中 文 信 息 学 报
第 14 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 14 No. 2

# 一种多字体特大字符集字符识别系统

高 涛 李明敬 李志峰
中国科学院自动化研究所文字识别实验室 北京 100080

**摘要** 多字体特大字符集字符识别是当前 OCR 技术研究的热点之一。本文利用一组在抗干扰和描述字符拓扑结构方面具有互补性的特征,基于 Support Vector 技术和可增长自组织神经网络模型,建立一种识别系统来处理该问题。其中包括一个利用 Supprt Vector 技术建立的 Optimal Margin 语言分类器,一个以可增长自组织神经网的粗分类器,结合统计和结构两种识别方法的三级汉字分类器,最后给出良好的实验结果,从而得到该识别系统为解决上述问题的有效方法之一的结论。

**关键词** 汉字识别 OCR 技术 自组织神经网络 Support Vector 技术

## A Multi-Font Huge-Set Character Recognition System

Gao Tao Li Mingjing Li Zhifeng
Institute of Automation, CAS Beijing 100080
Email: gt@hw.ja.ac.cn

**Abstract** Recognition of multi-font characters becomes favorable research area in OCR by now. In this paper, based on Support Vector techniques and SOFM neural network, by use of a grorp of features that are complementary in their description of geometrical and topological structure of character, we have proposed a recognition system. It include an Optimal Margin lingual classifier based on Support Vector techniques, and a three-step Chinese character classifier based on growing self-organizing neural network. This system adopts both of statistical template matching and structure analysis methods. At the end, following the experimental data, a conclusion is made that this system is feasible and effective.

**Keywords** Chinese character reongnition OCR techniques Growing self-organizing neural network Support Vector techniques

## 一、引言

OCR 技术自 60 年代迄今迅速发展,单字识别技术趋于成熟,市场上已出现大量的 OCR 商业产品,然而实际应用又提出更多需求。在不同的印刷、扫描质量下,多语种、多字体特大字符集的混识问题正是其中之一。已有的技术还不能很好解决这一问题,因而该问题成为当前 OCR 技术研究的热点之一。

研究和实验表明,特征的互补性对提高识别率有很重要作用。随着统计模板匹配和结构分析两种方法的趋于结合,统计方法所用的特征也含有字符的结构信息[1]。另外,不同特征对字符的位置、旋转、粗细、扭曲等变化也有不同的适应性[2]。如果在识别的不同阶段采用不同的特征,这些特征具有字符信息描述和抗干扰性能上的互补性,将会使识别效果得到显著提高。一般而言,粗分可选用抗变形和抗噪声能力效强的特征,细分选用区分能力强的特征。这里,我们使用字符的轮廓特征作为粗分特征,使用局部结构特征作为细分特征。

对于多语种、多字体特大字符集的混识问题,类别数目的巨大使得最优分类器的设计很困难。语言分类器为一两类器,其作用是将汉字与符号分开,以便单独设计汉字和符号分类器,使问题相对简化。本文使用 BP 网络作为符号分类器[3];对汉字,本文给出三级分类器。一级粗分使用可增长自组织神经网模型,二级分类使用结构匹配法,三级分类使用 LVQ4 分类器。这样,该系统结合统计和结构两种方法的优越之处,使得分类器有优良性能。

## 二、识别系统

图 1 所示为系统的框图,图中输入模式经过语言分类器,如果是英文符号,则进入 BP 神经网络符号分类器,得到输出结果;如果是汉字则进入 SOFM 神经网络的一级粗分类器和结构匹配方法建立的二级分类器,这时如果输出可信度大于一定的域值,则输出识别结果,否则,进入 LVQ4 分类器,进行三级分类。以下各节将介绍系统的主要组成部分。

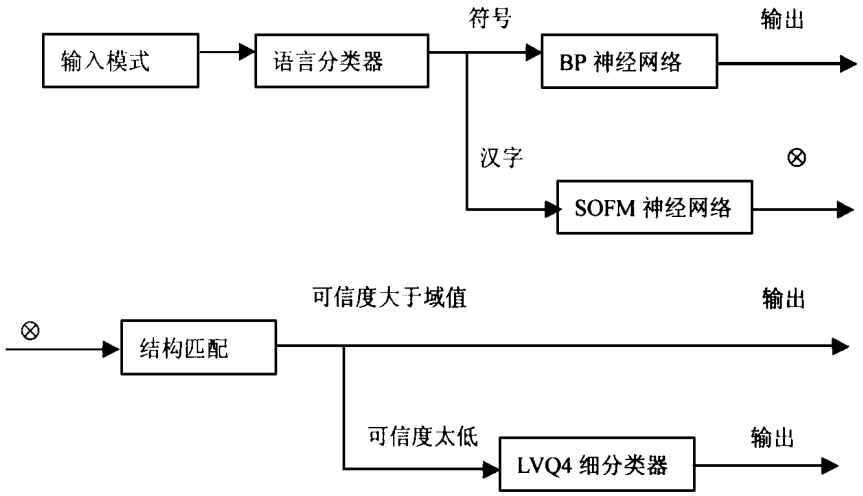


图 1　系统框图

### 2.1　预处理

二值输入图象的预处理包括平滑、归一化和去噪声等。这里用 Blurring 技术代替平滑[4]。由于汉字的形体结构复杂,很多字符的笔画构不成连通关系,所以在去除孤立噪声时,对滤波窗口大小的选择要适中,这里取图 2 所示的一组窗口[5]。

在窗口中,1 指当前像素,沿短线相连求各像素灰度值之和,满足 sum < = 1 条件时当前像点为噪声点。在使用时,由于有滤除宽度为 1 或 2 的笔画线的可能,所以还要加一些黑像素点游程判断。

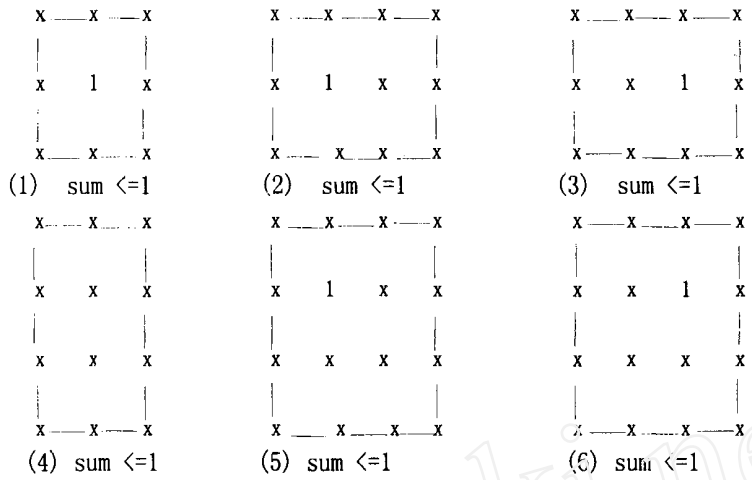


图 2 滤波窗口

### 2.2 特征提取

系统的识别率很大程度上将最终取决于特征的区分能力。A. K. Jain 等人曾对字符特征提取的方法进行归类[2],同时也对其稳定性行了分析。研究者早期曾尝试过提取在形变下能保持不变的特性,如几何位移不变量[6,7]。然而由于计算复杂,很少在实用系统中使用。

Bokser 首先在商业 OCR 产品中使用 Zoning 技术[8]。字符图象这时被分成 $m \times n$ 子区域,提取字符图象在这些区域中的特征作为该局部区域的特征,随后,研究者又提出了具有模糊边界的 Zoning 技术[9]。我们这里根据字符的质心,将字符图象划分为大小不同的区域,每一区域的边界重叠一个像素。

系统采用汉字的轮廓特征作一二级粗分[10]。沿四边将图象划分为 4 ×16 个子区域,在每个子区域,从边界开始沿划分方向扫描,遇到字符外轮廓或 1/2 字符宽时结束。累计各区域中的扫描游程及游程差分,得到 64 维一级分类和 64 维二级分类特征向量。三级分类特征由 Mesh 特征的 Kirsch 方向特征加权组合而成。将图象划分为 16 ×16 个子区域,累计各区域中的黑像素点数得到 256 维的 Mesh 特征,将图象划分为 8 ×8 的子区域,累计区域中各像素点的四个方向梯度值,归一化后得到 256 维方向数特征[11]。

汉字的变化主要包括位置、笔画厚度、旋转、笔画畸变等方面。外轮廓相对内轮廓较清晰,其特征对位置和笔划厚度变化有较好的适应能力。如果假设版面的倾斜校正在版面分析阶段有较好的处理,使用轮廓特征作粗分会得到较好的效果。Mesh 特征虽然区分能力不是最强,但它抗噪能力较强,特别是对印刷的浓淡有较好的适应性,所以用它作为细分特征的一部分。

### 2.3 基于 Support Vector 技术的语言分类器

汉字和英文、数字以及标点符号在几何和拓扑结构上存在较大的差异,研究者因而在识别前端加一个语言分类器,用以区分汉字和符号,以便将输入模式送入单独设计的符号分类器或汉字分类器。虽然系统将增加一个单元,然而,两类问题在模式识别技术中相对简单的多,而且分类器的设计技术已很成熟,所以可将原问题相对简化。

对线性可分的两类问题,M. A. Aizerman 和 E. M. Breaverman 等人已给出 Optimal Margin 分类器的设计公式。并且证明在泛函空间 $\Phi$ 中,分界面为一超平面 $D(x) = w \cdot \Phi(x) + b$,其中 $w$ 可被表述为训练集中的一部分样本向量的线性组合,这部分样本向量被称为 Support

Vectof[12]。Optimal Margin 分类器可确保得到两类之间的最大 Margin。后来的研究者又对分类器的设计作了大量工作,将之应用到线性不可分的情况。Bernhard E. Boser 和 Isabelle M. Guyon 等人最近以给出一个较适用的设计算法[13]。

假设训练集为分属于 $A$ 和 $B$ 两类的 $p$ 个样本:$(x_1, y_1)$, $(x_2, y_2)$, $(x_3, y_3)$ ……$(x_p, y_p)$,则

$$y_k = -1, \quad \text{如果 } x_k \text{ 属于 } B$$
$$y_k = 1, \quad \text{如果 } x_k \text{ 属于 } A \tag{1}$$

决策函数应有果

$$D(x) > 0, \quad \text{如果 } x_k \text{ 属于 } A$$
$$D(x) <= 0 \quad \text{如果 } x_k \text{ 属于 } B \tag{2}$$

如将 $D(x)$ 表述为 $\sum a_{kl} K(X_k, x) + b$ 形式,$K$ 为希尔伯特空间核,则最后问题可转化为式(3)所示的标准二次规划形式。

$$\max J(a, b) = \sum_k a_k (l - b_k) - A \cdot H \cdot A/2$$
$$a_k >= 0 \qquad k = 0, 1, 2, \cdots\cdots, p \tag{3}$$

为 $-P \times P$ 维矩阵,$H_{KL} = y_k y_l K(x_k, x_l)$。实际设计时,由于汉字较多,可以先用一些简单的特征将笔画较多、结构复杂的汉字从训练集中去除,用来降低 $H$ 矩阵的维数。

### 2.4 基元可增长的自组织神经网模型的粗分

自组织神经网络模型自 Kohonen 提出后[14,15],因其能较好的给出模式在高维向量空间中的概率分布估计,受到广泛的研究和应用。然而对于特大字符集字符识别,由于类别太多,训练样本数目太大,自组织网络建立的全局优化值仍然很难得到保证,而且这种最优值仅是量化意义下的最优,与分类误差不尽相同。Frotzke 在此基础上修改了 Kohonen 模型中神经元的相邻关系[16],并增加了神经元的分裂和删除操作,从而使得神经网规模可以增长。这里有别于原模型,我们采用式(4)所示的度量控制神经元的分裂[17]。

$$CDM = \frac{1}{N} \sum_{i=1}^{n} \{ k(i) - \max_j [k(j/i)] \} \tag{4}$$

其中,$K(i)$ 是第 $i$ 个神经元在竞争学习中捕获的样本总数,$N$ 参与训练的样本总数,$K(j/i)$ 是第 $j$ 类在第 $i$ 个神经元中的样本数目。

### 2.5 基于结构匹配方法的二级粗分

汉字识别就方法而言可分为统计模板匹配和结构分析两种方法。随着字符识别技术的发展,二者渐趋于结合。基于字符图象所提出的高维特征向量可以表述字符的几何及拓扑结构,特征向量中的每一元素可看作为字符结构基元,基元之间的关系可以通过其在向量中的位置关系表示出。这样,在特征向量的基础上可以完成结构匹配,并可使用 DP 规划或弹性匹配的方法[18]。

使用 Zoning 技术后,由局部图象特征所组成的特征向量也可看成是一个结构图,只不过它反映的是进行了非线性变换后的图象结构[19]。匹配过程可看作是在特征向量空间中计算相似度,仅形式上区别于通用的相似度量[20]。基于此,我们定义字符图象整体匹配度为:

$$S(X, X') = \sum_{i=1}^{n} f_i(\sigma_i, d(x_i - x_i'))$$

其中 $d$ 为距离度量函数,可取欧氏距离或绝对值距离。$i$ 为 $d$ 度量下特征向量第 $i$ 个分量的

方差，$f$ 为第 $i$ 个分量上的匹配分布函数，满足 $d=0$ 时，$s=1$，可采用线性函数或高斯函数形式。

2.6 LVQ4 细分类器

LVQ4 实际上是一结合了模拟退火算法的有监督训练算法[21]，通过训练窗口的选择微调模板的位置。在细分类中，主要问题之一是确定每个字符对应的模版。模版数目太少，不足以覆盖该字符的变化；模版数目太多，字符类别之间的分界重叠区域面积太大，识别错误率反而也会上升。以下算法在 LVQ 算法的基础上，动态调整各类别所建的模版数目。

假设当前训练集为 $S=\{X_1 \mid X \in R^n\}$，$c_1, c_2, \cdots, c_i, \cdots c_p$ 为样本所属的字符类别，则 $S$ 被划分为 $\{S_1, S_2, \cdots\cdots, S_p\}$。

下面为细分模版的生成算法。

(1) 选定初始模版向量 $w_1, w_2, \cdots, w_i, \cdots, w_p$ 这里取字符类别中样本的均值作为初始模版。

$$w_i = \frac{1}{N_i}\sum_{i=1}^{N_i} X_i \qquad X_i \in c_i \tag{6}$$

(2) 使用 LVQ4 算法对模版位置进行一定过程的微调。

(3) 按现在模版的位置将训练样本划分，假设为 $\{S'_1, S'_2, \cdots\cdots, S'_2\}$，计算样本的总体误识率，如果低于预设阀值，则停止；否则，进行(4)到(5)的增加模版操作。

(4) 计算各字符类别在学习中的分类错误率，选择分类错误最大的类别 S′。作为待分裂的子集。

(5) 以最大最小聚类方法对 S 中的样本进行重新划分，生成新的模版。

(6) 至(1)重复训练。

## 三、实验结果与分析

这里我们给出在此识别系统上的两组实验结果。实验一识别范围包括 5401 个一级繁体汉字，120 个符号，共 5521 个字符，汉字的字体有宋体、仿宋、楷体、黑体、圆体、隶书，对同一字体又包括特、粗、标、细四种变化。使用在 4 种不同扫描条件下的 3 号字作为训练集，使用在相应扫描条件下的 5 号字作为测试集。这样，细体在扫描浓度很淡的情况近似为实际工作中‘断’的现象，而粗和特则近似为‘糊’的现象。英文符号经语言分类器与汉字分开。表 1 所示为汉字识别结果，总体识别率为 99.33 %。宋体、仿宋、黑体和圆体四种最高。楷体稍低，隶书最低。对文本测试结果也表明，系统有较好的适应性。

表 1 繁体识别实验数据

| 字体 | 宋体 | 仿体 | 黑体 | 楷体 | 圆体 | 隶书 | 各体综合 |
|---|---|---|---|---|---|---|---|
| 识别率 | 99.39 % | 99.43 % | 99.40 % | 99.12 % | 99.41 % | 98.90 % | 99.33 % |

实验二的识别范围包括 4394 个繁体汉字，5164 个简体汉字和 92 个符号，共 6992 个字符。汉字字体包括宋、仿、楷、黑、圆、隶变、魏碑、行楷。每一字体又包括特、粗、标、细四种变化。同样用 3 号字作训练，用 5 号字作测试。表 2 所示为识别结果。总体识别率为 99.15 %，宋、仿、黑、圆、四种最高，和实验一比较基本没有变化。楷体、隶书也同样，新加入的魏碑为 98.7 %，行楷为 98.5 %。应用于汉王公司将要推出的名片识别系统中，有较好的实际识别效果。

表 2　简繁体混识实验数据

| 字体 | 宋体 | 仿体 | 黑体 | 楷体 | 圆体 |
|---|---|---|---|---|---|
| 识别率 | 99.38 % | 99.43 % | 99.40 % | 99.11 % | 99.40 |
| 字体 | 隶书 | 魏碑 | 行书 | 各体综合 | |
| 识别率 | 98.87 % | 98.71 % | 98.55 % | 99.15 % | |

## 四、结论

本文在以 Support Vector 技术、可增长自组织神经网络和 LVQ 算法的基础上,利用具有互补性几种特征,建立一种识别系统,用以解决实际应中对简、繁汉字及英文字符的混识问题。该系统占用内存空间在 4.5M 左右,在 PII233 机器上识别速度可达到 60～70 字/秒。该识别核心已经应用在汉王"名片自动识别和管理"产品中,经试验和实际使用检验,识别性能已达到较高实用水平,而且对印刷和扫描质量有较好的适应性。所以可以说,该方法是解决多体大字符集汉字识别的可行方法之一。

此外,对实际问题中误识字符的分析表明,因字形相近引起的错误,除了可使用后处理模型将相似字区分外,还可以采用多种特征的集成作为分类特征。噪声的存在使得字符特征产生扰动甚至变形,也因而成为引起识别错误的主要原因之一。所以,还要在去噪、归一化等预处理技术方面作进一步的工作。

## 参 考 文 献


[1] Shunji Mori, Ching Y Suen, Kazuhiko Yamamoto. Historical ReviewofOCR Research And Development. Proceedings of the LEEE, July 1992, 80(7): 1029 - 1057

[2] Oivind Due Trier, Jain A K, Torfinn Tax. Feature Extraction Methods For Character Recognition—A survey. Pattrn Tecognition, 1996, 29(4): 662

[3] Tumelhart D E, McClelland J L. Parallel Distributed Processing. MIT Press, 1998, 1

[4] Ken-ichi Maeda, Yoshiaki Kurosawa, Haruo. Hand-Printed KANJI Recognition by Pattern Matching Method. CH1801 - 0/82/0000/0789, 1982, IEEE

[5] Ray S. A Heuristic Noise Reduction Algorithm Applied to Hand-Written Numeric Characters. 067 - 8655/881988, Elsevier Science Publishers B. V. (North-Holland)

[6] Jain A K, Chitra Doral. Practicing Vision: Integration, Evaiuation and Application. Pattern Recognition, 1997, 30(2): 183 - 196

[7] Hu M K. Visual Pattern Recognition by Movement Invariant. IRE Trans. Inform. Theory Vol. IT - 8. Feb. 1962, 179 - 187

[8] Bokser M. Omnidocument Technologies. Proc. IEEE 80, July 1992, 1066 - 1078

[9] Cao J, Ahmadi M, Shridhar M. Handwritten Numeral Recognition with Multipie Features and Multistage Classifiers. IEEEInt. Symp. CircuitsSyst. 6, London, 30May - 2 June 1994, 323 - 326

[10] Hirai S, Sakai K. Development of a High Performance Chinese Characters. Proc. 2. ICPR, 1980, 867 - 871

[11] Seong-Whan Lee, Jeong-Seon Pak. Nonlinear Shape Normalization Nethods for the Recognition of Large-Set Handwritten Character. Pattern Recognition, 1994, 27(7): 895 - 902


(下转 54 页)

Segments. Pattern Recognition ,1998 ,31(10)
[7] Kasturi R ,Bow S ,El-Masri W *et al*. A System for Interpretation of Line Drawings. IEEE Trans on PAMI ,1990 ,12(10)
[8] Ramachandran K. A coding method for vector representation of engineering drawings. Proc IEEE ,1980 ,68 :813 - 817
[9] **郭承恩. 表格理解及其实用化[硕士学位论文]. 北京 :中国科学院自动化研究所** ,1998
[10] Casey R ,Ferguson D ,Mohiuddin K *et al*. Intelligent Forms Processing System. Machine Vision and Applications ,1992 ,5 :143 - 155
[11] Tang Y Y ,Lee S ,Suen C Y. Automatic Document Processing :A Survey ,Pattern Recognition ,1996 ,29 (12)
[12] Fan K C ,Lu J M ,Chen G D. A Feature Point Clustering Approach to the Recognition of Form Documents. Pattern Recognition ,1998 ,31(9)

(**上接** 36 **页**)

[12] Aizerman M A ,Braverman E M ,Rozonoer L I. Theoretical Foundation of the Potential Function Method in Pattern Recognition Learning. Automation and Remotecontrol ,1964 ,25 :821 - 837
[13] Boser B E ,Guyon I M ,Vapnik V N. A Training Algorithm for Optimal Margin Classifiers.
[14] Kohonen T. The Self-Organizing Map. Proceedings of the IEEE ,Sep 1990 ,78(9)
[15] Kohonen T. Self-Organization and Associative Memory. New York ,Springer-Verlag
[16] Fritzke B. Growing Cell Structures-A Self-Organizing Network for Unsupervised and Supervised Learning. Neural Ntwork ,1993 ,8(6) :1441 - 1460
[17] Kohn F ,Nakano G M. A Class Discriminability Measure Based On Feature Space Partitioning. Pattern Recognition ,1996 ,29(5) :873 - 887
[18] Yamada H. Contour DP Matching Method and its Applications to Hand-Printed Chinese Characters Recognition. Proc. 7 ,IJCPR ,1984 ,389 - 392
[19] Mori S. Foundation of Chinese Characters and Figures Recognition Techniques. Tokyo ,Ohm ,1993
[20] Fu K S. Application of Pattern Recognition. CRC Press ,Inc. ,Boca Raton ,Florida ,1982
[21] Lee S W ,Song H H. Optimal Design of Reference Models for Large-Set Handwritten Character Recognition. Pattern Rattern Recognition ,1994 ,27(9) :1267 - 1274