一种适用于超多类手写汉字识别的新改型 Adaboost 算法

丁晓青，付　强
（清华大学电子工程系智能技术与系统国家重点实验室，北京 100084）

［摘要］ 提出一种适用于超多类手写汉字识别的新改型 Adaboost 算法，采用基于描述性模型的多类分类器（modified quadratic discriminant function，MQDF）作为 Adaboost 基元分类器，可直接进行多类分类，无需将多类问题转化为多个两类问题处理，其训练复杂度大大低于已有的多类 Adaboost 算法。算法提出根据广义置信度更新样本权重，实验证明这种算法适用于大规模多类分类问题。为了降低算法的识别复杂度，提出从所有训练后得到的 Adaboost 基元分类器组中选择一个最优的基元分类器作为最终分类器的方法进行删减。在 HCL2000 及 THOCR－HCD 数据集上进行实验证明，所提改型 Adaboost 算法提高了识别率的有效性，该算法的相对错误率比现有最优算法分别下降了 14.3 %，8.1 % 和 19.5 %。

［关键词］ 多类 Adaboost 算法；手写汉字识别；广义置信度；改进的二次鉴别函数
［中图分类号］ TP312 ［文献标识码］ A ［文章编号］ 1009－1742(2009)10－0019－06

1 前言

Boosting 是提高分类器识别性能的一般性框架，Adaboost 是 Boosting 算法的典型代表，并在近年来得到越来越多的关注。它在不同的样本分布下，根据相同的学习算法构建多个“弱学习器”，并采用加法模型将这些“弱学习器”集成为“强学习器”。在每一轮的样本权重更新中，都是提高识别错误样本的权重，降低识别正确样本的权重，这使得后续基元分类器能够更加关注那些被前级分类器错分的样本。Adaboost 原本用于处理两类问题，之后又出现了很多用于处理多类问题的扩展算法，如 Adaboost. M1，Adaboost. M2，Adaboost. MH，Adaboost. OC 以及 Adaboost. ECC 等[1～5]。虽然这些算法已被应用于数字识别等小类别数多类分类问题，但对于汉字识别这种大类别分类问题，大部分多类 Adaboost 算法由于其过高的训练复杂度而难以被应用。笔者提出了一种新的改型多类 Adaboost 算法，它具有较低的训练复杂度，可以有效提高手写汉字的识别率。

总的来说，Adaboost 算法中有两个关键要素。a. 样本权重的更新算法；b. 基元分类器。样本权重更新算法可分为两大类：一类是根据基元分类器的识别率进行权重更新，如 Discrete Adaboost；另一类是根据样本的识别可信度进行权重更新，如 Real Adaboost。一般来说，第二类方法比第一类方法更加精细，因此具有更好的性能。但是，第二类算法中需要估计识别结果的后验概率，在很多实际应用中，识别结果后验概率的估计往往比较复杂，也比较困难。

多类 Adaboost 算法中使用的基元分类器可以分为两类。a. 采用多类分类器进行直接的多类分类，如 Adaboost. M1；b. 将多类问题转化为多个两类问题，进而应用多个两类分类器实现多类分类。现有大部分多类 Adaboost 算法均属于第二类。目前

［收稿日期］ 2009－08－24
［基金项目］ 国家自然科学基金资助项目（60472002）；国家“八六三”高科技研究发展计划（2006AA01Z115）
［作者简介］ 丁晓青（1939－），女，江苏睢宁县人，清华大学教授，博士生导师，研究方向为图像处理、模式识别、计算机视觉；
E－mail：dingxq@tsinghua.edu.cn

主要有 3 种策略实现将多类问题转化为多个两类问题,即一类对一类策略(1v1);一类对余类策略(1vL);编码策略(Output Code)。Adaboost. M2 和 Adaboost. MH 采用一类对余类策略;Adaboost. OC 和 Adaboost. ECC 采用编码策略。记训练样本总数为 N,类别总数为 C。设训练复杂度与训练中所使用的样本总样次成正比,则一类对一类及一类对余类策略的训练复杂度约为 $O(N\times C)$;编码策略的训练复杂度至少为 $O(N\times\log(C))$。对于汉字识别来说,$C=3\ 755$,因此一类对一类以及一类对余类策略都由于过高的训练复杂度而难以被应用。虽然编码策略的训练复杂度比其他两种策略低很多,但是在大类别分类问题中,如何构造合理有效的码表仍未被很好的解决。总之,将多类分类转化为多个两类分类来处理汉字识别仍存在很多问题。因此,笔者算法中采用多类分类器作为基元分类器。

文章第一节对此算法进行概述,比较了所提算法与现有算法的不同;第二节介绍算法的具体流程;第三节介绍基元分类器组的删减;第四节和第五节分别为实验和结论。

2 改型 Adaboost 算法概述

基于描述性模型的多类分类器(modified quadratic discriminant function,MQDF)分类器由于其优良的识别率、简单的训练流程、较低的训练复杂度以及稳定的表现被广泛应用于汉字识别领域[6]。文章算法中采用 MQDF 作为基元分类器,并根据广义置信度进行权重更新。它与传统多类 Adaboost 算法相比主要有如下特点:

2.1 采用的基元分类器类型不同

1)大部分现有 Adaboost 算法都采用诸如线性分类器、SVM 等基于鉴别性模型的分类器作为基元分类器;在处理多类问题时大都是将多类分类转化为多个两类分类去处理,训练复杂度较高。

2)文章算法采用基于描述性模型的多类分类器 MQDF 作为基元分类器。它可以直接进行多类分类,无需将多类问题转化为多个两类问题处理,具有较低的训练复杂度,这使得它可以被用于汉字识别这种大类别分类问题。

2.2 样本权重更新算法不同

1)Adaboost. M1 是一种采用多类分类器作为基元分类器的典型算法。与 Discrete Adaboost 相同,它根据基元分类器的识别率进行样本权重更新,忽略了不同样本识别结果可信度的差异。但 MQDF 与传统 Adaboost 算法中的"弱学习器"不同,它的分类性能比较强:对于任意书写汉字的识别率已经接近 90 %,对于较规整书写汉字的识别率甚至达到 95 % 以上。通过实验发现,根据基元分类器识别率更新样本权重的做法不适合 MQDF 这种分类性能很好的基元分类器。

2)文章借鉴两类分类中 Real Adaboost 算法的思想,根据各样本识别结果的可信度对其进行权重更新。但是 Real Adaboost 需要估计样本识别结果的后验概率,这通常比较困难,计算也比较复杂。笔者引进一种可简便计算的广义置信度作为评价识别结果优劣的度量,并根据广义置信度对样本进行权重更新。实验证明这种方式简便有效。

当使用笔者的改型 Adaboost 算法训练得到一组基元分类器后,为了提高识别速度以满足实际应用的需要,从基元分类器组中挑选一个最优的基元分类器作为最终的分类器。

3 改型多类 Adaboost 算法流程

由于文章算法与 Real Adaboost 具有紧密联系,因此先给出 Real Adaboost 的流程。然后将它推广到多类,并对权重更新系数的计算方法做一定改型。记训练样本集合为 $\{(x_i,y_i)\mid 1\leqslant i\leqslant N\}$,$\boldsymbol{x}_i$ 为特征向量,y_i 为其所属类别标号;记 T 为迭代总轮数。

3.1 Real Adaboost 算法流程

在 Real Adaboost 算法中,$y_i\in\{-1,+1\}$,它的流程如下:

1)初始化:$t=1$;对样本赋初始权重 $w_t^i=\dfrac{1}{N}$。

2)对于 $t=1,2,\cdots,T$,在样本分布 w_t 下估计如下后验概率为:

$$p_t(y=1\mid x)\in[0,1] \tag{1}$$

根据式(1),得到本轮的基元分类器 $h_t(x)$

$$h_t(x)=\frac{1}{2}\log\frac{p_t(y=1\mid x)}{1-p_t(y=1\mid x)} \tag{2}$$

更新样本权重:

$$w_{t+1}^i=\frac{w_t^i\exp(-y_ih_t(x_i))}{Z_t} \tag{3}$$

式(4)中,Z_t 是归一化因子,它使得 w_{t+1} 是一个概率分布,即满足 $\sum_{i=1}^{N}w_{t+1}^i=1$。

3)最终的分类器为 $H(x)$。

$$H(x) = \text{sign}[\sum_{t=1}^{T} h_t(x)] \quad (4)$$

在 Real Adaboost 中，$h_t(x)$ 含有两层含义，其符号表示了识别结果的类别；其绝对值表示了识别结果的可信度。

3.2 改型 Adaboost 算法流程

在多类情况下，将第 t 轮的基元分类器记为 (h_t, s_t)。$h_t(x)$ 表示样本 x 的识别结果，$h_t(x) \in \{1,2,\ldots,C\}$；$s_t(x)$ 表示其广义置信度。

1）初始化：$t=1$；对样本赋初始权重 $w_t^i = \frac{1}{N}$。

2）对于 $t=1,2,\cdots,T$，在样本分布 w_t 下训练得到第 t 级基元分类器（h_t，s_t）。更新样本权重：

$$w_{t+1}^i = \frac{w_t^i}{Z_t} \times \begin{cases} \exp[-s_t(x_i)] & \text{若 } h_t(x_i) = y_i \\ \exp[s_t(x_i)] & \text{其他情况} \end{cases} \quad (5)$$

式(5)中，Z_t 是归一化因子，它使得 w_{t+1} 是一个概率分布，即满足 $\sum_{i=1}^{N} w_{t+1}^i = 1$。

3）对于待测样本 x，将累计广义置信度最大的类别作为其识别结果，如式(6)所示：

$$H(x) = \underset{1 \le r \le C}{\text{argmax}} \sum_{t=1}^{T} s_t(x)\delta(h_t(x) = r) \quad (6)$$

$$\delta(h_t(x) = r) = \begin{cases} 1 & \text{若 } h_t(x) = r \\ 0 & \text{其他情况} \end{cases} \quad (7)$$

3.3 基元分类器训练流程

文章使用的原始特征是梯度方向直方图特征[7]。原始特征维数是 392 维，过高的特征维数往往会影响分类器的泛化能力；此外原始特征中含有大量与鉴别无关的信息，这些无用信息会对识别产生干扰。为了去除原始特征中那些与鉴别无关的信息，提高分类器的识别率及泛化能力，需要采用特征压缩方法从原始特征中提取鉴别特征，降低特征维数。

最常用的特征压缩方法是线性鉴别分析方法（linear discriminant analysis，LDA）。但 LDA 假设各类样本具有相同的协方差矩阵，这对于汉字识别并不合适。尤其是在文章算法中，各类样本分布在不断变化，这一假设更加无法保证。文献[8]提出了适用于异方差情况下的特征压缩算法，即改进的异方差线性鉴别分析方法（modified heteroscedastic linear discriminant analysis，M－HLDA）。在笔者等实验中，LDA 和 M－HLDA 均被使用。

通过特征压缩方法，可以将原始特征向量进行维数压缩，成为特征向量。然后在特征向量集合上训练 MQDF 分类器[9]。对于特征向量 $\boldsymbol{x}$，可根据式(8)计算得到 MQDF 分类器输出的识别距离。其中，$g_i(x)$ 表示 x 到 ω_i 类的识别距离，$\boldsymbol{m}_i$ 是类 ω_i 的均值向量，λ_{ij} 和 φ_{ij} 分别为类 ω_i 协方差矩阵 $\sum_i$ 的第 j 个特征值（按照从大到小排序）和对应的特征向量，d 是特征维数，q 是截断维数，σ 是一个常数。各类均值向量及协方差矩阵可由最大似然估计得到。

$$g_i(x) = \frac{1}{\sigma^2}\{\left\|x - m_i\right\|^2 - \sum_{j=1}^{q}(1 - \frac{\sigma^2}{\lambda_{ij}})[\varphi_{ij}^T(x - m_i)]^2\} + \sum_{j=1}^{q}\log\lambda_{ij} + (d-q)\log\sigma^2 \quad (8)$$

3.4 广义置信度计算方法

根据 Real Adaboost 算法，在多类情况下样本的识别可信度可由式(9)进行计算。

$$s_t(x) = \frac{1}{2}\log\frac{p_t(h_t(x) \mid x)}{1 - p_t(h_t(x) \mid x)} \quad (9)$$

式(9)中，$h_t(x)$ 是基元分类器的识别结果；$p_t(h_t(x) \mid x)$ 是估计所得的识别结果的后验概率。在多类情况下，$p_t(h_t(x) \mid x)$ 很有可能会小于 0.5，这会使得式(9)计算得到的权重更新系数为负数。根据权重更新算法，对于那些权重更新系数为负数的样本来说，如果他们被识别正确，则会增加其权重；如果识别错误，则会降低其权重。这显然与 Adaboost 的基本出发点相矛盾。为了避免这种矛盾，需要保证权重更新系数 $s_t(x)$ 为非负值，但式(9)并不能保证这一点。此外，式(9)中后验概率的估计也比较困难。因此，利用式(9)来计算权重更新系数 $s_t(x)$ 在多类情况下存在一定的问题。

笔者引入一种广义置信度作为权重更新系数 $s_t(x)$，它既能保证 $s_t(x)$ 为非负值，又无需估计后验概率，实验证明此方法简单有效。广义置信度的定义如下：若存在一个函数 $e(h(x) \mid x)$，以及一个单调递增函数 $f(\cdot)$，满足式(10)，则称 $e(h(x) \mid x)$ 为 x 的广义置信度。笔者等使用的广义置信度计算方法如式(11)所示[10]，其中 $g_k(x)$ 表示 x 的第 k 候选识别结果的识别距离。

$$e(h(x) \mid x) = f(p(h(x) \mid x)) \quad (10)$$

$$s(x) = e(h(x) \mid x) = 1 - \frac{g_1(x)}{g_2(x)} \quad (11)$$

4 分类器组的删减

通过改型 Adaboost 算法,得到 T 个基元分类器,每个基元分类器均包含一个特征降维 LDA 或 M-HLDA 矩阵,以及一个 MQDF 分类器。基元分类器与目前汉字识别中广泛使用的 LDA + MQDF 方法具有相同的识别复杂度。因此,改型 Adaboost 算法的整体识别复杂度是传统 LDA + MQDF 的 T 倍。为了保持与传统算法相同的识别复杂度,对训练得到的分类器组进行删减。根据 T 个基元分类器在训练集上的识别性能,挑选一个最优的基元分类器作为最终的分类器。在识别时仅使用最优基元分类器进行识别。记第 t 个基元分类器在训练集上的识别率为 CR_t。最优基元分类器所对应的序号 $\hat{t}$ 根据式(12)得到。

$$\hat{t} = \underset{1\leqslant t\leqslant T}{\arg\max}\{t \mid (CR_t - CR_{t-1} \geqslant Th) \&(CR_t > CR_i, \text{若 } t > i)\} \tag{12}$$

5 实验

5.1 手写汉字样本库

文章使用两个手写汉字样本库,HCL2000 和 THOCR-HCD。HCL2000 样本库由北京邮电大学在国家"八六三"计划的资助下收集,该样本库收集了由不同年龄、职业和文化程度的人书写的 1 600 套汉字样本,每套样本中均包含 GB2312-1980 中的3 755 个一级汉字的字符图像。笔者获得了该数据库的前 1 000 套样本,将其中的 700 套样本(标号为 xx001 ~ xx700)作为训练集,而另外的 300 套样本(标号为 hh001 ~ hh300)作为测试集使用。HCL2000 的示例样本图像如图 1 所示。THOCR-HCD 样本库由清华大学电子工程系智能图文信息处理研究室历年来收集的手写汉字样本组成。同 HCL2000 库一样,THOCR-HCD 库也只包含 GB2312-1980 中的 3 755 个一级汉字样本,但无论是样本规模还是样本采集的时间跨度都更大,包含的样本风格也更多。THOCR-HCD 样本库分为 10 个子集,按照其质量标为好、中和差三级(见表 1)。在试验中,HCD4 和 HCD9 作为两个测试集,其他1 870套样本作为训练集。HCD4 为测试集 1,含有 100 套书写质量一般的样本,HCD9 为测试集 2,含有 20 套书写非常潦草的样本,示例样本图像如图 2所示。

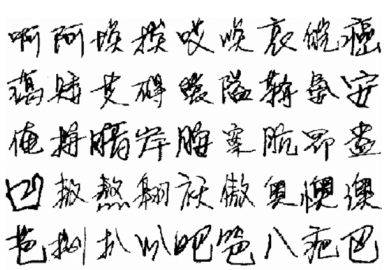

图 1 HCL2000 库样本图像示例

Fig. 1 Samples of HCL2000 database

表 1 THOCR-HCD 库子集信息

Table 1 THOCR-HCD database subset information

THOCR-HCD 子集	样本数量	样本质量
HCD-1	100 × 3 755	好
HCD-2	500 × 3 755	好
HCD-3	107 × 3 755	中
HCD-4	100 × 3 755	中
HCD-5	300 × 3 755	中
HCD-6	300 × 3 755	中
HCD-7	300 × 3 755	中
HCD-8	100 × 3 755	差
HCD-9	20 × 3 755	差
HCD-10	172 × 3 755	差

图 2 THOCR-HCD 库样本图像示例

Fig. 2 Samples of THOCR-HCD database

5.2 算法识别率实验

使用 LDA + MQDF 作为基元分类器,用改型 Adaboost 算法在 THOCR-HCD 库上进行 40 轮迭代训练,所得的 40 个基元分类器在训练集和两个测试集上的识别率分别如图 3、图 4 和图 5 所示;在 HCL2000 库上进行 40 轮迭代,所得 40 个基元分类器在训练集和测试集上的识别率分别如图 6 和图 7 所示。在每幅图中,均有两条识别率曲线,分别表示

了前 T 个基元分类器进行集成后的整体分类器的识别率和各单独基元分类器的识别率。

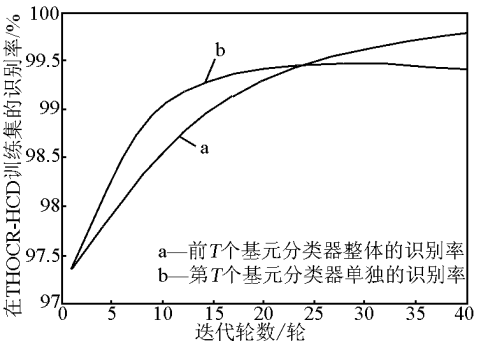

图 3　分类器在 THOCR – HCD 训练集的识别率

Fig. 3　The recognition rate on THOCR – HCD training set

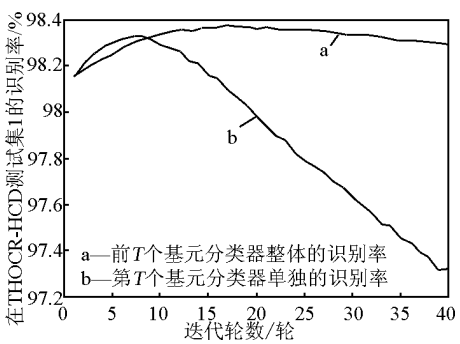

图 4　分类器在 THOCR – HCD 测试集 1 的识别率

Fig. 4　The recognition rate on THOCR – HCD testing set 1

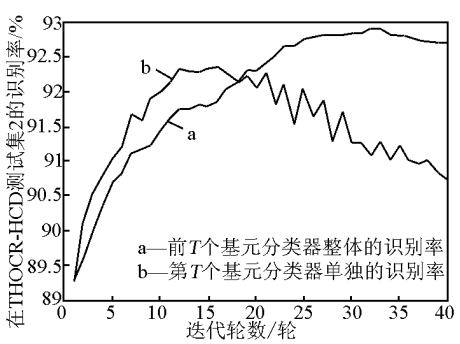

图 5　分类器在 THOCR – HCD 测试集 2 的识别率

Fig. 5　The recognition rate on THOCR – HCD testing set 2

以上各实验测得的识别率曲线具有如下共同点。**a**. 在训练集上,前 T 轮整体分类器的识别率随迭代轮数的增加而持续增加,而各轮单独基元分类器的识别率则是先增加后降低。b. 在测试集上,前 T 轮整体分类器和各轮单独基元分类器的识别率随着 T 的增加均呈现先增加后降低的趋势;前 T 轮整体分类器的识别率经过一个峰值后缓慢下降,而各轮单独基元分类器的识别率经过峰值后下降迅速。c. 无论是在训练集还是测试集上,前 T 轮整体分类器识别率的峰值均大于单独基元分类器识别率的峰值,这两个峰值又均明显高于未经 Adaboost 进行性能提升的分类器的识别率,即第 1 轮基元分类器的识别率。

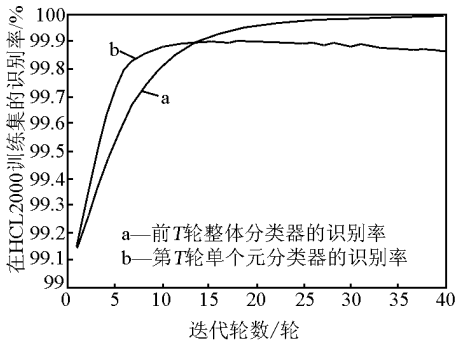

图 6　分类器在 HCL2000 训练集的识别率

Fig. 6　The recognition rate on HCL2000 training set

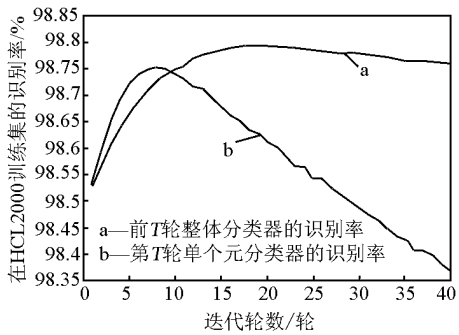

图 7　分类器在 HCL2000 测试集的识别率

Fig. 7　The recognition rate on HCL2000 testing set

随后,根据式(12)挑选一个最优的基元分类器作为最终的分类器。以 THOCR – HCD 为例,由图 3 中测得的各单独基元分类器的识别率曲线,最终选择第 9 轮基元分类器为最终的分类器。此外,还采用了 M – HLDA + MQDF 作为基元分类器进行了实验。表 2 列出了不同识别算法的识别率,其中,Boosted LDA + MQDF 和 Boosted M – HLDA + MQDF 都是仅包含一个最优基元分类器,因此与传统 LDA

+MQDF 方法具有相同的识别复杂度。实验结果验证了文章算法的有效性,其中 Boosted M-HLDA+MQDF 具有最优的识别结果;与现有的 M-HLDA+MQDF 方法相比,它的相对错误率分别降低了 14.3 %,8.1 %和 19.5 %。

表 2 各算法识别率比较

Table 2 The algorithms' recognition rate comparison

测试集	算法的识别率/%			
	LDA+MQDF	Boosted LDA+MQDF	M-HLDA+MQDF	Boosted M-HLDA+MQDF
HCL2000 测试集	98.53	98.75	98.67	98.86
THOCR-HCD 测试集 1	98.14	98.29	98.28	98.42
THOCR-HCD 测试集 2	89.29	92.01	91.40	93.08

6 结语

1)采用基于描述性模型的多类分类器作为基元分类器,因此可直接进行多类分类,无需将多类问题转化为多个两类问题,大大降低了训练复杂度,可用于大类别分类问题。

2)引入式(11)所示的广义置信度作为评价识别结果优劣的度量,并根据它进行权重更新。实验说明这种处理方式简便有效。

3)首次将 Adaboost 算法应用到超多类手写汉字识别中来,对其他大类别分类问题有借鉴作用。

4)在不同手写汉字数据库上进行实验,在保持与传统分类算法具有相同识别复杂度的前提下,显著提高了识别率。实验结果显示笔者等算法对于书写质量差的样本,如 THOCR-HCD 测试集 2,效果尤其显著。

参考文献

[1] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting[J]. The Annals of Statistics, 2000, 28(2):337-407

[2] Schapire R E, Singer Y. Improved boosting algorithms using confidence-rated predictions [J]. Machine Learning, 1999, 37 (3): 297-336

[3] Freund Y, Schapire R E. A decision-theoretic generalization of on-line learning and an application to boosting[J]. Journal of Computer and System Sciences, 2006, 55(1): 119-139

[4] Guruswami V, Sahai A. Multiclass learning, boosting, and error-correcting codes [A]. Proceedings of the twelfth Annual Conference on Computational Learning Theory [C]. Santa Cruz, USA: ACM, 1999:145-155

[5] Schapire R. Using output codes to boost multiclass learning problems[A]. Proceedings of the fourteenth International Conference on Machine Learning[C]. Nashville, USA: Morgan Kaufmann Publishers Inc, 1997:313-321

[6] Liu C L, Fujisawa H. Classification and learning for character recognition: comparison of methods and remaining problems[A]. Proceedings of the First IAPR TC3 Workshop on Neural Networks and Learning in Document Analysis and Recognition[C]. Seoul, Korea: IEEE, 2005:1-7

[7] Liu H L, Ding X Q. Handwritten character recognition using gradient feature and quadratic classifier with multiple discrimination schemes[A]. Proceedings of the Eighth International Conference on Document Analysis and Recognition [C]. Seoul, Korea: IEEE, 2005:19-25

[8] Liu H L, Ding X Q. Improve handwritten character recognition performance by heteroscedastic linear discriminant analysis[A]. The Eighteenth International Conference on Pattern Recognition [C]. Hongkong, China, 2006:880-883

[9] Kimura F, Takashina K, Tsuruoka S, etc. Modified quadratic discriminant functions and its application to Chinese character recognition [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 1987, 9(1): 149-153

[10] Lin X F, Ding X Q, Chen M, etc. Adaptive confidence transform based classifier combination for Chinese character recognition [J]. Pattern Recognition Letters, 1998, 19 (10):975-988

(下转 31 页)