

文章编号:1001-9081(2005)10-2418-04

基于神经网络和模糊匹配算法的手写汉字预分类研究

卢 达¹, 浦 炜¹, 陈琦伟¹, 谢铭培²

(1. 常熟理工学院 物理与电子科学系, 江苏 常熟 215500; 2. 复旦大学 计算机科学系, 上海 200433)

(ld47@cslg.edu.cn)

摘要: 对手写汉字识别问题, 提出了一种在识别之前对手写汉字预分类的新方法, 该方法用 Neocognitron 网提取字符笔画特征, 然后采用有监督的扩展 ART 神经网络(SEART)产生一定数量的预分类组并通过基于模糊相似测量的匹配算法进行预分类。实验表明, 该方法用于手写汉字分类效果良好, 预分类正确率达到 98.22%。

关键词: 手写汉字预分类; 人工神经网络; 有监督的扩展 ART; 模糊匹配算法

中图分类号: TP391 **文献标识码:**A

Study on preclassification for handwritten Chinese character based on neural net and fuzzy matching algorithm

LU Da¹, PU Wei¹, CHEN Qi-wei¹, XIE Ming-pei²

(1. Department of Physics and Electric Science, Changshu Institute of Technology, Changshu Jiangsu 215500, China;

2. Department of Computer Science, Fudan University, Shanghai 200433, China)

Abstract: To settle the recognition task of handwritten Chinese characters, the authors put forward a method for handwritten Chinese character preclassification before character recognition. In this method, Neocognitron was used in extracting stroke features, then uses the Supervised Extended ART (SEART) to create some preclassification groups, and uses matching algorithm of fuzzy prototypes of similarity measurement for character preclassification. The experiment shows this method is effective when used for handwritten Chinese character classification and characters of the testing set can be distributed into correct preclassification classes at a rate of 98.22%.

Key words: handwritten Chinese character preclassification; artificial neural network; supervised extended ART; fuzzy matching algorithm

0 引言

汉字识别是字符智能识别的重要组成部分, 手写汉字识别对于我国中文信息化发展具有重要的意义。理论和实践告诉我们, 特征选取和分类器设计是获取一个有效的字符识别系统的关键, 手写汉字识别的研究也基本上围绕这两方面进行努力。

汉字本身具有类别规模大、结构复杂、相似模式多的特点, 手写汉字更是由于书写习惯因人而异, 使字形变化极为严重。从国内外手写汉字识别的研究状况看, 面对手写汉字字形严重变化造成的特征的发散分布, 用高斯分布来描述及使用最小马氏距离二阶分类器来识别已显得明显不足。选取稳定、典型的手写字符类别密切相关的特征、利用统计模式识别方法, 根据对大量训练样本学习所获得的有关识别样本概率分布的知识, 设计与此概率分布相匹配的分类器, 以及多种分类器的集成已成为当前手写汉字识别研究所采取的基本方法, 相关的实验结果也说明这些方法的有效性^[1~3]。然而可以识别超多类手写汉字、且具有高抗干扰和高鲁棒识别性能

分类器的获得, 目前还有一定困难, 因此, 手写汉字识别研究仍存在如何提高识别的准确性、可靠性和识别效率的问题。但是, 在手写汉字比较识别之前的字符预处理过程中, 如何将神经网络的大规模分类能力来“分隔”庞大的手写汉字库, 再结合模板匹配能有效提取特定形状的结构能力来完成手写汉字样板的预分类处理, 减少字符类别, “缓解”待识别手写汉字数量巨大问题, 缩小后阶段分类、识别范围, 也是提高手写汉字识别水平值得尝试、研究的一个工作。

本文在作者前期完成基于模糊相似测量的字符无监督分类^[4]和将模糊模型相似测量用于小类别数汉字识别研究^[5]的基础上, 提出了一种基于神经网络和模糊匹配算法的手写汉字分类法作为手写汉字预分类处理。该方法首先在采用 Fukushima 提出的自组织、竞争学习的 Neocognitron 多层神经网络^[6]提取具有笔画方向、笔画密度特征的基础上, 通过有监督扩展 ART 神经网络(SEART)作为分类器产生字符预分类组^[7], 这一步的目的用来将待识别汉字库分成若干个“子”字库, 降低后续预分类步骤的计算代价, 并充分利用该网络分类器优越的容错能力以确保该阶段较高的预分类正确率。在

收稿日期:2005-07-16 基金项目:江苏省教育厅自然科学基金资助项目(02KJD54001)

作者简介: 卢达(1947-), 男, 江苏常熟人, 教授, 主要研究方向: 模式识别和图像处理; 浦炜(1973-), 男, 江苏常熟人, 讲师, 主要研究方向: 图像处理和 CAD; 陈琦伟(1978-), 男, 江苏常熟人, 硕士, 主要研究方向: 通信和信息系统、信号处理等; 谢铭培(1937-), 男, 福建泉州人, 教授, 主要研究方向: 人工智能、图像识别、计算机控制等。

此基础上,我们将汉字二值图像采用模糊逻辑方法转换成基于非线性加权相似函数的模糊样板,并通过模糊相似测量的匹配算法完成对前阶段产生的每个预分类组的汉字样板的预分类。这是一种直接利用字符图像全局信息进行汉字统计分类识别的方法,理论上讲可基本解决分类识别抗干扰的鲁棒性识别问题。该方法能有助于减小噪声的影响,限制样板库容量的迅速增加,缩小后阶段字符的分类识别范围。经实验,取得了良好的分类效果。

1 字符预分类特征的提取

1.1 预处理

汉字由笔画构成,可以利用笔画的特性进行手写汉字的预分类。手写汉字的笔画差别由两部分组成:1)笔画属性差别,即不同汉字中同类型笔画的方向、宽度、方向变化等方面的差别;2)笔画分布差别,即同一汉字在不同人书写时,对应笔画在字符图像中的长短、位置等方面的差别。为尽可能选择对分类最有效的特征和在满足识别需要的足够的互信息熵的情况下进行特征维数压缩的同时能有效表示笔画属性和笔画分布差别,我们使用笔画方向和笔画密度特征。

系统的初始输入为手写汉字二值图像,在特征提取之前,需对二值图像平滤、去噪、细化,并进行归一化,获取 72×72 点阵的固定大小的图像。

由于汉字周边区域含有丰富的笔画特征信息,我们采用 Lee 和 Sheu 的分割法^[8]将归一化的汉字图像分成 9 个相同大小的方块,如图 1 所示。我们着重 9 个方块中的四周部分,并从图 1 中各阴影区域提取笔画特征。图 1 中的中央阴影区域与周边阴影区域交叠宽度为 5 个像素,以避免书写位置的变化。

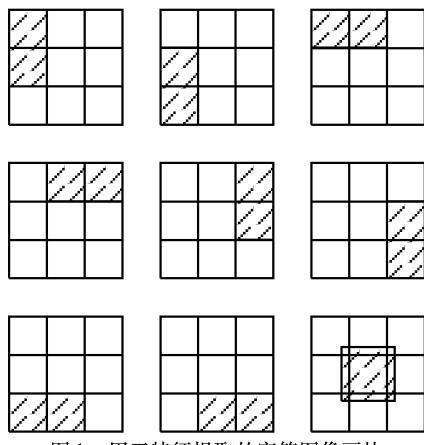


图 1 用于特征提取的字符图像画块

1.2 笔画特征提取

用于预分类的笔画方向、笔画密度特征由 Fukushima 提出的 Neocognitron 多层网络提取。Neocognitron 是一个自组织、竞争学习的多层神经元网络,其最大特点之一是通过连续多层网络对其输入层的图像信息进行分级特征提取(Hierarchical feature extraction),如图 2 所示。其中首先由 US1 层抽取笔画方向特征,以后通过各层分别提取其他特征,很适合手写汉字的笔画特征提取。

Neocognitron 中的 US1 层有 12 个属性,现把它们分为 8

类(S1~S8),如图 3 所示。这 8 类可视为字符预分类的 8 个笔画方向特征。用 3×3 窗口从左到右,从上到下扫描字符图像,计及图 1 各阴影划块中 8 类的数目,即形成 $8 \times 9 = 72$ 维笔画方向特征,然后定义连续像素小于等于 4 为短笔画,连续像素大于 4 为长笔画,并由 Neocognitron 的 US2 层提取 36 维笔画密度特征。这样,总计为 $72 + 36 = 108$ 维笔画特征。

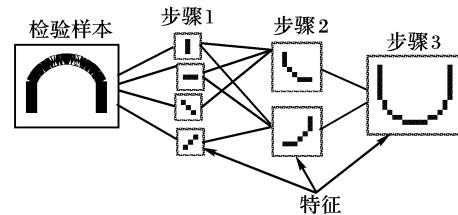


图 2 Neocognitron 网分级特征抽取示意图

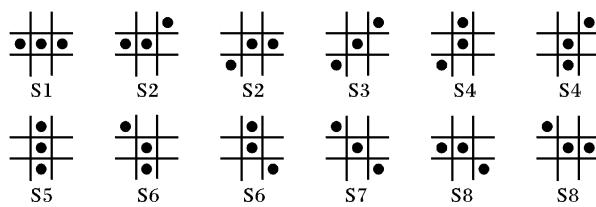


图 3 US1 层中的 12 个属性和相应的 8 个类

2 字符预分类

2.1 SEART 网

由前面提取的 108 维笔画特征送至有监督的扩展 ART(SeART)网,在其输出节点产生各预分类组,每个预分类组即为一个待识别手写汉字的“子”图像库。采用 SeART 网的目的是“分隔”庞大的待识别手写汉字库,减小后续采用模糊相似测量的匹配算法进行字符样板预分类的计算代价,确保较高的预分类正确率。

我们采用的 SeART 网络结构如图 4 所示,下面二层是带有补码的基本模糊 ART,但有不同的连接权值及匹配控制。底层 F1 层为输入层, F2 层为类别表示层,两层节点之间采用双向链接。顶层 F3 层为 Grossberg 的 Outstar 作为输出层。此层的作用是学习对应 F2 层类别节点的目标输出并完成节点的网络输出和目标输出的匹配程度,计算得到类别输出。若两者失配则对输入样本向量来说必须选择另一类别,这由匹配控制逻辑单元来实现匹配程度控制功能。

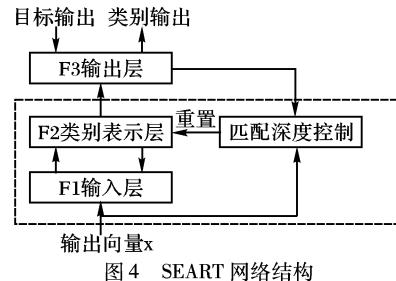


图 4 SeART 网络结构

我们的 SeART 网的输入节点数等于提取的笔画特征数,即为 108,而 SeART 网的输出节点数等于相应预分类集合中的字符类别数。

设 F1、F2 层的节点数分别为 N 、 M (SeART 的输入节点数等于输入字符图像的抽取的特征数,即 $N = 108$,SeART 的输出节点数等于相应预分类字符类别数), b_{ij} 为 F1 层到 F2

层的连接权值, ω_{ij} 为 F2 层到 F1 层的连接权值, t_i 为 F2 层到 F3 层的连接权值, $i = 1, \dots, N; j = 1, \dots, M$ 。 λ 为匹配算子, 设输入向量为 X , 类别输出向量为 O , 目标输出向量为 S , 则 $\lambda = \frac{(O \oplus S)/M}{(X \oplus S)/N}$ 。SEART 网学习算法如下:

1) 连接权值初值, 令: $b_{1j}^{(0)} = \dots = b_{Nj}^{(0)} = \theta_j, j = 1, \dots, M$, 其中 θ_j 的排序为 $\theta_N < \dots < \theta_1 < \frac{1}{\alpha + N}$, 其中 $0 < \alpha \ll 1$;

$$\omega_{jl}^{(0)} = \dots = \omega_{jN}^{(0)} = 1, j = 1, \dots, M; t_1 = \dots = t_M = 1$$

2) 输入一二值字符样本 $X = (x_1, \dots, x_N), x_i \in \{0, 1\}$ 。

3) 求 F2 层各节点的加权组合值 $y_j = \sum_{i=1}^N b_{ij} x_i, j = 1, \dots, M$ 。

4) 根据“胜者为王”学习规则选取获胜节点 $J = \arg(\max_{j=1, \dots, M} y_j)$ 。

5) 求 F2 层到 F3 层的加权组合值 $O = \sum_{j=1}^M t_j y_j$ 。

6) 匹配度控制。设定 λ 的上下限 λ_2, λ_1 , 若通过计算所得的 $\lambda \in [\lambda_1, \lambda_2]$, 则匹配成功并转(7); 否则匹配不成功, 并置 y_j 为 -1, 转 4) 继续搜索。

7) 修正连权权值:

$$\text{若 } j = J, \text{ 则 } \omega'_{ji} = \omega_{ji} x_i, b'_{ij} = \frac{\omega_{ji}}{\alpha + \sum_{i=1}^N \omega_{ji} x_i};$$

$$\text{若 } j \neq J, \text{ 则 } \omega'_{ji} = \omega_{ji}, b'_{ij} = b_{ij}, i = 1, \dots, N; j = 1, \dots, M.$$

8) 转 2) 再输入新样本。

2.2 模糊匹配算法

在 SEART 网产生一定数量“子”字库, 即预分类组的基础上, 本阶段将各“子”字库中字符二值图像采用模糊逻辑方法转换成基于非线性加权相似函数模糊样板及基于模糊相似测量的匹配算法完成字符预分类。以下主要介绍匹配算法和相关的规则分类, 模糊样板的转换详见参考文献[4]。

根据模糊相似理论, 一二值图像的模糊样板 λ 可用有序对的一矩阵表示, 即 $\lambda = \{(p_{ij}, x_{ij})\}$ 。其中 p_{ij} 代表一像素, x_{ij} 代表 p_{ij} 的图像的隶属值。若 λ 为连体二值图像 $\{A_1, A_2, \dots, A_m\}$ 的集合组成的一模糊样板, 二值图像的权函数为 $\{\omega_{A_1}, \omega_{A_2}, \dots, \omega_{A_m}\}$, 则:

1) λ 中各元素的隶属值为:

$$x_{ij} = \frac{\sum_{a_{ij} \in A_1} a_{ij}}{m} \quad (1)$$

式中: a_{ij} 为集合中的元素。

2) λ 的权函数为:

$$\gamma_{ij} = \frac{\sum_{\omega_{ij} \in \omega_{A_1}} \omega_{ij}}{m} \quad (2)$$

式中: ω_{ij} 为 $\omega_{A_1}, \omega_{A_2}, \dots, \omega_{A_m}$ 中的元素。

3) 二个模糊样板 $\lambda_1 = \{p_{ij}, x_{ij}^{(1)}\}$ 和 $\lambda_2 = \{p_{ij}, x_{ij}^{(2)}\}$ 相似测量的相关系数为:

$$\zeta(\lambda_1, \lambda_2) = \frac{\sum (x_{ij}^{(1)} \wedge x_{ij}^{(2)} - \frac{1}{2} \gamma_{ij}^{(1)} x_{ij}^{(2)} - \frac{1}{2} \gamma_{ij}^{(2)} x_{ij}^{(1)})}{\sqrt{\sum x_{ij}^{(1)2} \sum x_{ij}^{(2)2}}} \quad (3)$$

式中: $\gamma_{ij}^{(1)}, \gamma_{ij}^{(2)}$ 为 λ_1, λ_2 的权函数, \wedge 为交运算符号, $x_{ij}^{(n)2} (n = 1, 2)$ 表示自交运算 $x_{ij}^{(n)} \wedge x_{ij}^{(n)}$ 。

2.2.1 匹配算法

两给定模糊图形 A 和 B 若相似, 它们应该有相似的几何性质, 因此相似测量可通过图形矩心匹配计算。由于噪声的存在, 计算过程中必须考虑容差, 算法描述如下:

1) 计算 A、B 的矩心 C_A, C_B :

$$C_A = \left[\frac{\sum j x_{ij}^{(A)}}{\sum x_{ij}^{(A)}}, \frac{\sum i x_{ij}^{(A)}}{\sum x_{ij}^{(A)}} \right]$$

和

$$C_B = \left[\frac{\sum j x_{ij}^{(B)}}{\sum x_{ij}^{(B)}}, \frac{\sum i x_{ij}^{(B)}}{\sum x_{ij}^{(B)}} \right] \quad (4)$$

2) 根据最小距离 $C_A C_B$ 计算 $\zeta(A, B), \zeta(A, B)$ 由图形 A 作 α, β 移动推导而得:

$$\zeta_{\alpha\beta}(A, B) = \frac{\sum (x_{ij}^{(A)} \wedge x_{i+\alpha, j+\beta}^{(B)} - \frac{1}{2} \gamma_{ij}^{(A)} x_{i+\alpha, j+\beta}^{(B)} - \frac{1}{2} \gamma_{i+\alpha, j+\beta}^{(B)} x_{ij}^{(A)})}{\sqrt{\sum x_{ij}^{(A)2} \sum x_{ij}^{(B)2}}} \quad (5)$$

式中: $\alpha = \text{round}(x_{C_B} - x_{C_A}), \beta = \text{round}(y_{C_B} - y_{C_A})$, round 代表旋转变换。

若相关系数 ζ 高于一阈值, 比如高于 0.95, A 和 B 可认为相同。

3) 若 ζ 在临界范围内, 比如 0.85 ~ 0.95, 则 $\zeta_{\alpha\beta}(A, B)$ 的值由 α, β 在如下范围内求得:

$$|\alpha - (x_{C_B} - x_{C_A})| \leq 1$$

和

$$|\beta - (y_{C_B} - y_{C_A})| \leq 1 \quad (6)$$

若 $x_{C_B} - x_{C_A}$ 和 $y_{C_B} - y_{C_A}$ 的值不是整数, 则 α, β 需作三个以上位置的匹配运算, 其中若有一次 ζ 大于 0.95, 则 A 和 B 为相同类, 否则, 两图形为不同类。第三步的模糊推论是基于临界范围内相似性的不确定, 若噪声引起的图形矩心的失真可通过移动一图形来获得较好的匹配。

2.2.2 规则分类

若待识字符模糊样板放在同一任意序列中, 其匹配是低效率的。我们对预分类第一阶段得到的四个预分类组分别采用分级树(hierarchical-tree)法, 该分类法与顺序分类相比有二个优点: 一是匹配过程的搜索时间比顺序分类短; 二是具有平行处理能力, 即分级树的各节点可同时处理。考虑接近根节点处因库容量迅速增加可能使匹配难于进行, 采用对两字符模糊样板 λ_1, λ_2 由公式(5) 对它们作相似测量的同时, 将相关系数 ζ 分成相等测量 E 和不等测量 I:

$$E = \frac{\sum x_{ij}^{(1)} \wedge x_{ij}^{(2)}}{\sqrt{\sum x_{ij}^{(1)2} x_{ij}^{(2)2}}} = \frac{\sigma_{\lambda_1 \cap \lambda_2}}{\sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}}};$$

$$I = \frac{\gamma_{ij}^{(1)} x_{ij}^{(2)} + \gamma_{ij}^{(2)} x_{ij}^{(1)}}{2 \sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}}} \quad (7)$$

式中: σ_{λ_1} 、 σ_{λ_2} 、 $\sigma_{\lambda_1 \cap \lambda_2}$ 为模糊样板 λ_1 、 λ_2 、 $\lambda_1 \cap \lambda_2$ 的基。

设: ζ_t 为相似测量的阈值, λ_1 为输入模糊样板, λ_2 为文本库中的模糊样板, 由 $\sigma_{\lambda_1} > \sigma_{\lambda_2}$, 若当 $\lambda_1 \supset \lambda_2$, 即 $x_{ij}^{(1)} \geq x_{ij}^{(2)}$ 时为最佳相等测量。若下式成立, 则 λ_2 为可能匹配样板:

$$E(\lambda_1, \lambda_2) = \frac{\sigma_{\lambda_1 \cap \lambda_2}}{\sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}}} = \frac{\sigma_{\lambda_2}}{\sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}}} = \sqrt{\frac{\sigma_{\lambda_2}}{\sigma_{\lambda_1}}} \geq \zeta_t \quad (8)$$

同理, 当 $\sigma_{\lambda_1} < \sigma_{\lambda_2}$, 若以下不等式成立, λ_2 为可能匹配样板:

$$\sqrt{\frac{\sigma_{\lambda_1}}{\sigma_{\lambda_2}}} \geq \zeta_t \quad (9)$$

因此, 规则 1 包括式(8)、(9) 的使用。

规则 1: 若 $\zeta_t^2 \sigma_{\lambda_1} \leq \sigma_{\lambda_2} \leq \sigma_{\lambda_1} / \zeta_t^2$ 成立, λ_2 是 λ_1 的可能匹配样板。

由于相似样板具有相似特征, 另外的分类规则都基于样板的特征, 叙述如下。

规则 2: 若两模糊样板宽度差值大于阈值 ω_t 时, 两模糊样板不可能匹配。

规则 2 没有考虑高度是因为相同字符的样板具有相同的高度。

规则 3: 若两模糊样板矩心左区域或右区域的列数差值大于阈值 C_1 , 则两模糊样板不可能匹配。

规则 4: 若两模糊样板矩心以上区域或矩心以下区域的行数之差大于阈值 C_2 时, 则两模糊样板不可能匹配。

规则 5: 设 γ_i^λ 为模糊样板 λ 第 i 列的隶属值之和, 两模糊样板 λ_1 、 λ_2 可能匹配的条件是:

$$\zeta_t \sqrt{\sigma_{\lambda_1} \sigma_{\lambda_2}} \leq \sum (\gamma_i^{\lambda_1} \wedge \gamma_{i+a}^{\lambda_2})$$

而 $|\alpha - (x_{c_B} - x_{c_A})| \leq 1$, 式中: \wedge 为最小值符号。

在我们的系统中, 各样板库根据样板的基分类, 用规则 1, 2, 3, 4, 5 滤去不可能匹配的样板后得到可匹配样板的集合, 最后应用基于外形推测的相似测量获得在二维图形匹配中的可匹配样板。

3 实验结果

实验采用清华大学智能技术与系统实验室采集的脱机手写汉字样本 smp001 ~ smp020 等 20 套样本集, 未经归一化处理。各套样本分别由不同的人书写, 每套样本包含 3 755 个国标一级汉字, 共计 $3755 \times 20 = 75100$ 个手写汉字字符。样本集经预处理后, 奇数样本集用于训练, 偶数样本集用于测试。实验结果见表 1、表 2。

从实验数据看, 各手写汉字样本集的预分类率变化不大, 说明该系统比较稳定, 并取得了预分类率 98.22% 较好的分类效果, 但仍有 1.73% 左右的字符不能正确分类和 0.05% 的字符落入不确定字符类内。因此, 如何确定 SEART 网类别输出数及学习算法中控制匹配度的上、下限 λ_2 、 λ_1 及如何调整模糊样板匹配算法中的相关系数 ζ 将是我们今后要进一步研

究的课题。

表 1 测试结果

测试集	
字符样板数	37 550
正确分类字符的样板数	36 882
不确定字符样板数	19
分类错误字符样板数	649
平均预分类率	98.22%

表 2 各测试集的分类结果

样本集	预分类率 (%)	不确定字符率 (%)	错分字符率 (%)
002	98.45	0.05	1.50
004	98.63	0.05	1.32
006	96.95	0.09	2.96
008	98.81	0.05	1.14
010	98.24	0.10	1.66
012	97.74	0.06	2.20
014	98.19	0.06	1.75
016	97.97	0.00	2.03
018	98.57	0.01	1.42
020	98.65	0.03	1.32

4 结语

如何将 Neocognitron 网良好的特征提取功能和 ART 网的自适应、自组织聚类功能及如何将模糊技术用于手写汉字识别过程中的预分类阶段来缩小后阶段的分类、识别范围, 提高识别率是一个值得探讨、研究的问题。本文对这个问题进行了研究, 通过 Neocognitron 网提取笔画特征后分别用 SEART 网和模糊样板的匹配算法进行手写汉字的预分类。实验表明, 该方法取得较满意的分类效果。但在优化笔画特征、网络学习算法及模糊样板匹配算法中的相关参数的选取、调整等方面还有待进一步的研究和改进。

参考文献:

- [1] 杨波, 赵学军, 乔进, 等. 一种识别手写字符的多分类器集成方法[J]. 重庆大学学报, 2001, 24(3): 114~116.
- [2] 蔡自兴, 成浩. 一种基于骨架特征和神经网络的手写体字符识别技术[J]. 计算技术及自动化, 2001(13): 59~65.
- [3] 吴天雷, 马少平. 基于重叠动态网格和模糊隶属度的手写汉字特征抽取[J]. 电子学报, 2004, 32(2): 186~190.
- [4] 卢达, 钱忆平, 谢铭培, 等. 基于模糊模型相似测量的字符无监督分类法[J]. 计算机学报, 2002, 25(4): 423~429.
- [5] 卢达, 浦炜, 钱忆平, 等. 基于模糊模型相似测量的小类别数数字及数字识别[J]. 计算机工程与应用, 2000, 36(8): 78~80.
- [6] FUKUSHIMA K, WAKE N. Handwritten alphanumeric character recognition by the Neocognitron [J]. IEEE transactions on neural network, 1991: 355~365.
- [7] CARPENTER G-A, GRESSBERG S. The ART of adaptive pattern recognition by a self-organizing neural network[A]. IEEE computer, 1988, 21: 77~88.
- [8] LEE H-M, SHEN C-C. A handwritten Chinese characters recognition method based on primitive and fuzzy features via SEART neural net model[C]. IEEE Int. Conf. Syst. Man Cybern. 1995. 1939~1944.