# 基于特征点的汉字字体识别研究

王　恺[①]　　靳简明[②]　　史广顺[①]　　王庆人[①]
[①](南开大学机器智能研究所　天津　300071)
[②](NEC 中国研究院　北京　100084)

**摘　要**：该文提出了整体分析法和个体分析法的概念，并在分析它们各自适用范围的基础上，指出个体分析法更适合于解决印刷体汉字字体识别。在此基础上，提出一种基于特征点的个体分析法来解决汉字字体识别问题，与以往方法相比，该方法具有 3 个优点：识别可信度可控；处理速度快；适用于多语混排情况。实验结果表明，该方法有效解决了印刷体汉字字体识别问题，其性能大大优于以往方法。

**关键词**：字体识别；光学字符识别；特征点
**中图分类号**：TP391.43　　**文献标识码**：A　　**文章编号**：1009-5896(2008)02-0272-05

# Chinese Font Recognition Based on Feature Point

Wang Kai[①]　　Jin Jian-ming[②]　　Shi Guang-shun[①]　　Wang Qing-ren[①]
[①](*Institute of Machine Intelligence*, *Nankai University*, *Tianjin* 300071, *China*)
[②](*NEC Laboratories*, *Beijing* 100084, *China*)

**Abstract**: Global analysis method and individual analysis method are proposed in this paper. By analyzing their traits, it is concluded that individual analysis method is more suitable for machine-printed Chinese font recognition. A feature point based individual analysis method is proposed to resolve Chinese font recognition problem. Compared with previous methods, there are mainly three advantages: The recognition reliability is controllable; the processing speed is fast; it is suitable for multi-lingual document image processing. Experimental results show that the proposed method is more effective than previous methods.

**Key words**: Font recognition; Optical character recognition (OCR); Feature point

## 1　引言

我国自 70 年代末 80 年代初开始进行汉字识别方面的研究工作，经过二十多年来的努力，成熟的中文 OCR 软件已经应用于实际中，为中文书籍的电子化做出了巨大贡献。然而，在中文 OCR 中还存在一些亟待解决的问题：一方面，现有的中文 OCR 系统往往将所有字体混合识别，随着待识别字体的增多，必然会造成误识率的上升和识别速度的下降；另一方面，复杂版面的恢复，实现文档的所见即所得。这两方面都涉及到了字体识别问题：在中文 OCR 系统中加入字体识别模块，根据字体识别结果将图像送入相应字体的字符识别器中，这可以很好地解决上述第一个问题；字体信息是版面恢复的内容之一，正确的字体信息有助于提高版面恢复的精度。

然而，汉字字体识别这一研究课题尚未引起学者们的足够重视，仅有少数文章进行过这方面的研究工作。本文认为，以往关于汉字字体识别的研究工作可以分为两类：整体分析法和个体分析法。

(1)整体分析法　在整体分析法中，以整块文字区域图像作为处理对象，经过频域变换获取到用于字体分类的特征。当前，采用这种方法的研究工作较多。比如，文献[1]基于多尺度非冗余小波纹理分析抽取字体分类特征；文献[2]利用 Gabor 滤波器提取文字区域的全局纹理特征作为字体分类特征；文献[3]利用小波包对文字区域图像作多级分解，提取用于字体分类的纹理特征；文献[4]和文献[5]基于经验模式分解从文字区域图像中抽取用于字体分类的特征。

(2)个体分析法　在个体分析法中，以单个汉字的字符图像作为处理对象。比如，文献[6]对单个汉字的字符图像进行小波分解，并在变换图像上提取小波特征，该方法在不知道汉字内容的前提下，识别单个汉字的字体。

一般来说，整体分析法不需要切分出单独的字符图像，非常适用于难以进行字符切分的情况。然而，印刷体汉字的切分并不困难[7]，整体分析法的这一优势在汉字字体识别中无法体现。此外，与个体分析法相比较，采用整体分析法进行印刷体汉字的字体识别，还存在以下两点不足之处：(1)从统计学的角度来说，个体分析法可以通过多个汉字投票表决来确定字体，并且随着参与字体识别的个体数目的增多，其分类可信度能够持续上升；而整体分析法的分类可信度则很难提升。(2)文献[1–5]均未考虑中文文档中夹杂着英文的情

2006-07-31 收到，2007-05-07 改回

况，实际上，随着全球一体化，多语文档的出现越来越普遍，文档中其它语种的存在必然会对整体分析法的性能造成很大的负面影响。

因此，对于汉字字体识别来说，个体分析法更为适用。目前，仅有文献[6]采用个体分析法，但文献[6]中的方法也存在一些问题：(1)在未知汉字内容的情况下识别字体，认为每个汉字对字体识别所起的作用是确定性的；而实际上，不同字体中不同汉字的相似度不同，对字体识别所起的作用大小也必然是随机变化的。不认识到这一点，难以构造出一个具有高稳定性的汉字字体识别器。(2)对每一个汉字图像都通过小波分解抽取特征，这大大增加了计算复杂度，难以应用于实际中。

鉴于以往工作的不足，本文提出一种基于特征点的个体分析法来解决汉字字体识别问题。特征点方法在文献[8]中首先应用于汉字识别中，并通过大量实验证明了该方法具有较好的性能。之后，特征点方法被应用于相似字的判别中，表现出了优良的性能[9–11]，说明特征点方法具有对细微差别进行辨别的能力，可以较好地解决相似字的混识问题。同时，不同字体同一汉字的整体结构是一致的，仅在一些细节上存在着差别，这与相似字的判别类似，因此，将特征点的方法用于汉字字体识别完全可行。本文提出的基于特征点的个体分析法具有：稳定性高、简单高效、通用性好的优点，所以比以往方法具有更好的实用性。

本文的内容组织如下：第 2 节讨论多字符字体识别原理，第 3 节介绍基于特征点的字体识别，第 4 节给出基于汉字字体识别的汉英混排 OCR 系统框架，第 5 节是实验结果及分析，第 6 节是对本文工作的总结。

## 2　多字符字体识别原理

不同字体中不同汉字的相似度不同，对字体识别所起的作用大小也必然是随机变化的。为了能够反映这种变化，每个汉字 $h$ 的第 $i$ 个字体识别模板中都包含一个 $n$ 维向量：

$$\boldsymbol{\alpha}_{hi}=[p(\theta_{hi}\mid w_1),p(\theta_{hi}\mid w_2),\cdots,p(\theta_{hi}\mid w_n)]^{\mathrm{T}} \tag{1}$$

其中 $n$ 为待识别字体数目，$\theta_{hi}$ 表示具体识别模板，$p(\theta_{hi}\mid w_j)$ 表示输入汉字 $h$ 属于字体 $w_j$ 时匹配模板为 $\theta_{hi}$ 的概率。$p(\theta_{hi}\mid w_j)$ 需要通过对实际样本的学习获得，同其它大多数基于样本的学习方法一样，训练样本越多，所得到的结果越逼近真正分布。

设用于字体识别的汉字集合为 $h_a(a=1,2,\cdots,m)$，与之匹配的字体识别模板分别记为 $\theta_{h_a i_a}$，则根据贝叶斯公式即可得字体识别结果为 $w_j$ 的概率为

$$\begin{aligned}p(w_j\mid\theta_{h_1i_1},\theta_{h_2i_2},\cdots,\theta_{h_mi_m})&=\frac{p(\theta_{h_1i_1},\theta_{h_2i_2},\cdots,\theta_{h_mi_m},w_j)}{p(\theta_{h_1i_1},\theta_{h_2i_2},\cdots,\theta_{h_mi_m})}\\&=\frac{\prod_{k=1}^{m}p(\theta_{h_ki_k}\mid w_j)\cdot p(w_j)}{\sum_{r=1}^{n}\left(\prod_{k=1}^{m}p(\theta_{h_ki_k}\mid w_r)\cdot p(w_r)\right)}\end{aligned} \tag{2}$$

为保证能够以最大的概率得到正确的结果，应该判断这些汉字所属字体为

$$w_c=\arg\max_{r=1,\cdots,n}(p(w_r\mid\theta_{h_1i_1},\theta_{h_2i_2},\cdots,\theta_{h_mi_m})) \tag{3}$$

通过增加用于字体识别的汉字数目 $m$，可以提高字体识别可信度，只有可信度高于预先定义的期望可信度时，才返回字体识别结果。因此，与以往方法相比，本文方法具有更高的稳定性。

## 3　基于特征点的字体识别

### 3.1 预处理

由于扫描分辨率、字号等的不同，会造成同一汉字的图像大小不同，为了获得稳定的汉字结构信息，须在处理前将汉字图像归一成为同一尺寸。

### 3.2 稳定点图生成

根据式(4)和式(5)，得到汉字 $h$ 的稳定黑点图 $B_h$ 以及稳定白点图 $W_h$。不失一般性，以 1 表示黑，0 表示白。

$$B_h=\bigcap_i I_{h_i} \tag{4}$$

$$W_h=\bigcup_i I_{h_i} \tag{5}$$

式(4)的物理意义是：对于一个汉字 $h$，如果它的所有训练字样 $h_i$ 在点 $(x,y)$ 处的值都为 1，那么在该汉字的稳定黑点图中该点取值为 1，否则取值为 0；式(5)的物理意义是：对于一个汉字 $h$，如果它的所有训练字样 $h_i$ 在点 $(x,y)$ 处的值都为 0，那么在该汉字的稳定白点图中该点取值为 0，否则取值为 1。

### 3.3 黑特征点抽取

首先介绍黑特征点抽取过程中用到的 3 个操作，如图 1 所示。

(a) 原字符图像　(b) 脱壳处理后图像

(c) 对 (b)H 操作后图像　(d) 对 (b)V 操作后图像

图 1　脱壳处理，H 操作和 V 操作示意图

(1)脱壳处理　先以行序优先方式扫描 $B_h$，当黑点与白点水平相邻时，若删除该黑点不破坏字符连通性，则删除该黑点。再以列序优先方式扫描 $B_h$，当黑点与白点竖直相邻时，若删除该黑点不破坏字符连通性，则删除该黑点。

(2)H 操作　以行序优先方式扫描 $B_h$，如果连续黑点个数大于一定数值，则保留两端黑点并将中间黑点变为白点，

如图 2 所示。

(3)V 操作　以列序优先扫描 $B_h$，如果连续黑点个数大于一定数值，则保留两端黑点并将中间黑点变为白点，如图 3 所示。

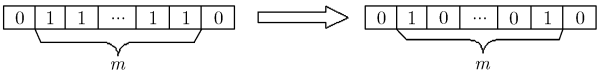


图 2　H 操作

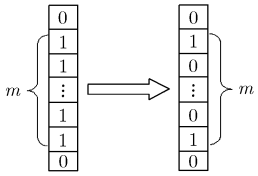


图 3　V 操作

设 $C_h$ 是 $B_h$ 两次脱壳处理后的点阵图。令 $B_1=\mathrm{V}(\mathrm{H}(C_h))$，$B_2=\mathrm{H}(\mathrm{V}(C_h))$，由式(6)得到只包含候选黑特征点的点阵图 $B_0$。

$$B_0(i,j)=\begin{cases}1, & B_1(i,j)=1\text{或}B_2(i,j)=1\\ 0, & \text{其他}\end{cases} \tag{6}$$

以行序优先方式扫描 $B_0$，选择大小为 $m\times m$ 的窗口，若窗口中候选特征点的数目多于两个，则将它们合并成一个特征点，位于原来几个候选特征点的重心上。

### 3.4 白特征点抽取

根据 $W_h$ 把背景点区域分隔为若干凸区域，每个凸域的中心点作为白特征点。

### 3.5 基于特征点的汉字字体识别

一幅汉字图像包含两方面的信息：字符识别信息和字体识别信息。因此，在抽取特征点的过程中，同样可以从两方面来进行：

(1)抽取字符识别特征点　将所有字体的汉字混合，抽取特征点时只考虑如何区分字符，而不考虑区分字体。

(2)抽取字体识别特征点　对不同字体的每一汉字分别抽取特征点，并选出那些能够反映字体差异的特征点。如图 4 所示是字体识别特征点抽取示例，其中抽取出来的特征点是为了区分宋体与其它 6 种字体。从图中可以明显看出，黑体、楷体、隶书和魏体这 4 种字体与特征点的匹配程度很低；仿宋和幼圆这两种字体与特征点的匹配程度虽然看上去稍高一些，但实际上，仿宋与 3 个特征点无法匹配，而幼圆与 4 个特征点无法匹配；对于宋体来说，则可以与所有特征点匹配。可见，仅通过 7 个特征点，即可以很好地区分宋体和其它 6 种字体。

对于单幅汉字图像，首先通过字符识别特征点匹配方法

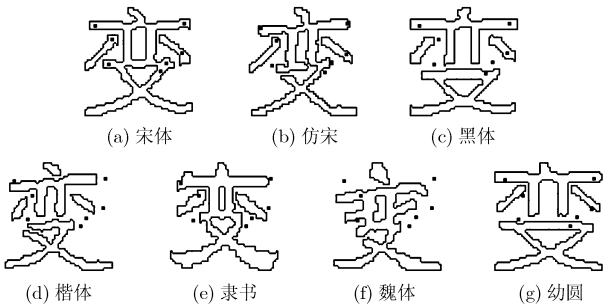


图 4　字体识别特征点示例图

识别字符内容，然后根据该字符对应的字体识别特征点识别字体。

本文提出的基于特征点的汉字字体识别方法，在系统具体实现时只需在空间域中进行特征点匹配，因此，比以往基于频率域的方法更加快速、实用。

## 4　基于汉字字体识别的汉英混排 OCR 系统框架

基于汉字字体识别的汉英混排 OCR 系统框架如图 5 所示，在文献[7]中提出的汉英混排 OCR 系统中加入了汉字字体识别模块。

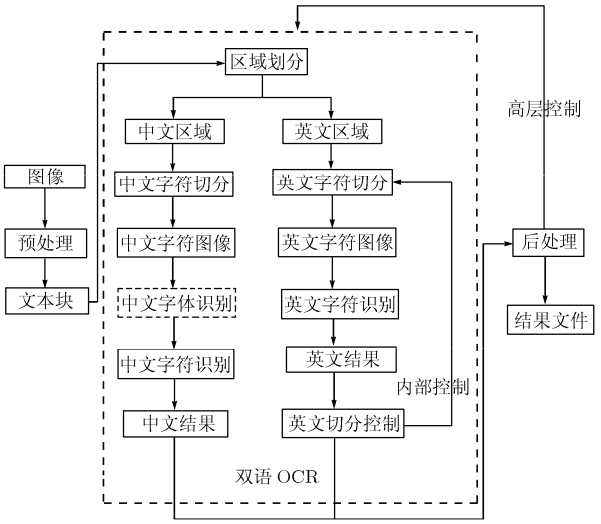


图 5　基于汉字字体识别的汉英混排 OCR 系统

由于汉字保持“单摆浮隔”(即每个汉字所占宽度相等、相邻字符保持间隔)式排版，中文文本行具有明显的全局特性——字符中心的等间距性(图 6)。大量实验表明：基于这一特性，可以以很高的正确率切分出汉字图像[7]。

这一系统流程，保证了仅将汉字图像送入汉字字体识别模块，有效克服了由于受其它语种文字干扰而造成汉字字体

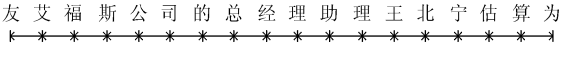


图 6　汉字字符中心的等间距性

识别模块性能下降这一问题。在当前已有的汉字字体识别研究工作中，只有本文考虑了多语混排情况，因而设计出来的方法也具有更好的通用性。

## 5　仿真实验

本文主要针对宋、仿宋、黑、楷、魏、隶书、幼圆 7 种字体进行实验，字符集为国标一级汉字集，共 3755 个汉字，对每种字体每个汉字分别采 36 个不同的训练样本，为了保证结果能够更接近真实分布，在采样过程中：(1)采用不同的分辨率；(2)采用不同的亮度；(3)采用不同的倾斜度。

根据采集条件的不同，最终训练集包含 36 组样本。从中选出 18 组样本按照 3.1 节至 3.4 节中描述的方法抽取字符识别特征点和字体识别特征点，并构建模板库；然后，对于每一个汉字 $h$，用相应的属于字体 $w_j$ 的字体识别模板分别对另外 18 组样本进行评测，统计出与模板 $\theta_{hi}$ 匹配的概率 $p(\theta_{hi} \mid w_j)$，$j=(1,2,\cdots,7)$。

本实验设定期望正确率须在 99%以上，用于字体识别的最大汉字数目为 20。为了验证方法的有效性，根据以下标准进行采样：

(1)测试集为激光打印样张和书籍杂志的扫描图像；

(2)扫描分辨率包括 300DPI 和 400DPI；

(3)包括宋、仿宋、黑、楷、魏、隶书、幼圆 7 组样本，每组样本包含 100 个段落，每个段落中混杂着不同比例的英文；

(4)测试样本质量与文献[1–5]中给出的示例样本质量相似(含少量噪音)。

本实验的字体识别结果如表 1 所示，期望正确率为根据式(2)和式(3)估计出来的正确分类概率，实际正确率表示实际得到的结果，平均长度表示参与字体识别的汉字数目的均值。可见，本方法能够达到很高的正确率，并且对用于字体识别的汉字数目的要求也较低，在实际应用中完全可以满足。

**表 1　本文方法的字体识别正确率(%)**

| 字体 | 宋 | 仿宋 | 黑 | 楷 | 魏 | 隶书 | 幼圆 |
|---|---|---|---|---|---|---|---|
| 期望正确率 | 99.53 | 99.54 | 99.57 | 99.39 | 99.58 | 99.30 | 99.54 |
| 实际正确率 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 平均长度 | 4.2 | 2.4 | 5.7 | 1.8 | 5.7 | 1.6 | 4.9 |
| 速度(秒/段) | 0.30 | 0.18 | 0.43 | 0.10 | 0.42 | 0.09 | 0.36 |

文献[6]中并未给出多字符汉字字体识别的准确率，文献[1]中得到的结果不稳定，因此，本实验只与文献[2–5]中的结果进行了比较，如表 2 所示。可见，本文所提出的方法大大优于文献[2–5]中的方法。

**表 2　本文方法与以往方法的实验结果比较(字体识别的准确率%)**

| 字体 | 宋 | 仿宋 | 黑 | 楷 | 魏 | 隶书 | 幼圆 | 平均 |
|---|---|---|---|---|---|---|---|---|
| 本文 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| 文献[2] | 100 | 92.8 | 100 | 98.4 | – | 100 | 99.6 | 98.47 |
| 文献[3] | 97.4 | – | 99.6 | 96 | 94.4 | 99.2 | – | 97.32 |
| 文献[4] | 97.2 | 93.3 | 96.5 | 95.4 | – | 98.6 | 97.4 | 96.4 |
| 文献[5] | 97.9 | 94.5 | 97.4 | 96.3 | – | 99.1 | 97.9 | 97.2 |

## 6　结束语

本文提出了整体分析法和个体分析法的概念，并在分析它们各自适用范围的基础上，指出个体分析法更适合于解决印刷体汉字字体识别这一问题。

在分析以往研究工作不足的基础上，本文提出一种基于特征点的个体分析法来解决汉字字体识别问题，该方法主要具有以下 3 个优点：

(1)稳定性高：可以通过调整用于字体判别的汉字数目，控制汉字字体识别的期望可信度；

(2)简单高效：直接在空间域中通过特征点比对方法就能够以较高的可信度判别出输入汉字的字体，避免了由空间域到频率域变换所引起的高计算代价；

(3)通用性好：在当前已有的汉字字体识别研究工作中，只有本文考虑了多语混排情况，因而设计出来的方法也具有更好的通用性，能够有效克服文档中其它语种文字对汉字字体识别的干扰。

实验结果表明，本文提出的基于特征点的个体分析法有效解决了印刷体汉字字体识别这一问题，其性能大大优于以往方法。

## 参 考 文 献

[1] 曾理, 唐远炎, 陈廷槐. 基于多尺度小波纹理分析的文字种类自动识别. 计算机学报, 2000, 23(7): 699–704.
Zeng Li, Tang Yuan-yan and Chen Ting-huai. Multi-scale wavelet texture-based script identification method. *Chinese Journal of Computers*, 2000, 23(7): 699–704.

[2] Zhu Yong, Tan Tieniu, and Wang Yunhong. Font recognition based on global texture analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, 23(10): 1192–1200.

[3] 王洪, 汪同庆, 刘建胜, 朱永权, 黄甫征声. 基于小波包纹理分析的字体识别方法. 光电工程, 2002, 29(S1): 62–65.
Wang Hong, Wang Tong-qing, Liu Jian-sheng, Zhu Yong-quan,and Huangfu Zheng-sheng. Method for character font recognition based on wavelet package texture analysis. *Opto-electronic Engineering,* 2002, 29(S1): 62–65.

[4] 杨志华, 齐东旭, 杨力华, 吴力军. 基于经验模式分解的汉字字体识别方法. 软件学报, 2005, 16(8): 1438–1444.

Yang Zhi-hua, Qi Dong-xu, Yang Li-hua, and Wu Li-jun. A Chinese font recognition method based on empirical mode decomposition. *Journal of Software*, 2005, 16(8): 1438–1444.

[5] Yang Zhihua, Yang Li-hua, and Suen Ching Y. A new method of recognizing Chinese fonts. Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition, Seoul, Korea, 29 August – 1 September 2005: 962–966.

[6] 陈力, 丁晓青. 基于小波特征的单字符汉字字体识别. 电子学报, 2004, 32(2): 177–180.
Chen Li and Ding Xiao-qing. Font recognition of single Chinese character based on wavelet feature. *Acta Electronica Sinica*, 2004, 32(2): 177–180.

[7] 王恺, 王庆人. 中英文混合文章识别研究. 软件学报, 2005, 16(5): 786–798.
Wang Kai and Wang Qing-ren. Research on Chinese/English mixed document recognition. *Journal of Software*, 2005, 16(5): 786–798.

[8] 张炘中, 闫昌德, 刘秀英. 汉字识别的特征点法及其一种应用. 中文信息学报, 1987, 11(3): 13–19.
Zhang Xin-zhong, Yan Chang-de and Liu Xiu-ying. A method of chinese recognition based on characteristic dot Matching. *Journal of Chinese Information Processing*, 1987, 11(3): 13–19.

[9] 刘维一, 盛益强, 乔辉, 方志良. 特征点匹配法实现汽车牌照的快速识别. 光电子· 激光, 2002, 13(3): 274–276.
Liu Wei-yi, Sheng Yi-qiang, Qiao Hui, and Fang Zhi-liang. Characteristic dot matching method of realizing rapidly recognition of the number plate. *Journal of Optoelectronics. Laser*, 2002, 13(3): 274–276.

[10] 邢向华, 顾国华. 基于模板匹配和特征点匹配相结合的快速车牌识别方法. 光电子技术, 2003, 23(4): 268–270.
Xing Xiang-hua and Gu Guo-hua. Method of quickly recognizing vehicle plate based on pattern matching and characteristic dot matching. *Optoelectronic Technology*, 2003, 23(4): 268–270.

[11] 靳简明. 数学公式图像处理研究. [博士论文], 南开大学, 2003.
Jin Jian-ming. Research on typeset mathematical expression image processing. [Ph.D.dissertation], Nankai University, 2003.

王 恺： 男，1979 年生，博士，讲师，主要研究方向为文档图像处理、人工神经网络.

靳简明： 男，1977 年生，博士，副研究员，主要研究方向为自然语言处理、文档图像处理.

史广顺： 男，1978 年生，博士，副教授，主要研究方向为文档图像处理、掌纹识别、机器翻译.

王庆人： 男，1944 年生，教授，博士生导师，主要研究方向为文档图像处理、文字识别、机器人学、计算机博弈、软件开发技术.