。

连通区域合并共由以下 3 个部分组成：

a. 任意两个连通区域位置关系的判定，若为非上下关系，则进行合并，否则保留。

上下关系判定条件为：

i）$\min(y_i^1)>\max(y_i^2)$；

ii）$\min(y_i^2)>\max(y_i^1)$；

iii）$\{\max\left(y_i^1\right)>\max\left(y_i^2\right),\min\left(y_i^1\right)<\max\left(y_i^2\right),$

$\left(\max\left(y_i^1\right)+\min\left(y_i^1\right)\right)/2>\max\left(y_i^2\right)\}$；

iv）$\{\max\left(y_i^2\right)>\max\left(y_i^1\right),\min\left(y_i^2\right)<\max\left(y_i^1\right),$

$\left(\max\left(y_i^2\right)+\min\left(y_i^2\right)\right)/2>\max\left(y_i^1\right)\}$。

若 C_1、C_2 满足以上任一条件，即可判定为上下关系。

b. 面积小于某一阈值的连通区域就近合并。

c. 面积最小的连通区域就近合并。

5）根据封闭区域、极小值处分割、横向分割、突变地方等条件，对存在粘连的部分进行分割。

6）根据图像的连通区域位置，水平投影，封闭区域等条件再结合网格识别出上部、中部、下部。

3 实验与评论

3.1 实验步骤与结果

1）取出GB2312规定的二级汉字基本集中的324个上中下结构的汉字，将该汉字制作成位图图像，汉字字体统一为黑体，以该图像作为处理对象。

2）图像的预处理。

3）依据上中下结构汉字的结构的网格识别规则，设计算法及编写程序，建立上中下结构汉字的结构识别系统。上中下汉字结构识别系统如图 5 所示。

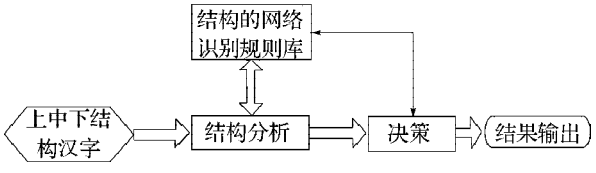

图 5 上中下结构汉字的结构识别系统

Fig. 5 Describing system for structure of Chinese character about up-center-down structure

4）对GB2312规定的二级汉字基本集中的324个上中下结构的汉字样本集进行测试，记录测试结果。

5）对比人工识别和计算机识别的异同，分析不同部分产生的原因。

实验实现了利用网格对上中下结构的汉字的结构识别，GB2312规定的二级汉字基本集上中下结构汉字共324个，正确地识别出303个，有21个汉字的结构不能正确识别，结构的识别正确率达到93.5 %。

3.2 分析与评论

实验过程中的识别错误主要由以下原因造成：

1）组成汉字的各部件之间粘连严重，候选分割位置处存在封闭区域。

2）汉字结构复杂，难以确定上部、中部、下部各部之间的精确界限。如图 6 所示。

图 6 不能正确切分的汉字

Fig. 6 Chinese character of wrong segment

4 结论

1）实验结果表明，基于网格的上中下结构汉字的结构识别方法有效，这为汉字结构信息化研究提供了一条新途径，也将促进汉字信息化的发展。

2）网格可以作为汉字结构识别的工具。对二级汉字基本集左中右结构汉字的结构识别率超过92 %。

3）对难以确定各部界限的汉字结构的识别，尚需进行深入研究。

参考文献：

[1] 张晓明. 二十世纪汉字字形结构研究[J]. 语言教学与研究，2004（5）：75-79.

[2] GF 3001-1997，信息处理用GB13000.1字符集汉字部件规范[S].

[3] 李晓辉，吴 蓓，董 武，等. 基于部件特征的分类方法以及在汉字识别中的应用[J]. 微电子学与计算机，2003，10：17-19.

[4] 皮佑国，牟总斌. 在计算机中描述汉字的网格及其描述方法：中国，200410015239.2[P]. 2004-12-29.

[5] Liang Tian-cai，Qiu Zhi-wen，Pi You-guo. Simple Grid Based on Cognitive Mechanism and Application Research on Description for Structure of Chinese Character[C]//The 26th Chinese Control Conference. Zhangjiajie: IEEE Computer Society，2007：689-693.